



University of Pennsylvania  
**ScholarlyCommons**

---

Publicly Accessible Penn Dissertations


---

2019

## Systems Biology Of Gene Regulation Across Scales: From Single Molecules To Cellular Identities

Ian Alexander Mellis  
*University of Pennsylvania*

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Genetics Commons](#), and the [Molecular Biology Commons](#)

---

### Recommended Citation

Mellis, Ian Alexander, "Systems Biology Of Gene Regulation Across Scales: From Single Molecules To Cellular Identities" (2019). *Publicly Accessible Penn Dissertations*. 3594.  
<https://repository.upenn.edu/edissertations/3594>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/3594>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# Systems Biology Of Gene Regulation Across Scales: From Single Molecules To Cellular Identities

## Abstract

Gene regulation takes many forms and is responsible for phenotypes at the scale of individual molecules up through the scale of complex tissue functions. At the smallest level, single-base modifications of individual mRNA molecules transcribed from the same gene can lead to functionally different protein products. In the first chapter of this thesis, I develop a new method, inoFISH, and associated analytical tools to visualize and quantify RNA editing with single molecule resolution in single mammalian cells. Using this new method in conjunction with mathematical modeling I show that the heterogeneity of single-cell mRNA editing rates across a population depends on the gene of interest. Further, I characterize subcellular localization patterns of edited and unedited mRNAs. At the other end of the spectrum, the regulation of transcriptome-wide patterns of gene expression can underpin cellular identities. In the second chapter of this thesis I develop a new experimental design and analytical framework for prioritizing lists of transcription factors that can be used for directed changes of cellular identity. With Perturbation Panel Profiling (P3), I show that cardiomyocyte lineage-driving transcription factors are more frequently up-regulated, or “perturbable”, than other highly expressed transcription factor genes. I subsequently demonstrate that a known cocktail of cardiomyocyte-perturbable transcription factors enables cardiac transdifferentiation of several types of human fibroblasts. Lastly I extend perturbability-based selection of transcription factors to another biological context, i.e., fibroblast reprogramming to pluripotency. I show that fibroblast-perturbable factor knockdown often enables more efficient fibroblast reprogramming. Together, my thesis makes critical steps toward understanding and engineering gene regulation through the development of a diverse array of methods, experimental designs, and analytical frameworks.

## Degree Type

Dissertation

## Degree Name

Doctor of Philosophy (PhD)

## Graduate Group

Genomics & Computational Biology

## First Advisor

Arjun . Raj

## Keywords

gene expression, reprogramming, RNA editing, systems biology, transcriptomics, transdifferentiation

## Subject Categories

Genetics | Molecular Biology

SYSTEMS BIOLOGY OF GENE REGULATION ACROSS SCALES:  
FROM SINGLE MOLECULES TO CELLULAR IDENTITIES

Ian A. Mellis

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

---

Arjun Raj, Ph.D.  
Professor of Bioengineering

Graduate Group Chairperson

---

Benjamin F. Voight, Ph.D.  
Associate Professor of Genetics and Systems Pharmacology and Translational Therapeutics

Dissertation Committee:

Junhyong Kim, Ph.D., Patricia M. Williams Term Professor and Chair of Biology; Adjunct  
Professor of Computer and Information Science

Marisa S. Bartolomei, Ph.D., Perelman Professor of Cell and Developmental Biology

Christopher D. Brown, Ph.D., Associate Professor of Genetics

Tuuli Lappalainen, Ph.D., Assistant Professor of Systems Biology, Columbia University;  
Junior Investigator and Core Member, New York Genome Center

SYSTEMS BIOLOGY OF GENE REGULATION ACROSS SCALES:  
FROM SINGLE MOLECULES TO CELLULAR IDENTITIES

COPYRIGHT

2019

Ian Alexander Mellis

This work is licensed under the  
Creative Commons Attribution-  
NonCommercial-ShareAlike 4.0  
License

To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>

*In memory of my grandparents*

# ACKNOWLEDGMENTS

To my family and friends: your support and love have made this possible. I am forever grateful for you.

Mom and Dad, thank you for everything. You have provided me with countless opportunities, and you are amazing role models. Eric, thank you for being such a supportive brother. There have been many challenging times over the last few years, which I couldn't have overcome without you. Emily, you are the most amazing partner in life I can imagine. Thank you for putting up with my long hours in lab and rants about science, and for looking out for me when I forget how to do that myself.

To the members of the Raj lab: you are some of the kindest, smartest, and most thoughtful people I have had the privilege of working with. Thank you for everything. Sara, Uschi, Ally, Eduardo, Ben, Yogesh, and everyone else in lab, you have all taught me so much and I'm so happy to call you my friends.

To my mentors over the years: thank you for taking a chance on me, for teaching me so much about how to do science, and for enabling me to get to this point. Walter Muller, Aris Economides, David Friendewey, Wen Fury, Ron DePinho, David Ratner, Junhyong Kim, Marisa Bartolomei, Casey Brown, Tuuli Lappalainen, and so many others, you have provided invaluable guidance, and I am so grateful for it.

And to Arjun. You are an amazing mentor and a brilliant scientist. I have learned so much over these last few years, and my life in science wouldn't be the same without your guidance.

# ABSTRACT

## SYSTEMS BIOLOGY OF GENE REGULATION ACROSS SCALES: FROM SINGLE MOLECULES TO CELLULAR IDENTITIES

Ian A. Mellis

Arjun Raj, Ph.D.

Gene regulation takes many forms and is responsible for phenotypes at the scale of individual molecules up through the scale of complex tissue functions. At the smallest level, single-base modifications of individual mRNA molecules transcribed from the same gene can lead to functionally different protein products. In the first chapter of this thesis, I develop a new method, inoFISH, and associated analytical tools to visualize and quantify RNA editing with single molecule resolution in single mammalian cells. Using this new method in conjunction with mathematical modeling I show that the heterogeneity of single-cell mRNA editing rates across a population depends on the gene of interest. Further, I characterize subcellular localization patterns of edited and unedited mRNAs. At the other end of the spectrum, the regulation of transcriptome-wide patterns of gene expression can underpin cellular identities. In the second chapter of this thesis I develop a new experimental design and analytical framework for prioritizing lists of transcription factors that can be used for directed changes of cellular identity. With Perturbation Panel Profiling (P<sup>3</sup>), I show that cardiomyocyte lineage-driving transcription factors are more frequently up-regulated, or “perturbable”, than other highly expressed transcription factor genes. I subsequently demonstrate that a known cocktail of cardiomyocyte-perturbable transcription factors enables cardiac transdifferentiation of several types of human fibroblasts. Lastly I extend perturbability-based selection of transcription factors to another biological context, i.e., fibroblast reprogramming to pluripotency. I show that fibroblast-perturbable factor knockdown often enables more efficient fibroblast reprogramming. Together, my thesis makes critical steps toward understanding and engineering gene regulation through the development of a diverse array of methods, experimental designs, and analytical frameworks.

# TABLE OF CONTENTS

<b>ACKNOWLEDGMENTS.....</b>	<b>IV</b>
<b>ABSTRACT .....</b>	<b>V</b>
<b>TABLE OF CONTENTS.....</b>	<b>VI</b>
<b>LIST OF TABLES.....</b>	<b>IX</b>
<b>LIST OF ILLUSTRATIONS.....</b>	<b>X</b>
<b>INTRODUCTION.....</b>	<b>1</b>
One base can change everything: RNA editing is important but hard to study.....	2
The whole transcriptome matters: the state of identifying lineage-driving transcription factors.....	4
<b>CHAPTER 1: VISUALIZING ADENOSINE-TO-INOSINE RNA EDITING IN SINGLE MAMMALIAN CELLS .....</b>	<b>8</b>
Adenosine deamination RNA editing changes base pairing .....	8
Competitive hybridization as a strategy for detecting RNA editing <i>in situ</i> .....	10
Validating inoFISH of GRIA2 in human neural lineage cell lines .....	12
Comparing editing rate estimates of inoFISH and other methods.....	17
Identification of candidate inoFISH-compatible RNA editing sites with RNA-seq .....	21
Subcellular localization of edited and unedited transcripts in neural lineage cell lines .....	24
Association of edited and unedited GRIA2 transcripts with nuclear paraspeckles .....	31
Visualization of unedited NUP43 transcripts at the site of transcription.....	33
GRIA2 and NUP43 RNA editing levels across a population .....	33



<b>CHAPTER 2: PERTURBATION PANEL PROFILING (P<sup>3</sup>) IDENTIFIES TRANSCRIPTION FACTORS THAT ENHANCE DIRECTED CHANGES OF CELLULAR IDENTITY .....</b>	<b>37</b>
Most current protocols for directed changes of cell identity.....	37
Parallel small-molecule perturbations of fibroblasts and cardiomyocytes.....	39
A potency-based drug dosing scheme for perturbation panel profiling (P <sup>3</sup> ).....	42
Most P <sup>3</sup> perturbations induce differential expression profiles while gross cellular phenotypes remain stable.....	47
Most highly expressed genes are differentially expressed after at least one perturbation .....	54
Cardiomyocyte lineage-driving transcription factors are up-regulated in more perturbation conditions than other highly expressed transcription factors .....	56
Overexpression of a known cocktail of transcription factors perturbable in cardiomyocytes enables cardiac transdifferentiation of fibroblasts .....	59
Knockdown of transcription factors that are perturbable in fibroblasts often enhances fibroblast reprogramming to iPSC.....	65
<b>CONCLUSIONS AND FUTURE DIRECTIONS .....</b>	<b>70</b>
<b>APPENDIX A: STATISTICAL ANALYSIS OF RNA EDITING LEVELS BASED ON INOFISH EXPERIMENTAL RESULTS .....</b>	<b>75</b>
<b>1 Detecting the presence of any RNA editing .....</b>	<b>75</b>
<b>2 Modeling RNA editing levels in a single experiment.....</b>	<b>75</b>
2.1 Goal: estimating editing level .....	76
2.2 Modeling inoFISH detection efficiency .....	76
2.3 Modeling inoFISH experiments .....	79
2.4 Population-wide parameter estimation.....	80
2.5 Performing simulated detection experiments .....	85
2.6 Parametric bootstrapping a confidence interval for $\hat{E}$ .....	86
<b>3 Modeling differences between RNA editing levels .....</b>	<b>86</b>
3.1 Bootstrapping a confidence interval for the difference between editing levels.....	87
<b>4 Incorporating multiple inoFISH experimental replicates .....</b>	<b>87</b>
4.1 Alternative summary statistics on experimental replicate results .....	88

4.2 Detecting differences in summary statistics with multiple replicates in two conditions.....88

**5 Single-Cell Editing Level Analysis.....89**

**APPENDIX B: MATERIALS AND METHODS FOR CHAPTER 1 ..... 91**

**APPENDIX C: MATERIALS AND METHODS FOR CHAPTER 2..... 99**

**REFERENCES ..... 109**

# LIST OF TABLES

## Chapter 1

- Table 1.1: Sequences of oligonucleotides used in Chapter 1.

## Chapter 2

- Table 2.1: Drugs used for perturbations in Chapter 2.
- Table 2.2: Sequences of oligonucleotides used in Chapter 2.
- Table 2.3: Sequences of shRNAs used for knockdowns.

# LIST OF ILLUSTRATIONS

## Chapter 1

- Figure 1.1: Chemical Structures of Adenosine and Inosine
- Figure 1.2: Overview of the inoFISH probe design strategy
- Figure 1.3: Sanger sequencing of GRIA2 RT-PCR product
- Figure 1.4: inoFISH discriminates between adenosine and inosine
- Figure 1.5: Expanded controls for GRIA2 inoFISH
- Figure 1.6: False colocalization analysis of GRIA2 detection probes with SFPQ probes
- Figure 1.7: Estimated mean editing levels of GRIA2 in SH-SY5Y cells
- Figure 1.8: Comparison of inoFISH with targeted and off-target detection probes
- Figure 1.9: GRIA2 editing levels estimated by inoFISH and traditional methods
- Figure 1.10: target selection and inoFISH probe design pipeline
- Figure 1.11: Sanger sequencing of EIF2AK2 and NUP43 RT-PCR products
- Figure 1.12: Subcellular localization analysis of editing targets with inoFISH
- Figure 1.13: Mean nuclear and cytoplasmic editing level estimates
- Figure 1.14: GRIA2 mRNA nuclear retention analysis
- Figure 1.15: Expanded controls for EIF2AK2 and NUP43 inoFISH
- Figure 1.16: Cyanoethylation reduces number of detected inosines
- Figure 1.17: GRIA2 inoFISH results with NEAT1 colocalization
- Figure 1.18: NUP43 inoFISH results with transcription site localization
- Figure 1.19: Single-cell analysis of inoFISH results

## Chapter 2

- Figure 2.1: P<sup>3</sup> perturbation panel profiling for developing transdifferentiation protocols
- Figure 2.2: Samples processed for GM00942 and iCard-942 P<sup>3</sup>

- Figure 2.3: Differentially expressed genes per drug for GM00942 fibroblasts - pilot
- Figure 2.4: Comparison of differentially expressed genes per drug and median drug IC50
- Figure 2.5: Differentially expressed genes per drug at 3 different IC50-based doses
- Figure 2.6: Comparison of differentially expressed genes per drug and drug target expression level
- Figure 2.7: Representative images from videos of iCard-942
- Figure 2.8: Heatmap with hierarchical clustering of P<sup>3</sup> expression profiles
- Figure 2.9: Differentially expressed genes per drug in P<sup>3</sup> samples
- Figure 2.10: Down-sampling analysis of iCard-942 RNAtag-seq data
- Figure 2.11: Cumulative fraction of genes differentially expressed in at least 1 condition
- Figure 2.12: Perturbability of transcription factor genes in iCard-942 P<sup>3</sup>
- Figure 2.13: Comparison of conditions in which up-regulated with average expression
- Figure 2.14: TNNT2 and NPPA smFISH results for iCard-942 and immHCF
- Figure 2.15: smFISH assessment of 7F-mediated transdifferentiation of GM00942
- Figure 2.16: smFISH scans results of 7F-mediated transdifferentiation of GM00942
- Figure 2.17: smFISH assessment of 7F-mediated transdifferentiation of GM11169 and immHCF
- Figure 2.18: Overlap in genes detected as differentially expressed in each P<sup>3</sup> drug condition across iCard-942 and GM00942
- Figure 2.19: Perturbability of transcription factor genes in GM00942 P<sup>3</sup>
- Figure 2.20: hiF-T iPSC reprogramming efficiency following knockdown of fibroblast perturbable transcription factors

# INTRODUCTION

In its myriad forms gene regulation generates diversity in the function of the genome at many scales of biological organization: from single-molecule-scale copies of transcripts in single cells to transcriptome-scale patterns of gene expression across tissues. Scientific questions about gene regulation at these different scales come with different accompanying challenges. In this thesis I identify and resolve some of these diverse challenges for two different questions about the function of gene regulation at two different levels of biological organization.

At the level of single molecules in single cells, RNA editing is a form of gene regulation that generates diversity in the base composition of transcripts arising from the same gene. These tiny, yet precise, differences are sometimes associated with life-or-death outcomes for single cells but are extremely challenging to study due to a lack of sufficiently sensitive experimental techniques. In the first chapter of this thesis I develop a new experimental technique and improve associated analytical tools to enable unprecedented single-molecule resolution of RNA editing in single cells *in situ*. I apply this technique to the study of several genes in neural lineage cell lines. These genes include GRIA2, which, when its RNA editing is disturbed, is implicated in the loss of neurons in models of neurodegenerative diseases.

At the level of the entire transcriptome across tissues, transcription factor-driven gene expression patterns can dictate cellular identity. A major goal in the field of cellular engineering is efficiently directing cells toward a cell identity of interest on demand, e.g., toward pluripotency in the case of fibroblast reprogramming to induced pluripotent stem cells. However, an outstanding question in the field is whether we can establish a general procedure for identifying those transcription factors capable of driving cells toward a cell identity of choice on demand. In the second chapter of this thesis, I develop a new experimental pipeline and analytical framework for the identification of

transcription factors that enhance directed changes of cellular identity when overexpressed or suppressed.

## ONE BASE CAN CHANGE EVERYTHING: RNA EDITING IS IMPORTANT BUT HARD TO STUDY

Of the more than one hundred known types of RNA editing resulting in chemical base modification, one of the most common events is adenosine deamination, which results in a base identity switch from adenosine to inosine (A-to-I).[\(Piskol et al., 2013; Sakurai et al., 2014\)](#) A-to-I editing, among other effects, changes the complementarity profile of the affected base from adenosine's affinity for thymidine to inosine's affinity for cytosine; inosine functions as a guanosine analog (Fig. 1a). Researchers have studied RNA editing, including A-to-I editing, in the context of structural stabilization and regulation of abundant non-coding RNAs (ncRNAs; e.g., tRNAs) for decades. However, in addition to affecting the secondary structure of an A-to-I edited RNA molecule, an inosine resulting from editing could also modulate codon identity, alternative splicing, or other post-transcriptional regulation of an mRNA.[\(Flomen et al., 2004; Sommer et al., 1991\)](#)

ADAR enzymes, which catalyze A-to-I editing, are important for normal mammalian physiology: ADAR mutants have well-characterized phenotypes.[\(Higuchi et al., 2000; Rice et al., 2012\)](#) Phenotypes resulting from loss of A-to-I editing are not solely due to defects in abundant ncRNAs. One of the best known examples of conserved ADAR-mediated A-to-I editing of a mammalian mRNA is that of GRIA2, whose editing results in a non-synonymous Q607R change in GluR2 subunit of the glutamate receptor, thereby modulating the receptor's Ca<sup>++</sup> permeability.[\(Sommer et al., 1991\)](#) Electrophysiological studies show excessive divalent cation

permeability of neurons in model organisms with perturbed editing of GRIA2. Phenotypically, GRIA2 editing-deficient model rat forebrain neurons are more susceptible to ischemic insult. ([Peng et al., 2006](#)) In humans with ALS, some disease-affected motor neurons appear to die as a result of glutamate excitotoxicity and have perturbed GRIA2 A-to-I editing rates. ([Hideyama et al., 2012](#)) Research on mouse models has provided evidence suggestive of a causal role of deficient GRIA2 editing in ALS-like pathology: ADAR knockout mice demonstrate an ALS-like phenotype, which is rescued upon exogenous expression of edited GRIA2. ([Hideyama et al., 2010](#)) In addition to GRIA2, serotonin receptor 2C (5-HT2C) is a well-known A-to-I editing target, with clear effects on receptor activity resulting from A-to-I editing. ([Flomen et al., 2004](#))

Since the 1990s, the research community has used RT-PCR-based techniques to study editing rates in these canonical example transcripts. ([Sommer et al., 1991](#)) However, little is known about how editing affects the localization of these mRNAs to specific compartments of the cell, which is important for understanding the factors that regulate their translation. For example, specific edited RNAs, such as CTN-RNA, appear to be sequestered in paraspeckles until the cell is exposed to stress. ([Prasanth et al., 2005](#)) Overall, multiple experiments have provided conflicting evidence about the subcellular localization of edited transcripts, including about GRIA2 with respect to nuclear retention, affecting how these genes could be post-transcriptionally regulated in stressful environments. ([Jepson et al., 2011](#); [Kumar and Carmichael, 1997](#); [Prasanth et al., 2005](#); [Savva et al., 2012](#); [Wong et al., 2003](#))

Additionally, recent transcriptome-wide surveys of A-to-I editing in mammals suggest that there may be many more mRNA targets than were previously known. ([Bahn et al., 2012](#); [Porath et al., 2014](#); [Ramaswami et al., 2012, 2013](#); [Sakurai et al., 2010, 2014](#)) These studies have put forward more than 20,000 candidate ADAR targets, most but not all in *Alu* repeat regions. Several hundred non-*Alu* candidate editing sites are conserved between orthologous loci in humans,



mice, chimpanzees, and rhesus macaques, suggesting that they may in fact be functional in mammals. [\(Ramaswami and Li, 2014\)](#) However, as above, little about the basic biology of ADAR-mediated mRNA A-to-I editing is conclusively known.

In combination, these results provide strong motivation for developing a new generation of precise techniques for studying RNA editing, and for using these tools to begin the quantification of mRNA editing rate perturbation in diseases such as ALS.

In chapter 1 of this thesis I apply a variant of the single-molecule FISH (smFISH) method to characterizing the subcellular localization of A-to-I edited transcripts and to identifying the distribution in editing rates in populations of neural lineage cells. [\(Levesque et al., 2013; Mellis et al., 2017; Raj et al., 2008\)](#)

## THE WHOLE TRANSCRIPTOME MATTERS: THE STATE OF IDENTIFYING LINEAGE-DRIVING TRANSCRIPTION FACTORS

Cellular transdifferentiation, i.e., transforming a cell of one type into another type, is a long-standing goal in the biomedical field. However, the biology of many terminally differentiated cell types (e.g., fibroblasts and cardiac myocytes) is such that they are phenotypically stable in the sense that they retain their cell type even when subjected to perturbation. A striking but commonplace example of this stability is growing cells on tissue culture dishes, in which cells are clearly not in their native context, but still retain recognizable features of being, say, a fibroblast or cardiac myocyte.

Decades of work have shown that one can, however, directly change cellular identity through manipulation of genetic material. In the 1950s, Briggs and King and Gurdon et al. demonstrated that frog blastula whole-nuclear transplantation could transform eggs into cells resembling zygotes, capable of complete development to sexually mature individuals. ([Briggs and King, 1952](#); [Gurdon et al., 1958](#)) Relatedly, in the 2000s, Sul et al. showed that transplantation of the whole transcriptome of a rat astrocyte into a rat neuron was sufficient to transform the neuron into an astrocyte-like cell. ([Sul et al., 2009](#))

The work of Davis et al. showed that one can, further, transform one cell type into another through ectopically inducing expression of a small set of genetic factors ([Davis et al., 1987](#)) known to regulate larger gene expression programs. The celebrated work of Yamanaka ([Takahashi and Yamanaka, 2006](#)) showed that one can similarly turn a terminally differentiated cell into a pluripotent stem cell. These genetic factors are often transcription factors, which appear to induce reprogramming through the regulation of large genetic programs akin to those in the target cell type. However, a major issue in the field is knowing which factors to test for their potential for reprogramming to a fate of interest. Given that many of them work best in a “cocktail” consisting of multiple factors, brute force genome-wide screening is not feasible, leaving us with the requirement to narrow the list of potential factors down before testing them experimentally.

Thus far in the field, there have been two broad sets of approaches to identifying putative reprogramming factors. The approach taken by Yamanaka and many subsequent studies has been to take a limited pool of factors known to be involved in the maintenance of pluripotent stem cells, and then to methodically test combinations of subsets of those factors to see which ones were most critical to inducing cells to reprogram. Such an approach is no doubt powerful and rooted in years of painstaking developmental biology, but a fundamental problem is that it is limited in scope to factors that have a known role in development. In principle, however, many forms of transdifferentiation have no real analogy in developmental biology, and as such, it is

possible that there are many other factors that could potentially induce transdifferentiation that are completely independent of those involved in the developmental trajectory. Identification of such unknown factors could have a huge impact on our ability to transdifferentiate cells.

Recent studies have tried to use molecular profiling of cells to reveal these factors (e.g., CellNet ([Cahan et al., 2014](#); [Morris et al., 2014](#); [Radley et al., 2017](#)), Mogrify ([Rackham et al., 2016](#)), and others ([D'Alessio et al., 2015](#); [Tomaru et al., 2014](#))). These studies use primarily transcriptome data (microarray, CAGE, or RNA sequencing) to identify transcription factors with predicted gene regulatory activity that is highly specific to a particular cell type. The underlying hypothesis in these studies is that these *lineage-specific* factors might also be *lineage-driving* in the sense that their expression can induce cells to transform into the type for which they are most specific. However, in practice, this strategy has met with fairly limited success. For instance, CellNet used thousands of profiles to try to identify factors for a B-cell to macrophage transdifferentiation protocol, but it didn't identify the most important known factor in the transformation: C/EBP $\alpha$  overexpression. Mogrify also was a limited success: it did find some previously known factors, but failed to identify others. In the years subsequent to publication, to my knowledge, neither has been used to prospectively identify new transdifferentiation protocols.

A central conceptual limitation of these approaches is that they only identify factors whose expression or activity is specifically associated with the cell type, irrespective of how that cell type is able to maintain its overall phenotypic stability. In principle, there may be sets of factors that are neither cell type-specific nor associated with a cell type's development that are nonetheless important for maintaining its cellular identity. How, then, might one identify these additional factors, whose role in cell identity maintenance suggests that they are promising candidates for inclusion in a transdifferentiation protocol?

In the second chapter of this thesis, I develop an experimental and analytical pipeline for identifying transcription factors for use in directed changes of cell identity through their association with cellular responses to a variety of perturbations. I subsequently test several candidate factors identified in studies of perturbed cardiac myocytes and dermal fibroblasts in cardiac transdifferentiation and iPSC reprogramming experiments.

# CHAPTER 1: VISUALIZING ADENOSINE-TO-INOSINE RNA EDITING IN SINGLE MAMMALIAN CELLS

Conversion of adenosine to inosine is a frequent type of RNA editing, but important details about its biology remain unknown due to a lack of imaging tools. We developed inoFISH to directly visualize and quantify adenosine-to-inosine edited transcripts *in situ*. We found that editing of *GRIA2*, *EIF2AK2*, and *NUP43* is uncorrelated with nuclear localization and paraspeckle association. Further, *NUP43* exhibits constant editing levels between single cells while *GRIA2* levels vary.

## ADENOSINE DEAMINATION RNA EDITING CHANGES BASE PAIRING

Many RNA species are modified to contain non-canonical bases, a process known as RNA editing. The most prevalent type is adenosine-to-inosine editing ([Bass and Weintraub, 1987](#)), wherein adenosine deaminases (e.g., ADARs) enzymatically modify an adenosine base to an inosine base (**Figure 1.1**), disruption of which leads to defects in hematopoiesis ([Liddicoat et al., 2015](#)) and neurological function ([Higuchi et al., 2000](#)). It has been speculated that adenosine-to-inosine RNA editing influences subcellular localization patterns like nuclear retention ([Chen and Carmichael, 2009](#); [Kumar and Carmichael, 1997](#); [Prasanth et al., 2005](#); [Zhang and Carmichael, 2001](#)), but the lack of visualization tools has left this and other hypotheses untested. We thus developed inosineFISH (inoFISH), a fluorescence *in situ* hybridization-based method for directly imaging adenosine-to-inosine RNA editing events with single-molecule resolution.

Figure 1.1

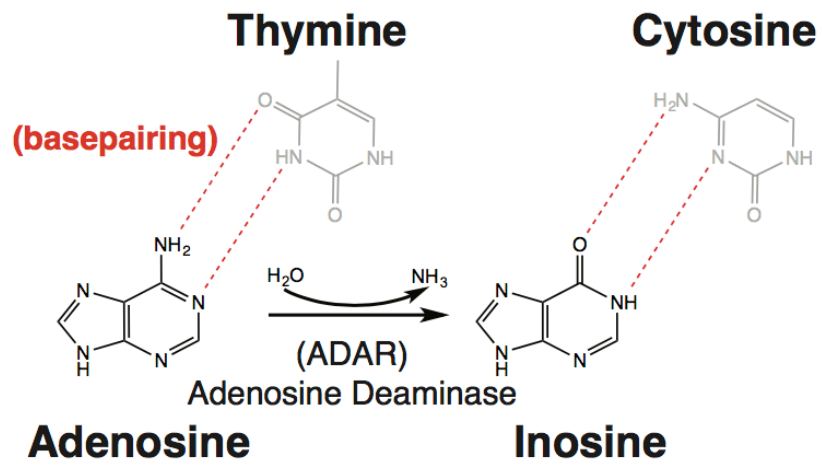


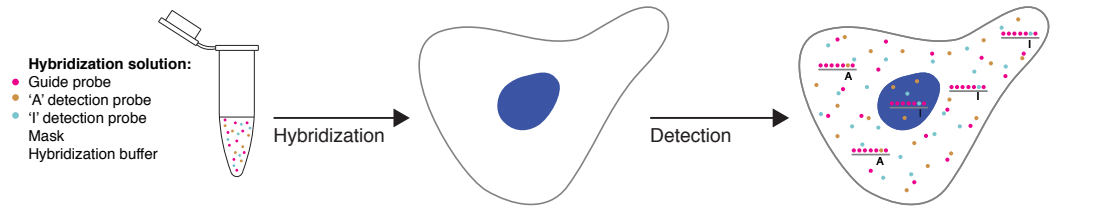
Figure 1.1: Chemical structures of adenosine and inosine. Inosine base-pairs with cytosine.

## COMPETITIVE HYBRIDIZATION AS A STRATEGY FOR DETECTING RNA

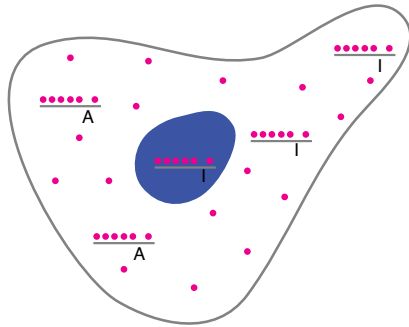
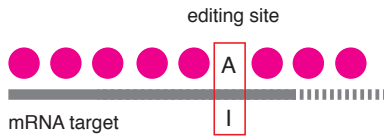
### EDITING *IN SITU*

Discriminating edited from unedited RNA via RNA fluorescence *in situ* hybridization (RNA FISH) is difficult because it relies on the hybridization of oligonucleotide probes to visualize the target of interest ([Raj et al., 2008](#)). Short oligonucleotides bind nonspecifically while long oligonucleotides cannot discriminate single-base differences. We thus used a 'toehold probe' strategy ([Levesque et al., 2013](#)) to reduce the initial hybridization region of our detection probes in order to confer selectivity based on single-nucleotide differences (**Figure 1.2**). Our scheme took advantage of the fact that inosine preferentially binds to cytosine rather than thymine by using two detection probes that compete to target the unedited, adenosine-bearing sequence using a thymine, and the edited, inosine-bearing sequence using a cytosine. Upon specific binding, the "mask" sequence is released by strand displacement to stabilize hybridization. However, single oligonucleotides are still prone to nonspecific binding, so we simultaneously used smFISH (the "mRNA guide" probe) to target a constant region of mRNA, coupled to a unique fluorophore (**Figure 1.2**). The mRNA guide showed us where to look for specific detection probes.

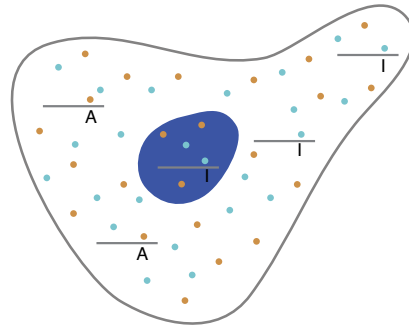
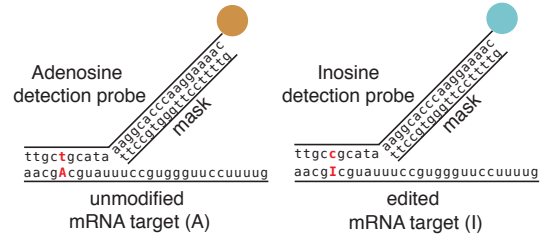
Figure 1.2: Overview of the inoFISH probe design strategy



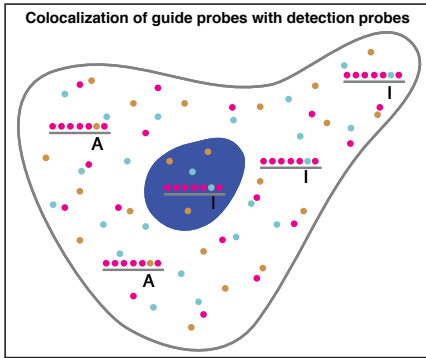
Guide probe = smFISH



Adenosine detection probe : Inosine detection probe : mask  
 1 : 1 : 3



- Mask probe binds constant region on both detection probes
- Mask probe temporarily shortens the binding region of detection probes
- Nucleotide-specific detection probes compete for binding on target sequence
- Short binding region and competition allow for single base resolution
- Mask dissociates by strand displacement to stabilize hybridization





## VALIDATING INOFISH OF GRIA2 IN HUMAN NEURAL LINEAGE CELL LINES

To test whether inoFISH could visualize adenosine-to-inosine editing, we chose the canonical, well-studied example of the Glutamate receptor 2 transcript (*GRIA2*). (*GRIA2* editing is critical for neuronal function([Seeburg et al., 2001](#)) and defects in *GRIA2* editing have been associated with ALS([Yamashita et al., 2012](#)).) We confirmed that *GRIA2* was edited by comparing genomic DNA and cDNA sequence in SH-SY5Y cells (**Figure 1.3**), and verified that it is a viable target for smFISH based on probe design constraints. ([Raj et al., 2008](#)) Combining four biological replicates, 10.53% of mRNA guides uniquely colocalized with adenosine or inosine detection probes, with 5.25% and 5.28% of *GRIA2* guides colocalizing with the adenosine-detection and inosine-detection probes respectively **Figure 1.4a,b**). The estimated mean editing level for *GRIA2* was 57.3% (95% confidence interval: 45.1%, 69.5%, full statistical model in **Appendix A**).

To confirm that detection probes did not colocalize with guide probes by random chance, we measured the rate of random colocalization by computationally shifting guide spots by 5 pixels in both the X and Y direction (“Pixel-shift”), thereby moving them outside the range of any true colocalization events (see **Appendix B; Figures 1.4b, 1.5**). Pixel-shift analysis reduced colocalization to 1.83% and 1.16% for adenosine and inosine, respectively, showing that most of colocalization events were specific. (Substituting an unrelated guide probe yielded similar results; **Figure 1.6**.) To check for dye-specific effects, we swapped fluorophores on the detection probes (**Figures 1.4b, 1.5**), revealing variation in the estimated mean editing level of 22% (**Figure 1.5**, and **Appendix A**). Together, these findings show that inoFISH can measure editing levels provided that one checks for dye-specific biases in detection probe sets.

Figure 1.3: Sanger sequencing of GRIA2 RT-PCR product

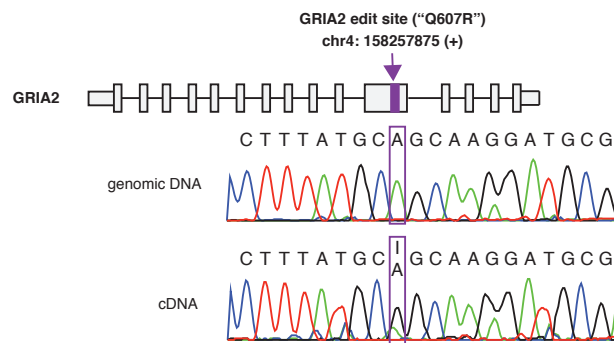


Figure 1.3: Sanger sequencing of GRIA2 RT-PCR product

Figure 1.4

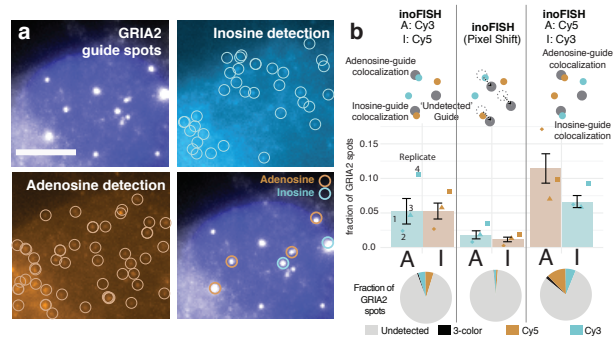


Figure 1.4: inoFISH discriminates between inosine and adenosine. (a) Fluorescence micrographs of Cal Fluor 610-labeled guide probe detecting GRIA2 mRNA (upper left), Cy5-labeled inosine detection probe (upper right) and Cy3-labeled adenosine detection probe (lower left), colocalized (lower-right). Scale bar, 5  $\mu$ m. (d) GRIA2 inoFISH results in SH-SY5Y cells (4 biological replicates), including pixel-shift and dye-swap controls; inoFISH probe detection efficiencies per-replicate (points) and mean +/- s.e.m.. Full summary of guide spot labels (pies; mean over all replicates).

Figure 1.5

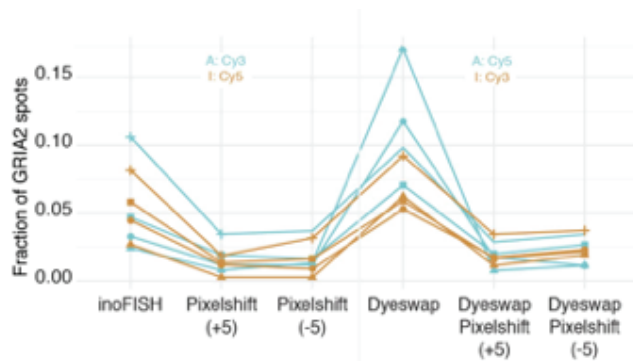


Figure 1.5: Expanded controls for GRIA2. InoFISH results for each replicate, including pixel-shift and dye-swap controls for GRIA2 in SH-SY5Y cells.

Figure 1.6

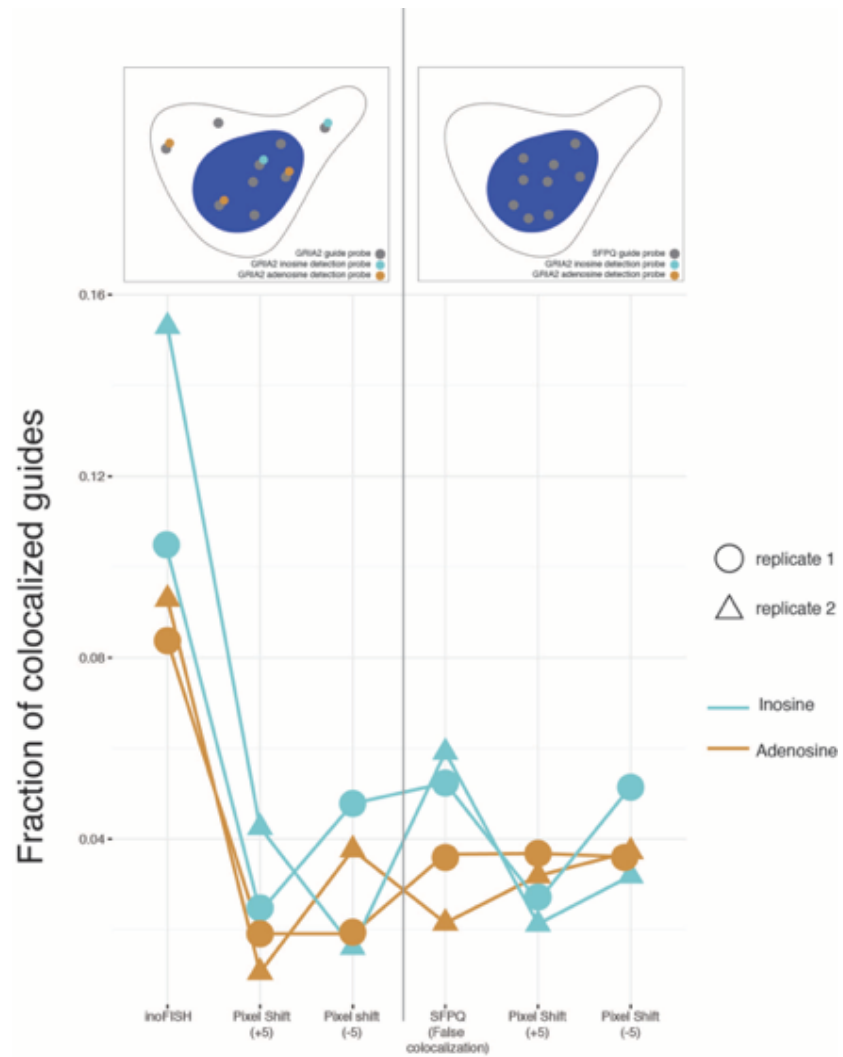


Figure 1.6: False-colocalization of GRIA2 detection probes with SFPQ guide probes. (top) Schematic representation of the experiment. Control experiment is normal GRIA2 inoFISH with GRIA2 detection probes. On the right, the guide probe has been replaced to target SFPQ with the GRIA2 detection probes. (bottom) Fraction of colocalized guides with GRIA2 (left) and SFPQ (right).

To verify that inoFISH signals were specific to inosine bases and not adenosine or guanosine bases, we altered the frequency of inosines in two different ways. *GRIA2* mRNA is primarily edited by the enzyme ADAR2, so we used siRNA to knock down *ADAR2* mRNA levels by 60% in SH-SY5Y cells ([Melcher et al., 1996](#); [O'Connell et al., 1997](#)). We observed a concomitant reduction in mean estimated *GRIA2* editing level from 65% to 14% (Parametric bootstrap  $p = 0.0004$ , see **Appendix A; Figure 1.7a,c**). We also chemically modified inosine bases with acrylonitrile on the N<sup>1</sup> position to prevent base pairing to cytosine ([Sakurai et al., 2010](#); [Yoshida and Ukita, 1968](#)), reducing observed editing level from 52.1% to 13.5% (Parametric bootstrap  $p = 0.0006$ ) (**Figure 1.7b,d**).

Additionally, we designed a guanosine-carrying “false detection” probe, which should not bind to either the edited or unedited transcript; it did not bind more than expected by chance (**Figure 1.8**). These results show that inoFISH specifically discriminates adenosine and inosine bases.

## COMPARING EDITING RATE ESTIMATES OF INOFISH AND OTHER METHODS

We validated inoFISH estimates of editing levels by comparing them to three established population-based methods (**Figure 1.9**). We generated *GRIA2* cDNA and estimated editing ratio either through Sanger sequencing or by digesting with a restriction enzyme specific to cDNA from the edited transcript ([Paschen et al., 1994](#)). We also cloned and sequenced individual *GRIA2* cDNA molecules. We found editing ratios of 59%, 54.9% and 50%, respectively, consistent with the 57.3% mean estimated editing level (95% confidence interval: [45.1%, 69.5%]) measured by inoFISH. Publicly available RNA-sequencing data from untreated SH-SY5Y cells also revealed *GRIA2* editing (see **Appendix B** and **Figure 1.9**).

Figure 1.7

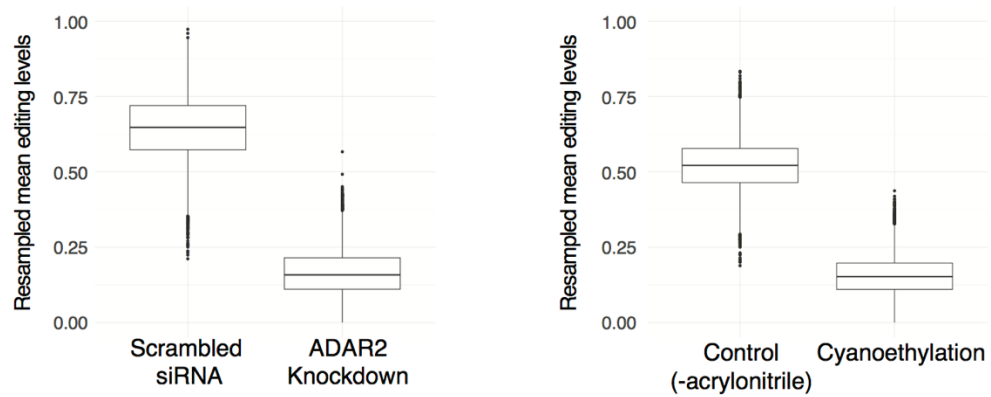


Figure 1.7: Resampled mean GRIA2 editing level estimates from ADAR2 knockdown and cyanoethylation experiments. (left) Resampled mean editing levels of scrambled siRNA vs ADAR2 siRNA knockdown. (right) Resampled mean editing levels of samples treated by cyanoethylation and control samples (-acrylonitrile). 10000 resampled mean estimated editing level values per boxplot.

Figure 1.8

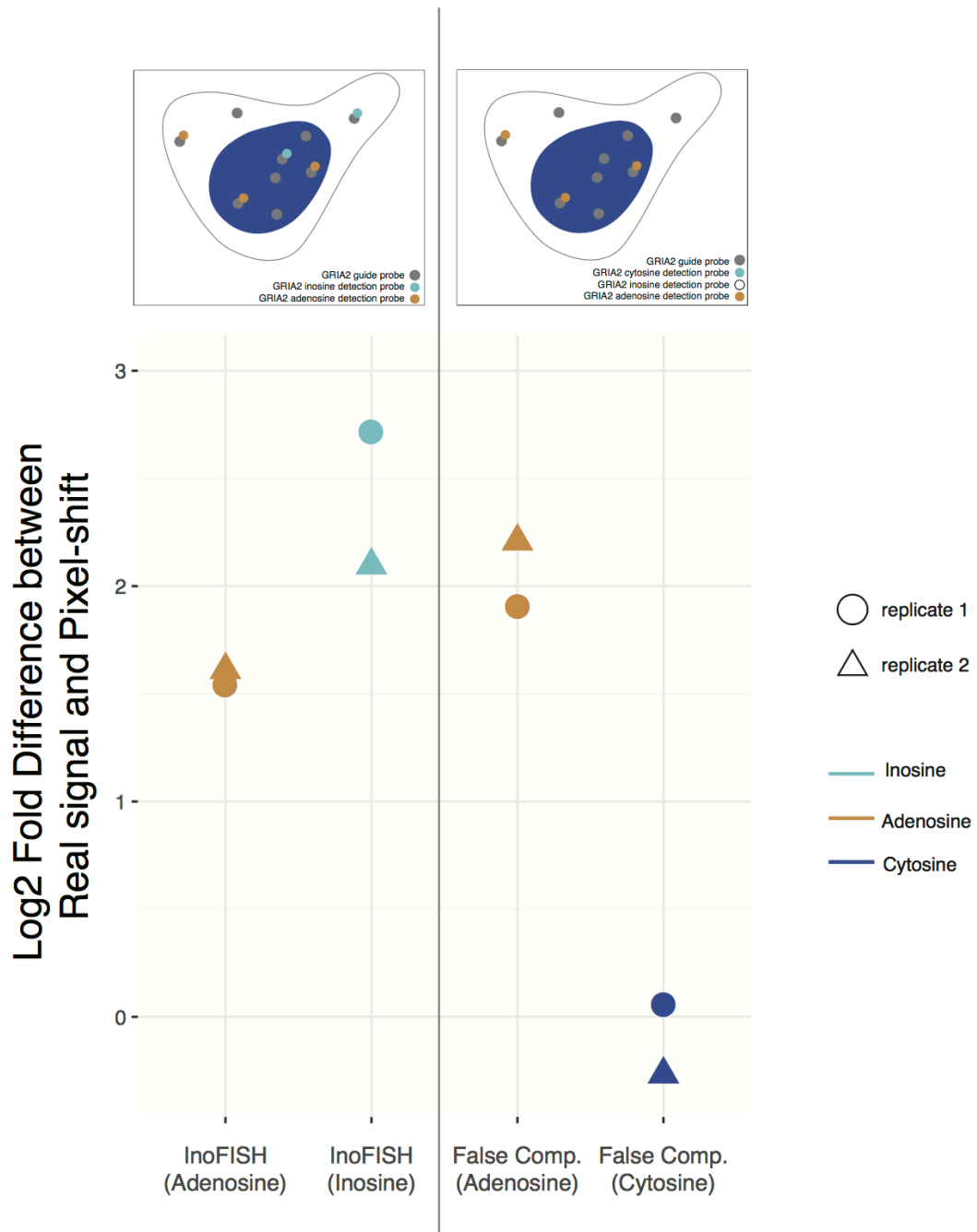
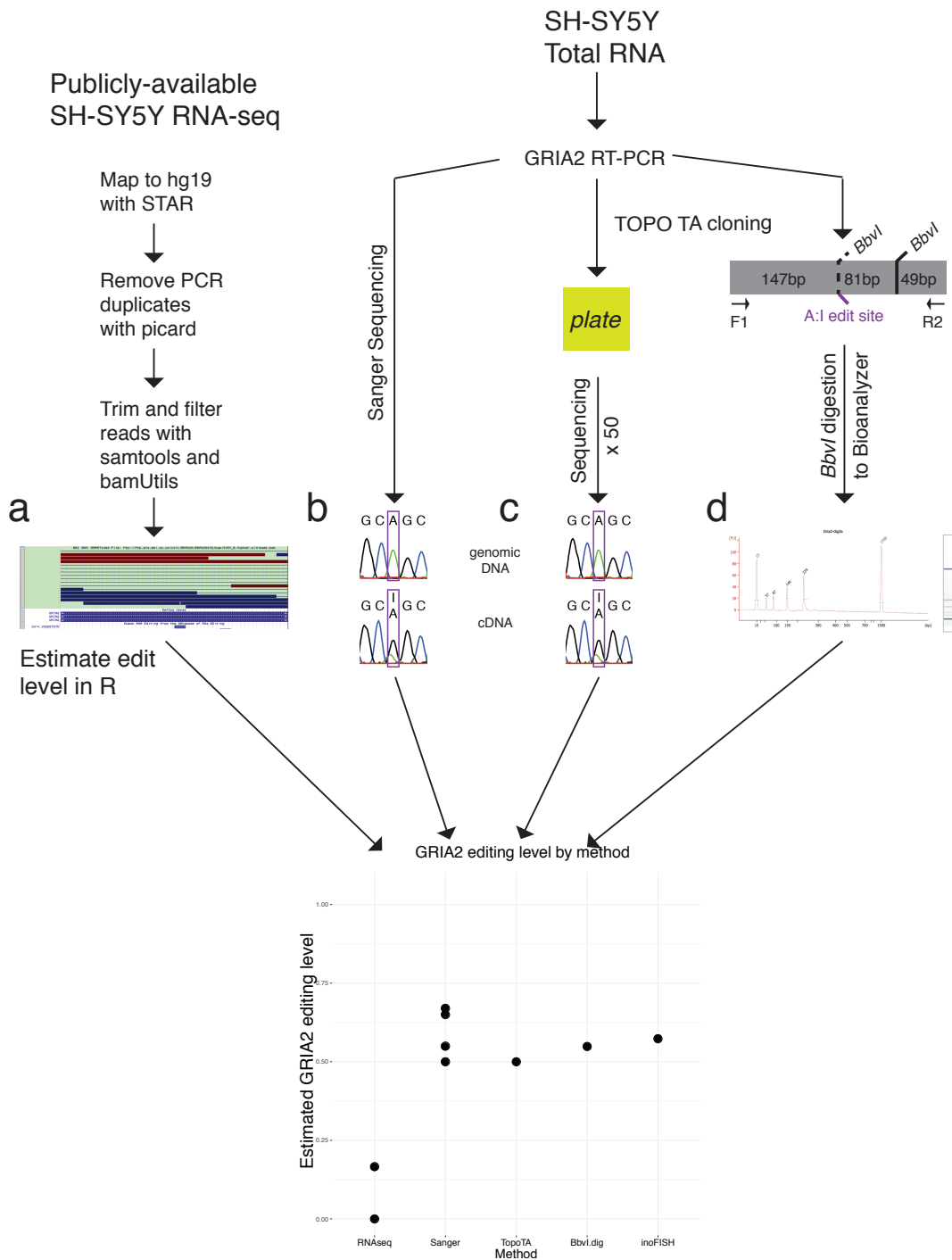


Figure 1.8: Comparison of inoFISH with targeted and off-target detection probes.(top) Schematic representation of the experiment. Control experiment is normal GRIA2 inoFISH with GRIA2 detection probes. On the right, the guide probe has been rcolocalized with an adenosine detection probe and a cytosine detection probe that doesn't target either nucleotide. (bottom) Fraction of colocalized GRIA2 guides by normal inoFISH (left) and inoFISH with false competition (right).



Figure 1.9: GRIA2 editing levels estimated by inoFISH and traditional methods



## IDENTIFICATION OF CANDIDATE INOFISH-COMPATIBLE RNA EDITING SITES WITH RNA-SEQ

Besides the well-studied editing target *GRIA2*, we wanted to test inoFISH on editing targets that are less commonly studied in the literature. For these we referred to the literature, to the RADAR database of RNA editing, and to publicly available RNA-seq data for screening. We identified *NUP43* and *EIF2AK2* as targets 1) with conserved editing sites across humans, chimps, and mice; 2) that are studied by a small number of research groups; 3) that are candidate editing targets in published transcriptome-wide adenosine-to-inosine editing screens; and 4) that have editing sites amenable to inoFISH probe designs. (**Figure 1.10**)

Then we wanted to check if these transcripts were also edited to any extent in the neural lineage cell lines we were studying: SH-SY5Y and U-87 MG. Therefore we downloaded publicly available RNA-seq data for these cell lines and conducted our own assessment of A-to-I editing at the annotated editing sites. We identified three sites amenable to inoFISH probe design across these two genes, 1 in *NUP43* and 2 in *EIF2AK2*. We validated as editing targets by Sanger sequencing of genomic DNA and cDNA (**Figure 1.11**). For two of these sites we were able to design inoFISH detection probe sets that gave colocalization with guide probes above pixel-shift estimates of random chance.

Figure 1.10

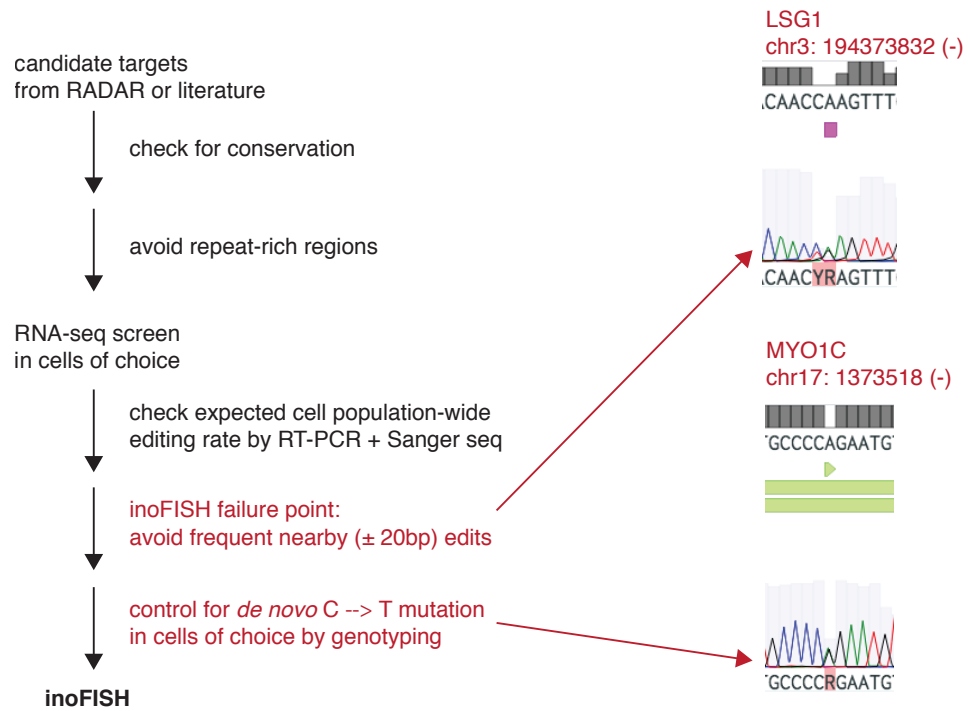


Figure 1.10: Outline of experimental design pipeline for inoFISH, including representative results (LSG1 RT-PCR/Sanger and MYO1C genotyping/Sanger) at critical steps in the target selection process.

Figure 1.11

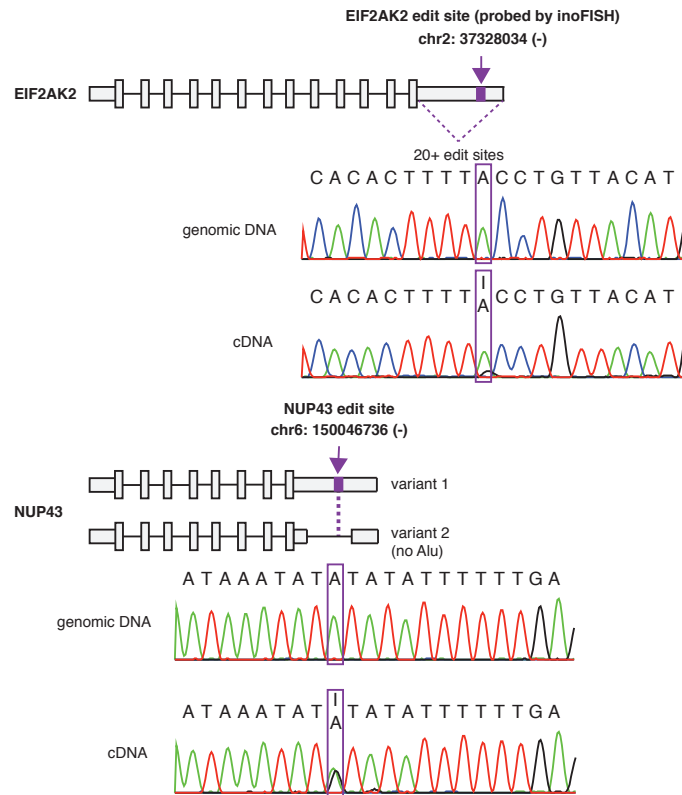


Figure 1.11: Sanger sequencing of EIF2AK2 and NUP43 RT-PCR products

## SUBCELLULAR LOCALIZATION OF EDITED AND UNEDITED TRANSCRIPTS IN NEURAL LINEAGE CELL LINES

We next measured the subcellular localization of edited and unedited transcripts. Previous studies used cell fractionation to show that unmodified RNAs exist in both the nucleus and the cytoplasm, whereas hyper-edited RNAs—but not selectively edited RNAs—were predominantly nuclear ([Kumar and Carmichael, 1997](#)), ([Zhang and Carmichael, 2001](#)). Other studies have shown that mRNAs containing Alu repeats, which are prone to adenosine-to-inosine editing, are inefficiently exported to the cytoplasm ([Chen and Carmichael, 2009](#)).

We therefore looked for associations between editing status and subcellular localization of *GRIA2* transcripts. We classified *GRIA2* transcripts as nuclear if they overlapped with the nuclear stain DAPI. Estimated *GRIA2* editing levels were roughly equal in both cellular compartments ( $p = 0.38$ ; **Figure 1.12a, 1.13a**). (Uncharacteristically of most mRNAs, 93.4% of *GRIA2* transcripts localized to the nucleus, though they were still translated; **Figure 1.14**)

We also used inoFISH to visualize localization of adenosine-to-inosine editing in two additional targets: the hyper-edited transcript *EIF2AK2* ([Wang et al., 2013](#)) (**Figure 1.12b**) as well as the Alu-bearing *NUP43* ([Chen and Carmichael, 2009](#)) (**Figure 1.12c**) in U-87 MG cells (**Figure 1.10**), which we found that 6.91% and 5.57% of *EIF2AK2* guide spots colocalized with adenosine- and inosine-specific detection spots, respectively (**Figures 1.12b, 1.15**), giving a population-wide mean editing level estimate of 36.4% (95% confidence interval: [20.4%, 53.1%]). For *NUP43*, 11.3% and 12.4% of guide spots colocalized with the adenosine- and inosine-specific detection spots (**Figures 1.12c, 1.15**), giving a population-wide editing level estimate of 53.2% (95% confidence interval: [45.1%, 61.2%]). In both cases, the editing level did not vary between nucleus and cytoplasm ( $p=0.18, 0.81$  for *EIF2AK2, NUP43*, respectively); (**Figures 1.12b,c,**

1.13). Note that the three inoFISH targets studied had detection efficiencies of 10%, 12% and 24%; the reasons for this variability is unknown, but it is within previous bounds([Levesque et al., 2013](#); [Shaffer et al., 2015](#)). As before, cyanoethylation reduced the percentage of inosine-detection probe colocalization with guide probe for both *EIF2AK2* and *NUP43*, again showing specificity (**Figure 1.16**).

Figure 1.12

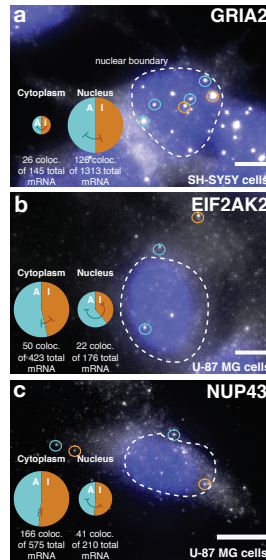


Figure 1.12: Analysis of subcellular localization using inoFISH. Nuclear localization analysis reveals no significant differences (parametric bootstrapping) in editing levels for each target between nucleus and cytoplasm. (a) GRIA2 ( $n = 4$  biological replicates;  $p = 0.38$ ), (b) EIF2AK2 ( $n = 3$ ;  $p = 0.18$ ) and (c) NUP43 ( $n = 2$ ;  $p = 0.81$ ) transcripts: representative overlays and fractions of labeled transcripts found to be unedited or edited (inlay).

Figure 1.13

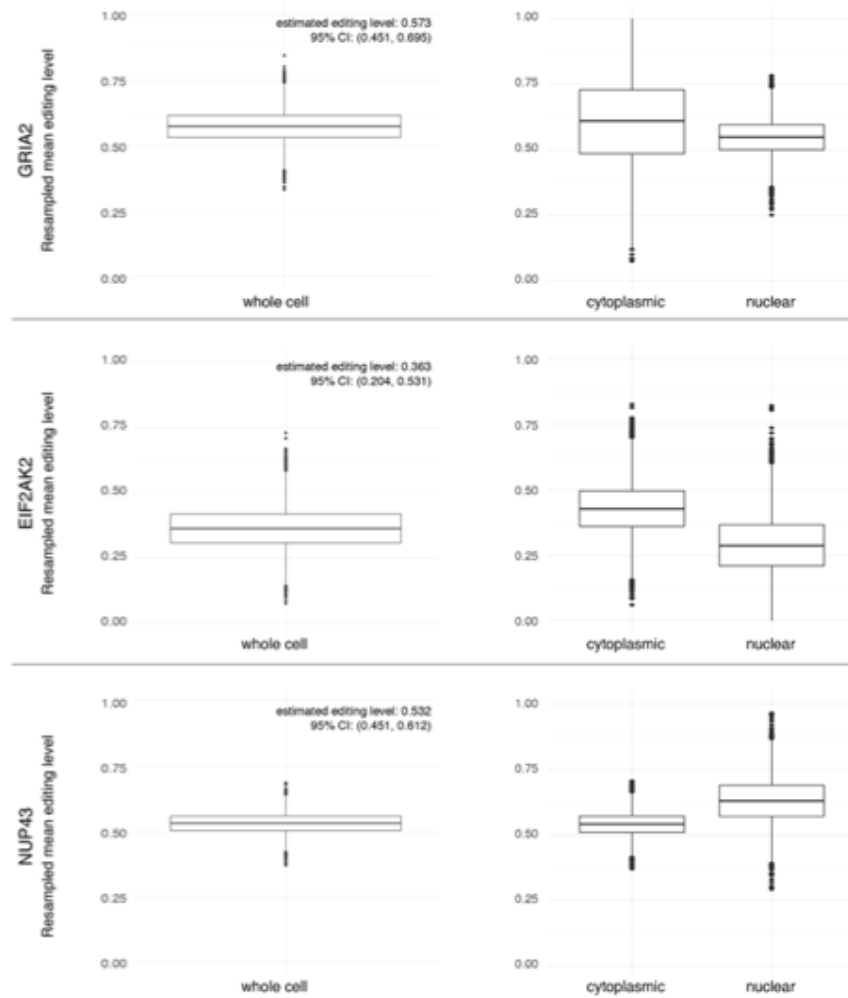


Figure 1.13: Mean editing level estimate distributions for all three targets. Boxplots of 10000 parametric bootstrapped samples of editing level per boxplot for GRIA2 (top), EIF2AK2 (middle), and NUP43 (bottom). Mean editing level models resampled irrespective of subcellular localization (left) and considering nuclear vs cytoplasmic localization (right).



Figure 1.14

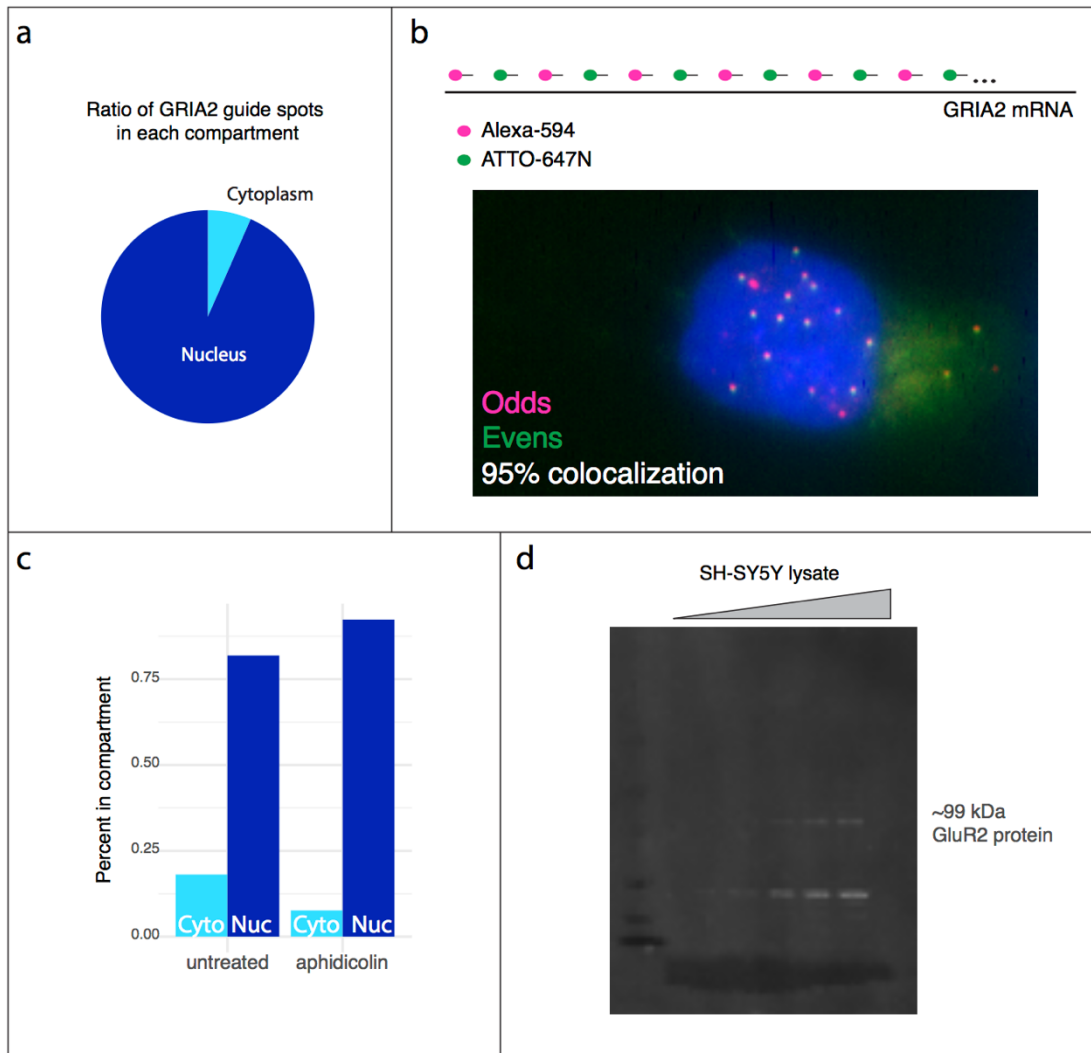


Figure 1.14: GRIA2 mRNA nuclear retention analysis. (a) Most GRIA2 transcripts are retained in the nucleus in SH-SY5Y cells. (b) GRIA2 guide probe specificity assessed by odds-evens probe set division experiment. (c) Nuclear retention of GRIA2 upon treatment with aphidicolin, a cell cycle inhibitor. (d) Western blot of GluR2 protein in SH-SY5Y cell lysate.

Figure 1.15

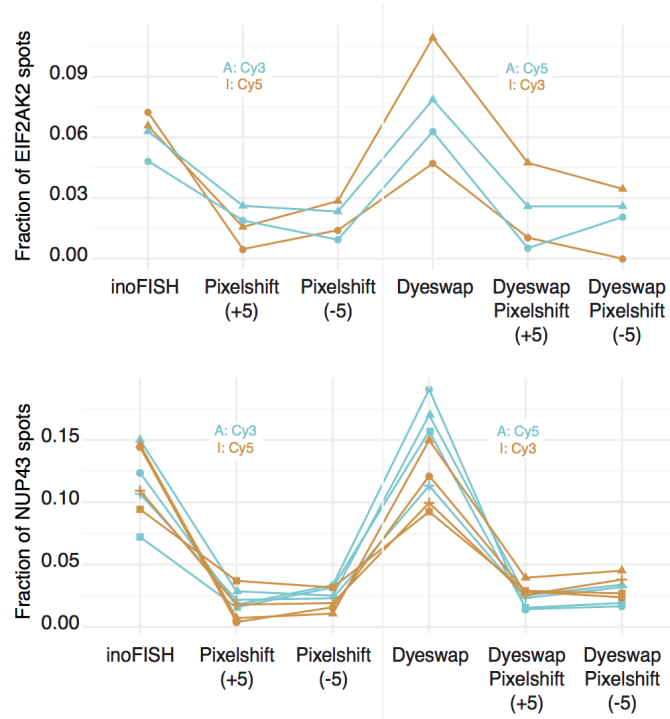


Figure 1.15: Expanded controls for EIF2AK2 and NUP43 inoFISH. InoFISH results for each replicate, including pixel-shift and dye-swap controls for EIF2AK2 in U-87 MG cells (top) and NUP43 in U-87 MG cells (bottom).

Figure 1.16:

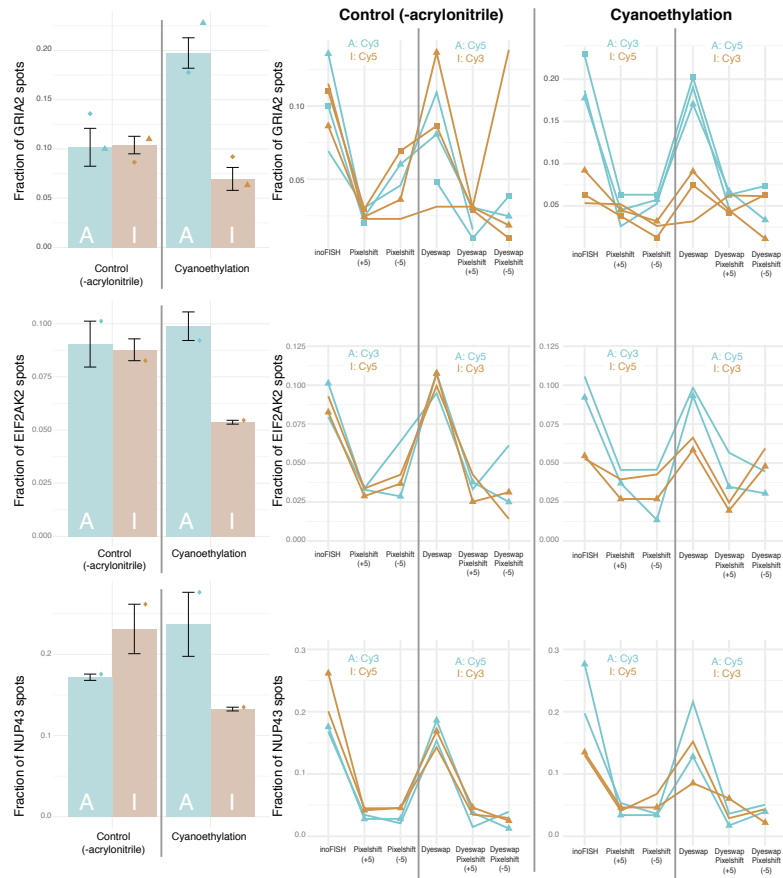


Figure 1.16: Cyanoethylation reduces number of detected inosines in situ. inoFISH mean  $\pm$  s.e.m after cyanoethylation treatment from a minimum of 2 biological replicates (left) including pixel-shift and dye-swap controls per replicate (right) for GRIA2 in SH-SY5Y cells (top), EIF2AK2 in U87 MG cells (middle), and NUP43 in U87 MG cells (bottom).

## ASSOCIATION OF EDITED AND UNEDITED *GRIA2* TRANSCRIPTS WITH NUCLEAR PARASPECKLES

InoFISH also allowed us to test whether edited transcripts are trafficked to nuclear paraspeckles([Prasanth et al., 2005](#)). We performed inoFISH together with single-molecule RNA FISH of *NEAT1* RNA, a marker of nuclear paraspeckles([Sunwoo et al., 2009](#)), in SH-SY5Y cells (**Figure 1.17**), revealing that 8.57% of all *GRIA2* transcripts colocalized with paraspeckles (**Figure 1.17**). Simulations (see **Appendix B**) showed that the observed rate of *GRIA2*-paraspeckle association was 1.7-fold greater than expected by random chance (Simulation of *GRIA2*-paraspeckle association rate null distribution for one representative replicate  $p < 0.001$ , see Online Methods).

We then used inoFISH to determine whether edited or unedited *GRIA2* transcripts were preferentially associated with paraspeckles. We found no significant differences in the editing status in paraspeckles for *GRIA2* in SH-SY5Y cells ( $p = 0.44$ , determined via simulation, for one representative replicate; **Figure 1.17**), demonstrating that edited *GRIA2* transcripts in SH-SY5Y cells do not necessarily preferentially associate with paraspeckles.

Figure 1.17

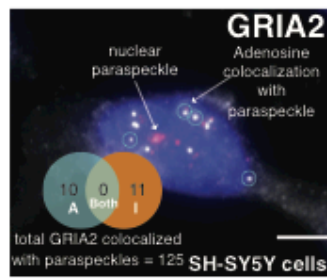


Figure 1.17: GRIA2 inoFISH results with NEAT1-colocalization pooled over  $n = 4$  biological replicates: representative overlays and counts of inoFISH-labeled, paraspeckle-associated transcripts. No significant difference between population-wide editing level and paraspeckle-associated editing level (simulation of conditional null per replicate;  $p = 0.44$  for one representative replicate)

## VISUALIZATION OF UNEDITED NUP43 TRANSCRIPTS AT THE SITE OF TRANSCRIPTION

We next used inoFISH to determine whether adenosine-to-inosine editing is co-transcriptional or post-transcriptional. Introns mark transcription sites, and colocalization of edited transcripts with intron signal would suggest that editing can occur co-transcriptionally. We concurrently performed *NUP43* inoFISH with single-molecule FISH targeting *NUP43* introns in 212 U-87 MG cells (**Figure 1.18**), observing 17 total transcription sites, of which 5 transcription sites contained unedited *NUP43* and none containing edited *NUP43* (**Figure 1.18**). This result does not rule out co-transcriptional editing of *NUP43* altogether, but does suggest that some *NUP43* editing may be post-transcriptional.

## GRIA2 AND NUP43 RNA EDITING LEVELS ACROSS A POPULATION

We also looked for evidence of fluctuations in editing level from cell to cell. We simulated inoFISH results in the cases of uniform (**Figure 1.19a**) or variable editing levels in single cells (**Figure 1.19b**). We found that *GRIA2* editing in single cells was not consistent with the constant editing level model, suggesting per-cell heterogeneity in *GRIA2* editing levels (**Figure 1.19c**). *NUP43* editing in U-87 MG cells, however, was consistent with the constant-editing level model (**Figure 1.19d**). Thus single-cell fluctuations in the level of editing may occur in a target-specific manner.

InoFISH provides a direct method for visualizing adenosine-to-inosine RNA editing in single cells with single-nucleotide resolution. Cell population-wide studies lack the resolution to provide information such as subcellular localization and cell-to-cell variability of RNA editing. This new tool will enable researchers to answer basic questions about edited RNA species and will enable a deeper understanding of the biology of adenosine-to-inosine editing.

Figure 1.18

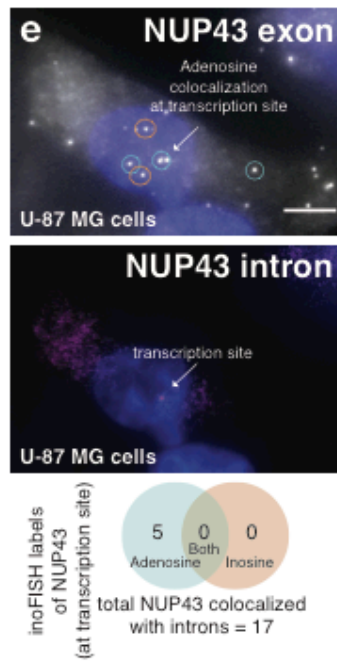


Figure 1.18: NUP43 inoFISH results with transcription site localization analysis (n = 2): representative images and counts of inoFISH-labeled, transcription site-associated NUP43 transcripts.



Figure 1.19

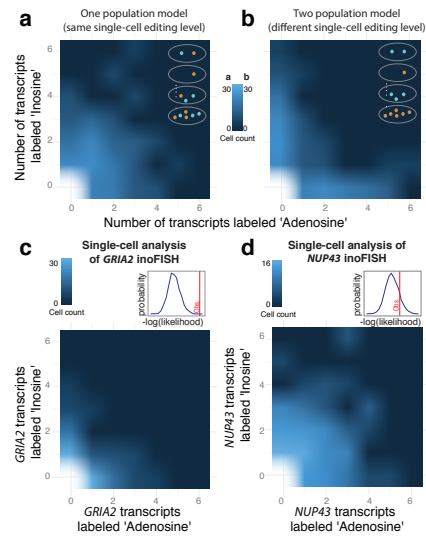


Figure 1.19: Single-cell analysis of inoFISH. Simulated inoFISH results assuming (a) binomially-distributed per-cell counts of edited and unedited transcripts; or (b) two populations of cells, one population with 95% editing and the other with 5% editing, mixed in proportion according to the population-wide editing level (c) Single-cell analysis of GRIA2 inoFISH results pooled over all 4 replicates and simulation of the exact conditional null distribution of  $-\log(-\text{likelihood})$  of the data under the binomial model specified in (a)(inset) (d) Single-cell analysis of NUP43 inoFISH results pooled over all 4 replicates (left) and simulation of the exact conditional null distribution of  $-\log(\text{likelihood})$  of the data under the binomial model specified in (a) (inset, see Appendix A).

## CHAPTER 2: PERTURBATION PANEL PROFILING (P<sup>3</sup>) IDENTIFIES TRANSCRIPTION FACTORS THAT ENHANCE DIRECTED CHANGES OF CELLULAR IDENTITY

MOST CURRENT PROTOCOLS FOR DIRECTED CHANGES OF CELL IDENTITY  
ARE INEFFICIENT

The directed conversion of one terminally differentiated cell type into another cell type without passing through a pluripotent intermediate stage, a process known as transdifferentiation, has a variety of scientific and translational implications. Over the last few years there have been many important discoveries of new transdifferentiation methods, however, transdifferentiation protocols for cell types of interest are usually very difficult to engineer, even *in vitro*. In particular, many existing transdifferentiation protocols change relatively few cells of the starting cell type and only do so in a way that incompletely reprograms to the target cell type ([Becker et al., 2017](#); [Fu and Srivastava, 2015](#)). Conversion of adult human fibroblasts to cardiac myocytes, for example, has been studied by several groups for years, but remains inefficient using genetic factors (i.e., gene overexpression or knockdown) alone ([Mohamed et al., 2017](#); [Nam et al., 2013](#)). Relatedly, despite more than a decade of study, the reprogramming to pluripotency of many terminally differentiated cell types also stubbornly remains inefficient. ([Cacchiarelli et al., 2015](#); [Takahashi and Yamanaka, 2016](#))

Overexpression of transcription factors often occupies the core of transdifferentiation protocols, with the idea being that these factors coordinate larger gene expression programs that are

specific to the target cell type. Strategies for identifying sets of transcription factor genes to be overexpressed or repressed for transdifferentiation fall into two major classes: testing combinations of a limited number of transcription factors known to influence the development of the target cell type, and screens of individual additional factors ([Cao et al., 2016](#); [Fu et al., 2013](#); [Nam et al., 2013](#); [Takahashi and Yamanaka, 2006](#); [Vierbuchen et al., 2010](#); [Zhou et al., 2017](#)). In recent years there have been several improvements to these strategies, each of which have enhanced the efficiency and fidelity of different transdifferentiation protocols. Several groups have considered predicted TF gene regulatory activity and specificity of TF gene expression in the cell types of interest to augment the limited sets of TFs to be tested combinatorially ([Cahan et al., 2014](#); [Morris et al., 2014](#); [Rackham et al., 2016](#); [Tomaru et al., 2014](#)). Others have made clever use of single-cell techniques and CRISPR-based epigenome engineering to improve the throughput of larger, more unbiased factor screens ([Black et al., 2016](#); [Duan et al., 2018](#); [Liu et al., 2018](#); [Parekh et al., 2018](#)). These new approaches have led to some improvement in the efficiency and accuracy of transdifferentiation toward several cell types across all three germ layers, but there is still plenty of room for improvement ([Guo and Morris, 2017](#)).

Here we propose a complementary experimental and analytical strategy, Perturbation Panel Profiling (P<sup>3</sup>), for the selection of transcription factor genes for transdifferentiation and other cellular reprogramming protocols. Different cell types display unique and complex phenotypes, but are unified by one major property: individually, each cell type is often resilient to a variety of minor perturbations. For example, *in vivo* myocytes remain myocytes and fibroblasts remain fibroblasts when challenged with changes in their environments, such as when people get sick, exercise, or take medications. That is, many mature cell types stabilize their identity in response to perturbation. *In vitro*, much the same can be seen when manipulating nonessential components of primary cell culture conditions. Can we use this common property of different cell

types—phenotypic stability to perturbation—to prospectively identify transcription factors that are capable of specifying a cell type of interest?

## PARALLEL SMALL-MOLECULE PERTURBATIONS OF FIBROBLASTS AND CARDIOMYOCYTES

Our first major experimental goal was to observe gene expression patterns transcriptome-wide in the cell types of interest exposed to a wide variety of perturbations. Specifically, we set out to conduct RNA-seq on hundreds of parallel bulk samples of each cell type of interest, cultured under standard (i.e., control) and tens of perturbed conditions. Ultimately we hoped to enhance directed changes of cell identity of human fibroblasts, either to cardiomyocyte-like states or more generally to other cell identities, like pluripotent stem cells. For our perturbation panels we used samples of human fibroblasts (GM942, Coriell) and genetically matched iPSC-derived cardiomyocytes (iCard-942) (**Figure 2.1**). We selected small molecule perturbagens in the SelleckChem Bioactive Library targeting any kinases or G protein-coupled receptors (GPCR) in order to induce *in vitro* perturbations that lead to changes in intracellular signal transduction (**Table 2.1**). Additionally, in order to simulate as many different types of perturbation as possible while still using small molecule drugs, we minimized redundant primary target set overlap, using information about signaling pathways and corresponding targets provided by the manufacturer, SelleckChem. We settled on a list of 100 drugs, administered individually or in pairs, for a total of 75 different perturbation culture conditions (**Figure 2.2**).

Figure 2.1

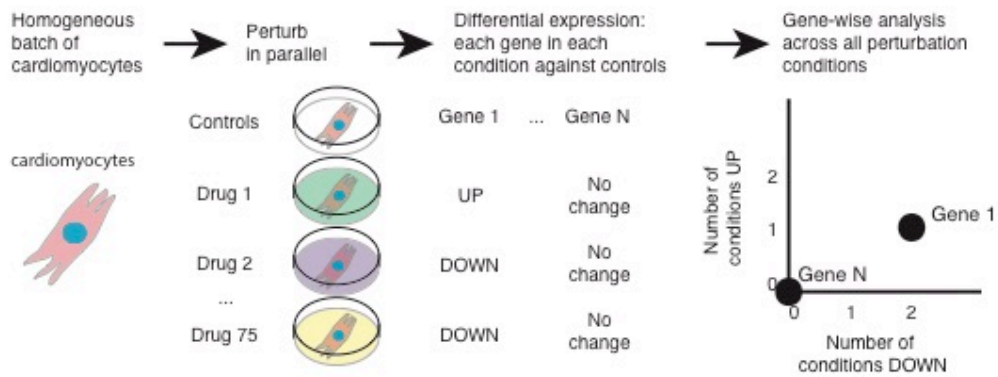


Figure 2.1: P<sup>3</sup> perturbation panel profiling for developing transdifferentiation protocols. In PerturbID perturbation panel profiling we expose parallel cultures of cells of a single cell type to small molecule drugs or to an equivalent volume of DMSO controls. We then perform RNAtag-seq on these samples in parallel for high-quality transcriptome-wide gene expression profiling, followed by differential expression analysis of each perturbation condition against controls. We integrate PerturbID results across all conditions to quantify the “perturbability” of each gene, as indicated by the number of conditions in which a gene is dysregulated.

Figure 2.2

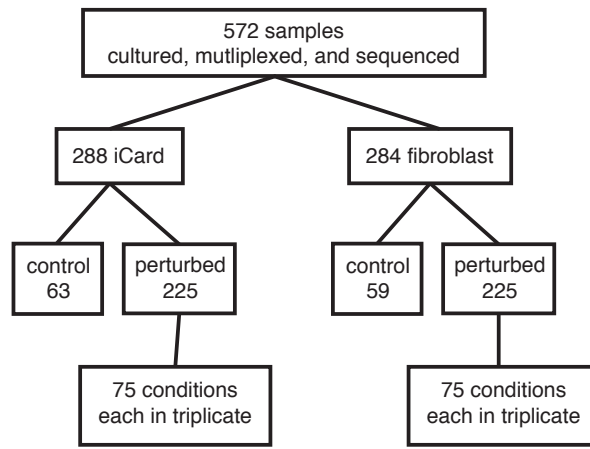


Figure 2.2: GM00942 fibroblast and iPSC-derived iCard-942 samples processed for P<sup>3</sup>

## A POTENCY-BASED DRUG DOSING SCHEME FOR PERTURBATION PANEL PROFILING (P<sup>3</sup>)

We first developed a dosing scheme that enabled screening of a large number of drugs that would perturb fibroblasts and cardiomyocytes in culture. In this scheme we wanted to avoid wasted samples that are either unperturbed by too low a drug dose or lethally perturbed by too high a dose. Therefore, we started by dosing parallel cultures of GM00942 dermal fibroblasts with a pilot panel of 12 drugs at a fixed concentration (100 nM) for 2 days, each in triplicate, along with vehicle controls (see **Appendix C**). We prepared RNA sequencing libraries in parallel for all samples using RNAtag-seq ([Shishkin et al., 2015](#)), and we quantified the number of differentially expressed genes in each drug condition relative to vehicle controls. We observed that only 7 of the 12 perturbations induced any differential expression of genes at a concentration of 100 nM (**Figure 2.3**).

We next wondered if the extent of gene expression perturbation (i.e., number of differentially expressed genes induced) was correlated with the potency of the drug. Therefore, we mined the literature for the annotated IC<sub>50</sub>s of each drug for each of its known targets. We observed anticorrelation between the median annotated IC<sub>50</sub> of each drug with the number of differentially expressed genes it induced (**Figure 2.4**). More potent drugs have lower IC<sub>50</sub>s, so we concluded that a strategy for consistently delivering a drug at a high enough dose to elicit a response might be achieved by tuning the dosage its known IC<sub>50</sub>s.

Figure 2.3

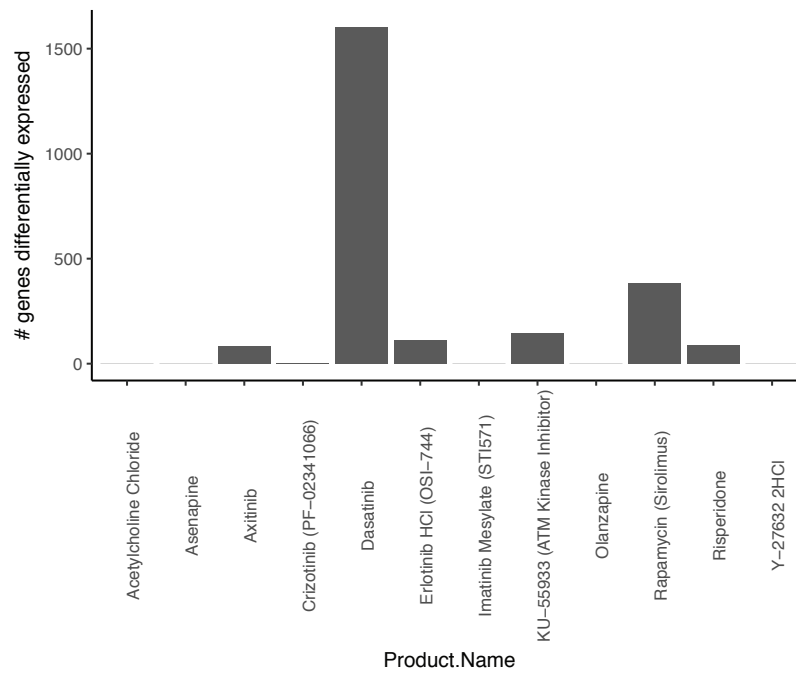


Figure 2.3: Number of differentially expressed genes detected relative to DMSO controls in GM00942 fibroblasts.



Figure 2.4

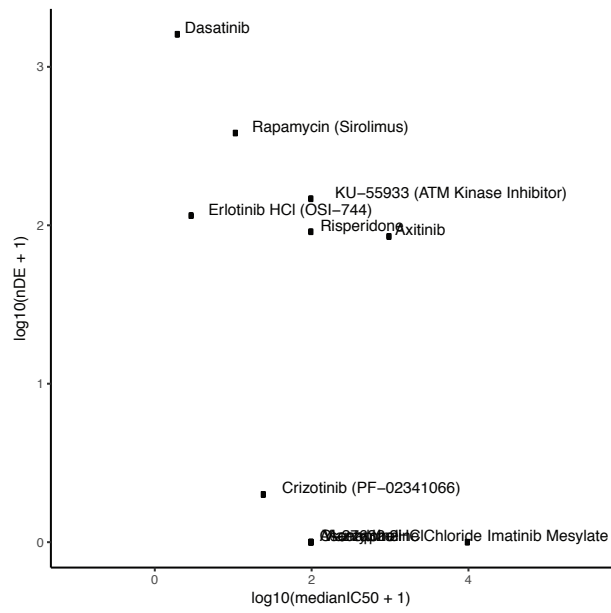


Figure 2.4: Comparison of differentially expressed genes per drug and the median IC50 for each drug relative to its annotated targets

Following up on this, we set out to identify whether there might be a predictable ratio of a drug's known IC50s that we could deliver to elicit a perturbation that was both significant enough to induce differential gene expression and mild enough to not grossly change the cells. That is, we also wanted to avoid killing the cells, altering their morphology, or in the case of cardiomyocytes, inhibiting their stereotypical ability to spontaneously beat in 2D culture. Therefore, we next dosed parallel cultures of GM00942 dermal fibroblasts with a new panel of 12 drugs, each at 3 drug-specific doses for 2 days (1x, 20x, and 100x the median annotated IC50 of that drug), along with vehicle controls. We administered the same doses of the same drugs to parallel cultures of iPSC-derived cardiomyocytes of the same genetic background (See **Appendix C**). We again performed highly parallelized RNA-seq with RNAtag-seq and quantified differential gene expression for each drug at each dose relative to vehicle controls of the same cell type.

We observed that 9 of 12 drugs induced some differential gene expression in fibroblasts at 100x IC50 and 10 of 12 drugs induced any differential gene expression in iCards at that same dose, while generally preserving gross morphology and iCard beating activity (**Figure 2.5**). At lower doses, these drugs tended to induce fewer differential gene expression events. Therefore, we concluded that for a larger set of small molecules with similar classes of targets (i.e., GPCRs and kinases), we should be able to frequently induce a moderate-strength differential gene expression response by administering these drugs at 100x the median annotated IC50.



We next wondered whether there was any information about the drugs that failed to induce any differential expression that we might use to avoid such lack of perturbation. Therefore we asked whether gene expression of a drug's targets in the cell type of interest might be correlated with the presence or absence of differential gene expression. We found that the targets of the drugs that failed to induce differential gene expression in fibroblasts were not expressed in fibroblasts (**Figure 2.6**). Therefore, we also decided to restrict the list of P<sup>3</sup> drugs to those whose targets are expressed in both GM00942 fibroblasts and iCard-942 cells (See **Appendix C**).

## MOST P<sup>3</sup> PERTURBATIONS INDUCE DIFFERENTIAL EXPRESSION PROFILES WHILE GROSS CELLULAR PHENOTYPES REMAIN STABLE

After 4 days of perturbation or DMSO-only exposure, we took transmitted light videos of iCard-942 cultures to check whether cells were still beating and morphologically similar to controls as a proxy for maintained cell identity, and found that 63 out of 75 conditions had wells with beating cells and without evident cell death. Similarly, we took transmitted light images of GM942 fibroblast perturbation cultures and found that 62 out of 75 conditions had wells with fibroblastic cells and without evident cell death (**Figure 2.7**). In sum, we developed a strategy for delivering small molecule perturbagens in a relatively highly parallelized fashion to samples of the same cell type in a format compatible with bulk RNA sequencing without obviously altering their grossest phenotypes.

Figure 2.6

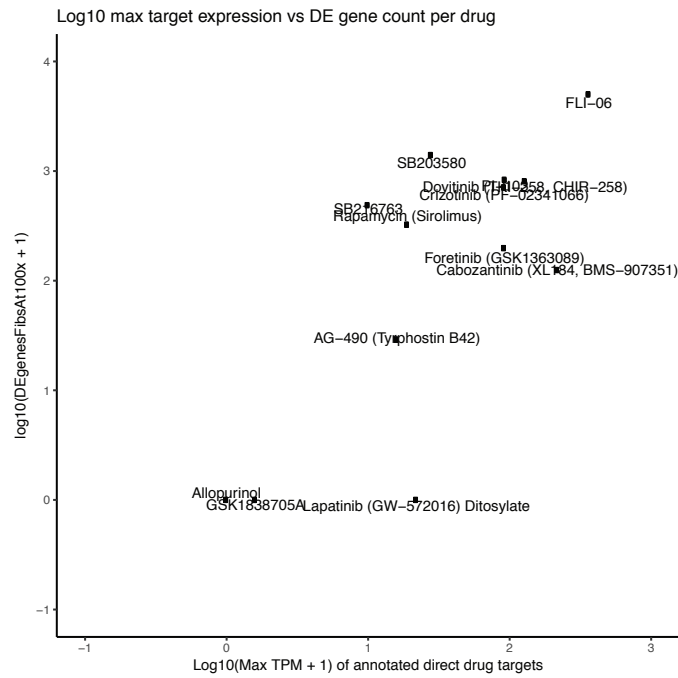


Figure 2.6: Comparison of number of differentially expressed genes in GM00942 fibroblasts against the average expression level of drug target in GM00942 fibroblasts

Figure 2.7

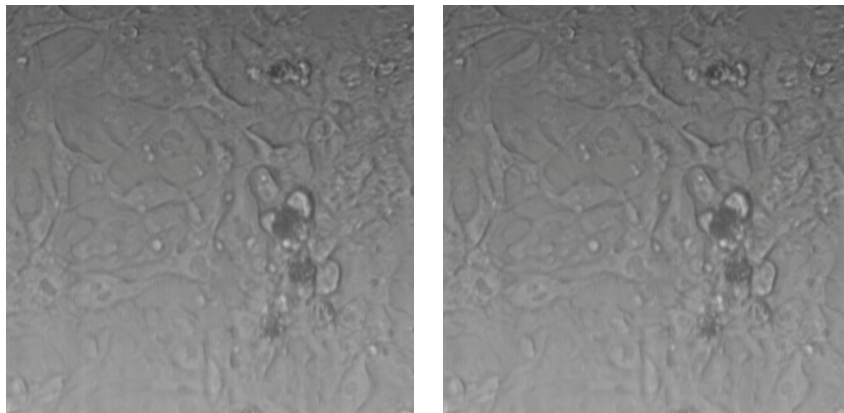


Figure 2.7: Representative frames from quality control videos for iCards-942 cells in 96-well format. Note the stretching in cells during contraction, or "beating", activity.

We next sought to conduct transcriptome-wide gene expression profiling of the hundreds of perturbed (and control) samples of cardiomyocytes and fibroblasts. We performed RNAtag-seq in batches of 96 samples, spreading DMSO controls evenly across each batch.

In our exploratory analysis, we clustered gene expression profiles across the 458 quality-controlled samples (454, excluding HeLa outgroup) clustered first by cell type, and then often by perturbation condition (**Figure 2.8**). We identified hundreds to thousands of differentially expressed genes in most perturbed conditions relative to corresponding cell type controls (**Figure 2.9, Appendix C**). Note that the number of detected differentially expressed genes is power-limited by relatively low sequencing depth. However, for genes with high expression, i.e., mean RPM > 20, we estimated that we can detect most differential expression events at the current sequencing depth (**Figure 2.10**). (We intentionally chose relatively low sequencing depth as a trade-off with an increased number of perturbation conditions; this allowed us to focus on dysregulation patterns of highly expressed genes across a large number of samples.) This shows that our perturbation panel did in fact elicit perturbation of gene expression levels, even when the cell's type, or grossest phenotypes, remained stable.

Figure 2.8

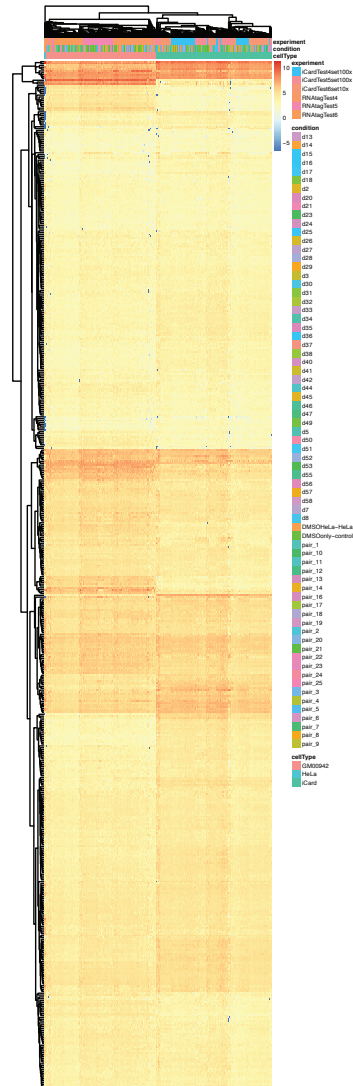


Figure 2.8: Heatmap with hierarchical clustering of all quality-controlled samples based on  $\log_2(\text{TPM})$  values of all genes expressed at 20RPM in both GM00942 and iCard-942 cells.



Figure 2.9

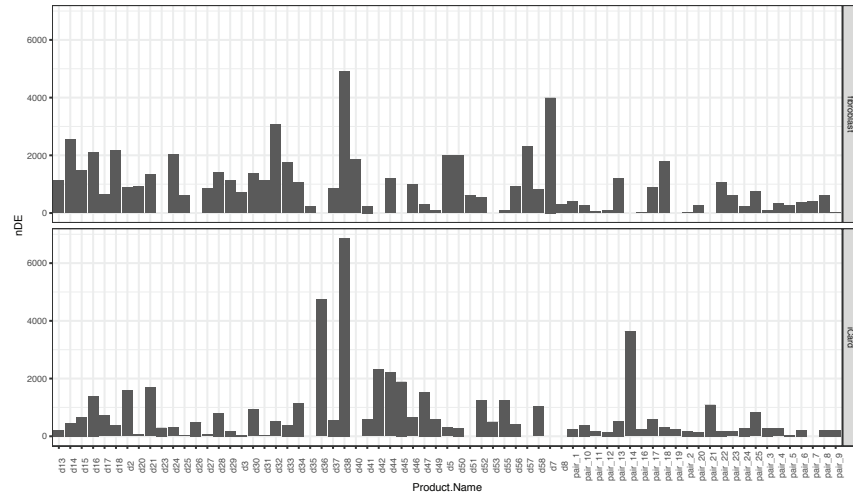


Figure 2.9: Differentially expressed gene counts per drug in GM00942 fibroblasts and iCard-942.

Figure 2.10

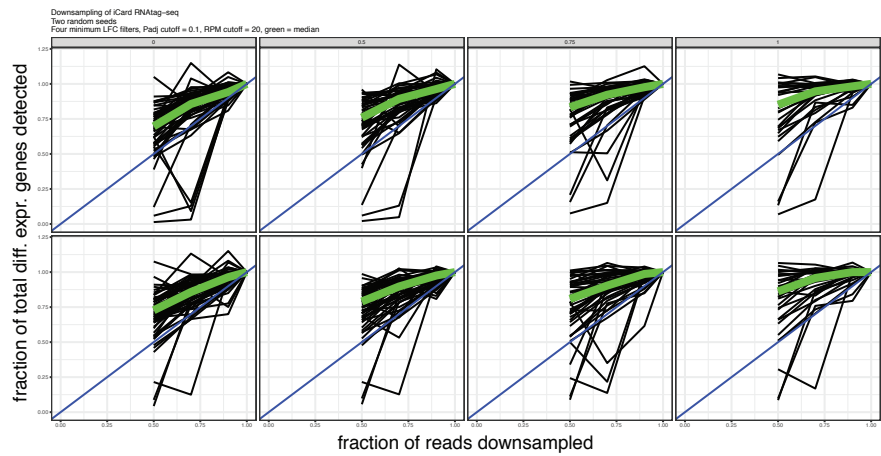


Figure 2.10: Downsampling analysis of iCard-942 RNAtag-seq data. Counts per sample were randomly downsampled to 50%, 70%, and 90% of the total dataset, using two different random seeds, and re-calculated. Each black line represents the fraction of differentially expressed genes still detected at the downsampled read fraction. The green line represents the median fraction of total differentially expressed genes per drug at that downsampled read fraction. Results are shown for four minimum log<sub>2</sub>FoldChange filters (0, 0.5, 0.75, and 1). Drugs are only included in this analysis if they have 200 or more differentially expressed genes detected in the full dataset.

## MOST HIGHLY EXPRESSED GENES ARE DIFFERENTIALLY EXPRESSED AFTER AT LEAST ONE PERTURBATION

We then wondered whether only a small subset of genes were dysregulated in each cell type across the 60+ perturbation conditions, or if most genes were dysregulated in response to some perturbation. To check this, in each cell type for each gene that was highly expressed in controls of that cell type, we counted the number of perturbations causing differential expression vs. controls. We found that the vast majority of highly expressed genes in each cell type were differentially expressed in at least 1 quality-controlled perturbation condition relative to that cell type's controls (**Figure 2.11**), suggesting that most highly expressed genes can be dysregulated by some of the signaling perturbagens in our library.

Figure 2.11

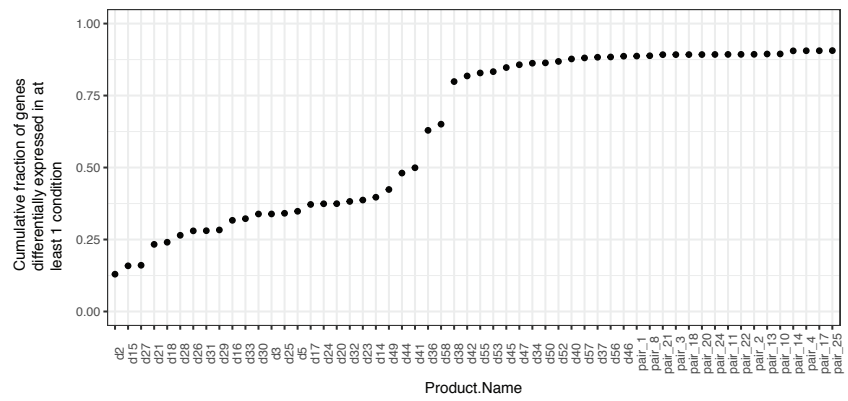


Figure 2.11: Cumulative fraction of genes differentially expressed in at least one condition as one consider more drugs in the dataset. Only genes with minimum average expression of 10 TPM shown.

## CARDIOMYOCYTE LINEAGE-DRIVING TRANSCRIPTION FACTORS ARE UP-REGULATED IN MORE PERTURBATION CONDITIONS THAN OTHER HIGHLY EXPRESSED TRANSCRIPTION FACTORS

Given that most genes can be differentially expressed in at least one perturbation condition, we wondered if some genes were dysregulated in more conditions than other genes. In particular, we wanted to know if known lineage-driving transcription factor genes were dysregulated following a relatively low or high number of different perturbations, i.e., whether these known lineage-driving TF genes are more or less “perturbable” than other highly expressed TF genes. If high or low perturbability is a feature of known lineage-driving TF genes, perhaps it could be used to prospectively identify other TFs that may also play a role in specifying or maintaining lineage identity (and therefore might be useful in transdifferentiation experiments). We pre-registered a set of 14 cardiomyocyte lineage-driving TF genes prior to analyzing our RNAtag-seq data (**Figure 2.12**). For each of these genes, we counted the number of perturbations that resulted in differential expression in cardiomyocytes vs. controls. Although we had initially hypothesized that lineage-driving TFs would be less perturbable than other highly expressed TF genes, we found the opposite to be true: lineage-driving TFs are highly perturbable.

Specifically, they are up-regulated across many different perturbation conditions (**Figure 2.12**). This effect is not explained by differences in the power to detect differential expression due to differences in average expression levels; i.e., lineage-driving TFs are more perturbable than other TF genes expressed at the same level in control samples (**Figure 2.13A**). Further, it is not the case that lineage-driving TF genes are up-regulated more dramatically on average than other genes differentially expressed in a similar number of conditions (**Figure 2.13B**).

Figure 2.12

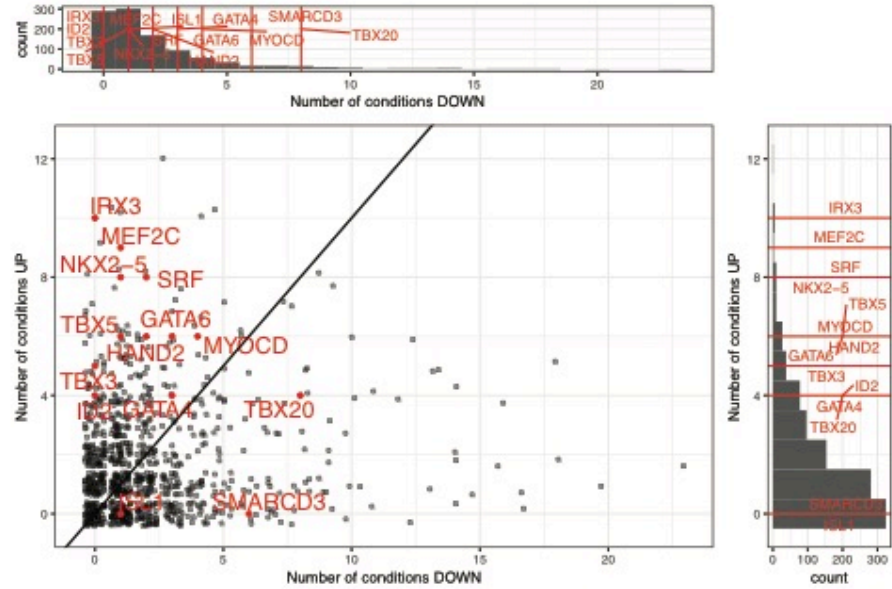


Figure 2.12: Perturbability of transcription factor genes in iCard-942. Only genes with average expression of 20 RPM or greater are shown. Pre-registered set of cardiomyocyte lineage-driving transcription factor marker genes in red. All non-marker transcription factor genes >20 RPM in grey. x-coordinate: number of drug conditions in which a gene is differentially expressed down; y-coordinate: number of drug conditions in which a gene is differentially expressed up.

Figure 2.13

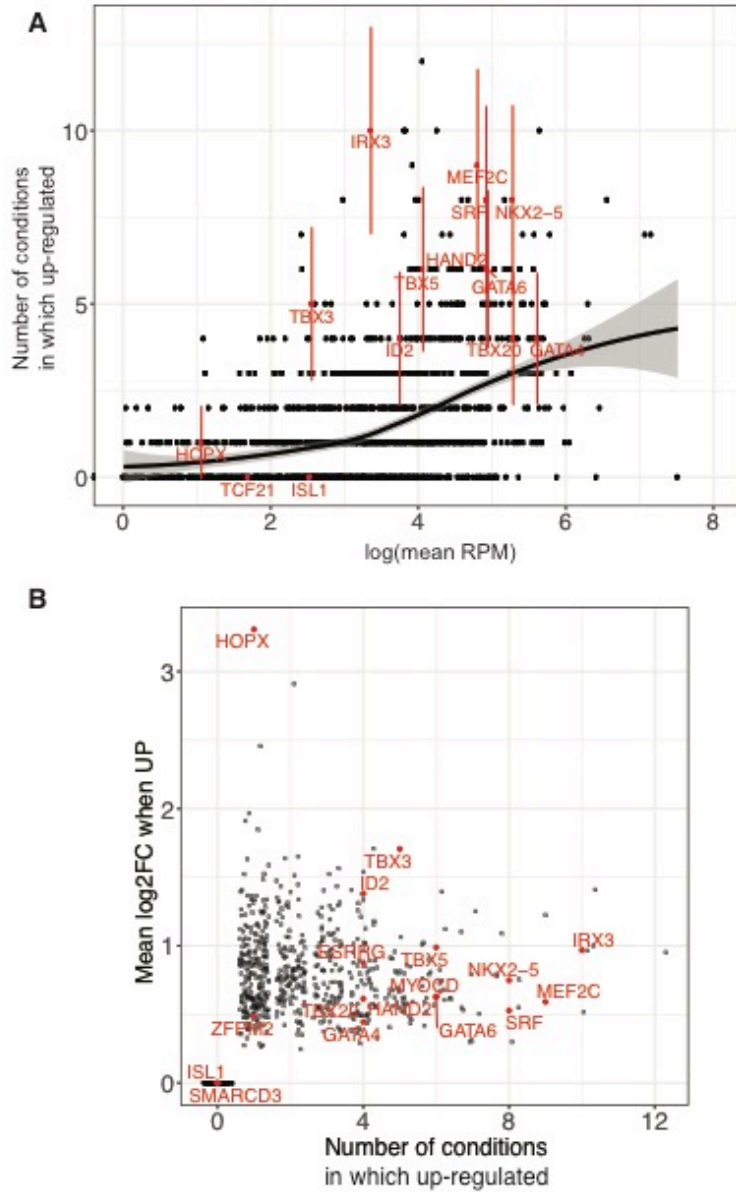


Figure 2.13: A) Number of conditions in which up-regulated vs. average expression in log(RPM). Error bars are standard deviation of 1000 bootstrap-resampled replicates. B) Average Log2(Fold change) per gene across all conditions in which that gene is up-regulated vs. number of conditions in which up-regulated.

## OVEREXPRESSION OF A KNOWN COCKTAIL OF TRANSCRIPTION FACTORS PERTURBABLE IN CARDIOMYOCYTES ENABLES CARDIAC TRANSDIFFERENTIATION OF FIBROBLASTS

We then wondered whether we could confirm that known lineage-driving transcription factors, most of which were highly perturbable, could induce transdifferentiation in the GM00942 dermal fibroblasts used for P<sup>3</sup> profiling. To do this we used pMXs retroviruses to overexpress a cocktail of 7 transcription factors (7F: GATA4, MEF2C, TBX5, MESP1, ESRRG, MYOCD, and ZFPM2) previously shown to transdifferentiate other human fibroblasts to a cardiomyocyte-like state.

A major challenge for this project was figuring out how to assess whether cells had entered a cardiomyocyte-like state after the process of transdifferentiation. I found that many published methods for such assessments, including cardiac Troponin immunofluorescence (IF), Troponin promoter-GFP reporters, and Troponin promoter-calcium transient reporters had too high background (IF) or too low signal (reporters) to be analyzed efficiently in thousands of mostly negative cells. These methods do work in iCard-942 cells, which are >95% positive. Eventually we developed a protocol to make post-transdifferentiation samples compatible with smFISH and its associated high-magnification fluorescence imaging (See **Appendix C**).

In order to use smFISH to assess transdifferentiation, we designed smFISH probe sets for markers specific to mature cardiomyocytes, NPPA and TNNT2. ([Mohamed et al., 2017](#)) These probes show high expression of both markers in iCard-942 cells and extremely low expression in cardiac fibroblasts (**Figure 2.14**).



Figure 2.14

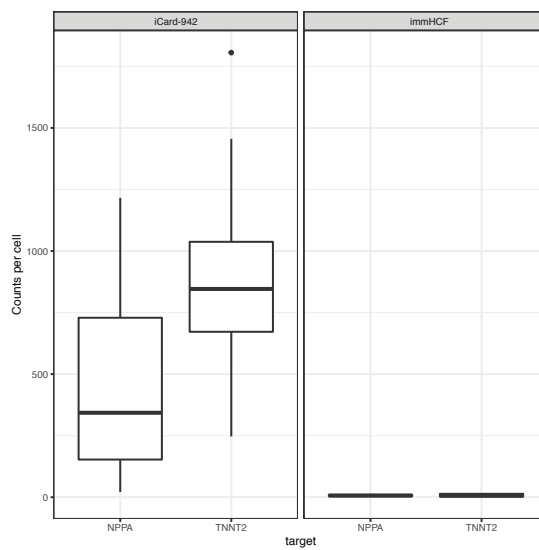


Figure 2.14: TNNT2 and NPPA smFISH counts per cell. Average of 33 cells per cell type. iCard-942 cells and immHCF, immortalized human cardiac fibroblasts.

We cultured 7F- and vehicle control-treated cells under progressively reduced serum conditions for 24 days following infection and at day 24 observed a small fraction of cells dramatically and concordantly up-regulate both NPPA and TNNT2 (**Figure 2.15**). Large-scale scans of approximately 1000 cells per sample suggested that approximately 1% of cells treated with 7F either marker gene expression above 25 copies per cell in a single focal plane at 60X magnification. 0 such cells were observed in vehicle controls (**Figure 2.16**). Therefore, we can conclude that some P<sup>3</sup>-identified frequently up-regulated transcription factors, i.e., members of 7F, were sufficient to induce cardiac transdifferentiation of cells of the genetic background in which they were identified.

We next tested whether 7F could induce transdifferentiation of fibroblasts from other sources, as well. Therefore we overexpressed 7F under similar conditions in two additional types of fibroblasts: an immortalized cardiac fibroblast cell line (immHCF) previously shown by one group to transdifferentiate with low efficiency upon 7F overexpression and primary cardiac fibroblasts (GM11169) that have not been studied in this context before. We observed rates of transdifferentiation in both of these cell lines comparable to the dermal fibroblasts used for P<sup>3</sup> analysis. This rate is consistent with published results in immHCF, as well (**Figure 2.17**). ([Mohamed et al., 2017](#)) Therefore, a cocktail of perturbable transcription factors identified in iCard-GM942 enable transdifferentiation to a cardiomyocyte-like state in fibroblasts of different genetic backgrounds, as well. This lends further support to the idea that this cocktail is generally sufficient for inducing cardiac transdifferentiation.

Figure 2.15

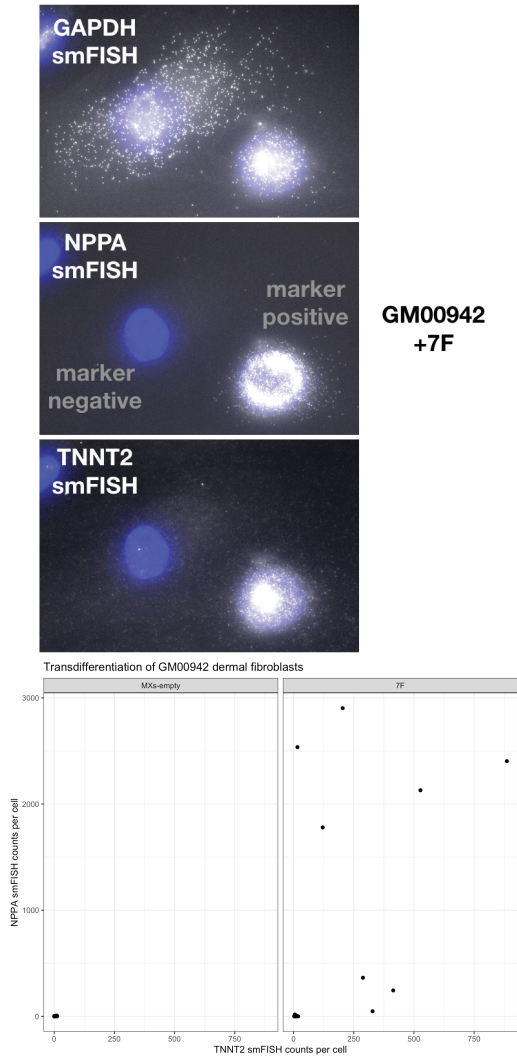


Figure 2.15: smFISH assessment of 7F-mediated transdifferentiation of GM00942 fibroblasts. (top) representative images of GAPDH, NPPA, and TNNT2 smFISH demonstrating a cell that expresses marker genes NPPA and TNNT2 at high levels and one cell that does not. (bottom) smFISH counts per cell of two marker genes for cells after receiving either vehicle control (MXs-empty) or 7F.

Figure 2.16

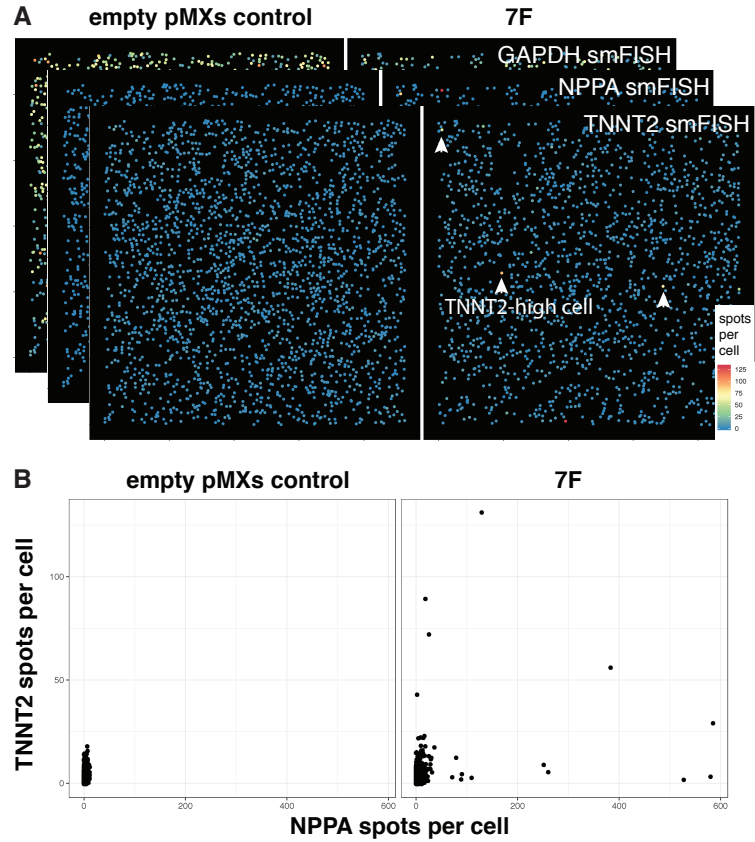


Figure 2.16: A) Reconstruction and per-gene quantification of smFISH scans of GM00942 cells exposed to vehicle control (MXs-empty; 2267 cells) or 7F (1693 cells). One z-slice per cell. Examples of cell with high expression of TNNT2 marked with white arrowheads. B) Comparison of single-cell TNNT2 and NPPA expression levels for all cells in smFISH scans.

Figure 2.17

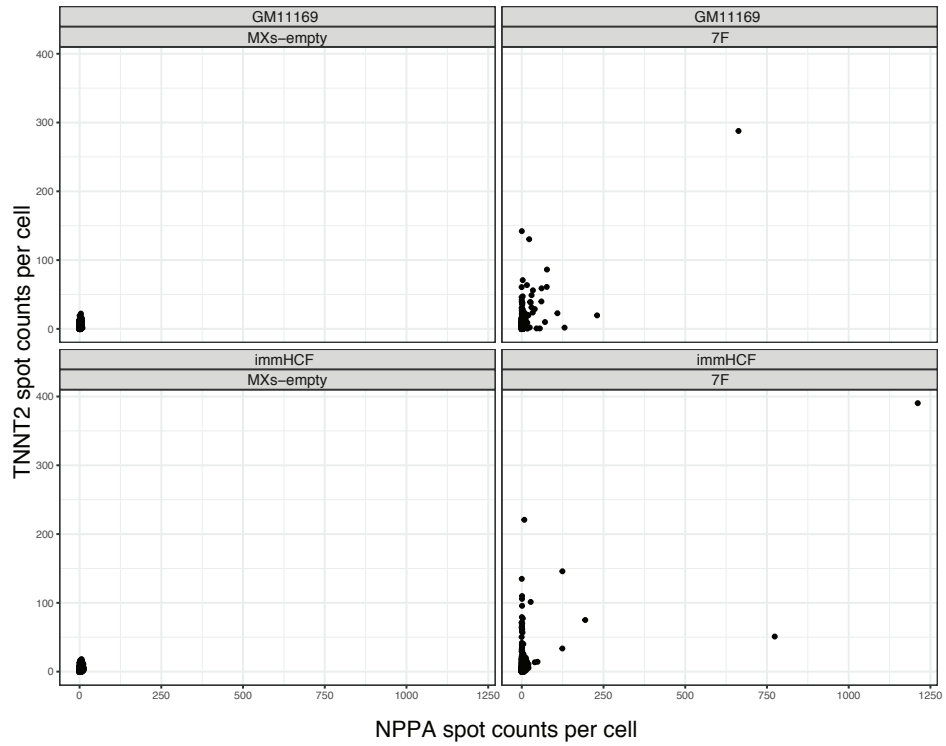


Figure 2.17: smFISH assessment of transdifferentiation of immHCF and GM11169 human cardiac fibroblasts after exposure to vehicle control (MXs-empty) or 7F. Average of 1151 +/- 51 cells per condition.

Next, we wondered whether any cocktail comprised of highly cardiomyocyte-perturbable transcription factors would be capable of inducing cardiac transdifferentiation of fibroblasts. Therefore we selected a set of seven transcription factors from the total list of 47 such factors that were highly perturbable and highly expressed in iCard-942 cells (7UP: SP3, ZBTB10, ZBTB44, NFIA, SSH2, ZNF770, and ZFP91. See **Appendix C** for details on how these factors were selected). We tested whether overexpression of this set of transcription factors, previously unstudied in the context of cardiac transdifferentiation, followed by culture in progressively reduced serum conditions would lead to up-regulation of cardiomyocyte markers. We did not observe a dramatic increase in TNNT2 and NPPA expression after overexpression of 7UPs in fibroblasts.

## KNOCKDOWN OF TRANSCRIPTION FACTORS THAT ARE PERTURBABLE IN FIBROBLASTS OFTEN ENHANCES FIBROBLAST REPROGRAMMING TO IPSC

We wondered whether we could use perturbation panel profiling to identify transcription factors that are presumably important for driving or maintaining another cell lineage, beyond those for cardiomyocytes. Therefore, we exposed GM00942 fibroblasts in culture to the same perturbation panel in culture as GM942-iCards and performed RNAtag-seq after several days of perturbation. We observed similar numbers of detectably differentially expressed genes in fibroblasts as in iCards across the perturbation panel (**Figure 2.9**).

Our first question was whether the same genes are differentially expressed in each condition in both iCards and fibroblasts. We found little overlap in the specific genes that were detectably differentially expressed in response to the same drug in each cell type (**Figure 2.18**).

Figure 2.18

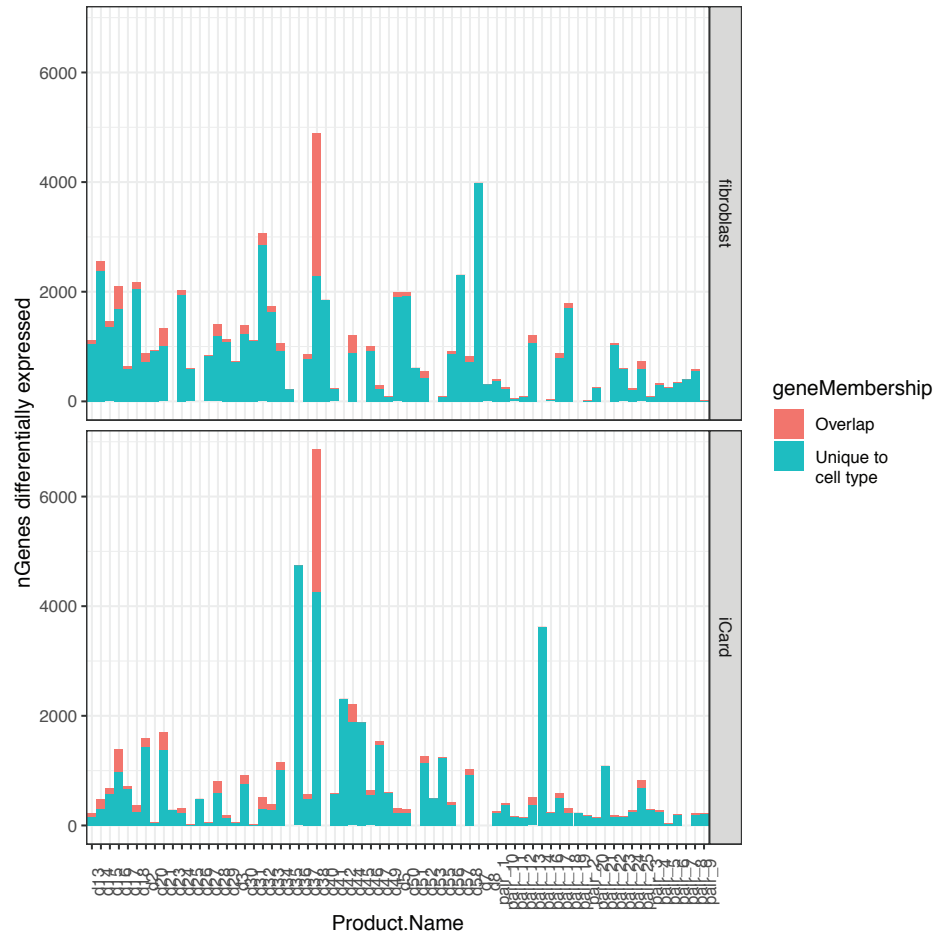


Figure 2.18: Overlap between differentially expressed genes per drug across GM00942 fibroblasts and iCard-942.

Since we found that cardiomyocyte lineage-driving transcription factor genes are frequently up-regulated in perturbed cardiomyocytes, we wondered whether transcription factors that are up-regulated in fibroblasts across many perturbation conditions could be useful in directed changes of fibroblasts to a new cell identity. Specifically, our hypothesis was that genes frequently up-regulated in fibroblasts may be useful for fibroblast identity maintenance and therefore their suppression could make fibroblasts more “reprogrammable”. We chose 16 transcription factors that were frequently up-regulated in fibroblasts and used shRNAs to knock down their expression levels in hiF-T cells prior to doxycycline-inducible Yamanaka Factor (Oct4, Sox2, Klf4, Myc; OSKM) over-expression (**Figure 2.19**). Of these 16 transcription factors, we found that at least 8 reproducibly increased the frequency of fibroblast reprogramming, as measured by counting Alkaline Phosphatase-positive colonies after 3 weeks of OSKM induction (**Figure 2.20**). This success rate, 8/16, is comparable to the validation rate of a recent pooled shRNA screen of epigenetic regulators of hiF-T reprogramming ([Cacchiarelli et al., 2015](#)), 9/23. Overall, this suggests that transcription factor perturbability is useful for identifying genes useful for directed changes of fibroblast identity, as well.

In summary, we developed P<sup>3</sup> for profiling gene expression differences in cell types of interest following a broad panel of small molecule perturbations. We found that known cardiomyocyte lineage-driving genes are frequently up-regulated after perturbation in cardiomyocytes. Further, we demonstrated that a known cocktail of cardiomyocyte-perturbable transcription factors enables transdifferentiation of multiple types of human fibroblasts. Additionally, we extended perturbability to the identification of fibroblast-perturbable factors, the knockdown of which improves the efficiency of fibroblast reprogramming to pluripotency. Overall, these results suggest that the high perturbability of lineage-driving transcription factors can be used to prioritize factors to include in protocols for directed changes of cellular identity.



Figure 2.19

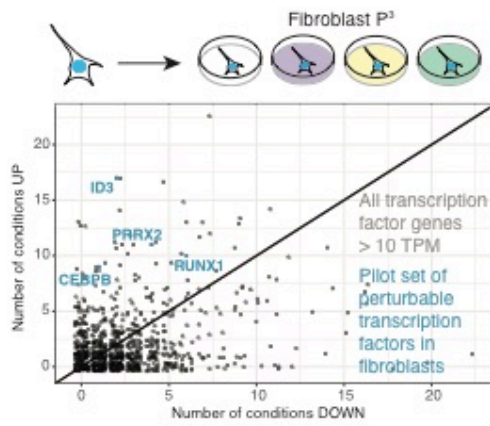


Figure 2.19: P<sup>3</sup> profiling of GM00942. All TFs with average expression > 10TPM shown in grey. Candidate fibroblast-perturbable factors in blue.

Figure 2.20

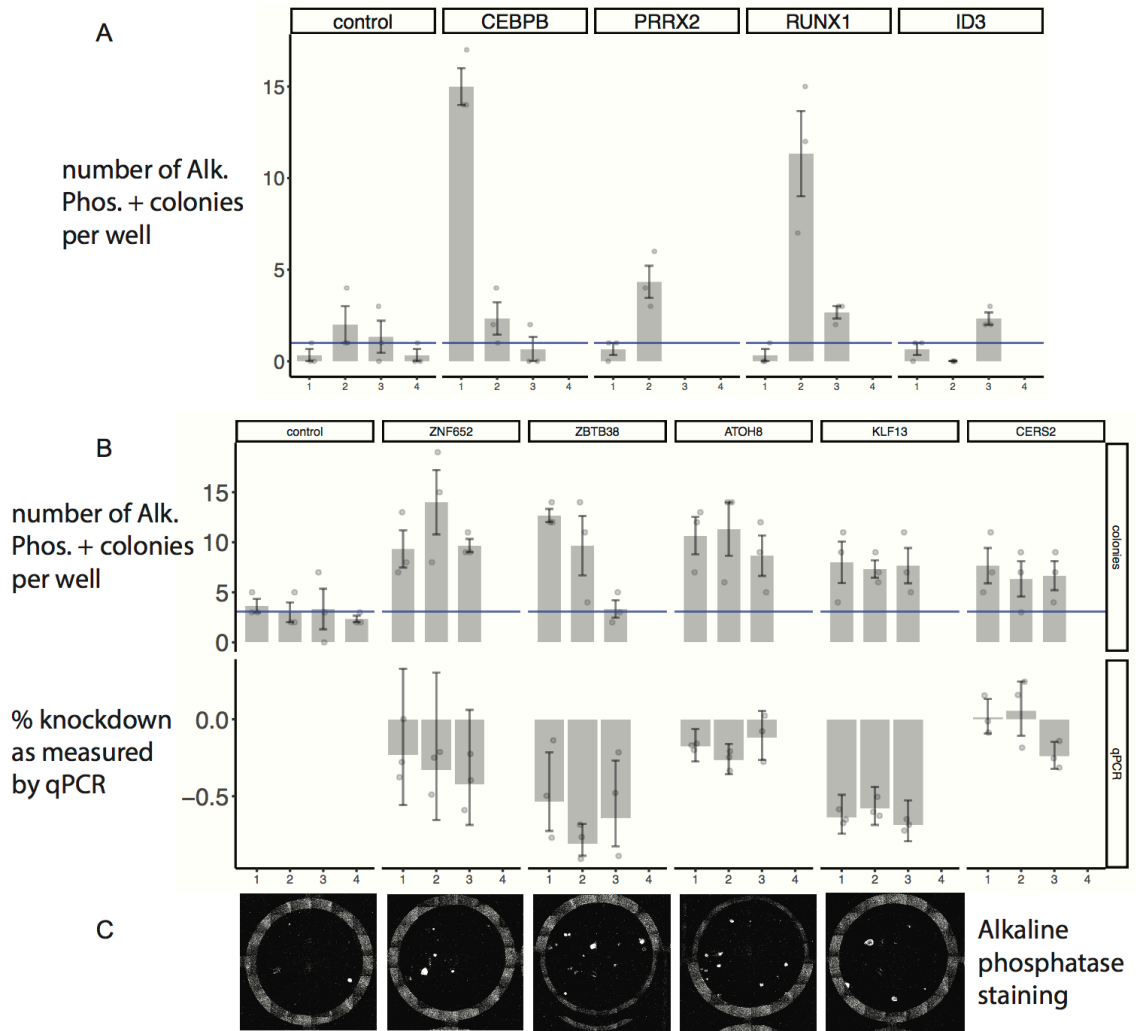


Figure 2.20: iPSC reprogramming efficiency of hiF-T cells after knockdown of fibroblast-perturbable factors. Control shRNAs (1 = empty backbone, 2 = scrambled negative control, 3 = GFP reporter, 4 = Luciferase-targeting negative control). (A,B) Controls shown in same row as factors tested are matched per batch. Alkaline Phosphatase-positive colony counts per well counted in triplicate per shRNA. Three shRNAs per factor tested. Representative RT-qPCR based assessment of knockdown efficiency. (C) Representative alkaline phosphatase staining.

## CONCLUSIONS AND FUTURE DIRECTIONS

The components of this thesis represent some of the many different ways in which we can build tools, design experiments, and construct analytical frameworks for quantitative models of gene regulatory processes and the effects of these processes on cellular phenotypes. I focused on two disparate topics of interest to me: RNA editing and cellular identity. While different, these two topics are both grounded in generating cellular diversity through regulation of a common genome.

In the first chapter I developed a relatively general method for visualizing and quantifying A-to-I RNA editing of single mRNA molecules with subcellular resolution. Prior to this work, no one had actually “seen” an edited mRNA before. The state of the art for detecting A-to-I RNA editing involved RT-PCR-based methods, such as restriction digestion of PCR products or RNA-seq, which sacrifice resolution in both single-cell quantification and localization. ([Higuchi et al., 2000](#); [Liddicoat et al., 2015](#)) Alternatively, related methods included standard smFISH with immunofluorescence for enzymes involved in the process of RNA editing, which visualize a transcript's physical association with an enzyme rather than its sequence identity (i.e., whether it is actually edited). ([Jepson et al., 2011](#)) Ultimately, in the case of GRIA2, the field required a more precise tool for studying this sequence-level difference in single cells.

Now we have seen edited mRNAs and their unedited cohorts in single human cells. With inoFISH, we can measure differences in localization and quantities of edited and unedited transcripts arising from the same gene within individual cells *in situ*. Using this new technique, we observed nuclear and cytoplasmic localization of both edited and unedited transcripts in the same cells. For the three genes and two cell lines studied, editing levels were not significantly different between these two cellular compartments. inoFISH also revealed high physical association of

GRIA2 with nuclear paraspeckles and that this association was possible for both edited and unedited transcripts. Additionally, we demonstrated that our method could discriminate the editing status of transcripts at transcription sites, and that we only observed unedited NUP43 transcripts at their transcription sites (though we of course could not demonstrate a conclusive absence of edited transcripts at these sites). Lastly, we showed that the distribution of single-cell editing levels can differ by gene: NUP43 editing levels are homogeneous across a population of cells, while GRIA2 editing levels are more variable.

In a broader context, inoFISH enables testing of hypotheses about whether single-cell RNA editing levels are correlated with single-cell-level phenotypes, such as permeability to divalent cations or cell death. In combination with recent advances in tissue clearing strategies for related smFISH methods, inoFISH might be used to visualize the editing status of GRIA2 in human brains with and without ALS lesions. ([Symmons et al., 2019](#))

Furthermore, inoFISH is a proof of principle that smFISH-based methods could be used to discriminate the editing status of transcripts in single cells in general. For other interesting types of RNA editing, such as pseudouridylation, we can imagine designing related techniques for single-molecule resolution of editing status at sites of interest. One could imagine combining established techniques for selective chemical labeling of pseudouridine bases by CMC derivatization with a new click chemistry-based technique for amplified smFISH signal to specifically hybridize clampFISH probes to edited transcripts. ([Rouhanifard et al., 2018](#); [Sakurai et al., 2014](#))

In the second chapter of this thesis, I developed an experimental and analytical pipeline for identifying transcription factors to be used in directed changes of cellular identity based on their expression levels following a broad panel of perturbations (P<sup>3</sup>). We used a drug-specific dosing scheme that allowed us to choose drugs and deliver them at a dose that would reliably perturb, but not kill, the cells of interest. Our IC50-based dosing scheme is consistent with observations of transcriptional responses after dosing curves of small molecules in other human cell lines, as well. (Srivastan et al., unpublished)

We validated our approach by showing that we could re-identify members of the previously characterized 7F transcription factor cocktail sufficient for cardiac transdifferentiation of fibroblasts. [\(Fu et al., 2013\)](#) In particular, we showed that transcription factors that drive a cellular identity, in this case cardiomyocyte, are highly “perturbable” in cardiomyocytes. Prior to collecting these results we had hypothesized that the opposite would be true, i.e., that lineage-driving factors would be less perturbable than other similarly expressed genes. Nonetheless, members of 7F are often up-regulated after several kinds of small molecule perturbation, whereas other highly-expressed transcription factors are not as frequently up-regulated by many kinds of perturbations. Following up on this, we developed a new smFISH-based assessment of cardiac transdifferentiation efficiency and applied it to verify 7F-based transdifferentiation in three types of human fibroblasts.

We then extended the discovered association between cardiomyocyte lineage-driving capacity of transcription factors and high perturbability to rational manipulation of fibroblast cell identity. That is, we identified transcription factors based on their high perturbability in fibroblasts and hypothesized that for them the inverse of 7F would be true. If 7F overexpression drove cells toward cardiomyocyte identity, then maybe fibroblast-perturbable factor knockdown could drive

cells away from fibroblast identity. Therefore, we tested knockdown of 16 fibroblast-perturbable factors in fibroblasts prior to iPSC reprogramming and observed that 8 of them enhanced the efficiency of conversion to pluripotency. This success rate, 8/16, is comparable to the validation rate of the top hits from a pooled screen of reprogramming efficiency enhancement by hundreds of epigenetic modifiers conducted in the same cell line (9/23 tested; [Cacchiarelli et al., 2015](#)).

Overall, these results provide insight into both practical and theoretical aspects of the regulation of cellular identity from two perspectives. First, from a practical, engineering point of view, the P<sup>3</sup> experimental design and analysis pipeline could be useful for prioritizing the sets of transcription factors to overexpress or suppress when developing a new transdifferentiation or reprogramming protocol. Second, theoretically, we can infer from these results that for maintenance of cell identity in general, it may be that, in some way, it is useful to reinforce the expression of particular transcription factors to stabilize larger gene expression programs. Intriguing work in yeast suggests that in new environments cells can stochastically tune and then reinforce the expression of fitness-enhancing genes.[\(Freddolino et al., 2018\)](#) Extended to a given cell identity in its niche in the body, fitness enhancement could come from best fulfilling its needed role in that niche.[\(Arendt et al., 2016\)](#)

Alternatively, the observation of perturbability in P<sup>3</sup> experiments might just be coincidental (and convenient for engineering) rather than *per se* functional. If, for example, lineage-driving transcription factors are downstream of a larger number of signal transduction pathways in a cell type than non-lineage-driving transcription factors, then perhaps there are therefore more ways to elicit a differential expression response in these genes.

In the future, I hope that P<sup>3</sup> proves to be useful not only in developing more efficient protocols for the manipulation of fibroblast and cardiomyocyte cellular identity, but also protocols for engineering any cell type of interest. Additionally, I look forward to comparing the results of P<sup>3</sup> with orthogonal methods for the discovery of factors that enhance directed changes of cellular identity. For instance, in the Raj lab we have already started an exciting new project, which combines cellular barcoding and a form of RNA FISH to retrospectively isolate the rare fibroblasts that go on to successfully reprogram to iPSC after Yamanaka factor induction. It will be interesting to see if fibroblast-perturbable factors are lower in cells primed to reprogram relative to cells that are destined to fail to reprogram.

In summary, inoFISH and P<sup>3</sup> are two projects that focused on building quantitative models of gene regulation, each at different scales of biological organization. As such, we needed to overcome different types of challenges to complete them. In the case of inoFISH, we had a straightforward, traditional small-scale model of single-cell variability for a small number of interacting molecules. In the case of P<sup>3</sup>, however, we were faced with a more complicated system: the set of all transcription factors. Therefore, we developed an analytical framework that was not just like an inoFISH-scale model but 1000-times larger. Through the lens of perturbability we were able to separate out lineage-driving transcription factors from others of similarly high expression level, and to use them in directed changes of cellular identity. Hopefully these projects, considered side-by-side, together demonstrate the utility of diverse methods, experimental designs, and models in the systems biology of gene regulation as this field develops in a world with access to ever higher-throughput and more precise genetic and biochemical techniques. ([Mellis and Raj, 2015](#))

# APPENDIX A: STATISTICAL ANALYSIS OF RNA EDITING LEVELS BASED ON INOFISH EXPERIMENTAL RESULTS

Please note that much of this analysis is based on the Supplementary Note in Levesque, et al., 2013. This analysis originally appeared in the Supplementary Note of Mellis et al., 2017.

## 1 DETECTING THE PRESENCE OF ANY RNA EDITING

The most basic question we can answer with inoFISH is whether or not there is any editing at a given site of interest in a given transcript of interest.

Consider an experiment with  $T$  total transcripts of a species of interest expressed across all  $N$  cells observed. Of these  $T$  transcripts, we observe  $O_j$  colocalizing with an Inosine detection probe spot. Upon pixel-shift analysis, we see that the probability of guide colocalization with an inosine detection probe spot purely by chance is  $S_j$  (range 0 - 1).

For simplicity, we will also assume that the editing status detection label (only false-positive here) of each transcript (of all  $T$  transcripts) in an inoFISH experiment is independent of the editing status label of any other transcript.

Therefore, we can model  $O_j$  as being drawn from a binomial distribution with parameters  $S_j$  and  $T$ . Our null hypothesis is that any observed colocalizations are false-positive colocalizations. The probability of observing at least  $O_j=x$  spots under the null hypothesis is:

$$P(O_j \geq x | S_j, T) = \sum_{j=x}^T Tj \cdot S_j^j (1-S_j)^{T-j} \quad (1)$$

## 2 MODELING RNA EDITING LEVELS IN A SINGLE EXPERIMENT



The next level of detail we might want from an inoFISH experiment is a quantitative estimate of the editing level of a given site of interest in a given transcript of interest. That is, what fraction of these transcripts are edited at this site?

Consider an experiment with  $T$  total transcripts expressed across all  $N$  cells observed. We observe some number  $O_I$  of them to be labelled Inosine, i.e., edited, and some fraction  $O_A$  labelled Adenosine, i.e., unedited. The remaining  $U$  transcripts are unlabelled.

## 2.1 GOAL: ESTIMATING EDITING LEVEL

We will model the editing level  $E$  (range 0 - 1), which is the probability that a transcript in this population is in fact edited at this point in time. The  $T$  total transcripts are composed of  $T_I$  edited transcripts and  $T_A$  unedited transcripts, where  $T_I + T_A = T$ . The editing level of the transcript of interest at the site of interest is:

$$E = \frac{T_I}{T} \quad (2)$$

The goal of our analysis, therefore, is to estimate  $E$ . Here we outline a framework for calculating a MLE for  $\hat{E}$ , including upper and lower bounds on the confidence interval for  $\hat{E}$ .

## 2.2 MODELING INOFISH DETECTION EFFICIENCY

Since we don't directly have access to  $T_I$  in an experiment, we need some other way to estimate the fraction in Eq. 2. As outlined above, inoFISH gives us the ability to observe some transcripts labelled as unedited ( $O_A$ ) and some transcripts labelled as edited ( $O_I$ ). We also know our expected false-positive colocalization rate for each detection probe through pixel-shift analysis. The fraction of all  $T$  spots colocalizing with Adenosine detection probes after pixel-shift is  $S_A$ . The fraction of all  $T$  spots colocalizing with Inosine detection probes after pixel-shift is  $S_I$ . We will exploit our measurements of  $T, O_A, O_I, S_A$ , and  $S_I$  to estimate  $E$ .

Before we do that, however, we have to make assumptions about  $O_A$  and  $O_I$ . Specifically, we need to make an educated guess about whether they equally scale with respect to the true values they are designed to measure ( $T_A$  and  $T_I$ ). Formally, we will make assumptions about the probe-specific true-positive detection efficiencies, which we will denote as  $d_A^{true}$  and  $d_I^{true}$ .  $d_A^{true}$  is the fraction of all  $T_A$  unedited transcripts correctly labelled as Adenosine. Similarly,  $d_I^{true}$  is the fraction of all  $T_I$  edited transcripts correctly labelled as Inosine.

### 2.2.1 A QUICK ILLUSTRATION OF DETECTION EFFICIENCY

Note that  $d_A^{true}$  and  $d_I^{true}$  may be equal while  $O_A$  and  $O_I$  are not (even if  $S_A=S_I$ , as well). For example, consider an experiment in which  $T=100$  total transcripts, of which  $T_A=90$  and  $T_I=10$  and  $d_A^{true}=d_I^{true}=10\%$ . If  $S_A=S_I=2\%$ , then we would expect to observe:

$$\text{False-positive Adenosine labels} = S_A T_A + S_I T_I = 0.02(90) + 0.02(10) = 2$$

$$\text{True-positive Adenosine labels} = d_A^{true} T_A = 0.1(90) = 9$$

$$\text{False-positive Inosine labels} = S_I T_A + S_I T_I = 0.02(90) + 0.02(10) = 2$$

$$\text{True-positive Inosine labels} = d_I^{true} T_I = 0.1(10) = 1$$

Which gives  $O_A=2+9=11$  and  $O_I=2+1=3$ .

### 2.2.2 EMPIRICAL ESTIMATES OF DETECTION EFFICIENCY

In an inoFISH experiment, the fraction of guide spots colocalizing with an Adenosine or Inosine detection probe spot is the total detection efficiency,  $d^{total}$ .

$$d^{total} = \frac{O_A + O_I}{T}$$

We can decompose the detection efficiency into true-positive and false-positive signal. The false-positive colocalization events are entirely estimated by pixel-shift analysis:

$$\begin{aligned} d^{total} &= d^{true} + d^{false} \\ &= d^{true} + S_A + S_I \end{aligned}$$

Therefore, in any experiment, we can estimate an overall fraction of likely true-positive colocalization events:

$$\hat{d}^{true} = \frac{O_A + O_I}{T} - (S_A + S_I)$$

This total true-positive fraction can itself be decomposed into Adenosine and Inosine fractions:

$$d^{true} = \frac{d_A^{true} T_A + d_I^{true} T_I}{T}$$

If  $d_A^{true} = d_I^{true} = d$ ,

$$d^{true} = \frac{d_A^{true} T_A + d_I^{true} T_I}{T}$$

$$d^{true} = \frac{d(T_A + T_I)}{T}$$

$$d^{true} = \frac{d(T)}{T}$$

$$\Rightarrow d = \hat{d}^{true}$$

If  $d_A^{true} \neq d_I^{true} = d$ , the estimation of editing level is a bit more complicated (as we will explore below).

In the case of the inoFISH-like method SNP FISH (Levesque et al., 2013), which allows for visualization of genetically encoded single-nucleotide variation in transcripts, we can directly estimate individual detection probes' detection efficiencies with homozygous genetic controls.

However, since RNA editing is a post-transcriptional regulatory process, we know of no way to ensure that all endogenously-transcribed copies of a transcript of interest are edited. This limits our ability to directly measure  $d_I^{true}$  alone. Similarly, we cannot directly measure  $d_A^{true}$ ; ADAR protein family members, which catalyze RNA editing, often have redundant targets and knockout of some family members is lethal. Hence, we also cannot directly measure the detection efficiency of the unedited detection probe on endogenously-transcribed RNAs that are all surely unedited.

In experiments evaluating genetically-encoded single-nucleotide variation in transcripts, rather than post-transcriptional single-nucleotide variation, we can design simple panels of controls to systematically quantify detection efficiencies for each detection probe. In our lab we have observed detection efficiencies of detection probes in such control experiments ranging from 5% to 60%. Further, in the majority of probe sets we have examined, the two detection probes have equal detection efficiencies. In an inoFISH experiment, equal detection efficiencies would mean that  $d_A^{true} = d_I^{true}$ . In the few cases where detection efficiencies differ between a probe set's two detection probes, we have not observed any differing by more than about 8%.

We will present a statistical framework for calculating  $\hat{E}$  and its confidence interval under different assumptions about the detection efficiencies of inoFISH detection probes.

### 2.3 MODELING INOFISH EXPERIMENTS

Our general strategy for finding a MLE and confidence interval bounds on  $\hat{E}$  will be to model inoFISH experiments based on our observations. From this model we can solve for a MLE and computationally simulate similar experimental observations to draw bounds on the confidence interval for  $\hat{E}$ .

### 2.3.1 MODELING INOFISH WITH A MULTINOMIAL

In order to find a MLE, we need a functional form for data likelihood. For simplicity, we will assume that the editing status of each transcript (of all  $T$  transcripts) in an inoFISH experiment is independent of the editing status of any other transcript's editing status. Further, we will assume that whether or not any transcript is detected at all is independent the detection of any other transcript. Therefore, we can model an inoFISH experiment as  $T$  transcripts drawn from a multinomial with probabilities as follows:

$$P(O_A=x, O_I=y|T) = \text{multinomial}(\alpha, \beta, \gamma)$$

Where for any individual transcript  $r \in T$ ,

$$\alpha = P(r \in O_A)$$

$$\beta = P(r \in O_I)$$

$$\gamma = P(r \in U)$$

### 2.4 POPULATION-WIDE PARAMETER ESTIMATION

Consider an experiment, as above, with the following findings: Across  $N$  cells, there are  $T$  total transcripts of interest, of which  $O_A=x$  are observed unedited and  $O_I=y$  are observed edited. Upon pixel-shift analysis, we see that the probability of false-positive colocalization with Adenosine

detection probes is  $S_A = s_A$  and the probability of false-positive colocalization with Inosine

detection probes is  $S_I = s_I$ . With knowledge of  $T, O_A, O_I, S_A,$  and  $S_I$ , we can also calculate  $d_A^{true}$

and  $d_I^{true}$ , as above.

By the law of conditional probability, since  $P(r \in T_A) + P(r \in T_I) = 1$  for any transcript  $r$ ,

$$\begin{aligned} P(r \in O_A) &= P(r \in O_A | r \in T_A)P(r \in T_A) + P(r \in O_A | r \in T_I)P(r \in T_I) \\ &= -d_A^{true} E + d_A^{true} + s_A \end{aligned}$$

$$\begin{aligned} P(r \in O_I) &= P(r \in O_I | r \in T_I)P(r \in T_I) + P(r \in O_I | r \in T_A)P(r \in T_A) \\ &= d_I^{true} E + s_I \end{aligned}$$

These will be two of our three probability parameters for the multinomial distribution simulating an inoFISH experiment:

$$P(r \in O_A) = -d_A^{true} E + d_A^{true} + s_A = \alpha$$

$$P(r \in O_I) = d_I^{true} E + s_I = \beta$$

All  $U = T - O_A - O_I$  transcripts not labelled as either Adenosine or Inosine are undetected, so:

$$P(r \in U) = 1 - \alpha - \beta$$

$$= (d_I^{true} - d_A^{true})E - d_A^{true} - s_A - s_I + 1 = \gamma$$

This means that we can model inoFISH experiments that are consistent with our observations as follows:

$$P(O_A=x, O_I=y|T) = \text{multinomial}(\alpha, \beta, \gamma)$$

$$= \frac{T!}{x!y!(T-x-y)!} \alpha^x \beta^y \gamma^{T-x-y}$$

We can find the maximum likelihood estimate of E by differentiating the data likelihood with respect to E and finding the root. Further, since the logarithm is a monotonically increasing function, we can alternatively solve for the maximum of the log-likelihood (which is simpler in this case).

$$0 = \frac{\partial}{\partial E} \ln \left( \frac{T!}{x!y!(T-x-y)!} \alpha^x \beta^y \gamma^{T-x-y} \right)$$

$$0 = \frac{x(-d_A^{true})}{\alpha} + \frac{y(d_I^{true})}{\beta} + \frac{(T-x-y)(d_I^{true} - d_A^{true})}{\gamma}$$

From here, we can solve for  $\hat{E} \in [0, 1]$ .

#### 2.4.1 EQUAL DETECTION EFFICIENCIES

If  $d_A^{true} = d_I^{true} = d^{true}$ , as is usually the case for probe sets of the sort used in inoFISH, the functional form of  $\hat{E}$  is rather simple:

$$0 = \frac{-x d^{true}}{\alpha} + \frac{y d^{true}}{\beta} + \frac{(T-x-y)(0)}{\gamma}$$

$$\frac{x d^{true}}{\alpha} = \frac{y d^{true}}{\beta}$$

$$x(d^{true}E+s_I) = y(-d_A^{true}E+d_A^{true}+s_A)$$

$$\hat{E} = \frac{y(d^{true}+s_A)-xs_I}{(xd^{true}+yd^{true})} \quad (3)$$

This result (Eq. 3) makes intuitive sense: it is the fraction of all labelled transcripts observed to be inosine, correcting for false-positives.

For example, consider a simple experiment in which we have the following observations:

$$T=100, O_A=15, O_I=25$$

Upon pixel-shift analysis, we get:

$$S_A=0.03, S_I=0.03$$

As in section 2.2.2, if we assume equal detection efficiencies we get:

$$\hat{d}^{true} = \frac{O_A+O_I}{T} - (S_A+S_I)$$

$$\hat{d}^{true} = \frac{15+25}{100} - (0.03+0.03) = 0.34$$

Therefore, in this experiment we get an estimated editing level:

$$\begin{aligned} \hat{E} &= \frac{y(d^{true}+s_A)-xs_I}{(xd^{true}+yd^{true})} \\ &= 0.647 \end{aligned}$$

Note that on first glance the results might suggest an editing level estimate of about  $25/(15+25)=0.625$ . The MLE, however, accounts for false-positive colocalizations as measured by



pixel-shift for each detection probe channel. In this case, there are the same number of expected false-positive colocalizations for each detection probe. When these are subtracted out in the MLE, the proportional difference between  $O_A$  and  $O_I$  is accentuated, giving a slightly higher MLE for editing level ( $0.647 > 0.625$ ).

#### 2.4.2 UNEQUAL DETECTION EFFICIENCIES

If  $d_A^{true} \neq d_I^{true}$ , as has been observed in some cases of probe sets like those used in inoFISH, we can still solve for a maximum likelihood estimate of  $E$ , but the functional form is not as compact.

$$\begin{aligned}
0 &= \frac{x(-d_A^{true})}{\alpha} + \frac{y(d_I^{true})}{\beta} + \frac{(T-x-y)(d_I^{true}-d_A^{true})}{\gamma} \\
0 &= \frac{x(-d_A^{true})}{-d_A^{true}E+d_A^{true}+s_A} + \frac{y(d_I^{true})}{d_I^{true}E+s_I} + \frac{(T-x-y)(d_I^{true}-d_A^{true})}{(d_I^{true}-d_A^{true})E-d_A^{true}-s_A-s_I+1} \\
0 &= x(-d_A^{true})(d_I^{true}E+s_I)((d_I^{true}-d_A^{true})E-d_A^{true}-s_A-s_I+1)+ \\
&\quad y(d_I^{true})(-d_A^{true}E+d_A^{true}+s_A)((d_I^{true}-d_A^{true})E-d_A^{true}-s_A-s_I+1)+ \\
&\quad (T-x-y)(d_I^{true}-d_A^{true})(-d_A^{true}E+d_A^{true}+s_A)(d_I^{true}E+s_I) \\
0 &= aE^2+bE+c, \text{ where} \\
a &= d_I^{true}(-T+x+y) \left(-d_A^{true}+d_I^{true}\right) d_A^{true}-x d_I^{true} \left(d_A^{true}\right)^2+x \left(d_I^{true}\right)^2 d_A^{true} \\
&\quad -y d_I^{true} \left(d_A^{true}\right)^2+y \left(d_I^{true}\right)^2 d_A^{true}
\end{aligned}$$

$$\begin{aligned}
b = & -y s_A (d_I^{true})^2 + s_A d_I^{true} (T-x-y) (-d_A^{true} + d_I^{true}) - x d_I^{true} d_A^{true} - y d_I^{true} d_A^{true} + \\
& 2 s_I x d_I^{true} d_A^{true} + s_I y d_I^{true} d_A^{true} + x s_A d_I^{true} d_A^{true} + 2 y s_A d_I^{true} d_A^{true} - y (d_I^{true})^2 d_A^{true} + \\
& s_I (-T+x+y) (-d_A^{true} + d_I^{true}) d_A^{true} + d_I^{true} (T-x-y) (-d_A^{true} + d_I^{true}) d_A^{true} - \\
& s_I x (d_A^{true})^2 + x d_I^{true} (d_A^{true})^2 + 2 y d_I^{true} (d_A^{true})^2 \\
c = & y s_A d_I^{true} - s_I y s_A d_I^{true} - y s_A^2 d_I^{true} + s_I s_A (T-x-y) (-d_A^{true} + d_I^{true}) - s_I x d_A^{true} + \\
& s_I^2 x d_A^{true} + s_I x s_A d_A^{true} + y d_I^{true} d_A^{true} - s_I y d_I^{true} d_A^{true} - 2 y s_A d_I^{true} d_A^{true} + \\
& s_I (T-x-y) (-d_A^{true} + d_I^{true}) d_A^{true} + s_I x (d_A^{true})^2 - y d_I^{true} (d_A^{true})^2
\end{aligned}$$

The root that falls within [0,1] will be the physically realizable solution (our  $\hat{E}$ ). The quadratic formula gives us the roots:

$$\hat{E} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (4)$$

## 2.5 PERFORMING SIMULATED DETECTION EXPERIMENTS

For a given experiment, we know  $T, O_A, O_I, S_A,$  and  $S_I$ . As discussed above, from here we can model our experiment with a multinomial distribution. We can simulate the population-wide results, i.e.,  $O_A=x, O_I=y, U=T-x-y$ , of one such inoFISH experiment by drawing from the multinomial above according to:

$$P(O_A=x, O_I=y | T) = \text{multinomial}(\alpha, \beta, \gamma)$$

After drawing  $O_A=x, O_I=y, U=T-x-y$ , we can re-calculate  $\hat{E}$  (call this  $\hat{E}'$ ) using the method above.

## 2.6 PARAMETRIC BOOTSTRAPPING A CONFIDENCE INTERVAL FOR $\hat{E}$

We can re-sample from this multinomial distribution many times (e.g., 100000 draws) and calculate the range of values of  $\hat{E}'$  seen in 95% of the samples. This method is known as parametric bootstrapping.

The 2.5%ile of  $\hat{E}'$  is the lower bound  $\hat{E}_{LB}$  on the 95% CI for  $\hat{E}$ . The 97.5%ile of  $\hat{E}'$  is the upper bound  $\hat{E}_{UB}$  on the 95% CI for  $\hat{E}$ .

## 3 MODELING DIFFERENCES BETWEEN RNA EDITING LEVELS

Consider two different inoFISH experiments with sets of population-wide findings

$\{T^{(1)}, O_A^{(1)}, O_I^{(1)}, S_A^{(1)}, S_I^{(1)}\}$  and  $\{T^{(2)}, O_A^{(2)}, O_I^{(2)}, S_A^{(2)}, S_I^{(2)}\}$ . Are the estimated editing levels,  $\hat{E}^{(1)}$  and  $\hat{E}^{(2)}$ , observed in these two experiments different from one another? We can parametrically bootstrap this, too, as follows.

1. Model both experiments with multinomials, as above. Call these distributions  $M^{(1)}$  and  $M^{(2)}$ .
2. Draw a sample from  $M^{(1)}$  and calculate  $\hat{E}^{(1)}$
3. Draw a sample from  $M^{(2)}$  and calculate  $\hat{E}^{(2)}$
4. Calculate  $\hat{E}^{(1)} - \hat{E}^{(2)} = \delta'$ . Store this result in the vector  $\Delta'$ .
5. Repeat steps 2-4 100000 times.

To check the significance level of your favorite difference  $a$ , calculate the fraction of entries  $\delta' \in \Delta'$  that satisfy  $\delta' \geq a$ .

### 3.1 BOOTSTRAPPING A CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN EDITING LEVELS

Alternatively, one could just report the 95% CI for the magnitude of the difference  $\Delta$  between editing rates between the experiments modeled by  $M^{(1)}$  and  $M^{(2)}$ . The lower bound of this CI  $\Delta_{LB}$  is the 2.5%ile of  $\Delta'$ . The upper bound of this CI  $\Delta_{UB}$  is the 97.5%ile of  $\Delta'$ .

## 4 INCORPORATING MULTIPLE INOFISH EXPERIMENTAL REPLICATES

In order to have a more accurate estimate of the editing rate, we care about the average editing rate across multiple replicates. The following strategy for estimating average editing level across multiple replicates can be applied to any number of replicates,  $R$ .

For example, if we have three replicates ( $R=3$ ) with results:  $\{T^{(1)}, O_A^{(1)}, O_I^{(1)}, S_A^{(1)}, S_I^{(1)}\}$ ,  $\{T^{(2)}, O_A^{(2)}, O_I^{(2)}, S_A^{(2)}, S_I^{(2)}\}$ , and  $\{T^{(3)}, O_A^{(3)}, O_I^{(3)}, S_A^{(3)}, S_I^{(3)}\}$ . We think that the mean of the editing levels across these three replicates is a good summary statistic for describing the editing level in the biological and experimental conditions we are observing.

$$\bar{E} = \text{mean}(E^{(1)}, E^{(2)}, E^{(3)}) \quad (5)$$

Our maximum likelihood estimate of  $\bar{E}$  is simply the mean of the editing rate across all three replicates,  $\bar{E}$ . We can draw a confidence interval on  $\bar{E}$  using a similar algorithm to an individual  $\hat{E}$ . In the case of averaging over three replicates:

1. Model all three experiments with multinomials, as above. Call these distributions  $M^{(1)}$ ,  $M^{(2)}$ , and  $M^{(3)}$ .
2. Draw a sample from  $M^{(1)}$  and calculate  $\hat{E}^{(1)}$
3. Draw a sample from  $M^{(2)}$  and calculate  $\hat{E}^{(2)}$

4. Draw a sample from  $M^{(3)}$  and calculate  $\hat{E}^{(3)}$
5. Calculate the mean of these draws:  $\bar{E}' = \text{mean}(\hat{E}^{(1)}, \hat{E}^{(2)}, \hat{E}^{(3)})$ . Store this result.
6. Repeat steps 2-5 100000 times.

As above, the 2.5%ile is the lower bound on the confidence interval,  $\bar{E}'_{LB}$ . The 97.5%ile is the upper bound on the confidence interval,  $\bar{E}'_{UB}$ .

#### 4.1 ALTERNATIVE SUMMARY STATISTICS ON EXPERIMENTAL REPLICATE RESULTS

One might also reasonably want to report statistics on editing levels across multiple replicates other than the mean. For example,

1. The maximum might help you better understand if the editing level of the site of interest saturates at a given percentage in these experimental conditions.
2. The median editing level across multiple replicates may be a more stable average value if you want to de-emphasize outlying results.
3. The sample variance would give insight into the variability between replicates.

In order to use these, or any other summary statistics on replicate editing levels, substitute that statistic in step 5 of the algorithm in Section 4.

#### 4.2 DETECTING DIFFERENCES IN SUMMARY STATISTICS WITH MULTIPLE REPLICATES IN TWO CONDITIONS

To bring our population-wide analyses to a close, let's integrate multiple experimental replicates across multiple conditions. In a toy case, let's work with  $C=2$  conditions, each with  $R=3$  replicates.

We want to know if the mean editing rates are significantly different from each other  $\bar{E}^{(1)} - \bar{E}^{(2)} = \Delta$ . As above, we can model each replicate  $j \in \{1, 2, 3\}$  in each condition  $i \in \{1, 2\}$  with  $R \cdot C = 6$  multinomials  $M^{ij}$ .

Our estimate of  $\Delta$  is  $\bar{\Delta} = \bar{E}^{(1)} - \bar{E}^{(2)}$ . As above, we can use parametric bootstrapping to draw a confidence interval around this estimate or to check its significance.

1. Model all experiments with multinomials, as above. Call these distributions  $M^{(1,1)}$ ,  $M^{(1,2)}$ ,  $M^{(1,3)}$ ,  $M^{(2,1)}$ ,  $M^{(2,2)}$ ,  $M^{(2,3)}$ .
2. Draw a samples from condition 1 models  $M^{(1,1)}$ ,  $M^{(1,2)}$ ,  $M^{(1,3)}$  and calculate  $\bar{E}^{(1)}$
3. Draw a samples from condition 2 models  $M^{(2,1)}$ ,  $M^{(2,2)}$ ,  $M^{(2,3)}$  and calculate  $\bar{E}^{(2)}$
4. Calculate  $\bar{E}^{(1)} - \bar{E}^{(2)} = \delta'$ . Store this result in the vector  $\Delta'$ .
5. Repeat steps 2-5 100000 times.

As above, the 2.5%ile is the lower bound on the confidence interval,  $\bar{\Delta}_{LB}$ . The 97.5%ile is the upper bound on the confidence interval,  $\bar{\Delta}_{UB}$ .

## 5 SINGLE-CELL EDITING LEVEL ANALYSIS

inoFISH also enables quantitative analysis of editing levels with single-cell resolution. The most obvious question we might ask about editing levels in single cells is whether all cells in the population appear to have the same editing level, or if they instead appear to have different editing levels.

Each cell has a variable and relatively small number of transcripts. Therefore our resolution of single-cell editing levels is much more limited than population-wide. However, if we make a simplifying assumption, we can nonetheless use our inoFISH data to answer this question. In particular, we will assume that all of our labels are true-positive.

Consider an experiment with  $N$  cells and population-wide measurements as above, along with an estimate of population-wide editing level  $\hat{E}$ . For  $i \in \{1, \dots, N\}$ , cell <sub>$i$</sub>  has  $O_I^{(i)}$  transcripts labelled as edited and  $O_A^{(i)}$  labelled as unedited. Call  $O_T^{(i)} = O_I^{(i)} + O_A^{(i)}$ .

If all  $N$  cells have the same level of editing, then each cell's count of labelled edited transcripts ( $O_I^{(i)}$ ) will appear to be drawn from a binomial distribution with parameters  $O_T^{(i)}$  and  $\hat{E}$ . The data likelihood of observing these results in all  $N$  cells is:

$$P(D|\tau) = \prod_{i=1}^N P(O_I^{(i)} | O_T^{(i)}, \hat{E}) = \prod_{i=1}^N O_T^{(i)} O_I^{(i) \hat{E}} (1 - \hat{E})^{O_T^{(i)} - O_I^{(i)}} \quad (6)$$

We can use this formula to bootstrap a distribution for data likelihood values (as in previous sections). If the observed value of  $P(D|\tau)$  falls outside of the bulk of the simulated values, then we can conclude that the experimental results are unlikely to come from our simple binomial model. That is, some other source of variability must also contribute to our results. The most intuitive explanation for this is that there is some cell-to-cell difference in editing level.

## APPENDIX B: MATERIALS AND METHODS FOR CHAPTER 1

### **inoFISH protocol.**

The step-by-step inoFISH protocol (description below) can be found

<https://media.nature.com/original/nature-assets/nmeth/journal/v14/n8/extref/nmeth.4332-S3.pdf>.

**Cell culture.** We grew human neuroblastoma cells (SH-SY5Y, ATCC CRL- 2266) in a 1:1 mixture of Eagle's Minimum Essential Medium and F12 Medium supplemented with 10% FBS and 50 U/mL penicillin and streptomycin. We grew human glioblastoma cells (U-87 MG, ATCC HTB-14) in Eagle's Minimum Essential Medium supplemented with 10% FBS and 50 U/mL penicillin and streptomycin. Note: Some SH-SY5Y and U87 cells can be autofluorescent when grown on glass slides, this is improved when the cells are ~70% confluent.

*Verification of RNA editing of candidate targets.* As described below, we used RT-PCR of total RNA and PCR of genomic DNA in cell lines of interest to further check that candidate editing sites were in fact RNA edited in our cell lines.

*Optimization of inoFISH targets.* We then chose targets for inoFISH by checking RT-PCR and genomic DNA PCR Sanger sequencing results for each candidate site to ensure that there would be no additional polymorphisms in transcripts, resulting either from RNA editing or single-nucleotide polymorphisms, in the regions flanking the editing site up to 30-bp up- or downstream. We ultimately designed inoFISH probe sets against one editing site in *NUP43* and two editing sites in *EIF2AK2* (both using the same guide probe set) that appeared to be amenable to inoFISH guide and detection probe set design. We were able to verify inoFISH probe binding with detection efficiencies greater than expected by random colocalization for the one *NUP43* candidate editing site and one of the two *EIF2AK2* editing sites. (See below for experimental methods.)

**Genotyping of edited regions.** We extracted genomic DNA from SH-SY5Y cells and U-87 MG cells using the Qiagen DNeasy Blood & Tissue kit. We used Platinum Taq (Invitrogen cat. #10966-018) for PCR amplification of the genomic regions of interest for each target, following



manufacturer's recommendations for reaction component concentrations. We PCR-amplified two biological replicates, each with two technical PCR replicates. For *GRIA2*, we used primers *GRIA2-F1* and *GRIA2-R2* (**Table 1.1**). For *EIF2AK2*, we used *EIF2AK2\_20-F1* and *EIF2AK2\_20-R1*, and for *NUP43* we used *NUP43-F1* and *NUP43-R1* (**Table 1.1**). We confirmed PCR product sizes by gel electrophoresis, using a 1.5% agarose gel in TAE. Then, we treated these PCR products with ExoSAP-IT (Affymetrix 78200) according to manufacturer's instructions, and submitted them for Sanger sequencing at the University of Pennsylvania DNA Sequencing facility.

**Estimation of editing efficiency by RT-PCR and Sanger sequencing.** We extracted total RNA from SH-SY5Y and U-87 MG cells using miRNeasy kits (Qiagen 217004) according to manufacturer's instructions. Then, we reverse-transcribed target transcripts around editing sites of interest using Superscript III First strand RT kit (ThermoFisher 18080044) according to manufacturer's instructions. In separate reactions for RNA from each cell type, we used both oligo-dT and transcript-specific primers for reverse-transcription. Briefly, we used 50 ng of RNA per reaction for reverse transcribed with either oligo-dT or transcript-specific primers (**Table 1.1**). Then, we performed PCR with transcript-specific primers (**Table 1.1**) using Platinum Taq (Invitrogen cat. #10966-018) according to manufacturer's instructions. We completed biological replicates, each with technical PCR replicates for these reactions. We confirmed product sizes by gel electrophoresis on 1.5% agarose gels in TAE. Then, we treated these products with ExoSAP-IT (Affymetrix 78200) according to manufacturer's instructions and submitted for sequencing at the University of Pennsylvania DNA Sequencing facility.

**Estimation of editing efficiency by clonal analysis of *GRIA2* RT-PCR product.** The amplified *GRIA2* cDNA was cloned into a vector using the TOPO TA cloning kit (Thermo), transformed into chemically competent *Escherichia coli* cells, and plated on LB plates with 0.1 mg/mL ampicillin. We isolated DNA from >20 Individual colonies and submitted it for sequencing at the University of Pennsylvania DNA Sequencing Facility. We performed sequence alignment at the editing site using MAFFT in Benchling to determine the ratio of edited and unedited transcripts.

**Estimation of editing efficiency by restriction digest and bioanalyzer analysis.** Edited and unedited *GRIA2* cDNAs yield distinct restriction fragment patterns upon digestion with *BbvI* ([Whitney et al., 2008](#)). Edited *GRIA2* cDNA yields two DNA fragments upon digestion (225 bp and 46 bp), and unedited *GRIA2* cDNA yields three DNA fragments (145 bp, 80 bp, and 46 bp). Following *BbvI* digestion (NEB R0173S) of *GRIA2* cDNA, according to manufacturer's instructions, we submitted digestion products for fragment sizing analysis on an Agilent 2100 Bioanalyzer at the University of Pennsylvania DNA Sequencing Facility.

**RNA probe design and synthesis.** For each of the validated editing sites, we designed probes by matching free energies of hybridization as specified in Levesque et al. (2013b). We optimized mask oligonucleotides to leave 8-base-pair (bp) overhangs for each of the detection probes and pooled all five together to act as the complete allele-specific probe. We provide all oligonucleotide sequences in Table 1.1. We coupled 3' amine-labeled adenosine- and inosine-detection probes to NHS-Cy3 or NHS-Cy5 fluorophores (GE Healthcare) and purchased respective guide probes labeled with Cal fluor 610 (Biosearch Technologies). We coupled probes targeting ADAR1, ADAR2 and NEAT1 mRNA to NHS-Atto700. We purified dye-coupled probes by high-performance liquid chromatography.

**inoFISH procedure.** We grew cells on glass coverslides until ~70% confluent. We washed the cells twice with 1X PBS, then fixed for 10 minutes with 4% formaldehyde/1X PBS at room temperature. We aspirated off the formaldehyde, and rinsed twice with 1X PBS prior to adding 70% ethanol for storage at 4°C or inoFISH after a one hour permeabilization in 70% ethanol. We incubated our cells overnight at 37°C in hybridization buffer (10% dextran sulfate, 2× SSC, 10% formamide) with 100 nM concentration of guide probe, 24 nM concentration of the adenosine- and inosine-detection probes and 72 nM concentration of the mask probe, ensuring excess mask for complete hybridization to the detection probes. The following morning, we performed two washes in wash buffer (2X SSC, 10% formamide), each consisting of a 30-min incubation at 37°C. After the second wash, we rinsed once with 2X SSC/DAPI and once with anti-fade buffer (10 mM Tris (pH 8.0), 2X SSC, 1% w/v glucose). Finally, we mounted the sample for imaging in

an anti-fade buffer with catalase and glucose oxidase (Raj et al. 2008) to prevent photobleaching. We performed RNA FISH on cell culture samples grown on a Lab-Tek chambered coverglass using 50  $\mu$ L of hybridization solution spread into a thin layer with a coverslip and placed in a parafilm-covered culture dish with a moistened Kimwipe to prevent excessive evaporation.

**Imaging.** We imaged each samples on a Nikon Ti-E inverted fluorescence microscope using a 100 $\times$  Plan-Apo objective (numerical aperture of 1.40) and a cooled CCD camera (Andor iKon 934). For 100 $\times$  imaging, we acquired z-stacks (0.3  $\mu$ m spacing between stacks) of stained cells in five different fluorescence channels using filter sets for DAPI, Cy3, Calfluor 610, Cy5, and Atto 700. The filter sets we used were 31000v2 (Chroma), 41028 (Chroma), SP102v1 (Chroma), 17 SP104v2 (Chroma) and SP105 (Chroma) for DAPI, Atto 488, Cy3, Atto 647N/Cy5 and Atto 700, respectively. A custom filter set was used for Alexa 594/CalFluor610 (Omega). We tuned the exposure times depending on the dyes used: 4 seconds for each guide probe, 4000 msec for each of the detection probes, 5000 msec for the NEAT1 probe, and 7000 msec for ADAR1 and ADAR2 probes. We also acquired images in the Atto 488 channel with a 1000 msec exposure as a marker of autofluorescence.

**Image analysis.** We first segmented and thresholded images using a custom Matlab software suite (downloadable at <https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Home>). Segmentation of cells was done manually by drawing a boundary around non-overlapping cells. The software then fits each spot to a two-dimensional Gaussian profile specifically on the Z-plane on which it occurs in order to ascertain subpixel-resolution spot locations. Colocalization took place in two stages: In the first stage, guide spots searched for the nearest-neighbor detection probes within a 2.5-pixel (360-nm) window. We ascertained the median displacement vector field for each match and subsequently used it to correct for chromatic aberrations. After this correction, we used a more stringent 1.5-pixel (195-nm) radius to make the final determination of colocalization. In order to test random colocalization due to spots occurring randomly by chance, we took our images and shifted the guide channel by adding 5 pixels (1.3  $\mu$ m) to the X and Y coordinates and then performing colocalization analysis.

**Autofluorescence subtraction.** For U-87 MG cells, we controlled for punctate autofluorescence by imaging with the 41028 (Chroma) filter set, the 'gfp channel', which we have previously found to be sensitive for autofluorescence in this cell line (data not shown). We performed colocalization as previously described between guide spots and any spot-like autofluorescence called in the gfp channel. In R, we excluded spots colocalizing with this autofluorescence from all inoFISH analyses.

**Subcellular localization.** *Nuclear localization.* We extracted a DAPI nuclear mask as previously described ([Raj et al., 2008](#)). We call a spot as localized to the nucleus if the guide spot X and Y coordinates overlap with the 2D nuclear mask.

*Localization to transcription sites.* We visualized *NUP43* introns by probing with intron-specific probes coupled to Atto 700 and imaging with SP105 filter set. We used the txnSiteGUI2 interface within rajlabimagetools to manually curate calls of exon-intron spot colocalization.

*Localization to paraspeckles.* We visualized paraspeckles by probing with NEAT1-specific probes coupled to Atto 700 and imaging with SP105 filter set. We used the txnSiteGUI2 interface within rajlabimagetools to manually curate calls of transcript-paraspeckle association.

**In situ cyanoethylation.** Cyanoethylation was performed similarly to previous descriptions ([Sakurai et al., 2010](#); [Yoshida and Ukita, 1968](#)). We aspirated the 70% ethanol off of the fixed cells and added cyanoethylation solution (1.1 M triethylammoniumacetate (pH 8.6) resuspended in 100% ethanol) with or without 1.6 M acrylonitrile at 70 °C for 15 min. Use large volume to prevent drying from evaporation. Remove from heat after 15 min (30 min incubation abolishes guide probe signal) and wash twice with wash buffer (2× SSC, 10% formamide) before beginning inoFISH procedure.

**Statistical analysis.** *Detection efficiency.* For each label (edited or unedited) in each experiment we calculated the mean fraction of transcripts colocalized with a spot of that label over all replicates (excluding 3-color spots). For complete details of this analysis, please see Appendix A.

*Population-wide editing level estimation by inoFISH.* We define the population-wide editing level estimate as the average over all replicates of the inferred fraction of uniquely labelled guide spots

labelled as edited. For a complete description of our estimation of population-wide editing level, please see the Supplementary Note.

*Paraspeckle-transcript association rates.* In MATLAB, we simulated the exact conditional null distribution of paraspeckle-transcript association rates for each experiment under the null hypothesis that a paraspeckle and a nuclear-localized transcript will only colocalize by chance. For each cell in each experiment, we conditioned on (1) the shape and size of that cell's nucleus, (2) the locations of all paraspeckles in that nucleus, and (3) the number of transcripts of interest (*GRIA2* or *EIF2AK2*) retained in the nucleus. In order to efficiently simulate these distributions, rather than using txnSiteGUI2 as above, we generated 2D masks for paraspeckle locations and called paraspeckle-transcript association when a randomly placed transcript spot overlapped with this mask. We selected the mask size as 25 pixels per paraspeckle spot called--roughly the mean paraspeckle size--based on our inspection of paraspeckles while calling spots (as in Image analysis). For each experiment, we simulated draws from the exact conditional null distribution 1000 times. A raw p-value for paraspeckle-transcript association rate is equal to the fraction of simulations with a higher paraspeckle-transcript association rate. We similarly simulated exact conditional null distributions for paraspeckle-edited-transcript and paraspeckle-unedited-transcript association rates.

*Single-cell editing level distributions.* In R, we assessed single-cell spot counts after inoFISH colocalization as reported by rajlabimagetools (<https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Home>; in Image analysis), as well as after autofluorescence subtraction (for U-87 MG data). We simulated the null distribution of data likelihood under a null model wherein all cells sharing the same effective editing level: for an experiment with overall estimated editing level equal to  $p_e$  (above), let  $n_e^j$  be the number of edited transcripts detected in cell  $j$  and  $n_u^j$  be the number the number of unedited transcripts detected in cell  $j$ . Under the null model,  $n_e^j$  is drawn from a Binomial with  $(n_e^j + n_u^j)$  draws and probability  $p_e$ . We simulated single-cell label counts for cells by drawing from these conditional null distributions for each cell 100000 times. We then compared the negative log-likelihood of the observed data,

combined over all replicates, with the distribution of negative log-likelihoods of each simulation iteration. A p-value of 0.12 indicates that 12% of the simulated iterations had a negative log-likelihood that was greater than the observed data. Note that the  $-\log(\text{likelihood})$  density plots in Fig. 3 are subsampled to 3000 of the aforementioned 100000 such iterations per plot, in order to facilitate figure generation. For complete details of this analysis, please see the Supplementary Note.

**siRNA knockdowns of ADAR2.** Briefly, we used Lipofectamine RNAiMax to transfect SH-SY5Y cells with Silencer Select siRNAs targeting ADARB1 (ADAR; ID:s1011, Ambion) and a Negative Control siRNA (#1, Ambion) for 72 hrs, verifying knockdown via RNA FISH of ADAR2.

**SFPQ-guided GRIA2 inoFISH.** We performed GRIA2 inoFISH, as described above, but substituted a smFISH probe set for SFPQ ([Cabili et al., 2015](#)) for the GRIA2 guide. Like GRIA2 transcripts, SFPQ transcripts are localized to the nucleus in SH-SY5Y cells. In parallel we performed regular GRIA2 inoFISH on a sample of cells from the same passage. We counted the number of GRIA2 mRNA per cell in the regular sample, and subsampled the SFPQ “guide” spots from that distribution of mRNA counts per cell. In this way, we could more directly compare colocalization rates with GRIA2 detection probes between SFPQ-guided and GRIA2-guided experiments. We then performed colocalization as described above.

**Cell cycle inhibitor.** We measured nuclear retention of *GRIA2* mRNA by inhibiting transcription for 24 hr by applying aphidicolin at 1 ug/ml.

**Reproducible analyses.** Scripts for all analyses presented in this paper, including all data extraction, processing, and graphing steps are freely accessible at the following url:

<https://www.dropbox.com/sh/j5umuneita1nck9/AAA4W4I648gIUUhePJfXyaRaa?dl=0> .

All imaging and other non-RNA-sequencing data associated with this paper are also freely available at the following url:

<https://www.dropbox.com/sh/vwnwrmgg72o75c/AACsFK6VbJHY2S5MK8JLR2JNa?dl=0> . For

publicly available RNA-seq data discussed in Supp. Fig. 1, please see EBI ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>) accession numbers E-MTAB-2690 (SH-SY5Y samples

“SY5Y\_A” and “SY5Y\_B”) and E-MTAB-1875 (U-87 MG samples “s\_2\_78”, “s\_2\_82”, and “s\_2\_88”).

## APPENDIX C: MATERIALS AND METHODS FOR CHAPTER 2

### **Cell culture**

Unless otherwise noted, all cell culture incubations below were performed at 37°C, 5% CO<sub>2</sub>. We tested intermittently for mycoplasma contamination.

#### **GM00942 human dermal fibroblast culture**

We cultured GM00942 human dermal fibroblasts (Coriell, GM00942; XX donor, normal-appearing tissue) according to the distributor's instructions, on tissue culture-treated dishes in E-MEM (QBI 112-018-101) + 10% FBS (Life Technologies 16000044, lot 1802004) + Pen/Strep.

#### **GM11169 human cardiac fibroblast culture**

We cultured GM11169 human cardiac fibroblasts (Coriell, GM11169; XX donor) on tissue culture-treated dishes in DMEM w/Glutamax + 9% FBS (Life Technologies 16000044) + P/S.

#### **HEK293FT culture**

We expanded HEK293FT cells in DMEM w/Glutamax + 9% FBS + P/S.

#### **hiF-T culture**

We cultured hiF-T cells as previously described prior to hiF-T-iPSC reprogramming experiments([Cacchiarelli et al., 2015](#)). Briefly, we expanded hiF-T cells in growth medium on TC plastic dishes coated with Attachment Factor (Fisher S006100), and split cells 1:3 when they reached 60-70% confluency. hiF-T growth medium (GM) is DMEM/F-12 w/ Glutamax (Life Tech. 10565018) + 10% ES-FBS (Life Tech. 16141079) + 1x 2-Mercaptoethanol (Life Tech. 21985023) + 1x NEAA (Invitrogen 11140050) + P/S + 0.5µg/mL Puromycin + 16ng/mL rhFGF-basic (Promega G5071).

#### **Immortalized human cardiac fibroblast (immHCF) culture**

We cultured human immortalized human cardiac fibroblasts (HCFs) as previously described([Mohamed et al., 2017](#)). We received HCFs from Deepak Srivastava. In brief, we expanded HCFs on gelatin-coated (Millipore ES-006-B) tissue culture dishes in iCM medium (per



500mL: 350mL DMEM w/Glutamax + 85mL Medium 199 + 50mL FBS + 5mL Non-essential amino acids + 10mL Pen/strep).

### **Platinum-A (Plat-A) retroviral packaging cell line culture**

We cultured Platinum-A cells (Cell Biolabs RV-102) according to the manufacturer's instructions. We expanded these cells in DMEM w/Glutamax + 9% FBS + P/S + 10ug/mL Blasticidin + 1ug/mL Puromycin.

### **Cellular reprogramming**

#### **Derivation of cardiomyocytes (iCards) from GM942-iPSCs**

We derived cardiomyocytes from GM942-iPSCs, called iCard-942 cells, as previously described ([Laflamme et al., 2007](#); [Palpant et al., 2015](#); [Zhu et al., 2010, 2011](#)). Briefly:

#### **Seeding of GM942-iPSCs**

We thawed GM942-iPSCs and grew them in feeder-free conditions on Geltrex-coated (ThermoFisher, cat. A1413301) dishes in iPS-Brew (Miltenyi, cat. 130-104-368) for 4-5 days, until ~75% confluency. Then, we split and seeded GM942-iPSCs into Geltrex-coated 12-well plates at a density of roughly 1e6 cells per well in iPS-Brew + 2µM Thiazovivin (Sigma, cat. SML1045-5MG). After 24 hours, we changed the culture medium to iPS-Brew + 1µM Chiron 99021 (Cayman Chemical, cat. 13122) and then incubated the cells for a further 24 hours.

#### **Differentiation to iCard-942 cells**

Starting on Day 0, we incubated GM942-iPSCs for 18 hours in RPMI (Life Technologies, cat. 11875-119) + 100ng/ml Activin A (R&D systems, cat. 338-AC- 010) + 2% B-27 (minus insulin; Life Technologies, cat. 17504-044). Next, on Day 1 we changed the medium to RPMI + 2% B-27 (minus insulin) + 5ng/ml BMP4 (Peprotech, cat. AF-120-05ET) + 1uM Chiron 99021 and incubated these cells for 48 hours. On day 3 we changed the medium to RPMI + 2% B-27 (minus insulin) + 1uM Xav 939 (Tocris Bioscience, cat. 3748) and incubated cells for 48 hours. On Day 5 we changed the medium to RPMI + 2% B-27 (minus insulin) and incubated the cells for 72 hours.

From Day 8 - 12 we changed the medium to RPMI + 2% B-27 (including insulin) + 1% pen/strep, and replaced medium every other day.

### **Glucose-free medium selection steps for cardiomyocytes**

In order to enrich the culture for cardiomyocytes, we subjected these cells to two glucose-free selection steps. On Day 12 we started the first selection step by replacing the medium with RPMI glucose free (ThermoFisher, cat. 11879020) + 2% B-27 (including insulin) + 1% Pen-Strep and incubated for 72 hours. On day 15 we replated the cells onto Geltrex-coated dishes at  $6.3 \times 10^5$  cells/cm<sup>2</sup> in RPMI + 20% FBS (Seradigm, lot 050B14) + 1 $\mu$ M Thiazovivin and incubated them for 24 hours. On Day 16 we changed the medium to RPMI + 2% B-27 (including insulin) and incubated them for 48 hours to recover. On Day 18 we started the second selection step, again by replacing the medium with RPMI glucose free (ThermoFisher, cat. 11879020) + 2% B-27 (including insulin) + 1% Pen-Strep and incubated for 72 hours.

### **Splitting iCard-942 cells into 96-well plates**

On Day 21 we split the double-selected iCard-942 cells into Geltrex-coated 96-well plates at a density of  $1 \times 10^5$  cells per well in RPMI + 20%FBS + 1 $\mu$ M thiazovivin and incubated overnight. On Day 22 we changed the medium to RPMI + 2% B-27 (with insulin) + pen/strep and incubated for 48 hours. By Day 23 or 24 we expected to observe recovery of beating activity among a majority of the cardiomyocytes in all of the wells. If we did not observe beating activity we changed the RPMI + 2% B-27 (with insulin) + pen/strep medium on Day 24, incubated for another 48 hours, and checked again for beating activity on Day 26. If the cardiomyocytes did not regain beating activity by Day 26 we did not proceed to perturbation culture.

### **Perturbation culture**

#### *GM00942 fibroblasts*

For GM00942 fibroblasts, we called the day on which they were seeded in 96-well plates Day -2 of perturbation. We split 95% confluent 10cm tissue culture-treated dishes of cells into 96-well plates at a density of roughly  $\sim 1.5 \times 10^4$  cells per well in EMEM + 10% FBS + Pen/strep and

incubated for 48 hours. On Day 0 we replaced the medium (250  $\mu$ L) in each well and added 0.7  $\mu$ L of drug stock in DMSO (see **Table 2.1**) or of DMSO (for control cultures). This kept total DMSO concentration of the perturbation culture medium below 0.3% for all conditions. We incubated cells for 48 hours, and then on Day 2 replaced medium and re-adding a fresh dose of drug stock at the same volumes as Day 0. We incubated cells for a further 48 hours before taking images of each well (below) and extracting RNA (below) on Day 4.

#### *iCard-942*

For iCard-942 cells, we called the first day on which we observed beating activity in the majority of wells of each 96-well plate Day 0 of perturbation. On Day 0 we replaced the medium with 250 $\mu$ L RPMI + 2% B-27 (with insulin) + pen/strep and added 0.7 $\mu$ L of drug stock in DMSO (See **Table 2.1**) or of DMSO (for control cultures), again keeping total DMSO concentration of perturbation culture medium below 0.3% for all conditions. We incubated cells for 48 hours, and then on Day 2 replaced medium and re-adding a fresh dose of drug stock at the same volumes as Day 0. We incubated cells for a further 48 hours before taking videos of each well (below) and extracting RNA (below) on Day 4.

#### **Transdifferentiation of immortalized HCFs to induced cardiomyocyte-like cells**

We performed transdifferentiation of immortalized HCFs to induced cardiomyocyte-like cells (iCMs) as previously described ([Mohamed et al., 2017](#)). Briefly, on day -3 we plated HCFs in 12-well culture vessels (indicated cell number per experiment) in iCM medium and Plat-A cells in 10cm dishes (4e6 cells per dish) in DMEM w/Glutamax + 9% FBS without any antibiotics. On day -2 we transfected each dish of Plat-A cells with 10ug of one indicated pMXs expression plasmid in 500uL Optimem + 35uL Fugene HD. On day 0, we collected viral supernatants and pooled them as needed for replicate conditions, filtered them through 0.45 $\mu$ m filter units, and transduced HCFs. For transductions we used 6 $\mu$ g/mL polybrene, a 30min 930 x g spin, and overnight incubation at 37C. On day 1 we replaced transduction medium with iCM medium. On day 4 we replaced iCM medium with 75% iCM medium/25% Reprogramming medium (RPMI 1640 + B-27 + P/S), on day 7 with 50% iCM medium/50% Reprogramming medium, on day 11 with 25% iCM

medium/75% Reprogramming medium, and on day 14 with Reprogramming medium alone. We then changed reprogramming medium daily until analysis on the indicated day per experiment, usually day 24. On day 24, we fixed cells in two formats for analysis. Some 12-well TC plastic wells were fixed in place using 3.7% formaldehyde and permeabilized at least overnight with 70% Ethanol in 4C, while others were dissociated with Accutase and transferred to Concanavalin-coated (Sigma C0412) 8-well Lab-Tek chambers. After 90-120 minutes, transferred samples were fixed in these 8-well chambers using 3.7% formaldehyde and permeabilized at least overnight with 70% Ethanol in 4C. Samples in 12-well wells were processed for FISH imaging by excising them from their 12-well plate after fixation with a heated 20mm cork borer and processing them for FISH or immunofluorescence as described below.

### **Cloning of transcription factor genes and TurboGFP into pMXs**

In order to drive overexpression of perturbable transcription factor genes and TurboGFP, we cloned cDNA for genes of interest into pMXs-gw (Addgene 18656; a gift from Shinya Yamanaka) using BP and LR Clonase II (Invitrogen). We amplified cDNA of targets of interest using attB-target-specific primers (**Table 2.2**). We used standard tools to verify sequence identity of the plasmid backbone and gene insert, such as restriction digestion and Sanger sequencing. We amplified attB-TurboGFP off of the SHC003 plasmid (Sigma SHC003).

### **Indirect functional titering of pMXs retroviral vectors**

Since our expression vectors do not contain selectable or fluorescent markers and pMXs retroviral vectors only transduce dividing cells, we indirectly titered each experimental replicate's batch of virus by co-transducing parallel samples of HCFs with pMXs-DsRed Express (Addgene 22724; a gift from Shinya Yamanaka) and pMXs-TurboGFP (see "Cloning" above) produced using the same batch of Plat-A cells under the same conditions. In order to estimate the fraction of cells that are infected at least once per factor, we considered the infection rates of these fluorescent pMXs vectors. By comparing the fraction that are co-infected with both against the fraction that are infected with each factor at all, we can infer the fraction of cells dividing in the population during the transduction period and the fraction of those cells that receive at least one

copy of any individual expression vector. We make the simplifying assumption that among dividing cells infection events are independent of each other. Therefore, the ratio of the fraction that are DsRed+ and GFP+ to the fraction that are DsRed+ (or GFP+) is approximately the square of the transduction rate for any individual virus. E.g., for 30% of cells being DsRed+ and 24.3% being DsRed+ and GFP+,  $24.3/30 = 0.81$ , which gives 90% transduction rate for each individual virus.

### **shRNA-based knockdown of transcription factors in hiF-T cells**

We conducted knockdown of individual transcription factors using shRNAs essentially as previously described ([Cacchiarelli et al., 2015](#)). In brief, we acquired cloned pLKO.1, pLKO.1-shRNA, and pLKO.1-TurboGFP plasmids from the University of Pennsylvania High-Throughput Screening Core (**Table 2.3**). We verified shRNA and backbone sequence with Sanger sequencing. We packaged shRNA lentivirus using pMD2.G (Addgene 12259; a gift from Didier Trono) and psPAX2 (Addgene 12260; a gift from Didier Trono) in HEK293FT cells, and filtered viral supernatant through 0.22 $\mu$ m filter units prior to infecting hiF-T cells. We infected hiF-T cells at an MOI of approximately 1 (for a transduction efficiency of ~70%) with 4 $\mu$ g/mL polybrene and 30 min 930 x g centrifugation. Since hiF-T cells are already Puromycin-resistant, we were unable to perform an antibiotic selection step after infection with these pLKO.1-puro-based shRNA plasmids.

#### *Verification of knockdown efficiency following shRNA transduction*

We performed RT-qPCR on RNA extracted from samples of the hiF-T cells that we used in reprogramming experiments. We used Superscript III Reverse Transcriptase for first-strand cDNA synthesis and Power SYBR qPCR Master Mix with gene-specific primer pairs for qPCR on an Applied Biosystems 7300 system. We performed all statistical analysis using custom scripts in R (see “Code accessibility” below for all scripts) and calculated knockdown efficiency using the  $\Delta\Delta$ Ct method.

### **hiF-T reprogramming to pluripotency**

We performed hiF-T reprogramming experiments as previously described. Briefly, after shRNA transduction on day -7, we expanded cells in hiF-T GM without puromycin for one week. On Day -1 we seeded CF-1 Irradiated MEFs on uncoated 24-well plates (Corning) at a density of  $2.5 \times 10^5$  cells per well in hiF-T GM without puro. On Day 0, we seeded  $10^4$  hiF-T cells per 24-well plate well. On Day 1 we began Yamanaka factor induction by switching media to hiF-T GM with  $2 \mu\text{g}/\text{mL}$  doxycycline and without puromycin. On Day 3 we switched media to KSR Medium (KSRM): DMEM/F-12 w/ Glutamax (Life Tech. 10565018) + 20% Knockout Serum Replacement (Life Tech. 10828010) + 1x 2-Mercaptoethanol (Life Tech. 21985023) + 1x NEAA (Invitrogen 11140050) + P/S + 8ng/mL rhFGF-basic +  $2 \mu\text{g}/\text{mL}$  Doxycyclin. We changed KSRM daily, and analyzed cells on day 21.

### **High-throughput RNA extraction**

We used RNeasy-96 kits (Ambion AM1920) for RNA extraction without the optional DNase step, according to manufacturer's instructions.

### **RNAtag sequencing**

We conducted highly parallelized bulk RNA sequencing with RNAtag-seq as previously described, using all components and steps in the published protocol. ([Shishkin et al., 2015](#)) We ordered the specified 32 barcoded DNA oligos for RNAtags from Biosearch Technologies and indexed primers for library amplification and reverse transcription from IDT. We sequenced all RNAtag-seq libraries in batches of 96 samples on an Illumina NextSeq 550 using 75 cycle high-output kits (Illumina 20024906).

### **RNAtag-seq data processing**

We demultiplexed RNAtag-seq reads using custom scripts, courtesy of Edward Wallace. ([Wallace and Beggs, 2017](#)) We aligned RNAtag-seq reads to the human genome (hg19) with STAR v2.5.2a and counted uniquely mapping reads with HTSeq v0.6.1. ([Shaffer et al., 2017](#)) We

performed all downstream analysis in R v3.6.1 using packages `yaml_2.2.0`, `DESeq2_1.24.0`, `SummarizedExperiment_1.14.0`, `DelayedArray_0.10.0`, `BiocParallel_1.18.0`, `matrixStats_0.54.0`, `Biobase_2.44.0`, `GenomicRanges_1.36.0`, `GenomeInfoDb_1.20.0`, `IRanges_2.18.1`, `S4Vectors_0.22.0`, `BiocGenerics_0.30.0`, `e1071_1.7-2`, `magrittr_1.5`, `ggrepel_0.8.1`, `ggplot2_3.2.0`, `tibble_2.1.3`, `tidyr_0.8.3`, and `dplyr_0.8.3` and their associated dependencies.

### **Gene expression perturbability**

As a measure of gene expression perturbability we used the count of the number of conditions in which a gene was differentially expressed relative to cell type DMSO controls. For most analyses we used any change with a DESeq2 adjusted p-value less than 0.1, but also conducted analyses with additional filters, such as minimum absolute values of `log2FoldChange`. We also considered other measures of perturbability, as well, which are not included in the manuscript above, details of which are available upon request.

### **Prioritization of highly perturbable genes for use in transdifferentiation and reprogramming experiments**

We chose to follow up on highly perturbable genes in transdifferentiation and reprogramming experiments. In order to narrow down the list of transcription factors, we considered those 1) with average expression of 50 TPM or greater in controls of the cell type of interest, 2) with at least 4 conditions in which they are up-regulated, and 3) that are up-regulated in more conditions than they are down-regulated. Then we selected from these lists genes that were relatively unstudied in the context of cardiac transdifferentiation and iPSC reprogramming, as far as we could find in the literature.

### **Live cell Tra-1-60 imaging**

In a pilot reprogramming experiment without shRNA transduction, we conducted live-cell staining of hiF-T-iPSC colonies Tra-1-60 with TRA-1-60 Alexa Fluor™ 488 Conjugate Kit for Live Cell Imaging (Life Tech. A25618) according to the manufacturer's instructions.

### **Alkaline phosphatase staining with colorimetry**

We used the Vector Red Substrate kit (Vector Labs SK-5100) to stain hiF-T-iPSC colonies after fixation on day 21 of reprogramming experiments. We fixed wells in 24-well format using 3.7% formaldehyde for 3 minutes, and followed the manufacturer's instructions.

### **Immunofluorescence**

We performed immunofluorescence for several markers. For Tra-1-60 immunofluorescence of hiF-T-iPSC samples that had already been stained with Vector Red, we blocked and permeabilized in 5% BSA + 0.1% Triton X-100 in PBS at room temperature for 30 min. Then we washed samples in PBS and used Stemgent 09-0068 at 1:200 in 5% BSA + 0.1% Triton X-100 for 2 hours at room temp. We washed samples in PBS and stained with DAPI prior to imaging. For cardiac Troponin immunofluorescence of iPSC-derived cardiomyocytes and transdifferentiated samples, we fixed samples in 3.7% formaldehyde for 10 min at room temp, washed in PBS, and permeabilized with 70% Ethanol overnight at 4C. Independent of smFISH or after the smFISH protocol, we performed immunofluorescence with Abcam ab45932 primary (1:200) with goat anti-rabbit-Alexa 594 (1:200) secondary and Fisher MA5-12960 primary (1:200) with donkey anti-mouse-Alexa 488 (1:200) secondary. We used samples in 3% BSA + 0.1% Tween 20 for blocking/binding buffer. Primary antibody incubations of 1 hour and secondary incubations of 30 min, both at room temperature. Samples were washed with PBS and stained with DAPI prior to imaging.

### **Single-molecule RNA FISH**



We incubated our cells overnight at 37°C in hybridization buffer (10% dextran sulfate, 2× SSC, 10% formamide) with standard concentrations of RNA FISH probes (**Table 2.2**).[\(Padovan-Merhar et al., 2015\)](#) The following morning, we performed two washes in wash buffer (2X SSC, 10% formamide), each consisting of a 30-min incubation at 37°C. After the second wash, we rinsed once with 2X SSC/DAPI and mounted the sample for imaging in and 2X SSC.[\(Raj et al., 2008\)](#) We performed RNA FISH on cell culture samples grown on a Lab-Tek chambered coverglass using 50 µL of hybridization solution spread into a thin layer with a coverslip and placed in a parafilm-covered culture dish with a moistened Kimwipe to prevent excessive evaporation.

**Imaging.** We imaged each sample on a Nikon Ti-E inverted fluorescence microscope using a 60× Plan-Apo objective and a Hamamatsu ORCA Flash 4.0 camera. For 60× imaging of complete cells, we acquired z-stacks (0.3 µm spacing between stacks). For 60× imaging of a large field of cells with one plane each, we used Nikon Elements tiled image acquisition with perfect focus. All image of stained cells were in different fluorescence channels using filter sets for DAPI, Cy3, Alexa 594, and Atto 647N. The filter sets we used were 31000v2 (Chroma), 41028 (Chroma), SP102v1 (Chroma), 17 SP104v2 (Chroma) and SP105 (Chroma) for DAPI, Atto 488, Cy3, Atto 647N/Cy5 and Atto 700, respectively. A custom filter set was used for Alexa 594/CalFluor610 (Omega). We tuned the exposure times depending on the dyes used: 400 ms for probes in Cy3 and Alexa 594, 500 ms seconds for each probe in Atto 647N, and 50 ms for DAPI probes. We also acquired images in the Atto 488 channel with a 400 ms exposure as a marker of autofluorescence.

### **Image Processing**

smFISH analysis of image scans and stacks was done as previously described using rajlabimagetools changeset 775fd10

<https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Home> in MATLAB v2019a.[\(Shaffer et al., 2017\)](#)

## REFERENCES

- Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., et al. (2016). The origin and evolution of cell types. *Nat. Rev. Genet.*
- Bahn, J.H., Lee, J.-H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* 22, 142–150.
- Bass, B.L., and Weintraub, H. (1987). A developmentally regulated activity that unwinds RNA duplexes. *Cell* 48, 607–613.
- Becker, J.S., McCarthy, R.L., Sidoli, S., Donahue, G., Kaeding, K.E., He, Z., Lin, S., Garcia, B.A., and Zaret, K.S. (2017). Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol. Cell* 68, 1023–1037.e15.
- Black, J.B., Adler, A.F., Wang, H.-G., D'Ippolito, A.M., Hutchinson, H.A., Reddy, T.E., Pitt, G.S., Leong, K.W., and Gersbach, C.A. (2016). Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. *Cell Stem Cell*.
- Briggs, R., and King, T.J. (1952). Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proc. Natl. Acad. Sci. U. S. A.* 38, 455–463.
- Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Bjaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J.L., and Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* 16, 20.
- Cacchiarelli, D., Trapnell, C., Ziller, M.J., Soumillon, M., Cesana, M., Karnik, R., Donaghey, J., Smith, Z.D., Ratanasirintrao, S., Zhang, X., et al. (2015). Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. *Cell* 162, 412–424.
- Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* 158, 903–915.
- Cao, N., Huang, Y., Zheng, J., Spencer, C.I., Zhang, Y., Fu, J.-D., Nie, B., Xie, M., Zhang, M., Wang, H., et al. (2016). Conversion of human fibroblasts into functional cardiomyocytes by small molecules. *Science* 352, 1216–1220.
- Chen, L.-L., and Carmichael, G.G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol. Cell* 35, 467–478.
- D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M., et al. (2015). A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. *Stem Cell Reports*.

- Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* 51, 987–1000.
- Duan, J., Li, B., Bhakta, M., Xie, S., Zhou, P., Munshi, N., and Hon, G. (2018). Rational reprogramming of cellular states by combinatorial perturbation.
- Flomen, R., Knight, J., Sham, P., Kerwin, R., and Makoff, A. (2004). Evidence that RNA editing modulates splice site selection in the 5-HT<sub>2C</sub> receptor gene. *Nucleic Acids Res.* 32, 2113–2122.
- Freddolino, P.L., Yang, J., Momen-Roknabadi, A., and Tavazoie, S. (2018). Stochastic tuning of gene expression enables cellular adaptation in the absence of pre-existing regulatory circuitry. *eLife Sciences* 7, e31867.
- Fu, J.-D., and Srivastava, D. (2015). Direct reprogramming of fibroblasts into cardiomyocytes for cardiac regenerative medicine. *Circ. J.* 79, 245–254.
- Fu, J.-D., Stone, N.R., Liu, L., Spencer, C.I., Qian, L., Hayashi, Y., Delgado-Olguin, P., Ding, S., Bruneau, B.G., and Srivastava, D. (2013). Direct reprogramming of human fibroblasts toward a cardiomyocyte-like state. *Stem Cell Reports* 1, 235–247.
- Guo, C., and Morris, S.A. (2017). Engineering cell identity: establishing new gene regulatory and chromatin landscapes. *Curr. Opin. Genet. Dev.* 46, 50–57.
- Gurdon, J.B., Elsdale, T.R., and Fischberg, M. (1958). Sexually mature individuals of *Xenopus laevis* from the transplantation of single somatic nuclei. *Nature* 182, 64–65.
- Hideyama, T., Yamashita, T., Suzuki, T., Tsuji, S., Higuchi, M., Seeburg, P.H., Takahashi, R., Misawa, H., and Kwak, S. (2010). Induced loss of ADAR2 engenders slow death of motor neurons from Q/R site-unedited GluR2. *J. Neurosci.* 30, 11917–11925.
- Hideyama, T., Yamashita, T., Aizawa, H., Tsuji, S., Kakita, A., Takahashi, H., and Kwak, S. (2012). Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiol. Dis.* 45, 1121–1128.
- Higuchi, M., Maas, S., Single, F.N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P.H. (2000). Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* 406, 78–81.
- Jepson, J.E.C., Savva, Y.A., Jay, K.A., and Reenan, R.A. (2011). Visualizing adenosine-to-inosine RNA editing in the *Drosophila* nervous system. *Nat. Methods* 9, 189–194.
- Kuhlwilm, M., Davierwala, A., and Pääbo, S. (2013). Identification of putative target genes of the transcription factor RUNX2. *PLoS One* 8, e83218.
- Kumar, M., and Carmichael, G.G. (1997). Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl. Acad. Sci. U. S. A.* 94, 3542–3547.

- Laflamme, M.A., Chen, K.Y., Naumova, A.V., Muskheli, V., Fugate, J.A., Dupras, S.K., Reinecke, H., Xu, C., Hassanipour, M., Police, S., et al. (2007). Cardiomyocytes derived from human embryonic stem cells in pro-survival factors enhance function of infarcted rat hearts. *Nat. Biotechnol.* 25, 1015–1024.
- Levesque, M.J., Ginart, P., Wei, Y., and Raj, A. (2013). Visualizing SNVs to quantify allele-specific expression in single cells. *Nat. Methods* 10, 865–867.
- Liddicoat, B.J., Piskol, R., Chalk, A.M., Ramaswami, G., Higuchi, M., Hartner, J.C., Li, J.B., Seeburg, P.H., and Walkley, C.R. (2015). RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science*.
- Liu, Y., Yu, C., Daley, T.P., Wang, F., Cao, W.S., Bhate, S., Lin, X., Still, C., Liu, H., Zhao, D., et al. (2018). CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. *Cell Stem Cell* 0.
- Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P.H., and Higuchi, M. (1996). A mammalian RNA editing enzyme. *Nature* 379, 460–464.
- Mellis, I.A., and Raj, A. (2015). Half dozen of one, six billion of the other: What can small- and large-scale molecular systems biology learn from one another? *Genome Res.* 25, 1466–1472.
- Mellis, I.A., Gupte, R., Raj, A., and Rouhanifard, S.H. (2017). Visualizing adenosine-to-inosine RNA editing in single mammalian cells. *Nat. Methods*.
- Mohamed, T.M.A., Stone, N.R., Berry, E.C., Radzinsky, E., Huang, Y., Pratt, K., Ang, Y.-S., Yu, P., Wang, H., Tang, S., et al. (2017). Chemical Enhancement of In Vitro and In Vivo Direct Cardiac Reprogramming. *Circulation* 135, 978–995.
- Morris, S.A., Cahan, P., Li, H., Zhao, A.M., Roman, A.K.S., Shivdasani, R.A., Collins, J.J., and Daley, G.Q. (2014). Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet. *Cell* 158, 889–902.
- Nam, Y.-J., Song, K., Luo, X., Daniel, E., Lambeth, K., West, K., Hill, J.A., DiMaio, J.M., Baker, L.A., Bassel-Duby, R., et al. (2013). Reprogramming of human fibroblasts toward a cardiac fate. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5588–5593.
- O’Connell, M.A., Gerber, A., and Keller, W. (1997). Purification of human double-stranded RNA-specific editase 1 (hRED1) involved in editing of brain glutamate receptor B pre-mRNA. *J. Biol. Chem.* 272, 473–478.
- Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Mol. Cell* 1–36.
- Palpant, N.J., Pabon, L., Roberts, M., Hadland, B., Jones, D., Jones, C., Moon, R.T., Ruzzo, W.L., Bernstein, I., Zheng, Y., et al. (2015). Inhibition of  $\beta$ -catenin signaling respecifies anterior-like endothelium into beating human cardiomyocytes. *Development* 142, 3198–3209.

- Parekh, U., Wu, Y., Zhao, D., Worlikar, A., Shah, N., Zhang, K., and Mali, P. (2018). Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. *Cels 0*.
- Paschen, W., Hedreen, J.C., and Ross, C.A. (1994). RNA editing of the glutamate receptor subunits GluR2 and GluR6 in human brain tissue. *J. Neurochem.* 63, 1596–1602.
- Peng, P.L., Zhong, X., Tu, W., Soundarapandian, M.M., Molner, P., Zhu, D., Lau, L., Liu, S., Liu, F., and Lu, Y. (2006). ADAR2-dependent RNA editing of AMPA receptor subunit GluR2 determines vulnerability of neurons in forebrain ischemia. *Neuron* 49, 719–733.
- Piskol, R., Peng, Z., Wang, J., and Li, J.B. (2013). Lack of evidence for existence of noncanonical RNA editing. *Nat. Biotechnol.* 31, 19–20.
- Porath, H.T., Carmi, S., and Levanon, E.Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* 5, 4726.
- Prasanth, K.V., Prasanth, S.G., Xuan, Z., Hearn, S., Freier, S.M., Bennett, C.F., Zhang, M.Q., and Spector, D.L. (2005). Regulating gene expression through RNA nuclear retention. *Cell* 123, 249–263.
- Rackham, O.J.L., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., FANTOM Consortium, Suzuki, H., Nefzger, C.M., Daub, C.O., et al. (2016). A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.*
- Radley, A.H., Schwab, R.M., Tan, Y., Kim, J., Lo, E.K.W., and Cahan, P. (2017). Assessment of engineered cells using CellNet and RNA-seq. *Nat. Protoc.* 12, 1089–1102.
- Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* 5, 877–879.
- Ramaswami, G., and Li, J.B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 42, D109–D113.
- Ramaswami, G., Lin, W., Piskol, R., Tan, M.H., Davis, C., and Li, J.B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods* 9, 579–581.
- Ramaswami, G., Zhang, R., Piskol, R., Keegan, L.P., Deng, P., O’Connell, M.A., and Li, J.B. (2013). Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods* 10, 128–132.
- Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szykiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., et al. (2012). Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. *Nat. Genet.* 44, 1243–1248.
- Rouhanifard, S.H., Mellis, I.A., Dunagin, M., Bayatpour, S., Jiang, C.L., Dardani, I., Symmons, O., Emert, B., Torre, E., Cote, A., et al. (2018). ClampFISH detects individual nucleic acid molecules using click chemistry-based amplification. *Nat. Biotechnol.*

- Sakurai, M., Yano, T., Kawabata, H., Ueda, H., and Suzuki, T. (2010). Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nat. Chem. Biol.* *6*, 733–740.
- Sakurai, M., Ueda, H., Yano, T., Okada, S., Terajima, H., Mitsuyama, T., Toyoda, A., Fujiyama, A., Kawabata, H., and Suzuki, T. (2014). A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. *Genome Res.* *24*, 522–534.
- Savva, Y.A., Rieder, L.E., and Reenan, R.A. (2012). The ADAR protein family. *Genome Biol.* *13*, 252.
- Seeburg, P.H., Single, F., Kuner, T., Higuchi, M., and Sprengel, R. (2001). Genetic manipulation of key determinants of ion flow in glutamate receptor channels in the mouse. *Brain Res.* *907*, 233–243.
- Shaffer, S.M., Joshi, R.P., Chambers, B.S., Sterken, D., Biaesch, A.G., Gabrieli, D.J., Li, Y., Feemster, K.A., Hensley, S.E., Issadore, D., et al. (2015). Multiplexed detection of viral infections using rapid in situ RNA analysis on a chip. *Lab Chip* *15*, 3170–3182.
- Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*.
- Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M., et al. (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods* *12*, 323–325.
- Sommer, B., Köhler, M., Sprengel, R., and Seeburg, P.H. (1991). RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* *67*, 11–19.
- Sul, J.-Y., Wu, C.-W.K., Zeng, F., Jochems, J., Lee, M.T., Kim, T.K., Peritz, T., Buckley, P., Cappelleri, D.J., Maronski, M., et al. (2009). Transcriptome transfer produces a predictable cellular phenotype. *Proc. Natl. Acad. Sci. U. S. A.* *106*, 7624–7629.
- Sunwoo, H., Dinger, M.E., Wilusz, J.E., Amaral, P.P., Mattick, J.S., and Spector, D.L. (2009). MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* *19*, 347–359.
- Symmons, O., Chang, M., Mellis, I.A., Kalish, J.M., Park, J., Suszták, K., Bartolomei, M.S., and Raj, A. (2019). Allele-specific RNA imaging shows that allelic imbalances can arise in tissues through transcriptional bursting. *PLoS Genet.* *15*, e1007874.
- Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* *126*, 663–676.
- Takahashi, K., and Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.* *17*, 183–193.
- Tomaru, Y., Hasegawa, R., Suzuki, T., Sato, T., Kubosaki, A., Suzuki, M., Kawaji, H., Forrest, A.R.R., Hayashizaki, Y., FANTOM Consortium, et al. (2014). A transient disruption of fibroblastic

- transcriptional regulatory network facilitates trans-differentiation. *Nucleic Acids Res.* 42, 8905–8913.
- Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Südhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* 463, 1035–1041.
- Wallace, E.W.J., and Beggs, J.D. (2017). Extremely fast and incredibly close: cotranscriptional splicing in budding yeast. *RNA* 23, 601–610.
- Wang, I.X., So, E., Devlin, J.L., Zhao, Y., Wu, M., and Cheung, V.G. (2013). ADAR regulates RNA editing, transcript stability, and gene expression. *Cell Rep.* 5, 849–860.
- Whitney, N.P., Peng, H., Erdmann, N.B., Tian, C., Monaghan, D.T., and Zheng, J.C. (2008). Calcium-permeable AMPA receptors containing Q/R-unedited GluR2 direct human neural progenitor cell differentiation to neurons. *FASEB J.* 22, 2888–2900.
- Wong, S.K., Sato, S., and Lazinski, D.W. (2003). Elevated activity of the large form of ADAR1 in vivo: very efficient RNA editing occurs in the cytoplasm. *RNA* 9, 586–598.
- Yamashita, T., Tadami, C., Nishimoto, Y., Hideyama, T., Kimura, D., Suzuki, T., and Kwak, S. (2012). RNA editing of the Q/R site of GluA2 in different cultured cell lines that constitutively express different levels of RNA editing enzyme ADAR2. *Neurosci. Res.* 73, 42–48.
- Yoshida, M., and Ukita, T. (1968). Modification of nucleosides and nucleotides. VII. Selective cyanoethylation of inosine and pseudouridine in yeast transfer ribonucleic acid. *Biochim. Biophys. Acta* 157, 455–465.
- Zhang, Z., and Carmichael, G.G. (2001). The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* 106, 465–475.
- Zhou, H., Morales, M.G., Hashimoto, H., Dickson, M.E., Song, K., Ye, W., Kim, M.S., Niederstrasser, H., Wang, Z., Chen, B., et al. (2017). ZNF281 enhances cardiac reprogramming by modulating cardiac and inflammatory gene expression. *Genes Dev.* 31, 1770–1783.
- Zhu, W.-Z., Xie, Y., Moyes, K.W., Gold, J.D., Askari, B., and Laflamme, M.A. (2010). Neuregulin/ErbB Signaling Regulates Cardiac Subtype Specification in Differentiating Human Embryonic Stem Cells Novelty and Significance. *Circ. Res.* 107, 776–786.
- Zhu, W.-Z., Van Biber, B., and Laflamme, M.A. (2011). Methods for the derivation and use of cardiomyocytes from human pluripotent stem cells. *Methods Mol. Biol.* 767, 419–431.