Tools for engineering multicellular systems through cell sorting and cell state detection

By

Casper Nørskov Enghuus

B.Sc.Eng, Human Life Science Engineering, Technical University of Denmark (2013)
M.Sc.Eng, Bioinformatics and Systems Biology, Technical University of Denmark (2015)

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Signature of Author .....
Casper Nørskov Enghuus
Microbiology Graduate Program
April 16, 2021

Certified by .....
Ron Weiss, PhD
Professor of Biological Engineering & Professor of Electrical Engineering and Computer Science, MIT
Thesis Supervisor

Accepted by .....
Jacquin C. Niles
Associate Professor of Biological Engineering
Chair of Microbiology Program

THESIS COMMITTEE:

Rudolf Jaenisch, PhD
Chairman, Thesis Committee
Professor of Biology, MIT

Ron Weiss, PhD
Thesis Supervisor
Professor of Biological Engineering & Professor of Electrical Engineering and Computer Science, MIT

Domitilla Del Vecchio,
Thesis Committee Member
Professor of Mechanical Engineering, MIT

Wilson Wong
Thesis Committee Member
Associate Professor of Biomedical Engineering
Boston University

Tools for engineering multicellular systems through cell sorting and cell state detection

By

Casper Nørskov Enghuus

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

The human genome, the genetic blueprint that every cell in our body follows, encodes approximately 20,000 genes. Through complex regulation of these genes, each cell is able to play the role it needs within our body. Synthetic biology, an emerging field in biology, seeks to expand on this blueprint and create cells with novel functions. The aim of this thesis is to provide methods that expands our ability to engineer and control multicellular systems by detecting and rewriting the cell state.

We first develop a method that enables the creation of a synthetic cell state to control morphogenesis. Using inducible expression of recombinases, we show this approach can induce a cell to commit to one of two mutually exclusive cell states. By regulating the expression of recombinases, we are able to control the distribution of cell states within an initially monoclonal and homogenous population of cells. We use the induction of a synthetic cell state to control morphogenesis by cell state-specific expression of homotypic cadherins which controls the cell's adhesive properties. This enables us to create a large number of different shapes and control morphogenesis.

Secondly, we develop a library-based approach for cell state-specific gene regulation. We design a set of 6,107 Synthetic Promoters with Enhanced Cell-State Specificity (SPECS), and identify several SPECS with spatiotemporal specificity during the programmed differentiation of stem cells, as well as SPECS that are highly specific for breast cancer and glioblastoma stem-like cells.

Thirdly, we develop a method that allows detection of endogenous gene expression without modifying the endogenous gene itself. We show that placing a regulatory RNA downstream of a terminator allows for expression of the regulatory RNA, and demonstrate this method for miRNAs and gRNAs.

Together, this thesis develops methods to create synthetic cell states that can be used to control morphogenesis, and provides tools to detect endogenous cell states which can serve as inputs to control gene regulatory networks.

Thesis Supervisor: Ron Weiss, PhD Title: Professor of Biological Engineering and Professor of Electrical Engineering and Computer Science, MIT

Acknowledgements

“Grad school is a marathon”. I have lost track of the number of times I have heard that cliché, as well as the number of times I have told it. Writing this thesis, I realize it is wrong. Grad school has certainly been a test of endurance, and I have had my share of highs and lows throughout the process. But if the metaphor is true then it implies a race, a lonely competition between you and others to be the first that crosses the finish line. As I write this thesis, I am about to cross the vaunted finish line. But the “marathon” that was grad school was not a race. It was not a lonely process. And it certainly did not feel like a competition.

The first steps of this “marathon” would not have been possible if not for my undergrad professor, Morten Sommer. He welcomed me into his lab, where I had the pleasure of receiving mentorship from him and one of his grad students, Hans Genee who taught me the fundamentals of molecular biology and synthetic biology. They gave me my first research experience and without their help, and them believing in me, I would never have found myself on the path of a PhD.

To Ron, who kept me company throughout this “marathon” or test of endurance, and always provided a path forward. You welcomed me into your lab and gave me the opportunity to dedicate myself to synthetic biology. From the first day I joined your lab, you have provided a home for me where I could grow and learn. Throughout my time in your lab, you have continuously helped by challenging my thinking, you have taught me to think like a synthetic biologist and engineer, and you even helped shape how I communicate. You have taught me a lot over the years and helped cement the values that now drives me. I am grateful for all the experience and all the help I have received from you over the years. They have been some of the most important lessons I have ever learned. Thank you.

I would especially like to thank my thesis committee members, Domitilla Del Vecchio and Rudolf Jaenisch. You helped me narrow down the important steps and prioritize my

research when I felt lost and without a path forward. Without your gracious feedback and advice, I would not be writing this thesis.

The opportunity to study at MIT would not have been possible without my friend and mentor from my Master's degree, Daniel B. Goodman. If not for your passion for science, your patience, and your unwavering support, I would not have found myself at MIT on the path towards a PhD. You taught me the power of combining computational and experimental science, and the skills to do so. You instilled in me a deep love for programming that has brought many joyful moments in and out of the lab. In fact, I owe the plots in this thesis to you.

My journey through grad school would not have been possible without my friends. A special thanks to Naomi Lin who always believed in me when I doubted myself. Your kindness, support, and adventurous spirit kept me going when I was struggling. I can always rely on you to challenge me, to call me out if I am not at my best, and you continue to help me grow. But perhaps the most important thing you taught me: climbing. This sport has been invaluable to me as a place to recover and rethink. Thank you for always having my back.

My time in the Weiss lab would not have been the same if not for Jesse Tordoff. Your positive energy and friendship made every day in lab a joy. You are a brilliant scientist, and I am grateful for the countless times I could bounce ideas off you, and have someone take a critical look at my data. Even though it took four years, I am glad that we finally found a project to collaborate on, and I owe an entire chapter in this thesis to you. I continue to be amazed that we shared a bench for four years, yet only worked next to each other three times! And to borrow a phrase from you: you are a beautiful angel Jesse.

To Jasmine Qin, a close friend and constant shoulder to lean on. You are always there for me when I struggle, and you help me get back on my feet when I fail. I am grateful for our countless conversations about leadership and vulnerability, how you help me reflect and see my experiences from a different perspective, and how you help me grow. I know I can always count on you to support me and I am grateful to have a friend like you.

To Xia Lapides who provided a refugee when COVID hit, our many long conversations on your porch during the Spring and Summer of 2020, and the countless loaves of sourdough and Danish rye bread you have provided. When the stress and anxiety of finishing a PhD, applying for jobs, and navigating a society that had closed down was getting too much, you provided a break and a space to relax and wind down. And to Oona, your puppy: never have I seen such unconditional love. She was truly a highlight of 2020.

Throughout my PhD, I relied on climbing as a place to relax, reflect, and rethink. I am grateful to the company of Ray and Sarah Li, Xia Lapides, and Naomi Lin for being excellent climbing partners. You provided countless fun, and challenging experiences in the various climbing gyms around Boston.

During my last year of my PhD, I had the opportunity to work with UrSure and help fight the HIV epidemic. I am deeply grateful to Shane Hebel, Eli Kahn-Woods, Caitlin Conyngham, and Giffin Daughtridge for this amazing experience and opportunity. I am deeply humbled by how passionate and mission-driven you are in everything you do, and I am thankful for your mentorship and support as I was learning everything about the field of HIV prevention. The work I did with you is among the work I am most proud of; thank you for welcoming me to the team and giving me this amazing opportunity.

To the many amazing people I had the opportunity to work with through the Consulting Club at MIT. Thank you to Jaclyn Mallard for believing in me and providing me the opportunity to join your consulting team: you gave me a life-changing opportunity! To Sophie Bertram, Wenbi Shcherbakov-Wu, Max Robinson, Stephanie Jones, Chensu Wang, Darryl Fong, Kevin Dervishi, and the many other consultants whom I have had the wonderful experience to work with to help companies in Boston and beyond.

Thank you to my collaborators for making so many of these projects possible. To Sebastian Palacios for teaching me how to work with stem cells, and work with our organoids; to Lior Nissim and Ming-Ru Wu for the amazing feedback, drive, and ideas to get our paper out; to Jeremy Gam for helping me identify my first project in the lab, and

to Fabio Callendo for helping me finish the project. To Matt Lima and Selam Mamo for your amazing help with experiments.

A very special thanks to the entire Weiss lab for making these past years such an incredible experience. To Nicholas Delateur, a good friend, and always there when I needed you. I am thankful for the countless breaks you provided, your perspective, and for always being there when I had something to discuss. To Noreen Wauford, and amazing, and helpful friend always willing to lend a hand. Congratulations for finally beating me to lab – on the one day where I came in at noon to spend five minutes to submit samples for sequencing. To Ross Jones, Allen Tseng, Bre DiAndreth, Jin Huh for providing me with countless DNA parts and making my research possible. Thank you to Roger Guevara-Flores, Cammie Haase-Pettingell, Olga Parkin, and Darlene Ray for your support over the years in keeping the lab running, scheduling meetings, handling orders, and doing so many things behind the scenes that made it so much easier to be a PhD student.

While I do not often get to visit Denmark, I am grateful to my friends and family there who always made it feel like home, and made it so easy to visit. A special thanks to Mette Halmø Rasmussen, Pia Ebbesen, and Mads Malte for countless good conversations, dinners, and a place to stay in Copenhagen. Thank you to my parents and sister for always being there to support me. I would not be where I am today if not for you.

Grad school was a lot of hard work, mentally, physically, and emotionally, and I would never have been able to do it alone. So many people played a role in my ability to complete grad school. I am grateful to each and every one of you, and you have my deepest thanks for helping me on this journey.

Table of contents

Abstract 3

Acknowledgements 5

List of figures 12

List of tables 14

Chapter 1. Introduction and background 15

1.1 Engineering cells for therapies 15

1.2 Engineering organoids as human disease models 16

1.3 Coupling endogenous and synthetic gene circuits 18

1.4 Thesis overview 20

Chapter 2. Engineered synthetic cell states for morphological control of cell aggregates 22

2.1 Summary of chapter 2 22

2.2 Introduction and aim 23

2.3 Stochastic recombinase expression for tissue imaging and lineage tracing 24

2.4 Cadherins control cell sorting 26

2.5 Circuit design 29

2.6 Cell states can be controlled by titrating inducers 32

2.7 Cell state bifurcations generates different morphological structures 37

2.8 Discussion 43

2.9 Methods 44

2.9.1 Plasmid construction 44

2.9.2 Cell culture and transfections 44

2.9.3 Pattern formation assay 45

2.9.4 Microscopy and image analysis 45

Chapter 3. Synthetic promoters for cell-type specific transcriptional regulation 53

3.1 Summary of chapter 3 53
3.2 Introduction and aim 54
3.3 SPECS show distinct activities in an organoid model 56
3.4 The combined pipeline identifies cancer-specific SPECS 62
3.5 SPECS identify glioblastoma stem-like cells 67
3.6 Discussion 69
3.7 Methods 70
3.7.1 SPECS Library Construction 70
3.7.2 Cell culture and cell lines 71
3.7.3 Virus production and cell line infection 72
3.7.4 Lentiviral library introduction to cells of interest 73
3.7.5 Flow cytometry 73
3.7.6 FACS sorting 73
3.7.7 Next-generation sequencing 74
3.7.8 Pre-processing of NGS data 74
3.7.9 Fluorescence estimation 75
3.7.10 Differentiation and infection of liver organoids 76
3.7.11 Shotgun cloning promoter identification 78
3.8 Supplementary Information 79
3.8.1 Supplementary Text 1 - Features for machine learning 89
3.8.2 Supplementary Text 2 - Identifying GSC specific promoters 89
Chapter 4. Transcriptional regulating using novel post-PAS RNA 90
4.1 Summary of chapter 4 90
4.2 Introduction and aim 90
4.3 Designing constructs for post-PAS RNA expression 92
4.3 miRNAs can be expressed downstream of a polyA signal 96
4.4 The effect on upstream gene expression can be minimized by removing a splice site 100
4.4 gRNAs can be expressed downstream of a polyA signal 108
4.5 Discussion 112

4.6 Methods 113
4.6.1 Plasmid construction 113
4.6.2 Cell culture and transfections 114
4.6.3 Flow cytometry 114
4.6.4 Data analysis 114
Chapter 5. Future Directions 119
Bibliography 121

List of figures

Figure 2-1: Schema for assembling larger multicellular structures 28

Figure 2-2: Schema for recombinase-based cell state control 30

Figure 2-3: GMM to identify cell populations 34

Figure 2-4: Distribution of cell states for the *XFP* circuit 35

Figure 2-5: Shapes formed by induction of the *Cadherin* circuit 39

Figure 2-6: Shapes formed by induction of the *Cadherin-p27^{Kip1}* circuit 42

Figure S2-1: Distribution of EBFP and EYFP cell states for the *XFP* circuit 47

Figure S2-2: Distribution of EBFP and mKate cell states for the *XFP* circuit. 48

Figure S2-3: Brightfield images for the shapes formed by induction of the *Cadherin* circuit 50

Figure S2-4: Brightfield images for the shapes formed by induction of the *Cadherin-p27^{Kip1}* circuit 52

Figure 3-1: The experimental and computational pipeline for identifying cell state-specific promoters 59

Figure 3-2: Synthetic promoters exhibit distinct temporal and spatial behavior in organoid cultures derived from iPSCs 60

Figure 3-3: Machine-learning based prediction model can efficiently predict cell state specificity 65

Figure 3-4: Promoter activities in glioblastoma stem-like cells (GSCs) and serum-cultured glioblastoma cells (ScGCs) 68

Figure S3-1: The activities of promoters identified by the Top 5% approach 79

Figure S3-2: Relationship between fluorescence and NGS read counts 81

Figure S3-3: Machine-learning features 82

Figure S3-4: Observed vs. predicted fluorescence for the 1st round of the machine learning predictions 84

Figure S3-5: Model performance and feature importance for the 2nd round of the machine learning predictions. 86

Figure 4-1: The mammalian termination signal 94

Figure 4-2: Construct design to test post-PAS RNAs 95

Figure 4-3: Post-PAS miRNA repression for different terminators 99

Figure 4-4: Post-PAS miRNAs might affect EYFP expression 102

Figure 4-5: miRNA structures 104

Figure 4-6: Removing splice sites from the miRNAs restore EYFP expression 105

Figure 4-7: Splice sites might not be necessary for miRNA expression 107

Figure 4-8: gRNAs can be expressed downstream of a terminator 110

Figure 4-9: Testing cryptic gRNA expression 111

Figure S4-1: Dynamic range of post-PAS miRNA repression for different terminators 116

Figure S4-2: EYFP/EBFP expression for post-PAS miRNAs 118

List of tables

Table S3-1: Cell state-specific promoters derived from the validation set 87

Table 4-1: Terminators tested for post-PAS RNA expression and the position of the post-PAS RNA cassette 98

Chapter 1. Introduction and background

1.1 Engineering cells for therapies

Humans develop from a single fertilized egg into an organism comprised of trillions of cells. Each cell, derived from the same ancestor, develop to become part of a tissue and perform a highly specialized functions such as muscle cells that enable movement, immune cells which protect the body against foreign molecules and cells, red blood cells that transport oxygen throughout the body, and the vast number of different cells that make up our brain and provides us with higher cognitive function. This diverse set of cellular functions in the human body is enabled through cell type-specific regulation of our approximately 20,000 genes. The exact combination of active genes, their expression levels, protein and mRNA location, post-transcriptional and post-translational modifications etc., all play into defining the exact cell state which in turn defines a cell's function.

Interference, changes, or unintended signaling within the endogenous gene regulation can have significant consequences. Human viruses are able to rewire and repurpose our cellular machinery to replicate their viral genomes with lethal consequences as in the case of HIV. Unintended signaling as observed in autoimmune diseases can have similar drastic consequences: in type I diabetes, the immune system recognizes the insulin-producing pancreatic β-cells and attacks them. The consequence is an inability to produce insulin and requires a lifetime of insulin therapy for survival. Cancer is an extreme example of dysregulation where accumulation of genomic damage within a cell can lead to uncontrolled cell divisions, tissue invasion by cancer cells, and avoidance of the body's ability to repair, contain, or kill dysregulated cells that threaten the body.

The observation that the cellular program is not static, but rather dynamic, and that Nature has already developed multiple tools to change it, implies that cells can be repurposed to our needs. Early human civilization had already adapted the natural processes of bacteria

and yeast for fermentation and baking, but with scientific advances, especially in the fields of genome engineering, systems biology, metabolic engineering, and developmental biology, we can now take a much more direct role in reprogramming and repurposing cells far beyond their current capabilities. This has enabled the repurposing of yeast to produce the precursor of the antimalarial drug artemisinin,[1,2] using engineered T cells as highly effective cancer treatments,[3,4] organoids as drug screening platforms,[5,6] vaccine development,[7,8] and regenerative medicine.[9,10]

Synthetic biology, an emerging discipline that draws on inspiration from electrical engineering and computer science, aims to engineer cells beyond what they were initially capable of. While early efforts of synthetic biology aimed to demonstrate that the cell could be treated as a microscopic computer that could be reprogrammed,[11,12] the field is now poised to expand on current advances in cell based therapies by increasing safety, efficacy, and reproducibility by optimizing the genetic program driving cell behavior.[13,14]

1.2 Engineering organoids as human disease models

Organoids are complex multicellular structures derived from stem cells that undergo differentiation and cell sorting *in vitro* to show an organ-like phenotype. This makes them attractive models for the study of human diseases, especially when there are no good animal models. One example is brain diseases where the brain from other primates, our closest relatives, show a significant difference in gene expression, tissue organization, and development compared to the human brain.[15] By using protocols to differentiate stem cells into an organoid with a brain-like phenotype, organoids provide an alternative to animal models for studying brain diseases,[16] including Alzheimer's[17–19] and schizophrenia,[20] that might more closely mimic the disease as it manifests in humans.

The diversity of cell types and structure of organoids make them significantly more realistic models of human diseases compared to simpler and more homogenous tissue

cultures. In addition, organoids offer to address certain shortcomings of animal models such as species-specific genetic differences, and differences in cell type composition and organization of the tissue. Importantly, by using patient-derived stem cells to develop the organoid, organoids can be used for personalized medicine.[6]

Despite the benefits of using organoids as human disease models, there are challenges that limits their use. The self-organizing process that is the very foundation of organoid development might also be its greatest weakness. While organoids require few external cues, the deep reliance on intra-organoid cell-cell signaling, cell autonomous differentiation, and self-organization makes it extremely sensitive to any changes in cell state or the few external cues provided, and high organoid-to-organoid variability is a significant challenge.[16,21] The effect of the cell state on organoid development is exemplified by the variability between organoids developed from different stem cell lines,[22] and even protocols with high reproducibility can show this effect.[23] Advances to more accurately reprogram and dedifferentiate somatic cells into induced pluripotent stem cells[24] (iPSCs) might address some of these shortcomings and increase reproducibility between organoids by reducing cell state variability. Other challenges and limitations include development of the proper cell types, correct cell sorting, lack of vasculature, and challenges with cell maturation to create organoids that mimic adult tissue.[25–27]

Carefully controlling gene expression when reprogramming somatic cells offer the potential to reduce cell state variability for the generation of more reproducible organoids.[24] In a similar vein, genetic reprogramming might offer the same advantages by programming a specific cell state, behaviors or outcomes. This has been applied when the endogenous signaling pathway is not known. Not knowing the signaling pathway that led to Nkx2-1 and Pax8 co-expression, Antonica *et al.*[28] used transient overexpression of Nkx2-1 and Pax8 to direct mouse embryonic stem cells into thyroid follicular cells that subsequently organized into thyroid follicles. Genetic reprogramming of a sub-population of cells has been applied to address the lack of vascularization in cortical organoids. By ectopically expressing ETV2 in a small population of cells within a developing cortical organoid, Cakir *et al.*[29] was able to reprogram these cells into endothelial cells that would

organize to form vasculature within the organoid. However, the cells to be differentiated would have acquired different cell states and be randomly distributed throughout the organoid potentially increasing the variability of the experiment, and potentially leading to challenges in transdifferentiation.[30] By factoring in the cell state and using carefully controlled gene expression, Saxena *et al.*[31] was able to differentiate iPSCs into glucose-sensitive insulin-secreting β-like cells. Together, these approaches highlight how ectopic gene expression can be used to program the formation of organoids or alter the subpopulations within them.

While induced differentiation might ensure the generation of desired cell types, the resulting organoid still depends on self-organization and autologous cell sorting which can be a separate challenge.[25,32] Genetic circuits have been developed that enables synthetic cell sorting. In a study by Toda *et al.*,[33] they used the cell-cell signaling system synNotch to control expression of cadherins. This enabled them to control cell sorting behavior and artificially induce symmetry breaking. Tordoff *et al.*[34] used a similar approach and showed that the differential expression of cadherins could be used to reproducibly form a large number of different shapes. Although this work was done in minimally adhesive cells which might not reflect cells within a developing organoid, it shows the potential that genetic circuits can be used to reliably control cell sorting through expression of adhesion-molecules, thereby enabling increased control over organoid development. By coupling the expression of specific adhesion-molecules to specific cell states and points in development, cell sorting can be made cell state-specific.

1.3 Coupling endogenous and synthetic gene circuits

Single-cell RNA seq has provided significant insights into different cell states and their regulation.[35,36] This has significantly impacted developmental and organoid biology by providing a more detailed understanding of the gene regulatory networks (GRNs) that define and drive different cell states during differentiation, as well as benchmarking

organoids to developing or adult tissue to understand their differences and limitations.[23,37–42] The increased understanding of endogenous cell states provides an opportunity to improve *in vitro* cell differentiation through better understanding of the central transcriptional pathways of cell differentiation and a more clearly defined end-state. By mimicking the endogenous GRNs driving differentiation, it might be possible to build genetic circuits to compose a synthetic GRN, and reprogram cells with this to facilitate or induce a cell state transition. The most simple form of this is the overexpression of one or more transcription factors which has led to a large number of organoids.[27–29,43–47] However, using more advanced synthetic GRNs to fine-tune the differentiation program to support and direct the cell through transitional cell states or to more accurately define the end-state, might lead to differentiated cells that more accurately mimic adult tissue.

Synthetic GRNs have significant potential in supporting or directing organoid development. However, they are unlikely to act in isolation, and the activity of specific components within the synthetic GRN will likely depend on the exact cell state. For instance, maturation factors such as MafA in the case of β-cell maturation, benefit from being expressed once the cell has committed to the endocrine lineage within the pancreas.[31] This necessitates methods to couple synthetic GRNs to endogenous GRNs. Several methods currently exist to use or provide outputs from endogenous GRNs that can serve as inputs into synthetic GRNs. These include endogenous promoters to recruit cell type-specific transcription factors;[48] transcriptional fusions of guide RNAs[49–51] to endogenous genes for CRISPR-based gene regulation;[52,53] cell state-specific miRNAs or synthetic miRNAs fused to endogenous genes;[54] and co-expressing proteins with endogenous genes.[55,56] While these methods can be used either individually or in combination to detect specific cell states, they often involve manipulation of endogenous genes which risks disrupting the endogenous GRN, or they might fail to factor in certain types of regulation such as chromatin state or distal enhancers that play an important role in endogenous gene regulation. As our understanding of endogenous GRNs continue to expand, it becomes increasingly important to develop minimally invasive methods to read and respond to changes in the cell state.

1.4 Thesis overview

The aim of this thesis is to develop methods to detect and manipulate cell states for engineering complex multicellular systems through controlled cell sorting.

In Chapter 2, we develop a recombinase-based method to define a synthetic cell state of a monoclonal and homogenous cell population based on externally provided inputs. Each of the two cell states is defined by the expression of a unique, homotypic cadherin. By titrating the inputs, we control the ratio between the different cell states. Through homotypic cadherin expression, we are able to induce controlled cell sorting between the two different cell populations. We find that this method can create a wide array of different shapes and enables spatial organization of an initially homogenous, monoclonal cell population.

In Chapter 3, we develop a method for cell-state specific gene expression through synthetic promoters. We build a library of synthetic promoters with enhanced cell state-specificity (SPECS) from two databases of transcription factor binding motifs. We apply this library to a liver bud-like organoid and show that we can identify SPECS that show differences in spatiotemporal characteristics and expression strength, as well as SPECS for breast cancer and glioblastoma stem-like cells.

In Chapter 4, we develop a second method to detect the cell state by expressing regulatory RNAs downstream of a terminator. We show that this method can be used to express both miRNA and gRNA without impacting upstream gene expression. By placing the regulatory RNA downstream of the gene, this method can couple the cell state to a synthetic gene regulatory network without requiring any changes to the gene itself. Current work is ongoing to apply this to directed stem cell differentiation.

Combined, this thesis aims to develop principles for engineering complex multicellular populations. Using external inputs, we are able to separate a homogenous and

monoclonal cell population into two different cell states, and enable these cell states to sort into larger multicellular structures. We then develop two different approaches that can be used to detect the cell state and initiate a synthetic gene regulatory network. These methods have the potential to detect a specific cell state within a developing organoid and initiate a program for cell sorting or differentiation, thereby enabling engineering of complex multicellular structures.

Chapter 2. Engineered synthetic cell states for morphological control of cell aggregates

This chapter is a collaboration with Dr. Jesse Tordoff and a continuation of previous work.[34]

2.1 Summary of chapter 2

Tissue is composed of different cells, and the function of the tissue is correlated to the different cell states within the tissue, their distribution, and their relative organization. As such, the ability to control and engineer cell states, their distribution, and their organization has important implications for tissue engineering. Despite this importance few tools exist to engineer multicellular systems. Here, we propose a recombinase-based method for controlling cell state and cell adhesiveness as a means to control morphogenesis. Using inducible recombinase expression, we are able to control the probability that a cell commits to one of two mutually exclusive cell states. We show that by titrating recombinase expression, we can control the distribution of cells in a given state. By coupling each cell state to the expression of a unique, homotypic cadherin, we enable cell sorting based on cadherin expression and create a set of distinct morphologies that depend on the adhesiveness of a cell, the total number of cells that express a specific cadherin, and the ratio between cells expressing different cadherins. Taken together, we provide a method to control the morphology of larger multicellular structures such as organoids, starting from a single monoclonal, and homogenous population of cells.

2.2 Introduction and aim

Around 3 billion years ago, the first signs of multicellularity arose and since then, multicellular organisms are thought to have evolved independently at least 25 times.[57] At the core of multicellularity is the ability to adhere, communicate, and coordinate. This has led to the evolution of tissues that are able to carry out specialized functions such as movement produced by the cooperation between neurons and muscles, protection from external factors by the skin, and cognitive functions enabled by our brains. While early morphology-based estimates suggested humans are comprised of approximately 200 different cell types, advances in single-cell RNA sequencing using molecular features has shown that this number is significantly higher[35] emphasizing the diverse functions and demand for cells to specialize required to sustain a larger multicellular organism.

Multicellularity is at the very core of our human biology, yet the tools we have to engineer multicellularity is limited. Advances in cell differentiation and the development of organoids have shown that complex multicellular organization can be recreated in the lab.[27,58,59] However, these approaches rely on the autonomous organization carried out by the cells, and is outside our direct control. While bioprinting and hydrogel scaffolds offer some control over organoid patterning,[26,60] they are limited to the initial conditions of the experiment.

In an effort to control cell state and organization dynamically, Matsuda *et al.*[61] used transgenic Delta-Notch signaling to create a lateral inhibition system capable of spatially bifurcating into Delta-positive and Notch-active cell populations. Building on this work, Toda *et al.*[33] used synthetic Notch signaling coupled with cadherin expression to create a lateral inhibition circuit capable of driving cell state bifurcation, ultimately leading to robust self-organization into multidomain structures. A limitation of Notch-based cell type bifurcations is that the cell type is dependent on the neighboring cells and subject to change as cells rearrange. The consequence is that larger structures in which one cell

becomes separate from another will not be stable. While this can be desirable in some scenarios, others require a permanent commitment to cell state; migrating or circulating cells such as immature immune cells are one example where the cell cannot rely on neighboring, healthy cells to define its cell state. To overcome this limitation, Tordoff *et al.*[34,62] mixed different cell types with stable cadherin expression and showed that by varying four parameters, cadherin expression, adhesion timing, cell population ratio, and size, a diverse set of morphological structures could be produced.

The aim of this chapter is to develop a method to controllably induce a synthetic cell state that can be used to drive cell sorting behavior as a function of cell adhesiveness, and the ratio between different cell states in order to drive precise development of morphological structures. The method is centered around a recombinase-based genetic switch regulating cell state and responding to external inputs. By controlling the external inputs, the divergence of cell states can be controlled. With this, we demonstrate the method can control cell-state bifurcations with distinct behaviors from a monoclonal and homogenous population of cells. This enables synthetic symmetry breaking, a unique and critical event in development of multicellularity, and permanent commitment to distinct cell states that enables self-organization into distinct morphological structures.

2.3 Stochastic recombinase expression for tissue imaging and lineage tracing

Multicellular organisms have demonstrated it is possible to create diversity from a homogenous population of cells. While this divergence in cell states is a carefully orchestrated process in some cells, for others cell state commitment is based on chance.[38] Common to both is the challenge to carefully and controllably decide exactly when and where any given cell commit to a particular state. In order to engineer biology, it is therefore critical to have methods to induce and stably maintain different cell states.

To create stable cell states, we apply a tool that has played an important role in genetic engineering: recombinases.[63] Recombinases can be used to insert, reverse, or delete DNA sequences.[64,65] This has led to a number of different applications including recording the cell state,[66] and complex computational logic.[67] A recent computational model by Appleton *et al.*[68] provided a framework for designing larger multicellular structures starting from a single cell. Their framework uses recombinase-based counters and adhesion molecules to drive the formation of smaller multicellular aggregates for subsequent modular assembly.

By using recombinases with orthogonal recombinase sites, and positioning the sites such that they are mutually exclusive, recombinases can be used to create genetic diversity in an otherwise monoclonal population. This strategy of mutually exclusive recombinase sites was applied in the “Brainbow” project to create cellular diversity which enabled single cell and lineage tracing by inducing recombinases to create a diverse set of heritable, fluorescent tags for each individual cell.[69–72] While the “Brainbow” project relied on recombination to occur at approximately equal probability for each of the possible outcomes, Wang *et al.*[73] developed a system to control the probability of recombination between two mutually exclusive pairs of recombinase sites by varying the distance between the sites. These efforts were focused on imaging or proof-of-concepts, and did not lead to any functional changes in cell behavior. Movahedi *et al*.[74] expanded on this work by using recombinases to create functional mosaicism in a mouse model by expressing dominant negative G protein-coupled receptors in a fraction of cells that could then be compared to differentially marked control cells in the same organism.

In this chapter, we expand on these recombinase-based approaches to create genetic diversity in a monoclonal and homogenous population of cells, and provide a set of recombinase-based circuits used to induce permanent and genetically encoded cell states. These cell states are mutually exclusive and irreversible, meaning that the cell can be in one state, and one state only, and that the other state(s) are not accessible once the initial cell state has changed. By regulating the expression of these recombinases using small molecule inducers, we are able to control the distribution of cell states.

2.4 Cadherins control cell sorting

Cell sorting is the process in which cells can physically rearrange themselves to form clusters of distinct cell populations, and it is driven by differences in tissue surface tension between populations of cells.[75] Tissue surface tension is in turn primarily determined by an interplay between cell-cell adhesion and cell cortex tension.[76,77] While the cytoskeleton impacts cortex tension, cadherins and integrins play a large role in the cell's adhesive properties. Here, we focus on cadherins as a driver of cell sorting.

Cadherins are a superfamily of glycoproteins that are involved in homotypic cell-cell adhesion, and are essential for holding cells together and creating tissue boundaries during development.[78,79] The homotypic propensity of cadherins and their ability to modify surface tension have previously been applied for synthetic morphogenesis, creating a vast array of structures including an inner sphere with an outer layer of cells,[80] segmented and separate populations,[79] and maze-like and intertwined populations.[81] These structures are the results of autonomous cell sorting that occurs over the course of hours when cells are randomly mixed in 2D or 3D. More recently, Toda *et al*.[33] showed that regulated cadherin expression can be used to induce symmetry breaking and mimic the type of cell-sorting that occurs during the early stages of gastrulation. In a systematic approach to understand the type of cell sorting that occurs, Tordoff *et al.*[62] modeled the effect total cell number and the ratio between cells with different levels of adhesion strength had on cell sorting. They demonstrated how large populations of cells show incomplete sorting that form predictable and reproducible patterns that remain stable for multiple days. In follow up work, Tordoff *et al.*[34] further explored the design principles and types of shapes that could be engineered by varying four parameters: the total number of cells, the ratio between different cell types, the type of cadherin expressed by each cell type, and the timing of adhesion.

To control cell sorting, we take advantage of the homotypic binding affinities of two different classic cadherins, E-cadherin (a type I cadherin encoded by Cdh1) and K-cadherin (a type II cadherin encoded by Cdh6), previously shown to have no heterotypic binding affinity.[79] By expressing them in Chinese hamster ovary (CHO) cells, cells known to have negligible native cadherin expression,[82] we can separate the cell populations as a function of the type of cadherin the cell has been engineered to express.

In this chapter, we combine the recombinase-based strategy for creating controllable genetic diversity in a monoclonal population of cells with the differential expression of cadherins to achieve functional diversity. Our method enables stable bifurcations in cell states, and we demonstrate how this can be used to drive the assembly of larger cell structures (Figure 2-1).

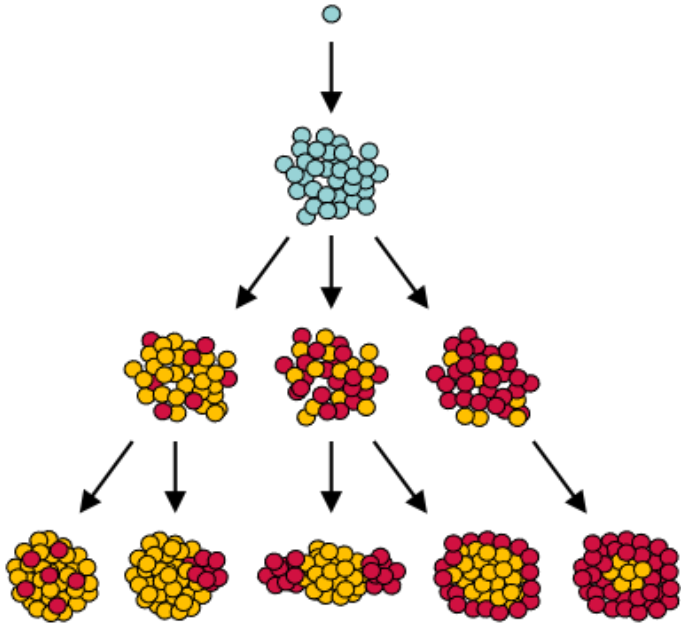


Figure 2-1: Schema for assembling larger multicellular structures. Starting from a monoclonal, homogenous population of cells (blue), we induce a change in cell state (red or yellow). Each cell state is defined by the expression of a unique, homotypic cadherin. By controlling the ratio between different cell states morphogenesis of larger multicellular structures can be controlled.

2.5 Circuit design

The circuit is designed around mutually exclusive recombinase sites that enables recombinase-mediated DNA excision to change state. By using mutually exclusive recombinase sites, the circuit can flip state once, and only once, and the final state depends on which recombinase is the first to complete recombination (Figure 2-2A). This is done by orienting the recombinase sites such that recombination results in excision of the DNA sequence between the recombinase sites, thereby deleting the recombinase sites required for the competing recombinase.

Recombination of the cell state circuits (Figure 2-2B) is facilitated by two serine recombinases, φC31 and Wβ, which have orthogonal attB and attP sites, works outside their original host,[65] and for with which we had good experience in our lab. The transcription of these recombinases is regulated by an inducible promoter that responds to either abscisic acid (ABA) or doxycycline (Dox), controlling φC31 and Wβ, respectively. By regulating the concentration of each inducer, the transcription rate of a given recombinase is increased leading to a higher concentration of recombinase within the cell. This in turn increases the probability a recombinase will be bound to a pair of recombinase sites and catalyze a change in cell state. In this way, titrating inducers changes the probability of the cell switching to a specific state. Three different circuits for synthetic cell states are tested: *XFP*, *Cadherins*, and *Cadherins-p27^{Kip1}*. For all circuits, the ground state or starting state, is characterized by the expression of EBFP. In addition to the ground state, each circuit has two mutually exclusive states defined by the expression of a different fluorescent protein, and the expression of a cadherin for the *Cadherin* and *Cadherin-p27^{Kip1}* circuit. One cell state of the *Cadherin-p27^{Kip1}* circuit has an additional component, co-expression of the cell cycle inhibitor p27^{Kip1}.[81]

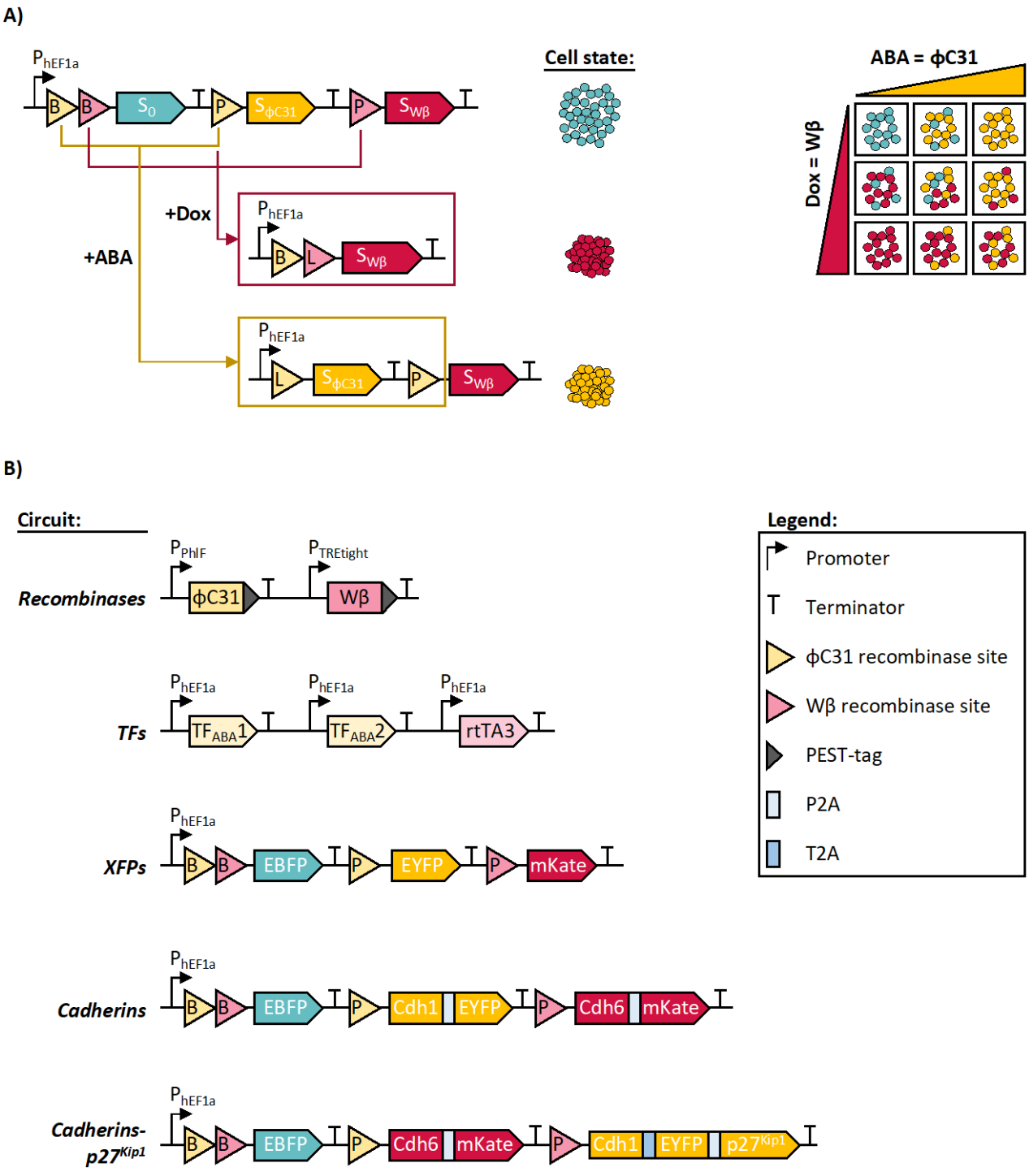


Figure 2-2: Schema for recombinase-based cell state control. (A) Recombinases bind their respective sites, attB and attP (triangles marked "B" and "P"; yellow = φC31, red = Wβ), and recombine them as shown ("L" indicates the recombined site), resulting in genomic rearrangement and an irreversible change in cell state, S. Here, the cell state is depicted as a change in color and adherence. Increasing the concentration of small molecule inducers increases transcription of the recombinase. An increased recombinase concentration increases the probability a recombinase is bound to its recombinase site with subsequent

recombination and a change in cell state. The right diagram illustrates the competition between recombinases and the resulting combination of cell states, as well as the low probability of recombination occurring at low levels of inducer. **(B)** Circuits used in this study. Cdh1 encodes E-cadherin, Cdh6 encodes K-cadherin.

2.6 Cell states can be controlled by titrating inducers

We first tested if recombinases with mutually exclusive sites could be used to generate two distinct populations from a single monoclonal population at different ratios determined by the addition of small molecule inducers. Using a CHO-K1 cell line with a landing pad (LP) in the Rosa26 locus,[83] we integrated the *Recombinase* and *TF* circuits by PiggyBac integration[84] followed by zeocin and blasticidin selection. The *XFP* circuit was integrated in this cell line using the BxB1 integrase. Following integration of the *XFP* circuit, we applied puromycin selection and FACS to establish a monoclonal cell line.

To test if the *XFP* circuit and inducible recombinase expression could be used to create distinct cell populations at controllable ratios, we added a combination of the small molecule inducers Dox and ABA to activate the Wβ and φC31 recombinase, respectively. After 72 hours of induction, we performed flow cytometry to quantify the distribution of cells states. Cell states were assigned by fitting a Gaussian Mixture Model to the fluorescent distributions (Figure 2-3) and considering cells with posterior probabilities for EBFP+, $EYFP_{High}$ or $mKate_{High}$ ≥0.5 to have remained in the EBFP state or committed to the EYFP or mKate state, respectively.

Figure 2-4 shows the proportion of cells that have committed to the different cell states after 72 hours. The results show that the *XFP* circuit can be used to create distinct and mutually exclusive cell states at ratios that can be controlled by the concentration of Dox and ABA. At lower concentrations of Dox and ABA, we observe a large proportion of cells classified as being in the EBFP+ state. We hypothesize committing to a cell state is a probabilistic event that depends on the concentration of recombinase, the kinetics of the recombination event for the given recombinase, and the distance between the recombinase sites. At low concentrations of both inducers, transcription of both recombinases is weak and the corresponding protein concentrations are low resulting in recombination being a low-probability event, and the majority of cells remain in the EBFP+ state. As the concentration of either inducer increases, the probability of changing to any cell state increases, and a larger fraction of cells leave the ground state.

At low and intermediate inducer concentrations, we observe a gradual commitment to a cell state (Figure 2-4, Figure S2-1, and Figure S2-2) as evidenced by the population of cells in the transition between an EBFP+ state towards the $EYFP_{High}$/EBFP- or $mKate_{High}$/EBFP- state.

While some cells are classified as $EYFP_{Low}$ or $mKate_{Low}$, closer inspection of Figure 2-4 reveals this is due to spectral bleed-through between EYFP and mKate.

Since cell state changes are permanent and inherited from mother to daughter cells, the recombinases contained a C-terminal PEST-tag which targets the protein for rapid degradation via the Ub-proteasome pathway or the UB-independent pathway[85] in an attempt to decrease background recombinase expression. Nevertheless, we still observe 14.2% of cells committing to the $mKate_{High}$ state and 5.9% of cells committing to the $EYFP_{High}$ state in the absence of inducer. Considering the recombinases are integrated via PiggyBac, it is likely that multiple copies of the circuit was integrated. Combined with potential leakiness from the promoters driving the recombinases, this can explain the observed background pre-commitment to a cell state.

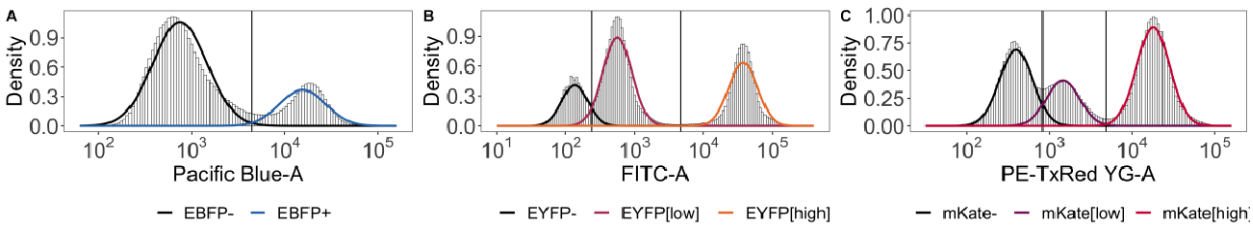


Figure 2-3: GMM to identify cell populations. Distribution of fluorescence as measured by **(A)** EBFP (Pacific Blue-A), **(B)** EYFP (FITC-A) and **(C)** mKate (PE-TxRed YG-A). A Gaussian Mixture Model was used to cluster the cells by fluorescence (colored lines), and cells were considered to belong to a given cluster if the prior for that cluster was ≥ 0.5. Vertical lines indicate the boundaries of that cluster and were calculated as the average between the highest value of the lower-intensity cluster and the smallest value of the higher-intensity cluster.

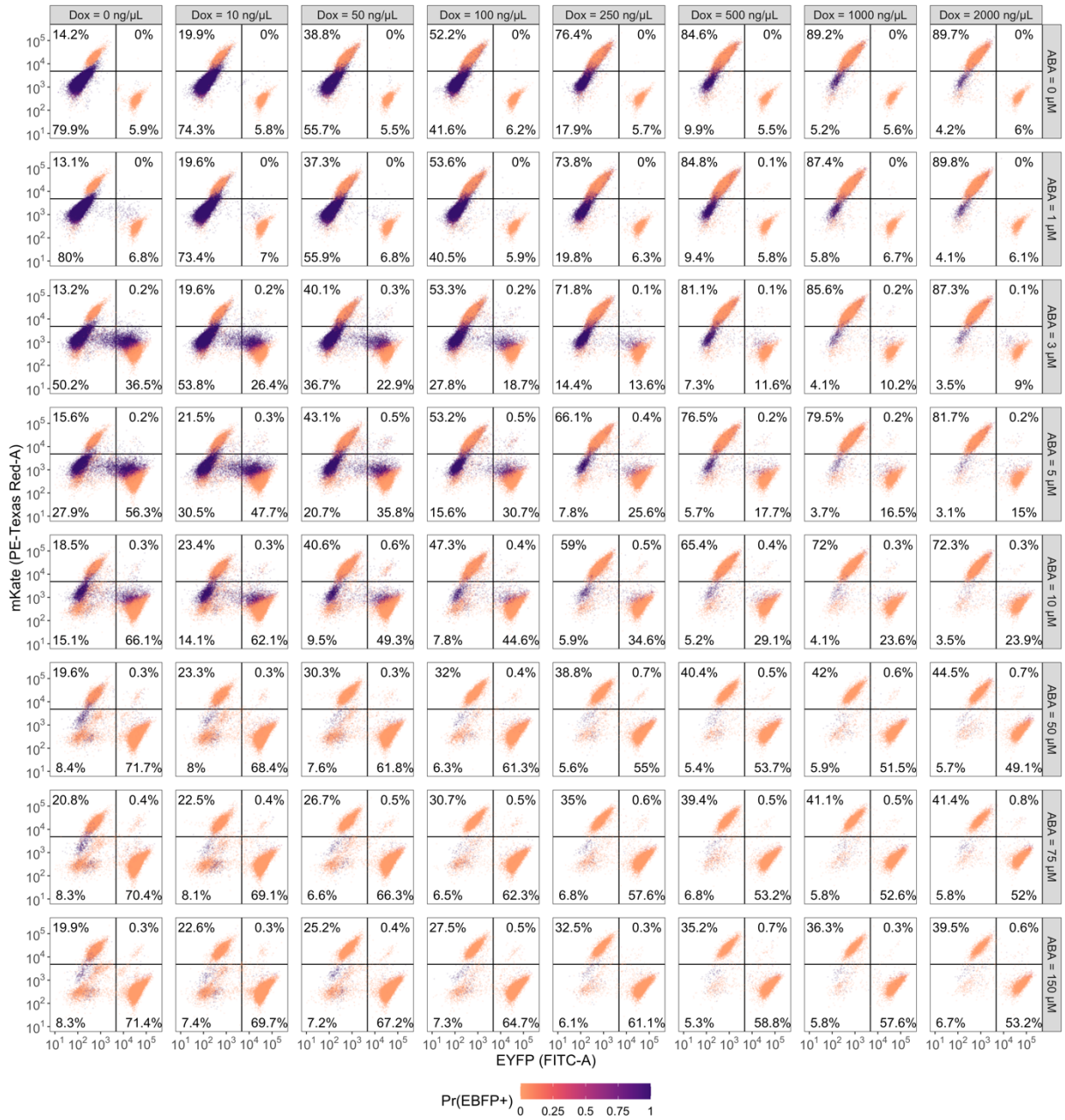


Figure 2-4: Distribution of cell states for the *XFP* circuit. Each quadrant in the subplots represent a unique cell state: EYFP-/mKate- (lower left), $EYFP_{High}$/mKate- (lower right), $EYFP_{High}$/$mKate_{High}$ (upper right), and EYFP-/$mKate_{High}$ (upper left). Quadrants are defined by the lower bound for the $EYFP_{High}$ (vertical) and the lower bound for the $mKate_{High}$ (horizontal) populations (refer Figure 2-3). The color shows the probability cells were EBFP+ as estimated by the GMM. Each subplot indicates a specific combination of ABA and Dox. The plot shows that the recombinase circuit can be used to control the distribution of cell states in the population. The high fraction of EBFP+ cells at low inducer concentrations indicates a change in cell state remained a low probability event. As a consequence, cells continue to change state throughout the experiment. The small fraction of cells that are negative for all three fluorophores indicates a low probability the circuit is either silenced or deleted. Induction of the φC31 recombinase (by ABA), might unintentionally lead to the mKate state, possibly by being able to recognize the Wβ recombinase sites. Note, the *low*

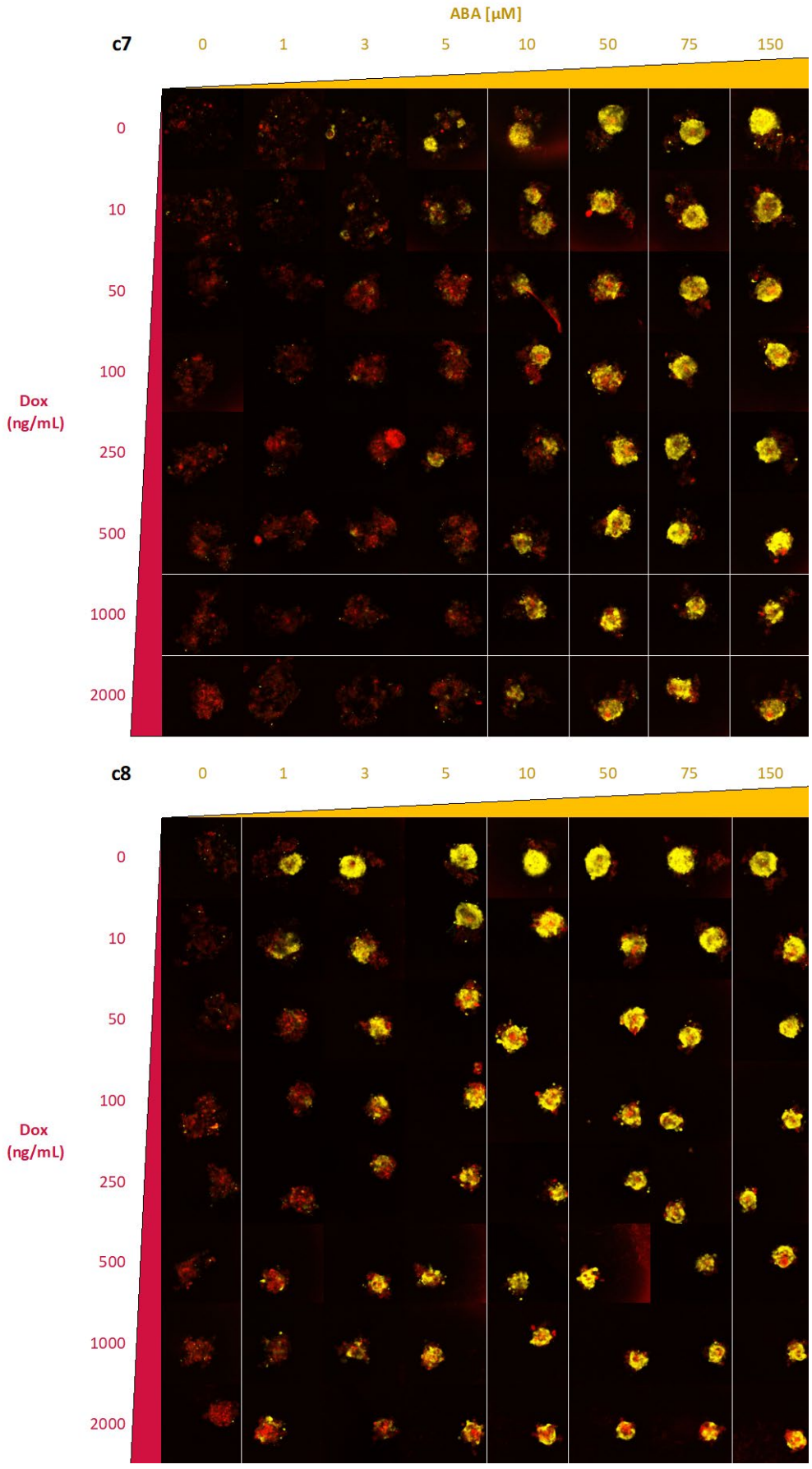
populations are not used to call cell state as it represents spectral bleed through as can be seen from the shift in the $mKate_{High}$ and $EYFP_{High}$ populations relative to the EBFP+ population in the lower left quadrant.

2.7 Cell state bifurcations generates different morphological structures

Differential cadherin expression was previously used to generate a diverse set of morphological shapes.[33,34] Rather than using the previous approaches that relies on a mix of populations, we sought to test if this was possible with a monoclonal and homogenous cell population which enables control of cell state at any point in time independent of the initial conditions for the experiment. As an additional layer of complexity, the time it takes for a cell to change state and achieve high cadherin expression might similarly play a role in cell sorting.[33,34]

Using the polyclonal cell line with the *TF* and *Recombinase* circuits integrated, we integrated the *Cadherin* circuit in an identical way to the *XFP* circuit, and created multiple monoclonal cell lines. Figure 2-5 shows the result from three independent cell lines after 72 hours of induction. We first observe that the inducer concentration required to initiate a change in cell state differs between cell lines, indicating the location and copy number of the *TF* and *Recombinase* circuit integration plays a role in the dynamics of the *Cadherin* circuit.

Secondly, we notice two different types of sorting: E-cadherin$^+$/EYFP$^+$ cells pack very tightly in contrast to K-cadherin$^+$/mKate$^+$ cells which only observe loose packing that is still tighter than for uninduced cells (Figure S2-3). Expression of K-cadherin with a P2A-mKate transcriptional fusion from the Rosa26 locus has previously been shown to facilitate tight packing of CHO cells.[34] We therefore hypothesize that Wβ-mediated recombination changes the DNA sequence such that either transcription or translation is reduced. As the E-cadherin$^+$/EYFP$^+$ cells appears to have stronger cadherin expression and are expected to be more adhesive cells, we observe them form a tight sphere with K-cadherin$^+$/mKate$^+$ cells on the outside, as we expect for a mixed population of strongly and weakly adherent cells.[62]


ABA [µM]
c7
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000
c8
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000

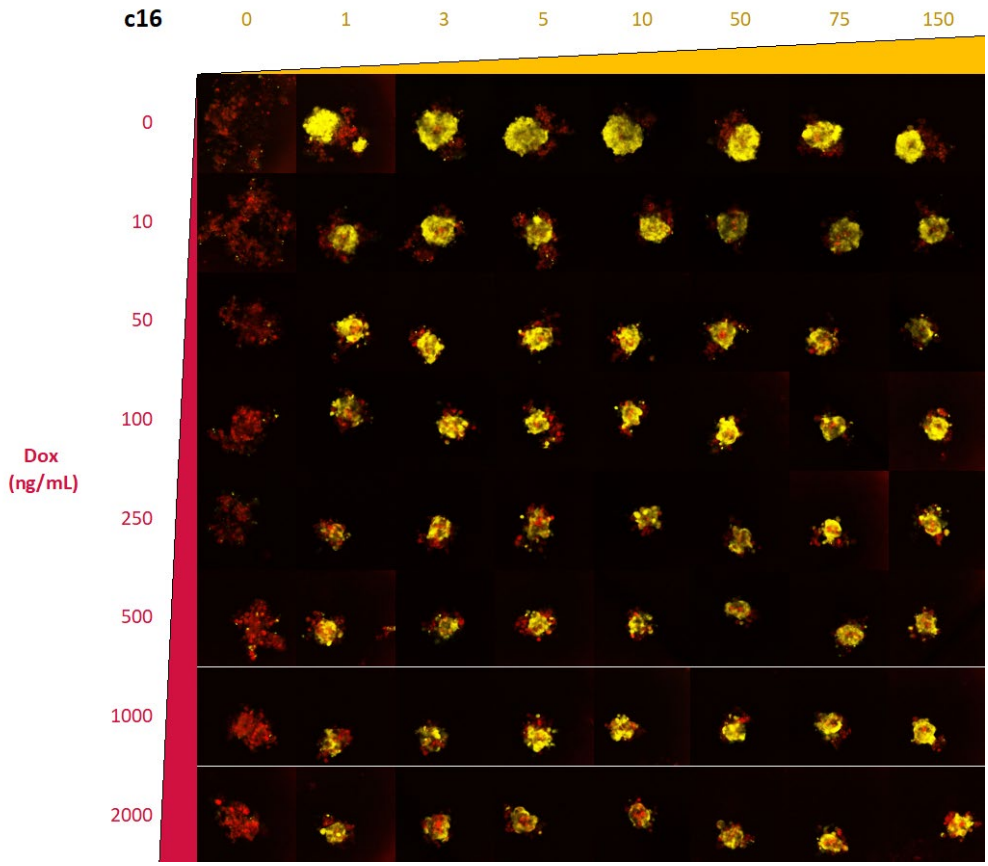
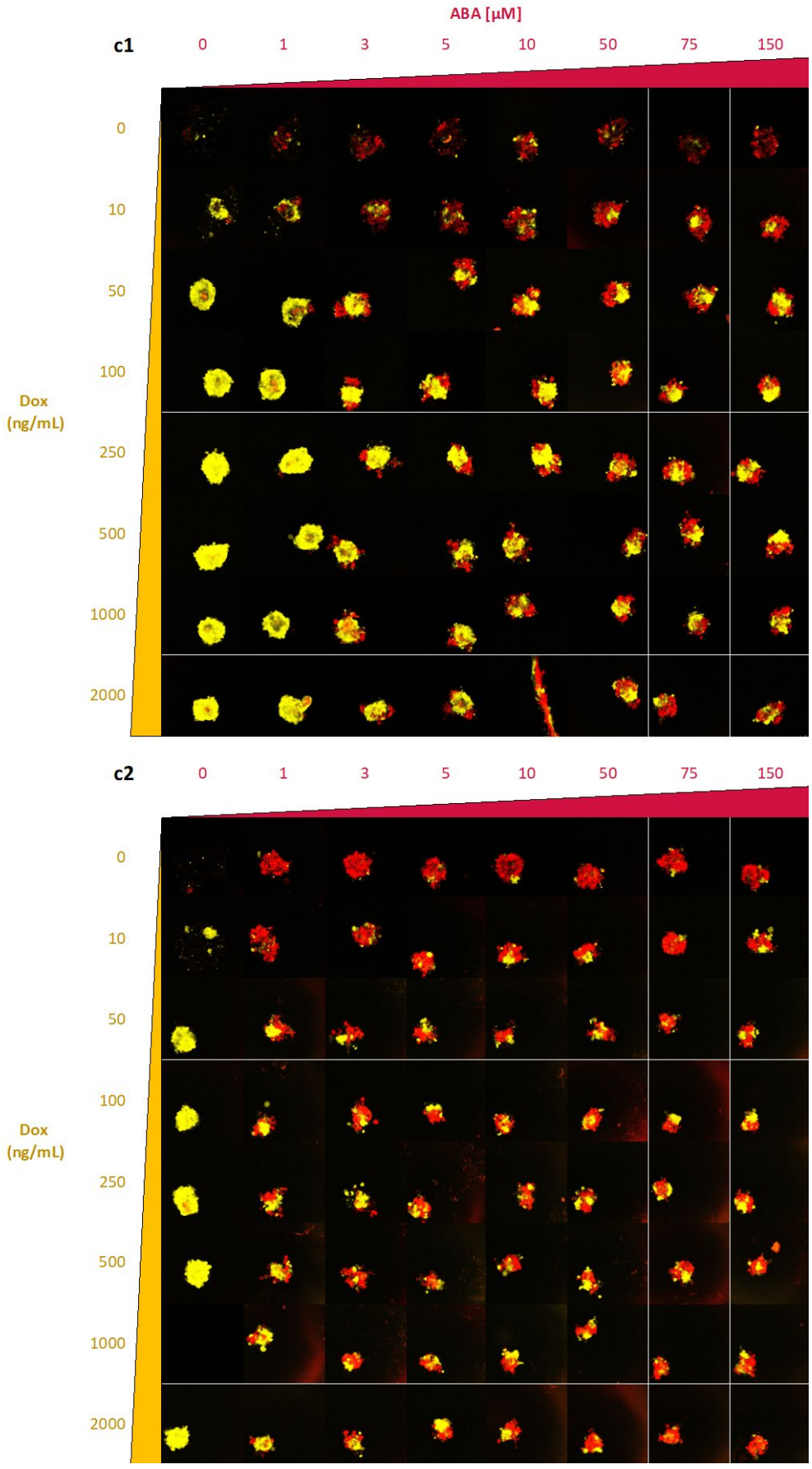


Figure 2-5: Shapes formed by induction of the *Cadherin* circuit. E-cadherin (encoded by Cdh1, yellow) and K-cadherin (encoded by Cdh6, red) is activated by ABA and Dox, respectively. Inducer determines the fraction of the population that switches to a given state. Cells sort based on the expression of the cadherin they express, and a number of different shapes can be generated. Cells expressing E-cadherin (yellow) pack tightly, while cells expressing K-cadherin pack loosely indicating poor expression. See Figure S2-3 for brightfield images. c7, c8, and c16 refer to individual clones.

Cell cycle regulation and cell death play an important role in development and the cell sorting observed for different organs.[86–88] We previously showed that $p27^{Kip1}$ could be used to inhibit the cell cycle in CHO cells.[34] To test the combined effect of cadherin-based cell sorting and cell cycle inhibition for one of the states, we co-express $p27^{Kip1}$ with E-cadherin by integrating the *Cadherin-$p27^{Kip1}$* circuit into the *TF* and *Recombinase* cell line as described for the other constructs tested. Figure 2-6 shows the CHO aggregates after 72 hours of induction for three different cell lines. As opposed to the *Cadherin* circuit, we observe tighter packing of the K-cadherin$^+$/mKate$^+$ cells, while the E-cadherin$^+$/EYFP$^+$ cells continue to pack tightly. This enables sorting as would be expected for two different, strongly adherent populations expressing mutually exclusive homotypic cadherins.[34,79] Many of these shapes are reproducible across cell lines, albeit at different levels of inducer combinations likely due to the different location and copy number of the *TF* and *Recombinase* circuits. Thus, the ratio between the different cell populations might remain an important factor in cell sorting. For instance, when a majority of cells have committed to an E-cadherin$^+$/EYFP$^+$ cell state, we frequently observe smaller K-cadherin$^+$/mKate$^+$ protrusions. At larger ratios of K-cadherin$^+$/mKate$^+$ cells, we observe engulfment of the E-cadherin$^+$/EYFP$^+$ cells, indicating that these might still be the most strongly adhering cells.

Surprisingly, the *Cadherin-$p27^{Kip1}$* resulted in two populations of strongly adherent cells indicating potential differences in adhesive properties of E- and K-cadherin. While we initially showed $p27^{Kip1}$ to inhibit the cell cycle,[34] this was not observed for the *Cadherin-$p27^{Kip1}$* cell line (data not shown), indicating that weak transcription or translation might still take place from the cell state controlled by Wβ recombination despite that the E-cadherin$^+$/EYFP$^+$ cells continued to pack tightly.


ABA [µM]
c1
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000
c2
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000

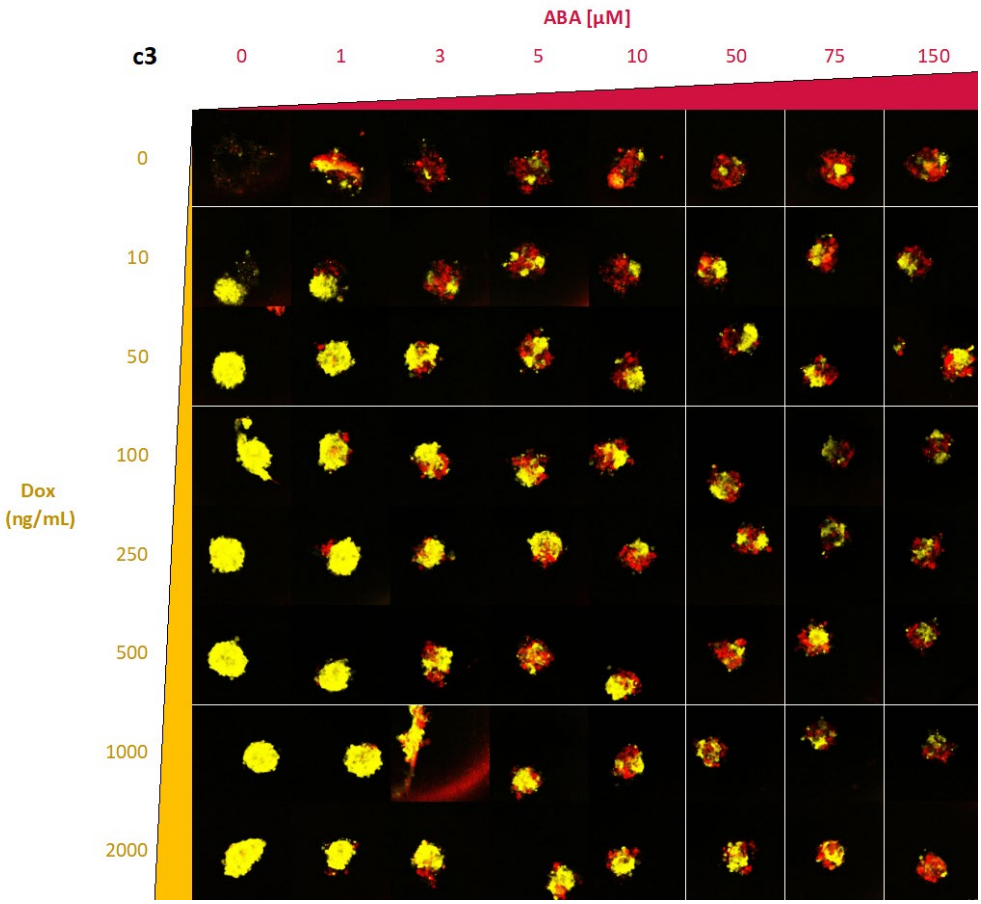


Figure 2-6: Shapes formed by induction of the *Cadherin-p27^{Kip1}* circuit. E-cadherin-T2A-EYFP-P2A-p27^{Kip1} (yellow) and K-cadherin-P2A-mKate (red) is activated by Dox and ABA, respectively. Inducer determines the fraction of the population that switches to a given state. Cells sort based on the expression of the cadherin they express, and a number of different shapes can be generated. Cadherin expression leads to tightly packed cells that form a variety of shapes depending on the ratio between the different cell states. See Figure S2-4 for brightfield images. c1, c2, and c3 refer to individual clones.

2.8 Discussion

Here, we use recombinases to induce cells to commit to one of two mutually exclusive cell states. These cell states are defined by either E-cadherin or K-cadherin expression which enables us to control cell sorting in a multicellular population as a function of cell state and their ratios. This has potential implications for organoids, self-organized multicellular systems grown from stem cells. These multicellular systems are valuable tools to understand and treat human disease. By mimicking organs, from their development to their function, organoids provide insight into disease mechanisms and can serve as drug screening platforms. The structure of organoids rely on cell-autonomous self-organization, and while this allows for complex tissue structures, it is extremely sensitive to external conditions and the state of the cell.[21,89,90] By creating synthetic cell states that controls cell sorting, our approach offers a method to control cell sorting within multicellular structures. Importantly, by coupling recombinase expression to the cell state, our approach is compatible with cell state-specific cell sorting at any point in time. While advances in bioprinting and hydrogels have enabled a significant amount of control over cell sorting within organoids *in vitro*,[26,60] these techniques are generally limited to setting up the initial conditions for patterning with few options to control cell sorting throughout organoid development.

While cell-cell adhesion plays an important role in morphogenesis, other factors can be used to control cell sorting. Differential growth has been shown to drive morphogenesis in development, including intestinal vili development[91] and gut tube formation.[92] We previously showed that the cell-cycle inhibitor $p27^{Kip1}$ could be used to inhibit CHO cell growth.[34] While the weak expression of $p27^{Kip1}$ was insufficient to significantly arrest the cell cycle, optimization of the recombination sites, mRNA stability, and translation might address this issue and allow us to further control morphogenesis by limiting cell proliferation. Taking cell cycle inhibition one step further, controlled neuronal cell death plays an important role in brain maturation,[93] and our recombinase switch could be modified to enable this as an additional input to cell sorting. A third driver of cell sorting is cell motility which has previously been engineered through inducible expression of the

cytoskeleton adaptor protein CRK-II15, and with constitutively active mutants of the GTPases RhoA and Rac1.[94] Cell motility might be particularly important when induced in larger cell aggregates if incomplete cell sorting is a concern.[62]

In future work, additional modulators and combinations thereof will be explored to expand the tools available to engineer morphogenesis and create tightly controlled, and highly reproducible multicellular structures with the potential to address current limitations in cell sorting within organoids.

2.9 Methods

2.9.1 Plasmid construction

Plasmids were constructed using a modified version of the hierarchical MoClo system.[95] Mouse Cdh1 and Cdh6 coding sequences were ordered from IDT without Type IIS restriction sites and inserted into L0 destination vectors. L1 expression vectors were assembled from L0s containing the different components for each transcriptional unit. L2 vectors were assembled from L1 vectors by Golden Gate cloning using SapI, into either a modified PiggyBac backbone compatible with SapI-based Golden Gate, or the SmallBOB backbone similarly modified to work with SapI-based Golden Gate.

2.9.2 Cell culture and transfections

CHO-K1 cells with an integrated LP in the Rosa26 locus[83] were maintained in Dulbecco's modified Eagle's medium F10 (DMEM/F-10) supplemented with 10% FBS. Cells with an integrated circuit were maintained with selection added to the media.

Cells were transfected with ViaFect (Promega) according to manufacturer's instructions. The *TF* and *Recombinase* circuits were integrated using 150,000 cells in a reverse transfection in a 12-well plate with 1 µg plasmid expressing the PiggyBac integrase, and 0.5 µg of each of the plasmids to be integrated. Selection (zeocin and blasticidin) started after 48 hours and continued for 7 days. The cell-state expressing plasmids were integrated into the *TF*- and *Recombinase*-expressing cell line in a similar manner but using a 24-well plate using 30,000 cells, and 250 ng each of the SmallBOB expression vector and circuit to be integrated. Selection (puromycin in addition to the other selection markers) was started after 24 hours. After 7 days, $EBFP^{+}/EYFP^{-}/$ $mKate^{-}$ cells were sorted to single cells using a BD FACS ARIA.

2.9.3 Pattern formation assay

To make 3D aggregates, a monoclonal cell line expressing the circuit to be tested was trypsinized. Cells were counted and diluted, and 100 cells were seeded in each well of a 96-well ultra-low attachment round bottom plate. Media containing inducer and selection (to maintain the circuit) was added to each well and the cells were centrifuged at 300 x g for 5 min. to bring the cells to the bottom of the well.

2.9.4 Microscopy and image analysis

Aggregates were imaged in a Leica TCS SP5 II Confocal Laser Scanning Microscope in an incubation chamber at 37°C and 5% CO_2. Each image represents a single Z slice through the aggregate at approximately the center, as estimated by where the edges of the aggregate were most sharply in focus. Image processing was done with FIJI as previously described.[34] Briefly, the image analysis is built around the Particle Analyzer plugin[96] for FIJI which is used to connect components on a binary image and group

adjacent pixels into clusters. These clusters are filtered by size, excluding anything smaller than 350 μm^2, as this threshold is much smaller than the area of any single cell and helps exclude small-particle noise. The binary mask used for the Particle Analyzer plugin is created by thresholding the fluorescent images for each fluorescent channel. This process was automated using Jython.

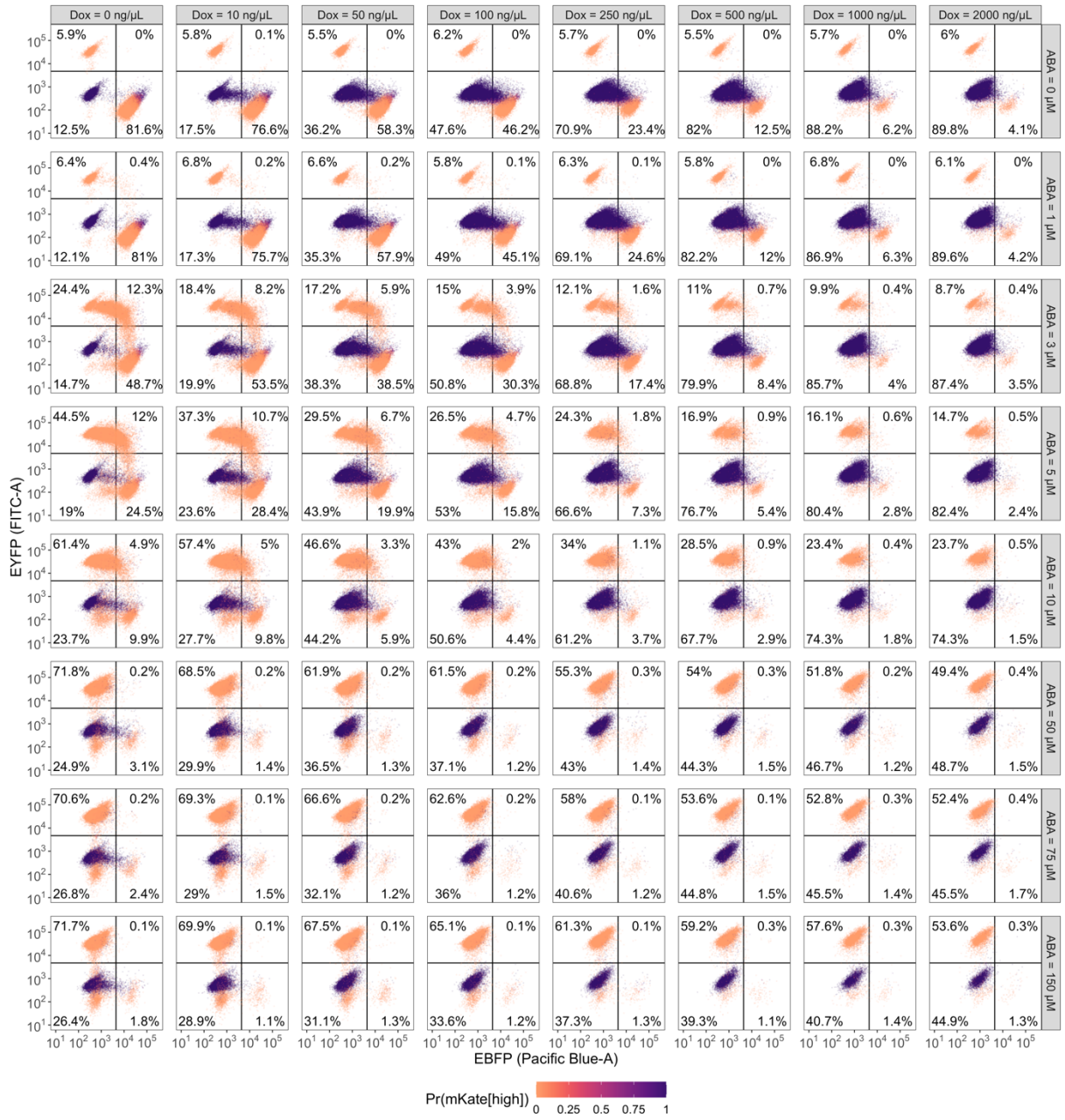


Figure S2-1: Distribution of EBFP and EYFP cell states for the *XFP* circuit. Each quadrant in the subplots represent a unique cell state: EBFP-/EYFP- (lower left), EBFP+/EYFP- (lower right), EBFP+/EYFP$_{High}$ (upper right), and EBFP-/EYFP$_{High}$ (upper left). Quadrants are defined by the lower bound for the EBFP+ (vertical) and the lower bound for the EYFP$_{High}$ (horizontal) populations (refer Figure 2-3). The color shows the probability cells were mKate$_{High}$ as estimated by the GMM. Each subplot indicates a specific combination of ABA and Dox. The number of cells in a transition from EBFP+ to EYFP$_{High}$ at intermediate concentrations of inducer indicates the switch in cell state continuously happens throughout the experiment, and a switch in cell state is more likely to happen at high concentration of inducer.

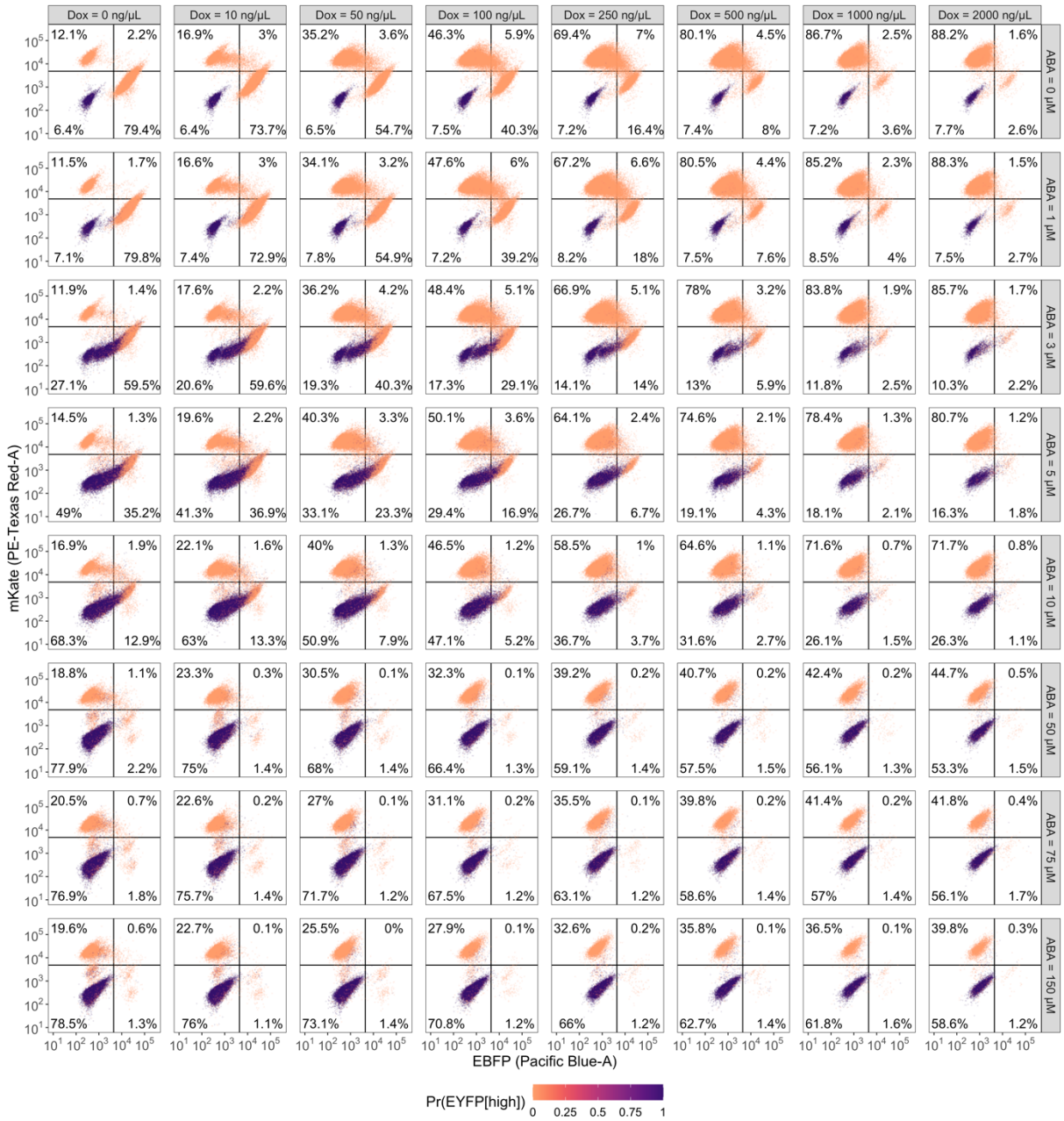
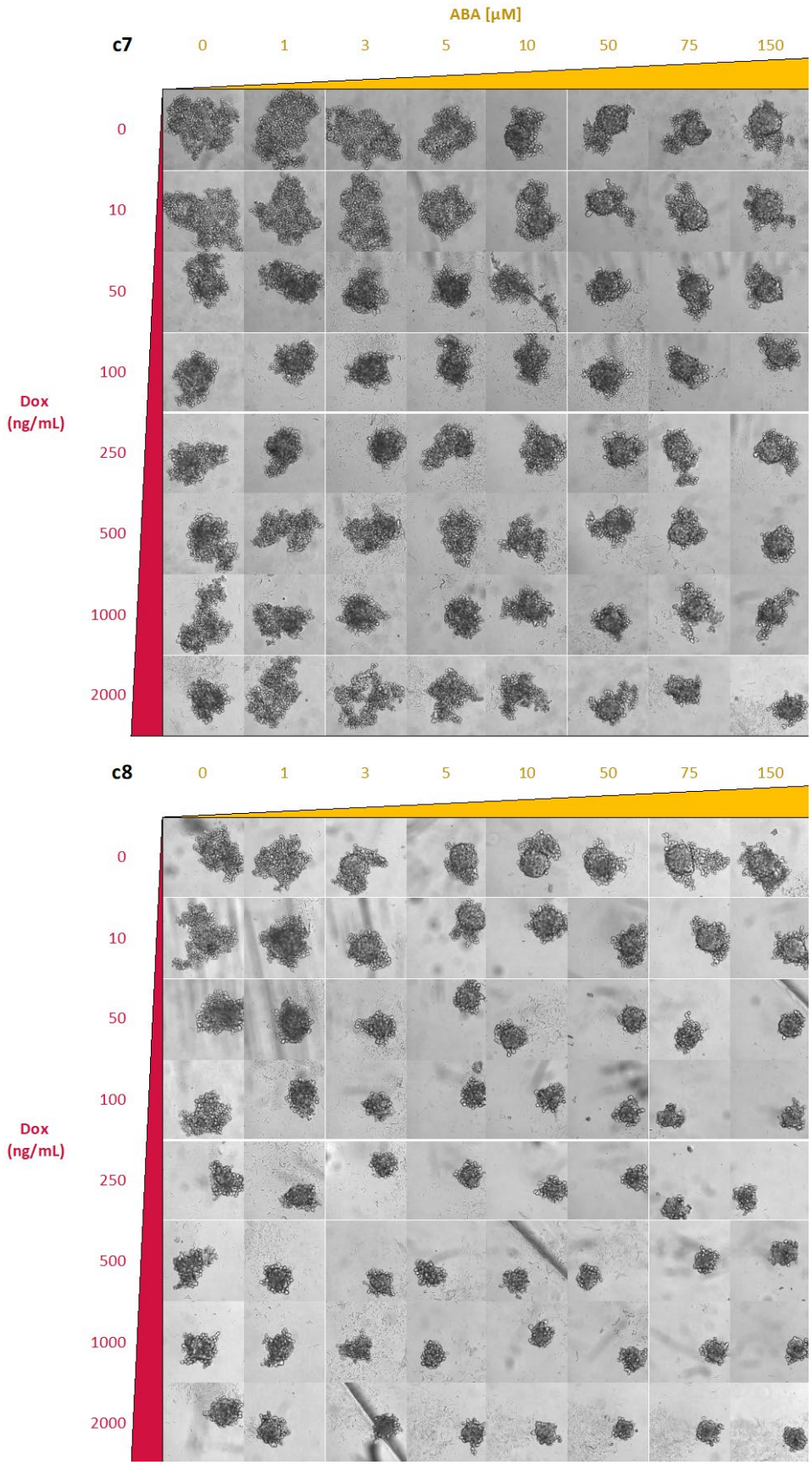


Figure S2-2: Distribution of EBFP and mKate cell states for the *XFP* circuit. Each quadrant in the subplots represent a unique cell state: Each quadrant in the subplots represent a unique cell state: EBFP-/mKate- (lower left), EBFP+/mKate- (lower right), EBFP+/mKate$_{High}$ (upper right), and EBFP-/mKate$_{High}$ (upper left). The quadrants are defined by the lower bound for the EBFP+ (vertical) and the lower bound for the mKate$_{High}$ (horizontal) populations (refer Figure 2-3). The color shows the probability cells were EYFP$_{High}$ as estimated by the GMM. Each subplot indicates a specific combination of ABA and Dox. The number of cells in a transition from EBFP+ to mKate$_{High}$ at intermediate concentrations of inducer indicates the switch in cell state continuously happens throughout the experiment, and a switch in cell state is more likely to happen at high concentration of inducer.


ABA [µM]
c7
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000
c8
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000

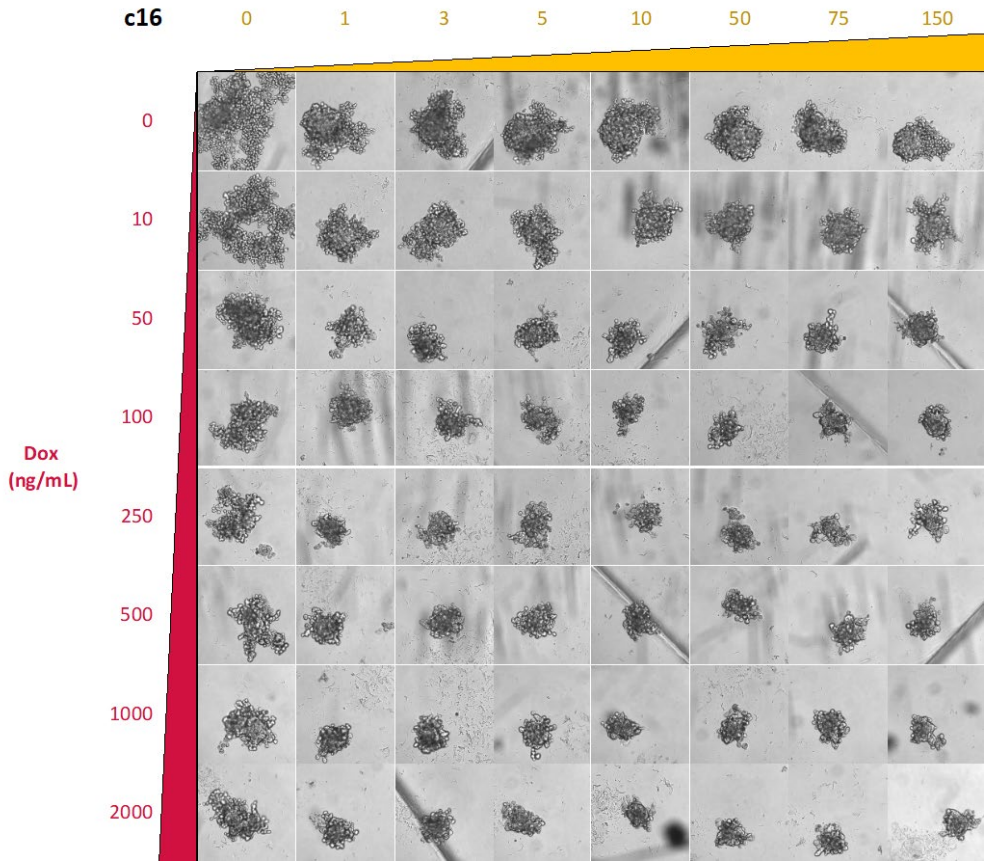
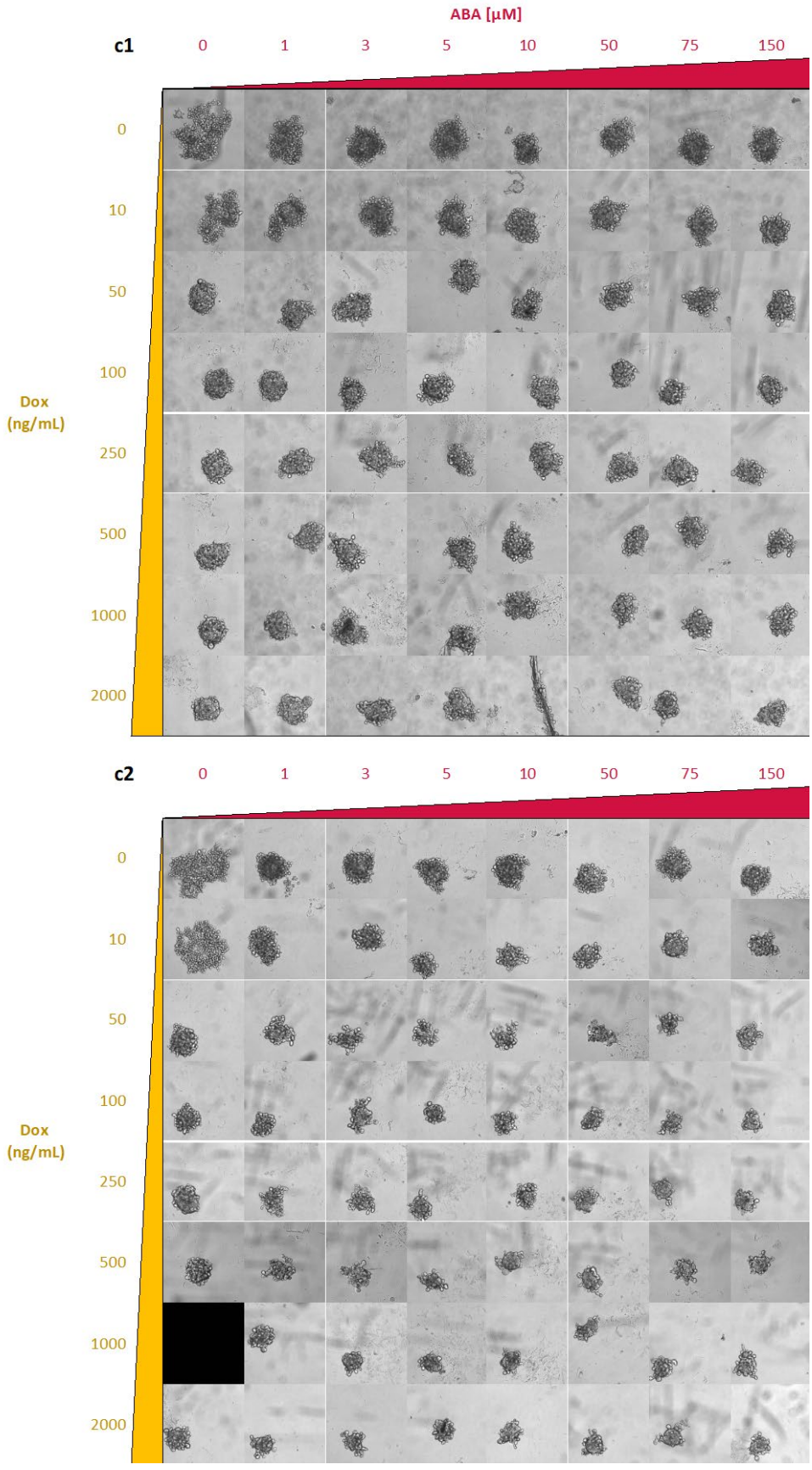


Figure S2-3: Brightfield images for the shapes formed by induction of the *Cadherin* circuit. E-cadherin (encoded by Cdh1) and K-cadherin (encoded by Cdh6) is activated by ABA and Dox, respectively. Inducer determines the fraction of the population that switches to a given state. Cells expressing E-cadherin form tightly packed cells, while cells expressing K-cadherin form loosely packed cells indicating poor expression. c7, c8, and c16 refer to individual clones.


ABA [µM]
c1
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000
c2
0
1
3
5
10
50
75
150
Dox (ng/mL)
0
10
50
100
250
500
1000
2000

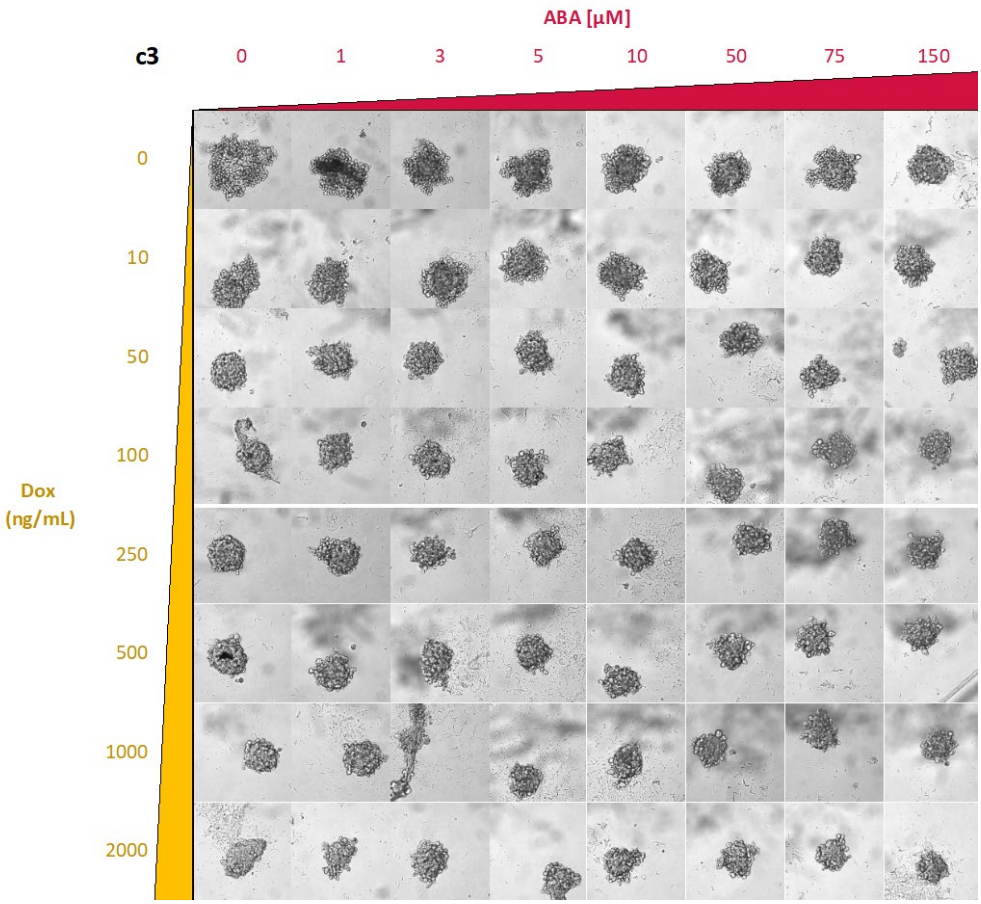


Figure S2-4: Brightfield images for the shapes formed by induction of the *Cadherin-p27^{Kip1}* circuit. E-cadherin-p27^{Kip1} and K-cadherin is activated by Dox and ABA, respectively. Inducer determines the fraction of the population that switches to a given state. Cadherin expression leads to tightly packed cells. c1, c2, and c3 refer to individual clones.

Chapter 3. Synthetic promoters for cell-type specific transcriptional regulation

This chapter is based off published work authored by me, Ming-Ru Wu, Lior Nissim, Doron Stupp, Erez Pery, Adina Binder-Nissim, Karen Weisinger, Sebastian R. Palacios, Melissa Humphrey, Zhizhuo Zhang, Eva Maria Novoa, Manolis Kellis, Ron Weiss, Samuel D. Rabkin, Yuval Tabach, and Timothy K. Lu.[97] A detailed list of author contributions is provided in the paper. Briefly, in collaboration with S.R.P, I designed and conducted the experiment testing SPECS in the liver bud-like organoid, and I analyzed the data for that section of the paper.

3.1 Summary of chapter 3

Cell state-specific promoters constitute essential tools for basic research and biotechnology because they activate gene expression only under certain biological conditions. Synthetic Promoters with Enhanced Cell-State Specificity (SPECS) can be superior to native ones, but the design of such promoters is challenging and frequently requires gene regulation or transcriptome knowledge that is not readily available. Here, to overcome this challenge, we use a next-generation sequencing approach combined with machine learning to screen a synthetic promoter library with 6107 designs for high-performance SPECS for potentially any cell state. We demonstrate the identification of multiple SPECS that exhibit distinct spatiotemporal activity during the programmed differentiation of induced pluripotent stem cells (iPSCs), as well as SPECS for breast cancer and glioblastoma stem-like cells. We anticipate that this approach could be used to create SPECS for gene therapies that are activated in specific cell states, as well as to study natural transcriptional regulatory networks.

3.2 Introduction and aim

Promoters are key regulatory DNA elements located upstream of a gene coding region. In combination with other regulatory DNA elements, such as enhancers and silencers, and epigenetic modifications, promoters regulate the timing and levels of gene expression.[98] In eukaryotes, promoter activity is trans-regulated by transcription factors (TFs). TFs recognize specific DNA sequences, bind them, and recruit general components of the transcriptional machinery necessary for transcription initiation. Therefore, promoter activity is regulated by the composition and activity of TFs in the cell. This regulation plays vital roles in many biological processes, whether in health or disease, such as cellular differentiation, organ development, and malignancy.[99] Many promoters are selectively active in specific cell states, such as a particular phase of the cell cycle, certain tissues, or abnormal states such as cancer.[100–102] These promoters can be utilized as simple and autonomous sensors to trigger the transcription of an output gene only under predetermined conditions. Such outputs include reporter genes for cell state diagnosis and effector genes that enable programmed cellular behavior, decision-making, and actuation. For example, cell state-specific promoters have been used to selectively express transgenes in muscle cells, to specifically target cancer cells, and to visualize and isolate antigen-stimulated primary human T cells.[103–106] Additionally, synthetic gene circuits have been designed to integrate the activity of multiple cell state-specific promoters to precisely diagnose and treat disease such as cancer,[107,108] diabetes,[109] and psoriasis.[110] Thus, cell state-specific promoters constitute an essential building block for genetic engineering and enable a wide range of applications in basic biological research, biomedicine, synthetic biology, and biotechnology.[111,112] Ideal cell state-specific promoters should exhibit high activation exclusively in the cellular condition of interest. Here we define the cell state specificity of a promoter as the ratio of its activity in the cell state of interest to its activity in the control cell state. For instance, we might tumorigenic cells to have one cell state versus the cell state of non-tumorigenic cells of the same lineage. Native promoters often exhibit modest cell state specificity. For

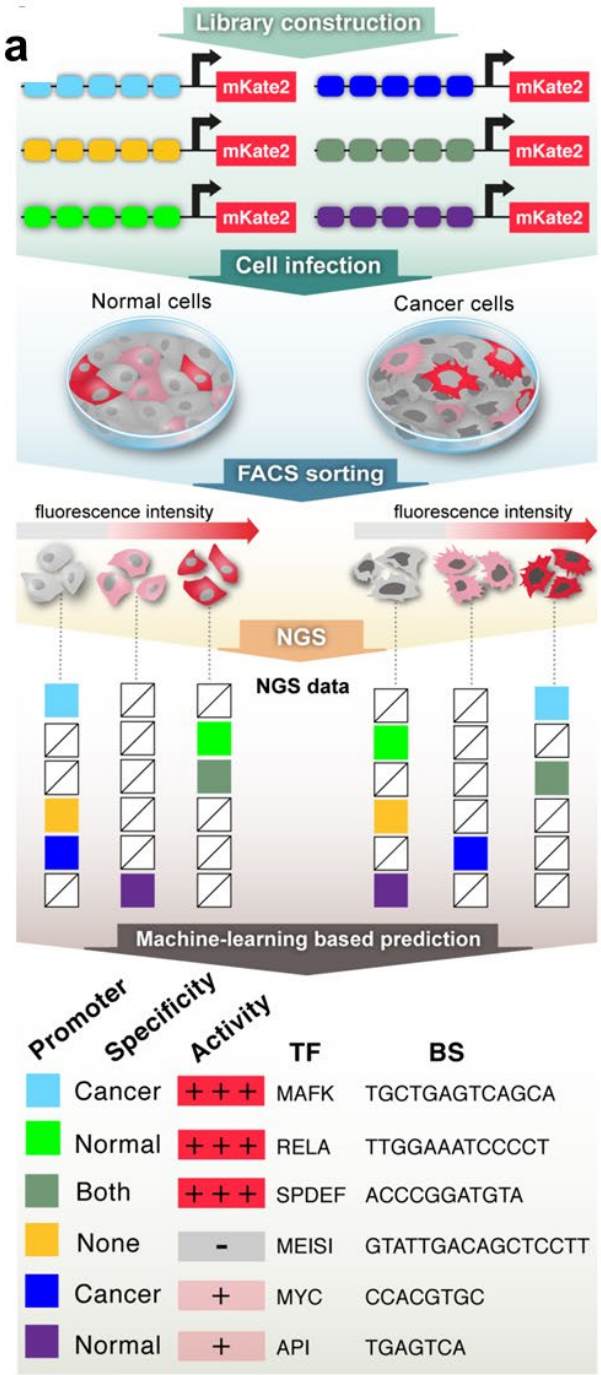
example, many native cancer-specific promoters also show considerable activity levels in normal cells.[113,114] This is likely due to native promoters typically containing a wide range of TF-binding sites (TF-BSs) that can be potentially bound and activated by numerous TFs belonging to multiple TF families.[108] Because it is very unlikely that a wide range of TFs will be active only in a particular cell state, native promoters generally exhibit considerable basal activity in multiple cell states and therefore have lower cell state specificity. Synthetic promoters with enhanced cell-state specificity (SPECS) were previously developed as alternatives to native ones. A typical design consists of tandem repeats of TF-BSs for one or a few TFs that are active only in the cell state of interest, encoded upstream of a minimal promoter that contains essential transcription initiation elements.[108,115–118] However, for these previous approaches, the promoters were generally built one by one by molecular cloning based on prior knowledge of gene regulation or the transcriptome of the cell state of interest, which is not always readily available. Additionally, even with suitable data at hand, this process often requires multiple design-build-test cycles to build adequate promoters.[108,115] Synthetic promoter library screens have also been developed to identify strong promoters or to study transcriptional regulation,[119–121] but these approaches were not specifically designed to identify SPECS. For example, most of these approaches utilized a library of random K-mers as TF-BSs.[119] However, most of these random K-mers are not functional TF-BSs and therefore library screening is more challenging, as it requires large-scale experiments to achieve sufficient coverage. Alternatively, in other studies, long 68bp K-mers, which are significantly larger than the average length of TF-BSs [≈10–13 bp[122,123]], were used. These long K-mers can be potentially bound by multiple different TFs,[120,121] which could confound efforts to make promoters that are responsive only to specific TFs.[120,121] Here we develop a high-throughput experimental and computational pipeline for efficient SPECS identification, which does not require any prior data of the cell state of interest. For this purpose, we design a library of synthetic promoters that corresponds to 6107 eukaryotic TF-BSs reported in two databases.[124,125] Each construct in the library comprises tandem repeats of a single TF-BS encoded upstream of an adenovirus minimal promoter to control the expression of mKate2 fluorescent protein. Our screening pipeline combines lentiviral library introduction, FACS cell sorting, next-generation sequencing,

and a machine-learning based computational analysis (Figure 3-1). We demonstrate the versatility of this approach by identifying a panel of SPECS in a variety of distinct biological settings, including: (i) SPECS that demonstrate spatial and temporal dynamics in an in vitro organoid differentiation model; (ii) SPECS that exhibit strong and specific activity in breast cancer cells vs. normal breast cells; and (iii) SPECS that distinguish differentiated bulk glioblastoma cells from glioblastoma stem-like cells derived from the same patient. The diversity of this library and the efficiency of our screening and computation pipeline enable efficient identification of SPECS for various biomedical applications.

3.3 SPECS show distinct activities in an organoid model

Organ differentiation requires tightly orchestrated spatiotemporal regulation of promoter activity.[126,127] In vitro organ differentiation models can be generated by programmed differentiation of induced pluripotent stem cells (iPSCs), which generates organoids comprising multiple cell types.[128,129] We therefore used one such model to examine whether screening our library of 6107 synthetic promoters (see Methods for details) could identify SPECS that distinguish between distinct normal cellular states.[27] For this purpose, we first infected the organoid with our SPECS library, followed by FACS sorting of mKate2 positive cells to enrich active promoters in the organoid culture, shotgun cloning of PCR-amplified promoter fragments, and a noise filtering process. As a result, we identified four promoters with distinct spatial and temporal behaviors in the organoid (see Methods for detailed screening process). To characterize the spatiotemporal activity of each identified promoter during the organoid differentiation process, we infected an entire iPSC population with a construct in which mKate2 expression is regulated by a single promoter. We then induced differentiation and measured mKate2 fluorescence levels using time-lapse confocal microscopy. Analysis of pixel intensities from microscope images showed that each identified promoter generated a distinct activity pattern during the organoid differentiation process (Figure 3-2). The promoter comprising RELA TF-BSs was strongly

and ubiquitously activated around day 11. The promoter comprising STAT disc5 TF-BSs was active only between days 3 and 7. The promoters comprising SPDEF and HIF1A TF-BSs were each active in only a small fraction of the organoid and demonstrated distinct timing and strength of expression. These results show that SPECS with diverse activity patterns can be identified in vitro in a complex 3D multicellular structure by our library. Thus, our library can be utilized to generate SPECS that distinguish among normal cell states.


a
Library construction
mKate2
Cell infection
Normal cells
Cancer cells
FACS sorting
fluorescence intensity
fluorescence intensity
NGS
NGS data
Machine-learning based prediction
Promoter
Specificity
Activity
TF
BS
Cancer
+++
MAFK
TGCTGAGTCAGCA
Normal
+++
RELA
TTGGAAATCCCCT
Both
+++
SPDEF
ACCCGGATGTA
None
-
MEISI
GTATTGACAGCTCCTT
Cancer
+
MYC
CCACGTGC
Normal
+
API
TGAGTCA

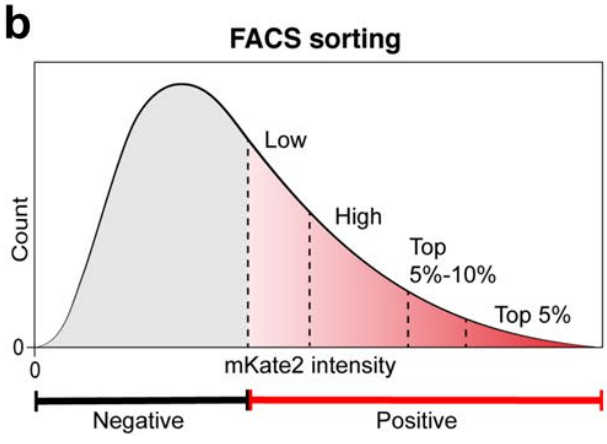

b
FACS sorting
Low
High
Top
5%-10%
Top 5%
Count
0
0
mKate2 intensity
Negative
Positive

Figure 3-1: The experimental and computational pipeline for identifying cell state-specific promoters. (a) The experimental pipeline consisted of infecting cells with synthetic promoter libraries encoded on lentiviruses, FACS sorting of cells into subpopulations according to fluorescence intensity, next-generation sequencing (NGS), and computational analysis to identify the promoters enriched in each subpopulation. From top to bottom, the promoters in the library contained tandem repeats of a single transcription factor (TF) binding site (BS) (colored boxes). Cells of different cell states (e.g., normal vs. cancer) were infected with the pooled library and then sorted by FACS into bins based on fluorescence intensity. For each bin, NGS was performed to determine the abundance of each promoter in each bin. Finally, a machine-learning based prediction was used to determine the activity of each promoter and its cell state specificity (e.g., light blue indicates that the promoter is specific to cancer cells whereas light green indicates that the promoter is specific to normal cells). **(b)** The cells infected with the promoter library were FACS sorted into five subpopulations according to fluorescence intensity (negative, low, high, top 5–10%, top 5%), followed by NGS and computational analysis to identify the promoters enriched in each subpopulation

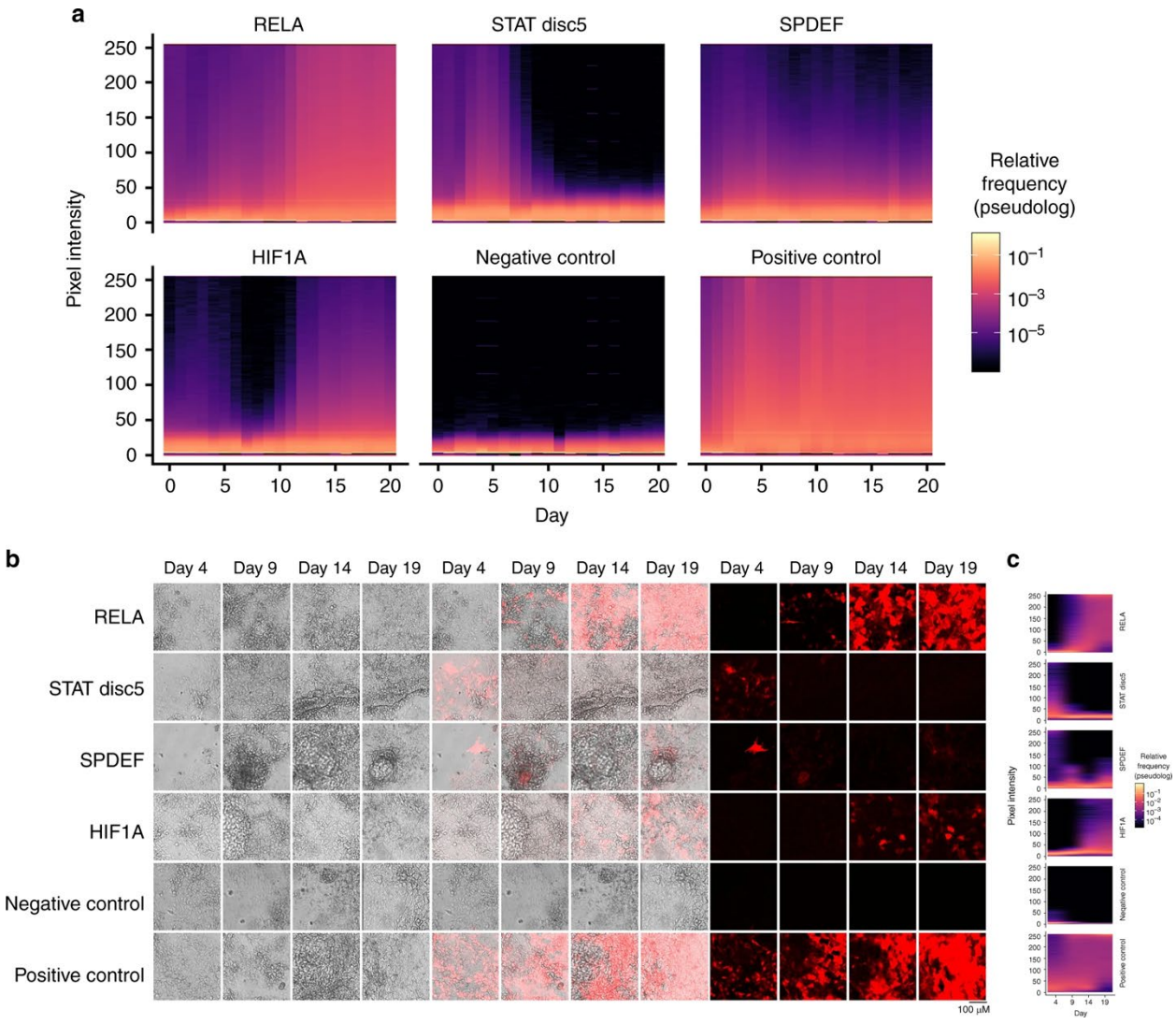


Figure 3-2: Synthetic promoters exhibit distinct temporal and spatial behavior in organoid cultures derived from iPSCs. (a) The heat maps show distinct temporal and spatial activities of four promoters across the time course of differentiation. The X-axis denotes the days post Dox-induced differentiation. The Y-axis denotes the fluorescence intensity as the pixel value of an 8-bit image (fluorescence intensity is equally divided into 256 bins, 0 being the lowest, and 255 being the highest). Heat map colors show the relative frequencies of pixel fluorescence intensity distribution in each bin with a log pseudocount to account for absent bins [(1 + number of pixels in each fluorescence intensity bin/number of total pixels)]. The distributions show the difference in the timing and strength of promoter activation, and the fraction of the image containing fluorescent cells. The negative control sample consisted of cells infected with a non-fluorescent protein; the positive control sample consisted of cells infected with a Ubiquitin C promoter expressing mKate2. **(b)** Representative fluorescence and bright field microscopy images show distinct temporal and spatial activities and differences in expression strength of the four promoters. The sub-regions exhibiting the strongest fluorescence signal for each promoter are shown. Left panel contains the bright field images (Days 4–19), middle panel contains the overlay images (Days 4–19), and right panel contains

the fluorescence images (Days 4–19). **(c)** The heat maps show the relative frequencies of pixel distribution in each fluorescence bin for the representative fluorescence microscopy images in **(b)**. N = 3 biological replicates

3.4 The combined pipeline identifies cancer-specific SPECS

Cancer-specific promoters constitute useful tools for basic biological research and biomedical applications.[102] However, most cancer-specific promoters reported in the literature generally exhibit only modest tumor specificity and are hard to find.[113,130] Therefore, we next examined whether we could identify SPECS with enhanced tumor specificity using our platform. As a proof- of-concept, we aimed to identify SPECS that distinguish the breast cancer cell line MDA-MB-453 (as a breast cancer model) from the non-tumorigenic breast cell line MCF-10A (as a model of normal breast cells).[131,132] To identify SPECS for MDA-MB-453, we infected the cells with our library, sorted the cells by FACS, and isolated the population consisting of the top 5% most fluorescent cells (Figure 3-1b**,** Top 5% population). We shotgun-cloned promoters extracted from DNA of the top 5% population and characterized their activity in both MDA-MB-453 and MCF-10A to identify SPECS that are exclusively active in MDA-MB-453. Of the 17 promoters that we isolated using this approach, 4 promoters had enhanced cancer specificity, showing 64-, 137-, 406-, and 499-fold activation in MDA-MB-453 compared to MCF-10A (Figure S3-1). All other promoters were either inactive in both MDA-MB-453 and MCF-10A cell lines or had substantial activity in both cell lines, constituting false positives from the pipeline under these experimental conditions. Although this Top 5% approach enables identification of SPECS, it is relatively low-throughput and may not be sufficient for finding SPECS in more challenging scenarios. Thus, we developed a comprehensive high-throughput SPECS screening pipeline to predict the activity of all the promoters in our library for each cell state. This pipeline was used to systematically and efficiently identify promoters with a range of absolute activity levels and activity patterns in these model cell lines (Figure 3-1 & Methods). In the first step, a library of synthetic promoters that regulate the fluorescent protein mKate2 was delivered into the cell lines of interest. Next, each cell line population was FACS sorted into five differential subpopulations according to promoter-activity levels, based on five distinct fluorescence intensity bins. Sorting the cells into multiple bins provided a more accurate description of promoter fluorescence distribution than just sorting into the fluorescence negative and positive bins (Figure 3-1b). We then calculated the counts of each promoter in each fluorescence bin by analyzing

data from next-generation sequencing (NGS). We then sought to compare the fluorescence measurements and counts for promoters identified in the Top 5% approach screening. We found that the promoter-count distribution across fluorescence bins approximated the actual promoter activity levels, measured by infecting an entire cell population with a single promoter regulating mKate2 (Figure S3-2). Therefore, we utilized these counts as inputs to machine learning regression models to achieve library-wide promoter activity predictions.

We collected data to train the models by measuring fluorescence for single promoters from the library. Promoters were chosen based on an approximate measure of activity resembling weighted averages (see Methods for more details). We chose 64 promoters predicted to have a range of activity in MDA-MB-453 and MCF-10A cells based on this heuristic, which together with the 17 promoters measured in the Top 5% random shotgun cloning approach, constituted a total of 81 promoters used to train the machine learning algorithms. Fluorescence levels and counts from the 81 promoters were fed as inputs (a 60–40% train-test split) to several machine learning regression algorithms (linear-regression based models, tree-based models, and support vector machines) with several feature engineering steps performed. Features, based on the relationships observed in comparing counts to fluorescence as described above, included counts, sum of counts, ratios between the bins, etc. (Supplementary Text 1**,** Figure S3-3). A generalized linear model (GLM) with elastic net regularization (GLMNET) was chosen based on performance[133] (Figure S3-5). This model was trained on the features as well as interaction terms to identify non-linear relationships (GLMNET-inter) (see Methods for more details). Based on this model, we picked additional 54 promoters with a wide dynamic range of predicted activity, including promoters with enhanced specificity to either cell state and promoters with various predicted fluorescence output levels as our validation set (Figure 3-3a). We then measured the fluorescence output levels generated by these promoters in both cell lines and found that the experimental data indeed validated the model. Of 12 predicted MDA-MB-453-specific promoters, 11 had over 10-fold greater activity in MDA-MB-453 compared to their activity in MCF-10A, and 6 of these 11 promoters exhibited more than 100-fold greater activity in MDA-MB-453 compared to

that in MCF-10A (Table S3-1). Overall, this model was highly predictive of promoter activity in both the held-out test set (R^2 = 0.81) and the separate 54-promoter validation set (R^2 = 0.77, Figure S3-4). A second model was trained using all 135 (81 + 54) promoters with similar performance on a held-out test set (R^2 = 0.77, Figure 3-3b). This second model was used to predict the promoter activities of the entire library. Overall, we found dozens of promoters with MCF-10A specificity and hundreds with MDA-MB-453 specificity (Figure 3-3c). Therefore, our experimentally validated promoters constitute only a small portion of the potential cell state-specific promoters in our library. Moreover, this approach enabled the identification of promoters with a wide dynamic range of activity (Figure 3-3a—promoters with light blue and orange color names). Moderately active promoters are essential for applications in which only temperate output levels are required, for example, to regulate an effector protein that is cytotoxic at high concentrations. These promoters can be chosen to be either cell state specific or not, based on the required experimental condition. Overall, while the Top 5% approach exhibited reasonable efficiency in this experimental setup, a combined library screen and machine-learning based computational approach provided efficient large-scale prediction of promoter activity in the cell lines of interest. We anticipate that this experimental-computational pipeline will be useful for finding cell state-specific promoters in more challenging experimental setups, for example, when numerous cell types or similar cell lines are involved.

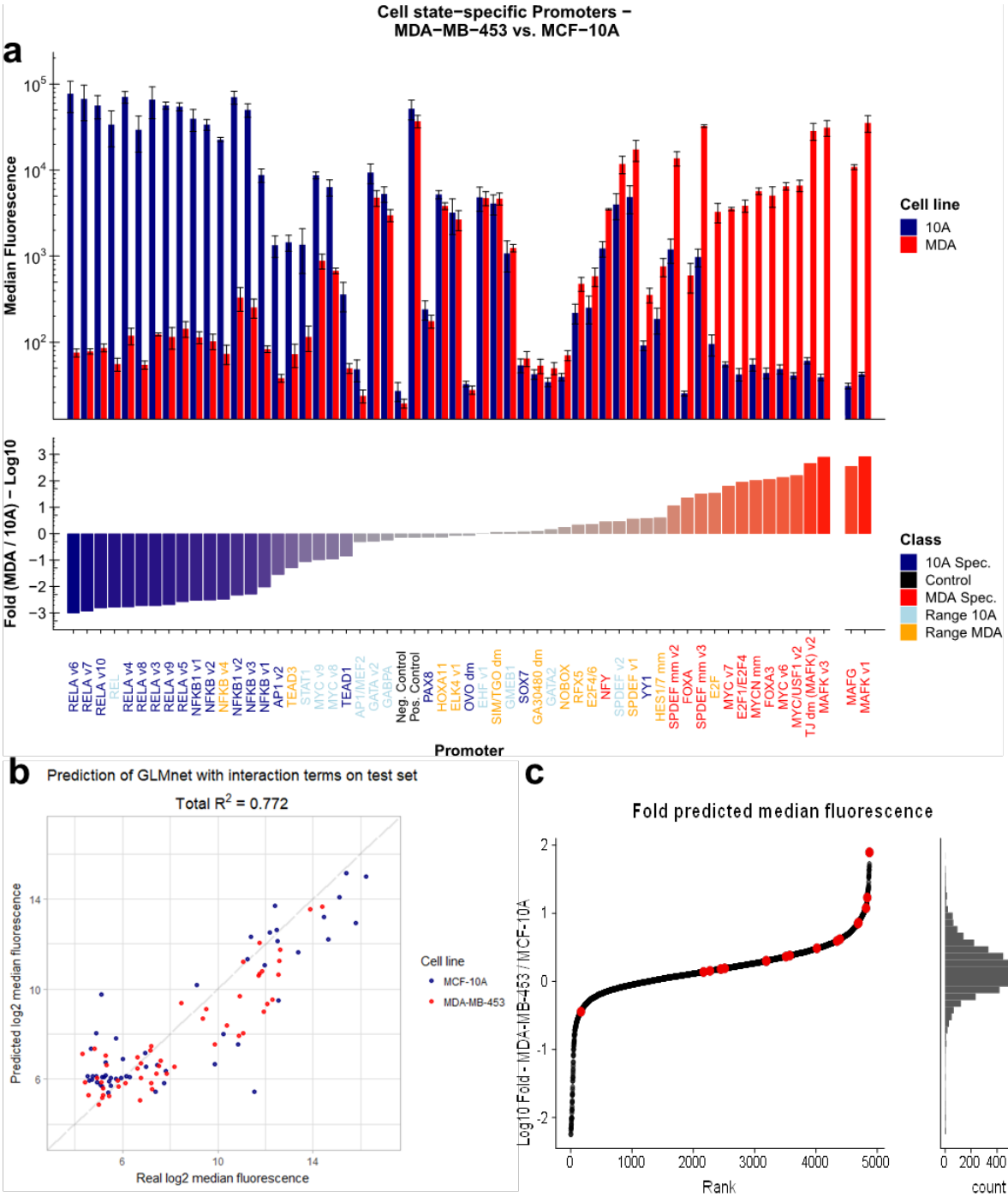


Figure 3-3: Machine-learning based prediction model can efficiently predict cell state specificity. (a) Validation guided by machine-trained algorithms. We selected 54 promoters predicted to be specific to either of the cell states of the cell lines MDA-MB-453 (a breast cancer cell line, MDA) or MCF-10A (non-tumorigenic breast cell line, 10A), or to have a range of fluorescence in either cell state (defined as four "classes" of promoters). Specific promoters showed up to ≈1000-fold difference in activity between cell states and exhibited activity as strong as that of a constitutive promoter (Ubiquitin C promoter) commonly used for gene expression (also used as the positive control, Pos. Control). The negative control sample

(Neg. Control) consisted of cells infected with a non-fluorescent protein. Names refer to the TF-BS in the promoter. All the promoters shown here are taken from the newly generated validation set, except for MAFK v1, which was identified by the Top 5% approach, and MAFG, which was taken from the training data. The dots represent the values of three biological replicates. **(b)** The machine-learning based prediction model achieved a Pearson R^2 of 0.77 between the prediction and true fluorescence measured by FACS (log_2 scaled) on a held-out test set. **(c)** Inspecting the predicted fold difference of all promoters in the library showed that there were plenty of promoters specific to each cell state. The Top 5% approach identified cell state-specific promoters (in red) in a significant manner (p = 0.0016, Wilcoxon rank sum test, two-sided). Error bars represent S.E.M., N= 3 biological replicates.

3.5 SPECS identify glioblastoma stem-like cells

We next applied our approach to identify promoters that specifically target cancer stem cells, which are generally resistant to radiation and chemotherapy.[134] For this purpose, we used a clinically relevant patient-derived glioblastoma cell model.[135] Glioblastoma stem-like cells (GSCs) were isolated from the dissociated tumor specimen of patient MGG4 by sphere culture in defined growth-factor supplemented media, while bulk differentiated MGG4 glioblastoma cells were isolated from the same tumor specimen by adherent culture in serum-containing media.[136] In contrast to serum-cultured glioblastoma cells (ScGCs), GSCs are highly tumorigenic and epigenetically distinct, and also express different transcription factors.[136–138] We introduced our SPECS library into both MGG4 GSCs and ScGCs and utilized FACS sorting, NGS, and computational analysis to identify GSC-specific promoters. From the computational analysis, we noticed that the coverage of our library was low, probably due to cell death caused by the FACS sorting. The low library coverage reduced our ability to accurately predict promoter activity. Nevertheless, several of the most important features identified by our machine learning model (Figure S3-5) were still calculable. These features were chosen based on having the largest coefficients in the MDA-MB-453 vs. MCF-10A model, leading to the highest contribution to the previous model predictions. Thus, this subset of features was used to manually identify potential SPECS. These features included total counts over all bins and counts in the negative bin, as well as a determination of which bin had maximal counts (see Supplementary Text 2 for detailed information). Using these features, we identified 30 candidate promoters potentially having distinct activity in the GSC vs. ScGC state of the MGG4 cells (Figure 3-4, upper panel). Among 15 promoters predicted to be ScGC-specific, five promoters showed higher activity in ScGCs compared to GSCs, ranging from 27-fold to 462-fold higher activity (Figure 3-4, lower panel). Among 15 promoters predicted to be GSC-specific, one promoter showed 100-fold higher activity in GSCs compared to ScGCs (Figure 3-4, lower panel). These promoters could be used for targeting glioblastoma cells that are resistant to traditional therapies in patients, as well as for basic biological studies of glioblastoma cancer stem cells.

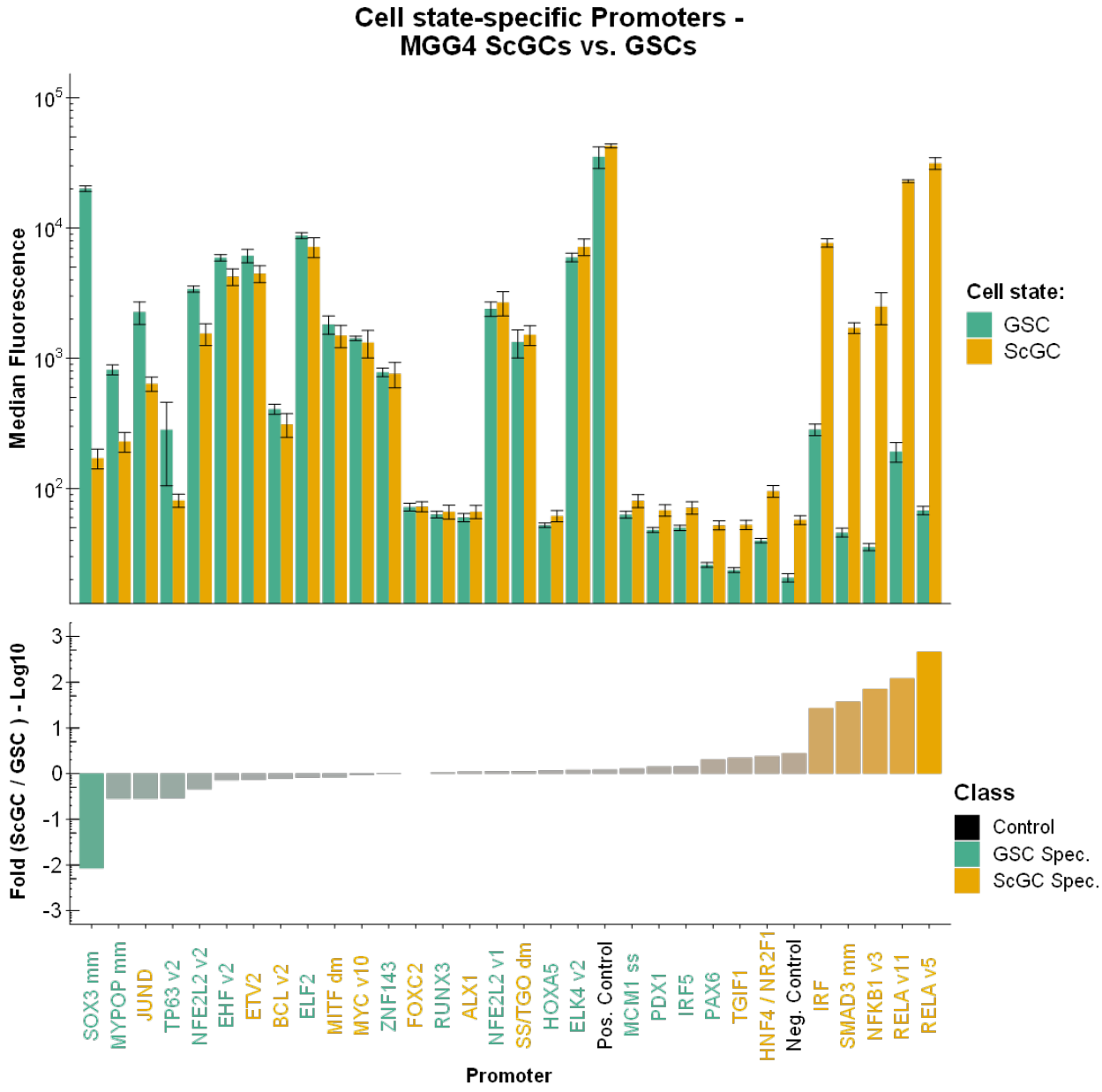


Figure 3-4: Promoter activities in glioblastoma stem-like cells (GSCs) and serum-cultured glioblastoma cells (ScGCs). Thirty promoters predicted to be specific to either MGG4 ScGCs or GSCs were validated (defined as two "classes" of promoters). Among the 15 promoters predicted to be ScGC-specific, five showed >10-fold higher activity in ScGCs compared to GSCs, ranging from 27-fold to 460-fold higher activity. Among the 15 promoters predicted to be GSC-specific, one showed 100-fold higher activity in GSCs compared to ScGCs. The upper panel depicts the median fluorescence intensity of each promoter. The blue bars denote the activity in MGG4 GSCs, and the yellow bars denote the activity in MGG4 ScGCs. The lower panel shows the $\log_{10}$ difference in activity between MGG4 ScGCs and GSCs for each promoter. The name on the X-axis denotes the TF-BS of each promoter. The dots represent the values of three biological replicates. Error bars represent S.E.M., N = 3 biological replicates. Source data are provided as a Source Data file

3.6 Discussion

In this study, we present a high-throughput screening and computational pipeline for the systematic discovery of SPECS with superior cell-state specificity. This pipeline enabled the identification of SPECS for a variety of cell states, including SPECS with: (i) distinct spatiotemporal activity in an organoid differentiation model; (ii) specificity for either a breast cancer or a normal breast cell line; and (iii) discrimination of stem-like glioblastoma cells from their differentiated counterparts. Two major advantages of using a fluorescent protein as an output for the SPECS library compared to using non-fluorescent protein are that promoter activity can be measured at the single cell level and that cells can be separated into distinct populations by FACS sorting based on promoter activity. This approach can be used to study promoter activity in living cells, tissues, or even entire organisms (if they are transparent, e.g., *C. elegans*) and track their activity for prolonged periods of time. We developed a machine-learning based prediction model to predict the activity of all the promoters in our library in each individual cell state. This approach enabled us to identify promoters showing a wide range of desired activities as well as promoters exhibiting very high cell state specificity. Similar approaches have been taken in studying transcriptional regulation of unicellular organisms but usually require a large number of cells and many fluorescence bins to achieve accurate estimations of promoter activity.[98,139] Our machine-learning based computational approach enabled us to use fewer fluorescence bins to achieve good accuracy in prediction, thereby facilitating screening while also allowing an accurate estimation of promoter activity in human cells. Several issues can be addressed to improve the pipeline. For example, the FACS sorting step can be cytotoxic to some cells, like primary GCSs, causing unwanted cell death; in this case, the pipeline requires large numbers of cells and yields low library coverage, hence making the computational prediction of promoter activity more challenging. In the future, gentler cell sorting methods and additional refinements of the prediction algorithms would improve the screening process. Furthermore, additional work is required to extend this approach to accommodate a wider range of cellular conditions. Our approach can efficiently screen for cells that can be cultured in vitro for a reasonable amount of time. However, further development is required to enable this screening approach to be used

for short-lived cell samples such as patient-derived tissues. In the future, this approach may be developed for high-throughput real-time analysis of TF activity, which is challenging to measure using current methods. Existing approaches such as RNA-seq or TF ChIP-Seq generally measure only TF expression levels or genome-wide binding profiles in dead cells or cell lysates. Our approach is essentially a massively parallel reporter assay for TFs following a thorough analysis of the exact TF that binds each synthetic promoter. Thus, this method can be used to isolate the regulatory effect of the binding of a single TF, while disregarding the regulatory effects of other transcriptional and post-transcriptional effectors. In summary, our high-throughput systematic approach efficiently identifies SPECS displaying up to a 1000-fold activity difference between cell states of interest and their counterparts. This approach can be used to find SPECS for a myriad of cell states and types. Our platform could be applied to the design of sensors for synthetic gene circuits, and could also be used for other applications in basic biological research, biotechnology, and biomedicine.

3.7 Methods

3.7.1 SPECS Library Construction

For the construction of the SPECS library, all position weighted matrices (PWMs) from two databases, The ENCODE project[124] and CIS-BP,[125] were downloaded. These databases contain binding motifs derived from direct binding assays (SELEX, HT-SELEX, PBM, ChIP-Seq, etc.) from several organisms. In order to create a consensus sequence for each PWM, the maximum probability nucleotide from each position of the PWM was taken. The reverse complement sequence of each consensus sequence was also used. The list contains 6107 unique motifs (including the reverse complement), derived from 1095 TFs (of which 665 are human) from 71 species. Each promoter consisted of parts shared by all promoters: plasmid backbone, global primers, and restriction sites. The variable parts were the TF-BS repeats. To create the variable part of the promoters, each

consensus TF-BS was repeated k times, where k is equal to 129 bp divided by the TF-BS length +3 bp (spacer), rounded down to the nearest integer. Each promoter was also associated with a 17 bp unique random barcode for later retrieval using the barcode as a primer. All the oligonucleotides containing the tandem TF-BSs in the synthetic promoter library were synthesized as a set of ≈150 bp pooled oligonucleotides by array-based DNA synthesis from Twist Bioscience (San Francisco, CA). These oligonucleotides were further cloned into lentiviral vectors with conventional restriction enzyme cloning, upstream of an adenovirus minimal promoter to control the expression of mKate2 fluorescent protein gene.

3.7.2 Cell culture and cell lines

MDA-MB-453, MCF-10A, and HEK-293T cells were obtained from the American Type Culture Collection, Rockville, MD (MDA-MB-453, Catalog #HTB-131; MCF-10A, Catalog #CRL-10317; HEK-293T, Catalog #CRL-3216). MDA-MB-453 and HEK-293T cells were cultured in DMEM (Life Technologies, Carlsbad, CA) supplemented with 10% fetal bovine serum (FBS; VWR, Radnor, PA; Catalog #95042–108), 1% Non-Essential Amino Acids (MEM/ NEAA; Hyclone; Catalog #16777–186), and 1% Pen/Strep (Life Technologies Catalog #15140–122) at 37 °C with 5% CO2. MCF-10A cells were cultured in MEGM BulletKit (Lonza, Walkersville, MD; Catalog #CC-3151 & CC-4136). All cell lines were banked directly after being purchased from vendors and used at low passage numbers. MGG4 GSCs[136,137] were cultured in neurobasal media (Thermo Fisher Scientific; Catalog #21103049) supplemented with 3mM L-Glutamine (Corning, Corning, NY; Catalog #25–005-CI), 1x B27 supplement (Thermo Fisher Scientific; Catalog #17504044), 0.5x N2 supplement (Thermo Fisher Scientific; Catalog #17502048), 2 µg/mL heparin (Sigma; Catalog #H3149), 20 ng/mL recombinant human EGF (R&D systems, Minneapolis, MN; Catalog #236-EG-200), 20 ng/mL recombinant human FGF-2 (PeproTech, Rocky Hill, NJ; Catalog #100–18B), and 0.5x Penicillin/Streptomycin/Amphotericin B (Corning; Catalog

#30–004-CI). MGG4 ScGCs (also referred to as FCS cells or DGCs) were cultured in DMEM with 10% FBS.

3.7.3 Virus production and cell line infection

Lentiviruses containing the synthetic promoter library were produced in HEK-293T cells using co-transfection in a six-well plate format. In brief, 12 µl of FuGENE HD (Promega, Madison, WI) mixed with 100 µl of Opti-MEM medium (Thermo Fisher Scientific, Waltham, MA) was added to a mixture of 4 plasmids: 0.5 µg of pCMV-VSV-G vector, 0.5 µg of lentiviral packaging psPAX2 vector, 0.5 µg of lentiviral expression vector of the library, and 0.5 µg of lentiviral expression vector constitutively expressing ECFP. During 20 min incubation of FuGENE HD/DNA complexes at room temperature, HEK-293T suspension cells were prepared and diluted to 3.6 × 106 cells/ml in cell culture medium. 0.5 ml of diluted cells (1.8 × 106 cells) were added to each FuGENE HD/DNA complex tube, mixed well, and incubated for 5 min at room temperature before being added to a designated well in a six-well plate containing 1 ml cell culture medium, followed by incubation at 37 °C with 5% CO2. The culture medium of transfected cells was replaced with 2.5 ml fresh culture medium 18 h post-transfection. Supernatant containing newly produced viruses was collected at 48-h post-transfection, and filtered through a 0.45 µm syringe filter (Pall Corporation, Ann Arbor, MI; Catalog #4614). For infecting target and control cells for primarily single copy vector integration, various dilutions of filtered viral supernatants were prepared to infect 5× 106 MDA-MB-453, MCF-10A, MGG4 GSC, and MGG4 ScGC cells in the presence of 8 µg/ml polybrene (Sigma) overnight. Five days after infection, the dilutions producing around or below 15% of cells expressing ECFP were selected for further expansion and sorting.

3.7.4 Lentiviral library introduction to cells of interest

By infecting the cells with different titrations of viruses and selecting the titration that gave around 15% infectivity based on the percentage of ECFP positive cells (see the above virus production and cell line infection section for details), we expected the integration of a single copy of the promoter in most of the infected cells. To ensure the reproducibility of our screening results, we maintained >100-fold coverage of each library member throughout the screening pipeline. Infected cells were further expanded and FACS sorted into five subpopulations based on distinct levels of mKate2 activity (Figure 3-1b).

3.7.5 Flow cytometry

To characterize fluorescent protein expression, cells were resuspended with DMEM and analyzed by a LSRII Fortessa cytometer (BD Biosciences, San Jose, CA). Data analysis was performed by FlowJo software (TreeStar Inc, Ashland, OR).

3.7.6 FACS sorting

To further characterize fluorescent protein expression and sort cells into different bins of fluorescence intensity, cells were resuspended with FACS buffer (PBS +1% FBS) and sorted by an BD Aria cell sorter (BD Biosciences, San Jose, CA). For the first sorting, cells were sorted into fluorescence positive and negative bins. The sorted fluorescence positive cells were continuously cultured and expanded for the second sorting. For the second sorting, fluorescence positive cells were sorted into top 5%, top 5–10%, high, and low fluorescence bins. The high and low fluorescence bins were created by equally splitting the remaining 90% of fluorescence positive cells into two halves.

3.7.7 Next-generation sequencing

For NGS library preparation, DNA from each sample was extracted and 250 ng of genomic DNA were used as template for PCR amplification with a global primer (Pi5) and a distinct primer (Pi7) for sample barcoding. Sequencing was performed at the MIT BioMicro Center facilities on an Illumina MiSeq machine to yield 150 bp single-end reads. Each lane was loaded with 12 samples to achieve approximately 1 × 106 reads per sample.

3.7.8 Pre-processing of NGS data

Fastq files were first inspected for quality control (QC) using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (version 0.11.5). Fastq files were then filtered and trimmed using fastx_clipper of the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) (version 0.0.14). Only reads containing the 3′ restriction site Asc1 created during the library construction were kept. The restriction site was trimmed leaving only the variable promoter sequence. FastQC was run again to inspect the quality after trimming. Trimmed fastq files were collapsed using fastx_collapser of the FASTX-Toolkit. The collapsed fasta file was used as an input for alignment in Bowtie2 with a very sensitive alignment mode and aligned against the library reference.[140] The resulting SAM file was filtered for mapped reads using SAMtools,[141] and the reads were then quantified by summing the counts of each unique promoter using an in-house R script. The reads were normalized by dividing all reads in the sample by a size factor estimated by DESeq2.[142] Correlation among technical and biological replicates for each of the NGS samples was calculated, with $R^2 \approx 0.8$ between technical replicates and $R^2 \approx 0.3$ between biological replicates. The promoters were then filtered, and only

promoters with counts in at least two replicates (biological or technical) in both cell lines were retained, leaving 4872 promoters total.

3.7.9 Fluorescence estimation

To estimate the fluorescence for all promoters in each of the cell lines, a machine learning approach was used. First, fluorescence data were collected for training, based on measurements of whole populations infected with a single promoter from the library. Promoters for the training set were chosen based on an approximate measure of fluorescence denoted as the activity score. The activity score was used to find promoters representing a broad spectrum of fluorescence values in each cell line to be used as training data, as we hypothesized that using random promoters would lead to mostly non-active promoters. This activity score (A) is a weighted-average-like heuristic, calculated by multiplying the mean fluorescence of each bin (as depicted in the gates) by the proportion of $\log_2$ transformed counts in each bin. It follows the Eq. (1.1) for some promoter labeled as *i*:

$$A_i = \frac{\sum_b \overline{y}_b n_{i,b}}{\sum_b n_{i,b}} \tag{1.1}$$

Where $\overline{y}_b$ is the mean fluorescence in some bin *b* and $n_{i,b}$ is the $\log_2$ normalized counts for that promoter for that bin. We identified 64 candidate promoters estimated to show a range of fluorescence activity in MDA-MB-453 and MCF-10A cells based on this activity score metric. Next, normalized counts, as well as fluorescence measurements for 81 promoters (64 + 17 from random top 5% shotgun cloning approach) in MDA-MB-453 and MCF-10A cell lines, were obtained for generating a machine-learning based predictive model. Fluorescence measurements were processed using flowCore in R to calculate the median fluorescence for each promoter.[143] The median fluorescence was $\log_2$ transformed to serve as the target value. Training was performed using a 60/40 train/test split and taking a five-times 5-fold repeated cross-validation using the caret package in

R.[144] Normalized counts were log_2 transformed and several features engineered based on the perceived counts-fluorescence relationship. Briefly, the number of counts per bin (and total) as well as relationships between bins were used as features. First degree interaction terms between features were included as well (Supplementary Text 1). We tested the performance of linear regression (lm), generalized linear model with elastic net regularization (GLMNET),[133] random-forest regression and SVM regression with a linear, polynomial or radial kernel. RMSE and R-squared values were used to evaluate the models on fitting log_2 median fluorescence on the training set, test set, and a separate biological validation. Performance was evaluated on cross-validation on the training set (Figure S3-5). A separate biological validation (54 promoters) was then incorporated into the data and the models trained for a second time using the same parameters. The updated models were evaluated on the new training and new test sets. The chosen model was GLMNET with interaction terms (GLMNET-inter) based on its performance on both data — with and without biological validation. The model trained on the data with the biological validation was then used to predict log_2 median fluorescence for all the library promoters in both cell lines. For MGG4 GSCs and ScGCs, fluorescence was estimated manually based on a subset of the metrics, which were calculable under the low coverage condition (Supplementary Text 2).

3.7.10 Differentiation and infection of liver organoids

The SPECS library was introduced into a liver bud-like organoid derived from GATA6-expressing iPSCs.[27] Five days before the promoter library transduction, 2D organoids were prepared by seeding 2.5×10^4 GATA6-expressing iPSCs in each well a of matrigel-coated, flat-bottom 24-well plate. iPSC differentiation was initiated by Doxycycline (Dox)-induced (1 μg/mL) GATA6 expression in mTeSR1 media (STEMCELL Technologies Vancouver, Canada) for 5 days.[27] On day 5, organoids were transduced with a 1:1 mixture of the SPECS library virus and an infection control UbCp-ECFP virus. The viral titer was serially diluted to ensure that <15% of the cells expressed the transduction marker. After

viral transduction, the media was switched to the non-pluripotency supporting media APEL2 (STEMCELL Technologies) for further organoid differentiation. Differentiation continued for a total of 16 days, after which organoids were dissociated to single cells with Accutase (STEMCELL Technologies) for FACS sorting of the mKate2 positive population by BD Aria FACS sorter (BD Biosciences). The genomic DNA was purified from the sorted mKate2 positive population, and the SPECS library region was amplified with standard PCR with 50 amplification cycles. The amplified promoters were cloned into a lentiviral vector backbone by standard restriction digestion cloning with enzymes AscI and SbfI. Colonies were randomly picked, and plasmid DNA was submitted for Sanger sequencing. Candidate promoters identified by Sanger sequencing were further validated for their spatial and temporal behavior in organoids. We discarded promoters with no detectable activity (false positives from the screening) or whose activity could not be replicated, which reduced the initial 37 promoters to a set of 4 with a distinct spatial and temporal behavior. We transduced undifferentiated GATA6-expressing iPSCs with lentivirus containing a single promoter driving mKate2 expression in biological triplicates. We seeded 3×10^5 GATA6-expressing iPSCs per well in a 12-well plate 2 days before lentiviral transduction. Cells were transduced with a 1:4 diluted viral supernatant with 2 μg/mL polybrene. Two days after viral transduction, transduced cells were dissociated and seeded at 2.5×10^4 cells/well in a 24-well plate (day 0). The following day, we initiated organoid differentiation by Dox as described above. Cell condition and mKate2 expression were tracked from day 0 to day 21 daily using a TCS SP5 II confocal microscope (Leica, Buffalo Grove, IL). Images were acquired as a tiled scan and automatically stitched together using the Leica Application Suite software. In-house Python and R scripts were used to apply a median filter to the red channel for noise reduction and image analysis.

3.7.11 Shotgun cloning promoter identification

Promoter plasmids created by shotgun cloning were sequenced by Sanger sequencing, and the sequencing output was aligned using Bowtie2 (version 2.2.9) with a very sensitive local alignment mode against the library reference.[140] An in-house script was used to identify mutated

3.8 Supplementary Information

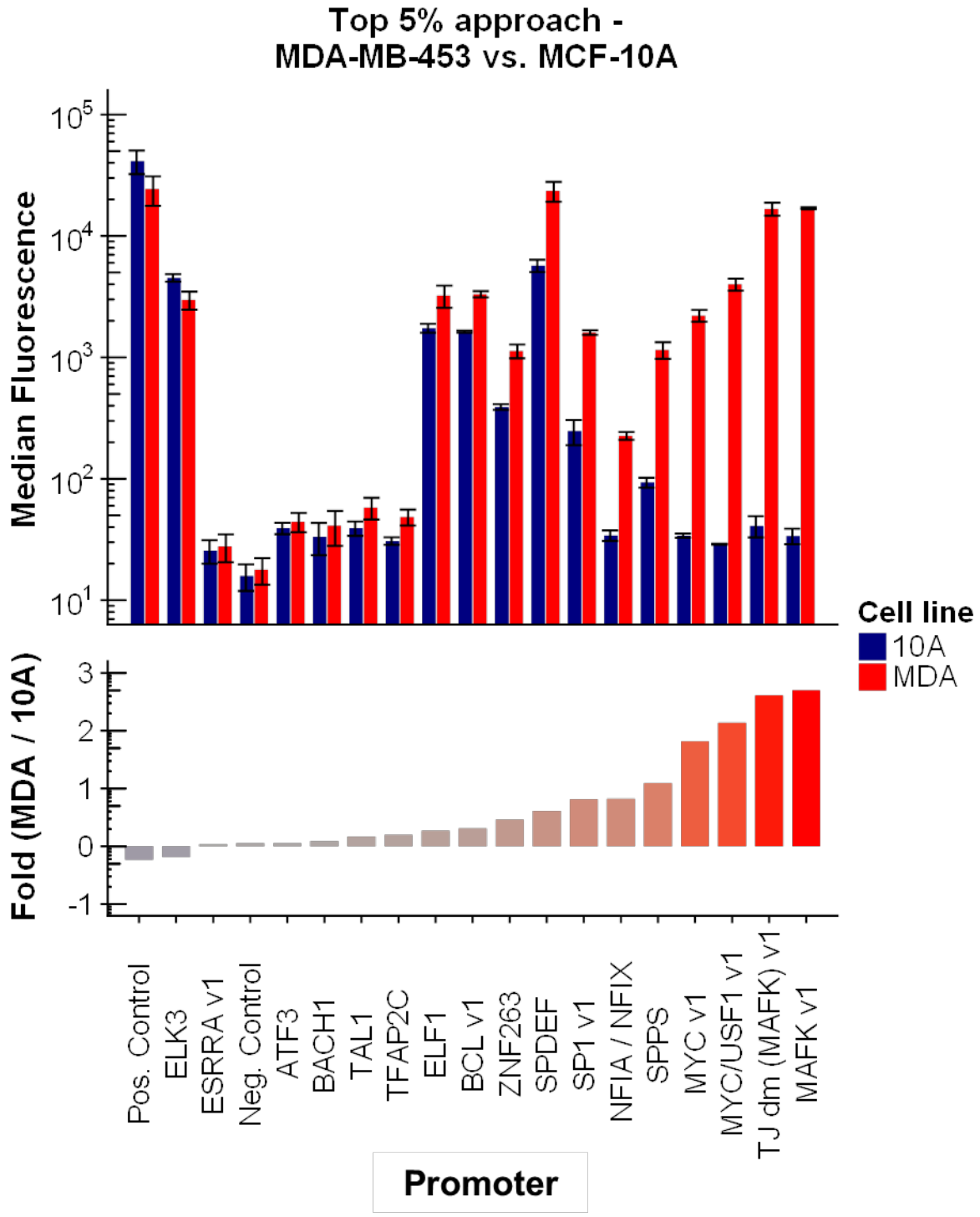


Figure S3-1: The activities of promoters identified by the Top 5% approach. Promoters identified by the Top 5% approach presented up to ≈500-fold activity difference between the breast cancer cell line MDA-

MB-453 (MDA) and the normal breast cell line MCF-10A (10A). Some of these identified promoters were highly active and achieved median fluorescence intensities comparable to that seen with a constitutive Ubiquitin C promoter (Pos. Control). The negative control sample (Neg. Control) consisted of cells infected with a non-fluorescent protein. The dots represent the values of three biological replicates. Error bars represent S.E.M., N = 3 biological replicates. Source data are provided as a Source Data file.

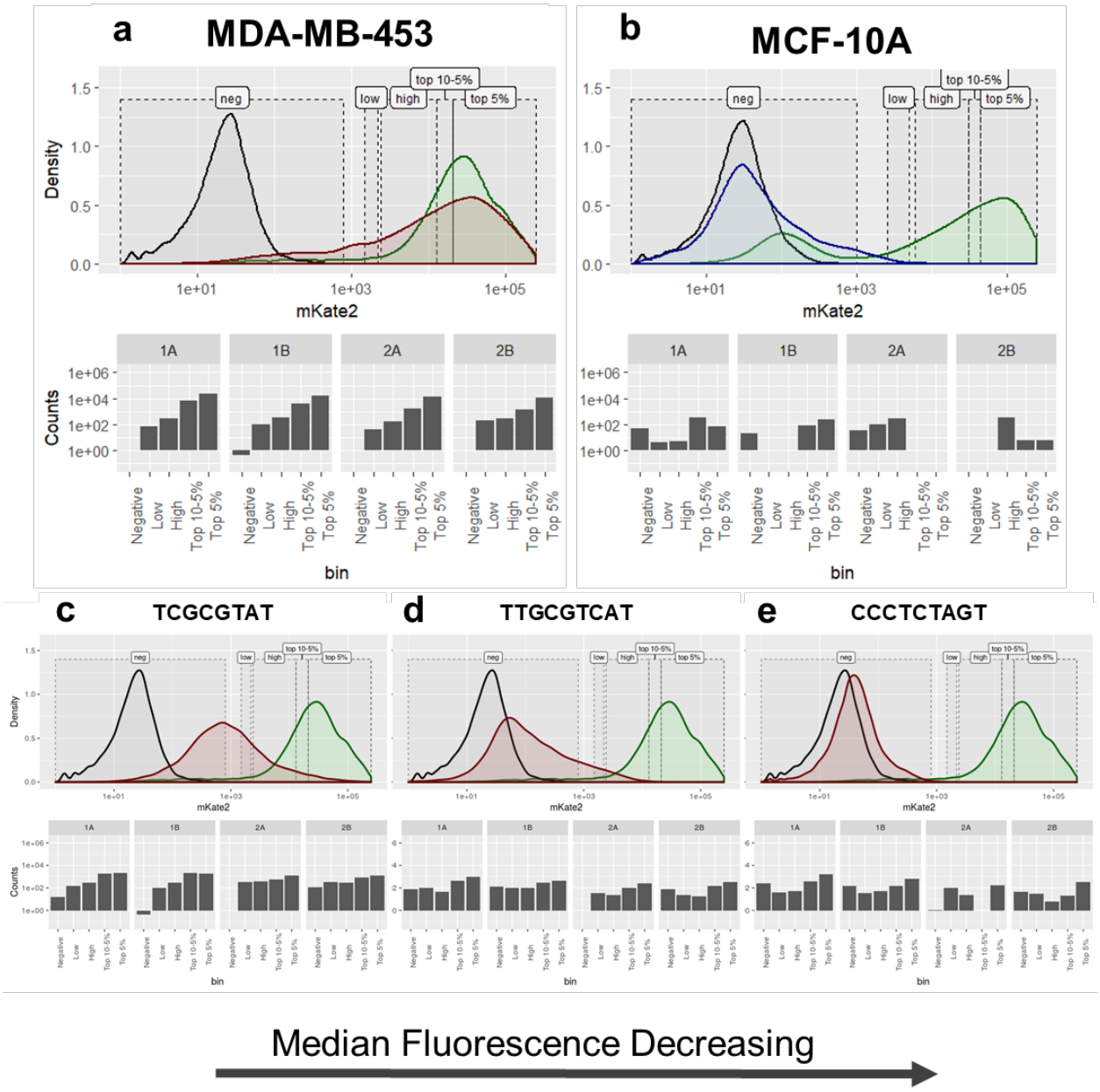


Figure S3-2: Relationship between fluorescence and NGS read counts. (a) Comparison of the fluorescence distribution and NGS normalized counts of a promoter that contains the MAFK V1 TF-BS–(TGCTGAGTCAGCA) from the Top 5% shotgun cloning approach. This promoter exhibited very high activity in MDA-MB-453 cells and very low activity in MCF-10A cells. Dashed boxes represent the FACS gate for each bin. The numbers 1 and 2 denote data from 2 independent screening experiments. The letters A and B denote PCR technical replicates amplified from the promoter locus from genomic DNA for NGS. We observed that the fluorescence distribution of this promoter in MDA-MB-453 (red line) was comparable to that of the positive control UbC promoter (green line) and was much higher than that of the negative control sample (grey line). There were much higher counts in the positive bins than in the negative bin, with the highest counts being in the top 5% bin. **(b)** On the contrary, the fluorescence distribution of the same promoter in MCF-10A cells (blue line) was similar to that of the negative control sample (grey line) and

much lower than that of the positive control sample (green line). **(c-e)** When three promoters with decreasing fluorescence intensities in MDA-MB-453 cells were compared (from Figure S3-2c to S3-2d to S3-2e; red lines in these 3 panels), there was a trend of decreasing total counts and a more uniform distribution of counts. A shift of counts from positive fluorescence bins to lower fluorescence or negative bins was also observed. For all panels, the negative control sample (grey line) consisted of cells infected with a non-fluorescent protein, and the positive control sample (green line) consisted of cells infected with a Ubiquitin C promoter expressing mKate2.

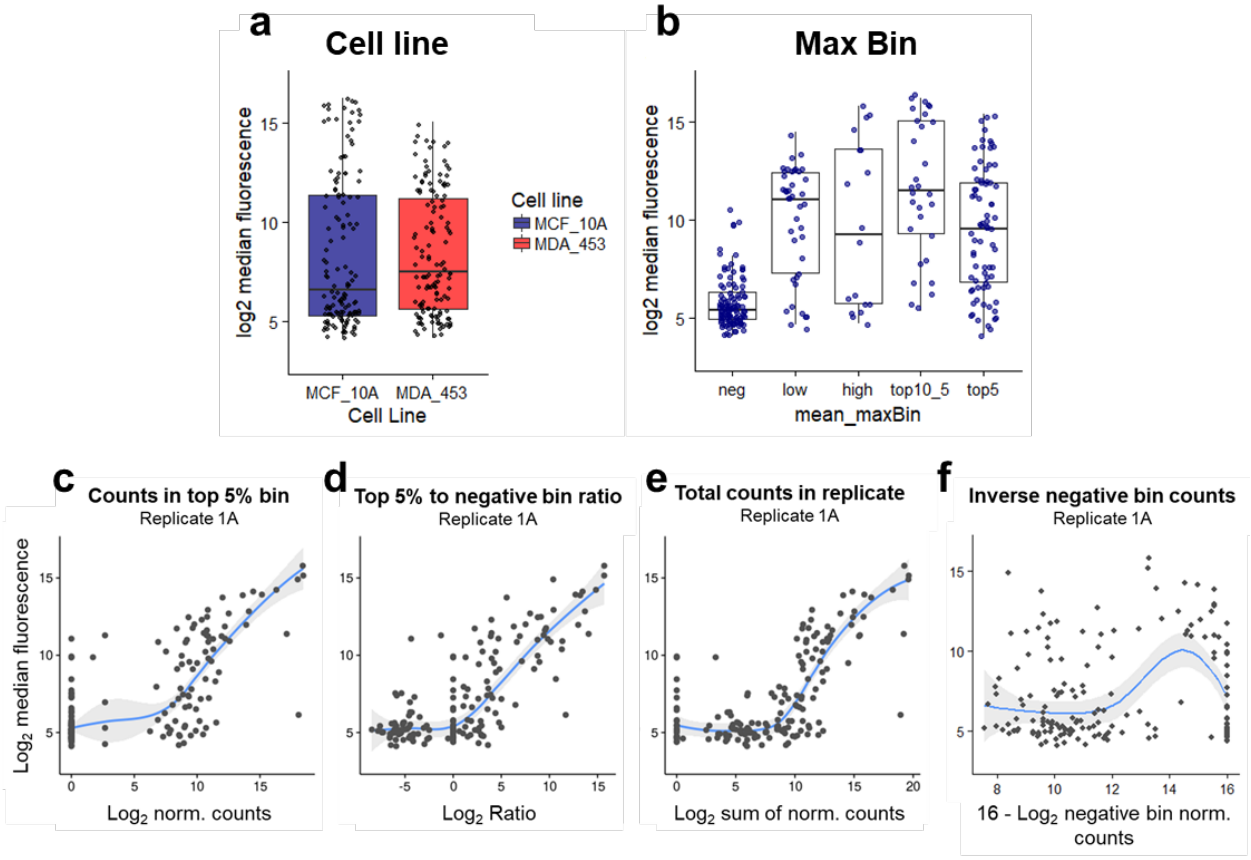


Figure S3-3: Machine-learning features. (a) In the model, the type of cell line is used as a categorical feature. As seen by the median, MDA-MB-453 cells had a greater fluorescence than MCF-10A cells; this difference contributed to model predictions. Each dot represents the fluorescence intensity of an experimentally tested promoter. **(b)** Maximal Bin (Max Bin), the bin having the most counts for a specific promoter, is another categorical feature. The negative and top 5% bins contained the most promoters tested, with the top 5% showing greater fluorescence by the median than the negative bin. All other bins contained much fewer promoters tested. Each dot represents the fluorescence intensity of an experimentally tested promoter. For **(a)** and **(b)**, the boxes denote the lower quartile, the median, and the upper quartile. Whiskers denote the minimum and maximum up to 1.5x interquartile range. **(c-f)** Feature values are shown for continuous features (X-axis) plotted against $\log_2$ observed median fluorescence (Y-axis) for all validated promoters. Each dot represents the data from a validated promoter. Features include: $\log_2$value of the normalized counts of promoters in the top 5% bin **(c)**; count ratio of the promoters in the

top 5% bin to the negative bin **(d)**; total normalized counts for the promoters from all the biological and technical replicates **(e)**; and the “reverse” (16 (max) – $\log_2$counts) of the negative bin counts **(f)**. These features show a monotonically increasing approximation for fluorescence, with counts in the negative bin showing an inverse relationship. A description of the axes is provided in each subplot, and the blue line is the loess regression with the grey area being the 95% confidence interval. Feature data were displayed for all experimentally tested promoters

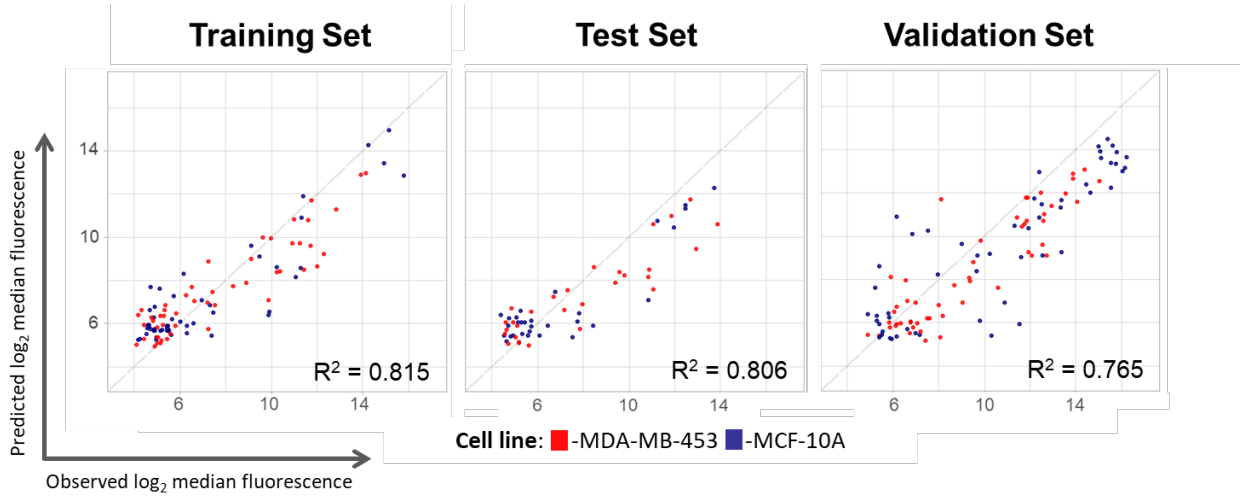
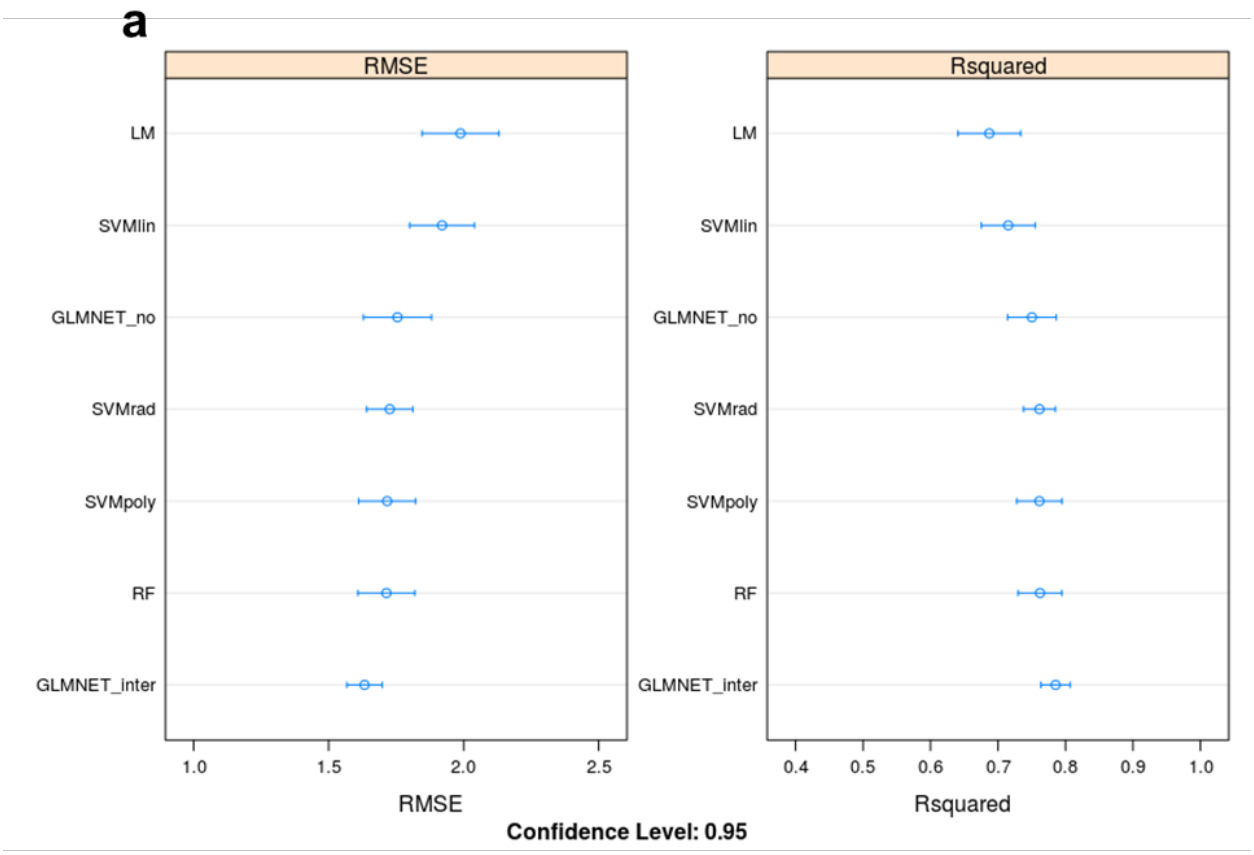


Figure S3-4: Observed vs. predicted fluorescence for the 1st round of the machine learning predictions. Observed fluorescence data were compared with the fluorescence predicted by the GLMNET-inter model from the 1st round of machine learning prediction. Training (left) and test (middle) sets achieved comparable R^2 values, signaling that the model did not overfit the data. When the model was tested on the validation set (right), it performed less well for the midrange of fluorescence ($\approx 2^8 - 2^{12}$), especially in MCF-10A cells. The X-axis denotes the observed $\log_2$ median fluorescence; the Y-axis denotes the predicted $\log_2$ median fluorescence (this GLMNET-inter model was trained with data from 81 promoters: 17 from the Top 5% approach and 64 selected from the activity score metrics). The validation set consisted of an additional 54 promoters selected from the prediction result from the machine learning algorithm (GLMNET-inter = generalized linear model with elastic net regularization, using features and interaction terms between features).


a
RMSE
Rsquared
LM
SVMlin
GLMNET_no
SVMrad
SVMpoly
RF
GLMNET_inter
1.0
1.5
2.0
2.5
0.4
0.5
0.6
0.7
0.8
0.9
1.0
RMSE
Rsquared
Confidence Level: 0.95

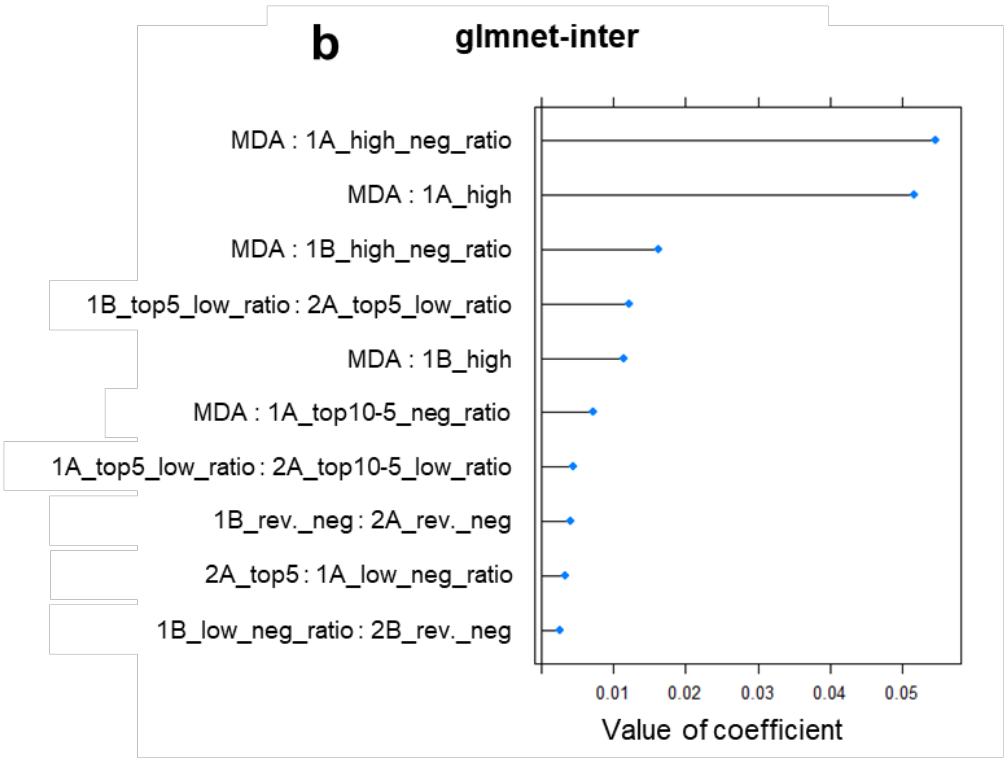

b
glmnet-inter
MDA : 1A_high_neg_ratio
MDA : 1A_high
MDA : 1B_high_neg_ratio
1B_top5_low_ratio : 2A_top5_low_ratio
MDA : 1B_high
MDA : 1A_top10-5_neg_ratio
1A_top5_low_ratio : 2A_top10-5_low_ratio
1B_rev._neg : 2A_rev._neg
2A_top5 : 1A_low_neg_ratio
1B_low_neg_ratio : 2B_rev._neg
0.01
0.02
0.03
0.04
0.05
Value of coefficient

Figure S3-5: Model performance and feature importance for the 2nd round of the machine learning predictions. (a) Models were compared based on RMSE and R-squared (R^2). Most of the models performed similarly with an R^2 of ≈0.75 and RMSE of ≈1.6-1.8 (with standard deviation of the data being 3.4), except for LM and SVMLin, which did not perform quite as well. GLMNET-inter was the best model, having a slight margin of performance over the others. Data presented in the plot contains the distribution of summary function (RMSE or R^2) for all resamples done for the model in the repeated CV when predicting the portion of the data which was left out. Model abbreviations: LM – linear regression; SVMLin – SVM with a linear kernel; GLMNET_no – generalized linear model with elastic net regularization (GLMNET) with no interaction terms; SVMRad – SVM with a radial kernel; SVMPoly – SVM with a polynomial kernel; RF – random forest regression; GLMNET_inter – GLMNET using features and interaction terms between the features; RMSE – root mean squared error, R^2 – Pearson's correlation coefficient. Error bars represent 95% CI. **(b)** Examining the coefficients of the features in the GLMNET-inter model, we identified as the most important features as: cell line (being MDA-MB-453), counts in high ("high") and top 5% bins ("top5"), high to negative bin ratio ("high_neg_ratio"), top 5% to low ratio ("top5_low_ratio"), or top 10%-5% to low ratio ("top10-5_low_ratio"), and the inverse negative ("rev._neg"). 1A-2B represent biological and technical replicates with the numbers denoting the biological replicates and the alphabets denoting the technical replicates.

		Predicted		Meassured		Fold	
Class	TFBS	MDA-453	MCF-10A	MDA-453	MCF-10A	MDA-453	MCF-10A
10A Spec.	RELA v6	41	22735	76	77233	0.00	1021.15
10A Spec.	RELA v7	66	13123	78	67400	0.00	858.96
10A Spec.	RELA v4	80	19205	119	70800	0.00	593.79
10A Spec.	RELA v8	50	13020	55	29367	0.00	538.51
10A Spec.	RELA v3	32	14349	123	66200	0.00	538.21
10A Spec.	RELA v9	40	36110	116	56500	0.00	488.47
10A Spec.	RELA v5	38	21284	144	54633	0.00	380.63
10A Spec.	NFKB1 v1	48	23583	114	39433	0.00	345.30
10A Spec.	NFKB v2	61	38401	103	33867	0.00	329.34
10A Spec.	NFKB1 v2	41	28701	331	70633	0.00	213.61
10A Spec.	NFKB v3	69	33599	254	50067	0.01	197.37
10A Spec.	NFKB v1	18	6315	83	8772	0.01	105.85
MDA Spec.	NFY	2640	28	3507	1227	2.86	0.35
MDA Spec.	SPDEF mm v2	5137	31	13749	1202	11.44	0.09
MDA Spec.	FOXA	4134	29	597	25	23.52	0.04
MDA Spec.	SPDEF mm v3	5262	48	32467	979	33.15	0.03
MDA Spec.	MYC v7	4764	41	3534	55	63.79	0.02
MDA Spec.	E2F1/E2F4	5038	51	3858	42	91.27	0.01
MDA Spec.	MYCN mm	5952	31	5700	55	103.64	0.01
MDA Spec.	FOXA3	3059	17	5043	44	114.61	0.01
MDA Spec.	MYC v6	4076	63	6508	49	133.63	0.01
MDA Spec.	MYC/USF1 v2	3091	22	6630	41	161.71	0.01
MDA Spec.	TJ dm (MAFK) v2	7016	25	28533	61	468.27	0.00
MDA Spec.	MAFK v3	2750	27	31133	39	794.89	0.00
Control	Neg. Control			19	27	0.71	1.40
Control	Pos. Control			37100	51667	0.72	1.39

Table S3-1: Cell state-specific promoters derived from the validation set. Of the 54 promoters from the validation set, 12 were predicted to be MDA-MB-453 specific and 12 to be MCF-10A specific. Of the 12 predicted to be MDA-MB- 453 specific, 11 were indeed specific (fold > 10X). For MCF-10A, 12/12 were specific. Predictions are derived from the GLMNET-inter trained on the 1st set (81 promoters) only.

Observed values are the average of median fluorescence in biological triplicates. Fold value is the activity fold-difference between the predicted cell state of interest / the other cell state.

3.8.1 Supplementary Text 1 - Features for machine learning

Features were engineered according to the observed relationship between $\log_2$ transformed counts and fluorescence. These Features included the numeric features – counts (in each bin and replicate), bin X to negative ratio (for each replicate), top 5-10% to low ratio (for each replicate), top 5% to low ratio (for each replicate), geometric mean of bin counts in replicates (for each bin), reverse negative counts (for each bin, using 16 - neg. where 2^{16} is above range measured for the FACS data), total replicate counts (for each bin) and total counts (in all replicates and bins). The categorical features included were cell line (MDA-MB453 or MCF-10A) and max. bin (negative, low, high, top 5-10%, top 5%) (Figure S3-3). Interaction terms between all features were also used in some of the models using R formula interface (as '*'). In the model chosen (GLMnet-inter) coefficients are regularized essentially performing feature selection by itself.[133]

3.8.2 Supplementary Text 2 - Identifying GSC specific promoters

Due to low coverage of the library in the MGG experiment, the machine learning model described above was not deployable. Some of the more important features (for feature importance see Figure S3-5), including total counts, counts in negative and the bin with most counts (i.e. most counts in negative bin) were still calculable. Thus, these features were used to rank the promoters and manually identify promising candidates ranking high on these metrics and showing reasonable coverage.

Chapter 4. Transcriptional regulating using novel post-PAS RNA

This chapter is based on ongoing work authored by me, Fabio Callendo, and Elvira Vitu.

4.1 Summary of chapter 4

Cells can be engineered as therapies or as disease models by introducing synthetic gene regulatory networks (GRNs) that reprograms them to perform a desired function. However, this requires precise control of the timing, population, and location in which the synthetic GRN is active. Sensors to detect the current cell state can serve as an important regulatory input to a synthetic GRN, but is limited by the risk of perturbing the cell state it is meant to detect. Here, we propose a strategy, post-PAS RNAs, that enables detection of the cell state through the transcription of cell state-defining genes, by co-expression of a desired regulatory RNA such as a miRNA or gRNA. By placing a regulatory RNA downstream of the terminator of the target gene, the effect on upstream gene expression can be minimized, thereby decreasing the risk of unintended cell state perturbations. Post-PAS RNAs can serve as inputs to synthetic GRNs and expand on the functionality of the cell. In future work, we will apply the post-PAS RNAs to directed differentiation of distinct sub-populations within an organoid.

4.2 Introduction and aim

The Central Dogma of molecular biology explains the flow of genetic information: from DNA through RNA, and into proteins that in turn carry out the majority of functions within

cells. Gene regulatory networks (GRNs) are an essential part of that structure by regulating the flow of information and maintaining cellular identity and function.[36,145] When this regulation is disrupted or overwritten, it can change the cell state or function of the cell. This provides an opportunity to engineer cells, and changing or disrupting endogenous GRNs has already been used to dedifferentiate human somatic cells into a pluripotent state,[146] drive differentiation and cell development,[27,31] and detect and destroy cancer.[54,147] More complex synthetic GRNs that factors in the cell state can further expand on these developments. For instance, by carefully controlling when, where, and under what circumstances a genetic program is active, therapeutic modalities such as CAR-T cells can be made safer by ensuring they are only acting on cancerous tissue. Similarly, this can be used for directed differentiation and maturation of organoids, which might increase the reproducibility and quality of the developed organoid.[31]

For a synthetic GRN to perform in a desired and predictable manner, it is critical that it takes into account the current cell state. That is, it is active in the right cell type and under the right conditions. This can be achieved by coupling the synthetic GRN to an endogenous GRN that is involved in a desired function, such as the expression of a transcription factor involved in differentiation, responses to external stress, or detection of pathogens or cancerous cells. Several methods exist which can facilitate this coupling between an endogenous GRN characterized by a cell-type specific response or state, and a synthetic GRN. Examples include protein fusions;[55,56] recruiting cell type-specific transcription factors using endogenous[48] or synthetic promoters (refer Chapter 3); co-expression of gRNAs[49–51] for CRISPR-based[52] GRNs; and using endogenous and cell state-specific miRNA to repress synthetic GRNs.[54] However, these methods are limited by several factors such as failing to account for certain types of transcriptional regulation including chromatin state and distal enhancers, affecting endogenous gene regulation by changing the 3'-UTR, or by changing the protein coding sequence.

In this chapter, we present a method that can be used to couple endogenous and synthetic GRNs. The method allows regulatory RNAs to be co-expressed with a gene without modifying the primary transcript. This method is based on the observation that

RNA polymerase II (RNA Pol II), the RNA polymerase responsible for transcribing most protein coding genes in mammalian cells, does not dissociate immediately from the template strand after encountering the terminator.[148,149] By placing a gRNA or miRNA downstream of the terminator (referred to as a post-PAS gRNA or miRNA) and flanking them by sequences that ensures their release from the flanking RNA, we can express both gRNAs and miRNAs conditioned on the expression of the upstream gene. Future research aims to use this approach to grow organoids with the differentiation of individual cells conditioned on the cell state. Expression of miRNAs and gRNAs conditioned on the expression of an endogenous gene can be used as an input to a synthetic GRN that drives cell differentiation. By using a post-PAS miRNA or gRNA, it is possible to rewire endogenous GRNs or couple them to synthetic GRNs while preserving the regulation of the input. For instance, by coupling a post-PAS gRNA or miRNA to Sox17, an endoderm marker,[150] a synthetic GRN can be selectively expressed in endoderm lineages.

4.3 Designing constructs for post-PAS RNA expression

Most protein coding genes in mammalian cells are transcribed from promoters regulated by RNA polymerase II, and a terminator serves as a signal to end transcription.[151] The mammalian terminator is characterized by the presence of an essential AAUAAA motif (or the close variants A[U/G]UAAA and UAUAAA; termed polyA signal or PAS) flanked by U rich upstream elements, G/U or U rich downstream elements, and a CA (or less frequently CG) dinucleotide cleavage signal (Figure 4-1).[151] Identification of a PAS is essential to initiate termination and post-translational processing of the primary mRNA, including synthesis of a polyA tail by polyadenylation at the 3'-end of the nascent mRNA to confer stability and signal nuclear export. While most experiments focus on the fate of the mRNA, experiments have shown that the RNA downstream of the PAS is transcribed and can even fold into catalytically active RNA.[152,153] This indicates there is a possibility that small non-coding RNA (ncRNA) can be placed and transcribed downstream of the PAS. However, two main barriers exist. First, it remains unclear if this RNA can be excised

as there is no cleavage signal, and secondly, the absence of a 5' G-cap and a polyA tail would make the RNA subject to rapid degradation by RNases. The observation that ribozymes can be co-transcriptionally active and play a role in termination[153] indicates that any ribozyme able to fold into its catalytically active structure might be able to release a ncRNA transcribed downstream of the PAS. While some ncRNAs such as intronic miRNAs are transcribed by RNA Pol II, they typically lack the 5' G-cap and the polyA tail as a consequence of post-transcriptional processing. To stabilize these RNAs, they might instead depend on protein interactions to sterically hinder RNases from degrading them prematurely.[154,155] Taken together, it might be possible to express ncRNAs downstream of a PAS if it is first released from the transcribed RNA, and secondly is able to interact with a protein or other RNA able to protect it from premature degradation.

Here, we propose two strategies for transcribing post-PAS RNAs (Figure 4-2): first, we propose a system for expressing synthetic miRNAs which relies on splice sites for miRNA release.[54] Secondly, we propose a system relying on self-cleaving ribozymes to excise a gRNA modeled after a method previously shown to work for gRNAs positioned in the 3'-UTR.[49] The first system is hypothesized to rely on splicing occurring after RNA pol II has committed to termination, and that the excised miRNA can be processed to the extent it will interact with its target mRNA. The second system relies on the ribozymes being able to fold into their catalytically active structures, and catalyze cleavage and excision of the gRNA. The gRNA must then be able to interact with dCas9-VPR. It is expected that the process of excising the post-PAS gRNA must be relatively fast, as the exonuclease Xrn2 might be involved in degrading the uncapped transcript that arises after cleavage of the primary mRNA.[148]

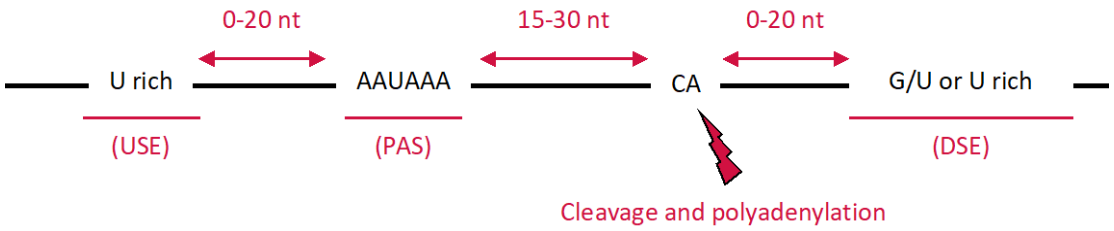


Figure 4-1: The mammalian termination signal. The polyA signal (PAS), AAUAAA, is necessary to initiate termination. This sequence is flanked by a U rich upstream element (USE), a G/U or U rich downstream element (DSE), and a CA dinucleotide which is necessary for, and signals where the mRNA is cleaved.

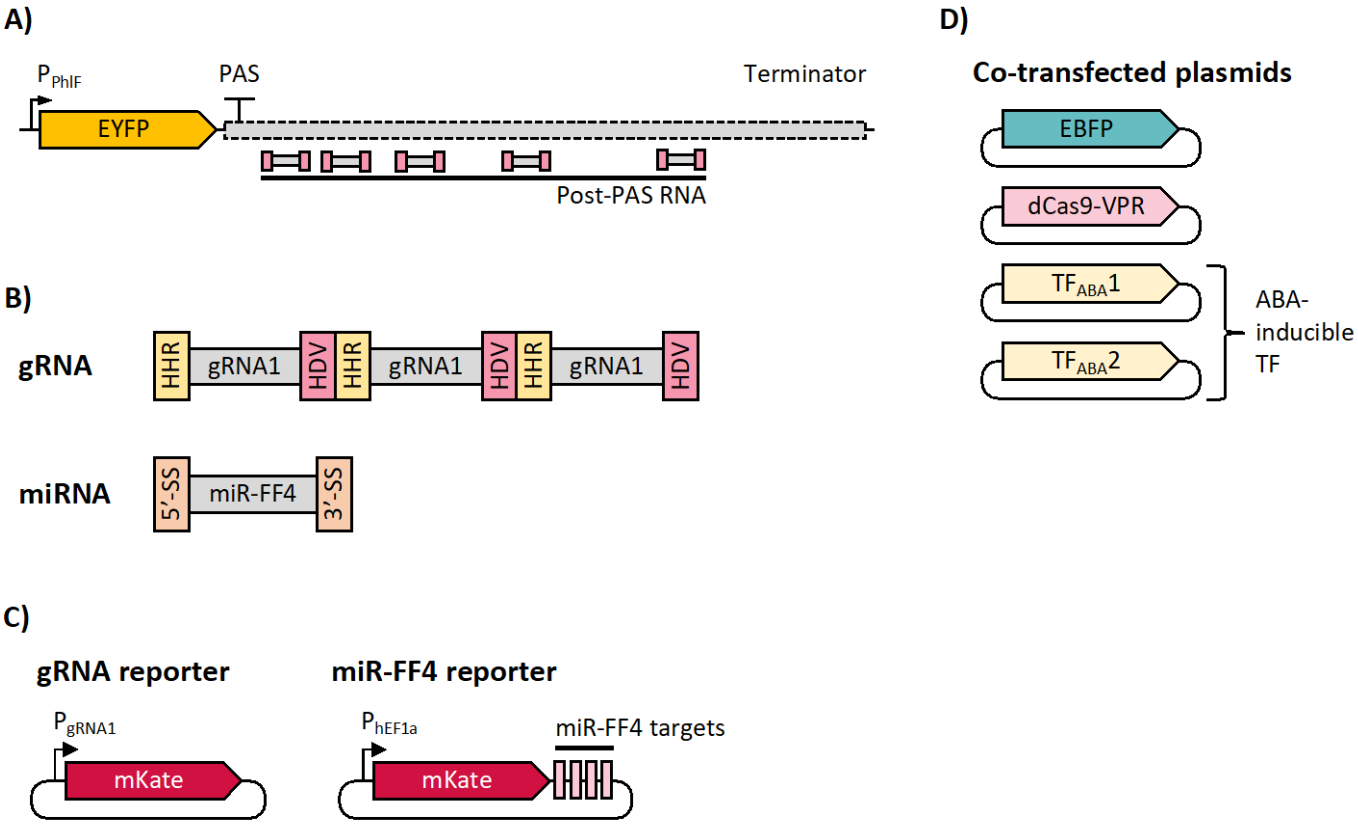


Figure 4-2: Construct design to test post-PAS RNAs. (A) Transcription of EYFP is regulated by an abscisic acid (ABA) inducible promoter (P_{PhlF}). An endogenous terminator (Hesx1, Meox1, Pyy, Sox17, or T [TBXT]), is PCR amplified from human genomic DNA and inserted downstream of EYFP (grey box). Each terminator spans ≈200 bp upstream of the annotated PAS and 3000 bp downstream of it. A single post-PAS RNA cassette is inserted in each construct at approximately 100, 200, 300, 500, and 1000 bp from the PAS, or in the 3'-UTR (not shown). **(B)** The gRNA cassette contains three repeats of a gRNA (termed gRNA1) flanked by an upstream hammerhead ribozyme (HHR) and downstream HDV ribozyme (HDV), both self-cleaving ribozymes used to release the gRNA from the RNA transcript. The miRNA cassette uses a 5' and 3' splice site (5'-SS and 3'-SS, respectively) to excise the synthetic miRNA, miR-FF4. **(C)** An mKate reporter construct is used to detect the presence of post-PAS RNA. Post-PAS gRNA expression is measured by mKate activation. The gRNA1-dCas9-VPR inducible promoter (P_{gRNA1}) is used to detect post-PAS gRNA expression through mKate activation. Post-PAS miRNAs are detected by repression of constitutively transcribed mKate by binding of miR-FF4 to the four miR-FF4 target sites located in the 3'-UTR of the mKate transcripts. **(D)** Each transfection contains a transfection marker (EBFP), and the ABA-inducible split transcription factor (TF_{ABA} 1 and 2). The transfection marker is used to adjust flow cytometry data for transfection efficiency. $TF_{ABA}1$ and $TF_{ABA}2$ dimerize in the presence of ABA and translocate to the nucleus where they activate P_{PhlF}. When testing post-PAS gRNAs only, dCas9-VPR is co-transfected.

4.3 miRNAs can be expressed downstream of a polyA signal

In humans, most miRNA sequences are located within intronic regions where they are transcribed as the primary transcript of human miRNAs (pri-miRNAs) by RNA pol II.[154] Following transcription, the pri-miRNA is cleaved by the Microprocessor complex (consisting of Drosha and DiGeorge syndrome chromosomal region 8 [DGCR8]) into a short precursor miRNA (pre-miRNA) with a hairpin structure.[154] This pre-miRNA is then transported into the cytoplasm where it undergoes another round of processing by Dicer to a 21-26 nucleotide miRNA that is composed of an antisense and sense strands. Either the sense or the antisense strand is then incorporated into the RNA-induced silencing complex (RISC). The mature-miRNA:RISC complex is capable of inducing RNA degradation or translational inhibition.[154]

Two different methods have been established for expression of small interfering RNAs, such as miRNAs, in mammalian cells. One relies on expression of short hairpin RNA which mimics the stem-loop structure of pre-miRNAs, and are expressed from RNA pol III promoters.[156] The second approach relies on placing a synthetic miRNA into the backbone of an endogenous miRNA.[157,158] By flanking the miRNA with splice site, the miRNA can be expressed as part of an RNA Pol II transcribed gene. While this does not alter the sequence of the mRNA, splicing is inherently connected to transcription, and introduction of splice sites can have unintended effects on gene expression.[159] Thus, current methods for miRNA expression might not be viable if coupled to endogenous genes whose regulation must be carefully preserved. By placing the miRNA downstream of the terminator, we hypothesize any effect the miRNA has on the upstream gene can be minimized while miRNA expression remains conditional on upstream gene expression.

To rapidly test expression of miRNAs downstream of a terminator, we used the plasmid-based system outlined in Figure 4-1. The test construct consists of EYFP expressed from an abscisic acid (ABA) inducible promoter (P_{PhlF}), a terminator (including a large upstream and downstream region; Table 4-1), and a single synthetic miRNA flanked by splice sites (referred to as the miRNA cassette) placed downstream of the PAS. The post-PAS

miRNA construct being tested is co-transfected with a reporter plasmid that constitutively expresses mKate and contains 3' binding sites for the post-PAS miRNA which targets the mRNA for degradation, two plasmids encoding a split transcription factor that dimerizes in the presence of ABA (the transcription factor has been modified to use VPR instead of VP16, courtesy of Dr. Allen Tseng),[160] and a transfection marker with constitutive EBFP expression used to adjust for transfection efficiency.

We tested if the post-PAS miRNA could be expressed conditioned on upstream gene expression. We chose to test five different terminators and varied the distance between the post-PAS miRNA and the PAS (Table 4-1). These five different terminators belong to genes that are specific to either endoderm or mesoderm lineages during development, and might be relevant for directed differentiation of early cell lineages during organoid development. Since the region both upstream and downstream of the PAS might affect termination, each terminator spans approximately 3200 bp starting 200 bp upstream of the PAS. miRNA expression was detected by repression of mKate which is constitutively expressed from the reporter plasmid. Figure 4-3 and Figure S4-1 shows an approximately 20-fold reduction in mKate expression correlated with the increased transcription of the upstream gene. While repression largely depends on the concentration of ABA, weak mKate repression occurs even in the uninduced state. This indicates the miRNA might drive its own expression, as this baseline repression is observed across multiple terminators and different positions within each one.

Distance appears to have a small effect on mKate repression, which decreases as the distance to the PAS increases. However, the dynamic range of post-PAS miRNA repression remains relatively constant independent of the terminator and the distance (Figure S4-1). Considering that the RNA Pol II was found to accumulate approximately 1 kb downstream of the PAS,[148] it is likely that transcription continues unhindered in this 1 kb window, and the distances tested are not large enough to cover the downstream region where a significant proportion of RNA Pol II transcription complexes completely abolish transcription and dissociates from the DNA.

Table 4-1: Terminators tested for post-PAS RNA expression and the position of the post-PAS RNA cassette.

Terminator	Genomic location	Distances tested (bp from 3'-end of PAS)
Hesx1	**3**:57,226,068-57,229,463	124
		204
		321
		528
		1002
Meox1	**17**:43,640,639-43,637,307	52
		106
		215
		520
		994
Pyy	**17**:43,949,693-43,952,944	75
		200
		328
		507
		1362
Sox17	**8**:54,460,646-54,463,905	104
		202
		301
		500
		1065
TBXT (T)	**6**:166,154,618-166,157,769	62
		501
		1034

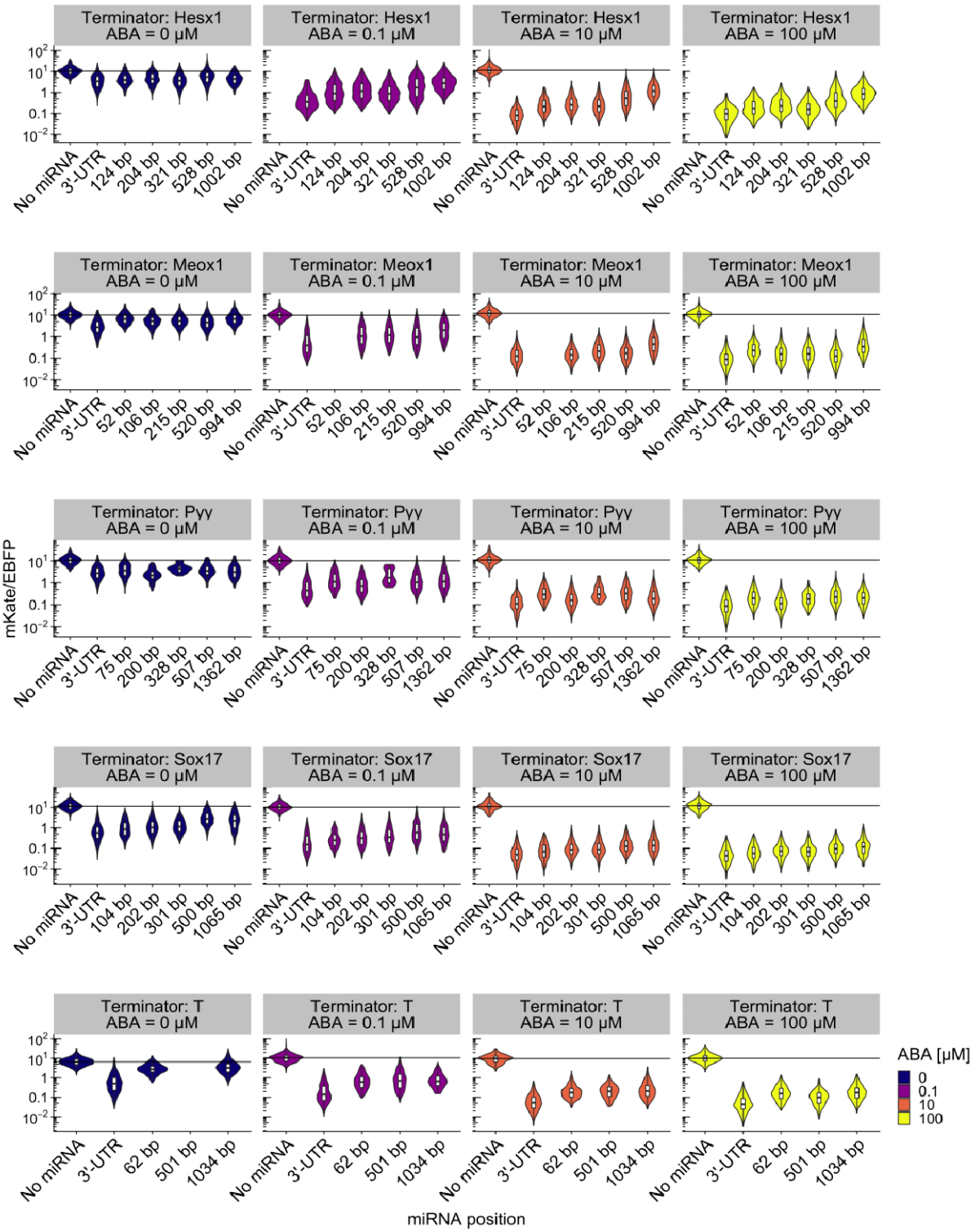


Figure 4-3: Post-PAS miRNA repression for different terminators. Different constructs testing either a post-PAS miRNA, a miRNA in the 3'-UTR, or no miRNA. miRNA expression was detected by repression of constitutively expressed mKate. The horizontal line indicates the median for the construct without a miRNA at the specified concentration of ABA. mKate repression is observed for all constructs but differs between terminators and position relative to the PAS. Note that repression is observed in the absence of ABA

indicating cryptic expression of the miRNA. mKate was normalized to the transfection marker, EBFP. Missing data indicates the sample was lost during preparation for flow cytometry.

4.4 The effect on upstream gene expression can be minimized by removing a splice site

We hypothesized that placing a miRNA downstream of the terminator would make EYFP expression independent of post-PAS miRNA expression. Calculating the ratio between EYFP expression from constructs with a miRNA to the EYFP expression of constructs without one, we observe that post-PAS miRNAs have an effect on upstream gene expression (Figure 4-4). The effect of a miRNA is dependent on the terminator and the position relative to the PAS. With the exception of the Hesx1 and Meox1 terminators, placing the miRNA in the 3'-UTR had a substantial effect on upstream gene expression. This effect was reduced as the miRNA was moved further downstream of the terminator. While most miRNAs had a negative effect on upstream gene expression, the Pyy terminator showed increased EYFP expression. However, it is worth noting that the Pyy construct without a miRNA had relatively low absolute EYFP expression (Figure S4-2).

Splicing is tightly coupled to transcriptional elongation and termination,[159,161,162] and although the miRNA is located downstream of the PAS, the presence of an intronic miRNA appears to affect upstream gene expression. Considering that miRNAs can be located within exons[163,164] and that processing is thought to occur co-transcriptionally and before splicing,[165] we hypothesized that the splice sites could be removed and ensure the upstream gene expression is unaffected by the post-PAS miRNA. We therefore chose to test an exonic miRNA strategy for post-PAS miRNA expression defined by the absence of splice sites. Additionally, we tested a second, minimal miRNA that consists of only a single hairpin structure[166] (referred to as the "Short" miRNA; Figure 4-5), to test if the altered upstream gene expression and the background miRNA expression were a product

of the structure of the miRNA itself. To avoid any confounding factors that might arise when expressing the genes from circular DNA, we linearized the post-PAS constructs by PCR and transfected the purified PCR products.

Removing the splice sites restore EYFP expression to levels similar to those when no miRNA is present indicating that splicing might have an effect on upstream gene expression when placed within the first 300 bp of the Sox17 PAS (Figure 4-6). Interestingly, we see a small decrease in EYFP expression when the post-PAS miRNA is placed 104 bp from the PAS. This indicates that placing the post-PAS miRNA too close to the PAS might interfere with termination in some manner.

Unexpectedly, placing an exonic miRNA in the 3'-UTR did not reduce EYFP levels for either type of miRNA. This could be explained if removing the splice sites yielded a miRNA that would not be processed, thereby leaving the mRNA unaltered. Contrary to our expectations,[167] we observe that the "Long" miRNA is processed and capable of mKate repression independent of location whereas the "Short" miRNA is only processed when placed in the 3'-UTR (Figure 4-7). This indicates that the miRNA processing machinery is capable of recognizing and processing miRNAs after termination has been initiated but that a minimal hairpin structure is insufficient. While RNA can mimic the function of a polyA tail,[168] this is unlikely to be the case. Alternatively, it might be possible that the close proximity of the miRNA to the cleavage site changes the location of polyadenylation. This remains an unlikely scenario however, given that ribozyme cleavage in the 3'-UTR required a synthetic polyA tail to rescue the mRNA.[169]

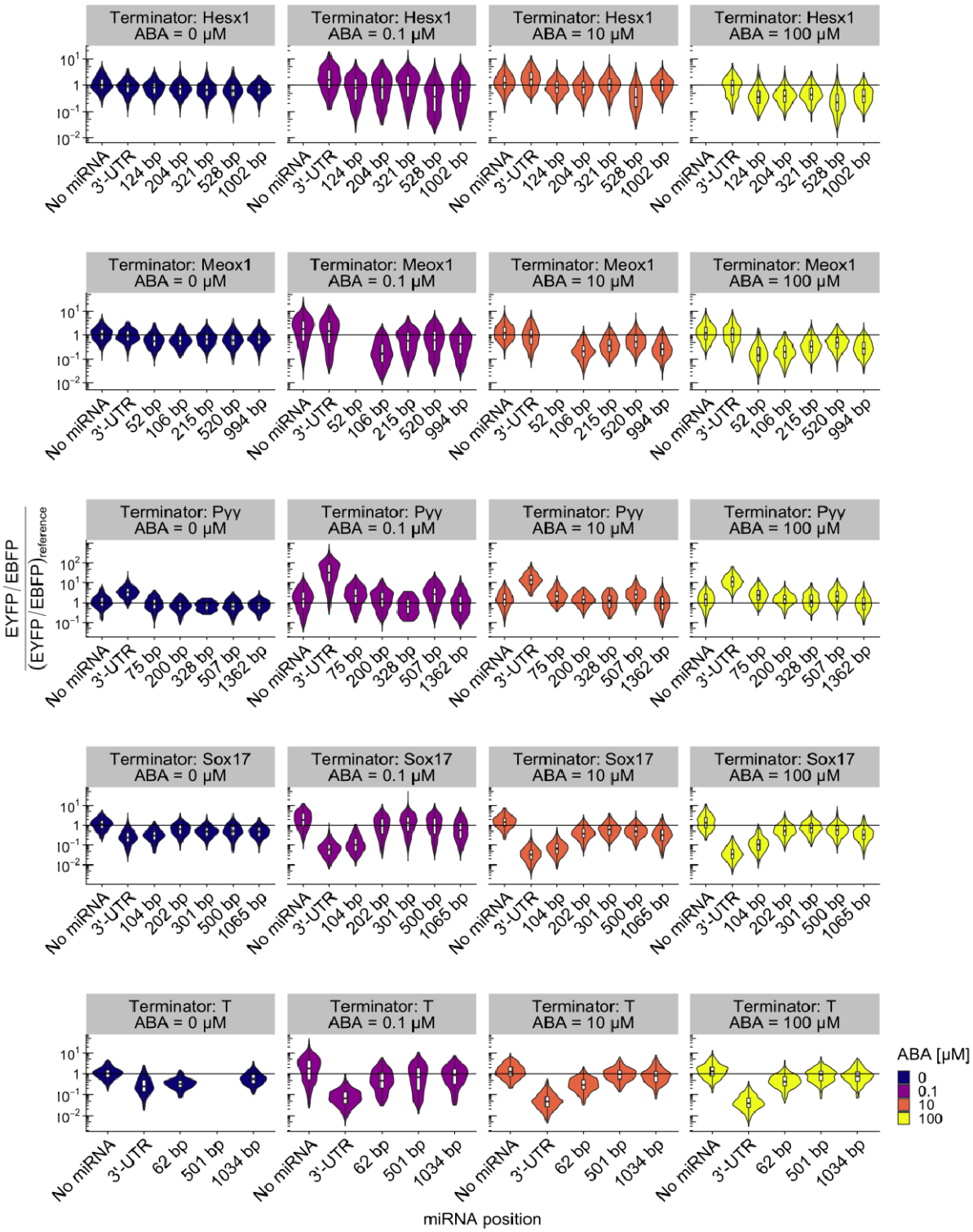


Figure 4-4: Post-PAS miRNAs might affect EYFP expression. Expression of EYFP for each construct was normalized to the geometric mean of EYFP expression of the construct "No miRNA" for each concentration of ABA. The plots shows that EYFP expression might be affected by the expression of a miRNA flanked by splice sites dependent on the terminator and position of the miRNA relative to the PAS. A

horizontal line at y=1 indicates no difference in EYFP expression relative to the “No miRNA” construct. Missing data indicates the sample was lost during preparation for flow cytometry.

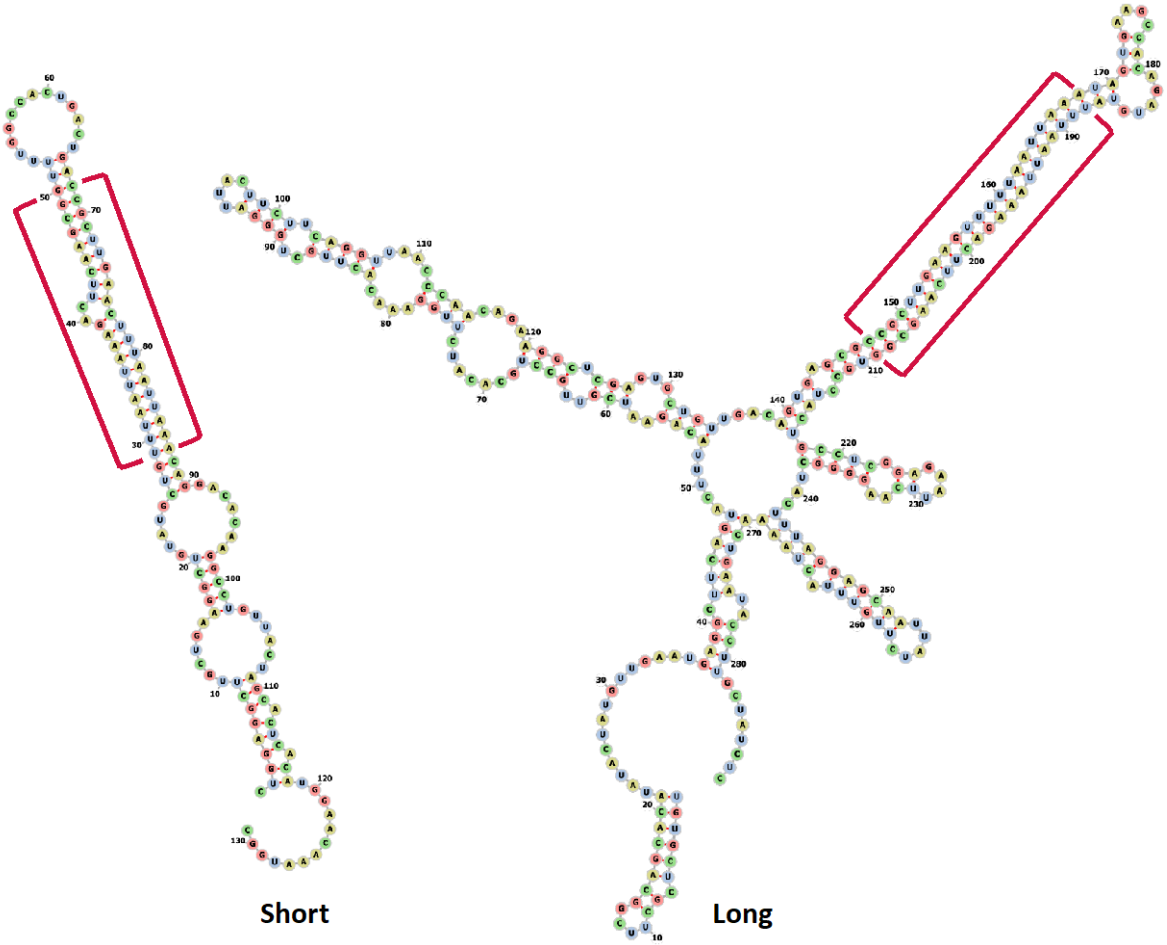


Figure 4-5: miRNA structures. Vienna RNAfold[170] was used to predict the structure of the "Short" minimal miRNA (left) and the "Long" miRNA (right). The miRNAs have identical sense and antisense strands (indicated by brackets).

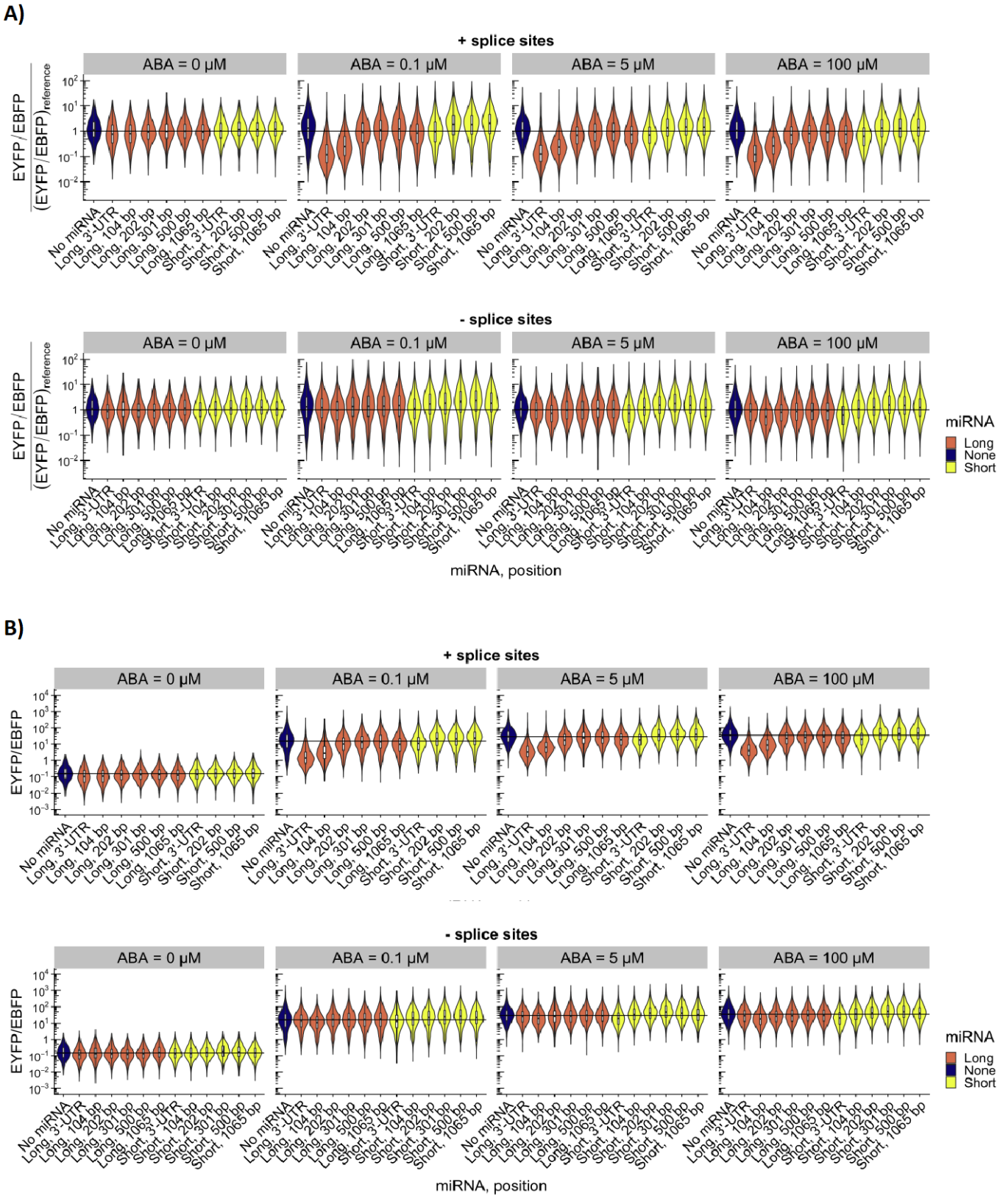


Figure 4-6: Removing splice sites from the miRNAs restore EYFP expression. The annotated splice site from both the "Long" and "Short" miRNAs were removed, constructs were linearized by PCR,

transfected, and EYFP expression was quantified and normalized to the EBFP expression from the transfection marker. **(A)** In the absence of the annotated splice sites, EYFP expression is restored as measured by ratio between EYFP expression of the “No miRNA” construct relative to constructs containing a miRNA. **(B)** EYFP/EBFP values for each construct shows that activation takes place and that absolute levels of EYFP/EBFP remains unchanged.

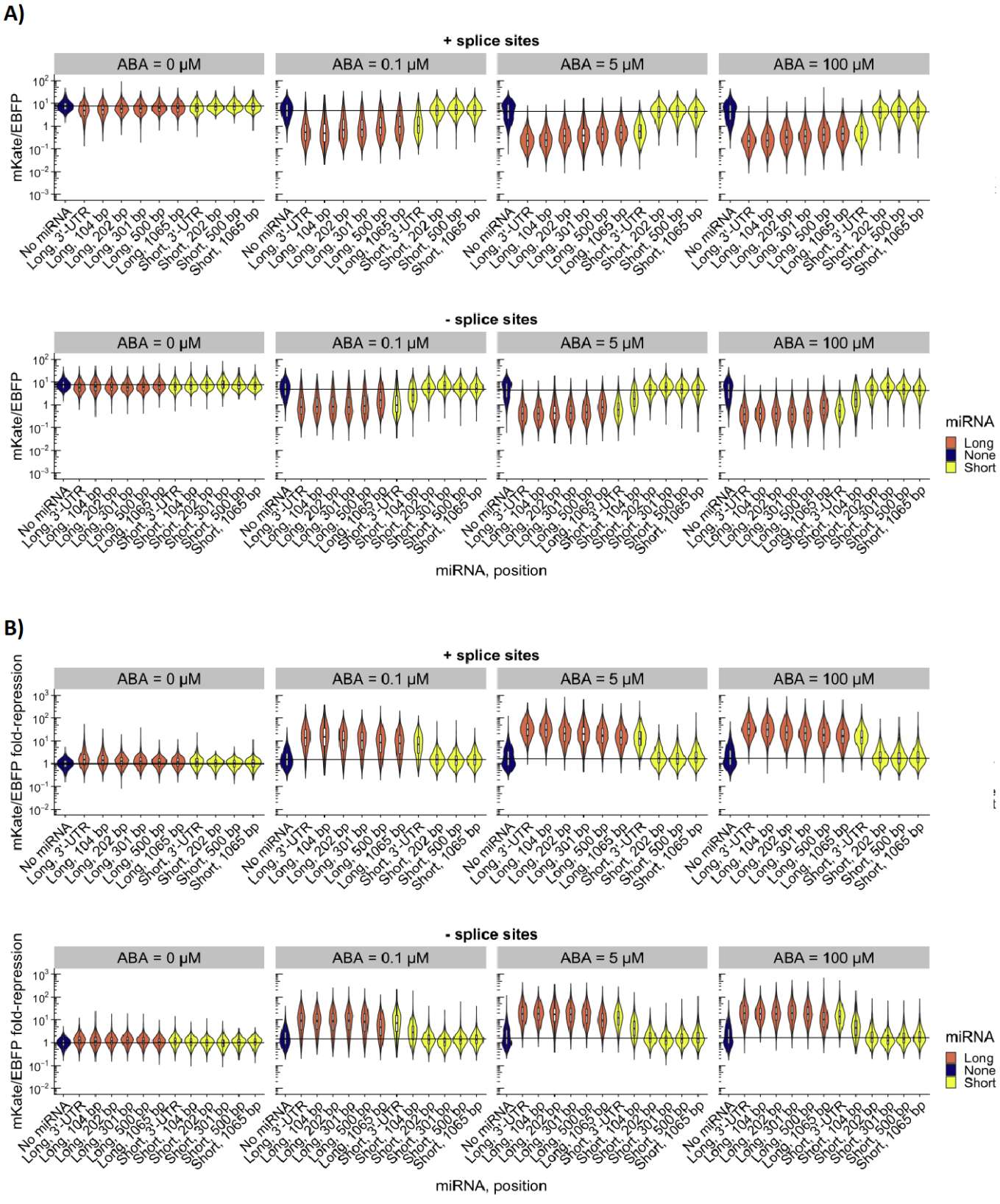


Figure 4-7: Splice sites might not be necessary for miRNA expression. The annotated splice site from both the "Long" and "Short" miRNAs were removed, constructs were linearized by PCR, transfected, and

mKate expression was quantified and normalized to EBFP expression from the transfection marker. **(A)** In the absence of the annotated splice sites, mKate is repressed by the "Long" miRNA, but not the "Short" miRNA unless placed in the 3'-UTR. This indicates that a minimal miRNA structure is insufficient for processing in the absence of splice sites **(B)** Quantification of the fold-repression relative to the "No miRNA" construct.

4.4 gRNAs can be expressed downstream of a polyA signal

Having shown that miRNAs can be expressed downstream of a terminator, we sought to test if post-PAS RNA expression can be expanded to other RNAs with a potential for gene regulation. gRNAs are short RNAs that can be used to direct the bacterial Cas9 protein to almost any DNA sequence of interest.[171] By fusing a catalytically inactive variant of the Cas9 protein, dCas9, to mammalian activators, the Cas9 protein can be directed and used to activate, rather than cut, almost any gene of interest.[52]

To test if gRNAs can be expressed downstream of a terminator, we adopted the method described by Nissim *et al.*[49] in which ribozymes are used to excise the gRNA. Briefly, the gRNA cassette consists of three repeats of a gRNA (gRNA1) flanked by an upstream hammerhead (HHR) and a downstream HDV (HDV) ribozyme with short (4 bp) spacers between each HHR-gRNA1-HDV repeat (Figure 4-2). Similar to the post-PAS miRNA constructs, EYFP was expressed from P_{PhlF}, and the gRNA cassette was placed downstream of the Sox17 PAS. To detect the gRNA, we co-transfected HEK293 cells with dCas9-VPR and an mKate reporter (Figure 4-2). If the gRNA is expressed, it can bind to dCas9-VPR and activate the minimal promoter (P_{gRNA1}) upstream of mKate that has been designed with binding sites for the dCas9-VPR:gRNA1-complex.

mKate expression increases as a function of ABA (Figure 4-8) and has a dynamic range of ≈10-100-fold activation relative to the uninduced state for each construct. The dynamic range of mKate activation depends significantly on the distance between the PAS and the post-PAS gRNA, likely due to saturation of mKate activation. In the uninduced state,

mKate remains ≈10-100 fold higher than the control where no gRNA is present. Together with the observation that EYFP expression is OFF in the absence of ABA, this points to the presence of a cryptic promoter significantly driving gRNA1 expression and thus mKate activation. Despite the presence of a cryptic promoter, the 10-100-fold activation of mKate in the presence of full ABA induction points to the post-PAS gRNAs being expressed downstream of a terminator.

Since cryptic expression occurred at every position tested within the Sox17 terminator, we hypothesized the cryptic promoter was located within the post-PAS gRNA cassette itself, most likely from one of the ribozymes. To test if any of the ribozymes contained a cryptic promoter, we constructed a range of plasmids (Figure 4-9A) replacing one or both of the ribozymes with Csy4 recognition sites. These 28 bp sites are recognized by the endonuclease Csy4 and cleaved, thereby releasing the gRNA.[49] Replacing the HDV ribozyme with a Csy4 recognition site bring mKate expression in the uninduced state close to the mKate levels observed for the control that lacks a gRNA1, indicating that the HDV ribozyme is the primary driver of cryptic gRNA expression. By removing both ribozymes entirely and replacing them with Csy4 recognition sites, mKate expression can be brought close to the mKate expression observed for the control. However, we note that the dynamic range decreases, and that the combination of a 5’ HHR and a 3’ Csy4 recognition site appears optimal if a higher dynamic range is more important than low background expression.

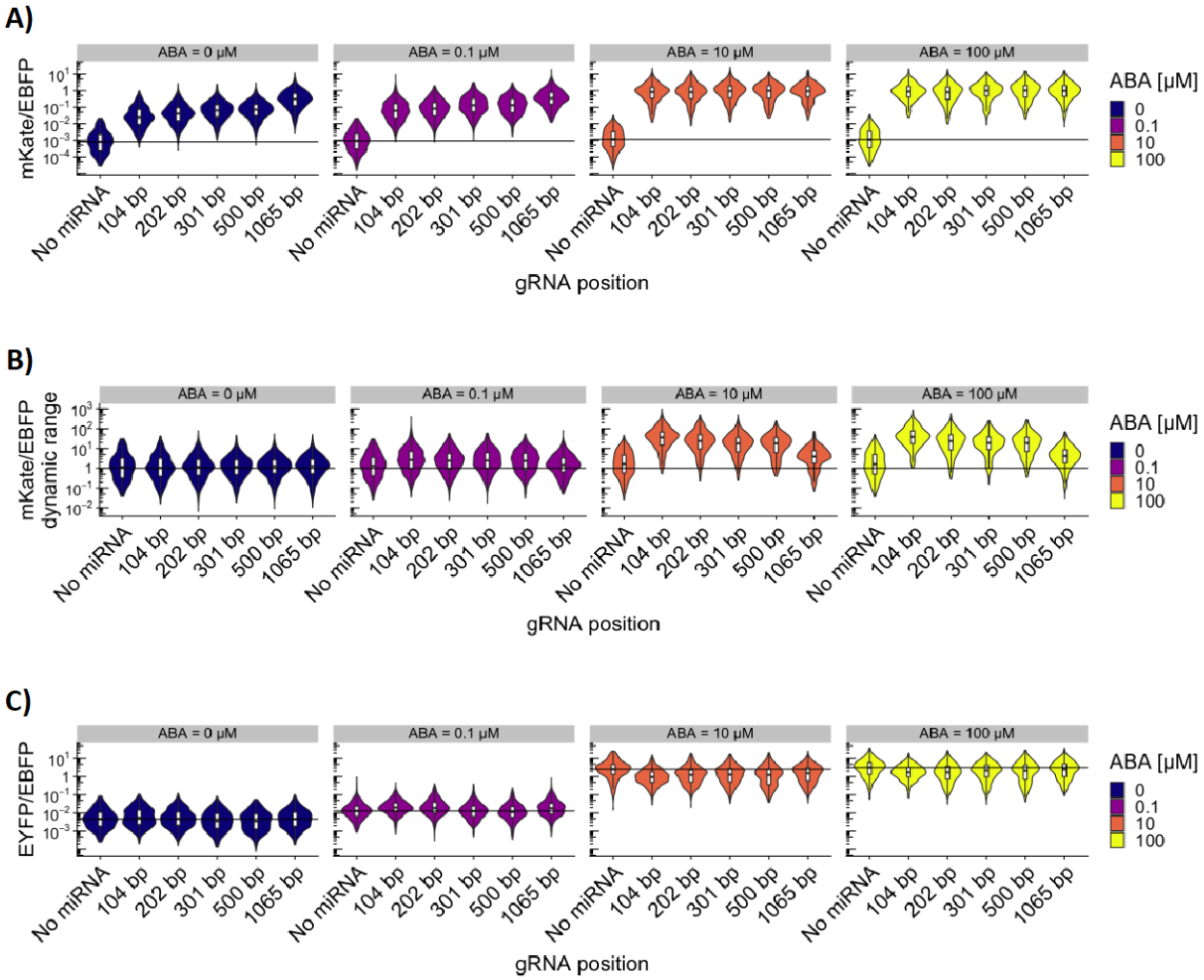


Figure 4-8: gRNAs can be expressed downstream of a terminator. (A) mKate normalized to the EBFP transfection marker shows that the gRNA is expressed but with a high degree of cryptic expression indicated by the significant mKate activation in the absence of the inducer, ABA **(B)** The dynamic range of mKate activation (the change in mKate activation for each construct relative to itself at ABA = 0) indicates the gRNA can be expressed downstream of a terminator, as expression changes as a function of the inducer, ABA **(C)** EYFP normalized to mKate shows the activation and change in upstream gene expression.

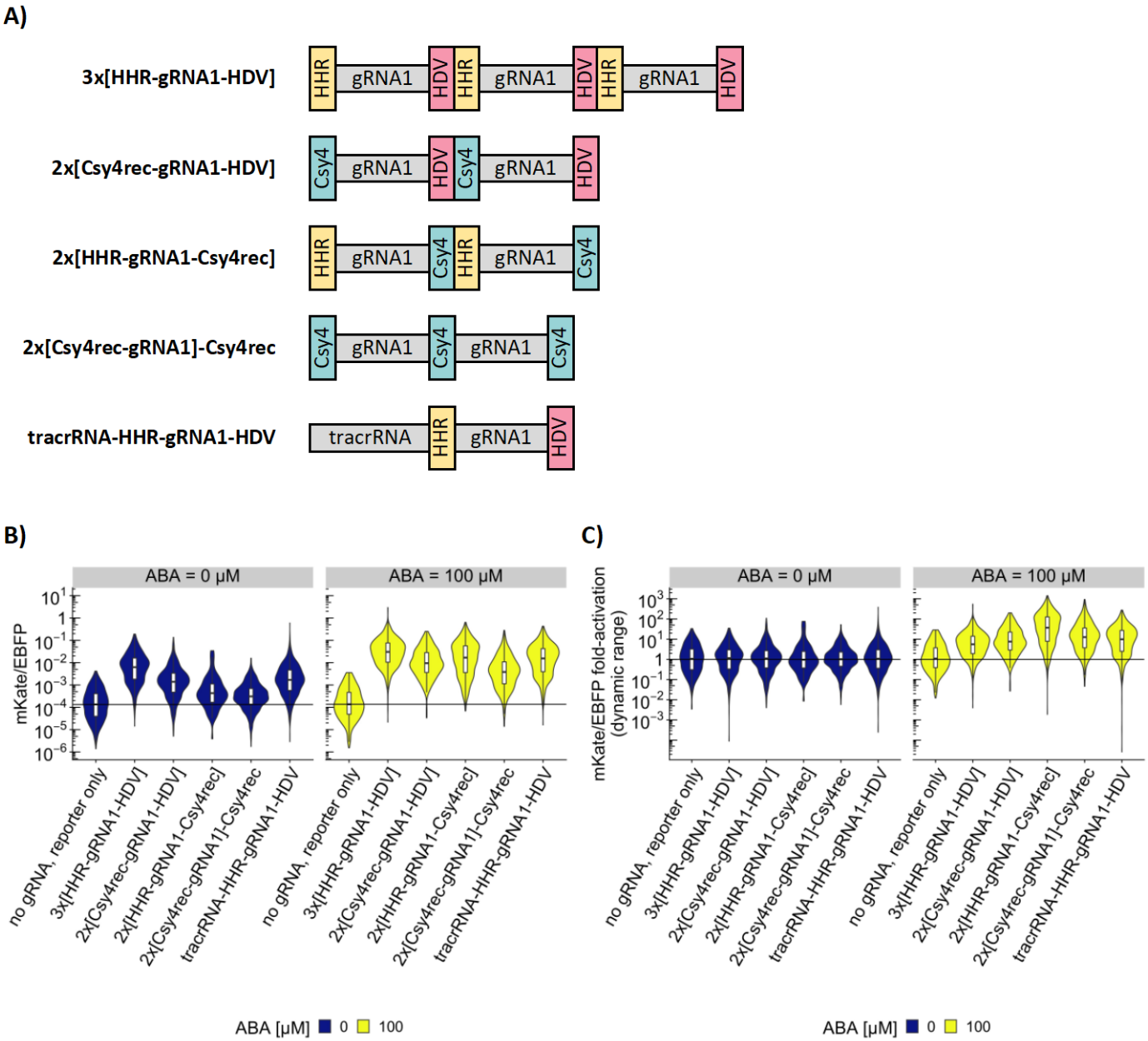


Figure 4-9: Testing cryptic gRNA expression. (A) Constructs designed to test cryptic gRNA expression. **(B)** gRNA expression was measured by mKate activation. By replacing the ribozymes with Csy4 recognition sites, cryptic gRNA expression can be reduced. **(C)** The dynamic range of mKate activation for each construct normalized to the expression of EBFP from the transfection marker. Csy4, Csy4 binding sites; HHR, hammerhead ribozyme; HDV, HDV ribozyme; tracrRNA, transactivating CRISPR RNA, the structural component of a gRNA that excludes the guiding sequence.

4.5 Discussion

Using a plasmid-based system, we show that miRNAs and gRNAs can be expressed downstream of a terminator within the first 1 kb without affecting upstream gene expression. This has the potential to enable us to co-express miRNAs and gRNAs with any endogenous gene without any modifications to the gene itself.

Both the 5'- and 3'-UTR play a significant role in gene regulation[172,173] and changes to these regions have the potential for disruptions to endogenous gene expression. Embryonic tissue tends to express shorter 3'-UTRs than differentiated tissue, and even across differentiated tissue, different cell-types are found to express different 3'-UTRs, indicating fine-tuning of gene regulation to the particular cell-type and development stage.[172] By moving the miRNA or gRNA downstream of the PAS, we reduce the chance of any unintended effect on upstream gene expression, while completely preserving the endogenous gene. Moreover, using a post-PAS RNA located downstream of the endogenous gene removes the need for editing the gene itself. This might be particularly important if engineering cells for therapeutic purposes.

In contrast to other experiments,[167] having an exonic miRNA located on the mRNA did not impact EYFP expression as a proxy for EYFP mRNA stability. The most likely explanation would be the presence of cryptic splice sites, although no pair of splice site acceptor and donor could confidently be predicted[174] and the previously observed effect of the splice sites on EYFP expression is absent. While it is possible the miRNA is not processed, the miRNA would in turn also be inactive[163] and unable to repress mKate in contrast to our observation. Further studies with genomic integration of post-PAS miRNA and examination of EYFP transcripts are necessary to understand our observation that a miRNA is present in the 3'-UTR without affecting upstream gene expression.

Our observation that exonic miRNAs can be expressed without splice sites and downstream of a gene has potential biological implications. It suggests that endogenous miRNAs could be expressed downstream of endogenous genes without their own promoter as transcription naturally continues for up to several thousand bases downstream of the terminator.[148] Thus, there could be either positive or negative selection towards miRNAs located in the proximity of terminators depending on the selective advantage of having them co-expressed with the upstream gene.

Finally, we demonstrate that gRNAs can be expressed downstream of a terminator, but that our current design suffers from high cryptic expression. We showed that endonucleases might be a viable strategy to replace the self-cleaving ribozymes as the means to excise the gRNA. Further studies led by Dr. Fabio Callendo and Dr. Elvira Vitu are ongoing to test and optimize this strategy.

Taken together, we show that small non-coding RNAs can be expressed downstream of a terminator. This offers a powerful new method to couple synthetic GRNs to endogenous GRNs by coupling an actuator in the form of a small regulatory RNA to the cell state as defined by the expression of one or more endogenous genes.

4.6 Methods

4.6.1 Plasmid construction

Plasmids were constructed using the hierarchical MoClo system.[95] Endogenous terminators were amplified from human genomic DNA by PCR and inserted into L0.T entry vectors. Post-PAS RNA was inserted into the terminators by PCR amplifying the different parts and using Golden Gate cloning to insert them. Expression vectors were assembled from entry vectors with the P_{PhlF} promoter, EYFP coding sequence, and the terminator (either a terminator with the post-PAS RNA to be tested or a LacZ flanked by

SbfI and AscI restriction sites). In cases where the terminator and post-PAS RNA to be tested contained BsaI restriction sites, the expression vectors with LacZ and the terminator were separately digested with SbfI and AscI, gel purified, then mixed and ligated to create the final plasmid with the terminator inserted.

4.6.2 Cell culture and transfections

HEK293 were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS. HEK293 cells were transfected with ViaFect (Promega) according to manufacturer's instructions. Each reaction was done with 20,000 cells/well in a 96-well plate. Each transfection was done with 100 ng total DNA/well with an equal mass of each plasmid or linearized PCR product. Cells were immediately induced with ABA and incubated for 48 hours after which cells were prepared for flow cytometry

4.6.3 Flow cytometry

Cells were washed with calcium- and magnesium-free DPBS, trypsinized, and resuspended in DMEM with 10% FBS. Suspended cells were centrifuged, and the supernatant was removed. Cells were prepared for flow cytometry by a final resuspension in calcium- and magnesium-free DPBS with 10% FBS. Cells were analyzed on a BD LSRII flow cytometer.

4.6.4 Data analysis

Flow cytometry data was adjusted for autofluorescence and spectral bleed through using the Cytoflow (v1.0) Python package. Adjusted values were exported to R and analyzed.

mKate and EYFP fluorescence was normalized to the EBFP transfection marker. Fold-change was calculated as the ratio between the sample relative to the geometric mean of the uninduced plasmid using the same terminator but lacking a post-PAS RNA insertion. Dynamic range was calculated as the ratio between the induced state and the geometric mean of the uninduced state for each construct. EYFP relative to baseline was calculated as the difference between the measured sample and the geometric mean of the construct without a post-PAS RNA insert.

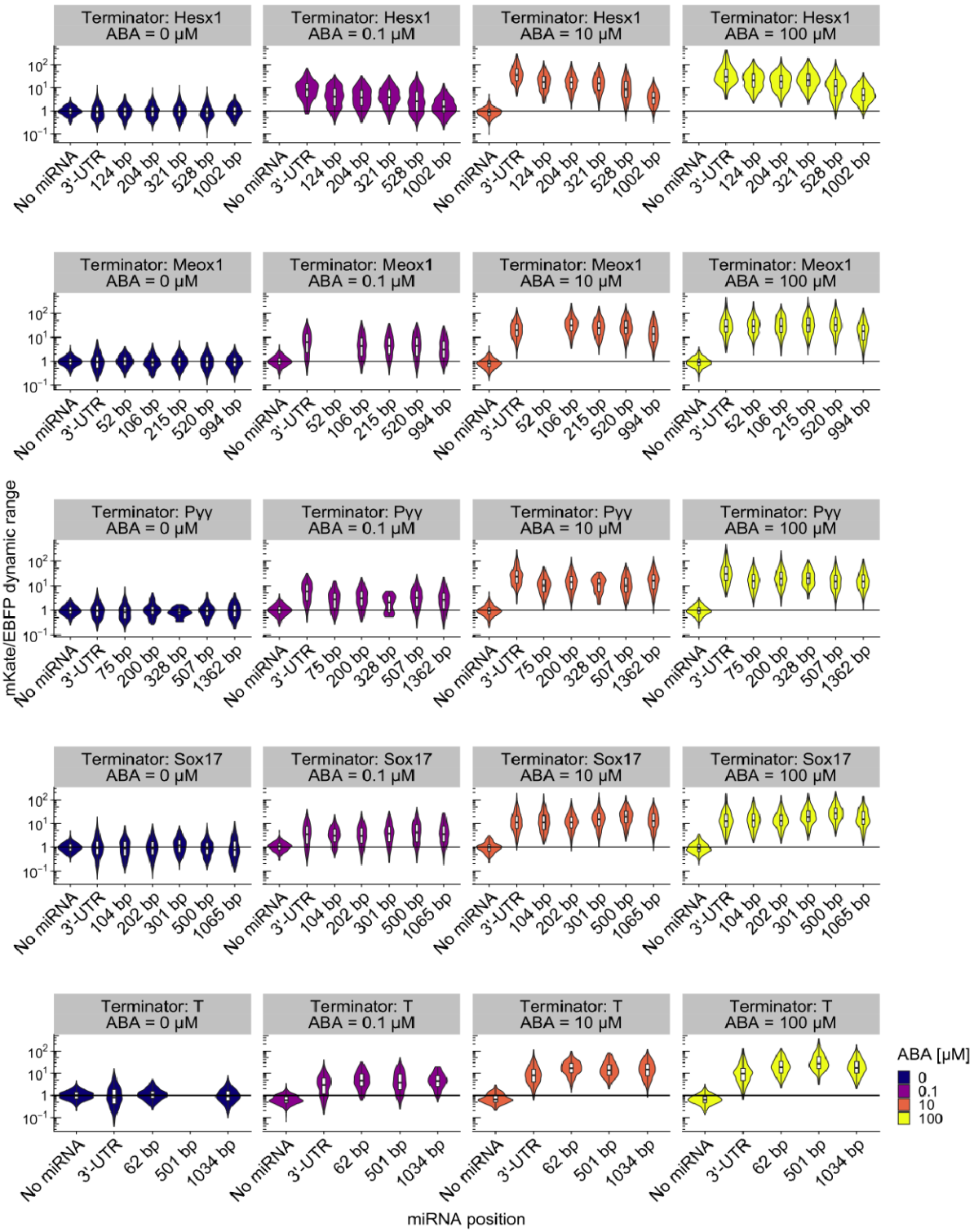


Figure S4-1: Dynamic range of post-PAS miRNA repression for different terminators. Different constructs testing either a post-PAS miRNA, a miRNA in the 3'-UTR or no miRNA were tested and miRNA expression was detected by repression of mKate which was constitutively expressed from a reporter plasmid. The dynamic range of repression was measured as the ratio

between the uninduced and induced state for each construct. The horizontal line at y=1 indicates no repression. Missing data indicates the sample was lost during preparation for flow cytometry.

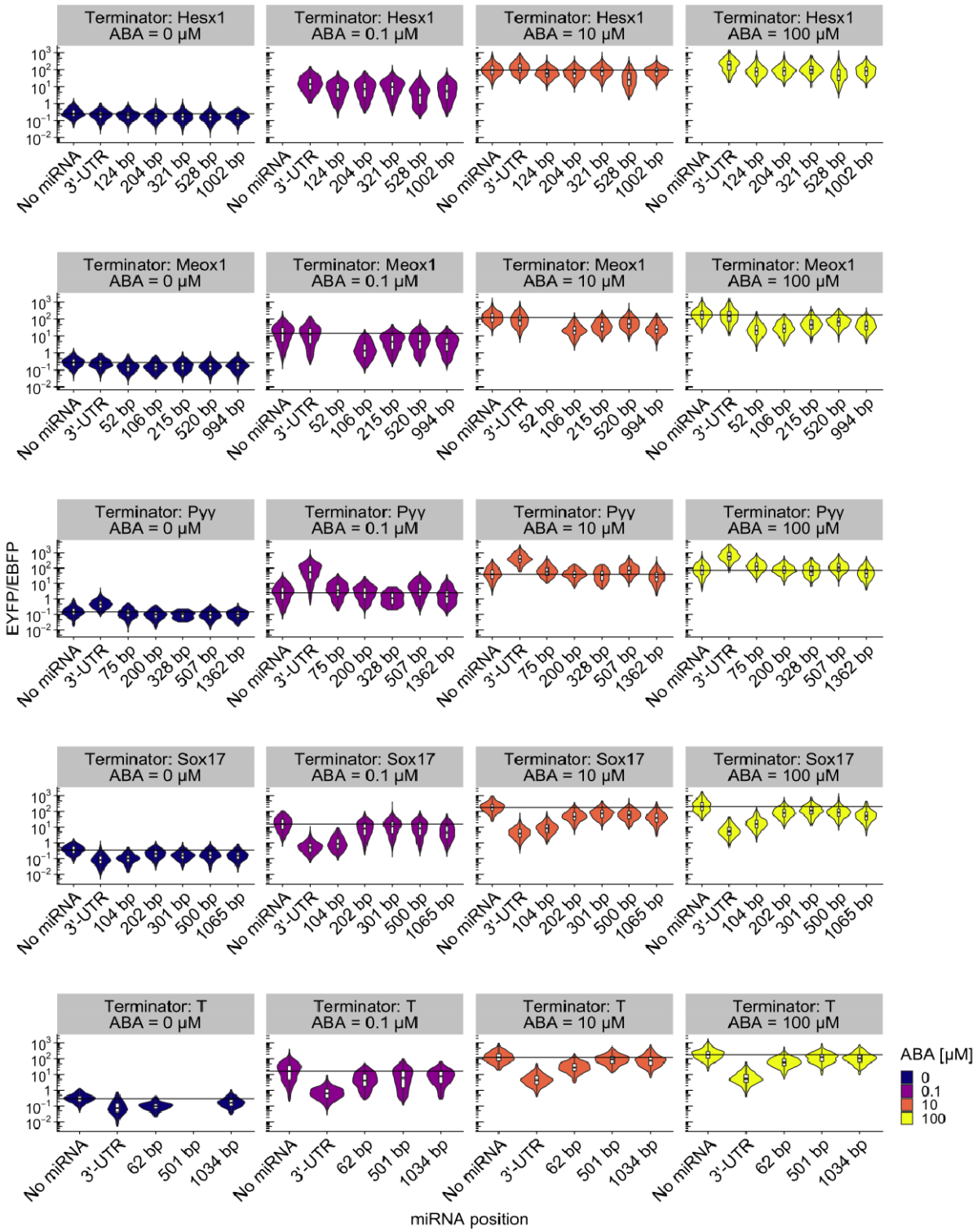


Figure S4-2: EYFP/EBFP expression for post-PAS miRNAs. Expression of EYFP for each construct was normalized to EBFP. The horizontal line indicates the median value of the reference construct, “No miRNA”. Missing data indicates the sample was lost during preparation for flow cytometry.

Chapter 5. Future Directions

Here, we have developed tools for (i) inducing a synthetic cell state to control morphogenesis (Chapter 2), (ii) detecting a complex cell state as defined by a cell's gene expression profile through synthetic promoters with enhanced cell-state specificity (SPECS; Chapter 3), and (iii) sensing gene expression through a post-PAS gene sensor (Chapter 4). Combined, these tools enable us to engineer complex multicellular systems such as organoids by controlling when and where synthetic gene regulatory networks (GRNs) are active, and the morphogenesis of larger multicellular assemblies.

The use of gene sensors and SPECS enable us to detect distinct cell lineages in a heterogenous cell population. Our gene sensor was shown to work for a range of terminators that are derived from genes specific to the mesoderm (Meox1 and T) and endoderm, specifically the developing foregut (Pyy), anterior endoderm (Hesx1), and endoderm more broadly (Sox17).[175,176] Applying the gene sensor to detect these cell lineages enables directed differentiation or controlled morphogenesis of them. For instance, ectopic GATA6 expression in induced pluripotent stem cells led to the formation of all three germ layers and subsequent development of a liver bud-like structure.[27] The liver and pancreas both develop from the foregut endoderm, and as such, it might be possible to direct the differentiation of the liver bud-like organoid towards a pancreatic fate. We hypothesize that inhibition of pro-hepatic markers such as HNF4α[27] by post-PAS miRNA expression, and upregulation of non-canonical Wnt signaling such as Wnt5a[177] by post-PAS gRNA expression can be used to direct endoderm cells towards a pancreatic fate. Following differentiation of cells to pancreatic progenitors, expression of Pdx1, Ngn3, or MafA might be further used to drive β-cell development and maturation.[31]

Lack of vasculature is another common issue of organoids,[26] but might be addressed by ectopic ETV2 expression.[29] The original method did not take cell state into account, thereby depending on transdifferentiation of some cells which could lead to poor or failed differentiation of some cells. A post-PAS gene sensor or SPECS to detect endothelial

progenitors might be used for directed differentiation of endothelial progenitors, possibly resulting in more robust differentiation that can be applied to multiple different organoids.

Organoids are organ models that partially capture development and leads to multicellular assemblies composed of cell lineages and architecture that ideally resemble the lineages and structural organization observed in human organs. While gastruloids are developed to mimic the earliest stages of development,[178] organoids might skip several of steps of early human development and instead model the development of specific tissues such as the liver, colon, or brain.[23,27,179] Our method to induce cell sorting, either as a function of adding inducer, or by coupling recombinase expression to a specific cell state or gene through the SPECS or post-PAS gene sensor, respectively, offer an opportunity to engineer the early morphology of the developing organoid. For instance, in the previously mentioned liver bud-like organoid that expresses all three germ layers,[27] sorting of these germ layers might change the development of the organoid. We hypothesize that using differential cadherin expression might be used to sort different germ layers and cell lineages such as the different parts of the foregut. By sorting cell types, gradients of secreted molecules might be better controlled and the effect different cell populations have on differentiation can be better understood. For instance, we hypothesize that the number of mesoderm cells, the mesoderm-to-endoderm ratio, and the distance between mesoderm and endoderm cells can influence how any individual cell develops as signaling between germ layers is critical for proper development.[150] By coupling recombinase expression to a gene sensor, this can be used to control ectopic expression of cadherins and modulation of cell sorting of different populations as defined through gene expression.

Future work combining directed differentiation and engineered cell sorting thus offer exciting opportunities to engineer organoids, with the aim of increasing the reproducibility and quality of developing organoids.

Bibliography

1. Ro, D.-K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940–943 (2006).

2. Dietrich, J. a *et al.* A novel semi-biosynthetic route for artemisinin production using engineered substrate-promiscuous P450(BM3). *ACS Chem. Biol.* **4**, 261–7 (2009).

3. Sadelain, M. Chimeric antigen receptors: Driving immunology towards synthetic biology. *Current Opinion in Immunology* **41**, 68–76 (2016).

4. Sadelain, M., Rivière, I. & Riddell, S. Therapeutic T cell engineering. *Nature* **545**, 423–431 (2017).

5. Broutier, L. *et al.* Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. *Nat. Med.* **23**, 1424–1435 (2017).

6. Berkers, G. *et al.* Rectal Organoids Enable Personalized Treatment of Cystic Fibrosis. *Cell Rep.* **26**, 1701-1708.e3 (2019).

7. Polack, F. P. *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **383**, 2603–2615 (2020).

8. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).

9. Macchiarini, P. *et al.* Clinical transplantation of a tissue-engineered airway. *Lancet* **372**, 2023–2030 (2008).

10. Oswald, J. & Baranov, P. Regenerative medicine in the retina: from stem cells to cell replacement therapy. *Ther. Adv. Ophthalmol.* **10**, 251584141877443 (2018).

11. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).

12. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).

13. Kitada, T., DiAndreth, B., Teague, B. & Weiss, R. *Programming gene and engineered-cell therapies with synthetic biology*. *Science* **359**, (2018).

14. Gramelsberger, G. Synthetic Morphology: A Vision of Engineering Biological Form. *J. Hist. Biol.* **53**, 295–309 (2020).

15. Pollen, A. A. *et al.* Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**, 743-756.e17 (2019).

16. Quadrato, G., Brown, J. & Arlotta, P. The promises and challenges of human brain organoids as models of neuropsychiatric disease. *Nat. Med.* **22**, 1220–1228 (2016).

17. Choi, S. H. *et al.* A three-dimensional human neural cell culture model of Alzheimer's disease. *Nature* **515**, 274–278 (2014).

18. Gerakis, Y. & Hetz, C. Brain organoids: a next step for humanized Alzheimer's disease models? *Mol. Psychiatry* **24**, 474–478 (2019).

19. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290.e17 (2017).

20. Brennand, K. J. *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473**, 221–5 (2011).

21. Quadrato, G. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48–53 (2017).

22. Capowski, E. E. *et al.* Reproducibility and staging of 3D human retinal organoids across multiple pluripotent stem cell lines. *Development* **146**, dev171686 (2019).

23. Velasco, S. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523–527 (2019).

24. Jones, R. D. Genetic devices for robust, context-independent control of gene expression levels in mammalian cells. (2020).

25. Cheah, P.-S., Mason, J. O. & Ling, K. H. Challenges and future perspectives for 3D cerebral organoids as a model for complex brain disorders. *Neurosci. Res. Notes* **2**, 1–6 (2019).

26. Brassard, J. A. & Lutolf, M. P. Engineering Stem Cell Self-organization to Build Better Organoids. *Cell Stem Cell* **24**, 860–876 (2019).

27. Guye, P. *et al.* Genetically engineering self-organization of human pluripotent stem cells into a liver bud-like tissue using Gata6. *Nat. Commun.* **7**, 1–12 (2016).

28. Antonica, F. *et al.* Generation of functional thyroid from embryonic stem cells. *Nature* **491**, 66–71 (2012).

29. Cakir, B. *et al.* Engineering of human brain organoids with a functional vascular-like system. *Nat. Methods* 1–7 (2019). doi:10.1038/s41592-019-0586-5

30. Francesconi, M. *et al.* Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. *Elife* **8**, 1–22 (2019).

31. Saxena, P. *et al.* A programmable synthetic lineage-control network that differentiates human IPSCs into glucose-sensitive insulin-secreting beta-like cells. *Nat. Commun.* **7**, 11247 (2016).

32. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D. A. In vivo reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* **455**, 627–632 (2008).

33. Toda, S., Blauch, L. R., Tang, S. K. Y., Morsut, L. & Lim, W. A. Programming self-organizing multicellular structures with synthetic cell-cell signaling. *Science (80-.).* **361**, 156–162 (2018).

34. Tordoff, J. & Weiss, R. Engineering Self-Assembling Living Structures with Mammalian Synthetic Biology. (2020).

35. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Aspects Med.* **59**, 114–122 (2018).

36. Fiers, M. W. E. J. *et al.* Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* **17**, 246–254 (2018).

37. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).

38. Pijuan-Sala, B., Guibentif, C. & Göttgens, B. Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nat. Rev. Mol. Cell Biol.* **19**, 399–412 (2018).

39. Gouti, M. *et al.* A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development. *Dev. Cell* **41**, 243-261.e7 (2017).

40. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science (80-.).* eaar5780 (2018). doi:10.1126/science.aar5780

41. Mohammed, H. *et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).

42. Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).

43. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* **28**, 2679–2692 (2014).

44. Lis, R. *et al.* Conversion of adult endothelium to immunocompetent haematopoietic stem cells. *Nature* **545**, 439–445 (2017).

45. Panciera, T. *et al.* Induction of Expandable Tissue-Specific Stem/Progenitor Cells through Transient Expression of YAP/TAZ. *Cell Stem Cell* **19**, 725–737 (2016).

46. Sugimura, R. *et al.* Haematopoietic stem and progenitor cells from human pluripotent stem cells. *Nature* **545**, 432–438 (2017).

47. Du, Y. *et al.* Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* **14**, 394–403 (2014).

48. Yang, G. *et al.* Integration-deficient lentivectors: an effective strategy to purify and differentiate human embryonic stem cell-derived hepatic progenitors. *BMC Biol.* **11**, 86

(2013).

49. Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P. & Lu, T. K. Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells. *Mol. Cell* **54**, 698–710 (2014).

50. Kiani, S. *et al.* CRISPR transcriptional repression devices and layered circuits in mammalian cells. *Nat. Methods* **11**, 723–6 (2014).

51. Xie, K., Minkenberg, B. & Yang, Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3570–3575 (2015).

52. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* **12**, 326–328 (2015).

53. Yeo, N. C. *et al.* An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat. Methods* **15**, 611–616 (2018).

54. Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R. & Benenson, Y. Multi-Input RNAi-Based Logic Circuit for Identification of Specific Cancer Cells. *Science (80-.).* **333**, 1307–1311 (2011).

55. Donnelly, M. L. L. *et al.* Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: A putative ribosomal 'skip'. *J. Gen. Virol.* **82**, 1013–1025 (2001).

56. Tang, W. *et al.* Faithful Expression of Multiple Proteins via 2A-Peptide Self-Processing: A

Versatile and Reliable Method for Manipulating Brain Circuits. *J. Neurosci.* **29**, 8621–8629 (2009).

57. Grosberg, R. K. & Strathmann, R. R. The Evolution of Multicellularity: A Minor Major Transition? (2007). doi:10.1146/annurev.ecolsys.36.102403.114735

58. Lancaster, M. A. & Knoblich, J. A. Generation of cerebral organoids from human pluripotent stem cells. *Nat. Protoc.* **9**, 2329–2340 (2014).

59. Eiraku, M. *et al.* Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature* **472**, (2011).

60. Yin, X. *et al.* Engineering Stem Cell Organoids. *Cell Stem Cell* **18**, 25–38 (2016).

61. Matsuda, M., Koga, M., Woltjen, K., Nishida, E. & Ebisuya, M. Synthetic lateral inhibition governs cell-type bifurcation with robust ratios. *Nat. Commun.* **6**, 1–12 (2015).

62. Tordoff, J. *et al.* Incomplete Cell Sorting Creates Engineerable Structures with Long-Term Stability. *Cell Reports Phys. Sci.* **2**, 100305 (2021).

63. Wirth, D. *et al.* Road to precision: recombinase-based targeting technologies for genome engineering. *Curr. Opin. Biotechnol.* **18**, 411–419 (2007).

64. Turan, S., Zehe, C., Kuehle, J., Qiao, J. & Bode, J. Recombinase-mediated cassette exchange (RMCE) - A rapidly-expanding toolbox for targeted genomic modifications. *Gene* **515**, 1–27 (2013).

65. Brown, W. R. A., Lee, N. C. O., Xu, Z. & Smith, M. C. M. Serine recombinases as tools for genome engineering. *Methods* **53**, 372–379 (2011).

66. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based State machines in living cells. *Science (80-.).* **353**, (2016).

67. Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol.* **31**, 448–452 (2013).

68. Appleton, E. *et al.* Genetic design automation for autonomous formation of multicellular shapes from a single cell progenitor. *bioRxiv* (2019). doi:10.1101/807107

69. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).

70. Cai, D., Cohen, K. B., Luo, T., Lichtman, J. W. & Sanes, J. R. Improved tools for the Brainbow toolbox. *Nat. Methods* **10**, 540–547 (2013).

71. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).

72. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56–62 (2007).

73. Wang, S. Z., Liu, B. H., Tao, H. W., Xia, K. & Zhang, L. I. A genetic strategy for stochastic gene activation with regulated sparseness (STARS). *PLoS One* **4**, 1–6 (2009).

74. Movahedi, K., Wiegmann, R., De Vlaminck, K., Van Ginderachter, J. A. & Nikolaev, V. O. RoMo: An efficient strategy for functional mosaic analysis via stochastic Cre recombination and gene targeting in the ROSA26 locus. *Biotechnol. Bioeng.* **115**, 1778–1792 (2018).

75. Schötz, E. M. *et al.* Quantitative differences in tissue surface tension influence zebrafish

germ layer positioning. *HFSP J.* **2**, 42–56 (2008).

76. Manning, M. L., Foty, R. A., Steinberg, M. S. & Schoetz, E. M. Coaction of intercellular adhesion and cortical tension specifies tissue surface tension. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12517–12522 (2010).

77. Heisenberg, C. P. D'Arcy Thompson's 'on Growth and form': From soap bubbles to tissue self-organization. *Mechanisms of Development* **145**, 32–37 (2017).

78. Halbleib, J. M. & Nelson, W. J. Cadherins in development: Cell adhesion, sorting, and tissue morphogenesis. *Genes Dev.* **20**, 3199–3214 (2006).

79. Katsamba, P. *et al.* Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11594–11599 (2009).

80. Foty, R. A. & Steinberg, M. S. The differential adhesion hypothesis: A direct evaluation. *Dev. Biol.* **278**, 255–263 (2005).

81. Cachat, E., Liu, W., Hohenstein, P. & Davies, J. A. A library of mammalian effector modules for synthetic morphology. *J. Biol. Eng.* **8**, (2014).

82. Niessen, C. M. & Gumbiner, B. M. Cadherin-mediated cell sorting not determined by binding or adhesion specificity. *J. Cell Biol.* **156**, 389–399 (2002).

83. Duportet, X. *et al.* A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Res.* **42**, 13440–13451 (2014).

84. Wilson, M. H., Coates, C. J. & George, A. L. PiggyBac transposon-mediated gene transfer in human cells. *Mol. Ther.* **15**, 139–145 (2007).

85. Chassin, H. *et al.* A modular degron library for synthetic circuits in mammalian cells. *Nat. Commun.* **10**, 2013 (2019).

86. Dehay, C. & Kennedy, H. Cell-cycle control and cortical development. *Nat. Rev. Neurosci.* **8**, 438–450 (2007).

87. Bilitou, A. & Ohnuma, S. I. The role of cell cycle in retinal development: Cyclin-dependent kinase inhibitors co-ordinate cell-cycle inhibition, cell-fate determination and differentiation in the developing retina. *Developmental Dynamics* **239**, 727–736 (2010).

88. Halevy, T., Biancotti, J.-C. C., Yanuka, O., Golan-Lev, T. & Benvenisty, N. Molecular Characterization of Down Syndrome Embryonic Stem Cells Reveals a Role for RUNX1 in Neural Differentiation. *Stem Cell Reports* **7**, 777–786 (2016).

89. Sasai, Y. Cytosystems dynamics in self-organization of tissue architecture. *Nature* **493**, 318–326 (2013).

90. Dahl-Jensen, S. & Grapin-Botton, A. The physics of organoids: A biophysical approach to understanding organogenesis. *Development* **144**, 946–951 (2017).

91. Shyer, A. E. *et al.* Villification: How the gut gets its villi. *Science (80-.).* **342**, 212–218 (2013).

92. Savin, T. *et al.* On the growth and form of the gut. *Nature* **476**, 57–63 (2011).

93. Vanderhaeghen, P. & Cheng, H. J. Guidance molecules in axon pruning and cell death. *Cold Spring Harb. Perspect. Biol.* **2**, a001859 (2010).

94. MacKay, J. L. & Kumar, S. Simultaneous and independent tuning of RhoA and Rac1 activity with orthogonally inducible promoters. *Integr. Biol.* **6**, 1–10 (2014).

95. Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, (2011).

96. Tinevez, J. Y. *et al.* TrackMate: An open and extensible platform for single-particle tracking. *Methods* **115**, 80–90 (2017).

97. Wu, M.-R. M.-R. *et al.* A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). *Nat. Commun.* **10**, (2019).

98. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* **15**, 453–468 (2014).

99. Lelli, K. M., Slattery, M. & Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* **46**, 43–68 (2012).

100. Hwang, A., Maity, A., McKenna, W. G. & Muschel, R. J. Cell cycle-dependent regulation of the cyclin B1 promoter. *J. Biol. Chem.* **270**, 28419–28424 (1995).

101. Saukkonen, K. & Hemminki, A. Tissue-specific promoters for cancer gene therapy. *Expert Opinion on Biological Therapy* **4**, 683–696 (2004).

102. Dorer, D. E. & Nettelbeck, D. M. Targeting cancer by transcriptional control in cancer gene therapy and viral oncolysis. *Adv. Drug Deliv. Rev.* **61**, 554–571 (2009).

103. Takeshita, F. *et al.* Muscle creatine kinase/SV40 hybrid promoter for muscle-targeted long-term transgene expression. *Int. J. Mol. Med.* **19**, 309–315 (2007).

104. Chen, X., Scapa, J. E., Liu, D. X. & Godbey, W. T. Cancer-specific promoters for expression-

targeted gene therapy: ran, brms1 and mcm5. *J. Gene Med.* **18**, 89–101 (2016).

105. Amit, D. *et al.* Transcriptional targeting of glioblastoma by diphtheria toxin-A driven by both H19 and IGF2-P4 promoters. *Int. J. Clin. Exp. Med.* **5**, 124–135 (2012).

106. Hooijberg, E., Bakker, A. Q., Ruizendaal, J. J. & Spits, H. NFAT-controlled expression of GFP permits visualization and isolation of antigen-stimulated primary human T cells. *Blood* **96**, 459–466 (2000).

107. Nissim, L. & Bar-Ziv, R. H. A tunable dual-promoter integrator for targeting of cancer cells. *Mol. Syst. Biol.* **6**, 1–9 (2010).

108. Nissim, L. *et al.* Synthetic RNA-Based Immunomodulatory Gene Circuits for Cancer Immunotherapy. *Cell* **171**, 1138-1150.e15 (2017).

109. Xie, M. *et al.* B-Cell-Mimetic Designer Cells Provide Closed-Loop Glycemic Control. *Science (80-.).* **354**, 1296–1301 (2016).

110. Schukur, L., Geering, B., Charpin-El Hamri, G. & Fussenegger, M. Implantable synthetic cytokine converter cells with AND-gate logic treat experimental psoriasis. *Sci. Transl. Med.* **7**, (2015).

111. Sedlmayer, F., Aubel, D. & Fussenegger, M. Synthetic gene circuits for the detection, elimination and prevention of disease. *Nat. Biomed. Eng.* **2**, 399–415 (2018).

112. Xie, M. & Fussenegger, M. Designing cell function: assembly of synthetic gene circuits for cell biology applications. *Nat. Rev. Mol. Cell Biol.* **19**, 1 (2018).

113. Selvakumaran, M. *et al.* Ovarian epithelial cell lineage-specific gene expression using the

promoter of a retrovirus-like element. *Cancer Res.* **61**, 1291–1295 (2001).

114. Bao, R., Selvakumaran, M. & Hamilton, T. C. Targeted gene therapy of ovarian cancer using an ovarian-specific promoter. *Gynecol. Oncol.* **84**, 228–234 (2002).

115. Cheng, J. K. & Alper, H. S. Transcriptomics-Guided Design of Synthetic Promoters for a Mammalian System. *ACS Synth. Biol.* **5**, 1455–1465 (2016).

116. Saxena, P., Bojar, D. & Fussenegger, M. Design of Synthetic Promoters for Gene Circuits in Mammalian Cells. *Methods Mol. Biol.* **1651**, 263–273 (2017).

117. Lipinski, K. S. *et al.* Optimization of a synthetic β-catenin-dependent promoter for tumor-specific cancer gene therapy. *Mol. Ther.* **10**, 150–161 (2004).

118. Martinelli, R. & De Simone, V. Short and highly efficient synthetic promoters for melanoma-specific gene expression. *FEBS Lett.* **579**, 153–156 (2005).

119. Schlabach, M. R., Hu, J. K., Li, M. & Elledge, S. J. Synthetic design of strong promoters. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2538–2543 (2010).

120. Gerber, A. *et al.* Blood-borne circadian signal stimulates daily oscillations in actin dynamics and SRF activity. *Cell* **152**, 492–503 (2013).

121. Gosselin, P., Rando, G., Fleury-Olela, F. & Schibler, U. Unbiased identification of signal-activated transcription factors by barcoded synthetic tandem repeat promoter screening (BC-STAR-PROM). *Genes Dev.* **30**, 1895–1907 (2016).

122. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).

123. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).

124. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).

125. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).

126. Goode, D. K. *et al.* Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Dev. Cell* **36**, 572–587 (2016).

127. Almalki, S. G. & Agrawal, D. K. Key transcription factors in the differentiation of mesenchymal stem cells. *Differentiation* **92**, 41–51 (2016).

128. Lancaster, M. A. & Knoblich, J. A. Organogenesisin a dish: Modeling development and disease using organoid technologies. *Science (80-.).* **345**, (2014).

129. Shi, Y., Inoue, H., Wu, J. C. & Yamanaka, S. Induced pluripotent stem cell technology: A decade of progress. *Nat. Rev. Drug Discov.* **16**, 115–130 (2017).

130. Bao, R. *et al.* Activation of cancer-specific gene expression by the survivin promoter. *J. Natl. Cancer Inst.* **94**, 522–528 (2002).

131. Holliday, D. L. & Speirs, V. Choosing the right cell line for breast cancer research. *Breast cancer Res.* **13**, 1–7 (2011).

132. Soule, H. D. *et al.* Isolation and Characterization of a Spontaneously Immortalized Human Breast Epithelial Cell Line, MCF-10. *Cancer Res.* **50**, 6075–6086 (1990).

133. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

134. Eyler, C. E. & Rich, J. N. Survival of the fittest: Cancer stem cells in therapeutic resistance and angiogenesis. *J. Clin. Oncol.* **26**, 2839–2845 (2008).

135. Wakimoto, H. *et al.* Maintenance of primary tumor phenotype and genotype in glioblastoma stem cells. *Neuro. Oncol.* **14**, 132–144 (2012).

136. Wakimoto, H. *et al.* Human glioblastoma-derived cancer stem cells: Establishment of invasive glioma models and treatment with oncolytic herpes simplex virus vectors. *Cancer Res.* **69**, 3472–3481 (2009).

137. Suvà, M. L. *et al.* Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157**, 580–594 (2014).

138. Rheinbay, E. *et al.* An Aberrant Transcription Factor Network Essential for Wnt Signaling and Stem Cell Maintenance in Glioblastoma. *Cell Rep.* **3**, 1567–1579 (2013).

139. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).

140. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

141. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

142. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion

for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

143. Hahne, F. *et al.* flowCore: A Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* **10**, (2009).

144. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).

145. Schaffer, A. E. *et al.* Nkx6.1 Controls a Gene Regulatory Network Required for Establishing and Maintaining Pancreatic Beta Cell Identity. *PLoS Genet.* **9**, (2013).

146. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells: Commentary. *Science (80-.).* **318**, 1917–1920 (2007).

147. Srivastava, S. & Riddell, S. R. Engineering CAR-T cells: Design concepts. *Trends Immunol.* **36**, 494–502 (2015).

148. Fong, N. *et al.* Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol. Cell* **60**, 256–267 (2015).

149. Zhang, H., Rigo, F. & Martinson, H. G. Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Mol. Cell* **59**, 437–448 (2015).

150. Zorn, A. M. & Wells, J. M. Vertebrate Endoderm Development and Organ Formation. *Annu. Rev. Cell Dev. Biol.* **25**, 221–251 (2009).

151. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770–

1782 (2011).

152. West, S., Gromak, N., Norbury, C. J. & Proudfoot, N. J. Adenylation and exosome-mediated degradation of cotranscriptionally cleaved pre-messenger RNA in human cells. *Mol. Cell* **21**, 437–443 (2006).

153. Teixeira, A. *et al.* Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. *Nature* **432**, 526–530 (2004).

154. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–24 (2014).

155. Kurata, M. *et al.* Highly multiplexed genome engineering using CRISPR/Cas9 gRNA arrays. *PLoS One* **13**, (2018).

156. Gu, S. *et al.* The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. *Cell* **151**, 900–911 (2012).

157. Poi Liu, Y., Haasnoot, J., ter Brake, O., Berkhout, B. & Konstantinova, P. Inhibition of HIV-1 by multiple siRNAs expressed from a single microRNA polycistron. *Nucleic Acids Res.* **36**, 2811–2824 (2008).

158. Fellmann, C. *et al.* An optimized microRNA backbone for effective single-copy RNAi. *Cell Rep.* **5**, 1704–1713 (2013).

159. Lee, Y. & Rio, D. C. Mechanisms and regulation of alternative Pre-mRNA splicing. *Annual Review of Biochemistry* **84**, 291–323 (2015).

160. Chang, M. M. *et al.* Small-molecule control of antibody N-glycosylation in engineered

mammalian cells. *Nat. Chem. Biol.* **15**, 730–736 (2019).

161. Harlen, K. M. & Churchman, L. S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat. Rev. Mol. Cell Biol.* **18**, 263–273 (2017).

162. Müller-McNicoll, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA-protein complexes. TL - 14. *Nat. Rev. Genet.* **14 VN-r**, 275–287 (2013).

163. Slezak-Prochazka, I. *et al.* Cellular Localization and Processing of Primary Transcripts of Exonic MicroRNAs. *PLoS One* **8**, (2013).

164. Sundaram, G. M. *et al.* 'See-saw' expression of microRNA-198 and FSTL1 from a single transcript in wound healing. *Nature* **495**, 103–106 (2013).

165. Kim, Y. K. & Kim, V. N. Processing of intronic microRNAs. *EMBO J.* **26**, 775–783 (2007).

166. Wang, T., Xie, Y., Tan, A., Li, S. & Xie, Z. Construction and characterization of a synthetic MicroRNA cluster for multiplex RNA interference in mammalian cells. *ACS Synth. Biol.* **2015**, 47 (2015).

167. Han, J. *et al.* Posttranscriptional Cross regulation between Drosha and DGCR8. *Cell* **136**, 75–84 (2009).

168. Wilusz, J. E. *et al.* A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. *Genes Dev.* **26**, 2392–2407 (2012).

169. Dower, K., Kuperwasser, N., Merrikh, H. & Rosbash, M. A synthetic A tail rescues yeast nuclear accumulation of a ribozyme-terminated transcript. *RNA* **10**, 1888–99 (2004).

170. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36**, (2008).

171. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–21 (2012).

172. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu. Rev. ofGenetics* **51**, 171–194 (2017).

173. Pickering, B. M. & Willis, A. E. The implications of structured 5 untranslated regions on translation and disease. *Semin. Cell Dev. Biol.* **16**, 39–47 (2005).

174. Brunak, S., Engelbrecht, J. & Knudsen, S. *Prediction of Human mRNA Donor and Acceptor Sites from the DNA Sequence*. *J. Mol. Biol* **220**, (1991).

175. Grapin-Botton, A. *Endoderm specification*. (2008). doi:10.3824/stembook.1.30.1

176. Beccari, L. *et al.* Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. *Nature* **562**, 272–276 (2018).

177. Rodríguez-Seguel, E. *et al.* Mutually exclusive signaling signatures define the hepatic and pancreatic progenitor cell lineage divergence. *Genes Dev.* **27**, 1932–1946 (2013).

178. Simunovic, M. *et al.* Molecular mechanism of symmetry breaking in a 3D model of a human epiblast. *Nat. Cell Biol.* **21**, 900–910 (2019).

179. Múnera, J. O. *et al.* Differentiation of Human Pluripotent Stem Cells into Colonic Organoids via Transient Activation of BMP Signaling. *Cell Stem Cell* **21**, 51-64.e6 (2017).

Tools for engineering multicellular systems through cell sorting and cell state detection

By

Casper Nørskov Enghuus

B.Sc.Eng, Human Life Science Engineering, Technical University of Denmark (2013)
M.Sc.Eng, Bioinformatics and Systems Biology, Technical University of Denmark (2015)

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Signature of Author .....
Casper Nørskov Enghuus
Microbiology Graduate Program
April 16, 2021

Certified by .....
Ron Weiss, PhD
Professor of Biological Engineering & Professor of Electrical Engineering and Computer Science, MIT
Thesis Supervisor

Accepted by .....
Jacquin C. Niles
Associate Professor of Biological Engineering
Chair of Microbiology Program

THESIS COMMITTEE:

Rudolf Jaenisch, PhD
Chairman, Thesis Committee
Professor of Biology, MIT

Ron Weiss, PhD
Thesis Supervisor
Professor of Biological Engineering & Professor of Electrical Engineering and Computer Science, MIT

Domitilla Del Vecchio,
Thesis Committee Member
Professor of Mechanical Engineering, MIT

Wilson Wong
Thesis Committee Member
Associate Professor of Biomedical Engineering
Boston University

Tools for engineering multicellular systems through cell sorting and cell state detection

By

Casper Nørskov Enghuus

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

The human genome, the genetic blueprint that every cell in our body follows, encodes approximately 20,000 genes. Through complex regulation of these genes, each cell is able to play the role it needs within our body. Synthetic biology, an emerging field in biology, seeks to expand on this blueprint and create cells with novel functions. The aim of this thesis is to provide methods that expands our ability to engineer and control multicellular systems by detecting and rewriting the cell state.

We first develop a method that enables the creation of a synthetic cell state to control morphogenesis. Using inducible expression of recombinases, we show this approach can induce a cell to commit to one of two mutually exclusive cell states. By regulating the expression of recombinases, we are able to control the distribution of cell states within an initially monoclonal and homogenous population of cells. We use the induction of a synthetic cell state to control morphogenesis by cell state-specific expression of homotypic cadherins which controls the cell's adhesive properties. This enables us to create a large number of different shapes and control morphogenesis.

Secondly, we develop a library-based approach for cell state-specific gene regulation. We design a set of 6,107 Synthetic Promoters with Enhanced Cell-State Specificity (SPECS), and identify several SPECS with spatiotemporal specificity during the programmed differentiation of stem cells, as well as SPECS that are highly specific for breast cancer and glioblastoma stem-like cells.

Thirdly, we develop a method that allows detection of endogenous gene expression without modifying the endogenous gene itself. We show that placing a regulatory RNA downstream of a terminator allows for expression of the regulatory RNA, and demonstrate this method for miRNAs and gRNAs.

Together, this thesis develops methods to create synthetic cell states that can be used to control morphogenesis, and provides tools to detect endogenous cell states which can serve as inputs to control gene regulatory networks.

Thesis Supervisor: Ron Weiss, PhD Title: Professor of Biological Engineering and Professor of Electrical Engineering and Computer Science, MIT

UNIVERSITY MICROFILMS INC.

D
PAO

MIT Doctoral Dissertation
PUBLISH ABSTRACT ONLY
INFORMATION

Abstract No. _____

DO NOT WRITE IN THIS SPACE

Vol/Issue _____

School Code _____

Advisor _____

PLEASE TYPE OR PRINT

PERSONAL DATA

1. Full name (as it appears on dissertation title page)
Enghuus (last) Casper (first) Noerskov (middle)

2. Year of birth (optional) _____

3. Present mailing address 79 Columbus Ave, Apt 1, Somerville MA 02143

Future mailing address _____

Effective date of future mailing address _____

Home telephone 857-829-7720 Business telephone _____

DOCTORAL DEGREE DATA

4. Full name of university conferring degree Massachusetts Institute of Technology

5. Degree awarded (check one) ☑ Ph.D. ☐ Sc.D.

6. Year degree awarded 2021

7. IMPORTANT: Attach a copy of your dissertation title page and abstract to thls form. Please be certain that the name of your dissertation supervisor is included on both.

8. Subject categories for your dissertation. Enter 4-digit code from 'Subject Categories' list found on the opposite side of this form, and write in the category selected. You may enter two additional codes and categories on the lines provided.

Code 0306 Category Biological Sciences

Code _____ Category _____

Code _____ Category _____

(Optional) List up to five additional words from your dissertation not already found in ***either*** your title ***or*** abstract which would be useful for database access.

a. _____ b. _____ c. _____

d. _____ e. _____

UMI-ProQuest Subject Categories

The ProQuest Dissertations and Theses (PQDT) database and the ProQuest/UMI citation indices are arranged by subject categories. Please select the one category below that best describes your field of research or creative work. You may add one or two additional categories on your UMI-ProQuest form that will also be associated with your work as secondary subjects.

Arts, Business, Education, Humanities, and Social Sciences

AREA, ETHNIC, AND GENDER STUDIES

African American studies	0296
African studies	0293
American studies	0323
Asian American studies	0343
Asian studies	0342
Baltic studies	0361
Black studies	0325
Canadian studies	0385
Caribbean studies	0432
Classical studies	0434
East European studies	0437
Ethnic studies	0631
European studies	0440
French Canadian culture	0482
Gender studies	0733
GLBT studies	0492
Hispanic American studies	0737
Holocaust studies	0507
Islamic culture	0512
Judaic studies	0751
Latin American studies	0550
Middle Eastern studies	0555
Native American studies	0740
Near Eastern studies	0559
North African studies	0560
Pacific Rim studies	0561
Regional studies	0604
Scandinavian studies	0613
Slavic studies	0614
South African studies	0654
South Asian studies	0638
Sub Saharan Africa studies	0639
Women's studies	0453

BUSINESS

Accounting	0272
Arts management	0424
Banking	0770
Business	0310
Entrepreneurship	0429
Finance	0508
Management	0454
Marketing	0338
Sports management	0430

COMMUNICATIONS AND INFORMATION SCIENCES

Communication	0459
Information science	0723
Journalism	0391
Library science	0399
Mass communication	0708
Technical communication	0643
Web studies	0646

FINE AND PERFORMING ARTS

Art criticism	0365
Art history	0377
Cinematography	0435
Dance	0378
Design	0389
Film studies	0900
Fine arts	0357
Music	0413
Performing arts	0641
Theater	0465
Theater history	0644

EDUCATION

Adult education	0516
Art education	0273
Bilingual education	0282
Business education	0688
Community college education	0275
Continuing education	0651
Curriculum development	0727
Early childhood education	0518
Education	0515
Education finance	0277
Education policy	0458
Educational administration	0514
Educational evaluation	0443
Educational leadership	0449
Educational psychology	0525
Educational technology	0710
Educational tests & measurements	0288
Elementary education	0524
English as a second language	0441
Foreign language instruction	0444
Gifted education	0445
Health education	0680
Higher education	0745
Higher education administration	0446
History of education	0520
Home economics education	0278
Industrial arts education	0521
Instructional design	0447
Language arts	0279
Mathematics education	0280
Middle school education	0450
Multicultural education	0455
Music education	0522
Pedagogy	0456
Performing arts education	0457
Philosophy of education	0998
Physical education	0523
Reading instruction	0535
Religious education	0527
School counseling	0519
Science education	0714
Secondary education	0533
Social sciences education	0534
Sociology of education	0340
Special education	0529
Teacher education	0530
Vocational education	0747

HUMANITIES

HISTORY

African history	0331
American history	0337
Ancient history	0579
Asian history	0332
Black history	0328
Canadian history	0334
European history	0335
History	0578
History of Oceania	0504
History of science	0585
Latin American history	0336
Medieval history	0581
Middle Eastern history	0333
Military history	0772
Modern history	0582
Russian history	0724
World history	0506

LANGUAGE & LITERATURE

African literature	0316
American literature	0591
Ancient languages	0289
Asian literature	0305
British and Irish literature	0593
Canadian literature	0352
Caribbean literature	0360
Classical literature	0294
Comparative literature	0295
English literature	0593
French Canadian literature	0355
Germanic literature	0311
Icelandic & Scandinavian literature	0362
Language	0679
Latin American literature	0312
Linguistics	0290
Literature	0401
Literature of Oceania	0356
Medieval literature	0297
Middle Eastern literature	0315
Modern language	0291
Modern literature	0298
Rhetoric	0681
Romance literature	0313
Slavic literature	0314

PHILOSOPHY AND RELIGION

Aesthetics	0650
Biblical studies	0321
Canon law	0375
Clerical studies	0319
Comparative religion	0618
Divinity	0376
Epistemology	0393
Ethics	0394
Logic	0395
Metaphysics	0396
Pastoral counseling	0397
Philosophy	0422
Philosophy of Religion	0322
Philosophy of science	0402
Religion	0318
Religious history	0320
Spirituality	0647
Theology	0469

LAW AND LEGAL STUDIES

Alternative dispute resolution	0649
Intellectual property	0513
International law	0616
Law	0398
Patent law	0562

SOCIAL SCIENCES

Archaeology	0324
Area planning and development	0341
Criminology	0627
Cultural anthropology	0326
Demography	0938
Economic history	0509
Economic theory	0511
Economics	0501
Economics, Commerce-Business	0505
Economics, Labor	0510
Folklore	0358
Forensic anthropology	0339
Geography	0366
Individual & family studies	0628
International relations	0601
Labor relations	0629
Military studies	0750
Organization theory	0635
Organizational behavior	0703
Peace studies	0563
Physical anthropology	0327
Political Science	0615
Public administration	0617
Public policy	0630
Recreation and tourism	0814
Social research	0344
Social structure	0700
Social work	0452
Sociolinguistics	0636
Sociology	0626
Transportation planning	0709
Urban planning	0999

INTERDISCIPLINARY

Alternative energy	0363
Biographies	0304
Climate change	0404
Cultural resources management	0436
Energy	0791
Food science	0359
Home economics	0386
Information technology	0489
Multimedia	0558
Museum studies	0730
Sustainability	0640
Textile research	0994
Wood sciences	0746

Behavioral, Natural, and Physical Sciences

AGRICULTURE

Agriculture	0473
Agronomy	0285
Animal diseases	0476
Animal sciences	0475
Fisheries and aquatic sciences	0792
Forestry	0478
Horticulture	0471
Plant pathology	0480
Plant sciences	0479
Range management	0777
Soil sciences	0481
Urban forestry	0281
Wildlife management	0286

ARCHITECTURE

Architecture	0729
Architectural engineering	0462
Landscape architecture	0390

BEHAVIORAL SCIENCES

Animal behavior	0602
Behavioral sciences	0384
Clinical psychology	0622
Cognitive psychology	0633
Counseling psychology	0603
Developmental psychology	0620
Experimental psychology	0623
Occupational psychology	0624
Personality psychology	0625
Physiological psychology	0989
Psychobiology	0349
Psychology	0621
Quantitative psychology and psychometrics	0632
Social psychology	0451

BIOLOGICAL SCIENCES

Biochemistry	0487
Bioinformatics	0715
Biology	0306
Biomechanics	0648
Biophysics	0786
Biostatistics	0308
Cellular biology	0379
Developmental biology	0758
Endocrinology	0409
Entomology	0353
Evolution & development	0412
Genetics	0369
Histology	0414
Limnology	0793
Microbiology	0410
Molecular biology	0307
Morphology	0287
Neurosciences	0317
Parasitology	0718
Physiology	0719
Plant biology	0309
Systematic biology	0423
Virology	0720
Zoology	0472

ECOSYSTEM SCIENCES

Ecology	0329
Macroecology	0420
Paleoecology	0426

ENGINEERING

Aerospace engineering	0538
Artificial intelligence	0800
Automotive engineering	0540
Biomedical engineering	0541
Chemical engineering	0542
Civil engineering	0543
Computer engineering	0464
Computer science	0984
Electrical engineering	0544
Engineering	0537
Geological engineering	0466
Geophysical engineering	0467
Geotechnology	0428
Industrial engineering	0546
Mechanical engineering	0548
Mining engineering	0551
Naval engineering	0468
Nanotechnology	0652
Nuclear engineering	0552
Ocean engineering	0547
Operations research	0796
Packaging	0549
Petroleum engineering	0765
Plastics	0795
Robotics	0771
System science	0790

ENVIRONMENTAL SCIENCES

Conservation biology	0408
Environmental economics	0438
Environmental education	0442
Environmental engineering	0775
Environmental geology	0407
Environmental health	0470
Environmental justice	0619
Environmental law	0439
Environmental management	0474
Environmental philosophy	0392
Environmental science	0768
Environmental studies	0477
Land use planning	0536
Natural resource management	0528
Water resources management	0595
Wildlife conservation	0284

GEOSCIENCES

Aeronomy	0367
Atmospheric chemistry	0371
Atmospheric sciences	0725
Biogeochemistry	0425
Biological oceanography	0416
Chemical oceanography	0403
Continental dynamics	0406
Geobiology	0483
Geochemistry	0996
Geographic information science and geodesy	0370
Geology	0372
Geomorphology	0484
Geophysics	0373
Hydrologic sciences	0388
Marine geology	0556
Meteorology	0557
Mineralogy	0411
Paleoclimate science	0653
Paleontology	0418
Petroleum geology	0583
Petrology	0584
Physical geography	0368
Physical oceanography	0415
Planetology	0590
Plate tectonics	0592
Remote sensing	0799
Sedimentary geology	0594

HEALTH AND MEDICAL SCIENCES

Aging	0493
Alternative medicine	0496
Audiology	0300
Dentistry	0567
Epidemiology	0766
Gerontology	0351
Health care management	0769
Health sciences	0566
Immunology	0982
Kinesiology	0575
Medical ethics	0497
Medical imaging and radiology	0574
Medicine	0564
Mental health	0347
Nursing	0569
Nutrition	0570
Obstetrics and gynecology	0380
Occupational health	0354
Occupational therapy	0498
Oncology	0992
Ophthalmology	0381
Osteopathic medicine	0499
Pathology	0571
Pharmaceutical sciences	0572
Pharmacology	0419
Physical therapy	0382
Public health	0573
Public health occupations education	0500
Speech therapy	0460
Surgery	0576
Toxicology	0383
Veterinary medicine	0778

MATHEMATICAL AND PHYSICAL SCIENCES

Acoustics	0986
Analytical chemistry	0486
Applied mathematics	0364
Astronomy	0606
Astrophysics	0596
Atomic physics	0748
Chemistry	0485
Condensed matter physics	0611
Electromagnetics	0607
High temperature physics	0597
Inorganic chemistry	0488
Low temperature physics	0598
Materials science	0794
Mathematics	0405
Mechanics	0346
Molecular chemistry	0431
Molecular physics	0609
Nanoscience	0565
Nuclear chemistry	0738
Nuclear physics	0756
Optics	0752
Organic chemistry	0490
Particle physics	0798
Physical chemistry	0494
Physics	0605
Plasma physics	0759
Polymer chemistry	0495
Quantum physics	0599
Statistics	0463
Theoretical mathematics	0642
Theoretical physics	0753