Using Visual Phenotypes to Dissect Sequence-Function Relationships and Complex Drug Responses

Nicholas Hasle

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Douglas Fowler (Chair)

Emily Hatch

Hao Yuan Kueh

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Using Visual Phenotypes to Dissect Sequence-Function

Relationships and Complex Drug Responses

Nicholas Hasle

Chair of the Supervisory Committee:

Douglas M. Fowler, PhD

Genome Sciences

Cellular morphology is a potent indicator of cellular function and dysfunction, but the relationships between morphology, genetic variants, and cellular state remain incompletely understood. In this thesis, I describe a method called Visual Cell Sorting, which can be used to systematically characterize cellular morphologies and other visual phenotypes of interest. In a Visual Cell Sorting experiment, automated imaging and phenotypic analysis directs selective illumination of Dendra2, a photoconvertible fluorescent protein expressed in live cells; these photoactivated cells are then isolated using fluorescence-activated cell sorting. Visual Cell Sorting can be used to characterize hundreds of genetic variants according to a visual phenotype and to discover genes that are responsible for maintaining homeostasis in response to drug treatment. Visual Cell Sorting's greatest strength is that a variety of downstream assays can be performed on the separated cells, which together can characterize a morphologic phenotype in a multimodal and systematic fashion.

Table of Contents

Dedication

Throughout high school and college, during classes and even in lab-based internships, a rosy picture of science emerged in my mind's eye. I believed that scientists did three things: (a) run experiments, (b) analyze data, and (c) present their results. I believed that these three activities were performed over and over again in an endless cycle of facile discovery and clever thinking. How naïve I was! Throughout my Ph.D., I have come to realize that scientists are far more courageous, tenacious and passionate than I gave them credit for. As I became a scientist myself, I was surprised to realize that published protocols do not always work, that months can go by without a single interpretable result, that the currency of scientific work is not discovery itself but rather the awe you inspire in others. Most importantly, I learned that science is laced with failure, and that humility is its best antidote.

With this in mind, I dedicate this thesis to, and thank profusely, all of the people who instilled within me the passion, tenacity, courage, and humility to complete my Ph.D. I'd like to thank my British schoolteachers, whose names I do not even remember, for instilling in me a deep respect for rules and protocols; my American grade school teachers like Coleen Besman, for showing me that teachers can care deeply for their students; my high school biology teachers (Caryn Abrey and Devin Parry) for fostering my passion for cellular biology; my university teachers, who somehow managed to keep me enthralled by biochemistry for three years; and my university tutor, Robert Gilbert, for teaching me most of the biochemistry I know and for supporting the sometimes-rather-odd ideas I proposed in my biochemistry essays. I would also like to thank my previous scientific mentors, including Mehmet Sarikaya, Ken Stuart, Julian Knight, David Rawlings, and Andy Scharenberg, for giving me the time and resources to learn how to do perform and interpret scientific experiments with confidence.

Over the course of several months, I performed the same multi-day experiment (trying to sort paclitaxel-treated cells for nuclear lobulation) thirty-three times. I attribute the tenacity required to keep going to my mom, who taught me that not going to sports practice is not an option, and to the swimming coaches and teammates I've had. Thank you, Tom Pardee, Justine Schluntz, Owen Wurzbacher, Liz Weitz, Aydan Sarikaya, Tom Booth, James Jurkiewicz, Tristan Goodfellow, Joe Northover, and everyone else who taught me that failure does not mean that you give up.

Thank you to all of the friends who gave me the courage to pursue and finish this degree: knowing that all of you – Justin Norden, Tom Booth, Hannah Thomas, Sanjay Srivatsan, Greg

Olson, and countless others who aren't named here – believed that I could do this helped me enormously when things weren't going my way.

A special thanks to Doug, whose most important lesson to me is that humility provides the most graceful defense against failure (or *perceived* failure). I could not have asked for a more powerful, enduring lesson – or a more communicative, dynamic, and tough-love oriented mentor.

A special thanks to my family, whose unconditional love is critical for having the courage to try things that may not work out. A special thank you to my mom for buying me all sorts of science-based encyclopedias, and for letting me devour them on the floor of her home office. A special thank you to both of my parents for showing me that success, however it is measured, does not *de facto* make someone more worthy of affection, kindness, or attention.

And finally, a special thank you to my wife, Vanessa, for being someone I look forward to seeing and laughing with every day, no matter how easy or hard my work has been. I don't think I could have done this without you.

Introduction

Structure-Function Relationships: A Hallmark of Biology

The discovery of cells can be traced back to Robert Hooke and Antoine van Leeuwenhoek, who in the mid-17th century used microscopes to examine plant material, fungi, pond water, and bodily fluids (Gest, 2004). Descriptions of what cells looked like dominated their scientific reports; after all, these descriptions were the primary way they could hope to understand the mysterious functions of the microscopic "globules" and "animalcules". In essence, these microscopists were following the steps of other biologists in hypothesizing the connections between biological form and function.

In the ensuing centuries, scientists continued to leverage form-function relationships to understand biological systems. Zoologists connected phenotypes such as beak shape and size to feeding behaviors and diet (Darwin, 1845). Physiologists uncovered the function of the kidney by examining the structural relationships between its blood vessels, tubules, and the ureter (Hierholzer *et al*, 1999). Biochemists used the crystal structure of hemoglobin to understand how it carried blood to various organs (Liddington *et al*, 1988). In these fields, visualizing structures provided critical biological insights in part because many of the structures in question – beaks, blood vessels, and alpha helices – could be approximated to behave like structures human encounter in everyday life, such as hands, pipes, and levers.

Cellular biologists have typically placed less emphasis on structure-function relationships. Some cellular structures, such as cilia, axons, and dendrites, could be ascribed structure-related functions. However, the relationships between cellular morphology (defined loosely as the sum of all cell structures for a given cell) and cellular function, to this day, remains incompletely understood. Part of the problem is that our structural understanding of cells is incomplete because, until recently, we have lacked technology that permits visualization of whole cells at protein-scale (~10 nM resolution; Oikonomou & Jensen, 2016)). An equally important challenge is that cellular morphology cannot be easily approximated by our lived experiences as humans: we do not encounter cell-like structures (described as "globules" by Leeuwenhoek) in our everyday life (Gest, 2004). As a result of this, it is challenging to generate hypotheses regarding what a ruffled cell membrane or a misshapen nucleus might tell us about a cell's function, for example.

Genes encode for the molecular components of the cell, thereby determining its form and function. Pathogenic variants often result in loss of a cellular function *and* a change in cellular

morphology; we can infer that the morphological and functional changes caused by the variant are therefore connected. Famous examples of this are sickle cell anemia and hereditary spherocytosis variants (Diez-Silva *et al*, 2010). Here, changes in the gene encoding for hemoglobin cause a change in red blood cell morphology from a biconcave shape to a sickled or spherical shape, respectively. The resulting cells become stuck or damaged as they pass through small capillaries; we can therefore infer that the biconcave shape of blood cells is important for their ability to squeeze through small diameter capillaries.

There are other genetic variants that are known to affect cellular morphology. For example, variants in the tumor suppressor *PTEN* can cause morphologic changes in cell lines and tissues (Mester & Eng, 2013; Tamura *et al*, 1998). Variants in *LMNA* can cause a variety of defects in the nuclear lamina, which are associated with a striking array of clinical phenotypes (Piekarowicz *et al*, 2017). However, the mechanism by which these genes cause these changes, and what they mean for cellular function, remains unknown. Understanding how a variety of variants across these proteins affect cellular morphology would facilitate the generation of hypotheses regarding protein structure-function and protein function-cell morphology relationships.

In 2014, Fowler and colleagues published a roadmap for using high throughput reverse genetic screens to elucidate how genetic variants affect protein function and cellular phenotypes (Fowler & Fields, 2014). Thereafter, scientists have published a multitude of studies that examine the relationship between selectable cellular phenotypes, such as growth or fluorescence intensity, and protein variants (Starita *et al*, 2015; Thyagarajan & Bloom, 2014). One example of this is VAMP-seq, which uses a fluorescent protein tag to quantify the abundance a cell-specific protein variant across hundreds of thousands of cells in a single experiment. By selecting for cells with different fluorescence intensities (and thereby different abundances of the protein of interest) and sequencing the selected the protein variants in those cells, Matreyek, Starita, and colleagues could ascribe abundance scores to thousands of *PTEN* and *TMPT* variants (Matreyek *et al*, 2018). Critically, however, screens such as VAMP-seq require a selection method; and because we have been unable to select for cells in high-throughput according to their visual phenotypes, screens on how *PTEN* and *LMNA* affect cellular and nuclear morphology have remained out of reach.

New Methods for Detecting Cell Morphology-Function Relationships

One rather simple explanation for our sparse understanding of cell morphology-function relationships is that they are rare; that is, there are few cellular morphologic phenotypes that

provide robust information about cellular function. High content imaging is an approach developed in the last two decades that has allowed us to test this hypothesis at higher scale. High content imaging leverages recent advances in microscope technology (microplates, automated microscopes, inexpensive dyes that label subcellular structures, increased data storage capacity, faster computers), genetics (oligonucleotide synthesis, mammalian expression vectors), pharmacology (chemical libraries of biologically-active compounds), and robotics. Here, cells are (1) seeded in thousands of microwells, (2) treated with genetic or chemical perturbations, (3) stained with dyes or antibodies that are related to a morphologic or visual phenotype of interest, and (4) imaged. The resulting images are subject to automated image analysis pipelines that quantify parameters related to the phenotype of interest (Boutros *et al*, 2015).

Until 2010, high content imaging was used on single, parametrically defined phenotypes of interest. For example, Mukherji (2006) and colleagues searched for cell cycle regulators by performing RNAi-knockdown of over 1,000 genes on cells and using total DAPI intensity distributions to find wells containing cells with cell cycle defects. In the past decade, however, groups have tried to move to an unbiased approach that comprehensively searches for cell morphology-function relationships (Fuchs 2010). Cell Painting, developed by Anne Carpenter's group at MIT, is a high content imaging workflow where (1) cells are fixed, (2) cell structures are stained with standard set of dyes (3) a widefield fluorescence microscope is used to profile the morphologies of the cells, and (4) sophisticated image analysis pipelines quantify various cellular morphologic parameters (Caicedo *et al*, 2017; Rohban *et al*, 2017). Dr. Carpenter's lab treated cells with various drug and genetic perturbations (known to impact cellular function) and used Cell Painting to calculate a representative "morphological profile" for each perturbation. They found that perturbations with similar effects on cellular function also had overlapping morphological profiles. Critically, these reproducible, function-related morphological profiles were related to subtle visual phenotypes that were not readily detected by scientists examining the images (Rohban *et al*, 2017). This finding demonstrates that cell morphology-function relationships are widespread, yet they can be subtle and non-obvious to humans.

Though these perturbation-based Cell Painting screens can provide critical information about structure-function relationships in cell biology, they suffer from two critical shortcomings. First, the perturbation-centric (i.e. function-centric) approach makes it challenging to ask questions that pertain to specific morphologies. An example of such a question might be, "What are the cellular functions associated with membrane ruffling in my biological system?". To

answer this question with Cell Painting, one must take the biological system in question and expose it to a vast array of perturbations, with the hope of finding a perturbation or set of perturbations that result in membrane ruffling. Second, the use of drug or genetic perturbations means that the functions in question are tied to a specific protein target or gene, rather than to a cellular state. This approach makes it challenging to link cell morphologies to cell states such as senescence, cellular differentiation, and drug resistance. Third, the “one perturbation, one well” paradigm in Cell Painting demands careful consideration of batch-related artifacts, including assigning the same perturbation to multiple wells and using appropriate batch-correction analyses. Given the drawbacks of Cell Painting, other methods that link cell morphology to function – and to cell state in particular – are needed.

Transcriptomics and proteomics have emerged as comprehensive and unbiased readouts of tissue and cell state, but it remains challenging to link these readouts to cellular morphologies (Lein *et al*, 2017). *In situ* methods such as Multiplexed, Error-Resistant Fluorescence In Situ Hybridization (MERFISH) and *in situ* sequencing are capable of deriving image-based transcriptomes on fixed cells using a confocal microscope (Chen *et al*, 2015; Feldman *et al*, 2019; Lee *et al*, 2014, 2015). In MERFISH, a set of combinatorial fluorescent probes are used to identify individual cell transcripts; by counting the number of spots for each transcript probed, it is possible to accurately quantify the expression of thousands of genes. *In situ* sequencing uses reverse transcription and rolling circle amplification to amplify the RNA corresponding to set of genes and Illumina sequencing chemistry to sequence a short stretch of the resulting cDNA to identify them. Though these technologies enable powerful transcriptome-morphology correlations to be made, their scale is limited, and they necessitate complex protocols, many rounds of imaging, and expensive disposable reagents. Furthermore, they cannot measure other potent indicators of cell state such as proteome-level protein abundances and metabolomics, and cellular phenotypes such as drug resistance.

Image-based cell selection methods, which seek to physically separate cells with different morphologies from one another, provide an alternative way to probe cell morphology-function relationships. The key advantage to cell selection methods is their downstream flexibility: by isolating cells of interest, it is possible to perform assays that measure cell phenotypes not easily measured using a microscope, such as drug resistance or cytokine production. Furthermore, it is in theory possible to apply almost any molecular assay downstream of cellular separation, including proteomics and metabolomics. The oldest and most widespread image-based cell selection method is micromanipulation, which involves manually picking cells of

interest using a microscope and micropipette. Though simple, micromanipulation is technically challenging; limited in throughput to 6 cells per minute; requires live, non-adherent cells; and cannot be automated (Zhang *et al*, 2014b).

In the past five years, new image-based cell isolation methods that require less technical expertise and can be performed on adherent, live cells have been developed. Such methods include Single Cell Magneto-Optical Capture (Binan *et al*, 2019), Photostick (Chien *et al*, 2015), and optical painting and fluorescence activated sorting (Kuo *et al*, 2016). These methods image live cells and "tag" cells with a phenotype of interest using microscope-controlled light. Tagging is achieved by photoactivation of (1) a fluorescent protein expressed in the cells (2) a fluorescent molecule attached to the cells, or (3) a metal binding moiety attached to the cells. A fluorescence- or magnet- activated cell sorting (FACS or MACS) step separates the tagged cells with a morphology of interest from those without it. Subsequently, the sorted cells can be subject to an assay that measures a non-image-based cell phenotype. Though widely applicable in theory, the methods published by Binan et al, Chien et al, and Kuo et al suffer from two critical shortcomings: they require custom-made hardware or non-commercially available reagents, and they are limited in throughput to less than 100 cells per experiment. The lack of throughput precludes the use of these technologies as a genetic screening tool, or as a tool that can be combined with standard transcriptomic and proteomic workflows.

Visual Cell Sorting: A Historical Perspective

In 2016, state-of-the-art microscope-based technologies struggled to comprehensively characterize a visual phenotype (e.g. cell morphology) of interest. *In situ* methods, which were capable of sequencing ~10 nucleotides per rolling circle amplicon, appeared to be technically challenging. Two *in situ* sequencing paper were published in 2013 and 2014 and no follow-up papers were published in the intervening two or three years (Ke *et al*, 2013; Lee *et al*, 2014). In the Shendure lab, a post-doctoral candidate had been working on *in situ* sequencing for two years and was unable to achieve accurate sequencing results of a control barcode after the first nucleotide. Furthermore, at the onset of my graduate work in 2016, I worked for months trying to get reproducible results for the reactions *in vitro*. Given the evident challenges getting the *in situ* sequencing protocol to work, and the lack of lab expertise in the reactions necessary to perform *in situ* sequencing, our lab decided to examine whether we could develop an improved and higher throughput image-based cell isolation method instead. Such a method could be used to perform selections in a genetic screen, or to apply any genomics method (e.g. proteomics, ATAC-seq, DNA sequencing) to cells with a visual phenotype (e.g. morphology) of interest.

To achieve higher throughput in image-based cell isolation, we decided to employ a digital micromirror device (DMD), which was used by Chien and colleagues (Chien *et al*, 2015). Digital micromirror devices are an array of thousands of micromirrors, which can each be switched on or off. If light is shined onto the device, only the micromirrors in the "on" position will direct light towards the objective; other light is discarded. The selective reflection of light results in patterned illumination at the plane of the image, thus allowing activation of cells to be parallelized and higher throughput to be achieved. The commercial availability of the digital micromirror device and its compatibility with a wide range of automated microscopes were also important features, as they gave us confidence that we would be able to complete the experiments as a lab without much trouble.

We next examined our options for photoactivatable or photoconvertible moieties that would mark the cells. A genetically encoded system seemed desirable, as it was less likely than a chemical system to be toxic to cells or expensive. This narrowed the search to proteins that reacted to light, including LOV2 photoswitches (Zimmerman *et al*, 2016), the photocleavable protein PhoCl (Zhang, Wei *et al*, 2017), photoactivatable transcription factor systems (Zhou *et al*, 2007) photoactivatable protein interaction systems (Toettcher *et al*, 2011), and photoconvertible fluorescent proteins; Table I.1). These systems differed in the wavelengths required for activation, the timescale required for activation, the irreversibility of activation, and the potential output mechanisms. Ultimately, we chose to use the photoconvertible fluorescent protein Dendra2 because it was rapidly and irreversibly activated, and because the readout (a switch from green to red fluorescence) was readily selected for using FACS.

Optogenetic system	**Mechanism**	**Activation light wavelength**	**Time to full activation**	**Half life of readout signal**	**Dynamic range**	**Notes**
Dendra2	Photoactivatable fluorescent protein	405nm	1 second	6 hours	~1500x	Activation performed using UV laser. Also activated at 488nm but less efficiently
PS-CFP	Photoactivatable fluorescent protein	405nm	5 seconds	*Not reported*	~1500x	Activation performed using blue LED. Same wavelength used for excitation and imaging.
LOV2	Photoswitchable alpha helix unfolding	450nm	2 minutes	30 seconds	NA	Reversible; not easily translated into fluorescence or MACs based selection
EL222	Photoactivatable transcription factor	465nm	9 hours	21 hours	~100x	Could induce transcription of FP or TM domain
OptoPanX	Photocleavable protein activates channel for cell-impermeable fluorophore	400nm	30 minutes	*Not reported*	~10x	Channel opening allows fluor to enter cell

PhyB-VP16	Light-mediated TF nuclear import	640nm	20 minutes	*Not reported*	~10x	Could be tied to transcription of FP or TM domain
LACE	Light-inducible VP16-dCas9	400-488nm	30 hours	12 hours	~500x	Could induce transcription of FP or TM domain

Table I.1. Genetically encoded photoactivation systems. References: Chudakov *et al*, 2007, 2004; Motta-Mena *et al*, 2014; Zhang, Wei *et al*, 2017; Polstein & Gersbach, 2015)

Collateral activation is a critical issue in image-based cell selection experiments. If light is used to mark cells of interest, great care must be taken to avoid illumination of adjacent cells. This "collateral activation" is caused by scattering of light by cells, dust, and glass; by small inaccuracies in laser or DMD alignment; or by cells growing on top of one another. The partially activated cells can complicate the subsequent selection process by smearing the signal demarcation between activated and unactivated cells; or worse, by causing full activation of cells that are not of interest. Our preliminary experiments were marred by collateral activation, and we fixed this problem with two strategies. First, we switched from working with HEK-293T cells, which can grow in clusters of cells, to U2OS cells, which are easier to image and do not grow on top of one another. Secondly, Dr. Emily Hatch suggested we move expression of Dendra2 from the whole cell to just the nucleusUltimately, this permitted activation of just the nucleus of cell, which increases the distance between potentially activated regions. This increased distance means that scattered light must travel much further to cause collateral activation; furthermore, a poorly aligned DMD will not necessarily result in activation of nearby cells. Further discussion of why moving Dendra2 to the nucleus is important is described in the Appendix.

Having picked a photoactivatable system and cell line, we next needed to determine when the analysis would be performed. We perceived two options: perform imaging and analysis of all cells, then go back and photoactivate the cells of interest; or image, analyze and photoactivate each cell in turn. We decided to do the latter, for two reasons: (1) should the experiment be done in live cells, the cells would likely move after they were imaged and before they were activated; and (2) to image all cells and then activate would require assuming that the microscope could align precisely to the location it had taken each image (to precision in the micron range), and we did not get assurance that this would be feasible if there were hours between when imaging and activation took place. This decision, in turn, led to problems regarding the timing of activation and sorting, which will be discussed in Chapter 1.

The final product of these efforts is a method that we call Visual Cell Sorting. We present it as an image-based cell selection method that facilitates our ability to test for relationships

between subcellular visual phenotypes, cellular morphology, genetic variants, and cell state. The key advantage of Visual Cell Sorting compared to other image-based selection methods is its throughput, which opens up new downstream applications that require tens of thousands of cells. Its ability to separate cells into up to four activation bins is another unique feature. In Chapter 1 of this thesis, I describe the method in detail, including its workflow, its high throughput (over 1000-fold higher than similar technologies) and its ability to separate cells into morphologic 4 bins. In Chapter 2, I describe how Visual Cell Sorting can be used to assess the functional impact of hundreds of genetic perturbations on a visually assessed phenotype, nuclear localization. Additionally, I argue that the throughput of Visual Cell Sorting is high enough to facilitate screens on thousands of genetic variants; and I discuss future experiments and analyses that will improve nuclear localization prediction algorithms. In Chapter 3, I describe how Visual Cell Sorting can be used to define the cell states that underlie morphologic changes observed in response to drug treatment. Furthermore, I relate the cell states identified to our current understanding of chemotherapy “pre-resistance”. In Chapter 4, I discuss other work I’ve done to allow genetic screen of the tumor suppressor *PTEN* for functions related to cell growth, DNA damage repair, and cell morphology. Additionally, I discuss why Visual Cell Sorting and other sort-based selection methods for performing genetic screens are likely to be supplanted by single cell phenotype-and-genotype approaches. In the Conclusions chapter, I argue that rather than being used for image-based genetic screens, Visual Cell Sorting is optimally suited to comprehensively profiling the cell states and functions associated with specific cellular morphologies.

Chapter 1: Separation of Cells by Visual Phenotypes with Visual Cell Sorting

Introduction

High content imaging (Boutros *et al*, 2015), *in situ* sequencing methods (Chen *et al*, 2015; Moffitt *et al*, 2016; Emanuel *et al*, 2017; Wang *et al*, 2019; Eng *et al*, 2019; Lee *et al*, 2014, 2015; Feldman *et al*, 2019), and Image-based cell selection methods (Binan *et al*, 2016; Chien *et al*, 2015; Binan *et al*, 2019; Kuo *et al*, 2016; David *et al*, 2017) have revolutionized the investigation of how genetic variants and gene expression programs affect cellular morphology, organization and behavior. These methods differ in fundamental ways, though all can provide information that links cellular and gene function with morphology.

One important application of these methods is visual genetic screening, in which a library of genetic variants is introduced into cells and the effect of each variant on a visual phenotype is quantified. In a classical high content visual genetic screen, each genetic perturbation occupies a separate well. New *in situ* methods, which employ sequencing by repeated hybridization of fluorescent oligo probes (Chen *et al*, 2015; Moffitt *et al*, 2016; Emanuel *et al*, 2017; Eng *et al*, 2019; Wang *et al*, 2019) or direct synthesis (Ke *et al*, 2013; Lee *et al*, 2014, 2015; Feldman *et al*, 2019) to visually read out nucleic acid barcodes, permit hundreds of perturbations to be assessed in a pooled format. For example, multiplexed fluorescent in-situ hybridization was used to assess the effect of 210 CRISPR sgRNAs on RNA localization in ~30,000 cultured human U-2 OS cells (Wang *et al*, 2019); and *in situ* sequencing was used to measure the effect of 963 gene knockouts on the localization of an NFkB reporter at a throughput of ~3 million cells (Feldman *et al*, 2019). Visual phenotyping methods can also dissect non-genetic drivers of phenotypic heterogeneity. Here, characterization of cells with distinct visual phenotypes can reveal different cell states – such as signaling pathway activities and gene expression profiles – that are associated with different cellular morphologies. For example, the photoactivatable marker technology Single-Cell Magneto-Optical Capture was used to isolate cells that successfully resolved ionizing radiation-induced DNA damage foci (Binan *et al*, 2019).

Despite their utility, current methods have limitations (Table 1.1). Some, such as high content imaging, require highly specialized or custom-built hardware. Others, like *in situ* sequencing, employ complex protocols, sophisticated computational pipelines, and expensive dye-based reagents. Methods that mark and sort for individual cells with a photoactivatable protein or compound are simpler and less expensive. However, these methods are either low

throughput (< 1,000 cells per experiment; (Binan *et al*, 2016, 2019; Chien *et al*, 2015; Kuo *et al*, 2016) or lack single-cell specificity (David *et al*, 2017). Furthermore, they cannot investigate more than one or two phenotypes per experiment.

To address these shortcomings, we developed Visual Cell Sorting, a flexible and simple high-throughput method that uses commercial hardware to enable the investigation of cells according to visual phenotype. Visual Cell Sorting is an automated platform that directs a digital micromirror device to mark single live cells that express a nuclear photoactivatable fluorescent protein for subsequent physical separation by fluorescence activated cell sorting (FACS). We demonstrate that Visual Cell Sorting enables visual phenotypic sorting into 4 bins; increases the throughput of cellular separation by 1,000-fold compared to other single cell photoconversion-based technologies (Binan *et al*, 2016, 2019; Chien *et al*, 2015; Kuo *et al*, 2016). Visual Cell Sorting requires simple, inexpensive, and commercially available widefield microscope hardware, routine genetic engineering, and a standard 4-laser FACS instrument to perform. As such, we envision that Visual Cell Sorting can readily be deployed to uncover the relationships between visual cellular phenotypes and their associated internal states, including genotype and gene expression programs.

Methods

General reagents, DNA oligonucleotides and plasmids

Unless otherwise noted, all chemicals were obtained from Sigma and all enzymes were obtained from New England Biolabs (Ipswich, MA). KAPA Hifi 2x Polymerase (Kapa Biosystems; Wilmington, USA; cat. no. KK2601) was used for all cloning and library production steps. E. coli were cultured at 37 °C in Luria broth. All cell culture reagents were purchased from ThermoFisher Scientific (Waltham, MA) unless otherwise noted. HEK 293T cells (ATCC; Manassas, VA; CRL-3216) and U-2 OS cells (ATCC HTB-96), and derivatives thereof were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 100 U/mL penicillin, 0.1 mg/mL streptomycin, and 1 ug/mL doxycycline (Sigma; St. Louis, MO), unless otherwise noted. For Visual Cell Sorting experiments, DMEM without phenol red was used to reduce background fluorescence. Cells were passaged by detachment with trypsin–EDTA 0.25%. All cell lines tested negative for mycoplasma in monthly tests. All synthetic oligonucleotides were obtained from IDT and their sequences can be found in Table EV3. All non-library-related plasmid modifications were performed with Gibson assembly. See the

Appendix Figure 1.1 and the Visual Cell Sorting publication (Hasle *et al.* 2020) for construction of the vectors used.

Cloning Visual Cell Sorting constructs

To create attB-H3-Dendra2, the Dendra2 open reading frame was obtained from Dendra2-Lifeact7 (a gift from Michael Davidson Addgene #54694) and cloned downstream of the H3 open reading frame from mEmerald-H3-23 (a gift from Michael Davidson Addgene #54115) and into the backbone of attB-EGFP-PTEN-IRES-mCherry (Matreyek *et al*, 2017).

To create attB-H3-Dendra2-P2A-H2B-miRFP703, attB-H3-Dendra2 and pH2B-miRFP703 (a gift from Vladislav Verkhusha, Addgene #80001) were combined and a P2A sequence included in the Gibson overhang regions between Dendra2 and miRFP.

To create pLenti-CMV-H3-Dendra2, the H3-Dendra2 reading frame in attB-H3-Dendra2 replaced the open reading frame in pLenti CMV rtTA3 Blast (w756-1), a gift from Eric Campeau (Addgene #26429).

Cell lines

U-2 OS cells (ATCC, HTB-96) expressing the Tet-ON Bxb1 landing pad (U-2 OS AAVS-LP Clone 11) were generated as previously described (Matreyek *et al*, 2017). To create H3-Dendra2- and H3-Dendra2/H2B-miRFP-expressing derivative cell lines, attB-H3-Dendra2 or attB-H3-Dendra2-P2A-H2B-miRFP703 were recombined into U-2 OS AAVS-LP Clone 11 cells, as previously described (Matreyek *et al*, 2017).

Lentivirus production

To produce lentivirus, HEK293T cells were plated in clear plastic 6 well plates (VWR, cat. no. 10062-892) at 4.5e5 cells per well. The next day, cells in each well were transfected with 1,125 ng psPAX2 (a gift from Didier Trono, AddGene #12260), 375 ng pMD2.G (a gift from Didier Trono, AddGene #12259), and 1,500ng of pLenti transfer vector using 6ul of FuGENE6 (Promega, cat. no. E2691) according to manufacturer's instructions. Media was replaced 24 hours after transfection and collected at 48 hours and 72 hours after transfection. Collected media was spun at 1000g for 5 minutes, then the viral supernatant was decanted and filtered using a 0.45um filter (VWR, cat. no. 28145-481). Finally, the virus was concentrated using PEG-it Virus Precipitation Solution (SBI, cat. no. LV810A-1) and stored at -80C.

Visual Cell Sorting: equipment and settings

A Lecia DMi8 Inverted Microscope was outfitted with Adaptive Focus; an Incubator i8 chamber with PeCon TempController 2000-1 and Oko CO_2 regulator set to 5%; a 6-line Lumencor Spectra X Light Engine LED; Semrock multi-band dichroic filters (Spectra Services, Ontario, NY; cat. no. LED-DA-FI-TR-Cy5-4X-A-000, LED-CFP/YFP/mCherry-3X-A-000); BrightLine bandpass emissions filters for DAPI (433/24 nm), GFP (520/35 nm), RFP (600/37 nm), and NIR (680/22 nm); a 20X 0.8 NA apochromatic objective; and a Mosaic3 Digital Micromirror Device affixed to a Mosaic SS 405 nm/1.1 W laser and mapped to an Ixon 888 Ultra EMCCD monochrome camera. The microscope and digital micromirror device were controlled with the Metamorph Advanced Image Acquisition software package (v7.10.1.161; Molecular Devices, San Jose, CA). The image size was ~560 x 495 µm. Image bit depth ranged from 12-16 bits, depending on the brightness of cells in the field of view.

Cells were plated and imaged on glass-bottom, black-walled plates (CellVis, Mountain View, CA; P06-1.5H-N, P24-1.5H-N, P96-1.5H-N) in phenol-red free media at 5% CO_2 and 37 °C using the 20X 0.8 NA objective. ~50-100 cells were imaged per field of view. To image unactivated Dendra2, 474/24 nm excitation and 482/25 nm emission filters were used. To image activated Dendra2, 554/23 nm excitation and 600/37 nm emission filters were used. To image miRFP, 635/18 nm excitation and 680/22 nm emission filters were used. Prior to imaging, the Auto Focus Control system was activated. Metamorph's Plate Acquisition module was used to collect images and run Metamorph jourhnals that analyzed cells and directed their selective photoactivation by the digital micromirror device. For more information about the Metamorph journals used to image and activate cells, see the Visual Cell Sorting publication (Hasle *et al.* 2020).

Visual Cell Sorting: cell preparation, imaging, analysis and photoactivation

An up-to-date version of this protocol can be found at protocols.io (https://www.protocols.io/view/visual-cell-sorting-beigjcbw).

1. 24 to 48 hours before imaging, plate cells onto 6-well glass bottom, black walled plates at a density of 50,000 to 200,000 cells per well.
2. Before imaging, wash cells with 1X DPBS and add complete media without phenol red.
3. Turn on the microscope and incubation chamber, set the CO2 regulator to 5%, and open Metamorph.

4. Place cells in microscope and bring cells into focus. Test imaging conditions (LED power, exposure time, etc.) for the desired channels.
5. Turn on Auto Focus Control. Using the Well Plate Acquire dialog box, image ~25-100 sites of experimental conditions (and controls, if applicable). Initialize a log file to collect phenotypic data. Using the Journal > Loop > Loop Through Images in Directory command, run the analysis journal on the images to collect the desired phenotypic information. The journal must include an "Integrated Morphometry – Measure" or a “Region Measurements” command to add phenotypic information for each cell to the log file. *Note: these specific images will not be used for activation; rather, this analysis serves to ensure that the phenotypes match what one would expect.*
6. Save the imaging conditions used for the Well Plate Acquire dialog box as a state file.
7. Close the log file. Check the distribution of phenotypes in experimental conditions and controls by running custom software (e.g. Python script) with the log file as input.
8. Load the site map. As of Metamorph v7.10.1.161, this must be done by:
    a. Closing Metamorph
    b. Replacing the *htacquir.cfg* file in the Metamorph application Groups > Metamoprh directory with an *htacquir.cfg* file that contains the site map. *htacquir.cfg* files that contain various site maps for 6- and 24- well plates used in our experiments can be found on the GitHub repository under the Metamorph directory.
    c. Reopening Metamorph and reloading the saved state file (load everything except for site map settings). *Note: in Metamorph v7.10.1.161, the site map can be contaminated by extra sites in the top left corner after this operation. Check the “Sites” tab of the Well Plate Acquire dialog box and remove any extra sites by left clicking.*
9. Center the well:
    a. Move the objective to the approximate center of well A1.

b. Under the Well Plate Acquire “Plate” tab, select “Set A1 Center …" > “Set A1 Center to Current”.
c. Under the “Sites” tab, move the objective to the top center site by right clicking.
d. Using the eyepiece and brightfield illumination settings, check whether the objective is centered at the top of the well. If not, manually change the A1 center settings (measured in microns) to move it in the desired direction.
e. Repeat steps (D) and (E) until the top center site of the site map is centered on the top.
f. Re-check that cells are in focus and that Auto Focus Control in “on”. Auto Focus Control can be turned off by the objective moving too far from the plate and hitting the plate holder.

10. Select the wells to be subject to Visual Cell Sorting under the “Plates” tab by left-clicking
11. Select appropriate journals to be run at the Start of Plate, After Imaging, and End of Plate under the “Journals” tab
    a. The “Start of Plate” journals (labelled “startup.jnl” in the GitHub) serve to add a delay to imaging, if necessary; set the 405 nm pulsetimes for the activations; set any phenotypic threshold values (e.g. NC ratios) for activation; etc.
    b. The “After Imaging” journals contain analysis and activation scripts that are performed after each image is taken
    c. The “End of Plate” journals turn off the laser to increase its lifetime
12. OPTIONAL: Re-align the digital micromirror device:
    a. Under Devices > Mosaic Targeted Illumination, click “Update Settings” in the Configuration tab
    b. Follow the instructions to re-calibrate the device
13. OPTIONAL: Run the experiment without the laser on to check that the correct cells are being identified and activated:

a. In the Well Plate Acquire dialog box, hit “Acquire”
b. Watch the first 5-10 sites of imaging, analysis, and marking cells for activation. *In the activation journals associated with this publication, nuclei subject to the three activation states (50, 200, and 800 ms) are outlined in three different colors.*

14. Turn on the laser
15. Hit “Acquire” to begin acquisition, analysis, and activation.

Visual Cell Sorting: FACS on microscope-activated cells

Cells activated on the microscope were analyzed using an LSR II (BD Biosciences; San Jose, CA) or sorted into bins according to their Dendra2 photoactivation state using a FACS Aria III (BD Biosciences). Raw .fcs files and code associated with this work are available on GitHub.

1. Trypsinize cells and resuspend in DPBS supplemented with 1-2% FBS or BSA
2. Make a gate for live cells using an SSC-A vs. FSC-A plot.
3. Within the live cell gate, make a gate for single cells using an FSC-W vs. FSC-A plot.
4. Within the single cell gate, make a gate for Dendra2-positive cells using a FITC-A histogram plot. *In some clonally derived lines, Dendra2 expression will silence over the course of weeks to months. If Dendra2-negative cells exceed 10%, we recommend resorting the population or returning to a lower passage stock.*
5. Create an activated (PE-YG-A) vs. unactivated (FITC-A) Dendra2 scatter plot. Draw gates for the activated populations of interest. *Activated populations will appear as diagonal clouds with higher PE-YG-A signals than a negative control.*
6. Create a ratio (PE-YG-A / FITC-A) histogram. Show the activated populations of interest (defined in Step 5) within the ratio histogram. Create sorting gates for each population.
7. Sort populations of activated cells according to the gates set on the ratio histogram plot.
8. Spin cells for 5 minutes at 300-500xg, then plate cells in warm, complete media.

9. Analyze data using FlowCytometryTools (v0.5.0) in Python (v3.6.5) or flowCore (v1.11.20) in R (v3.6.0).

Selective photoactivation of cells expressing miRFP

U-2 OS AAVS-LP Clone 11 cells with attB-H3-Dendra2 or attB-H3-Dendra2-P2A-H2B-miRFP recombined into the landing pad were counted and mixed in ratios ranging from 0.5% to 50% miRFP-expressing cells, then 40,000 cells of each mixture were seeded into three wells of a 24-well plate. The next day, cells were placed on the microscope and imaged, analyzed, and activated at 661 sites across each well of the plate, covering ~95% of the total well area. At each site, Dendra2 and miRFP were imaged with 2x2 binning; Metamorph's Count Nuclei module was used on the miRFP image to identify miRFP-expressing cells; and a binary with regions corresponding to miRFP-expressing cells was passed to the digital micromirror device, which subsequently activated the cells. Once all sites were imaged, analyzed, and activated, the cells were subject to flow cytometry to assess unactivated Dendra2, activated Dendra2, and miRFP expression. The experiment was repeated two additional times for a total of three replicates. For the Metamorph journals used to analyze and activate cells, see the GitHub repository. For more information about the gating scheme used for this experiment, see Appendix 1.1.

Photoactivation of cells for 0, 50, 200, and 800 milliseconds

U-2 OS AAVS-LP Clone 11 cells with attB-H3-Dendra2-P2A-H2B-miRFP recombined into the landing pad were seeded at 50,000 cells per well in a 6-well glass bottomed plate. The next day, cells were imaged for unactivated Dendra2 and miRFP at 100 sites (10x10 square) and quartiles of total miRFP intensity were measured using Metamorph. Then, cells across 661 sites in two wells were left unactivated or activated for 50ms, 200ms, or 800ms according to the miRFP intensity quartile to which they belonged (Q1 = 0-3803, Q2 = 3804-5839, Q3 = 7396-9674 , Q4 = 9674+). For the Metamorph journals used to analyze and activate cells, see the GitHub repository.

Testing for photoactivation-induced toxicity with Annexin V and DAPI

U-2 OS AAVS-LP Clone 11 cells with attB-H3-Dendra2 recombined into the landing pad were seeded at 20,000 cells per well in a 24-well plate. Over the next two days, cells across 400 sites (60% well coverage) in three replicate wells were segmented using the Count Nuclei module in Metamorph and activated for 800 ms. Forty-eight hours after the first well was activated, cells were trypsinized, stained with Annexin V (Thermo, cat. no. A23204) and DAPI

(Invitrogen, cat. no. D1306), and subjected to flow cytometry to assess unactivated Dendra2, activated Dendra2, Annexin V, and DAPI. Three wells of unactivated cells were heated at 50 °C for 10 minutes as a cell death positive control. The experiment was repeated two additional times for a total of three replicates. Data was analyzed using FlowJo (v10.5.3).

Testing for photoactivation-induced toxicity with RNA sequencing

U-2 OS AAVS-LP Clone 11 cells with attB-H3-Dendra2 recombined into the landing pad were seeded at 20,000 cells per well in 8 wells of a 24-well plate. Eighteen hours later, cells across 6 wells (678 sites per well; ~100% well coverage) were activated and then incubated for 0.5, 1.5, 2.5, 3.5, 4.5, or 6 hours (1 well each). Two wells were left unactivated. Dendra2 photoactivation was verified by flow cytometry, with the two unactivated samples were used as negative controls. Bulk RNA sequencing libraries were prepared as described previously (Cao et. al. 2017). Briefly, RNA was extracted from each sample using a Trizol/RNeasy Mini Kit (ThermoFisher, cat. no. 15596026, Qiagen; Germantown, MD; cat. no. 74104) then subjected to SuperScript IV First-Strand Synthesis (Thermo Fisher 18091050) and NEBNext Ultra II Directional RNA Second Strand Synthesis (NEB E7550), according to the manufacturer's instructions. cDNA was then tagmented with Nextera Tn5 (Illumina; San Diego, CA; FC-131-1024) and amplified/indexed by PCR with the NEBNext DNA Library Prep Kit (NEB E6040). Samples were sequenced using a NextSeq 500/550 75 cycle kit (Illumina, cat. no. TG-160-2005). Differential gene expression analysis of RNA sequencing data followed the standard DESeq2 workflow (Love *et al*, 2014). Briefly, differential gene expression testing was performed using a binary coding of photoactivation status in the DESeq2 design formula. Dispersion e stimates, log2 fold changes and adjusted p-values were all calculated using the DESeq () function with default parameters as specified in DESeq2.

Results

Visual Cell Sorting uses FACS to separate hundreds of thousands of cells by their visual phenotypes. Cells are first modified to express Dendra2, a green-to-red photoconvertible fluorescent protein (Chudakov *et al*, 2007) that will act as a phenotypic marker and enable downstream FACS sorting. Next, cells are imaged on an automated microscope. In each field of view, cells are identified and analyzed for phenotypes of interest. According to their phenotype, cells are illuminated with 405 nm light for different lengths of time using a digital micromirror device, resulting in different levels of red Dendra2 fluorescence (Figure 1.1A, Figure 1.2A). The

imaging, analysis, and photoactivation steps are performed at each field of view; and unlike previous photoactivatable marker-based methods, these steps are automated, allowing hundreds of thousands of cells to be assessed per experiment. Once all cells have been imaged, analyzed, and photoactivated, FACS is used to sort them into bins according to their level of Dendra2 photoactivation (Figure 1.1A).

We first sought to establish the single cell accuracy of Dendra2 photoactivation, and whether variable photoactivation states could be discerned by flow cytometry. We noticed that similar technologies use photoactivatable dyes or proteins localized to the whole cell body (Binan *et al*, 2016, 2019; Chien *et al*, 2015; Kuo *et al*, 2016). This localization strategy makes identifying the boundaries of the fluorescent signal difficult, which results in partial photoactivation or photoactivation of the marker in a cell adjacent to a cell of interest. With this in mind, we expressed Dendra2 in the nucleus either as a histone H3 fusion (H3-Dendra2) or with an upstream nuclear localization sequence (NLS-Dendra2x3). The boundaries of nuclear Dendra2 signal are easy to identify, permitting quantitative photoactivation of Dendra2 in the cells of interest; and the cytoplasm provides a spacer between the Dendra2 in different cells, reducing photoactivation of cells adjacent to the cells of interest.

To measure photoactivation accuracy, H3-Dendra2 positive cells co-expressing H2B-miRFP (Shcherbakova *et al*, 2016) were mixed with cells expressing H3-Dendra2 alone at decreasing ratios. We instructed the microscope to activate Dendra2 in cells harboring miRFP-positive nuclei, and then we quantified the co-occurrence of miRFP and activated Dendra2 florescence signals using flow cytometry (Figure 1.1B). The ratio of activated Dendra2 fluorescence to unactivated Dendra2 fluorescence (Dendra2 photoactivation ratio) accurately predicted whether a cell was miRFP-positive, even when the miRFP expressing cells were present at ~0.5% frequency, with average precision of 94% and recall of 80% (Figure 1.1C).

Previous photoactivatable marker-based methods have been limited to two photoactivation levels: activated and unactivated. To test whether we could encode more than one photoactivation level, and thus more than one phenotype, we exposed different cells in the same well to 405 nm light for 0, 50, 200, or 800 ms. Flow cytometry of the Dendra2 fluorescence distribution by showed four distinct levels of Dendra2 photoactivation, indicating that Visual Cell Sorting can sort four different visual phenotypes or four discrete bins of a continuous phenotype (Figure 1.1D). Furthermore, these four photoactivation levels can still be distinguished over 12 hours following activation (Figure 1.2B, left panel). To extend the amount of time that the photoactivation levels remain distinct from one another, we placed H3-Dendra2

expression under the control of a doxycycline-inducible promoter. By shutting off Dendra2 expression before the experiment, the 50, 200, and 800 ms photoactivation levels remained distinguishable for up to 24 hours (Figure 1.2B, right panel). Finally, we examined the effect of Dendra2 photoactivation on cell viability and function. Activated cells did not exhibit higher rates of apoptosis or cell death even two days after photoactivation, nor did we detect effects of photoactivation on gene expression (Figure 1.2C, D). These results indicate that Dendra2 photoactivation does not appreciably affect cell survival or gene expression programs.

Discussion

A major limitation of current microscopy-based experiments is the inability to isolate hundreds of thousands of phenotypically defined cells for further analysis. We developed Visual Cell Sorting, a microscope-based method that directs a digital micromirror device to irreversibly photoactivate a genetically encoded fluorescent protein in cells of interest, effectively translating a complex visual phenotype into one that can be sorted by FACS. In turn, this permits the characterization of a morphologic phenotype using any assay that can be applied to hundreds of thousands of live cells.

High throughput is a key advantage of Visual Cell Sorting, compared to other image-based cell selection methods. In our pooled image-based screen. The throughput is ~1,000-fold more than what could be achieved using other image-based selection methods. The high throughput affords Visual Cell Sorting with extra flexibility with respect to downstream experiments. For example, it enables the analysis of over one thousand genetic variants in a single experiment, as well as downstream applications that require tens of thousands of cells, including transcriptomic and proteomic workflows. By contrast, other image-based selection methods are limited to DNA sequences on clonal derivatives of sorted cells, low-throughput single cell RNA sequencing workflows (which suffer from low statistical power), and other experiments that test for the heritability of certain cellular phenotypes (such as DNA damage response; see Binan et al. 2019). Visual Cell Sorting throughput could be increased even further by analyzing cellular phenotypes at a lower magnification, by applying faster image analysis algorithms, or by shutting off Dendra2 expression before imaging to extend imaging time (Figure 1.2B).

A second key advantage afforded by Visual Cell Sorting is its photoactivation tunability, which permits recovery of up to four distinct cellular phenotypes in one experiment. This is in stark contrast to the other image-based cell selection methods, which have two activation states

(activated and unactivated), and which cannot make use of the unactivated sort bin because they do not have the throughput to image all cells in a well (Chien *et al*, 2015; Kuo *et al*, 2016; Binan *et al*, 2019). I believe the reason for Visual Cell Sorting's tunability has to do with the localization of the photoactivatable moiety, Dendra2, to the nucleus, and is discussed in the Appendix section titled, "Nuclear Dendra2 and Photoactivation Tunability".

Visual Cell Sorting's throughput also compares favorably to the *in situ* methods, especially those that seek to derive image-correlated single cell transcriptomes. As *in situ* experiments take much longer than Visual Cell Sorting experiments, it is fairest to compare per-day throughput rather than per-experiment throughput. A recent *in situ* image-based genetic screening method published by the Blainey lab showed impressive throughput, with 3 million tightly packed cells imaged and barcodes sequenced in 5 days (Feldman *et al*, 2019). Visual Cell Sorting has ~50% lower throughput in comparison: it would take approximately five Visual Cell Sorting experiments (each taking one day) to activate and sort for 1 to 3 million cells (depending on image analysis time).

Visual Cell sorting outperforms *in situ* methods that seek to derive single cell transcriptomes and connect them to visual or spatial phenotypes. Such *in situ* experiments take approximately 16 days to perform an experiment on hundreds of cells (Table 1.1). Combining the sorted cells output by Visual Cell Sorting with highly parallelized single cell RNA sequencing platforms allows morphology-labelled transcriptomes of *thousands* of cells to be derived in as little as three days. This corresponds to a per-day throughput that is ~50 fold higher. Therefore, Visual Cell Sorting performs particularly well when a complex molecular readout of cell state (e.g. transcriptomics) is desired, rather than a readout that can be captured with a barcode (e.g. a genetic variant).

A second key advantage of Visual Cell Sorting is that, in contrast to *in situ* methods, it does not require any expensive dye-based reagents such as oligo libraries or fluorescent-labelled oligos; customized hardware components; or complex workflows. Outfitting an automated wide-field microscope for Visual Cell Sorting requires just three inexpensive, commercially available components: a live cell incubation chamber, a digital micromirror device, and a 405 nm laser. *In situ* sequencing, demands ~$2,000 dollars of reagents (e.g. enzymes, fluorescent oligonucleotides) for a variant sequencing experiment on one million cells (Feldman *et al*, 2019); Visual Cell Sorting, by contrast, requires approximately $600.

Visual Cell Sorting suffers from a number of disadvantages, relative to *in situ* methods. Though Visual Cell Sorting can be combined with a great variety of downstream assays, it (and the other image-based cell selection methods) requires a pre-defined phenotype of interest. The method is therefore better suited for the biologist who is interested in a readily defined visual phenotype or cellular morphology, rather than one who seeks to understand general relationships between cellular structures (e.g. various microtubule configurations) and cellular state or genetic variants. By contrast, *in situ* methods can examine many different phenotypes in the same experiment, because the images can be re-analyzed for new phenotypes and correlated with the readout (i.e. the barcode or the transcriptome).

Unlike the *in situ* methods, cells in a Visual Cell Sorting experiment must be genetically engineered to express Dendra2, which is photoactivated by blue fluorescent protein (BFP) excitation wavelengths and emits at GFP and RFP wavelengths. This requirement limits the other fluorescent channels are available for imaging. However, miRFP (Shcherbakova *et al*, 2016) and mBeRFP (Yang *et al*, 2013) can be used in conjunction with Dendra2, allowing two additional compartments or proteins to be marked in each experiment. Moreover, new analytical approaches leveraging brightfield images may reduce the need for fluorescent markers (Ounkomol *et al*, 2018; Christiansen *et al*, 2018).

Visual Cell Sorting experiments are also limited to approximately twelve hours of imaging to avoid Dendra2 activation signal decay or cell overgrowth. The hours timescale required to execute a Visual Cell Sorting experiment makes it challenging to study transient phenotypes (e.g. cell-cycle dependent phenotypes) and to study behavioral phenotypes that require long time-lapse imaging and analysis times (e.g. neuronal firing patterns). Furthermore, decay of photoactivated Dendra2 may be more pronounced in rapidly dividing bacterial or yeast as activated Dendra2 is diluted by cell division. However, the workflow we present, with imaging at 20X magnification and image processing times of 3-8 seconds, is sufficient for the analysis of hundreds of thousands of human cells in one experiment.

Finally, Visual Cell Sorting, as described here, cannot be applied to whole tissues, as it requires live cells that can be imaged, segmented, and trypsinized into single cell suspensions. However, work by Hyeon-Jin Kim and Sriram Pendyala in the lab has shown the Visual Cell Sorting can be applied to fixed cells and perhaps even to fixed tissues.

Chapter 2: Visual Cell Sorting to Examine Sequence-Function Relationships

Introduction

With the Visual Cell Sorting workflow validated, we applied it as a selection method in a genetic screen. Throughout the 20th century, genetic screens were performed by treating whole organisms with a mutagen, looking for visual phenotypes of interest, and then mapping the phenotype to a genetic locus using genetic crosses. In the 21st century, with the advent of RNAi, CRISPR-Cas9, and other technologies that allowed genetic perturbations to be parallelized, reverse genetic screens have become popular (Morgens *et al*, 2016). These screens are performed by introducing hundreds to thousands of genetic perturbations in hundreds of thousands to millions of cells in a single experiment. At the outset of my graduate work, reverse genetic screens either examined a visual phenotype in microwell plates (i.e. high content imaging; Boutros *et al*, 2015) or examined a growth- or fluorescence-based phenotype a pooled format (Sanjana, 2017). Critically, no methods existed that allowed pooled, selection-based screening of a visual phenotype.

Pooled, image-based genetic screens offer a few key advantages over high content imaging screens. Firstly, they can leverage genetic libraries, which are introduced into a population of cells simultaneously. High content screens, by contrast, require preparing each genetic perturbation individually so that all the cells in each well receive the same perturbation. Secondly, pooled screens minimize well- and plate-based batch effects, where variables such as variation in seeding density and sample processing time can affect the visual phenotype of interest. Thirdly, they do not require the same robotics hardware as high content imaging screens (Boutros *et al*, 2015). With these advantages in mind, we set out to use Visual Cell Sorting as a selection tool for a pooled, image-based genetic screen for protein nuclear localization.

Nuclear localization sequences (NLS's) are short peptides that direct proteins to the nucleus, and NLS's are critical for the function of thousands of human transcription factors, nuclear structural proteins, and chromatin modifying enzymes. Many nuclear proteins lack an annotated NLS, and current NLS prediction algorithms cannot sensitively identify known NLS's without drastically decreasing their precision (Lin & Hu, 2013; Nguyen Ba *et al*, 2009). This shortcoming may arise because these NLS prediction algorithms rely on sequence alignments or amino acid frequencies of naturally observed NLS's, which are subject to discovery bias.

Unbiased assessment of nuclear localization sequence function is therefore needed to better identify novel NLS's and understand how they function.

The SV40 NLS was the first short peptide sequence discovered to enact nuclear import of protein cargos (Kalderon *et al*, 1984). It is a "monopartite" NLS, which contains a short lysine- and arginine- rich (KR-rich) central region that can bind to either the major or minor groove of the nuclear import protein importin alpha. Though low-throughput, single variant experiments have demonstrated that the KR-rich region is critical for NLS function (Hodel *et al*, 2001), it is not known how an arbitrary mutation affects the function of this NLS. Quantitative scores for every possible single amino acid substitution in the SV40 NLS would improve our understanding of NLS sequence-function relationships and improve monopartite NLS prediction models. Therefore, we used Visual Cell Sorting to evaluate a large library of NLS missense variants in a pooled, image-based genetic screen.

Methods

Cloning constructs and individual variants

To create attB-Nterm-CMPK-miRFP (the destination vector for the NLS library), a gBlock encoding an EcoRI site in-frame and upstream of CMPK (IDT; based off a previously published SV40 NLS construct (Kalderon *et al*, 1984)) was combined with the miRFP open reading frame from pH2B-miRFP703 and inserted into the backbone of attB-H3-Dendra2-P2A-H2B-miRFP703.

To create attB-NLS-CMPK-miRFP and all single, double, and triple amino acid variants, the attB-Nterm-CMPK-miRFP vector was digested with EcoRI for 2 hours at 37 C. Then, the digested plasmid and an oligo that contained the NLS (wild-type or variant of interest) and 55 C overhangs complementary to the edges of the cut site were incubated in a Gibson reaction in a one to three molar ratio and transformed, as per manufacturer's instructions.

Cloning the library of SV40 NLS variants

The library of all possible SV40 NLS missense variants was constructed using a Gibson cloning approach. Eleven primer pairs – 1 for each NLS codon, plus 2 codons upstream and 2 codons downstream of the NLS – were designed (The Reagents and Tools Table). For each pair, the forward primer contained a 3' annealing region (Tm ~55C), an NNK codon, and a 5' Gibson homology region (Tm ~55C). The reverse primer comprised of the reverse complement of the forward primer Gibson homology region. Each primer pair was used in a separate PCR reaction that included attB-NLS-CMPK-miRFP as the template, and 5ul of each reaction were

run on a 1% gel to check for product. The remaining 20ul was DpnI digested for 2 hours at 37 C to remove template plasmid, cleaned using DNA Clean & Concentrator-5 (Zymo Research D4013), subject to a 1-piece Gibson reaction, and transformed into chemically competent *E. coli*. Bulk transformant cultures were grown overnight and harvested using GenElute HP Plasmid DNA Midiprep Kit (Sigma, NA0200-1KT). DNA preps containing single codon variant were subsequently mixed such that each prep contributed an equal amount of DNA. The final library contained 346 NNK nucleotide variants which, due to codon degeneracy in the genetic code, encode for 209 single amino acid variants.

Recombination of single-variant SV40 NLS clones or the library into U-2 OS LLP-Blast/H3-Dendra2 Clone 4 cells

The SV40 NLS variant library or single-variant clones were recombined into U-2 OS LLP-Blast/H3-Dendra2 Clone 4 cells, as previously described in HEK 293Ts (Matreyek *et al*, 2017). Two recombination replicates were performed. To recombine NLS variants or the NLS library into cells, H3-Dendra2 expressing U-2 OS cells with the landing pad were subject to Lipofectamine 3000 (Thermo Fisher L3000015) transfections in 6 well plates, T-25 flasks, or T-75 flasks, according to manufacturer instructions, with the following specifications: plated cells at 0.1e5 cells/well (24 well plate), 0.6e5 cells/well (6 well plate), 1.4e6 cells/flask (T-25), or 4.2e6/flask (T-75); transfected with 0.75ul/3.75ul/10.4ul/31.2ul Lipofectamine 3000, 1ul/5ul/13.9ul/41.7ul P3000 reagent, 500ng/2500ng/7000ng/21000ng total DNA at a by-weight ratio of 1/3 pCAG Bxb1 and 2/3 attB plasmid(s). Cells were transfected immediately after plating. Twenty-four hours after transfection, media was replaced. Doxycycline was added 48h after transfection. BFP negative, miRFP positive, Dendra2 positive cells were sorted 5-8 days after transfection.

Visual Cell Sorting of cells expressing SV40 NLS library

Eighteen hours before imaging, 300,000 U-2 OS LLP-Blast/H3-Dendra2 Clone 4 cells with the attB-NLS-CMPK-miRFP library recombined into the landing pad were seeded into each well of a 6-well plate. The next day, cells were placed onto the microscope and imaged, analyzed, and activated across 2,949 sites (~100% well coverage) across two wells. At each site, Dendra2 and miRFP were imaged with 2x2 binning; Metamorph's Count Nuclei module was used on the Dendra2 image to identify nuclei and create a nuclear binary image; cytoplasm binaries were created by subjecting the nuclear binary to a dilate function and subtracting away the nuclear binary; each nucleus-cytoplasm binary pair was superimposed on the miRFP image and average pixel intensities were measured for each compartment; cells with an average nuclear or

cytoplasmic miRFP pixel intensity of less than 11,000 were filtered out; a nucleus-to-cytoplasm (N:C) ratio was calculated by dividing the average nuclear pixel intensity by the average cytoplasmic pixel intensity; nuclei with N:C < 0.964 were not activated at all, N:C 0.964 – 1.079 were activated for 50 ms, N:C 1.079 – 1.244 were activated for 200 ms, and N:C > 1.244 were activated for 800 ms. Once all sites were imaged, analyzed, and activated, the cells were subject to FACS and unactivated Dendra2 (FITC), activated Dendra2 (PE-YG), and miRFP (AlexaFluor-700) fluorescence intensities assessed. Cells were then sorted into four photoactivation bins (Figure 2B). A total of two Visual Cell Sorting technical replicates were performed on recombination replicate 1, and three were performed on recombination replicate 2. The details of replicate sorts for the NLS library can be found in Table 2.1. For an example of the gating scheme, see Appendix Figure 2.1.

Sorted SV40 NLS library genomic DNA preparation and sequencing

After sorting, cells in each Dendra2 photoactivation bin were grown in the absence of doxycycline until confluent in one well of a 6-well plate (~ 7 days), then pelleted and stored at -20 °C. DNA was extracted from cell pellets with the DNEasy kit (Qiagen, cat. no. 69504) using RNAse according to the manufacturer's instructions. gDNA was amplified using SV40_NLS_seq_f and SV40_NLS_seq_r (The Reagents and Tools Table) primers using Kapa Hifi (Kapa Biosystems, cat. no. KK2602) according to the manufacturer's instructions. Amplicons were cleaned using Ampure XP beads (Beckman Coulter; Brea, CA; cat. no. A63880), then subjected to an indexing PCR step using KAPA2G Robust (Kapa Biosystems, cat. no. KK5705) with primers P5 and an indexing primer (The Reagents and Tools Table). Amplicons were then run on a 1.5% agarose gel at 130 V for 40 min and the DNA in the 235bp band extracted using Freeze'N Squeeze DNA Gel Extraction Spin Columns (BioRad, cat. no. 7326165). Extracted DNA was sequenced on an Illumina NextSeq500 using SV40_NLS_Read1, SV40_NLS_Read2, and SV40_NLS_Index1 primers (The Reagents and Tools Table). Reads were trimmed and merged using PEAR (Zhang *et al*, 2014a). Sequences were quality-filtered and variants were called and counted by using Enrich2, as previously described (Rubin *et al*, 2017). The Enrich2 configuration file is available on the Fowler lab GitHub repository.

Calculating NLS variant localization scores

Jupyter v5.5.0 running Python v3.6.5 was used for analyses of the Enrich2 output. First, two filters were applied to remove low-quality variants: (1) a minimum nucleotide variant count cutoff of 5 in each bin in each replicate and (2) a requirement that the variant was accessible via NNK

codon mutagenesis. After filtering, remaining nucleotide variants encoding the same amino acid substitution were added to yield a sum of counts for that variant within each bin for each replicate. To generate raw quantitative scores (S_{raw}), a weighted average approach as previously described (Matreyek *et al*, 2018) was applied to the variant frequencies (f_{var}) across the 4 bins (*b1* – *b4*) in each replicate:

$$S_{raw} = \frac{0.25\ (f_{var_{b1}}) + 0.50\ (f_{var_{b2}}) + 0.75\ (f_{var_{b3}}) + f_{var_{b4}}}{f_{var_{b1}} + f_{var_{b2}} + f_{var_{b3}} + f_{var_{b4}}}$$

Raw scores were subsequently normalized such that variants with a wild-type raw score (S_{WT}) have a normalized score of 1 and variants with the median raw score of the bottom 10% of variants (S_{P10}) have a normalized score of 0:

$$S_{norm} = \frac{S_{raw} - median\ (S_{P10})}{S_{WT}}$$

A final round of frequency filtering for variants, which sought to increase score correlations without excluding too many variants, removed variants present at a frequency lower than 0.003% of reads in all bins. Then, the raw and normalized scores were recalculated for each replicate; and the mean and standard error of the normalized scores from the five replicates were calculated to produce final scores. An iPython notebook file with the code used to run the analysis is available on the Fowler lab GitHub repository.

Results

We used Visual Cell Sorting to evaluate a large library of NLS missense variants; sort cells according to the NLS function; and sequence the sorted cells (Figure 2.1A), with the hypothesis that the resulting data could be used to improve NLS prediction. To assess SV40 NLS variant function in mammalian cells, we constructed a fluorescent nuclear localization reporter similar to one described previously (Kalderon *et al*, 1984). Cultured U-2 OS H3-Dendra2 cells expressing the wild-type SV40 NLS fused to a CMPK-miRFP reporter had high levels of miRFP in the nucleus, relative to the cytoplasm. The degree of nuclear localization was calculated using a nucleus-to-cytoplasm miRFP intensity ratio (N:C ratio; Figure 2.2A). In contrast to the wild-type SV40 NLS-tagged reporter, cells expressing an untagged reporter had a low nucleus-to-cytoplasm ratio (Figure 2.1B).

We generated a library of 346 NLS nucleotide variants, corresponding to all possible 209 single amino acid missense variants. Cells expressing the library had a bimodal nucleus-to-cytoplasm ratio distribution, indicating that some variants preserved reporter nuclear localization while others disrupted its localization to different degrees (Figure 2.1B). We divided the library into four photoactivation levels spanning the nucleus-to-cytoplasm ratio range and used Visual Cell Sorting to sort cells into four bins (Figure 2.1B, dotted lines). A total of 637,605 cells were sorted across 5 replicates (Table 2.1). Microscopy on the sorted cells revealed that Visual Cell Sorting faithfully separated cells by the nuclear localization phenotype (Figure 2.1B, C). Deep sequencing revealed the frequency of each variant in every bin, and we used these frequencies to compute a quantitative nuclear localization score for 97% of the 209 possible single missense variants (see Fowler Lab GitHub; Rubin *et al*, 2017). Scores were subsequently normalized such that wild-type had a normalized score of 1 and the bottom 10% of scoring variants had a median normalized score of 0.

As expected, nuclear localization scores for synonymous variants were close to a wild-type-like score of one, and most missense scores were lower than one, indicating loss of nuclear localization sequence function (Figure 2.1C). Furthermore, the SV40 NLS was most sensitive to substitutions in its K/R motif (Figure 2.1D). Localization scores were reproducible (mean r = 0.73; Figure 2.2D), and individually assessed nucleus-to-cytoplasm ratios were highly correlated to the localization scores derived using Visual Cell Sorting (r^2 = 0.91; Figure 2.3A). Localization scores of individual variants were correlated with previously reported *in vitro* K_d values for binding to importin alpha (r = -0.76, Figure 2.4A). Thus, Visual Cell Sorting accurately quantified the effect of NLS variants on their nuclear localization function.

We mapped the median score at each position, which serves as a metric of position sensitivity to mutations, to the structure of the SV40 NLS. Residues that formed tight contacts (e.g. positions 5 through 8) with importin alpha had the lowest median score values. By contrast, residues whose R chains pointed away from importin alpha were less affected by mutations (e.g. positions 4 and 9). We noticed that a mutation to proline at position 2 of our library greatly enhances NLS function. Proline exists in the analogous position in the SV40 large T antigen, but it was not originally reported to be part of the SV40 NLS (Kalderon *et al*, 1984) so it was not part of the wild-type sequence in our construct. Examination of the crystal structure shows that this proline fits nicely into a pocket of the SV40 NLS (Figure 2.5). Therefore, we validated previous reports that the presence of a proline upstream of the K/R-rich region may be an important determinant of monopartite NLS function (Fontes *et al*, 2000).

The SV40 NLS is commonly used to localize recombinant proteins to the nucleus and is included in over 10% of all constructs deposited in AddGene (accessed June 2019). Thus, an optimized NLS could improve a wide range of experiments including CRISPR-mediated genome editing. We further investigated three variants that appreciably increased nuclear localization of the reporter compared to the wild-type SV40 NLS. Individually, these variants modestly improved nuclear localization, and a “superNLS” with three missense variants increased nuclear localization by 2.3-fold (Figure 2.3B, C).

Most NLS prediction algorithms use naturally occurring, individually validated NLS sequences to identify similar sequences in new proteins. By contrast, our data comprise a comprehensive set of NLS-like sequences with variable function. We trained a linear regression model to predict whether any given 11-mer functions as a monopartite NLS by using the experimentally-determined amino acid preferences (Bloom, 2014) at each NLS position, which were calculated with the localization score data. We evaluated our model using a test dataset, not used for training, of 30 NLS’s in 20 proteins. The resulting model more accurately predicted NLS’s than two previously-published linear motif scoring models (Lin & Hu, 2013; Nguyen Ba *et al*, 2009), particularly at a stringency where the majority of NLS’s are detected (Figure 2.3D). We used our model to annotate NLS’s in nuclear human proteins (Thul *et al*, 2017) according to two score thresholds: one for high confidence monopartite NLS (precision 0.88, recall 0.23) and one for candidate monopartite NLS’s (precision 0.51, recall 0.76). In total, we annotated 3,068 high confidence monopartite NLS’s and an additional 30,814 candidate monopartite NLS’s across 11,796 human nuclear proteins (available on the Fowler lab Github).

To substantiate that these represent bona-fide NLS sequences, we compared the top-scoring 11-mers in exclusively nuclear proteins to those in exclusively cytoplasmic proteins (Figure 2.4B, C). As expected, nuclear proteins had higher top-scoring 11mer sequences than cytoplasmic proteins (Wilcoxon rank sum p-value < 10^{-16}). Twenty-eight percent of the nucleus-only proteins contained an 11-mer with an NLS score higher than our high-confidence cutoff; only 11% of cytoplasmic proteins contained such a sequence. These results are consistent with our predictor identifying monopartite, SV40-like NLS’s in the human proteome.

Discussion

Here, we leveraged its high throughput to quantify the function of hundreds of nuclear localization sequence variants in a pooled, image-based genetic screen. We validated that the

K/R-rich region is critical for NLS function, and additionally validated that proline residues at the N-terminal side of the monopartite NLS improve function (Fontes *et al*, 2000; Conti *et al*, 1998). By combining single variants that individually improved NLS function, we created an eleven-residue superNLS (EPPRKKRKIGI) that could be used to improve CRISPR-mediated genome editing, fluorescent protein-based nuclear labelling, and other experiments that leverage nuclear recombinant proteins. We then used the variant scores to make an accurate, amino acid preference-based predictor of NLS function, which we applied to the human nuclear proteome and validated by comparing the top-scoring sequences between cytoplasmic and nuclear proteins. Interestingly, some cytoplasmic proteins contain putative NLS's, which could be explained by an NLS that becomes accessible to the nuclear import machinery after a signaling event (Beg *et al*, 1992) or a nuclear export signal located on the same protein that overwhelms an otherwise functional NLS (Marchand *et al*, 2019). Nuclear proteins without high-scoring sequences may harbor a non-SV40 type NLS or have an interaction partner with a functional NLS enables co-import into the nucleus.

Our NLS prediction model comes with substantial caveats. Firstly, applying the amino acid preferences from the SV40 NLS to all possible nuclear human proteins is unlikely to detect NLS's with a different import mechanism, such as bipartite NLS's PY-NLS's. Second, NLS's with exceptionally high binding affinities for importin alpha (which is essentially what our model predicts, given the large CMPK-miRFP cargo in our experiment) may not be selected for in human proteins; indeed, such proteins may interfere with the nuclear import machinery by, for example, resisting release from importin alpha in the nucleoplasm. Third, the model's amino acid preference inputs do not account for epistatic interactions that occur between positions.

To address these caveats, multiple approaches could be taken. Firstly, it could be extended to explicitly look for bipartite NLS's, which contain two monopartite NLS-like sequences separated by a spacer. One simple way this could be done is by looking for two sequences that have modestly high NLS prediction scores and are within 20 residues of one another. Secondly, a profile hidden Markov model (Yoon, 2009) could be used rather than the linear motif scoring model described. In such a model, each position's "match" state would be a continuous variable represented by the preference score of the amino acid at that position, thereby avoiding the standard model requirement that a sequence alignment of NLS's be made (which is a challenging endeavor). The profile hidden Markov would also be able to provide "weights" to each position by permitting an insertion and deletion at positions that do not contribute to NLS function as heavily. Thirdly, performing a screen that includes double mutations will be important

for detecting any epistatic interactions and modeling them appropriately. Lastly, additional pooled image-based screening experiments on other NLS's could be performed to update the model. It would be particularly informative to perform experiments on NLS's that bind karyopherin beta rather than importin alpha such as PY-NLS's (Soniat & Chook, 2015).

Whether Visual Cell Sorting could be performed on a larger library is a critical question, considering that most pooled screens seek to examine thousands rather than hundreds of genetic perturbations. To answer this question, one can examine the replicate correlations of the NLS experiment, as we know that the error of our score estimates will increase with increased library complexity. Unfortunately, although the NLS library was very small (covering a mere 11 positions, compared to hundreds of positions in similar libraries), our replicate correlation remained modest (mean Pearson's $r = 0.73$). However, the DMD was found to be slightly out of alignment by ~5 microns after the NLS experiments were completed, which would lead to inaccurate binning of variants (primarily partial activation of cells). The misalignment of the DMD also explains two other observations made during the experiment: namely, that the activation peaks from this experiment are not as discrete from one another as the peaks in the nuclear shape experiments discussed in Chapter 3 (compare Appendix Figures 2.1 and 3.1); and that many more cells were in lower activation bins than higher activation bins (see Appendix Figure 2.1). Therefore, I am hopeful that with proper DMD alignment, the Pearson's *r* of the repeated experiment would be substantially higher. Other factors that may have contributed to the low Pearson's *r* are errors associated with the analysis pipeline (e.g. segmentation errors) and genetic drift caused by variable cell viability after sorting.

Considering that we could fix the DMD alignment issue that was likely present in the first pooled screen experiment, how much larger could the libraries subject to Visual Cell Sorting become? To answer this question, we ought to look at the throughput of a typical 12-hour Visual Cell Sorting experiment and how it might be improved. In the NLS experiment, we analyzed approximately one million cultured human cells across 60 hours of imaging and sorting time, ultimately recovering ~650,000. That corresponds to a throughput of over 130,000 cells per experiment. Notably, the analysis pipeline for the NLS workflow is rather slow, at about six seconds per image. Using the neural net-based image analysis pipeline developed by Sriram, we could likely bring image analysis down to three seconds (while increasing accuracy), thereby increasing throughput by a factor of 1.5x (imaging and activation take approximately 3 seconds). Overall, then, the per replicate throughput could be approximately 200,000 sorted cells per replicate, a value on par with the sorted cell throughput that was performed using a FACS-

based, pooled genetic screen on thousands of PTEN variants (Matreyek *et al*, 2018). Therefore, it seems plausible that libraries as large as several thousand variants could be subject to a Visual Cell Sorting-based selection.

In conclusion, this chapter demonstrates that Visual Cell Sorting can be used to perform pooled, image-based genetic screens of hundreds to thousands of genetic variants. An immediate next step would be to perform additional NLS screens and improve the prediction algorithm. Other directions involve examining more complex subcellular localization phenotypes (e.g. subcellular localization of LMNA variants; West *et al*, 2016) or morphologic phenotypes (e.g. morphologic changes associated with PTEN expression; Tamura *et al*, 1998). However, as is discussed in the Chapter 4 and the Conclusions chapter, image-based genetic screens will likely be best performed using *in situ* methods, especially if measurement of multiple phenotypic parameters per variant is desired.

Chapter 3: Visual Cell Sorting to Dissect Complex Drug Responses

Introduction

To understand the responses of cells to perturbations, biologists typically compare a cellular function (e.g. growth, action potential) or molecular phenotype (e.g. protein abundance, post translational modification) between exposed and unexposed cells. Many of the gold standard methods used to make these measurements were developed decades ago and work well for studying partly characterized drugs or genes. However, if a drug's mechanism of action or a gene's function is unknown, it can be challenging to choose which phenotype or molecule to measure.

In the past 20 years, new technologies such as RNA sequencing, proteomics, and Cell Painting (Rohban *et al*, 2017) have offered a less biased assessment of drug mechanism of action or gene function. In these experiments, the transcriptome, proteome, or morphological profile of cells provides a rich and comprehensive readout of cellular state. By comparing the cell states readouts of cells subject to a perturbation to those of untreated cells, researchers can generate hypotheses regarding the perturbation's mechanism of action and how it impacts cellular function (Yang *et al*, 2020). Gold-standard, low-throughput methods such as Western blots and patch clamp experiments can then be applied to test the hypotheses.

Critically, these unbiased methods struggle to decode intra-perturbation morphologic heterogeneity, which has been observed in many model systems. For example, many genetic variants induce visual phenotypes in eukaryotic cells with incomplete penetrance (Mattiazzi Usaj *et al*, 2020), and clonal cells with different morphologies have been observed to have different drug sensitivities (Gupta *et al*, 2011). In both cases, we have a poor understanding of why the divergent morphologic responses exist and what they tell us about cellular homeostasis. Unfortunately, Cell Painting analysis pipelines compute summary statistics of the morphological profiles measured across a well, thereby precluding the study of heterogeneity at the single cell level (Bray *et al*, 2016). Similarly, proteomics and bulk RNA sequencing also average the signal of thousands of cells into single per-protein or per-gene expression values.

Single cell RNA sequencing (scRNAseq) is extraordinarily powerful at examining cellular heterogeneity and can discover the processes at play in a heterogeneous drug response. For example, Srivatsan and colleagues measured the effects of 188 biologically active compounds at multiple doses in multiple cell lines using a massively multiplexed version of scRNAseq called

sci-Plex (Srivatsan *et al*, 2020). The authors demonstrated that the expression of acetate-forming enzymatic pathways is heterogeneous after HDAC inhibitor treatment; and that the activities of these pathways are important determinants of cellular survival after treatment with HDAC inhibitors. However, single cell RNA sequencing's greatest strength – its high dimensionality – is also an Achille's heel: with so many variables (i.e. genes) measured in a single experiment, it is challenging to understand what variation is biologically important versus related to transcriptional noise or technical artifacts. Because of this, Srivatsan and colleagues could not report reproducible intra-perturbation heterogeneity for most of the compounds tested.

Visual Cell Sorting is uniquely capable of leveraging morphological heterogeneity to understand cellular responses to perturbations. Cells that receive the same drug dose or genetic variant can be sorted into morphologically defined bins and subject to a transcriptomics or proteomics workflow separately. By analyzing the differences between the bins, new cellular states and functions related to the perturbation can be discovered. In this chapter, I describe our efforts to demonstrate this Visual Cell Sorting capability using the chemotherapeutic compound paclitaxel. Paclitaxel is a chemotherapeutic agent that stabilizes microtubules and has been used to treat cancer for decades (Rowinsky *et al*, 1992). Even in a clonal population, a subset of cells treated with a low dose of paclitaxel adopt a lobulated nuclear morphology (Theodoropoulos *et al*, 1999). Though Theodoropoulos and colleagues associated this morphologic phenotype with defects in nuclear import, little else is known about the relationship between this abnormal nuclear morphology and cellular function; nor is anything known about its relationship to paclitaxel treatment.

Methods

Cloning

To create pLenti-CMV-NLS-Dendra2x3-P2A-H2B-miRFP, three PCRs of Dendra2 (template derived from Dendra2-Lifeact7) were performed: one with an N-terminal NLS appended on the forward primer and a Gly-rich linker on the reverse; one with the Gly-rich linker on the forward primer and a second, non-identical Gly-rich linker on the reverse primer; and one with the second Gly-rich linker on the forward primer and a stop codon on the reverse primer. These were combined with an attB construct backbone (Matreyek *et al*, 2017) to create attB-NLS-Dendra2x3. In a second cloning step, H2B-miRFP from pH2B-miRFP703 was appended

downstream to create attB-NLS-Dendra2x3-P2A-miRFP. Finally, the Dendra2x3-P2A-H2B-miRFP open reading frame was cloned into pLenti CMV rtTA3 Blast (w756-1).

To create pLenti-CMV-mBeRFP-NLS, a gBlock encoding codon-optimized mBeRFP (Yang *et al*, 2013) (IDT) was cloned into pLenti CMV rtTA3 Blast (w756-1) with an NES encoded into Gibson overhangs.

Cell line

hTERT RPE-1 cells (ATCC CRL-4000) and derivatives thereof were cultured in F12/DMEM supplemented with 10% FBS, 1 mM PenStrep, and 0.01 mg/mL hygromycin B.

To create the hTERT RPE-1 clonal line expressing NLS-Dendra2x3, mBeRFP-NES, and H2B-miRFP, lentiviruses encoding these constructs were added to the parental line, and single cell clones were similarly sorted and expanded in conditioned media in 96 well plates.

Time-lapse imaging of cells treated with paclitaxel

hTERT-RPE-1 cells expressing Dendra2-NLS, H2B-miRFP703, and mBeRFP-NES were plated at a density of 50,000 cells per well in 2-well μm-slide chambers (ibidi; Martinsried, Germany). Twenty-four hours after plating, the cell media was replaced with media containing 0.25 nM taxol. After the cell media change, the cells were imaged for 24 hours with a pass time of 10 minutes. Imaging was performed on a Leica DMi8/Yokagawa spinning disk confocal microscope with a 20x 0.8NA air objective at 37°C and 5% CO2. Images were captured with an Andor (Belfast, United Kingdom) iXon Ultra camera using Metamorph software. Videos were cropped and adjusted for brightness and contrast using ImageJ and Photoshop.

Visual Cell Sorting of cells treated with paclitaxel

RPE-1 NLS-Dendra2x3/H2B-miRFP/NES-mBeRFP Clone 3 cells were plated at 50,000 cells per well in a 6-well plate. After 24 hours, cells were treated with paclitaxel at a final concentration of 0.25 nM. After 30 hours of treatment, cells were placed on the microscope and imaged, analyzed, and activated across 2,204 sites (~75% coverage, avoided well edges) in 2 wells. At each site, Dendra2 was imaged with 1x1 binning; a custom nuclear segmentation pipeline that optimized detection of nuclear blebs, herniations, and other abnormalities was employed; Metamorph's MDA analysis was used to compute shape factors for nuclear binaries. Cells with nuclear shape factor < 0.65 were activated for 200 ms, and cells with nuclear shape factor > 0.65 were activated for 800 ms. Cells from each well were trypsinized and resuspended in DPBS supplemented with 1% BSA and 2% FBS. Using FACS, cells corresponding to 200 ms and 800 ms photoactivation were sorted using FACS (Figure EV4A) into a 1.5 mL tube

containing 1mL DPBS supplemented with 1% BSA. In Experiment 1, cells were sorted according to their nuclear phenotype (i.e. 200 ms cells in bin 1, 800 ms cells in bin 2; Fig S4A). Cells were imaged, activated, and sorted identically in Experiment 2, except that all activated cells were sorted into one bin (i.e. both 200 ms and 800 ms cells in bin 1; "unseparated paclitaxel-treated population"). For an example of the gating scheme, see Appendix Figure 3.1.

Extended description of the Visual Cell Sorting on cells treated with paclitaxel

To identify morphologically-normal and lobulated cells were imaged for unactivated Dendra2 (FITC channel; 100 ms). Then, a custom nuclear segmentation pipeline that optimizes detection of nuclear blebs, herniations, and other abnormalities was employed. First, a top hat filter with a maximum object area threshold of 5,000 pixels was applied to remove large autofluorescent objects, and a 3x3 low pass filter was applied to smooth nuclear fluorescence. To find nuclei, a flatten background filter (removal of objects < 20 pixels in size), Sobel edge detection kernel, and a sharpening kernel were used before applying Metamorph's "legacy heuristic" thresholding algorithm to create nuclear binaries. To clean the nuclear binaries, holes were filled; tunnels 1 pixel in width were filled in using a dilate function; holes were filled again; and then an erode function was used to reverse the enlarging effect of the dilate and edge detection steps. Finally, objects less than 20 pixels in size and greater than 400 pixels in size were discarded. Shape factors were computed for each remaining object. See the GitHub repository for the Metamorph journal that was used.

Single cell RNA sequencing of sorted, paclitaxel-treated populations

After sorting, cells were spun at 1,000 x g at 4 °C for 5 minutes, then all but 50-100 uL of supernatant was removed. Cells were counted and subjected to 10X Single-Cell RNA sequencing v2 (10x Genomics; Pleasanton, CA; cat. no. 120236, 12037) according to the manufacturer's instructions. 10x Cell Ranger version 2.1.1 was used to process lanes corresponding to the single cell libraries and map reads to the human reference genome build Hg19. Unique molecular identifier (UMI) cutoffs were chosen by 10x Cell Ranger software. Reads and cell numbers were normalized via downsampling by the aggregate function in 10x Cell Ranger. After normalization, cells had a median of 9,249 UMIs (Experiment 1, separated populations) or 16,932 (Experiment 2, unseparated population) per cell.

Analysis of single cell RNA sequencing data

Analysis of 10X CellRanger output files was done in RStudio v1.1.456 running R 3.6.0. Cell cycle scoring and annotations were performed with Seurat, as previously described (Butler *et al*, 2018). UMAP was performed with Monocle3 (Trapnell *et al*, 2014; Qiu *et al*, 2017). Mutual-

nearest neighbors batch correction was performed using the Batchelor package (Haghverdi *et al*, 2018) in the following order: unseparated cells from Experiment 2 were batch corrected with morphologically-normal cells from Experiment 1, and then lobulated cells from Experiment 1 were batch corrected. An R-markdown file with the code used to run the analysis is available on the GitHub repository.

Differentially Expressed Genes Analysis

Mutual nearest neighbors batch correction (Haghverdi *et al*, 2018) was used to align cells from Experiment 2 (normal and lobulated cells sorted into the same tube, one 10X lane) to cells from Experiment 1 (normal and lobulated cells sorted into separate tubes, two 10X lanes). Principle components 1 through 4, which were output by the batch correction algorithm, were used to train a logistic regression model for nuclear lobulation on the cells in Experiment 1. This model was applied to Experiment 2, resulting in each cell being assigned a lobulation score, which is high in lobulated cells in Experiment 1 and low in normal cells in Experiment 1. Then, a differentially expressed gene test was performed on the cells in Experiment 2 using lobulation score, Seurat-computed G1 score, and Seurat-computed G2/M score as covariates. For a detailed discussion of this analysis.

Extended description of the differentially expressed genes analysis

We noted that the Visual Cell Sorting-derived lobulated and normal single cell RNA transcriptomes appeared to be confounded by a batch effect, despite the fact that cells were derived from a single well, sorted on the same day, and processed side by side (Experiment 1; Supplementary Figure 4). Using SoupX (Young & Behjati, 2018), which applies a linear PCA transformation that is determined by the RNA in empty 10X emulsion droplets, we found that cell-free RNA was responsible for this effect (Supplementary Figure 4B). To confirm this hypothesis, we repeated the experiment but sorted lobulated (800 ms activated) and morphologically normal (200 ms activated cells) cells into the same bin (Experiment 2, “unseparated” population) and processed these cells in a single 10X lane. A UMAP embedding of the single cell transcriptomes derived from these unseparated cells showed a single cluster, confirming the batch effect in Experiment 1.

Although both SoupX and the mutual nearest neighbors algorithm (Haghverdi *et al*, 2018) applied to cells in Experiments 1 corrected the batch effect (Supplementary Figure 4B), it is not statistically appropriate to use batch-corrected gene expression values to conduct a differentially-expressed gene test (Haghverdi *et al*, 2018). As such, we sought to use mutual nearest neighbors batch-corrected principle components to label the unseparated cells in

Experiment 2 according to their similarity to the known lobulated or morphologically-normal cells in Experiment 1; and then use the raw gene expression values in Experiment 2 to conduct a differentially-expressed gene test. We noted that the first four principle components in the MNN batch correction output correlated with the visual phenotype (i.e. morphologically normal vs. lobulated) in Experiment 1. So, we performed on the cells in Experiment 1 a logistic regression to devise a score that distinguishes between morphologically normal and lobulated cells:

$$lobulation\ score \sim PC1 + PC2 + PC3 + PC4$$

This regression model was then applied to cells in Experiment 2 (unseparated cells, single 10X lane) and the model predictions, which we called the "lobulation scores", were extracted for each cell. Using Moncole3 (Trapnell *et al*, 2014; Qiu *et al*, 2017), we performed on the Experiment 2 gene expression matrix a DEG test using the lobulation scores and Seurat-computed cell cycle scores as covariates:

$$Gene \sim lobulation\ score + S\ score + G2M\ score$$

By doing the differentially expressed gene test using Experiment 2, in which lobulated and normal cells were sequenced together, we avoided any batch-related artifacts. This operation is analogous to the cluster-based analysis originally discussed by Haghverdi and colleagues (Haghverdi *et al*, 2018), but uses a principle component-derived score rather than principle component-derived clusters as cell labels.

Gene Set Enrichment Analysis

Gene set enrichment analysis was performed using the piano package (Väremo *et al*, 2013) in R on differentially expressed genes with a log2-normalized effect value (equivalent to the expected log2-fold change per unit increase in lobulation score) less than -0.1 and a q-value less than 0.01. The MSigDB Hallmarks and Canonical Pathways gene sets were used (Subramanian *et al*, 2005; Liberzon *et al*, 2015).

Results

We treated a telomerase-immortalized cell line derived from human retinal pigment epithelium, hTERT RPE-1, with paclitaxel and observed mitoses that sometimes resulted in nuclear lobulation that persists through the cell cycle (Videos EV1 and EV2, Hasle *et al*, 2020). In order to computationally define a cutoff for lobulated nuclei, we measured the shape factor, a circularity metric (Figure 3.1A), of nuclei in vehicle-treated cells and found that 95% of these

morphologically normal cells have a nuclear shape factor greater than 0.65. We then analyzed paclitaxel treated cells and found that 30% of paclitaxel treated cells had lobulated nuclei, defined by shape factor of less than 0.65 (Figure 3.1B).

Given that morphologic phenotypes are potent indicators of cell state (Rohban *et al*, 2017), we hypothesized that the change in nuclear morphology was accompanied by a distinct gene expression program. To test this hypothesis, we used Visual Cell Sorting to separate morphologically normal paclitaxel- or vehicle- treated cells (shape factor > 0.65) from those with lobulated nuclei (shape factor < 0.65). We then subjected each population of cells to single cell RNA sequencing (Figure 3.1C). Imaging, analysis, photoactivation, and FACS-based recovery (Figure 3.2A) of ~200,000 cells took less than 7 hours. Following FACS, we prepared sequencing libraries for approximately 6,000 single cell transcriptomes from each population. We observed an RNA sequencing batch effect that was completely attributable to different levels of cell-free RNA (Young & Behjati, 2018; Cao *et al*, 2017) in the lobulated and morphologically-normal cell sequencing preps (Figure 3.2B).

We used UMAP (McInnes *et al*, 2018) to visualize a low-dimensional embedding of the single cell transcriptomes. The distributions of normal and lobulated paclitaxel-treated cells in the UMAP embedding were similar, indicating modest differences in their transcriptomic states. Differences in cell-cycle phase (Butler *et al*, 2018) largely explained transcriptomic variation (Figure 3.1D). More lobulated cells than normal cells were in G1 (53% vs. 44%), suggesting that lobulated cells had an increased propensity to arrest in G1. Indeed, G1 arrest is known to occur after paclitaxel treatment in non-transformed cell lines (Trielli *et al*, 1996).

To understand the relationship between transcriptomic variation, lobulation, and cell cycle, we examined the top batch-corrected principal components of the single cell transcriptomes. We noticed that the first four principal components separated cells by nuclear morphology (Figure 3.2C). To discover the genes associated with nuclear lobulation while controlling for the cell-free RNA batch effect, we sequenced an unseparated paclitaxel-treated cell population and aligned their transcriptomes to those from morphologically-normal and lobulated cells (Haghverdi *et al*, 2018). We then derived a lobulation score for each cell via linear combinations of the four principle components that correlate with nuclear morphology (Figure 3.2D). Finally, we extracted genes associated with this lobulation score, which is higher in cells with lobulated nuclei, in the unseparated cells by using a differentially expressed gene test (Methods). As expected, there was less variance in the lobulation scores in DMSO-treated cells compared to their paclitaxel-treated counterparts.

In total, 1,065 genes were significantly associated with the lobulation score in the paclitaxel-treated cells, compared to 190 genes in the vehicle-treated cells (adjusted p-value < 0.01; Figure 3.2E; Supplementary Table 5). There was little overlap between the lobulated nucleus-associated genes in paclitaxel- and vehicle-treated cells, indicating treatment-specific morphologically defined cell states (IOU = 1.2%; Figure 3.2F). To our surprise, 82% of differentially expressed genes in the paclitaxel-treated cells were more highly expressed in normal-appearing cells. Unlike their vehicle-treated counterparts, morphologically-normal cells treated with paclitaxel upregulated the genes encoding microtubules (e.g. *TUBB4B, TUBA5A*; Figure 3.1E), a well-documented response to microtubule damage and paclitaxel treatment (Gasic *et al*, 2019). These cells also upregulated the chaperone clusterin (*CLU*) and its co-activator *HSPA5*, which together decrease paclitaxel-mediated apoptosis by stabilizing mitochondrial membrane potential (Li *et al*, 2013). Intrigued by the notion that the morphologically normal cells are resisting the effects of paclitaxel, we searched the literature for other genes upregulated in these cells and found that many of them, including *PRMT1* (Cho *et al*, 2012)*, ENO1* (Georges *et al*, 2011)*, STMN1* (Alli *et al*, 2007)*, LDHA* (Zhou *et al*, 2010)*, ANXA5* (Di Michele *et al*, 2009), and *HSPA8* (Sugimura *et al*, 2004), are associated with paclitaxel resistance in diverse cancers.

To better understand the gene expression program associated with normal nuclear morphology in the context of paclitaxel treatment, we looked for enrichment of genes in previously defined gene sets (Liberzon *et al*, 2015) covering a host of cellular processes (Supplementary Tables 6, 7). Morphologically normal cells upregulated all 8 members of the chaperonin containing TCP-1 complex (adjusted p value = 7.97e-15; Figure 3.1E), which is critical for tubulin folding and has been previously associated with paclitaxel resistance in ovarian cancer (Di Michele *et al*, 2009). Morphologically normal cells also upregulated the transcriptional targets of two paclitaxel resistance-associated signaling pathways (Parasido *et al*, 2019; Shafer *et al*, 2010): c-Myc (adjusted p value = 2.26e-36) and mTORC1 (adjusted p value = 1.09e-16; Figure 3.1F). By contrast, DMSO-treated cells had a far less significant association between nuclear morphology and the activity of these pathways (adjusted p values > 0.001). Together, these results suggest that the morphologically normal, paclitaxel-treated cells mounted a biosynthetic and proteostatic response to drug treatment, with remarkable similarities to the gene expression profiles observed in paclitaxel resistant cell lines and cancers.

We next tested whether the genes upregulated in normal-appearing, paclitaxel-treated cells represented a state of paclitaxel resistance. To do this, we compared our genes to those described by Wu and colleagues, who performed single cell RNA sequencing on an esophageal carcinoma parental line and a paclitaxel-resistant cell line derived by continuous treatment with paclitaxel for two months (Wu *et al*, 2018). There was remarkable concordance between the shared set of significant DEGs (Figure 3.3A; IOU = 0.14; Chi squared p value < 2.2e-16). Furthermore, when we compared genes that differed between DMSO- and paclitaxel-treated cells in our dataset to the Wu et al. dataset, we found far less overlap (Figure 3.3B; IOU = 0.08; Chi squared p value = 0.00042). This suggests that the gene expression program identified in the normal-appearing, paclitaxel treated cells is one that potentiates paclitaxel resistance.

Discussion

We leveraged Visual Cell Sorting's ability to recover live, phenotypically defined subsets of cells to investigate the heterogenous cellular response to paclitaxel treatment using single cell RNA sequencing. To our surprise, cells that resist the effect of paclitaxel on nuclear morphology appear to be counteracting the drug's effects at the molecular level with a gene expression program similar to paclitaxel-resistant cancers. This phenomenon, whereby a subset of clonal cells resists the effects of drug treatment with a protective gene expression program, is reminiscent of the "pre-resistance" reported in primary melanoma cells (Shaffer *et al*, 2017). The lack of a strong cell cycle signature for either the normal-appearing or lobulated cells suggests that the gene expression program is largely unrelated to the cell's position in the cell cycle when the perturbation was introduced.

"Pre-resistance" is defined by Shaffer et al. as a cellular state that (a) is stochastically adopted by cells in a population and (b) that provides the cell with some survival benefit when drug is added. Shaffer et al. cleverly used Luria-Delbruck fluctuation analysis to determine that pre-resistance (rather than genetically driven resistance) is a critical component of post-drug survival in their primary melanoma model system; they also provide evidence that pre-resistance is not cell cycle-related. In their system, pre-resistance was characterized by high expression of *EGFR* and other pro-survival signaling genes at levels that were 10-100x higher than the population average; such cells were detected at a frequency of ~0.1% and appeared to represent the tails of gene expression distributions.

The normal appearing cells in paclitaxel treated populations are in a more pre-resistant state than lobulated appearing cells. Compared to their lobulated counterparts, normal appearing cells upregulate the same genes as paclitaxel resistant cancers. Some of these genes, like *CLU*, have been mechanistically linked to resistance (Li *et al*, 2013). Furthermore, normal-appearing cells have an overall transcriptomic state similar to esophageal carcinoma cells selected for paclitaxel resistance over months (Wu *et al*, 2018), but they been exposed to drugs for less than 48 hours. This is not a time scale over which rare genetic resistance could sweep the population to the 50-70% frequency. The reason we see 50-70% frequency of normal appearing (i.e. more pre-resistant cells) rather than the 0.1% frequency observed by Shaffer and colleagues likely has to do with drug dose: we used a sublethal dose of paclitaxel, whereas they report using doses that kill most of the cells in culture (Shaffer *et al*, 2017). Our treatment regimen is therefore less selective; perhaps by treating cells with higher doses and selecting cells that remain normal appearing, we would be able to find more extreme pre-resistant states.

An alternative explanation for the divergent cell states we report is that genetic variants provide approximately half of cells with more resistance to paclitaxel. This hypothesis is less likely, as these genetic variants would (a) have arisen within this chromosomally stable cell line within the 10 weeks that it took to go from a single cell to our experiments and (b) somehow affect cellular state (e.g. noncoding mutations in multiple transcription factors) rather than the function of individual proteins (e.g. mutations in tubulin, which are reported much more commonly in cancers; Kavallaris, 2010). To help distinguish between genetic resistance and pre-resistance, one could apply mitochondrial lineage tracing (Ludwig *et al*, 2019) to the dataset, which would provide genetic relationships between paclitaxel-treated cells. If genetically related cells are more likely to have the same morphologic phenotype, our data would support a genetic resistance model.

If there is no evidence for a genetic cause of the resistant cell state, one could also use the data to examine whether the morphologically normal cell state described here *pre-exists* in the population, or whether it is *induced* by paclitaxel resistance. To do this, one would use mutual nearest neighbors batch correction (Haghverdi *et al*, 2018) to batch correct the DMSO- and the paclitaxel-treated samples from Experiment 2 together, and derive a “pseudodose” score (Srivatsan *et al*, 2020) for each cell with the same pipeline used to derive the lobulation score. This score will describe how similar each cell is to the typical paclitaxel-treated cell versus the typical DMSO-treated cell. Finally, one could compare the relationship between lobulation score and pseudodose. If pseudodose and lobulation score are positively or negatively correlated, it is

provides evidence that paclitaxel treatment induces the transcriptional state associated with lobulated or normal cells, respectively. If there is no correlation, it provides evidence that these states “pre-exist” in the DMSO treated population.

Though the analyses described above would support a model of genetic resistance, pre-existing pre-resistance, or induced pre-resistance, it is critical to confirm with follow-up wet lab experiments. To explicitly test for genetic resistance versus pre-resistance, one could sort paclitaxel-treated cells by morphology and re-treat them with a range of paclitaxel doses either directly after sorting or a week after sorting. This way, an EC50 curve could be generated for each set of morphologically defined cells at the two time points. Pre-resistance would result in morphologically normal cells having higher survival than lobulated cells directly after sorting but not one week after sorting. Genetic resistance, which should stay fixed after one week, would result in higher survival for morphologically normal cells at both time points.

To test for pre-existing versus induced pre-resistance, one could use a lentiviral barcoding approach (Weinreb *et al*, 2020). Here, one would introduce a complex library of barcodes into the RPE-I cells two days before paclitaxel treatment and then perform single cell RNA sequencing on the population before and after drug treatment. Weinreb and colleagues reported that daughter cells have highly correlated transcriptomes (a phenomenon that disappears after several cell divisions). By comparing cells with the same barcodes before and after treatment, you can determine whether treatment induced the gene expression program related to lobulation score, or whether it existed in the population prior to drug treatment.

Finally, it will be important to test whether the pathways identified in this chapter are mechanistically responsible for the pre-resistant state observed. To test this, one could treat cells with RNAi targeting *MYC* or with rapamycin before initiating paclitaxel treatment. If there is a shift towards more lobulated cells, it indicates that these genes may be critical for the morphology-related, pre-resistant cell state I report here. In a second experiment, one could build a fluorescent genetic reporter for c-Myc transcriptional activity and mTOR signaling and introduce them into the RPE-I cell line. By performing time-lapse microscopy shortly before and for ~48 hours after drug treatment, one could test whether pre-treatment reporter levels predict post-treatment morphologic phenotypes. These experiments may reveal the origins of pre-resistance, thereby improving our understanding of this phenomenon and perhaps facilitating the translation of these findings into clinical care.

Chapter 4: Towards Comprehensive Dissection of the Function of Thousands of *PTEN* Variants

Introduction

PTEN biology and its role in cancer and Mendelian disease

Phosphatase and tensin homolog (PTEN) is a tumor suppressor that affects cellular growth, morphology, and DNA repair (Box 1). PTEN is 403 amino acids in length and has two domains: an N-terminal phosphatase domain and a C-terminal "C2 domain" with homology to Tensin, a cytoskeleton-associated protein. Following the C-terminal domain, an approximately 50-residue unstructured region regulates PTEN activity. PTEN has both enzymatic and non-enzymatic molecular functions, acting as a lipid phosphatase, a protein phosphatase, and a protein scaffold. These molecular functions mediate PTEN's effects on cell growth, morphology, migration, genome stability, and DNA repair. It performs these functions at the plasma membrane, in cytoplasmic vesicles, and in the nucleus, including lipid phosphatase activity against phosphoinositol lipids; protein phosphatase activity against membrane-associated and nuclear proteins; and enigmatic non-enzymatic functions that primarily regulate nuclear protein activity (Box 1; Table 1).

PTEN's principal molecular function is as a phosphatase. The N-terminal domain contains a HCXXGXXR phosphatase catalytic motif, which is capable of dephosphorylating both lipid and protein substrates (Myers *et al*, 1997). The primary lipid target of PTEN is phosphoinositide-(3,4,5) phosphate (PIP3), an allosteric activator of the pro-growth and pro-survival Akt/PKB pathway (Myers *et al*, 1998). PTEN's protein phosphatase activity exhibits broad target specificity, including (1) membrane associated proteins, such as focal adhesion kinase (FAK) (Tamura *et al*, 1998) and protein tyrosine kinase 6 (PTK6) (Wozniak *et al*, 2017); (2) ER-localized proteins, such as PREX2 (Mense *et al*, 2015); (3) nuclear proteins, such as CREB (Gu *et al*, 2011); and (4) mitotic spindle-associated proteins, such as EG5 (He *et al*, 2016).

PTEN also has non-enzymatic molecular functions that involve direct interaction with and modulation of downstream protein effectors in the nucleus. Two clear examples of such scaffolding activity have been reported. For example, PTEN's interaction with the anaphase-promoting complex, APC/C, increases its ubiquitination of various cell cycle effectors (Song *et al*, 2011). Furthermore, PTEN interacts with RPA1 via its N-terminal phosphatase domain and promotes RPA1 protein stability via recruitment of OTUB1 to collapsed replication forks (Wang *et al*, 2015).

PTEN's diverse array of molecular targets and interactions is matched by its effects on many cellular phenotypes. PTEN knockout is associated with a number of cancer hallmarks, including uncontrolled cell growth, increased tissue invasion, and genome instability. The mechanisms behind each of these cellular phenotypes are multifaceted. For example, PTEN inhibits cell growth by decreasing Akt/PKB signaling (Myers *et al*, 1998), by suppressing CREB transcriptional activation (Gu *et al*, 2011), and by promoting APC/C complex activity (Song *et al*, 2011). It suppresses invasion by dephosphorylating FAK, PTK6, PREX2, and perhaps an unidentified lipid substrate (Tamura *et al*, 1998; Wozniak *et al*, 2017; Mense *et al*, 2015; Tibarewal *et al*, 2012). Finally, PTEN promotes DNA repair by stabilizing RPA1 and inducing FANC-mediated DNA repair at collapsed replication forks, among other uncharacterized mechanisms (Bassi *et al*, 2013; Vuono *et al*, 2016; Wang *et al*, 2015).

Genetic variation in *PTEN* causes human disease. For example, *PTEN* is either mutated or deleted in approximately 9% of all cancers. Germline *PTEN* variation causes PTEN hamartoma syndrome (PHTS), which is broadly characterized by macrocephaly, benign growths called hamartomas, and an increased risk of malignancy (Mester & Eng, 2013). Other *PTEN* germline variants are associated with a milder clinical course, or with Autism Spectrum Disorder (Butler *et al*, 2005). The relationship between PTEN sequence, molecular function, structure, impact on cellular phenotype, and these clinical outcomes are only partly understood. As such, it is currently challenging to predict the effect of most clinically observed *PTEN* variants either for cancer progression or in terms of outcomes of patients carrying germline *PTEN* variants.

Studies involving PTEN missense variants have helped resolve the relationships between molecular function and cellular phenotypes, as some missense variants can selectively impact individual molecular functions. Such separation-of-function variants specifically alter or abrogate some protein functions while preserving others. Comparing the molecular and cellular effects of separation-of-function variants can reveal how particular protein domains, surface pockets, and residues map to molecular functions and cellular phenotypes. Indeed, for the past 20 years, biologists have leveraged such missense variants to help dissect PTEN's functionality.

Multiplex assays of variant effects (MAVEs) are powerful tools for comprehensively determining how changes to protein variants affect protein function, which can be correlated with cell and disease phenotypes (Gasperini *et al*, 2016; Starita *et al*, 2017; Figure 4.1). MAVEs enable the simultaneous assessment of thousands of variants of a protein in a single experiment. Recently, MAVEs were used to assess two complementary properties of PTEN: lipid phosphatase function within yeast (Mighell *et al*, 2018) and intracellular abundance in human-derived cell lines

(Matreyek *et al*, 2018). By testing thousands of possible PTEN missense variants, these studies yielded new insights into the relationship between PTEN's sequence, structure, and function.

Mighell *et al.* adapted a humanized yeast model to assess PTEN lipid phosphatase activity at high throughput (Mighell *et al*, 2018). Expression of the human PI3K catalytic subunit p110 inhibits yeast cell proliferation, which can be rescued by the phosphatase activity of human PTEN (Rodríguez-Escudero *et al*, 2005). This assay was previously used to functionally characterize dozens of PTEN variants (Andrés-Pons *et al*, 2007; Rodríguez-Escudero *et al*, 2011). Mighell *et al.* incorporated a site-saturation mutagenesis library of PTEN into yeast and used high-throughput sequencing to measure the growth effects of each variant in the library in parallel. Of the 6,564 missense variants, which they scored with high confidence, 1,789 variants exhibited activities below the range of scores observed for synonymous variants. Of these, 1,249 had extremely low activities similar nonsense variants, with the remaining 540 variants exhibiting an intermediate phenotype.

The Fowler lab developed Variant Abundance by Massively Parallel Sequencing, or VAMP-seq (Matreyek *et al*, 2018), by fusing a library of PTEN single amino acid variants to EGFP. We used this assay to assess how each fused PTEN variant altered EGFP levels and thus PTEN abundance. We assessed 4,112 single-amino-acid variants of PTEN, identifying 1,138 variants with clear loss of intracellular abundance. Thermodynamic stability appeared to be a major correlate for PTEN intracellular abundance, as a large subset of loss-of-abundance variants were at hydrophobic residues present within the core regions of each domain or at residues making hydrogen bonds in the PTEN crystal structure.

The results of these MAVEs are consistent with previous literature. For example, many variants in PTEN's catalytic motif caused loss of lipid phosphatase activity in the yeast functional assay, but no change in steady state abundance. Variants abrogating serine/threonine phosphorylated sites in the C-terminal tail, such as S385A, exhibited increased activity and reduced abundance, as expected. Nonetheless, these experiments also come with some caveats: for example, the K254R and K266R sumoylation-defective variants do not exhibit reduced lipid phosphatase activity in the yeast assay, as would be expected. This may be due to the lack of a PTEN sumoylase orthologue in yeast.

MAVEs provide a comprehensive and nuanced understanding of how PTEN variants affect its molecular functions. For example, not all variants within the catalytic loop completely lost lipid phosphatase function, as computational approaches might predict. Furthermore, there were many

variants within the active site that exhibited normal abundance yet were completely inactive; these may represent dominant negative variants that act in a similar manner to C124S and G129E (Papa *et al*, 2014). As more PTEN MAVEs are completed, we expect that additional subtleties regarding PTEN's sequence-structure-function relationships will come to light.

U87MG cells have for two decades served as the principal cellular model of *PTEN* function. These cells are a glioblastoma line with homozygous null mutations in the *PTEN* gene; therefore, it is possible to introduce wild-type or a variant of *PTEN* into these cells and study their effects on various cellular phenotypes and molecular functions. In this chapter, I outline some preliminary experiments that validate a growth-based MAVE for selecting two PTEN functions in U87MG cells: growth and response to DNA damage. I also outline how I got the landing pad system (Matreyek *et al*, 2017) working in these cells. I finish with a road map for how to use standard selection based MAVEs and new single cell MAVEs to comprehensively dissect the functions of thousands of PTEN variants.

Methods

Cell culture

U87MG cells were cultured in DMEM supplemented with 10% FBS, 100 U/mL penicillin, and 0.1 mg/mL streptomycin. Passage number was kept below 25 for all experiments.

Cloning

Constructs that contained a 3rd generation lentivirus backbone and an mCherry-P2A-PTEN open reading frame with boosted expression using a WPRE were constructed using Gibson cloning. The template plasmids were G201A/G202A/G203A (from Kenny Matreyek; for the backbone and mCherry/PTEN ORF) and the TKOv3 LentiCRISPRv2 construct (AddGene #90294; for the WPRE). Information about the primers and templates used for the Gibson reaction are found below:

Fragment	**Target construct**	**Template**	**Primers**
Backbone (WT PTEN)	pNH073	G201A (KAM)	NH404/NH402
Backbone (C124S)	pNH074	G202A (KAM)	NH404/NH402
Backbone (Y68H)	pNH075	G203A (KAM)	NH404/NH402
WPRE	pNH073 pNH074	Addgene 90294	NH401/NH403

	pNH075		

KAPA 2x HiFi was used to perform the PCR. After a gel to ensure bands of the correct size were present, the reactions were digested with Dpn-I to remove template DNA. Then, the backbone and WPRE fragments were cloned together using a Gibson reaction. Individual clones for each target construct were picked and sequence confirmed using standard BigDye reactins.

To clone a barcode into the “classic” landing pad (pLenti-TetBxb1BFP-rtTA3_Blast; Matreyek *et al*, 2017), an oligo with Gibson overhangs (NH-landingpad_001) was made into a double stranded form using a Klenow extension with NH-landingpad_002 as a primer. In this 40ul reaction, there was 4ul of NEB Buffer 3.1, 4.5 ul of NH-landingpad_001 (stock concentration = 25 uM), 4.5ul of NH-landingpad_002 (stock concentration = 25 uM), and 27 ul of molecular biology grade water. The oligos were denatured at 98C for 3 minutes, then ramped down to 25C at a rate of 0.1 C per second; the thermocycler was subsequently paused at 1.35 ul of 1mM dNTPs and 1 ul of Klenow fragment enzyme (NEB) was added; the sample was vortexed and then placed back into the thermocycler and incubated at 25C for 15 minutes. A ZymoClean and Concentrate kit was subsequently used to clean the reaction. Next, 2 ug of pLenti-TetBxb1BFP-rtTA3_Blast was digested with KpnI-HF (NEB) according to manufacturer’s instructions; the reaction was confirmed using a 0.7% agarose gel; and then it was cleaned using a ZymoClean and Concentrate kit. After cleanup, the Klenow reaction product and the digestion product were subject to a Gibson reaction. Bacterial clones were sequence confirmed to have the barcode inserted directly 5’ to the blasticidin ORF. Nine out of ten clones sequenced had such a barcode. The resulting plasmid is named “LP (landing pad) classic barcoded”.

Lentiviral production

To produce lentivirus, 293T parental cells were seeded at a density of 0.4e6 cells per well in 6 well plates. One day later, cells were transfected with 1.5ug of pNH073/074/075, 1125ng of PsPax2 packaging vector, and 375ng of pMD-VSV-G envelope vector using 8 ul of FuGENE-6 per well. Cells were transfected according to manufacturer’s instructions. An mNeonGreen transfection control was included and checked using microscopy two days after transfection. Media was replaced one day after transfection and collected 2 and 3 days after transfection. Media was then spun at 300xg for 5 minutes at 4C, then decanted into a syringe with 0.45um cellulose acetate filter attached. Media was filtered, and then the PEG-it Virus Concentration solution was used as per manufacturer’s instructions to concentrate the virus ~500x.

Preliminary growth experiment

3.0e5 U87MG parental cells were transduced with 100 ul of pNH073/pNH074/pNH075 concentrated lentivirus in a 12 well plate. The complete media was supplemented with 0.8ug/mL polybrene and 2ug/mL doxycycline (final concentrations). Three days later, miRFP positive cells were sorted using the Aria III. Two days after the sort, cells were counted using the hemocytometer and mixed in 50-50 mixtures with untransduced cells. The relative proportion of miRFP-positive and untransduced cells were assessed using flow cytometry on the LSR-II at 3 days, 5, days, and 7 days after mixing the populations together.

Preliminary DNA damage experiment

1.5e5 U87MG parental cells were transduced with 50/30/30 ul of pNH073/pNH074/pNH075 concentrated lentivirus in a 12 well plate. The complete media was supplemented with 0.8ug/mL polybrene and 2ug/mL doxycycline (final concentrations). Three days later, the percentage of miRFP positive cells in each experiment was examined. Wortmannin (a PI3K inhibitor) and MMS (an alkylating agent) were added at concentrations of 2uM and 400uM, respectively. Fresh wortmannin was added every 12 hours, as it has a very short half life in cell culture. A negative control was also included, which only received doxycycline. After four days, the percentage of miRFP positive cells was assessed using flow cytometry on the LSR-II.

Derivation of U-87MG landing pad cells

U87MG cells do not grow well after sorting; as such, an alternative strategy was pursued to generate single cell clones. 1e5 U87MG parental cells were split, seeded into a 12 well plate, and transduced immediately with 40 ul of concentrated, barcoded “classic” landing pad virus (i.e. pLenti-TetBxb1BFP-rtTA3_Blast_barcode; virus was generated in the same was as explained under “Lentivirus Production” above; Matreyek *et al*, 2017)). At the same time as transduction, media was supplemented with 0.8 ug/mL polybrene and 2 ug/mL doxycycline (final concentrations). Two days later, cells were split into T-25s (doxycycline was re-added), and five days after that, they were sorted according to BFP positivity. Five days after sorting, cells were split into a 96 well plate at a concentration equivalent to ~0.75 cells per well. After three weeks, there were eight confluent 96 well plates. Clones were split out of the confluent 96 well plates and grown into a T-75. Cell gDNA was extracted using the DNeasy Blood & Tissue kits, a PCR using reaction using KAPA Robust 2X, KAM825, KAM112, and ~100ng of input DNA was performed according to manufacturer’s instructions (annealing temp 51 C, extension time 2 minutes), and then reactions with bands observed on a 1% agarose gel were sequenced using a BigDye according to manufacturer’s instructions. Clones with a single integration, denoted by

a single lentiviral barcode in the sequencing reaction, included C3, C7, C12, C4, C5, and C9. Landing pad line C3 was used for all subsequent experiments.

U87MG MaxCyte recombination protocol

To recombine attB constructs into U87MG cells, cells were split the day before transfection such that they are at 100% confluence on the day of transfection in a T-175. Cells were trypsinized and pelleted at 250xg for 5 minutes at room temperature. Cells were subsequently washed twice in 5mL of MaxCyte electroporation buffer. After electroporation buffer is added for the second wash, cells are gently resuspended completely and counted (approximately 4 million cells were counted per T-175). After the second wash, cells were resuspended in EP buffer at 1e8 cells/mL (assume 1e6 cells takes up ~5 ul of volume). Then 10ul of cells were added to tubes containing 500ng of pCAG-Bxb1 and 2400ng of attB-mCherry. mNeonGreen plasmid (3000ng) was used as a positive transfection control. Additional EP buffer was added to each tube to bring its total volume to 25 ul. Cells were then placed into OC25x3 cartridges and electroporated on the SH-5Y5Y setting. Cells were immediately plated into 6 well plates (no media added; just the 25ul of cells + DNA) and recovered in the 37C incubator for 25 minutes. Finally, 3mL warm complete media was added to cells. Cells were not split for at least 48 hours after transfection. Doxycycline (1 ug/mL) was added 24-48 hours prior to detection of recombination by flow cytometry.

Results

I first sought to test whether PTEN's plasma membrane-based lipid phosphatase activity could be selected for using growth assays in U87MG cells. To do this, U87MG cells were transduced with lentivirus that expressed miRFP and wild-type PTEN, an enzyme-null PTEN variant (C124S; Myers *et al*, 1998), or a nuclear-localized, low abundance PTEN variant (Y68H; Georgescu *et al*, 2000; Lobo *et al*, 2009). Cells were mixed with control cells that only expressed BFP. The proportion of miRFP and BFP was tracked over a week. As expected, cells expressed wild-type PTEN rapidly decayed in frequency compared to their BFP-expressing counterparts. Cells expressing the C124S variant also decayed over time but at a slower rate. Finally, the Y68H variant decayed at a rate between these two extremes (Figure 4.2A).

Next, inspired by work by Bassi and colleagues (Bassi *et al*, 2013), I tested whether the use of a PI3K inhibitor and a DNA damaging agent would allow us to select for PTEN's DNA damage repair functions. In this experiment, the PI3K inhibitor serves to bring PI3K activity to a low level, thereby depleting PTEN's PIP3 substrate. In turn, the PI3K inhibitor prevents PTEN's lipid phosphatase function, which may differ between variants, from affecting cellular growth

because all cells have the same low PIP3 concentration. As expected, when the growth experiment was repeated with Wortmannin (a PI3K inhibitor) and MMS (an DNA alkylating agent) added to the media, cells expressing any of the three PTEN variants showed higher growth than the PTEN-null, BFP expressing cells (Figure 4.2B). This was expected, as PTEN is known to have non-enzymatic activities that promote DNA damage repair (Hasle *et al*, 2019). This finding demonstrates that the DNA damage repair function of PTEN can be selected for using a combination of drugs.

With evidence that a growth-based selection and a drug-based selection could be used in U87MG cells to perform two distinct screens for PTEN function, I sought to introduce a functional landing pad into PTEN cells. After transducing cells with a landing pad system and selecting for single clones (Matreyek *et al*, 2017), I tried using chemical transfection to introduce control attB recombination plasmids into the cells. After failing to see evidence of recombination multiple times despite transfection rates of ~40% (Figure 4.3A), I tried using the MaxCyte electroporation system instead. This system worked, producing transfection rates of ~80% and recombination rates ranging from 2-15% (Figure 4.3B).

Future Directions

Library cloning

The library vector I propose using for the PTEN variant screens is outlined in Figure 4.4. In this vector, the IRES-miRFP cassette placed downstream of the PTEN open reading frame permits successful recombination and expression of PTEN to be assessed and sorted for using FACS. Critically, because the landing pad contains a doxycycline-inducible promoter upstream of the attB start site, it is possible to turn off PTEN expression until the selection experiments begin. This precludes the possibility that growth-inhibition by functional PTEN variants distorts the library composition before the experiment begins. The puromycin resistance gene, which is under the control of a constitutively active EF1a promoter, allows for selection of successfully recombined cells without induction of PTEN expression.

Kenny Matreyek and Jason Stephany have cloned a library of barcoded single amino acid PTEN variants (Matreyek *et al*, 2018). This library of variants was subsequently “filled-in” to include additional variants that were missing in the published library. To move this library into the landing pad vector, one could use directional cloning with the Xba-I and Xma-I sites present in the fill-in PTEN library. Once the library has been cloned in and recombined into the U87MG landing pad cells, it could be used for the experiments outlined below.

Growth selections

A simple growth selection would follow these steps: (1) the PTEN library is recombined into U87MG C3 landing pad cells; (2) recombinants are selected for in the absence of doxycycline (i.e. in the absence of PTEN expression) using puromycin; (3) PTEN expression is turned on using doxycycline; (4) after 24 hours, PTEN expressing cells (hopefully millions) are identified and sorted using the miRFP marker. Between five hundred thousand and one million of these cells are immediately frozen down for estimation of pre-selection variant frequencies (i.e. the input library to determine time point zero variant frequencies); (5) other cells are grown out, with samples of 500,000 to 1 million cells taken every 2-3 days and frozen down; and (6) variant frequencies in the input library and at each time point are estimated using next-generation sequencing. Enrich2 is used to estimate the effect of each variant on cell growth relative to wild-type (Rubin *et al*, 2017).

A growth-based selection for PTEN's DNA repair functions would be performed in an identical similar way, with the exception that cells are treated with a DNA alkylating agent and a PI3K inhibitor during the growth selection (i.e. starting directly after sorting). Two compounds that would be of particular interest are BKM120 (a specific PI3K inhibitor) and temozolomide (a DNA alkylating agent that is currently used to treat glioblastoma), because these two compounds are being tested for efficacy in clinical trials. Notably, the timescale of selection is likely to take longer for this experiment than for the growth function experiment, as the DNA damage may need to build over time before it causes catastrophic growth effects (see Figures 4.2 and 4.3; strength of negative selection is stronger in DMSO condition than positive selection in drug conditions). As such, I recommend doing the selection over 2 or 3 weeks instead of one week. Drug titration experiments on mixtures of PTEN expressing and PTEN null cells could be performed before the selection to determine the optimal drug doses and timescales of treatment.

Comparing the growth scores between these two experiments will be paramount to understanding how the lipid phosphatase (i.e. growth-inhibiting) and DNA repair functions of PTEN are related to one another. Unfortunately, the scores produced by Enrich2 have arbitrary units that are scaled according to the frequency wild-type cells, which will certainly grow at different rates in the two experiments. Brian Andrews in the Fields lab is currently developing a method that permits the true variant-dependent growth rate of cells to be estimated, rather than wild-type scaled scores with arbitrary units. His method requires that the population size of the cells under selection be recorded at each time point. To be able to use his method, I would use

the hemocytometer to calculate the number of cells at each time point. Notably, if the method works well it may be possible to forgo using the PI3K inhibitor and instead measure the difference in absolute growth rate between cells grown in the presence and absence of temozolomide. In doing so, you could account for the differences in cell growth due to the lipid phosphatase activity directly.

Single cell methods for assessing variant function

Though growth-based and FACS-based selections are a high-throughput way and flexible way to assess variant function, they suffer from three major limitations that hamper their ability to accurately assess variant function and to be scaled up. The first is their lack of dynamic range. For example, in a growth-based selection, if some variants are strongly selected against, their frequencies will drop to zero very quickly (e.g. around the first time point taken); this makes it difficult to assess exactly how deleterious that variant is. In sort-based selections, variants that end up in just the 1^{st} or 4^{th} sort bin (of four total sort bins) also have poorly estimated effects. The second problem is that they provide a limited number of readouts per variant per experiment. In most Fowler lab screens, barcode counts that are linked to the same amino acid variant are pooled together, resulting in a single readout of variant function per replicate performed. This often means that several replicates need to be performed to get an accurate estimate of variant function. Thirdly, selections can only be performed on one phenotype per experiment, limiting the number of phenotypes that can be examined in say, a single PhD.

Performing single cell phenotyping and genotyping addresses all three of these issues. Single cell phenotyping and genotyping experiments use a single cell method, such as *in situ* sequencing, single cell RNA sequencing, or single-cell antibody sequencing, to read out a variant barcode and perform multiplexed phenotypic readouts. By virtue of directly measuring phenotypes of single cells, these methods can be performed very shortly after variant expression is induced, thereby avoiding the possibility of variant dropout. Furthermore, because single cell methods provide continuous, quantitative data, they are less likely to suffer from the dynamic range issues that occur in data generated from four sort bins. It is also possible to derive a distribution of variant effects in each replicate of the experiment, thereby increasing the power to detect differences between variants. Finally, single-cell readouts can be multiplexed, unlike selection-based readouts, permitting multiple phenotypes to be assessed in a single experiment.

Feldman, Singh, and colleagues describe a relatively simple and high throughput *in situ* sequencing protocol for reading out variant barcodes (Feldman *et al*, 2019). They used the

protocol to examine the effects of thousands of CRISPR knockouts on the nuclear translocation of an NFkB reporter. The method could be adapted to studying PTEN's effect on morphology (Tamura *et al*, 1998). Such an experiment would involve (1) re-barcoding the PTEN library such that the region flanking its barcode region is identical to the one used by Feldman, Singh, and colleagues; (2) recombining the library into U87MG cells and selecting/sorting for successful recombinants; (3) plating cells onto fibronectin-coated glass wells; (4) fixing the cells and staining them with an actin stain and a nuclear stain; (5) performing *in situ* sequencing to score PTEN variants for their effect on cellular morphology. If multiple morphologic phenotypes were observed in the library, a score could be developed for each morphology.

The Shendure lab has developed a new, high-throughput single cell RNA sequencing method called single cell combinatorial indexing (sci-RNAseq; Cao *et al*, 2017). This method repeats the following steps 2-3 times: pool cells, split cells into 96- or 384-well plates, and perform a cDNA library preparation step that adds an index unique to the well. Because each cell in the output cDNA library is unlikely to end up being split into the same wells as any other cell, one can assume that each cDNA molecule with the same indices came from the same cell. This protocol allows single cell RNA sequencing to be performed on millions of cells in a single experiment, making its throughput high enough for large variant library screens.

Florence Chardon in the Shendure lab is using sciRNA-seq in a single cell phenotyping and genotyping experiment called DMS-Express. Here, she takes a library of cells expressing PTEN variants and subjects them to a mostly standard sciRNA-seq workflow. However, in the reverse transcriptase cDNA synthesis step, she adds a DMS-Express primer that targets the 3' UTR of *PTEN* variants in addition to the poly-dT primer, thereby capturing the variant barcode that is placed there. She therefore can retrieve the transcriptome and the variant barcode for each cell in the experiment. By examining where each variant lies in transcriptional space, she hopes to uncover new functional classes of PTEN variants.

Though sciRNAseq is a powerful way of assessing variant function in an unbiased manner, it suffers from the curse of high dimensionality, which makes it difficult to define biologically meaningful axes of transcriptomic variation. Furthermore, it is unlikely that different PTEN variants will lead to completely different transcriptome profiles; and because the whole transcriptome is captured, reaching a sequencing depth that accurately quantifies variant-specific genes changes could be challenging. As such, single cell experiments that measure a small number of PTEN-specific molecular phenotypes per experiment may make it easier to

assess the function of thousands of variants, some of which have only subtle effects on cellular function.

One single cell method that holds a lot of promise in this regard is single cell antibody sequencing. Here, antibodies tagged with oligonucleotide barcodes are mixed with cells of interest, and then a single cell RNA sequencing workflow that captures the antibody barcodes and the transcriptome is performed (Stoeckius *et al*, 2017). To use this as a single cell phenotyping and genotyping method, one could dispense with the transcriptome and instead capture a variant barcode with the DMS-Express primer; this would allow the effect of each variant on the levels of antibody-targeted protein concentrations to be assessed across millions of cells. By using antibodies that target pathways known to be critical for PTEN's various functions – phospho-PKB and phospho-FAK, for example – one could quantify the activity of multiple pathways across thousands of PTEN variants.

Discussion

Missense variants have been critical tools for understanding the complex relationships between PTEN's molecular functions and cellular effects. For example, in a recent review we identified eight variants that offered key insights into PTEN by altering molecular functions in unique, defined ways (Hasle *et al*, 2019). Three of these variants (Y138L, G129E, and C124S) abrogate protein phosphatase activity, lipid phosphatase activity, or both activities. These enzyme activity separation variants have been shown to impact multiple cellular phenotypes, which can occur in a context dependent manner, highlighting the complexity of PTEN function. The remaining five variants appear to affect constitutive subcellular localization (L42R, F21A) or post-translational modification (K254R, K289E, and Y240F). In contrast to the enzyme activity variants, the mechanism by which these variants affect migration and DNA repair has not been studied in detail; understanding their effects may shed light on how PTEN affects these critical cellular processes.

By expanding the repertoire of PTEN-related MAVEs to include single cell phenotyping and genotyping methods, dissection of PTEN along its many axes of function will become scalable. In turn, these datasets will hopefully resolve many variants of unknown significance, aid cancer risk predictions in patients with germline PTEN variants, and help provide cancer patients with accurate prognoses and efficacious therapies.

Conclusions

Elucidating the relationships between genetic variants, protein function, cellular function, and morphologic phenotypes is critical for understanding biological systems. Much like rashes, conjunctivitis, and palpable masses are important clues for clinicians understanding a patient's disease; so are changes in cellular structure important clues for cellular biologists. Visual Cell Sorting, as described in this thesis, is a helpful and extraordinarily flexible tool for leveraging the visual clues that cells provide. During my PhD, I used Visual Cell Sorting in two ways: to understand the function of genetic variants, and to understand the transcriptional basis of morphologic heterogeneity after drug response. These experiments confirmed that Visual Cell Sorting can be used to determine the effect of genetic variants on a visual phenotype at high throughput; and showed a strong relationship between cellular morphology and gene expression patterns that are relevant to a drug's mechanism of action.

In Chapter 1, I introduced the Visual Cell Sorting methodology and its strengths and weaknesses relative to other methods that can link morphology to visual phenomena. The primary advantage of Visual Cell Sorting is its ability to subject morphologically defined cells to multiple downstream assays. This allows the experimentalist to get a holistic understanding of how the morphological phenotype in question relates to cellular function. Another critical feature is that it separates live cells, which allows cells to be followed the perturbation to examine whether the morphology-function relationship observed is heritable or stochastic. The only other way that such questions can be answered is using cell tracking over long timescales (e.g. days to weeks), which is technically challenging and perhaps impossible for rapidly dividing cells; or indirectly via using lentiviral barcoding methods (Weinreb *et al*, 2020) and assuming that daughter cells have very similar transcriptomes.

Visual Cell Sorting has two critical limitations compared to *in situ* methods. First, it requires that the phenotype of interest be predefined and sorted into bins. This precludes an unbiased analysis of the relationship between genetic variants and morphology, for example. Furthermore, the binning of phenotypes comes at a loss of power because small differences in morphology may not be captured across bins. Second, Visual Cell Sorting has not been demonstrated to work in primary tissues. As such, using it to understand morphologic phenotypes in primary human tissue samples or in mouse models of human disease is not yet possible. However, it may be possible to use photoactivatable antibody stains to study primary tissues and two-photon microscopy and a Dendra2-expressing mouse line to study disease processes in mice.

In Chapter 2, I described how Visual Cell Sorting can be used to study how genetic variants impact a visual phenotype in a cell line. It would be relatively straightforward to apply the same process to study variants in medically relevant genes such as *LMNA* and *PTEN*. Cells expressing pathologic variants of these genes, which each cause Mendelian disease in humans, have been shown to exhibit morphologies that likely relate to the specific function(s) impacted by the variant (Tamura *et al*, 1998; Raharjo *et al*, 2001). By sorting for cells with these morphologies, one could elucidate the subtle relationships between protein function, cellular morphology, and clinical disease course. However, as was discussed in Chapter 4, it is likely that *in situ* sequencing is a better method to use for multiple reasons, including its larger dynamic range, its power to detect differences between variants, and its ability to examine multiple phenotypes per experiment.

In Chapter 3, I discussed the application of Visual Cell Sorting to studying heterogeneous drug responses. By performing transcriptomics on morphologically defined subpopulations of drug treated cells, I provided strong evidence that morphology and cell state are related. Strikingly, the relationship was opposite of what we had hypothesized: the cells with the mechanistically relevant cell state were the ones that remained normal appearing. These cells had a transcriptional program that strongly suggested they were pre-resistant to the drug administered, compared to the cells with the drug-induced morphology. In turn, the results show that the lobulated nuclear morphology observed is pathologic rather than homeostatic in nature.

One could imagine extending this type of experiment to identify “perturbation-modifying genes” for any drug associated with a morphologic phenotype. Here, it is critical to titrate the drug down to a level where cellular homeostatic systems are not overwhelmed. By using morphology to examine which cells successfully activate a homeostatic response to those that do not, it may be possible to identify pathways and genes that are critical for resisting the effect of the drug. In turn, the results provide information about the drug’s mechanism of action, as well as about the pathways involved in enacting homeostatic responses to the perturbation.

Using Visual Cell Sorting to identify perturbation-modifying genes could be extended to examine pathologic variants. The effect of pathogenic variants on cells is also dependent on whether cells activate homeostatic mechanisms to resist their effects. The homeostatic genes involved are termed “modifier genes” and are thought to be responsible for the incomplete penetrance observed in both cells (Mattiazzi Usaj *et al*, 2020) and mammalian systems (Klein *et al*, 2016). By expressing variants with incomplete penetrance in cells, separating them according to the gene-related morphology using Visual Cell Sorting, and characterizing each

population with an unbiased -omics scale method, it may be possible to identify a host of modifiers simply by using the existing transcriptional and possibly genetic variation in the cellular population. For these experiments, it is critical to use an expression system that permits very tight control of gene expression (Azizoglu & Stelling, 2019), as transcriptional dose could be an important confounder that overwhelms the effects of any perturbation-modifying genes.

The Visual Cell Sorting-based experiments that identify perturbation-modifying genes have multiple advantages over analogous CRISPR-based experiment (Ramkumar *et al*, 2018). Unlike a Visual Cell Sorting experiment, CRISPR-based experiments require building a library of CRISPR knockouts, which will be depleted for any essential perturbation-modifying genes. If they don't kill the cell, some gene knockouts may cause changes in cell state that are non-physiologic. The CRISPR-based experiments also require treating cells to the point of drug toxicity such that a growth phenotype can be achieved, and unlike transcriptomics and proteomics methods may miss the importance of splice isoforms or post-translational modifications.

In Chapter 4, I discussed efforts to get a library of PTEN variants expressed in a relevant cell line: U87MGs. For reasons that are unclear, the cell line was not amenable to chemical transfection-based recombination. However, using a high-performance electroporator resulted in recombination rates that are on par with previous PTEN variant effect screens (Matreyek *et al*, 2018). The use of this line will give access to previously published and intensively studied phenotypes such as DNA damage repair, cell morphology, and cell migration. Furthermore, using the line with new, single cell assays such as sci-RNAseq and antibody sequencing will permit variants to be linked to morphologic, transcriptomic and signaling phenotypes in new ways.

In 1665, the pioneer microscopist Robert Hooke described his observations regarding the morphologic characteristics of *Mucor* fungus species. He likened the fruiting bodies to flowers, inferring (correctly!) that they functioned as reproductive organs (Gest, 2004). In the past two decades, we have reinvigorated the study of morphology-function relationships in cells using new microscope and analysis technologies. However, the structures we study now – nuclei, endoplasmic reticulum, mitochondria – are largely unrelatable to objects encountered in the macroscopic world; we are, in essence, "blind" to what it could mean when they change shape.

To best understand what the abnormal morphologies mean, we need ways to directly connect them to molecular phenotypes – such as gene expression profiles, signaling pathways

activity, and genetic variants – that we understand better. In doing so, we can discover new axes of cellular function and dysfunction. My hope is that Visual Cell Sorting will permit mysterious structural changes in organelles like the nucleolus, the endoplasmic reticulum, and the mitochondrion to be characterized in terms of the functions and molecular phenotypes that underlie them. In doing so, we can return to a characteristically human way of pursuing biological research: to see something, marvel at it, and ask directly, "*what does it mean?"*

Figures

Figure 1.1

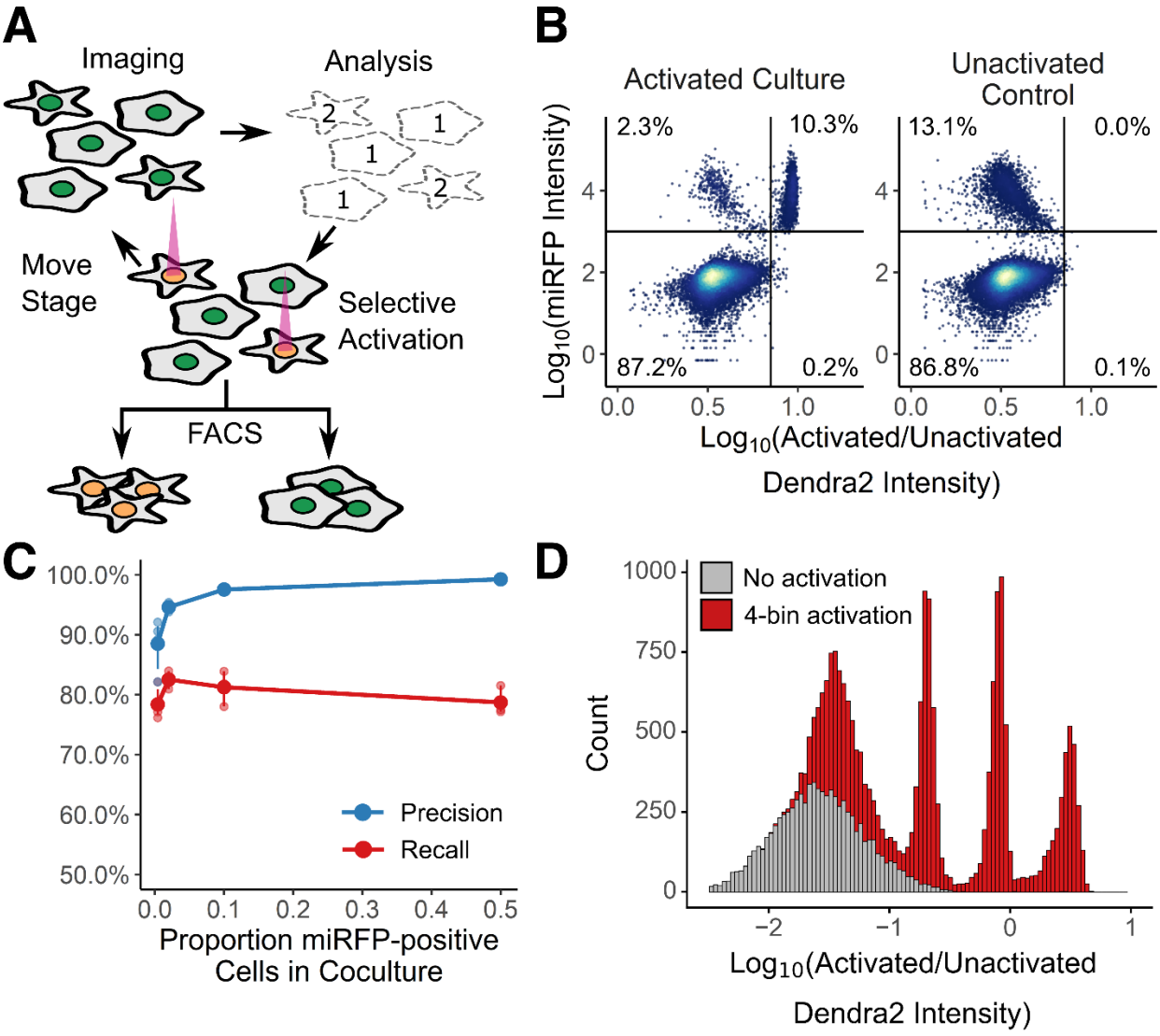


Figure 1.1. Visual Cell Sorting. (A) In an automated fashion, cells in a field of view are imaged and their phenotype classified. Cells of interest are illuminated with 405 nm light, which irreversibly photoactivates Dendra2 from its green to its red fluorescent state. The microscope then moves to a new field of view. These steps are repeated across an entire culture well. Then, fluorescence-activated cell sorting based on Dendra2 photoactivation is used to physically recover cells of interest. (B) To assess the photoactivation accuracy, U-2 OS cells expressing nuclear Dendra2 and miRFP, or nuclear Dendra2 alone, were co-cultured. The microscope was programmed to activate Dendra2 in cells expressing miRFP. Following photoactivation, miRFP expression and the ratio of activated to unactivated Dendra2 (left panel, n = 18,766 cells) were assessed with flow cytometry. In a second co-culture, Dendra2 was unactivated (right panel, n = 18,395 cells). Lines indicate gates for miRFP-positive cells and activated

Dendra2 cells, with the percentage of cells appearing in each quadrant indicated. (C) Same experiment as (B), except cells were mixed such that 0.5%, 4%, 12%, or 50% were miRFP positive. Precision and recall were computed; large solid, mean (N = 3 replicates); small points, individual replicate values; error bars, standard error from the mean. (D) U-2 OS cells in one well were illuminated with 405 nm light for 0, 50, 200, or 800 ms (red; N = 16,397). Cells in a second well were left unactivated (grey; N = 8,497). The ratio of activated to unactivated Dendra2 was determined by flow cytometry.

Figure 1.2

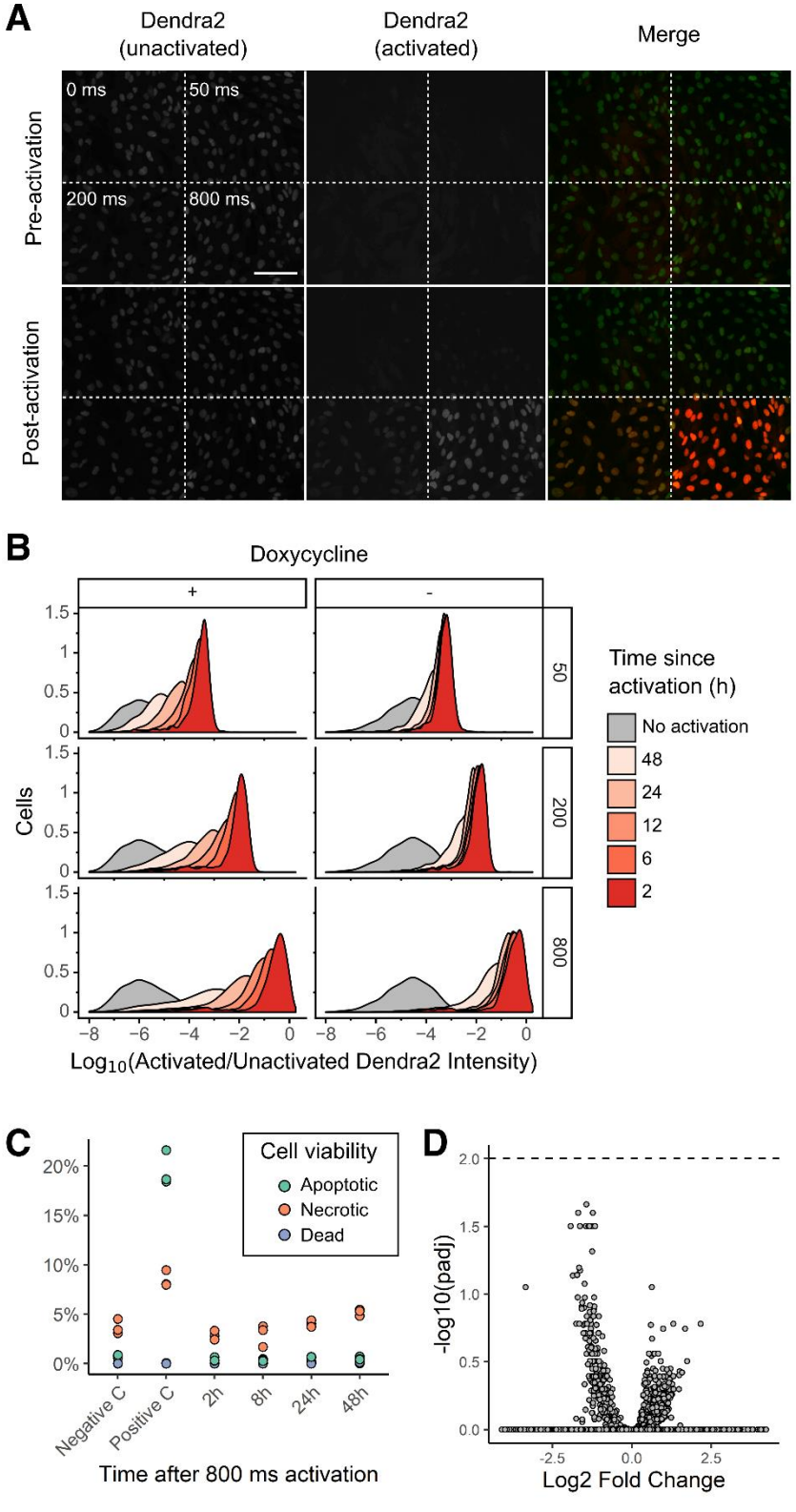

A
Dendra2 (unactivated)
Dendra2 (activated)
Merge
0 ms
50 ms
200 ms
800 ms
Pre-activation
Post-activation
B
Doxycycline
+
-
50
200
800
Cells
Time since activation (h)
No activation
48
24
12
6
2
Log10(Activated/Unactivated Dendra2 Intensity)
C
Cell viability
Apoptotic
Necrotic
Dead
20%
15%
10%
5%
0%
Negative C
Positive C
2h
8h
24h
48h
Time after 800 ms activation
D
-log10(padj)
Log2 Fold Change

Figure 1.2. Visual Cell Sorting, supplementary information. (A) RPE-I cells expressing NLS-Dendra2x3 were imaged in the unactivated and activated Dendra2 channels; then left unactivated or activated for 50, 200, or 800 ms; and re-imaged. Scale bar = 100 μm (B) U-2 OS cells expressing H3-Dendra2 under the control of a doxycycline inducible promoter were activated with 405 nm light for 50, 200, or 800 ms; incubated for various lengths of time; and then subject to flow cytometry to determine the degree of activated Dendra2 (left panel). To examine whether shutting off Dendra2 expression before the experiment increases photoactivation ratio stability, the experiment was repeated, but doxycycline was removed from the media before cells were placed under the microscope (right panel) (C) To examine the effect of Dendra2 photoactivation on cell viability, cells were activated for 800 ms and then apoptosis, necrosis, and death were assessed by flow cytometry using DAPI and Annexin-V (n = 10,000 cells) . Negative C, no photoactivation. Positive C, incubation of cells at 50 C for 10 min. The results of three independent replicates are shown.. (D) To test whether Dendra2 photoactivation affects gene expression, cells were activated for 800ms, incubated for 0.5, 1.5, 2.5, 3.5, 4.5, or 6 hours and subsequently subject to bulk RNA seq. Samples were compared to two separate replicates of unactivated cells. Volcano plot of differentially expressed genes shown. Dotted line, adjusted p-value of 0.01.

Figure 2.1

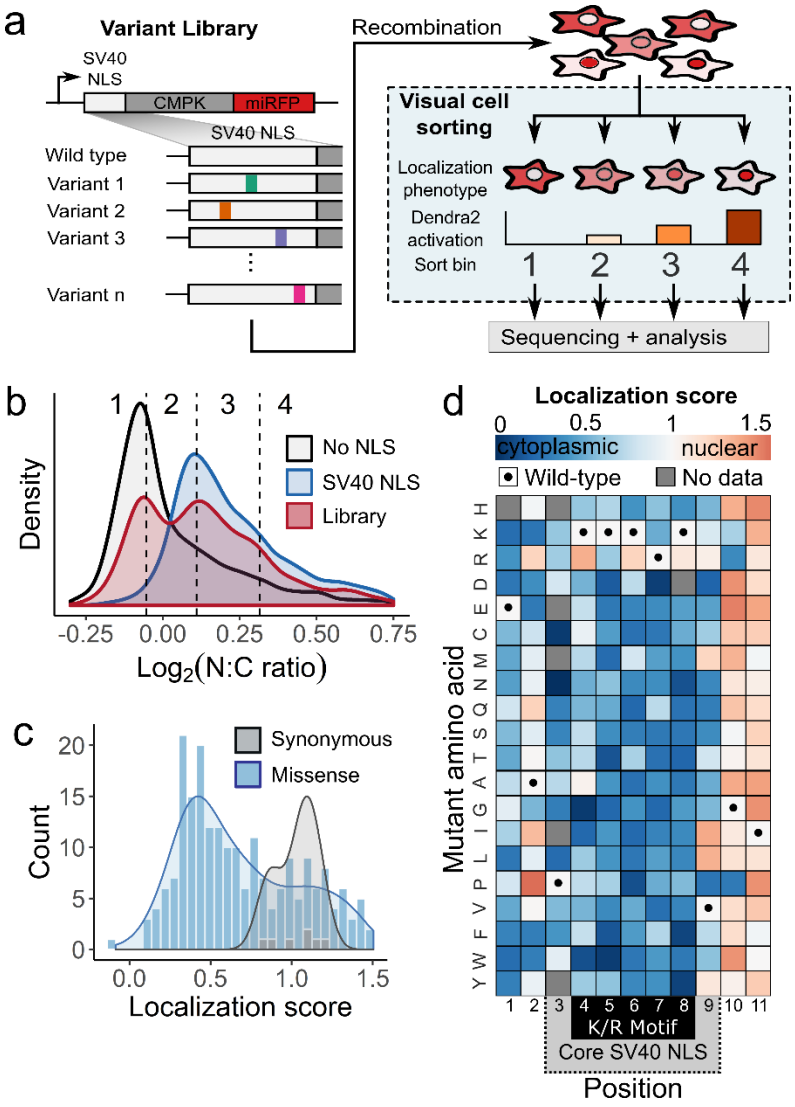


Figure 2.1. Visual Cell Sorting for pooled, image-based genetic screening. (A) A mutagenized simian virus (SV) 40 NLS library containing 346 unique nucleotide variants fused to a chicken muscle pyruvate kinase (CMPK) miRFP reporter was recombined into a U-2 OS H3-Dendra2 cell line. Visual Cell Sorting was performed to separate the NLS library expressing cells into four photoactivation bins according to the microscope-derived nucleus-to-cytoplasm ratio of the miRFP reporter. Each bin was deeply sequenced and analyzed to assign each amino acid variant a quantitative nuclear localization score. **(B)** U-2 OS H3-Dendra2 cells expressing either the NLS library, a wild type control or a no NLS control were imaged at 20X magnification and nucleus to cytoplasm (N:C) ratios measured. Curves, estimated kernel density of cells (N = 1,529, 3,269, and 3,931 cells for no NLS, SV40 NLS, and WT NLS, respectively); dotted lines, Visual Cell Sorting photoactivation gates with associated bin numbers. **(C)** Raw variant nuclear localization scores were calculated using a scaled weighted average of variant frequencies across the four sort bins. WT-like variants have a score of 1 and cytoplasm-localized variants a score of 0. Localization score, mean values of normalized scores from 5 replicates (N = 637,605 cells); curves, kernel density estimate of variant score distributions. (D) Nuclear localization scores of missense variants (N = 202) displayed as a heatmap. Gray boxes, variants not observed or scored in a single replicate; black dots, WT sequence; dotted gray area on the horizontal axis, SV40 NLS often used to localize recombinant proteins to the nucleus; black box, the five residue K/R-rich region.

Figure 2.2

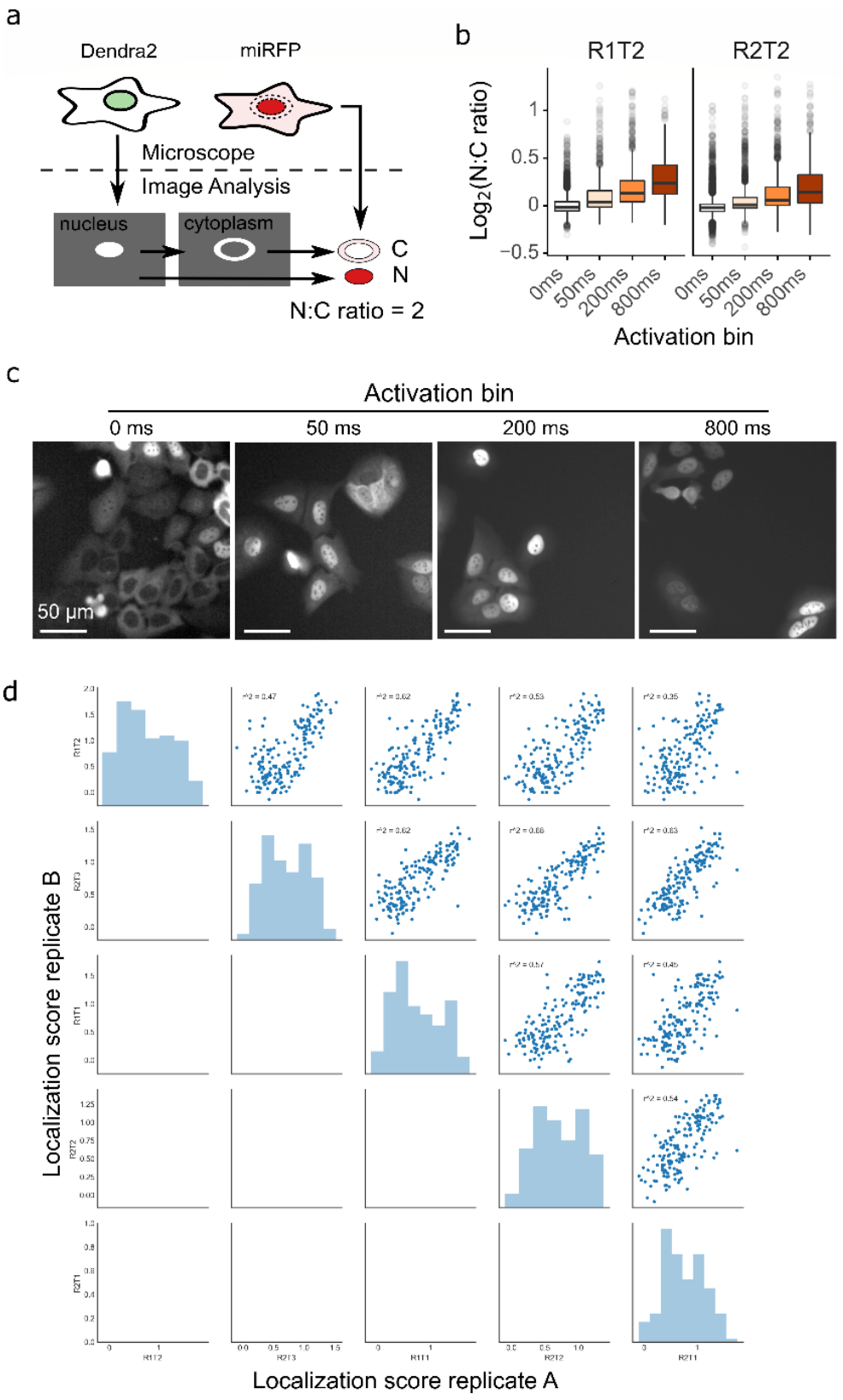

a
Dendra2
miRFP
Microscope
Image Analysis
nucleus
cytoplasm
C
N
N:C ratio = 2
b
R1T2
R2T2
Log2(N:C ratio)
1
0.5
0
−0.5
0ms
50ms
200ms
800ms
Activation bin
c
Activation bin
0 ms
50 ms
200 ms
800 ms
50 µm
d
Localization score replicate B
Localization score replicate A

Figure 2.2. Supplementary information regarding Visual Cell Sorting for pooled, image-based genetic screening. (A) Image analysis pipeline to calculate nucleus to cytoplasm (NC) ratio. Nuclei were segmented using the H3-Dendra2 signal. Cytoplasmic masks were created by dilating and then removing the nuclear mask. Mean miRFP intensity was measured within each mask and the nucleus to cytoplasm (N:C) ratio calculated. (B) After selective photoactivation on the microscope based on N:C ratio, cells were subject to fluorescence-activated cell sorting and sorted according to their nuclear localization phenotype. Two days after sorting, cells from each sort bin were re-imaged in the miRFP channel and the nucleus to cytoplasm ratio reassessed (n = ~1,500 per photoactivation bin). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. R1T2, recombination replicate 1, Visual Cell Sorting technical replicate 2 (C) Representative images of sorted cells. (D) Correlation plots of normalized scores calculated for each replicate. r^2, square of Pearson's correlation coefficient.

Figure 2.3

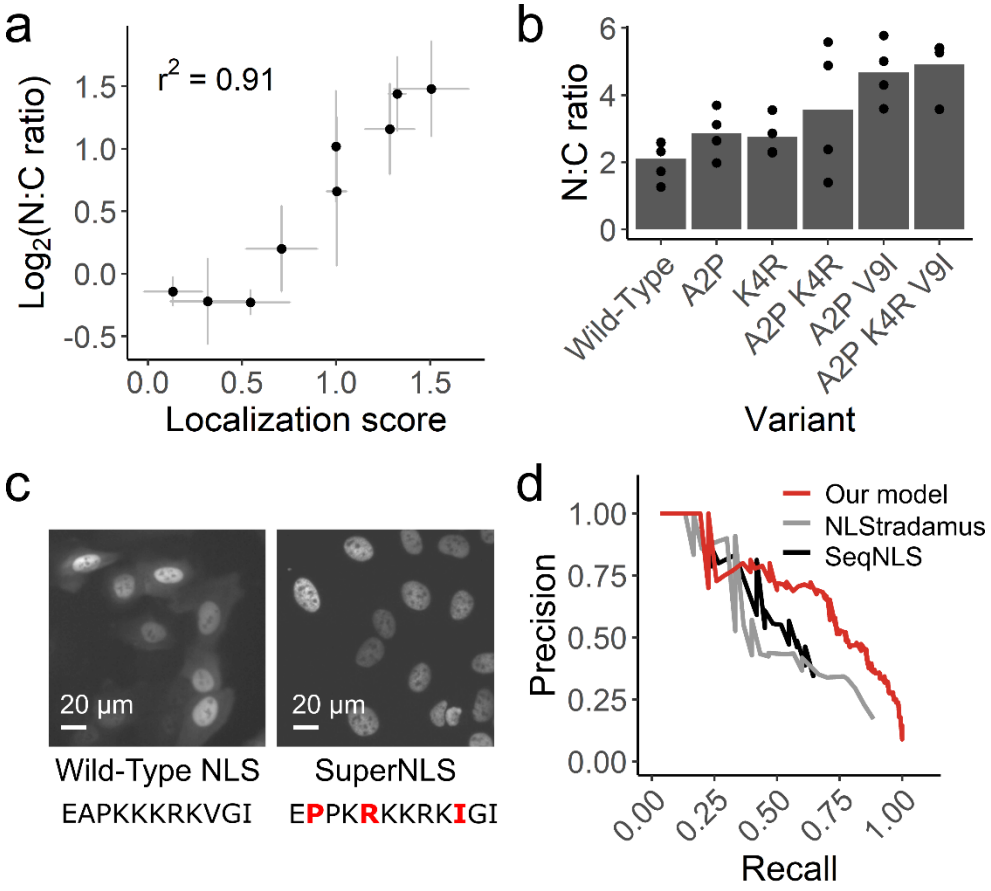


Figure 2.3. Visual Cell Sorting-derived variant scores accurately predict NLS function. (A) Nine NLS variants were individually expressed in the CMPK-miRFP reporter in U-2 OS H3- Dendra2 cells. The median nucleus to cytoplasm (N:C) ratio of cells expressing each variant was measured by microscope and compared to its localization score derived by Visual Cell Sorting. n ≥ 141 cells per variant per replicate. Bars, mean across at least three separate replicates. (B) SV40 NLS variants that appeared to enhance nuclear localization were individually tested both alone and in combination. NLS variants with up to three amino acid changes were expressed in U-2 OS H3-Dendra2 cells and imaged; the median N:C ratio was quantified across cells in the same well. n ≥ 527 cells per variant per replicate. (C) Representative images from cells expressing the wild-type SV40 NLS or the optimized superNLS fused to the miRFP reporter. Scale bars = 20 µm; red letters, amino acid differences from wild-type. (D) Nuclear localization scores derived from Visual Cell Sorting were used to generate a predictive model that was trained on UniProt NLS annotations. Precision/recall curves for our model and two other linear motif scoring models, NLStradamus (Nguyen Ba et al. 2009) and SeqNLS (Lin et al. 2013), on a test dataset (n = 30 NLSs) are shown.

Figure 2.4

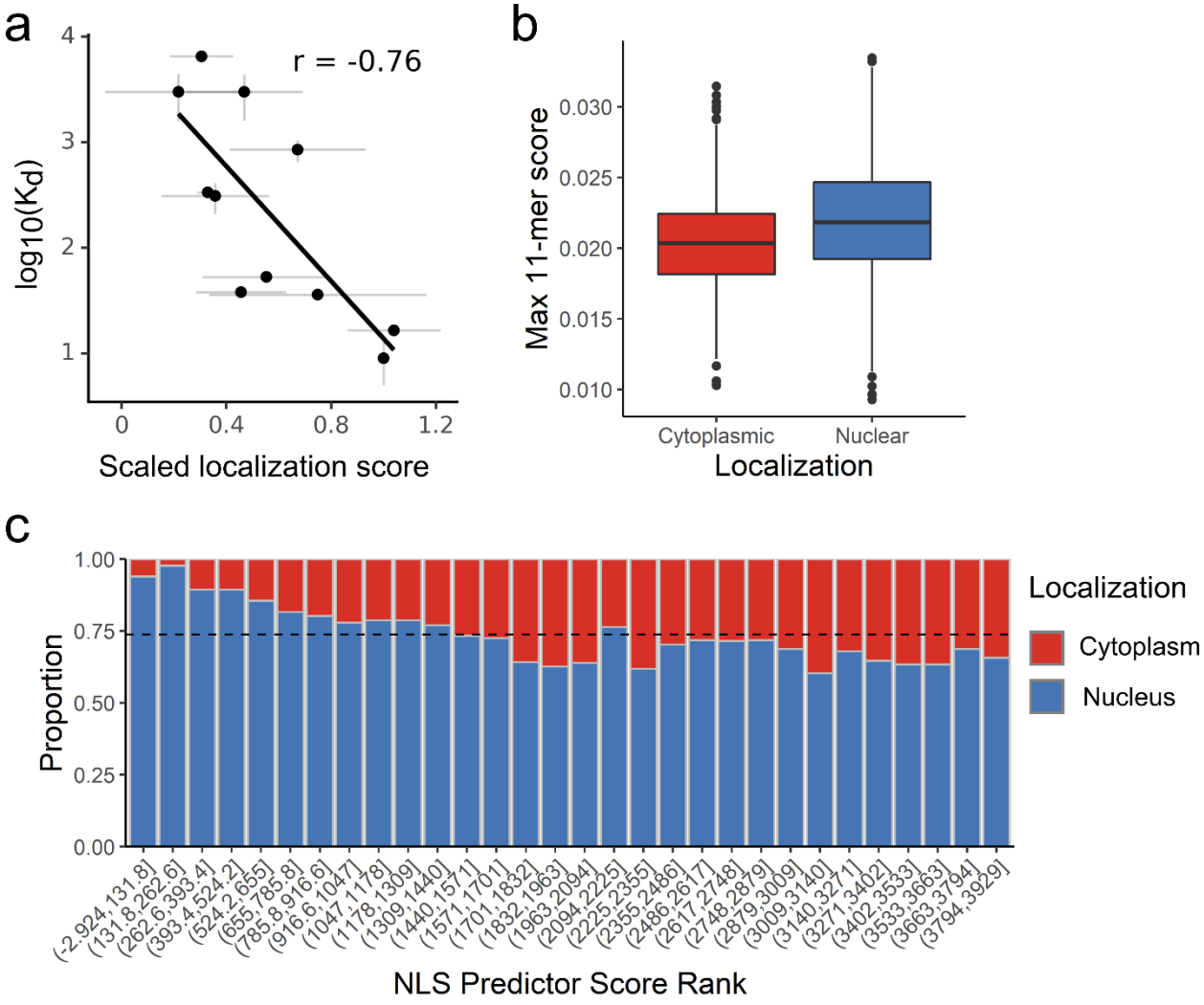


Figure 2.4. Supplementary information regarding Visual Cell Sorting-derived variant scores accurately predict NLS function. (A) Dissociation constants measuring binding between SV40 NLS variants and importin alpha, as reported by Hodel and colleagues (2001) were plotted against the variants' mean normalized scores. Grey bars, standard error from the mean. r, Pearson's correlation coefficient. (B) All 11-mers in proteins annotated as exclusively cytoplasmic or exclusively nuclear by the Human Protein Atlas were subject to our NLS prediction model; the top-scoring 11-mer within each protein was extracted for each group (N = 3,925 11-mers; Wilcoxon rank sum p value < 10^{-16}). (C) Each protein's top-scoring 11-mer was ranked and binned according to its score (n = 131 proteins per bin). Dotted line, expected proportion of nuclear proteins per bin if the model has no predictive power.

Figure 2.5

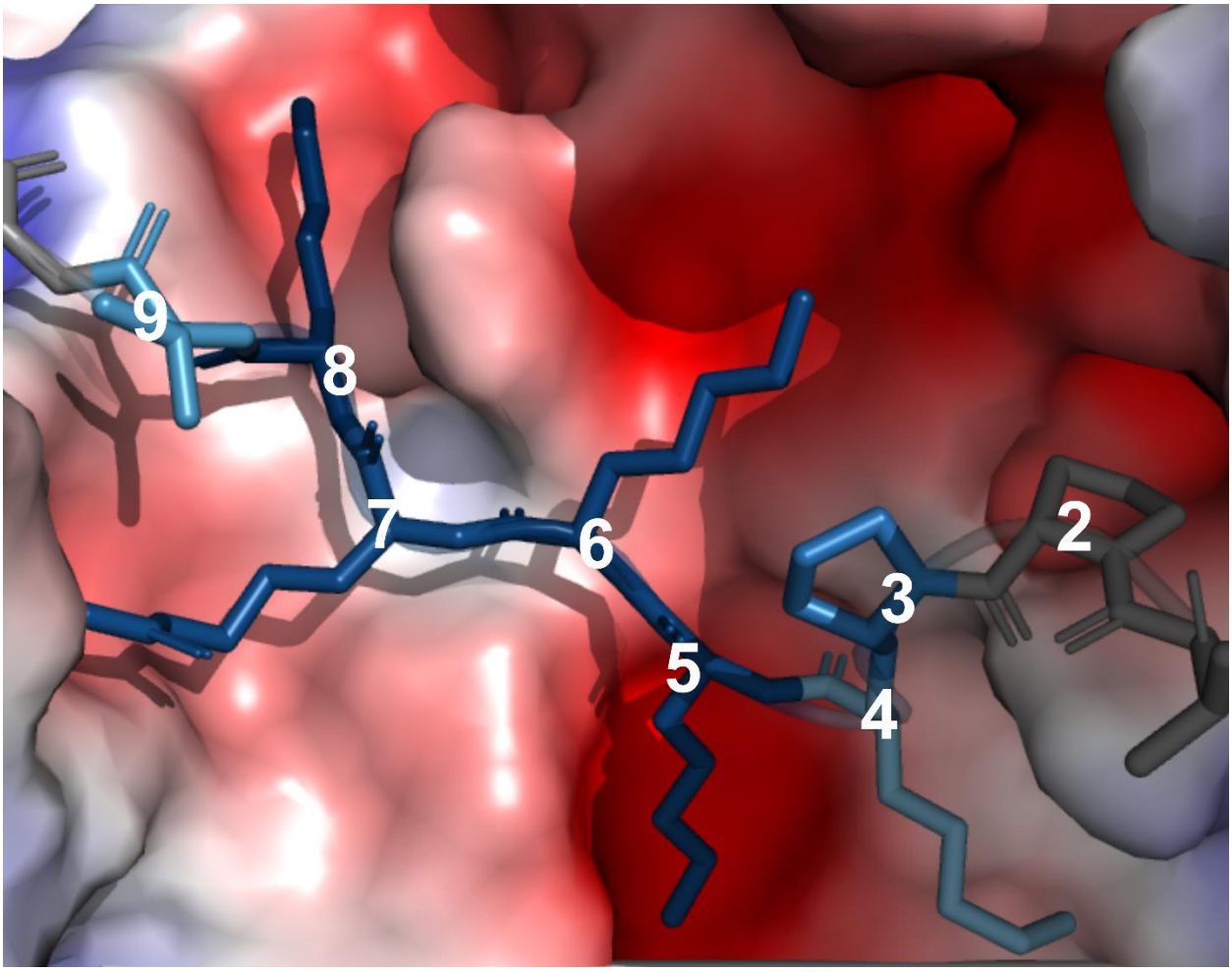


Figure 2.5. NLS structure mapped onto the major groove. Residues that were designated as “wild-type” residues in the NLS image-based, pooled genetic screen were colored according to the median effect of an amino acid substitution at their position. Residue 2 is a proline residue in the SV40-NLS, but is not included as part of the NLS in many annotations of the SV40 Large T Antigen; in our screen, the wild-type residue at this position was alanine.

Figure 3.1

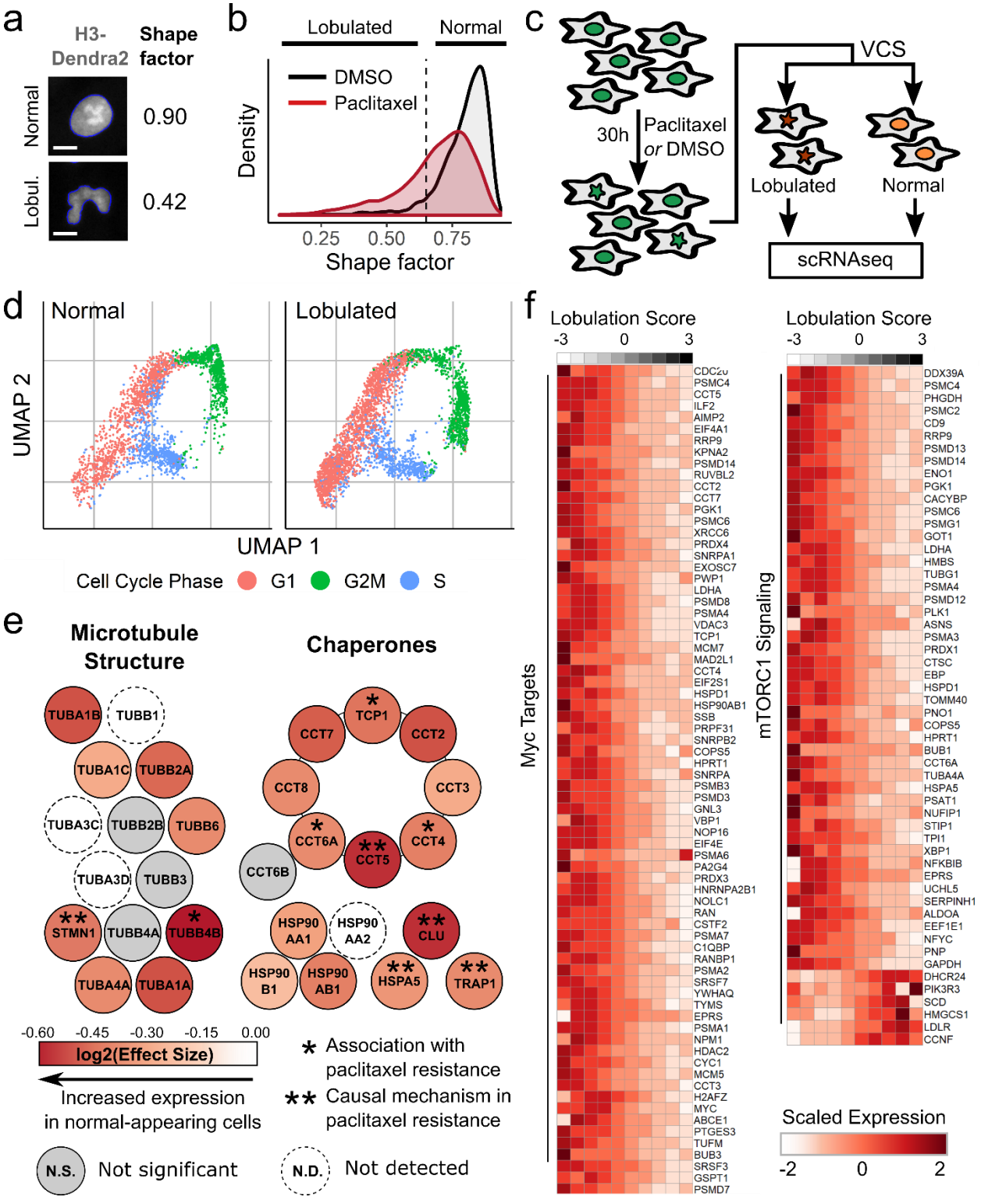


Figure 3.1. Visual Cell Sorting to dissect heterogeneous nuclear morphology following paclitaxel treatment. (a) RPE-1 NLS-Dendra2x3 cells were treated for 24 hours with 0.25 nM paclitaxel or DMSO

and imaged. The shape factor, which measures the degree of an object's circularity, was computed for each nucleus. One normal nucleus with a shape factor near one and one lobulated nucleus with a low shape factor are shown. The computationally determined boundaries of each nucleus are shown in blue; scale bar = 10 μm. **(b)** Shape factor density plots for vehicle (DMSO) and 0.25 nM paclitaxel-treated RPE-1 cells (n ≥ 3,914 cells per treatment). Dashed line, cutoff for lobulated nuclei (shape factor < 0.65). **(c)** RPE-1 cells were treated with 0.25 nM paclitaxel, then subjected to Visual Cell Sorting according to nuclear shape factor. Populations of cells with normal or lobulated nuclei were subjected separately to single cell RNA sequencing. **(d)** UMAP analysis of single cell RNA sequencing results of paclitaxel-treated cells. Expression of cell-cycle related genes were used to annotate each cell as being in G1, S, or G2/M. **(e)** A differential gene test was performed using as covariates cell cycle scores and a lobulation score, which is higher in lobulated cells compared to morphologically normal cells (Figure 3.2). Genes related to microtubule structure or various chaperone complexes are colored according to the expected $\log_2$ fold-change per unit increase in lobulation score (effect size); asterisks, genes associated with paclitaxel resistance (Di Michele *et al*, 2009; Ooe *et al*, 2007; Su *et al*, 2009; Alli *et al*, 2007; Dorman *et al*, 2016; Li *et al*, 2013). **(f)** Expression counts for genes associated with c-Myc and mTORC1 signaling were aggregated across cells binned according to their lobulation score, then log-normalized and rescaled. Higher lobulation scores correspond to a higher likelihood of nuclear lobulation.

Figure 3.2

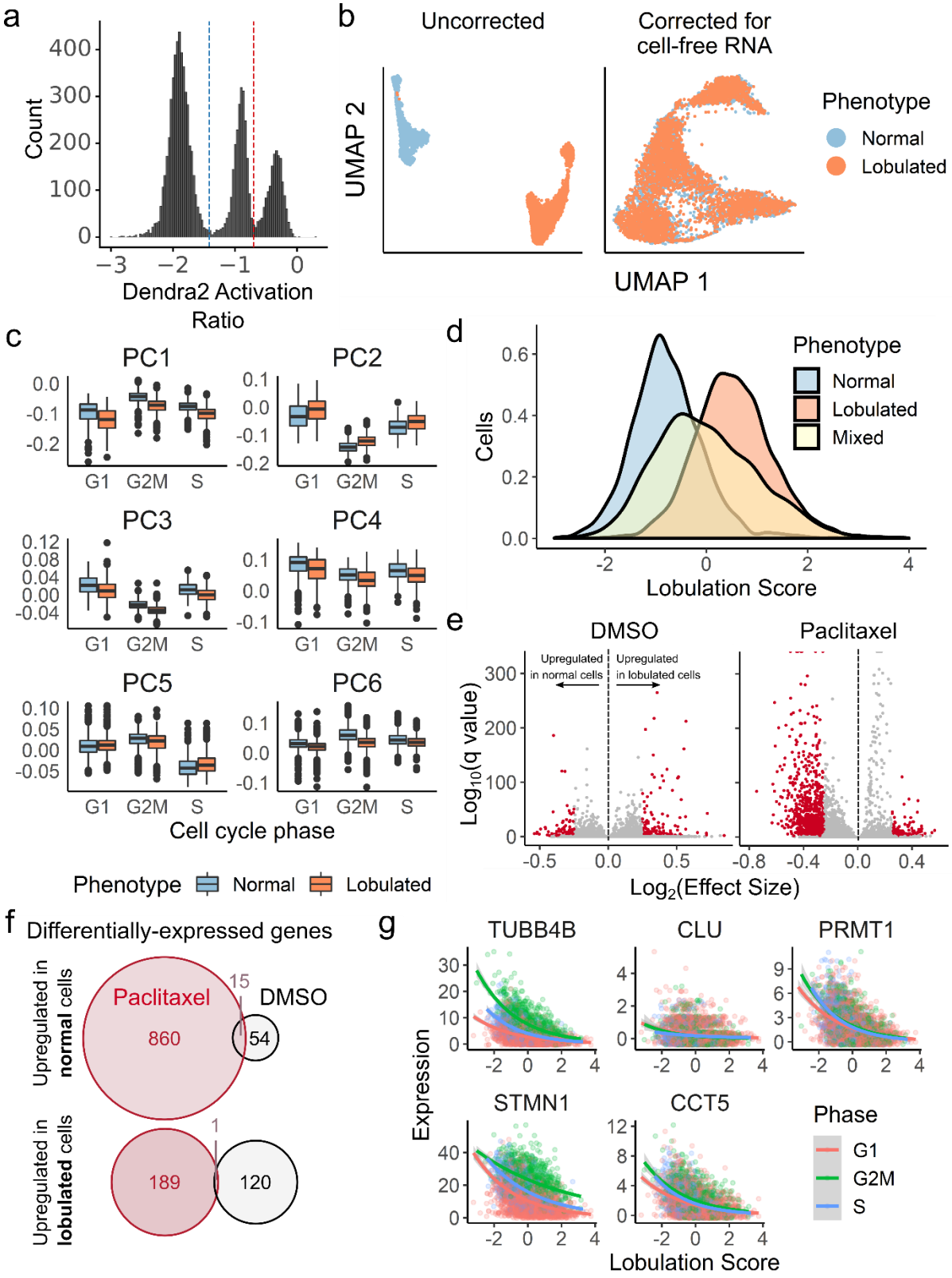
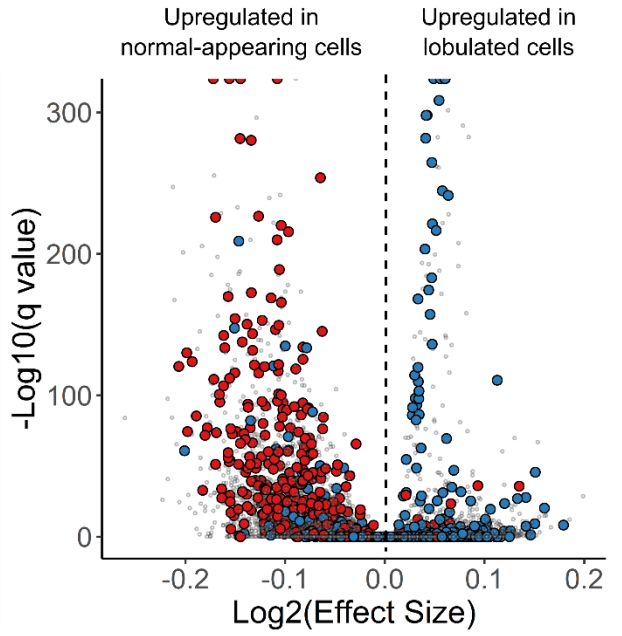

a
Count
Dendra2 Activation Ratio
b
Uncorrected
Corrected for cell-free RNA
UMAP 2
UMAP 1
Phenotype
Normal
Lobulated
c
PC1
PC2
PC3
PC4
PC5
PC6
G1
G2M
S
Cell cycle phase
Phenotype
Normal
Lobulated
d
Cells
Lobulation Score
Phenotype
Normal
Lobulated
Mixed
e
DMSO
Paclitaxel
Upregulated in normal cells
Upregulated in lobulated cells
Log10(q value)
Log2(Effect Size)
f
Differentially-expressed genes
Upregulated in normal cells
Upregulated in lobulated cells
Paclitaxel
DMSO
860
15
54
189
1
120
g
TUBB4B
CLU
PRMT1
STMN1
CCT5
Expression
Lobulation Score
Phase
G1
G2M
S

Figure 3.2. (a) Photoactivation gates for Visual Cell Sorting. Cells between the blue and red dotted lines represent putative normal nuclear shape factor cells activated with 405 nm light for 200 ms; and cells above the red dotted line represent putative low nuclear shape factor (lobulated) cells activated for 800 ms. Cells below the blue dotted line were not imaged or not activated **(b)** UMAP projection of the single cell transcriptomes derived from Visual Cell Sorting-separated lobulated and normal cells before and cell-free RNA correction with the algorithm used by SoupX (Young & Behjati, 2018) **(c)** Visual Cell Sorting-separated cells were aligned to an unseparated, treatment-matched population with the mutual nearest neighbors algorithm (Haghverdi *et al*, 2018). The first six principle components of the separated paclitaxel-treated cells, subset by nuclear phenotype and cell cycle stage, are shown. **(d)** Lobulation scores were generated using linear combinations of principle components 1-4. Scores for phenotypically normal and lobulated paclitaxel-treated cells (Experiment 1, N = 2,724 and 3,934, respectively) and unseparated paclitaxel-treated cells (Experiment 2, N = 4,066) are shown. **(e)** Volcano plots showing DEGs significantly correlated with the lobulation score in the unseparated, vehicle (DMSO) and paclitaxel-treated populations. Red points, significant DEGs with a $\log_2$(Effect Size), which estimates the expected log2 fold-change per unit increase in lobulation score, greater than 0.25 and a q-value less than 0.01. **(f)** Raw gene expression counts of selected significant DEGs of cells in the unseparated, paclitaxel-treated population versus the cells' lobulation scores. Colored lines, negative binomial regression model stratified by cell cycle stage.

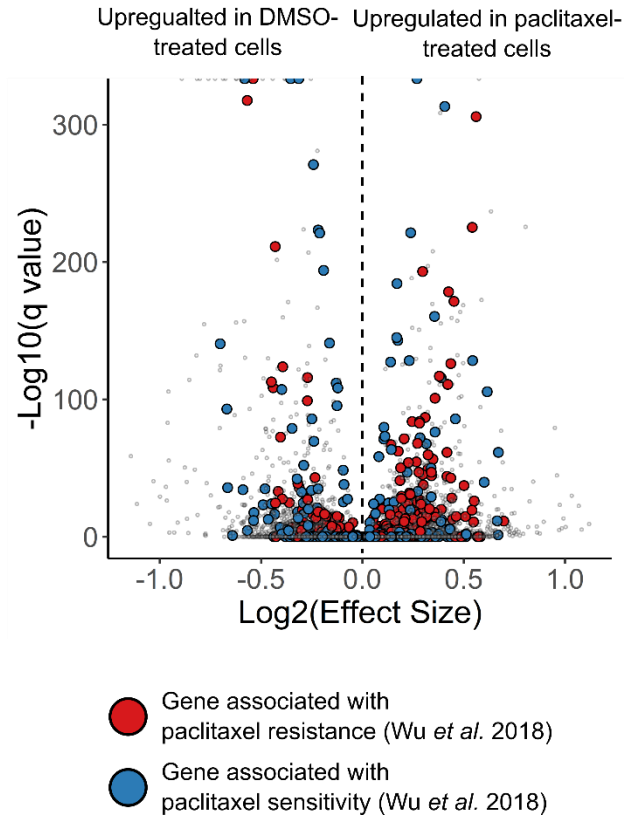
Figure 3.3

A



	Upregulated in normal appearing	Upregulated in lobulated
Sensitivity-associated	61	104
Resistance-associated	313	16

B



	Upregulated in paclitaxel-treated	Upregulated in DMSO-treated
Sensitivity-associated	72	67
Resistance-associated	70	126

Figure 3.3. Nuclear morphology is a better predictor of paclitaxel drug resistance than is acute drug treatment. The effect size and statistical significance of differentially expressed genes (DEGs) detected between cells with high and low lobulation scores (A) and between cells that were treated with paclitaxel or DMSO for 40 hours (B). Both DEG tests were performed using cells from Experiment 2. Genes also identified as being associated with paclitaxel resistance or sensitivity by Wu et al. are colored blue or red. Left, volcano plot: q value, BH-corrected p values, log2(effect size), expected log 2-fold change per unit increase in lobulation score, or between the average DMSO-treated and paclitaxel-treated cell. Right, confusion table of the significant DEGs (q value < 0.01) shared between the volcano plot shown on the left and those reported in paclitaxel-resistant esophageal carcinoma cells reported by Wu and colleagues (2018).

Figure 4.1

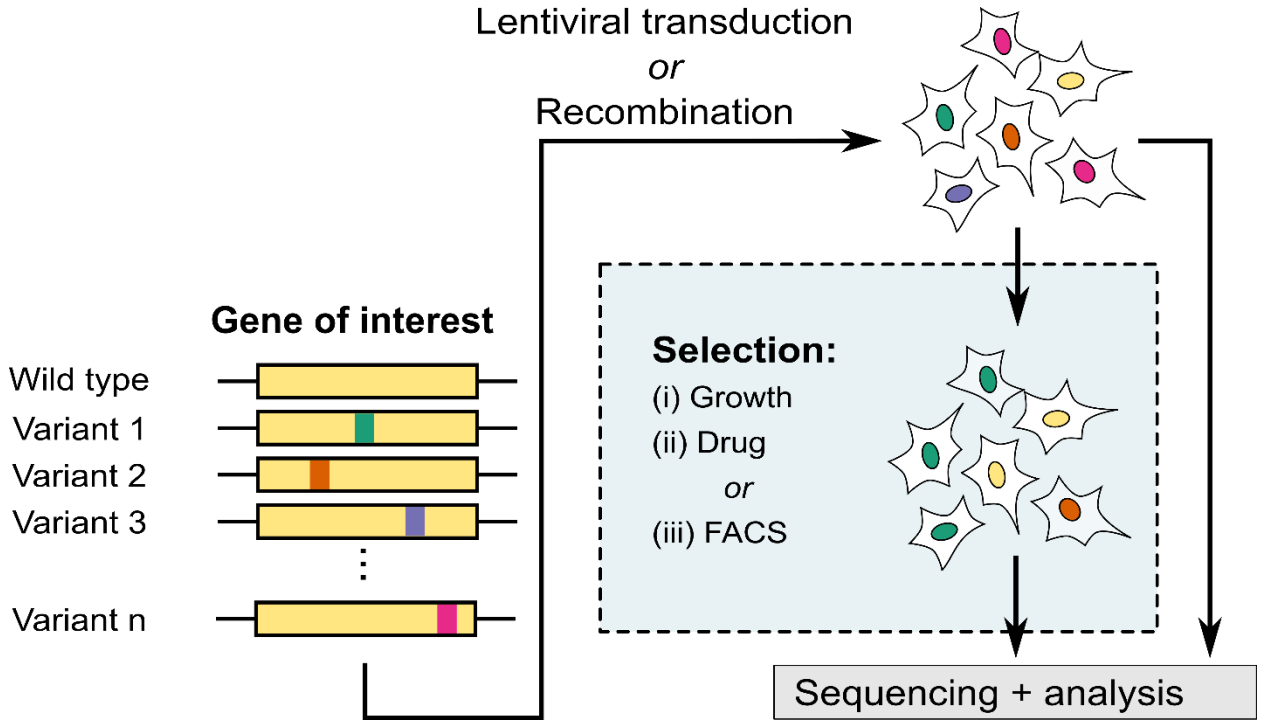


Figure 4.1. **Multiplexed Assay of Variant Effect (MAVE).** A variant library is generated for a gene or genomic region of interest, and the library is introduced into cells via lentiviral transduction or recombination such that each cell expresses a single genetic variant. A selection pressure whose strength is determined by the function of each genetic variant is applied to the population of cells. The variant frequencies of cells that were not and were subject to the selection pressure are determined using next generation sequencing; and a score for each variant is calculated using the ratio of the variant frequencies before and after selection.

Figure 4.2

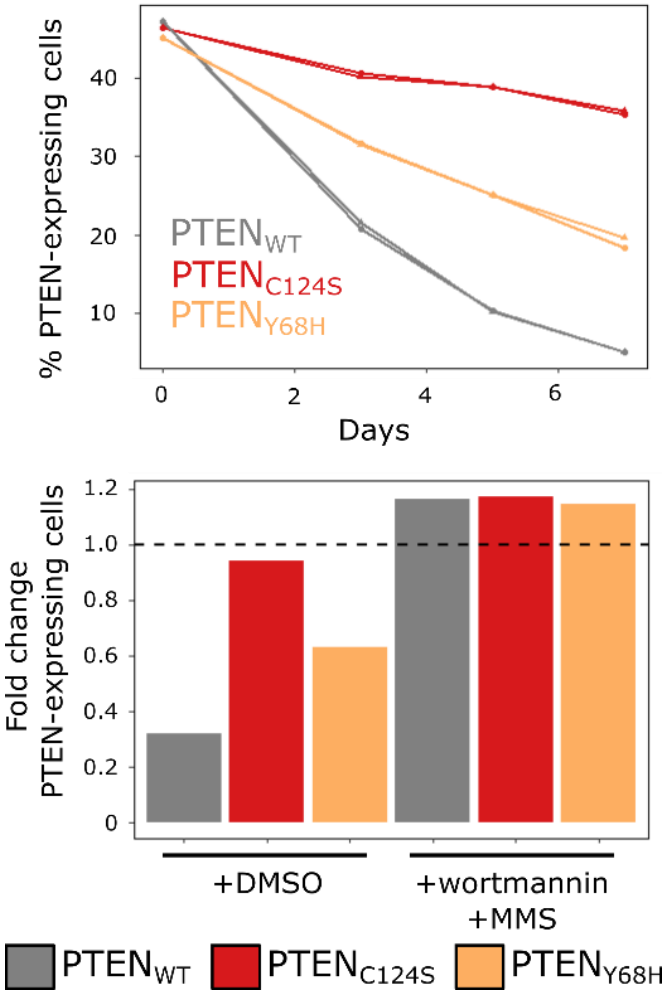


Figure 4.2. PTEN expression in U87MG cells affects cell growth and genome stability. (A) PTEN decreases cell growth in lipid phosphatase-dependent manner. U87MG cells were transduced with lentivirus containing PTEN variants linked to mCherry with a 2A peptide. On Day 0, each PTEN variant population was mixed at a 50:50 ratio with parental U87MG cells (PTEN-null). The percentage of PTEN-expressing cells in the population was tracked via flow cytometry for mCherry. Results of two replicates shown. **(B)** PTEN increases survival after DNA damage in lipid-phosphatase independent manner. U87MG cells were transduced with PTEN-2A-mCherry lentivirus. On Day 0, the proportion of PTEN-2A-mCherry-expressing cells was measured with FACS for, then DMSO or 2uM wortmannin (PI3K inhibitor) and 400uM MMS (alkylating agent) was added. PTEN-expressing proportion was re-measured on Day 4. Fold change calculated by dividing Day4 by Day0 mCherry+ proportions. Fold change of 1 is PTEN-null-like. C124S = lipid-phosphatase deficient Y68H = nuclear localized and low abundance.

Figure 4.3

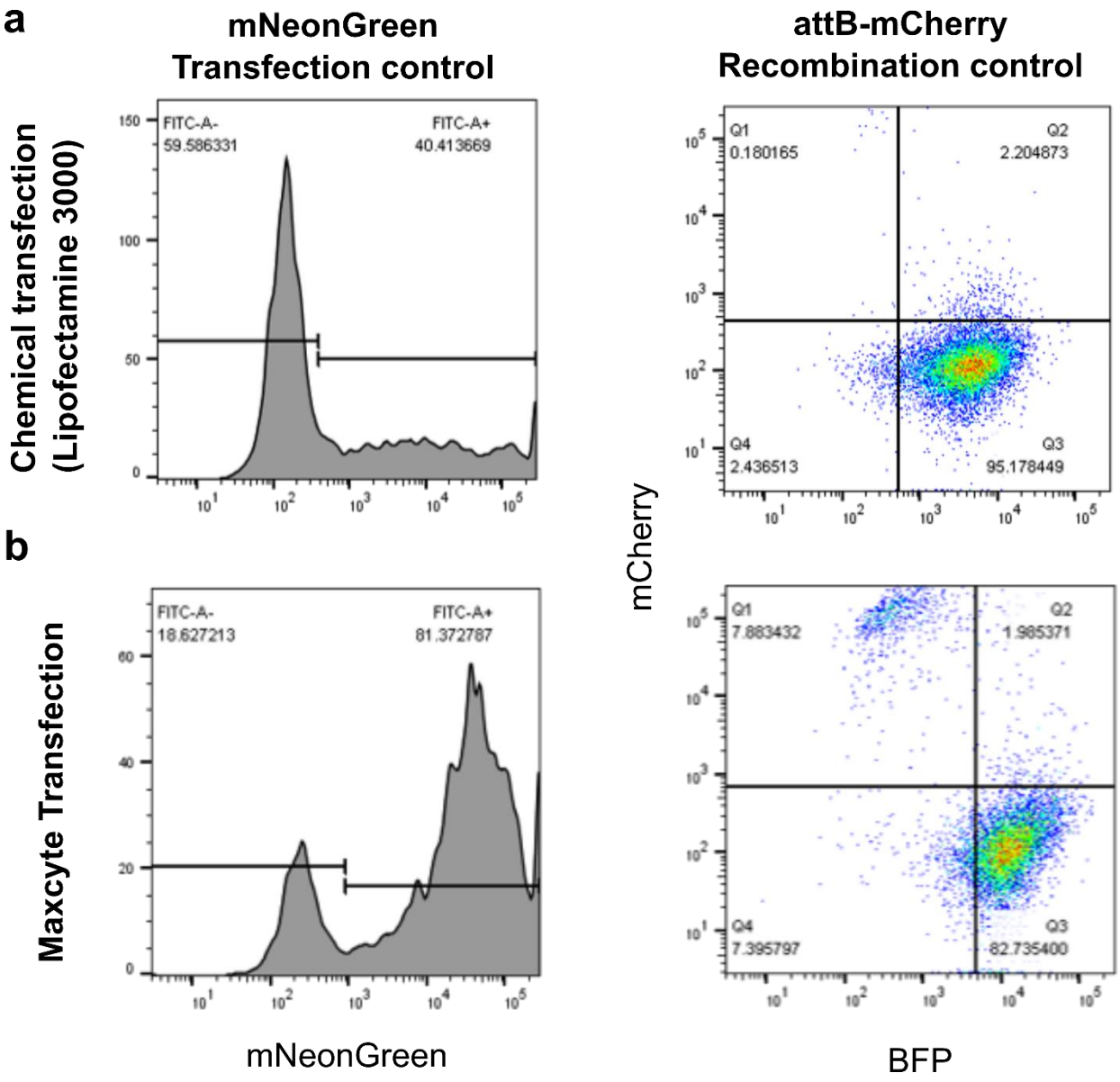


Figure 4.3. Transfection and recombination of U87MG Clone 3. Left; transfection efficiency as measured by an mNeonGreen positive control 2 days after transfection. Right; recombination efficiency (Q1, upper left quadrant) measured 6 days after transfection and two days after addition of doxycycline. **(a)** shows transfection with Lipofectamine 3000 according to manufacturer's instructions. **(b)** shows transfection with the MaxCyte as described in the Methods section of Chapter 4.

Figure 4.4

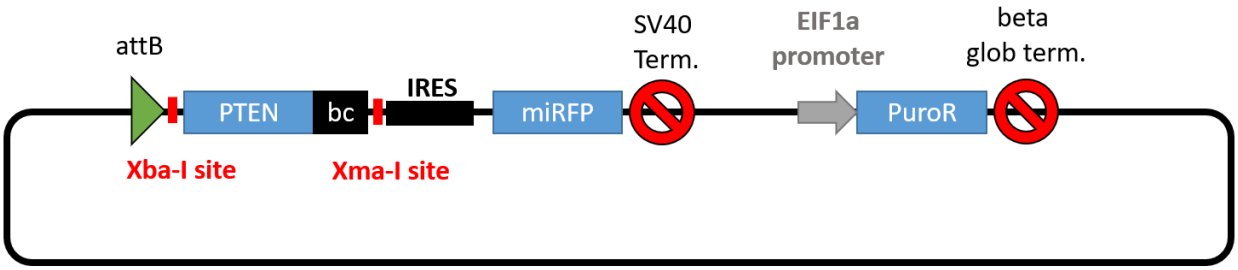


Figure 4.4. Recombination vector for use in *PTEN* MAVE experiments. attB, Bxb1 recombinase site; bc, variant barcode region; IRES, internal ribosomal entry site; SV40 Term., SV40 terminator sequence; PuroR, puromycin resistance marker; beta glob term, beta globin terminator sequence.

Tables

Table 1.1

Method type	Method	Recovery of live cell subpopulations?	Image-based genetic screening capabilities	Transcriptomics based on visual phenotypes	Number of recovery bins	Imaging modality	Single cell resolution?	Custom Hardware?	Dye-based disposable reagents?	Demonstrated scale / experiment (mammalian cells)	Time / experiment	Hardware requirements	Dye-based imaging reagents	Genetic engineering required	PMID
Multi-well	**High content imaging**	No	Yes	No	NA	Confocal, widefield	Yes	No	Yes	1e6 - 1e7	Hours - days	Specialized ultra-high throughput microscope; robotics for reagent preparation	Fluorescent probes	Fluorescent protein markers	26638068
***In situ* nucleic acid methods**	**MERFISH**	No	Yes	Yes	NA	Confocal, widefield	Yes	Yes	Yes	~8,000	~20 hours	Confocal scanning microscope or widefield microscope with custom-built microfluidics chamber	Array-derived probe library; fluorescent readout probes	None	25858977, 29083401, 31085639
	seqFISH+	No	No	Yes	NA	Confocal	Yes	Yes	Yes	200-500	*Not reported*	Confocal scanning microscope with custom-built microfluidics chamber	Array-derived probe library; fluorescent readout probes	None	30911168
	in situ sequencing (transcriptomics)	No	No	Yes	NA	Confocal	Yes	Yes	Yes	*Not reported*	16 days	Confocal scanning microscope	Next-generation sequencing kit-derived fluorescent probes	None	23852452, 24578530, 25675209
	***in situ* sequencing (pooled optical screens)**	No	Yes	No	NA	Widefield	Yes	No	Yes	~3,000,000	5 days	Standard widefield microscope	Next-generation sequencing kit-derived fluorescent probes	None	31626775
Photoactivation-based	**CLaP**	Yes	Yes	Yes	2	Confocal	Yes	No	Yes	100-1,000	*Not reported*	Confocal microscope	Biotin-4-fluorescein; Cy5-streptavidin	None	27198043
	Niche-seq	Yes	No	Yes	2	Two-photon	No	No	No	1,000 -10,000	*Not reported*	Two-photon microscope	None	PA-GFP stably expressed	29217582
	Optomagnetic cell capture	Yes	Yes	Yes	2	Confocal	Yes	Yes	Yes	5-30	*Not reported*	Confocal microscope, custom chamber	Biotin-4-fluorescein; magnetic bead coated with streptavadin	None	30969169
	Optical painting	Yes	Yes	Yes	2	Confocal	Yes	Yes	Yes	10-150	*Not reported*	Custom confocal microscope	Antibody-conjugated quantum dots made in-house	None	27118210
	Photostick	Yes	Yes	Yes	2	Widefield	Yes	No	Yes	10-100	*Not reported*	Widefield microscope with digital micromirror device	Cy3- or Cy5-SBED	None	25705368
	Visual Cell Sorting	Yes	Yes	Yes	4	Widefield	Yes	No	No	160,000	12-15 hours	Widefield microscope with digital micromirror device	None	H3-Dendra2 or NLS-Dendra2x3 stably expressed	Current Work

Table 2.1

Replicate	# cells bin 1 (no activation)	# cells bin 2 (50ms)	# cells bin 3 (200ms)	# cells bin 4 (800ms)	Total number of cells sorted	Total microscope time (hr)
R1T1	94122	16752	23761	12939	147574	7.6
R1T2	91770	23307	30495	16189	161761	13.4
R2T1	55237	9794	14636	8979	88646	10.3
R2T2	44305	6708	11721	7332	70066	9.1
R2T3	109122	15957	30135	14344	169558	14.8
				Grand total:	**637605**	**55.1**

Table 2.1. Number of cells sorted for each replicate of the image-based screen of SV40 NLS variants.

References

Alli E, Yang JM, Ford JM & Hait WN (2007) Reversal of stathmin-mediated resistance to paclitaxel and vinblastine in human breast carcinoma cells. *Mol. Pharmacol.* **71:** 1233–1240

Andrés-Pons A, Rodríguez-Escudero I, Gil A, Blanco A, Vega A, Molina M, Pulido R & Cid VJ (2007) In vivo functional analysis of the counterbalance of hyperactive phosphatidylinositol 3-kinase p110 catalytic oncoproteins by the tumor suppressor PTEN. *Cancer Res.* **67:** 9731–9739

Azizoglu A & Stelling J (2019) Controlling cell-to-cell variability with synthetic gene circuits. *Biochem. Soc. Trans.* **47:** 1795–1804

Bassi C, Ho J, Srikumar T, Dowling RJO, Gorrini C, Miller SJ, Mak TW, Neel BG, Raught B & Stambolic V (2013) Nuclear PTEN controls DNA repair and sensitivity to genotoxic stress. *Science (80-.).* **341:** 395–399

Beg AA, Ruben SM, Scheinman RI, Haskill S, Rosen CA & Baldwin AS (1992) IκB interacts with the nuclear localization sequences of the subunits of NF-κB: A mechanism for cytoplasmic retention. *Genes Dev.* **6:** 1899–1913

Binan L, Bélanger F, Uriarte M, Lemay JF, Pelletier De Koninck JC, Roy J, Affar EB, Drobetsky E, Wurtele H & Costantino S (2019) Opto-magnetic capture of individual cells based on visual phenotypes. *Elife* **8:** 1–21

Binan L, Mazzaferri J, Choquet K, Lorenzo LE, Wang YC, Affar EB, De Koninck Y, Ragoussis J, Kleinman CL & Costantino S (2016) Live single-cell laser tag. *Nat. Commun.* **7:** 1–8

Bloom JD (2014) An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit Article Fast Track. **31:** 1956–1978

Boutros M, Heigwer F & Laufer C (2015) Microscopy-Based High-Content Screening. *Cell* **163:** 1314–1325

Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC & Carpenter AE (2016) Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11:** 1757–1774

Butler A, Hoffman P, Smibert P, Papalexi E & Satija R (2018) Integrating single-cell

transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36:** 411–420

Butler MG, Dazouki MJ, Zhou XP, Talebizadeh Z, Brown M, Takahashi TN, Miles JH, Wang CH, Stratton R, Pilarski R, et al (2005) Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J. Med. Genet.* **42:** 318–321

Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, Vasilevich AS, Barry JD, Bansal HS, Kraus O, et al (2017) Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14:** 849–863

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, et al (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-.).* **357:** 661–667

Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348:** aaa6090

Chien M-P, Werley CA, Farhi SL & Cohen AE (2015) Photostick: a method for selective isolation of target cells from culture. *Chem. Sci.* **6:** 1701–1705

Cho JH, Lee MK, Yoon KW, Lee J, Cho SG & Choi EJ (2012) Arginine methylation-dependent regulation of ASK1 signaling by PRMT1. *Cell Death Differ.* **19:** 859–870

Christiansen EM, Yang SJ, Ando DM, Rubin LL, Nelson P, Finkbeiner S, Christiansen EM, Yang SJ, Ando DM, Javaherian A, et al (2018) In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell***:** 1–12

Chudakov DM, Lukyanov S & Lukyanov KA (2007) Tracking intracellular protein movements using photoswitchable fluorescent proteins PS-CFP2 and Dendra2. *Nat. Protoc.* **2:** 2024–2032

Chudakov DM, Verkhusha V V, Staroverov DB, Souslova EA, Lukyanov S & Lukyanov KA (2004) Photoswitchable cyan fluorescent protein for protein tracking. *Nat. Biotechnol.* **22:** 1435–1439

Conti E, Uy M, Leighton L, Blobel G & Kuriyan J (1998) Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin α. *Cell* **94:** 193–204

Darwin CR (1845) Proceedings of the Second Expedition. *J. Res. into Nat. Hist. Geol. Ctries. Visit. Dur. Voyag. H.M.S. Beagle round world, under Command Capt. Fitz Roy***:** 380

David E, Salame TM, Li H, De Giovanni M, Giladi A, Iannacone M, Shulman Z, Stoler-Barak L, Amit I, Biram A, et al (2017) Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science (80-.).* **358:** 1622–1626

Diez-Silva M, Dao M, Han J, Lim C-T & Suresh S (2010) Shape and Biomechanical Characteristics of Human Red Blood Cells in Health and Disease. *MRS Bull.* **35:** 382–388

Dorman SN, Baranova K, Knoll JHM, Urquhart BL, Mariani G, Carcangiu ML & Rogan PK (2016) Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol. Oncol.* **10:** 85–100

Emanuel G, Moffitt JR & Zhuang X (2017) High-throughput, image-based screening of pooled genetic-variant libraries. *Nat. Methods* **14:** 1159–1162

Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan G-C, et al (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*

Feldman D, Singh A, Schmid-Burgk JL, Carlson RJ, Mezger A, Garrity AJ, Zhang F & Blainey P (2019) Pooled optical screens in human cells. *Cell* **in press:** 383943

Fontes MR, Teh T & Kobe B (2000) Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-alpha. *J. Mol. Biol.* **297:** 1183–94

Fowler DM & Fields S (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods* **11:** 801–7

Gasic I, Boswell SA & Mitchison TJ (2019) Tubulin mRNA stability is sensitive to change in microtubule dynamics caused by multiple physiological and toxic cues. *PLoS Biol.* **17:** e3000225

Gasperini M, Starita L & Shendure J (2016) The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11:** 1782–1787

Georges E, Bonneau AM & Prinos P (2011) RNAi-mediated knockdown of α-enolase increases the sensitivity of tumor cells to antitubulin chemotherapeutics. *Int. J. Biochem. Mol. Biol.* **2:** 303–308

Georgescu MM, Kirsch KH, Kaloudis P, Yang H, Pavletich NP & Hanafusa H (2000) Stabilization and productive positioning roles of the C2 domain of PTEN tumor suppressor. *Cancer Res.* **60:** 7033–7038

Gest H (2004) The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, fellows of the Royal Society. *Notes Rec. R. Soc.* **58:** 187–201

Gu T, Zhang Z, Wang J, Guo J, Shen WH & Yin Y (2011) CREB is a novel nuclear target of PTEN phosphatase. *Cancer Res.* **71:** 2821–2825

Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C & Lander ES (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146:** 633–644

Haghverdi L, Lun ATL, Morgan MD & Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. **36:**

Hasle N, Matreyek KA & Fowler DM (2019) The impact of genetic variants on PTEN molecular functions and cellular phenotypes. *Cold Spring Harb. Perspect. Med.* **9:**

He J, Zhang Z, Ouyang M, Yang F, Hao H, Lamb KL, Yang J, Yin Y & Shen WH (2016) PTEN regulates EG5 to control spindle architecture and chromosome congression during mitosis. *Nat. Commun.* **7:** 1–13

Hierholzer K, Ullrich KJ, Henle W, Ludwig W, Heidenhain W, Fick W, Helmholtz W & Du Bois-Reymond W (1999) Origins of Renal Physiology (Dedicated to Carl Gottschalk) History of Renal Physiology in Germany during the 19th Century Key Words Renal physiology

Hodel MR, Corbett AH & Hodel AE (2001) Dissection of a nuclear localization signal. *J. Biol. Chem.* **276:** 1317–1325

Kalderon D, Roberts BL, Richardson WD & Smith AE (1984) A short amino acid sequence able to specify nuclear location. *Cell* **39:** 499–509

Kavallaris M (2010) Microtubules and resistance to tubulin-binding agents. *Nat. Rev. Cancer* **10:** 194–204

Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, hlby CW auml & Nilsson M (2013) In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10:** 1–6

Klein AD, Ferreira NS, Ben-Dor S, Duan J, Hardy J, Cox TM, Merrill AH & Futerman AH (2016)

Identification of Modifier Genes in a Mouse Model of Gaucher Disease. *Cell Rep.* **16:** 2546–2553

Kuo CT, Thompson AM, Gallina ME, Ye F, Johnson ES, Sun W, Zhao M, Yu J, Wu IC, Fujimoto B, et al (2016) Optical painting and fluorescence activated sorting of single adherent cells labelled with photoswitchable Pdots. *Nat. Commun.* **7:** 1–11

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Yang JL, Ferrante TC, et al (2014) Sequencing in Situ. *Science* **343:** 1360–1363

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, et al (2015) Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10:** 442–58

Lein E, Borm LE & Linnarsson S (2017) The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science (80-.).* **358:** 64–69

Li N, Zoubeidi A, Beraldi E & Gleave ME (2013) GRP78 regulates clusterin stability, retrotranslocation and mitochondrial localization under ER stress in prostate cancer. *Oncogene* **32:** 1933–1942

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP & Tamayo P (2015) The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1:** 417–425

Liddington R, Derewenda Z, Dodson G & Harris D (1988) Structure of the liganded T state of haemoglobin identifies the origin of cooperative oxygen binding. *Nature* **331:** 725–728

Lin J rong & Hu J (2013) SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. *PLoS One* **8:**

Lobo GP, Waite KA, Planchon SM, Romigh T, Nassif NT & Eng C (2009) Germline and somatic cancer-associated mutations in the ATP-binding motifs of PTEN influence its subcellular localization and tumor suppressive function. *Hum. Mol. Genet.* **18:** 2851–2862

Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15:** 1–21

Ludwig LS, Lareau CA, Ulirsch JC, Christian E, Muus C, Li LH, Pelka K, Ge W, Oren Y, Brack A, et al (2019) Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-

Cell Genomics. *Cell* **176:** 1325-1339.e22

Marchand C, Lemay G & Archambault D (2019) The Jembrana disease virus Rev protein: Identification of nuclear and novel lentiviral nucleolar localization and nuclear export signals. *PLoS One* **14:** e0221505

Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, et al (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50:** 874–882

Matreyek KA, Stephany JJ & Fowler DM (2017) A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45:**

Mattiazzi Usaj M, Sahin N, Friesen H, Pons C, Usaj M, Masinas MPD, Shuteriqi E, Shkurin A, Aloy P, Morris Q, et al (2020) Systematic genetics and single-cell imaging reveal widespread morphological pleiotropy and cell-to-cell variability. *Mol. Syst. Biol.* **16:**

McInnes L, Healy J & Melville J (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Mense SM, Barrows D, Hodakoski C, Steinbach N, Schoenfeld D, Su W, Hopkins BD, Su T, Fine B, Hibshoosh H, et al (2015) PTEN inhibits PREX2-catalyzed activation of RAC1 to restrain tumor cell invasion. *Sci. Signal.* **8:** 1–11

Mester J & Eng C (2013) When overgrowth bumps into cancer: The PTEN-Opathies. *Am. J. Med. Genet. Part C Semin. Med. Genet.*

Di Michele M, Della Corte A, Cicchillitti L, Del Boccio P, Urbani A, Ferlini C, Scambia G, Donati MB & Rotilio D (2009) A proteomic approach to paclitaxel chemoresistance in ovarian cancer cell lines. *Biochim. Biophys. Acta - Proteins Proteomics* **1794:** 225–236

Mighell TL, Evans-Dutson S & O'Roak BJ (2018) A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am. J. Hum. Genet.* **102:** 943–955

Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP & Zhuang X (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci.* **113:** 11046–11051

Morgens DW, Deans RM, Li A & Bassik MC (2016) Systematic comparison of CRISPR/Cas9

and RNAi screens for essential genes. *Nat. Biotechnol.* **34:** 634–636

Motta-Mena LB, Reade A, Mallory MJ, Glantz S, Weiner OD, Lynch KW & Gardner KH (2014) An optogenetic gene expression system with rapid activation and deactivation kinetics. *Nat. Chem. Biol.* **10:** 196–202

Myers MP, Pass I, Batty IH, Van Der Kaay J, Stolarov JP, Hemmings BA, Wigler MH, Downes CP & Tonks NK (1998) The lipid phosphatase activity of PTEN is critical for its tumor supressor function. *Biochemistry* **95:** 13513–13518

Myers MP, Stolarov JP, Eng C, Li J, Wang SI, Wigler MH, Parsons R & Tonks NK (1997) P-TEN, the tumor suppressor from human chromosome 10q23, is a dual-specificity phosphatase (cancer‾tyrosine phosphorylation‾signal transduction‾protein tyrosine phosphatase). *Biochemistry* **94:** 9052–9057

Nguyen Ba AN, Pogoutse A, Provart N & Moses AM (2009) NLStradamus: A simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* **10:** 1–11

Oikonomou CM & Jensen GJ (2016) A new view into prokaryotic cell biology from electron cryotomography. *Nat. Rev. Microbiol.* **14:** 205–220

Ooe A, Kato K & Noguchi S (2007) Possible involvement of CCT5, RGS3, and YKT6 genes up-regulated in p53-mutated tumors in resistance to docetaxel in human breast cancers. *Breast Cancer Res. Treat.* **101:** 305–315

Ounkomol C, Seshamani S, Maleckar MM & Collman F (2018) Label-free prediction of three-dimensional fluorescence images from transmitted light microscopy.

Papa A, Wan L, Bonora M, Salmena L, Song MS, Hobbs RM, Lunardi A, Webster K, Ng C, Newton RH, et al (2014) Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function. *Cell* **157:** 595–610

Parasido E, Avetian GS, Naeem A, Graham G, Pishvaian M, Glasgow E, Mudambi S, Lee Y, Ihemelandu C, Choudhry M, et al (2019) The Sustained Induction of c-MYC Drives Nab-Paclitaxel Resistance in Primary Pancreatic Ductal Carcinoma Cells. *Mol. Cancer Res.* **17:** 1815–1827

Piekarowicz K, Machowska M, Dratkiewicz E, Lorek D, Madej-Pilarczyk A & Rzepecki R (2017) The effect of the lamin A and its mutants on nuclear structure, cell proliferation, protein stability, and mobility in embryonic cells. *Chromosoma* **126:** 501–517

Polstein LR & Gersbach CA (2015) A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. *Nat. Chem. Biol.* **11:** 198–200

Qiu X, Hill A, Packer J, Lin D, Ma YA & Trapnell C (2017) Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14:** 309–315

Raharjo WH, Enarson P, Sullivan T, Stewart CL & Burke B (2001) Nuclear envelope defects associated with LMNA mutations cause dilated cardiomyopathy and Emery-Dreifuss muscular dystrophy. *J. Cell Sci.* **114:** 4447–4457

Ramkumar P, Kampmann M & Qian C (2018) CRISPR-based genetic interaction maps inform therapeutic strategies in cancer. *Transl. Cancer Res.* **7:** S61–S67

Rodríguez-Escudero I, Oliver MD, Andrés-Pons A, Molina M, Cid VJ & Pulido R (2011) A comprehensive functional analysis of PTEN mutations: Implications in tumor- and autism-related syndromes. *Hum. Mol. Genet.* **20:** 4132–4142

Rodríguez-Escudero I, Roelants FM, Thorner J, Nombela C, Molina M & Cid VJ (2005) Reconstitution of the mammalian PI3K/PTEN/Akt pathway in yeast. *Biochem. J.* **390:** 613–623

Rohban MH, Singh S, Wu X, Berthet JB, Bray M, Shrestha Y, Varelas X, Boehm JS & Carpenter AE (2017) Systematic morphological profiling of human gene and allele function via Cell Painting. *Elife***:** 1–23

Rowinsky EK, Onetto N, Canetta RM & Arbuck SG (1992) Taxol: the first of the taxanes, an important new class of antitumor agents. *Semin. Oncol.* **19:** 646–62

Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP & Fowler DM (2017) A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18:** 1–15

Sanjana NE (2017) Genome-scale CRISPR pooled screens. *Anal. Biochem.* **532:** 95–99

Shafer A, Zhou C, Gehrig PA, Boggess JF & Bae-Jump VL (2010) Rapamycin potentiates the effects of paclitaxel in endometrial cancer cells through inhibition of cell proliferation and induction of apoptosis. *Int. J. Cancer* **126:** 1144–1154

Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xian M, et al (2017) Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546:** 431–435

Shcherbakova DM, Baloban M, Emelyanov A V, Brenowitz M, Guo P & Verkhusha V V (2016) Bright monomeric near-infrared fluorescent proteins as tags and biosensors for multiscale imaging. *Nat. Commun.* **7:** 1–12

Song MS, Carracedo A, Salmena L, Song SJ, Egia A, Malumbres M & Pandolfi PP (2011) Nuclear PTEN regulates the APC-CDH1 tumor-suppressive complex in a phosphatase-independent manner. *Cell* **144:** 187–199

Soniat M & Chook YM (2015) Nuclear localization signals for four distinct karyopherin-β nuclear import systems. *Biochem. J.* **468:** 353–362

Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, Pliner HA, Jackson DL, Daza RM, Christiansen L, et al (2020) Massively multiplex chemical transcriptomics at single-cell resolution. *Science (80-.).* **367:** 45–51

Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J & Fowler DM (2017) Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101:** 315–325

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J & Fields S (2015) Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200:** 413–422

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R & Smibert P (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14:** 865–868

Su D, Smith SM, Preti M, Schwartz P, Rutherford TJ, Menato G, Danese S, Ma S, Yu H & Katsaros D (2009) Stathmin and tubulin expression and survival of ovarian cancer patients receiving platinum treatment with and without paclitaxel. *Cancer* **115:** 2453–2463

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102:** 15545–15550

Sugimura M, Sagae S, Ishioka SI, Nishioka Y, Tsukada K & Kudo R (2004) Mechanisms of paclitaxel-induced apoptosis in an ovarian cancer cell line and its paclitaxel-resistant clone. *Oncology* **66:** 53–61

Tamura M, Gu J, Matsumoto K, Aota S, Parsons RE & Yamada KM (1998) Inhibition of Cell Migration, Spreading, and Focal Adhesions by Tumor Suppressor PTEN. *Science (80-.).*

Theodoropoulos PA, Polioudaki H, Kostaki O, Dargemont C & Georgatos SD (1999) Taxol Affects Nuclear Lamina and Pore Complex Organization and Inhibits Import of Karyophilic Proteins into the Cell Nucleus Taxol Affects Nuclear Lamina and Pore Complex Organization and Inhibits Import of Karyophilic Proteins into the Cell Nucleus 1. **:** 4625–4633

Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, et al (2017) A subcellular map of the human proteome. *Science (80-.).* **356:** eaal3321

Thyagarajan B & Bloom JD (2014) The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife* **3:** e03300

Tibarewal P, Zilidis G, Spinelli L, Schurch N, Gray A, Perera NM, Davidson L, Geoffrey J, February NRL, Maccario H, et al (2012) PTEN Protein Phosphatase Activity Correlates with Control of Gene Expression and Invasion , a Tumor- Suppressing Phenotype , But Not with AKT Activity. *Science (80-.).* **5:** 1–12

Toettcher JE, Gong D, Lim WA & Weiner OD (2011) LIGHT CONTROL OF PLASMA MEMBRANE RECRUITMENT USING THE PHY – PIF SYSTEM. **:** 1–13

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS & Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32:** 381–386

Trielli MO, Andreassen PR, Lacroix FB & Margolis RL (1996) Differential taxol-dependent arrest of transformed and nontransformed cells in the G1 phase of the cell cycle, and specific-related mortality of transformed cells. *J. Cell Biol.* **135:** 689–700

Väremo L, Nielsen J & Nookaew I (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41:** 4378–4391

Vuono EA, Mukherjee A, Vierra DA, Adroved MM, Hodson C, Deans AJ & Howlett NG (2016) The PTEN phosphatase functions cooperatively with the Fanconi anemia proteins in DNA crosslink repair. *Sci. Rep.* **6:** 1–13

Wang C, Lu T, Emanuel G, Babcock HP & Zhuang X (2019) Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. *Proc. Natl. Acad. Sci. U. S. A.* **166:** 10842–10851

Wang G, Li Y, Wang P, Liang H, Cui M, Zhu M, Guo L, Su Q, Sun Y, Mcnutt MA, et al (2015) PTEN regulates RPA1 and protects DNA replication forks. *Cell Res.* **25:** 1189–1204

Weinreb C, Rodriguez-Fraticelli A, Camargo FD & Klein AM (2020) Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (80-.).* **367:**

West G, Gullmets J, Virtanen L, Li SP, Keinänen A, Shimi T, Mauermann M, Heliö T, Kaartinen M, Ollila L, et al (2016) Deleterious assembly of the lamin A/C mutant p.S143P causes ER stress in familial dilated cardiomyopathy. *J. Cell Sci.* **129:** 2732–2743

Wozniak DJ, Kajdacsy-Balla A, Macias V, Ball-Kell S, Zenner ML, Bie W & Tyner AL (2017) PTEN is a protein phosphatase that targets active PTK6 and inhibits PTK6 oncogenic signaling in prostate cancer. *Nat. Commun.* **8:**

Wu H, Chen S, Yu J, Li Y, Zhang X yan, Yang L, Zhang H, hou Q, Jiang M, Brunicardi FC, et al (2018) Single-cell Transcriptome Analyses Reveal Molecular Signals to Intrinsic and Acquired Paclitaxel Resistance in Esophageal Squamous Cancer Cells. *Cancer Lett.* **420:** 156–167

Yang J, Wang L, Yang F, Luo H, Xu L, Lu J, Zeng S & Zhang Z (2013) mBeRFP, an Improved Large Stokes Shift Red Fluorescent Protein. *PLoS One* **8:** 6–11

Yang X, Kui L, Tang M, Li D, Wei K, Chen W, Miao J & Dong Y (2020) High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. *Front. Genet.* **11:** 19

Yoon B-J (2009) Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* **10:** 402–415

Young MD & Behjati S (2018) SoupX removes ambient RNA contamination from droplet based single cell RNA sequenc-ing data. *bioRxiv***:** 303727

Zhang, Wei; Lohman, Alexander W.; Zhuravlova, Yevgeniya; Lu, Xiaocen; Wiens, Matthew D.; Hoi, Hiofan; Yaganoglu, Sine; Mohr, Manuel A.; Kitova, Elena N.; Klassen JS. & Pantazis, Periklis; Thompson, Roger J.; Campbell RE (2017) Optogenetic control with a photocleavable protein. *Nat. Methods* **14:** 391–394

Zhang J, Kobert K, Flouri T & Stamatakis A (2014a) PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614–620

Zhang K, Han X, Li Y, Li SY, Zu Y, Wang Z & Qin L (2014b) Hand-Held and Integrated Single-Cell Pipettes.

Zhou BY, Ye Z, Chen G, Gao ZP, Zhang YA & Cheng L (2007) Inducible and reversible transgene expression in human stem cells after efficient and stable gene transfer. *Stem Cells* **25:** 779–89

Zhou M, Zhao Y, Ding Y, Liu H, Liu Z, Fodstad O, Riker AI, Kamarajugadda S, Lu J, Owen LB, et al (2010) Warburg effect in chemosensitivity: Targeting lactate dehydrogenase-A re-sensitizes Taxol-resistant cancer cells to Taxol. *Mol. Cancer* **9:** 1–12

Zimmerman SP, Kuhlman B & Yumerefendi H (2016) Engineering and Application of LOV2-Based Photoswitches 1st ed. Elsevier Inc.

Appendix

Nuclear Dendra2 and Photoactivation Tunability

During my early experiments, I noticed three interesting patterns at the flow cytometer:

1. The ratio of activated to unactivated Dendra2 signals had much less variance than the raw activated Dendra2 signal
2. Moving Dendra2 from cell-wide expression to the nucleus resulted in lower variance of the ratio of activated to unactivated Dendra2 across cells activated under the same condition
3. Activating a small, constant area of Dendra2 in each nucleus (~5% of average nuclear size) produced lower variance activation ratios than did activating a larger, constant area (50% of average nuclear size). However, if the area was close to the size of the whole nucleus (~100% average nuclear size), the variance in activation ratios would decrease again.

This section of the Appendix seeks to explain these observations, and also explain how Visual Cell Sorting can so finely tune the activation of cells compared to other image-based cell selection methods. The explanation involves two phenomena that, surprisingly (to me, anyway), interact: (1) the ratio of activated to unactivated Dendra2 (2) the proportion of the Dendra2 in each cell that is illuminated by 405 nm light.

Even clonally derived cells expressing fluorescent markers have absolute fluorescence intensities that differ by up to ten-fold; after photoactivation, this leads to a highly variable red Dendra2 signal. To demonstrate this mathematically, if one assumes that Dendra2 photoactivation for a given laser strength and activation time occurs at an efficiency ***e*** (a variable related to laser intensity and illumination time), then Equation 1 shows that total red fluorescence intensity I_r for a cell after photoactivation is proportional to the total abundance of Dendra2 ***D*** in that cell:

$$I_r \propto eD \quad \text{(Equation 1)}$$

Thus, the final photoactivated (red) intensity is proportional to the abundance of Dendra2 in each cell and will vary by up to ten-fold for cells activated under the same conditions. In theory,

one could use the cell's green fluorescence intensity I_g to correct for the total Dendra2 abundance in cells; to calculate I_g, one must also take into account the fact that green Dendra2 signal is lost upon activation (because it is photoconverted to red):

$$I_g \propto D - eD \text{ (Equation 2)}$$

Additionally, one must be careful to update both Equations 1 and 2 if a proportion of the cell's Dendra2 ***p*** is illuminated with 405 nm light, rather than all of its Dendra2:

$$I_r \propto eDp \text{ (Equation 3)}$$

$$I_g \propto D - eDp \text{ (Equation 4)}$$

Taking the ratio of Equations 3 and 4 yields an activated to unactivated Dendra2 ratio (***I_r / I_g***) for each cell:

$$I_r / I_g \propto \frac{eDp}{D - eDp} \text{ (Equation 5)}$$

Here, the cell's Dendra2 abundance can be cancelled out, thereby solving the problem of uneven Dendra2 expression across cells:

$$I_r / I_g \propto \frac{ep}{(1 - ep)} \text{ (Equation 6)}$$

However, a new problem arises now. In Equation 6, the efficiency of Dendra2 photoactivaton ***e*** is a constant that stays the same across all cells. However, the proportion of Dendra2 activated ***p*** across cells can be different, especially if Dendra2 is expressed in the whole cell. Cytoplasmic Dendra2 leads to variable proportions of Dendra2 photoactivation because (1) segmentation using cytoplasmic fluorescence signals is challenging and it is likely that many cells would be incompletely segmented and (2) to prevent collateral activation, it would be important to activate an area that is well inside the cell's borders, and such an activation area may take up a higher proportion of some cells' areas than others. To summarize, if a different proportion of total cellular Dendra2 is activated across cells, this will lead to variability in the ratio of activated to unactivated Dendra2 signal.

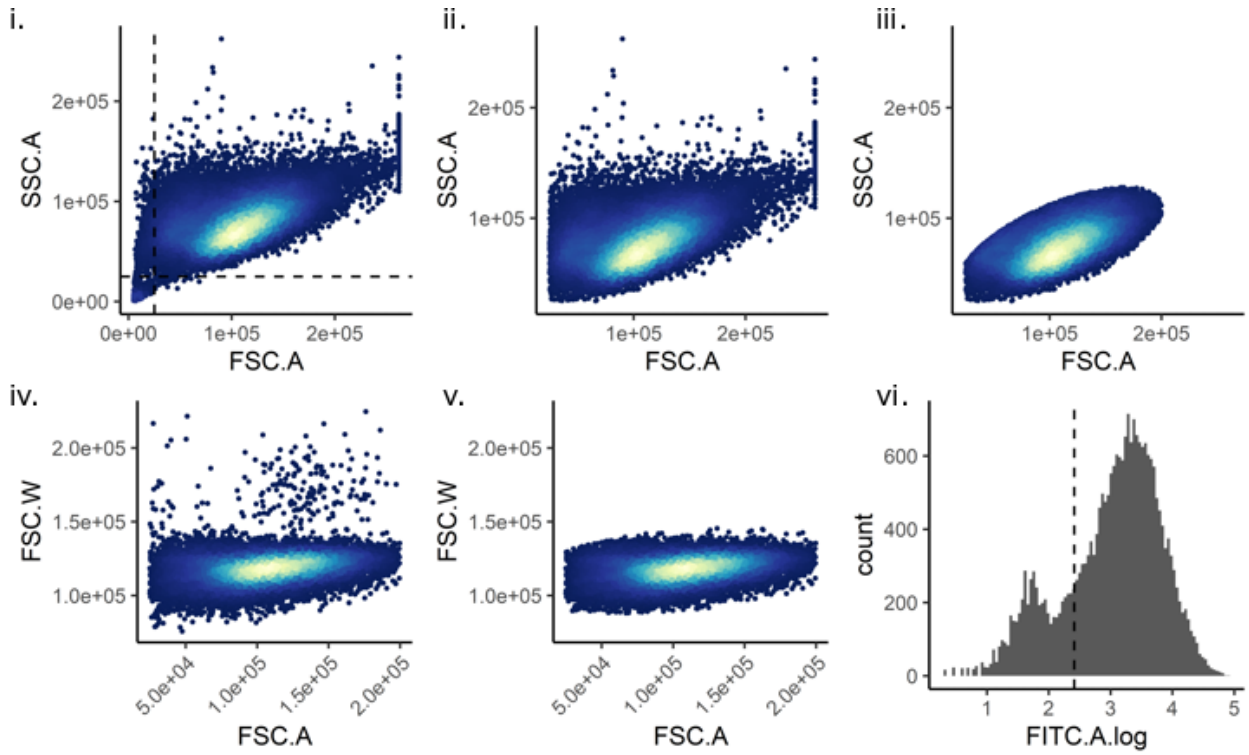
By contrast, nucleus-localized Dendra2 can ensure that 100% of the Dendra2 in each cell is subject to activation. This is because it is easy to segment, and there is no risk of collateral Dendra2 activation. Therefore, ***p*** always equals 1 and Equation 6 is simplified to:

$$I_r / I_g \propto \frac{e}{(1-e)}$$ (Equation 7)

Where ***e***, the efficiency of activation, is a constant parameter; in turn, this means that the ratio of activated to unactivated Dendra2 intensity will remain constant, even amongst cells with variable Dendra2 expression. Therefore, performing activation on nuclear Dendra2 and using an activated-to-unactivated intensity ratio allows for accurate identification of activated cells by flow cytometry. Furthermore, the lack of noise in ***D*** and ***p*** allows for fine tuning of the photoactivation signal by, for example, varying the illumination time.

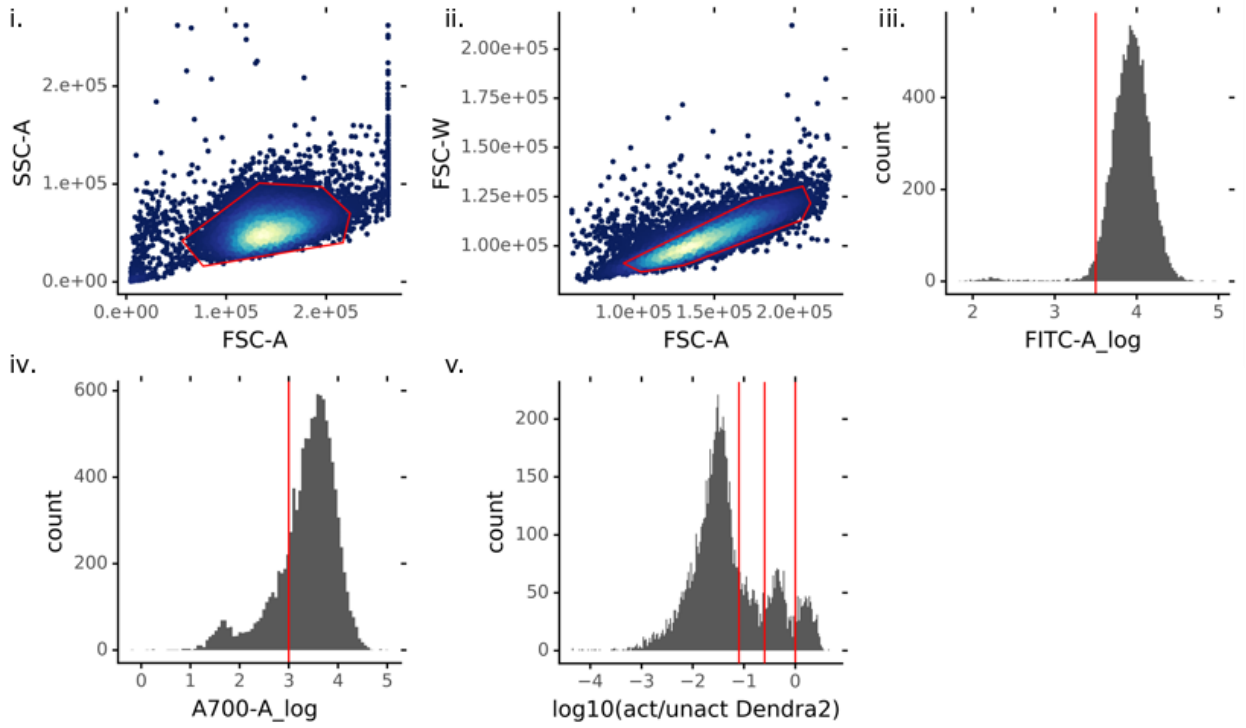
Appendix Figures

Appendix Figure 1.1



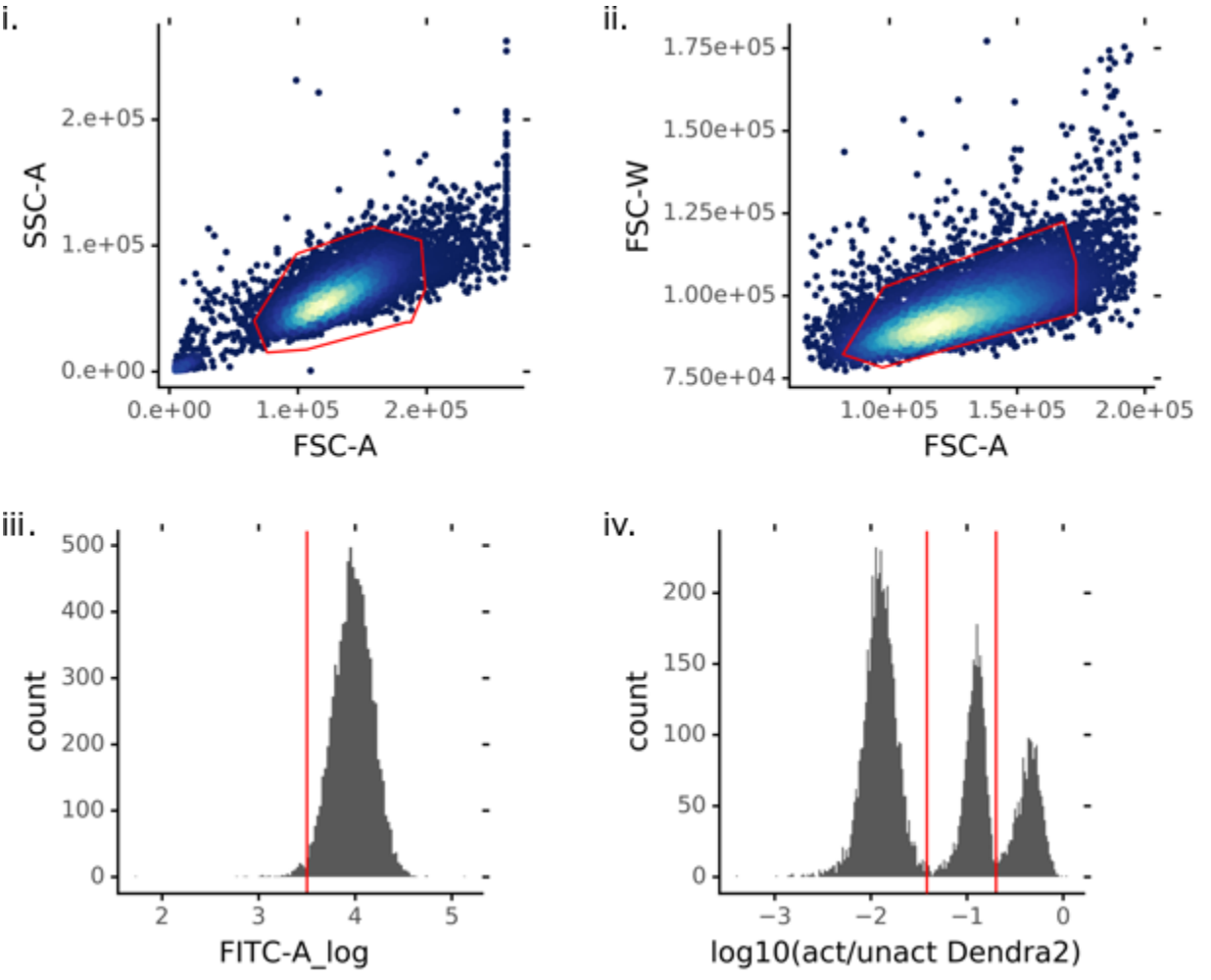
Appendix Figure 1.1. The gating scheme for selective photoactivation of cells expressing miRFP. Custom code using flowCore (v1.11.20) in R (v3.6.0) was used to gate the cells as follows. **(i)** Debris was removed using a SSC.A vs FSC.A plot. **(ii)** and **(iii)** A Mahalanobis distance filter was used to identify live cells on a SSC.A vs FSC.A plot. **(iv)** and **(v)** A Mahalanobis distance filter was used to identify single cells on a FSC.W vs FSC.A plot. **(vi)** Dendra2-positive cells were identified using a FITC plot.

Appendix Figure 2.1



Appendix Figure 2.1. The gating scheme for Visual Cell Sorting of cells expressing the SV40 NLS library. Using the BD FACSDiva software, cells were gated as follows. **(i)** Live cells were identified using a SSC-A vs FSC-A plot. **(ii)** Single cells were gated using a FSC-W vs FSC-A plot. **(iii)** Cells expressing Dendra2 were gated using a FITC-A plot. **(iv)** Cells expressing miRFP were gated using an AlexaFluor 700-A plot. **(v)** Cells were divided into four bins according to the ratio of activated / unactivated Dendra2.

Appendix Figure 3.1



Appendix Figure 3.1. The gating scheme for Visual Cell Sorting on cells treated with paclitaxel. Using the BD FACSDiva software, cells were gated as follows. (i) Live cells were identified using a SSC-A vs FSC-A plot (ii) Single cells were gated using a FSC-W vs FSC-A plot. (iii) Cells expressing Dendra2 were gated using a FITC-A plot. (iv) Cells were divided into three bins (0 ms, 200 ms, and 800 ms) according to the ratio of activated / unactivated Dendra2. Cells in the 200 ms and 800 ms bins were sorted.