Title
Cell communication across scales: Using single-cell technologies to explore cellular dynamics

Author
Nagle, Maeve Patricia

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Cell communication across scales: Using single-cell technologies to explore cellular dynamics

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

by

Maeve Patricia Nagle

2021

ABSTRACT OF THE DISSERTATION

Cell communication across scales: Using single-cell technologies to explore cellular dynamics

by

Maeve Patricia Nagle

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2021

Professor Roy Wollman, Chair

The ability of a cell to accurate interpret its environment and communicate cues within itself is crucial to survival. How cells orchestrate the transmission of large amounts of information within themselves is a key biological question. From surface-level receptors binding to myriads of extracellular molecules to internal changes in gene expression, cells must process vast amounts of inputs and outputs every minute. This work follows the innerworkings of cells to further understand how cells interpret and respond to their environments. By following the flow of information with a cell, we can parse out the mechanisms by which cell signaling controls response. In this dissertation, I first describe a new suite of *in situ* technologies that are offering a paradigm shift in the manner in which we can understand biology across scales both large, in terms of the makeup and function of organ systems, to the minute, in terms of intracellular localization of molecules. In the second chapter, I describe how one of these technologies, MERFISH, can pair with live-cell imaging of mouse macrophages to further understand their cellular responses to inflammatory signaling molecules. In the third chapter, I describe a method by which endogenous proteins can be genetically tagged and imaged with individual molecular

barcodes to further understand cellular signaling and protein dynamics. In the last chapter, I describe how the cell's chromatin environment is influenced more by *cis* regulatory elements than previously imagined. Combined, this dissertation explores several areas of cellular information transfer and communication.

The dissertation of Maeve Patricia Nagle is approved.

Guillaume Chanfreau

Jason Ernst

Jose Alfonso Rodriguez

Roy Wollman, Committee Chair

University of California, Los Angeles

2021

DEDICATION

To the love of my life, Eva Rosker, who supports me wholeheartedly.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I owe a great deal of gratitude to a large number of fellow scientists who have helped and supported me through my PhD journey. I want to thank my committee for their support. In particular, I want to thank Alex Hoffmann in particular for advice on all things immune and creating a fantastic QCBio environment. Of course, I want to thank my advisor Roy Wollman who lives the motto ‘science is fun and exciting.’

I would like to thank all the members of the fifth floor of Boyer, who have been fantastic colleagues. The Hoffmann lab especially has been incredible neighbors. Special thanks to Stefanie Leucke, who is the kindest and most competent collaborator anyone could ask for. All of the members of the Modelers and Microscopists group meetings have helped me fine-tune my scientific question asking and always provided a supportive and collaborative environment.

Thank you to all of the members of the Wollman Lab throughout my PhD. Thanks to Naomi Handly, Jason Yao, Thanutra (Bu) Zhang, Anna Pilko, Rob Foreman, Alok Maity, and Ryan Lannan who warmly welcomed me into the Wollman lab. Thank you to Zach Hemminger, Evan Maltz, Gaby Tam, and Jonathan Perrie for being by my side at the end of my Wollman lab journey.

VITA

2012 - 2016	B.S. in Chemistry, Georgia Institute of Technology, Atlanta, GA
2016 - 2018	M.S. in Biochemistry, Molecular and Structural Biology, University of California at Los Angeles, Los Angeles, CA

PUBLICATIONS

Nagle, M. P., Tam, G. S., Maltz, E., Hemminger, Z. & Wollman, R. Bridging scales: From cell biology to physiology using *in situ* single-cell technologies. *Cell Syst* **12**, 388–400 (2021)

CHAPTER 1

Bridging scales: from cell biology to physiology using in situ single-cell technologies

Abstract

Biological organization crosses multiple spatial scales: from molecular, cellular, to tissues and organs. The proliferation of molecular profiling technologies enables increasingly detailed cataloging of the components at each scale. However, the scarcity of spatial profiling has made it challenging to bridge across these scales. Emerging technologies based on highly multiplexed *in situ* profiling are paving the way to study the spatial organization of cells and tissues in greater detail. These new technologies provide the data needed to cross the scale from cell biology to physiology and identify the fundamental principles that govern tissue organization. Here, we provide an overview of these key technologies and discuss the present and future insights these powerful techniques enable.

Introduction

In biology, structure and function are tightly linked. For example, it is the structure of a protein that determines its function, and not simply its amino-acid composition. To solve a protein structure the x, y, and z coordinates of each atom are determined, the local organization identified (e.g. alpha-helix, beta-sheets) and the different domains of the proteins are defined. It is the detailed understanding of the spatial organization of the different amino acids that make up the protein that allows researchers to build a model that explains how its structure (and the dynamics of that structure) determines its function. Similarly, at the cellular level, a list of all the molecules in a cell is insufficient to understand a cell's function. Historically, the electron micrographs obtained by cell biology pioneers such as Palade and Porter in the 1950s were key to defining cellular organelles and determining the structural organization of the cell (Palade and Porter, 1954). In the 70 years that followed, modern cell biology connected these structural

insights to the molecular composition of a cell providing key understanding of how the spatial organization of the molecules that make up a cell determines the cell's function. At the next level, the connection between an organ's anatomy (i.e. structure) and its physiology (i.e. function) has always been a core perspective used to investigate tissues and organs. Histological sections observed using light microscopy have been a key tool that enabled understanding of organ function through insights into their microstructure. However, similar to Palade and Porter's electron micrograph, existing histological approaches lack sufficient molecular details. Histological staining is often based on a combination of non-specific dyes and a handful of molecular markers and does not provide sufficient information to fully understand the complex molecular and cellular structure of the organ. Therefore, while anatomical information is ubiquitous, the lack of spatio-molecular details limits the ability to connect a structure to its function across biological scales.

From dissociative to spatial measurements

Technological advances in single-cell measurements allow the cataloging of all cells into types, subtypes, and states. These catalogs provide key insights into the cellular composition of different organs. The most widespread single-cell technology is undoubtedly single-cell RNA sequencing (scRNA-seq) (Tang et al., 2009). Named "method of the year" for 2013 (2014), scRNA-seq has since become a fixture across many biology labs and has led to many new biological insights due to the ease of analyzing large numbers of cells in a short time frame. In scRNA-seq, cells are dissociated from each other, isolated, barcoded, and sequenced. Due to its dissociative nature, scRNAseq is especially suited for the task of cell classification. However, this dissociative approach loses the spatial context of cells. Therefore, while this technique provides an invaluable new vocabulary of cell type taxonomy, the lack of spatial information limits its use for organ-scale structure-function analysis.

In recent years, new technologies have been developed that measure the characteristics of single cells *in situ* (in the original site). These technologies link the detailed compositional information obtained through dissociative measurement with spatial histological measurements. These new measurement technologies have transformative potential as they provide the missing data on organs' molecular and cellular structures. By bridging the gap left by dissociative techniques *in situ* technologies provide a route to connect organs' functions to their molecular and cellular structure.

In this review, we discuss the main technologies for characterizing cells *in situ*. We additionally discuss the ways in which *in situ* measurements are contributing to our understanding of biological organization from the subcellular scale to the physiological scale. This review will not focus on the technical aspects of each technology, as previous reviews for scRNA-seq (Chen et al., 2019; Stark et al., 2019) and spatial technologies (Asp et al., 2020; Lundberg and Borner, 2019; Young et al., 2020) have thoroughly addressed these topics. Rather, we provide an overview of key approaches and how they can be used to bridge scales and connect organ cellular structure to its function.

In situ technologies

The fast pace of technology development in this space introduces some ambiguity related to terminology. In this review, we make a distinction between the establishment of a cell taxonomy, i.e. classification, and the creation of a cell atlas that requires spatial mapping of cell types in tissues and organs. Similarly, the term *in situ* technologies is ill-defined as *in situ* measurement technologies are as old as histology itself (Motta, 1998) and, depending on the definition, can include a vast range of measurements. In the scope of this review, we will use a more narrow definition of *in situ* measurements to focus on highly multiplexed spatial measurements of RNA and proteins. RNA measurements are based on *in situ* hybridization, *in*

situ sequencing, or RNA capture and cDNA barcoding. Protein measurements are based on antibodies that recognize a specific antigen that can be read either using many rounds of imaging or conjugation with metal ions that are read with a rastering mass spectrometer. Figure 1.1 provides an overview of current approaches for spatial *in situ* measurements.

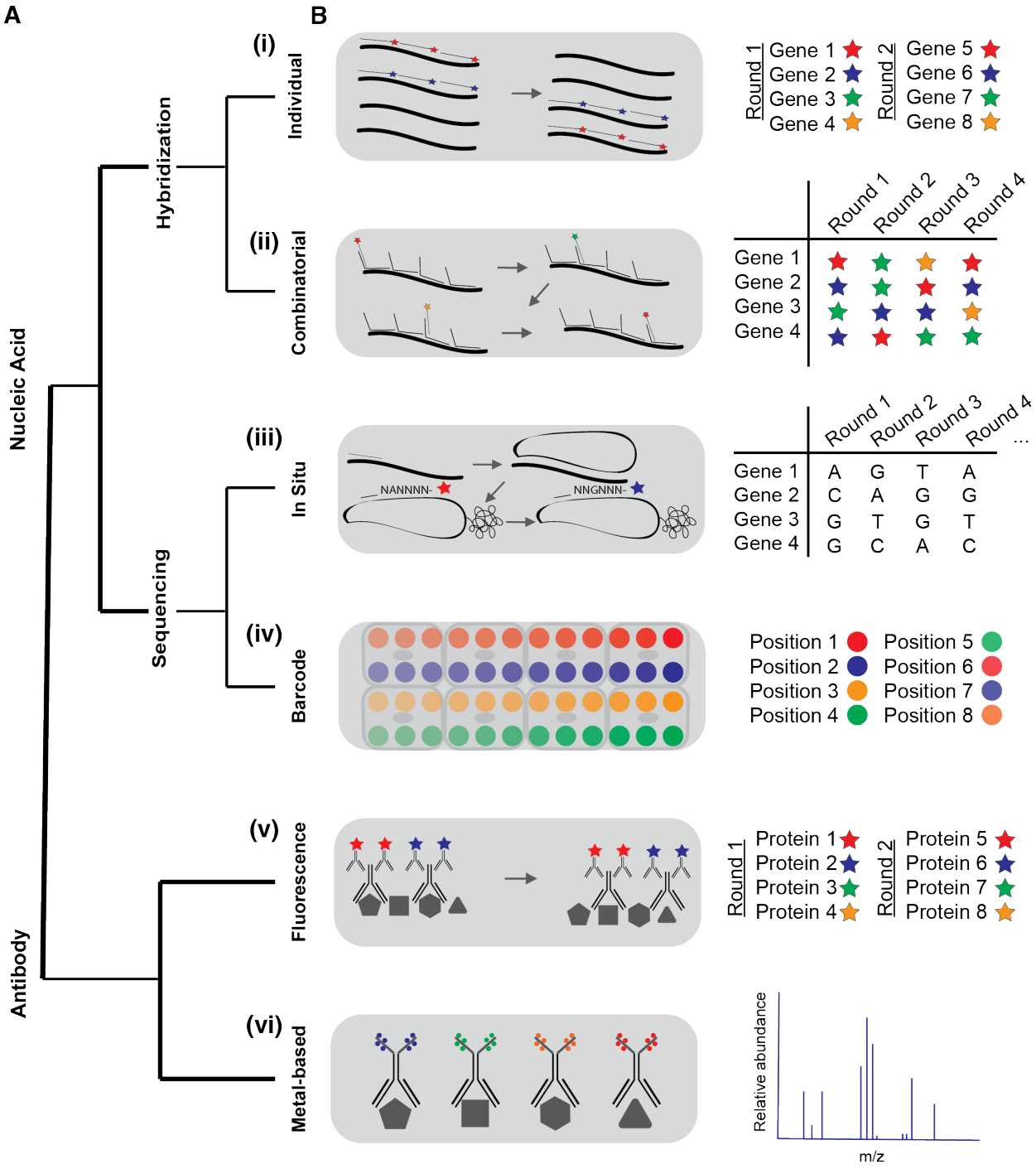


Figure 1.1: Overview of key *in situ* technologies. (A) Hierarchical classification of the main approaches used for *in situ* measurements. At the top level, methods are split depending if their main targets are nucleic acids or proteins. Nucleic acid approaches are divided based on the main readout mechanism, hybridization of fluorescent probes to the transcript of interest, or use sequencing to read out the transcript identity. Hybridization approaches are further split into individual approaches or combinatorial approaches. Sequencing approaches either measure RNA in the cell directly or measure DNA barcodes. Antibodies are frequently used to measure protein *in situ* and contain either fluorophore attachments that can be read by fluorescent imaging or metals that are read out by mass cytometry. (B) Schematic representations of key technologies. (i) Individual hybridization techniques, like smFISH,

employ many fluorescently-label DNA probes that bind to a transcript of interest. Each transcript appears as a diffraction-limited fluorescent spot in an image and is identifiable by its unique color. (ii) Combinatorial hybridization techniques utilize similar principles to individual hybridization but utilize consecutive binding of probes to the same molecules and sequential imaging to create a "barcodes" of fluorescent spots across imaging rounds that are used to determine a transcript's identity. An additional set of probes are used in combination hybridization that bind directly to an RNA transcript with overhangs for fluorescent readout probes to bind to. (iii) *In situ* sequencing involves the readout of an RNA transcript directly or of a barcoded primer used to amplify that transcript. The transcript is first reverse transcribed into cDNA, then that cDNA is amplified, frequently by rolling circle amplification. The amplified cDNA is either sequenced directly or the sequence of a specific primer that binds to the cDNA is sequenced. (iv) In *in situ* barcoding methods, a sample is applied to a slide covered with DNA-barcoded microbeads. The sample is lysed and the resulting RNA binds to the beads, which are then sequenced. The location of the RNA is mapped back to the known location of the DNA barcode sequence from the bead. (v) In each round of fluorescent-based antibody readout, proteins are bound to an antibody with a fluorescently labeled molecule attached, similar to in (i). Each protein is represented by a single-colored fluorescent spot in an image. (vi) Metal-based antibody readouts are similar to fluorescent-based antibody readouts but utilize unique metal atoms attached to antibodies instead of fluorescent molecules. These metal atoms are read out using a mass cytometer.

RNA hybridization

Single-molecule RNA fluorescence *in situ* hybridization (smFISH) (Femino et al., 1998; Raj et al., 2008) was the first widespread single-molecule *in situ* RNA measurement technology. smFISH counts the number of mRNAs transcribed from a gene of interest within a cell by using DNA probes specific to the mRNA target sequence. These DNA probes are attached to a fluorescent molecule and collectively create a single diffraction-limited spot in the position of the mRNA molecule. The number of diffraction-limited fluorescent spots in a cell is counted to determine the number of mRNA molecules present. Several techniques seek to improve upon the probe design of smFISH. A partial list of these extensions includes RNAscope ((Wang et al., 2012), which uses Z-shaped DNA probes to enhance specificity, osmFISH (Codeluppi et al., 2018) which is optimized for use in thin tissue sections such as brain slices, ExFISH (Chen et al., 2016) which uses expansion microscopy to further separate mRNA spots and make image analysis easier, and SABER-FISH which uses multi-part probes to enhance the signal from each mRNA (Kishi et al., 2019). Overall, the principle that is shared between smFISH and its many subsequent versions is that expression of pre-defined genes is measured in a targeted

manner with one measurement per gene. The high accuracy and mRNA capture rate (both >95%) have led smFISH to become the “gold standard” among validation techniques (Torre et al., 2018). However, the high accuracy comes at a price: smFISH-based approaches assign each gene a specific measurement (i.e. color), so there can only be as many genes measured in a single hybridization as there are non-overlapping fluorescent molecules available. Four rounds of hybridization with this approach using four different types of fluorescent probes can measure a maximum of 16 genes. This linear scaling limits the ability of smFISH-based approaches to provide full and detailed structural information.

Combinatorial FISH approaches address the key limitation of smFISH by increasing the number of genes that can be counted per experiment and thereby provide much more detailed information on the cellular composition in the tissue. These techniques include MERFISH (Chen et al., 2015; Moffitt et al., 2016a, 2016b; Wang et al., 2020; Xia et al., 2019a), seqFISH+ (Eng et al., 2019), and most recently split-FISH (Goh et al., 2020). These approaches, while very similar, differ in some of the details related to barcoding strategy and how they remove the fluorescently-tagged oligo probes. The core improvement over smFISH is that combinatorial FISH approaches utilize barcodes for each RNA to increase the measurement capacity. Each gene is given a ‘barcode’ that is a combination of colors, so the gene identity is uncovered by the data from every round of hybridization. This process scales exponentially, so four rounds of hybridization with four different types of fluorescent probes would allow for up to 256 genes to be analyzed, instead of 16. Using four dyes and eight rounds of hybridization (48 = 65,536), in principle an entire transcriptome can be measured. However, the use of RNA barcodes comes at a price. In the 48 scheme, any error in “calling” one of the four measurements needed to assign a gene identity to an RNA molecule will result in an incorrect assignment. Such errors have the potential to substantially reduce the accuracy of combinatorial FISH approaches. To address this limitation, the barcodes are typically chosen sparsely from a large set of possible

codes. This intentional reduction in chosen barcodes can substantially reduce the error rates of combinatorial approaches at a cost of an increase in the number of hybridization rounds. In typical combinatorial measurement, 24 rounds are used with each molecule having 4 measurements out of the possible 24 rounds. The sparsity of barcode assignment is such that 200-500 genes can be measured using 24 rounds of imaging. Both MERFISH and seqFISH+ were used to demonstrate that transcriptome scale (~10,000 genes) is possible but at a cost of a much higher number of measurements and overall reduced throughput (Chen et al., 2015; Eng et al., 2019). The flexibility of combinatorial FISH approaches is important as the complexity of the approach often requires tailoring measurements to specific samples and experiments. Optical crowding of many RNA spots per image can impede RNAs from being resolved and require integration of some smFISH rounds for highly expressed genes or a substantial increase in the number of measurements. As a result, large samples can require weeks of continuous imaging and can generate terabytes of image data. Overall, combinatorial FISH approaches provide a very powerful platform for targeted spatial RNA counting that can be tailored to the specific needs of a project.

Antibodies

Immunohistochemistry (IHC) has been used since 1942 to study the spatial location of proteins in a tissue (American Association of Immunologists, 1942). IHC involves adding labeled antibodies to a sample in order to visualize proteins and other molecules of interest. Despite the high specificity achieved by antibodies, IHC is difficult to multiplex. Only in the last two decades have a few approaches been successfully implemented to enable 30+ protein readouts in a sample. The first difficulty is the generation of validated high-quality antibodies. In practice, this is a non-trivial issue that has been partially addressed by both commercial and academic groups (Edfors et al., 2018) but is by no means a solved problem. The second difficulty relates to how the spatial distribution of these antibodies is read. To prevent cross-reactions and due to

the limited number of host animals used in antibody production, the use of primary and secondary antibodies, common in standard IHC, is difficult. This limits antibody selection to mostly primary antibodies that need to be read across multiple measurements. Here, we focus on solutions that address the readout problem. Overall the multiple attempts at “cracking” the multiplexing challenge can be divided into two types: 1) repeated imaging on a light microscope and 2) coupling antibodies to unique metal ions.

A straightforward way to increase the number of readouts is to use existing tools for fluorescence-based antibody detection and simply repeat them many times (Fig 1.1). For example, a set of antibodies would be added to a sample, imaged, then stripped away and replaced with a new set of antibodies. This idea is implemented in methods such as MxIF (Gerdes et al., 2013), CycIF (Lin et al., 2015, 2018), and 4i (Gut et al., 2018). The key distinction between the different variants is in how the multiple rounds of staining are achieved, i.e. are the antibodies themselves stripped from the sample, or are they simply quenched by photobleaching. An important advantage of these approaches is that since they are based on standard microscopy; they can also be coupled to live-cell imaging (Lin et al., 2015). Borrowing from the relative ease of repeated imaging after RNA hybridization a few methods, CODEX (Goltsev et al., 2018), DEI (Wang et al., 2017), and Immuno-SABER (Saka et al., 2019) use oligo-conjugated antibodies and fluidics systems almost identical to the one used by combinatorial FISH approaches.

An alternative approach for multiplexing antibody staining is based on changing the readout from a light microscope to a mass spectrometer. Mass cytometry imaging approaches have been developed to avoid some of the practical limitations encountered by attempts to multiplex IHC-based analysis. Specific implementations of imaging mass cytometry include Multiplexed Ion Beam Imaging (MIBI) (Angelo et al., 2014; Keren et al., 2019; Ptacek et al.,

2020) and Imaging Mass Cytometry (IMC) (Giesen et al., 2014; Ijsselsteijn et al., 2019). Each of these approaches uses secondary ion mass spectrometry to image antibodies tagged with isotopically pure elemental metal reporters. The main distinction between the two methods arises in sample ablation which leads to differences in image resolution and acquisition times between IMC and MIBI (Baharlou et al., 2019). Though these techniques can analyze up to 40 proteins in a sample at a given time, they are both limited by antibody availability and quality. Additionally, MIBI and IMC require specialized equipment to point-scan small fields, and therefore imaging large samples can be slow and costly.

Sequencing

The accessibility of DNA sequencing, achieved in part due to six orders of magnitude decrease in sequencing cost per base pair (Stark et al., 2017), motivated innovative approaches that leverage DNA sequencing while still preserving spatial information. The approaches that couple spatial information to RNA sequencing can be divided into three distinct categories: 1) separation of RNA based on their spatial location followed by sequencing, 2) use of spatially distinct DNA barcodes during library preparation, and 3) performing the sequencing reactions themselves *in situ*. The first two categories directly leverage existing sequencing technologies whereas the latter use many of the chemistry developed for sequencing however the readout itself is microscopy-based and shares many similarities to combinatorial FISH-based approaches.

Perhaps the most straightforward way to assign spatial information to RNA molecules is to only collect RNAs from a specific spatial domain. This concept is the basis of highly useful methods such as LCM-seq (Nichterwitz et al., 2016) and GEO-seq (Chen et al., 2017) that use laser capture microscopy to sequence a small number of cells at a time. A more systematic application of a spatial collection of RNA from distinct regions was applied using a method

called Tomo-seq (Burkhard and Bakkers, 2018) that uses cryosectioning to the tissue before sequencing. Photoactivation is another useful tool that was used to encode spatial information and capture RNAs in spatially distinct domains. Transcriptome in vivo analysis (TIVA) exposes live cells to multifunctional caged mRNA-capture molecule tags called TIVA that upon photocleavage hybridize to mRNAs within a cell allowing sequencing of RNAs from specific spatial position (Lovatt et al., 2014). A similar idea was implemented by ZipSeq (Hu et al., 2020) that used patterned light and three distinct colors to label cells according to their spatial position. Labeled cells are sorted and sequenced using standard scRNAseq tools. These spatial-specific capture approaches have been effective tools in understanding the organization of tissues. However, they suffer from an inherent tradeoff between resolution and throughput. While Tomo-seq allowed sequencing entire embryos, this was done in linear sections of 18-micron thickness. On the other extreme TIVA can be used for subcellular localization of RNA molecules however it can only process one location at a time. Therefore, while the approaches that are based on spatially restricted RNA collection provide important spatial information they stop short of enabling the cellular structure of organs and tissues.

To overcome the tradeoff between spatial resolution and throughput, an alternative approach is based on localized barcoding of cDNA during library preparation prior to sequencing. The key advantage of position-based barcoding is that once each region is labeled by a specific code the entire sample can be sequenced as one and using prior knowledge of the XY position of each barcode, the spatial position of all RNA molecules is reconstructed computationally. These approaches involve capturing RNA from tissue samples on a spatially barcoded bead array which is later sequenced. Both High Definition Spatial Transcriptomics (Salmén et al., 2018; Ståhl et al., 2016; Vickovic et al., 2019) and Slide-seq (Rodriques et al., 2019; Stickels et al., 2020) use this approach. While this technique cannot define cell boundaries, High Definition Spatial Transcriptomics can achieve two-micron resolution and

allows for fast, high-throughput processing (Vickovic et al., 2019). A key advantage of these approaches is that they leverage many of the experimental and computational tools developed for scRNASeq. In fact, popular analysis tools such as Seurat were able to add the spatial capture analysis despite the scarcity of datasets that used this approach partially due to its similarity to scRNAseq (Stuart et al., 2019). Spatially resolved sequencing is a promising approach, however, presently it suffers from low RNA capture efficiency. The low capture efficiency means that the capture bin (i.e. spatial domain of a single barcode) needs to be big enough to contain a sufficient number of RNA molecules. Furthermore, even if the capture chemistry will improve, similar to other capture-based approaches there is an inherent tradeoff between resolution, i.e. the size of a single capture bin and the number of bins. To allow subcellular information, capture bins need to be <100 micrometer2 which means that a standard tissue section of 100 mm2 will need 106 distinct barcodes, a non-trivial library to sequence.

In situ sequencing leverages the conceptual advances of DNA sequencing, but not the sequencing machines themselves. In situ sequencing converts RNA in a cell to cross-linked cDNA amplicons that are sequenced within a cell on a microscope. These molecules can either be the RNAs of interest themselves, as in FISSEQ (Lee et al., 2014, 2015), or an RNA barcode specific to transcripts of interest, like in ISS (Ke et al., 2013), STARmap (Wang et al., 2018), and Baristaseq (Chen et al., 2018). FISSEQ (Lee et al., 2014, 2015) cross-links DNA amplicons to a matrix to directly sequence the amplicon inside a cell. ISS, STARmap, and Baristaseq add barcoded oligos specific to targets of interest and sequence the barcodes to determine the presence of transcripts. Similar to combinatorial FISH approaches, barcode-based *in situ* sequencing requires an oligo library that targets genes of interest. While in principle *in situ* sequencing approaches can provide an unbiased view of RNA in tissues and organs, in practice this comes at a cost associated with the need to sequence many copies of highly abundant RNA molecules. The targeted methods have shown more robustness in their implementations

and have dominated over unbiased ones. Interestingly, given their targeted nature, the distinction between them and combinatorial hybridization-based approaches diminishes. This is exemplified in a new protocol called HybISS that merges the rolling circle amplification typical to *in situ* sequencing approaches with hybridization-based multi-round readout that is common in combinatorial FISH (Gyllborg et al., 2020).

Integrative in situ measurements

Integrative spatial multi-modal *in situ* approaches combine the measurements across modalities, i.e. RNA and protein. The integrative and multi-model data will likely enable a more comprehensive understanding of single-cell processes and functions. Many recent innovations in this direction point to an exciting future with complex datasets that span different data types. Techniques like Digital Spatial Profiling (DSP) (Merritt et al., 2020), RNAscope (Kann and Krauss, 2019), smFISH-IF (Tutucci and Singer, 2020), and ImmunoFISH (Kwon et al., 2020) combine FISH and immunofluorescence methods to measure RNA and protein levels within a single cell. RNAscope has additionally been paired with mass cytometry to read RNA and protein levels (Schulz et al., 2018). SABER-FISH also allows for the *in situ* measurement of DNA or RNA transcripts and can combine protein staining for simultaneous detection of a gene's transcript and protein levels (Kishi et al., 2019). Another venture involves reading out DNA and RNA within the same cell *in situ*. ClampFISH (Rouhanifard et al., 2018) probes can be used on both DNA and RNA sequences, allowing for the measurement of DNA and RNA in the same cell in the same experiment. Additionally, live-cell imaging has been combined with *in situ* transcriptomics to allow for mapping the transcriptional state of a cell to its phenotype. CycIF tracked the translocation of a YFP-FoxO3a reporter followed by the readout of seven additional protein levels (Lin et al., 2015). A recent paper analyzed calcium signaling response and gene expression of calcium signaling-related genes in over 5,000 cells (Foreman and Wollman, 2020). New avenues of spatial multi-omics are just now being explored and could have a great

impact on the construction of atlases with many maps. Furthermore, these technologies open up new avenues to study gene perturbations (Wang et al., 2019), cell lineage tracing (Chen et al., 2018; Frieda et al., 2017), and other aspects of functional DNA and RNA biology (Cai et al., 2020; Maiser et al., 2020).

Integrative reconstructions have been developed by combining large dissociative datasets with a smaller number of spatial measurements used as a “ruler”. For example, algorithms have been developed to infer the original spatial location of cells analyzed by scRNA-seq by correlating the level of key marker genes with levels of those genes found within *in situ* datasets (Achim et al., 2015; Satija et al., 2015). Other algorithms such as trendsceek and LIGER have also been used to integrate scRNA-seq data with spatial transcriptomics information (Edsgärd et al., 2018; Welch et al., 2019). The integration across spatial and non-spatial datasets enabled spatial reconstruction by combining laser capture microdissection, bulk sequencing those cells, and then reconstructing the whole tissue through spatial tissue reconstruction (Moor et al., 2018). These reconstruction-based approaches are very powerful as they merge the strengths of dissociative and spatial measurements. However, care needs to be taken in the interpretation of these reconstructions. The recovered maps are based on spatially stratified averaging of many cells. These averaging could mask additional spatial differences that are lost due to averaging. Therefore, the details of the reconstruction matter and care should be used in the interpretation of these measurements.

Challenges are truly opportunities

The ultimate *in situ* measurement technology will have sub-cellular resolution, high detection sensitivity, will be applicable to 3D volumes, compatible with multiple fixation protocols including FFPE, and provide highly multiplexed data on a wide range of molecular species. Given the inherent tradeoff between resolution, sensitivity, and throughput, none of the

technologies described above should be considered a “winner”. It is likely that many different technologies will be developed where each will be a “winner” for a specific subset of applications. As a result, *in situ* technologies contain a smorgasbord of different approaches, each with their own acronym and nuances. Despite the variety, these technologies face some similar challenges. The first challenge is the computational and data complexity. Despite the differences in methods, many computational steps, such as spot calling and cell segmentation (Littman et al., 2020), are shared across approaches. Development of standards and mature computational libraries that can allow the separation of the computational analysis from data acquisition will allow more cross-fertilization in this field. Currently, the Chan Zuckerberg Initiative (CZI) has begun building a unified data-analysis tool and file format called starfish to address this issue (Perkel, 2019). These standards will allow the use of modern machine learning methods that will invariably be key to solving many of these problems (Bannon et al., 2021; Chen et al., 2020a; Moen et al., 2019; Stringer et al., 2021). The second challenge relates to scale. MERFISH imaging of a volume comparable to a mouse brain would require more than a year of continuous imaging. Other technologies, such as spatial RNA barcoding, have a similar order of magnitude time requirements. To achieve cellular resolution for such volume requires ~1013 reads which, even on an advanced NovaSeq 6000 will take multiple years to sequence. The third challenge is the dissemination of these technologies to the scientific community. The complexity of many of these protocols makes the open-source / open-hardware model challenging. Many companies are actively working on bringing these innovations to market which will help. However, whether these efforts will democratize the best technologies remain to be seen. Finally, once spatial data is collected, how to fully analyze it and maximize the insights such data provides is very much an open research question. As was the case for single-cell biology, we anticipate that increase in data availability will result in further developments in statistical and bioinformatics methodology to analyze these rich and

interesting datasets. We are optimistic that these challenges will act as a catalyst for innovation and we expect further technological development in this space.

From cell biology to physiology

The technologies introduced above are paving the way for bridging the gap between intracellular, cellular, and physiological scales (Fig 1.2).

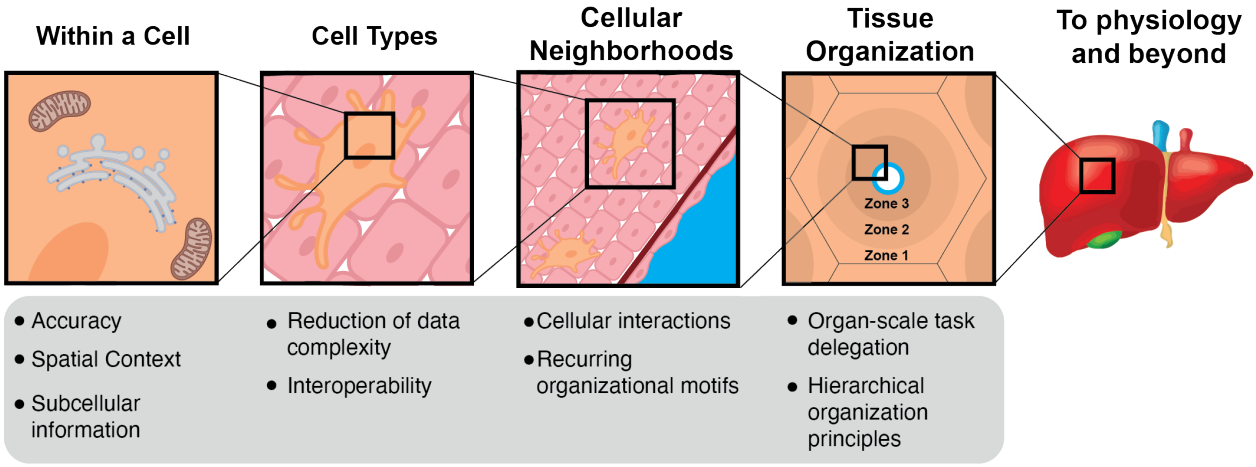


Figure 1.2: Bridging scales with *in situ* technologies. *In situ* technologies can reveal new biology across many scales of biology, including within a cell, cell types, cellular neighborhoods, tissue organization, and physiology. By bridging these scales, *in situ* technologies can provide insights into the structure-function relationship across multiple scales.

Within a cell

In situ measurements provide three key benefits over dissociative approaches in the analysis of single cells: 1) higher accuracy, 2) spatial context, and 3) subcellular information:

Accuracy: Many *in situ* techniques like MERFISH and seqFISH are more sensitive and less biased than their dissociative counterparts like scRNAseq. Therefore, for a broad range of biological questions that require accurate transcript numbers *in situ* technologies should be used. For example, analysis of gene expression variability is non-trivial using scRNAseq data

with sensitivities around 10%. Analysis of gene expression variability based on scRNAseq data requires accounting for this large measurement error with complex error models. Unfortunately, these are non-trivial and introduce a large number of additional assumptions, such as a high degree of transcriptional bursting (Jiang et al., 2017; Larsson et al., 2019), that are not always fully substantiated (Battich et al., 2015; Foreman and Wollman, 2020).

Spatial Context: The spatial context of *in situ* technologies allow for analysis of cellular heterogeneity in a much more physiological context. To fully understand the sources of cellular heterogeneity, we need to understand what factors influence cell state. Does spatial position in a tissue affect the variance of key genes? How does a cells' gene expression predict its present and future behavior? Efforts to track cells over time have revealed that understanding a cell's gene expression is insufficient to understand the choices that cells make (Weinreb et al., 2020). More information about a cell is therefore imperative to know in order to understand how a cell makes decisions. Recent work identified more than 40 genes in the mouse hippocampus to be cell subtype-specific spatial differentially expressed genes (spDEGs) (Littman et al., 2020). These results suggest that a spatial position can explain much of the heterogeneity seen using dissociative approaches.

Subcellular Information: A subset of *in situ* techniques are capable of discerning RNA and protein localization at the subcellular level. High resolution allows for the determination of expression patterns in organelles as well as the analysis of coexpression of genes by subcellular localization. MERFISH is one such technique and has characterized the RNA enrichment in the endoplasmic reticulum and the nucleus (Xia et al., 2019b) as well as the dendrites and axons of neurons (Wang et al., 2020). On the proteomics side, 4i allows subcellular detection of protein abundances (Gut et al., 2018). 4i goes further and determines that the subcellular spatial protein distribution between single cells that experience different cell

cycle states, microenvironments, or growth conditions affects the localization of EGFR upon the cell's exposure to EGF. Collectively, the accuracy, context, and resolution of many *in situ* technologies enable a more accurate picture of the biology of cells in a true physiological context.

Cell types: the building blocks of tissues

Classification of cells into (sub)types and states is an important step toward deciphering the structure/function relationship of tissues and organs. The two key advantages of classification of cells into (sub)types and state are 1) reduction of data complexity, i.e. a single cell type label can be used to replace a complex vector of transcriptome scale gene expression values. 2) interoperability between different experiments including across spatial and dissociate measurements, i.e the same nomenclature can be prescribed to cells across experiments. These benefits of cell classification systems and the existence of a large body of data from dissociative studies motivate many ongoing efforts to create robust cell classification systems (Trapnell, 2015; Yuste et al., 2020). However, the definition of a cell type and cell states is not consistent across fields, or even across researchers within a field. In addition, the appropriate criteria to use to classify cells are debated. This heterogeneity adds additional complications to cell classification, so it remains unclear if a single classification system will emerge or whether classification will have to be redefined for each analysis.

Three complementary and non-mutually exclusive views of cell types have been used as frameworks to determine cellular classification systems: landscape, microenvironment, and task (Fig 1.3). The first view, famously referred to as the Waddington landscape (Waddington, 1957a), suggests that intracellular biological regulatory networks are configured such that they can exist in a finite number of steady states. In the landscape point-of-view, cellular classification is molecular in origin and depends on stability analysis in high-dimensional phase

space (Ferrell, 2012; Trapnell, 2015). While inputs to the cell during its developmental trajectory can influence cell fate decisions, these transitions are still encoded by the underlying regulatory network and therefore the classification is focused on a cell's internal state (Waddington, 1957b). The second view is that a cell type is defined by its microenvironment: the chemical, mechanical, and biological cues surrounding the cell. The cell is influenced and shaped by its neighbors, its resources, and its environmental cues. The third view is that classification of cells into types has to follow the functional tasks cells are required to perform for the organism as a whole. Under this view, there are key cell archetypes, each specialized in a specific task. Each individual cell performs one or a few of these tasks and its molecular state will match the tasks it performs (Korem et al., 2015). Not only are the three views of landscape, microenvironment, and task non-mutually exclusive, they are in fact complementary and are likely different views of roles and states of cells in a multicellular organism. For example, the transition from monocyte to macrophage is guided by an internal epigenetic regulatory network (Álvarez-Errico et al., 2015). Macrophages can polarize to perform different tasks based on stimulatory cytokines (Murray, 2017) while at the same time are heavily influenced by the tissue microenvironment (Lavin et al., 2014). Combining multiple viewpoints to create one (or more) cell classification system is an important stepping stone in analyzing the cellular structure of tissues and organs.

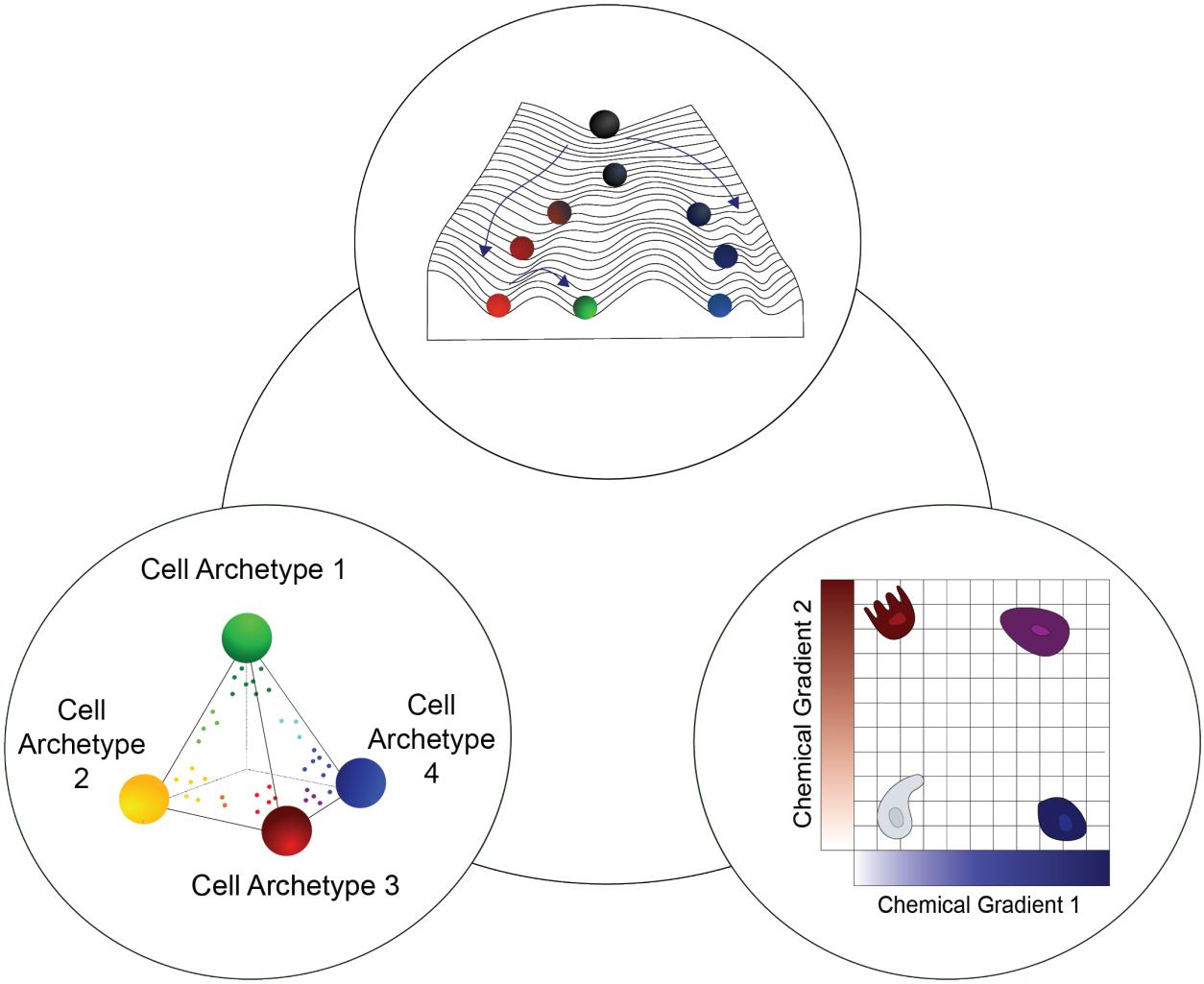


Figure 1.3: Geometrical representation of cell types. Three complementary views of the concept of cell type. These concepts are non-mutually exclusive and represent complementary views. (Top) The Waddington landscape uses the geometrical analogy of landscape. In this view, a cell type is a specific valley in 'cell space'. As pluripotent cells differentiate they pass through the landscape to reach their final position. This view is largely focused on the intracellular epigenetic and gene regulatory networks that define the possible valleys in the landscape. (Left) The task-based view proposes that each cell performs one or a few tasks. Each task (cell archetype) is represented as a vertice on a high dimensional polyhedron. The specific tasks each cell performs will determine its position within the polyhedron. (Right) The microenvironment view proposes that cell types are defined by the chemical, mechanical, and biological cues surrounding a cell. The cartoon shows a simplified view with two signaling gradients and the position of the cell in that space will determine its type.

In situ measurement technologies are well suited to generate and utilize cell classification systems. MERFISH and seqFISH were used to categorize the organization of predefined cell types within the brain (Chen et al., 2015; Littman et al., 2020; Shah et al., 2016). Other *in situ* technologies, such as *in situ* sequencing leverage the existing cell type taxonomies to overcome low RNA detection efficiency and still provide key cell type information (Qian et al.,

2020). Rather than solely focusing on existing classification systems, work based on seqFISH in combination with scRNAseq redefined cell types based on a Hidden Random Markov Field analysis of expression domains (Zhu et al., 2018). With further improvement of cell segmentation algorithms, it is likely that morphological information could be incorporated into classification models based on *in situ* measurements. Together with existing spatial information, it is expected that *in situ* technologies will play a key role in further refinement and development of cell type and state classification.

Cellular neighborhoods and communities

Cellular neighborhoods, the local spatial distribution of different cell types on the scale of hundreds of micrometers, are poorly understood. However, such length scales likely play an important role in bridging the gap between individual cell function and complex organ function. A good analogy for cellular communities is urban planning for human residential neighborhoods. A typical neighborhood with many houses will also have a coffee shop, a grocery store, and will be served by major roads and key public transportation. Similarly, cellular communities will have many cells of a few types that are needed for the specific organ (i.e. neurons in the brain, hepatocytes in the liver), but will also have resident macrophages, mast cells, and fibroblasts and will be in proximity to blood vessels. The number and spatial distribution of these specialized cell types have major implications for the function of the organ in their ability to relay information and perform their function (Bagnall et al., 2018). A good example of these principles come from recent cell-type mapping in the brain were *in situ* multiplexed RNA FISH uncovered a high spatial self-affinity of ependymal cells as well as spatial self-avoidance of inhibitory neurons, microglia, and astrocytes (Codeluppi et al., 2018). The paper also found that endothelial cells were found within roughly 65 microns of all other cell types. Another principle that will likely help identify cellular communities is communication between cells. Direct measurement of communication between cells is challenging, but a useful proxy is ligand-

receptor interactions in neighboring cells (Browaeys et al., 2020). Work that used multiplexed RNA measurement in brain slices (Eng et al., 2019) found that endothelial cells next to microglia in the olfactory bulb express endoglin and activin A receptor mRNA while the microglia expressed TGFB ligand mRNA (Eng et al., 2019). By contrast, endothelial cells adjacent to microglia in the cortex expressed Lrp1 and Pdgfb mRNA. Collectively such studies bring an intriguing hypothesis that there are key principles that could be generalized to identify community-level 'rules' of cellular patterning. What exactly are these rules and what are the molecular mechanisms used to implement them, e.g. the chemical gradient (Lander et al., 2009) and differential adhesion (Tsai et al., 2020), are key open questions.

In situ technologies coupled with new analysis approaches are well-positioned to make valuable contributions to our understanding of cellular communities. The highly multiplexed and inherently spatial nature of *in situ* measurement technologies makes them an ideal tool to acquire the data needed to understand cellular community organization. However, data collection is only the first step in identifying the rules and principles that govern cellular community organization. New bioinformatics and statistical tools will be required to allow researchers to convert the raw data on molecular distributions of RNA and proteins into insights. The rich literature of statistical learning including ideas related to community detection in multi-layer networks (Mucha et al., 2010) and concepts from topic modeling (Blei et al., 2003) will accelerate the development of these much-needed statistical analysis tools for cellular community organization. Initial implementation of ideas from topical modeling to cellular communities is very promising (Chen et al., 2020c). Overall, while the amount of work done so far to understand cellular neighborhoods remains small, the iterative development of data collection and analysis tools presents enticing prospects for understanding the role of the microenvironment in cellular organization.

Principles of tissue organization

How do multiple cellular communities interact together to form complex organs is a fundamental question. It is unclear whether there are few fundamental principles that can explain cellular self-organization across multiple organs or whether the way that multiple cellular communities synergy is organ dependent. It is likely that the organization of complex organs such as the brain or the liver that perform many distinct functions are different from each other and from simpler tissues such as the intestine or cornea. Nonetheless, the lack of complete data on cell type, RNA, and protein distribution across entire organs makes answering such questions difficult.

The liver was the first organ to be studied in depth with *in situ* approaches. The largest internal organ in the body, it performs roughly 500 tasks, including bile production, fat metabolization, vitamin and mineral storage, and blood filtration (Ben-Moshe and Itzkovitz, 2019). These tasks are non-homogeneously carried out by different subsets of cells within the liver. While it has long been understood that the various functions of the liver are not all carried out in the same spaces, *in situ* approaches have allowed researchers to dive further into the detailed arrangement of cell types and functions throughout the liver. Halpern et al. showed that roughly half of the hepatocyte genes, the main cells of the liver, are expressed in a zonated manner (Halpern et al., 2017). A subsequent study showed that liver endothelial cells are also highly zonated, with more than 30% of their genes expressed in a zonated manner (Halpern et al., 2018). Using spatial mapping, a high-resolution, global expression map of liver zonation was created that showed tasks that are high-energy are carried out in the highly oxygenated periportal locule layers where hepatocytes can more readily generate ATP through respiration (Halpern et al., 2017). This conclusion supports theoretical results on spatial task allocation in organs (Adler et al., 2019). While the work on the liver has shown exciting insights into its spatial organization it is still nascent and does not differentiate between the different lobes of the

liver, begging the question of whether each lobe shows additional sub-specializations. Outstanding questions of the liver organization still remain, including whether specific cell types including Kupffer cells (Bykov et al., 2004) and hepatic stellate cells (Friedman, 2008) are spatially heterogeneous. As technology develops, we anticipate exciting findings on the spatial organization of the liver's 500 tasks. The example of the liver shows how much was already learned, yet at the same time how much more there is to discover on tissue spatial organization using *in situ* measurement technologies.

To physiology and beyond

The ultimate goal of biomedical research is to improve our understanding of human biology and how it is disrupted during disease. From a translational point of view, it is often an organ function that is impacted by disease. In situ measurement technologies are poised to provide new insights on normal physiology and importantly provide detailed information on what goes wrong in a disease state. In situ techniques have been applied to a subset of organ diseases. Systematic charting of the brain (Moffitt et al., 2018; Shah et al., 2017; Zhang et al., 2020), heart (Asp et al., 2019), liver (Ben-Moshe and Itzkovitz, 2019; Halpern et al., 2017, 2018), intestine (Moor et al., 2018), and bone marrow (Baccin et al., 2020) are starting to uncover the single-cell architecture of multiple tissues and organs. Spatial Transcriptomics has been used to study ALS (Maniatis et al., 2019), prostate cancer (Berglund et al., 2018), melanoma (Thrane et al., 2018), and Alzheimer's disease (Chen et al., 2020b) while MIBI-TOF and imaging mass cytometry have been applied to the study of breast cancer (Jackson et al., 2020; Keren et al., 2018). Systematic efforts to scale *in situ* mapping to tumors are ongoing (Rozenblatt-Rosen et al., 2020).

Spatial maps of prostate cancer transcriptomes have shown prostate cancer samples have high heterogeneity across the tumor and that distinct cancer expression regions can

extend beyond the boundaries of annotated tumor areas (Berglund et al., 2018). These findings suggest using spatial information of tumors is important in classification schemes to rank tumor severity and could be used to predict further 'high risk' areas of potential cancer growth. Similarly, an analysis of triple-negative breast cancer by MIBI-TOF found that the spatial organization of infiltrating immune cells inside solid tumors is predictive of patient survival, where patients with more compartmentalized immune cells inside tumors fared better than patients where immune cells were well mixed within the tumor (Keren et al., 2018). Spatial transcriptomics has also been used in other diseases to track how disease progression occurs molecularly. A recent study of ALS quantified over 11,000 genes in mice and over 9,000 genes in humans to show that microglial dysfunction occurs well before ALS symptom onset and this dysfunction is mediated by the phagocytosis-related genes TREM2 and TYROBP (Maniatis et al., 2019). These technologies are pushing our frontier of understanding and even show areas where *in situ* analyses can suggest improvements in current medical practices. On a longer timescale, it is possible that *in situ* measurements will become an important diagnostic tool.

Conclusion

Spatial Biology is still an emerging field driven in large part by new *in situ* technologies. The ability of these approaches to provide rich spatially defined datasets about molecular and cellular distribution across multiple spatial scales poise these technologies to make critical contributions to our understanding of the inherent relationship between structure and function at multiple levels. The future of this field is quite bright, with applications ranging from understanding disease progression and how the structure of organs such as the brain and the liver relates to their functions. As was the case with single-cell biology, an increase in the adoption of these technologies will increase data availability and will result in innovation in data analysis. Such iterative improvements in data acquisition and analysis will provide key insights

that will allow researchers to bridge the gap from molecular and cellular biology to complex human physiology. The best is yet to come.

Acknowledgements

Results in this chapter were adapted from a manuscript published in *Cell Systems*:

Nagle, M. P., Tam, G. S., Maltz, E., Hemminger, Z. & Wollman, R. Bridging scales: From cell biology to physiology using *in situ* single-cell technologies. *Cell Syst* **12**, 388–400 (2021)

References

Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. *33*, 503–509.

Adler, M., Korem Kohanim, Y., Tendler, A., Mayo, A., and Alon, U. (2019). Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. Cell Syst.

Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M., and Ballestar, E. (2015). Epigenetic control of myeloid cell differentiation, identity and function. Nat. Rev. Immunol. *15*, 7–17.

American Association of Immunologists (1942). The Demonstration of Pneumococcal Antigen in Tissues by the Use of Fluorescent Antibody. The Journal of Immunology *45*, 159–170.

Angelo, M., Bendall, S.C., Finck, R., Hale, M.B., Hitzman, C., Borowsky, A.D., Levenson, R.M., Lowe, J.B., Liu, S.D., Zhao, S., et al. (2014). Multiplexed ion beam imaging of human breast tumors. Nat. Med. *20*, 436–442.

Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., et al. (2019). A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. Cell *179*, 1647–1660.e19.

Asp, M., Bergenstråhle, J., and Lundeberg, J. (2020). Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. Bioessays e1900221.

Baccin, C., Al-Sabah, J., Velten, L., Helbling, P.M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L.M., Trumpp, A., and Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat. Cell Biol. *22*, 38–48.

Bagnall, J., Boddington, C., England, H., Brignall, R., Downton, P., Alsoufi, Z., Boyd, J., Rowe, W., Bennett, A., Walker, C., et al. (2018). Quantitative analysis of competitive cytokine signaling predicts tissue thresholds for the propagation of macrophage activation. Sci. Signal. *11*.

Baharlou, H., Canete, N.P., Cunningham, A.L., Harman, A.N., and Patrick, E. (2019). Mass Cytometry Imaging for the Study of Human Diseases-Applications and Data Analysis Strategies.

Front. Immunol. *10*, 2657.

Bannon, D., Moen, E., Schwartz, M., Borba, E., Kudo, T., Greenwald, N., Vijayakumar, V., Chang, B., Pao, E., Osterman, E., et al. (2021). DeepCell Kiosk: scaling deep learning-enabled cellular image analysis with Kubernetes. Nat. Methods *18*, 43–45.

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of Transcript Variability in Single Mammalian Cells. Cell *163*, 1596–1610.

Ben-Moshe, S., and Itzkovitz, S. (2019). Spatial heterogeneity in the mammalian liver. Nat. Rev. Gastroenterol. Hepatol. *16*, 395–410.

Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat. Commun. *9*, 2419.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat. Methods *17*, 159–162.

Burkhard, S.B., and Bakkers, J. (2018). Spatially resolved RNA-sequencing of the embryonic heart identifies a role for Wnt/β-catenin signaling in autonomic control of heart rate. Elife *7*.

Bykov, I., Ylipaasto, P., Eerola, L., and Lindros, K.O. (2004). Functional Differences between Periportal and Perivenous Kupffer Cells Isolated by Digitonin-Collagenase Perfusion. Comp. Hepatol. *3 Suppl 1*, S34.

Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X., et al. (2020). RIC-seq for global *in situ* profiling of RNA-RNA spatial interactions. Nature *582*, 432–437.

Chen, F., Wassie, A.T., Cote, A.J., Sinha, A., Alon, S., Asano, S., Daugharthy, E.R., Chang, J.-B., Marblestone, A., Church, G.M., et al. (2016). Nanoscale imaging of RNA with expansion microscopy. Nat. Methods *13*, 679–684.

Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front. Genet. *10*, 317.

Chen, J., Suo, S., Tam, P.P., Han, J.-D.J., Peng, G., and Jing, N. (2017). Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nat. Protoc. *12*, 566–580.

Chen, J., Ding, L., Viana, M.P., Lee, H., Filip Sluezwski, M., Morris, B., Hendershott, M.C., Yang, R., Mueller, I.A., and Rafelski, S.M. (2020a). The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.

Chen, W.-T., Lu, A., Craessaerts, K., Pavie, B., Sala Frigerio, C., Corthout, N., Qian, X., Laláková, J., Kühnemund, M., Voytyuk, I., et al. (2020b). Spatial Transcriptomics and In situ Sequencing to Study Alzheimer's Disease. Cell *182*, 976–991.e19.

Chen, X., Sun, Y.-C., Church, G.M., Lee, J.H., and Zador, A.M. (2018). Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. Nucleic Acids Res. *46*, e22.

Chen, Z., Soifer, I., Hilton, H., Keren, L., and Jojic, V. (2020c). Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments. J. Comput. Biol.

Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I., and Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osmFISH. Nat. Methods *15*, 932–935.

Edfors, F., Hober, A., Linderbäck, K., Maddalo, G., Azimi, A., Sivertsson, Å., Tegel, H., Hober, S., Szigyarto, C.A.-K., Fagerberg, L., et al. (2018). Enhanced validation of antibodies for research applications. Nat. Commun. *9*, 4130.

Edsgärd, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. Nat. Methods *15*, 339–342.

Eng, C.-H.L., 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.

Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of single RNA transcripts *in situ*. Science *280*, 585–590.

Ferrell, J.E., Jr (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. Curr. Biol. *22*, R458–R466.

Foreman, R., and Wollman, R. (2020). Mammalian gene expression variability is explained by underlying cell state. Mol. Syst. Biol. *16*, e9146.

Frieda, K.L., Linton, J.M., Hormoz, S., Choi, J., Chow, K.-H.K., Singer, Z.S., Budde, M.W., Elowitz, M.B., and Cai, L. (2017). Synthetic recording and *in situ* readout of lineage information in single cells. Nature *541*, 107–111.

Friedman, S.L. (2008). Hepatic stellate cells: protean, multifunctional, and enigmatic cells of the liver. Physiol. Rev. *88*, 125–172.

Gerdes, M.J., Sevinsky, C.J., Sood, A., Adak, S., Bello, M.O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R.J., et al. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. Proc. Natl. Acad. Sci. U. S. A. *110*, 11982–11987.

Giesen, C., Wang, H.A.O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J.M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat. Methods *11*, 417–422.

Goh, J.J.L., Chou, N., Seow, W.Y., Ha, N., Cheng, C.P.P., Chang, Y.-C., Zhao, Z.W., and Chen, K.H. (2020). Highly specific multiplexed RNA imaging in tissues with split-FISH. Nat. Methods.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Cell *174*, 968–981.e15.

Gut, G., Herrmann, M.D., and Pelkmans, L. (2018). Multiplexed protein maps link subcellular

organization to cellular states. Science *361*.

Gyllborg, D., Langseth, C.M., Qian, X., Choi, E., Salas, S.M., Hilscher, M.M., Lein, E.S., and Nilsson, M. (2020). Hybridization-based *in situ* sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. Nucleic Acids Res.

Halpern, K.B., Shenhav, R., Matcovitch-Natan, O., Toth, B., Lemze, D., Golan, M., Massasa, E.E., Baydatch, S., Landen, S., Moor, A.E., et al. (2017). Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature *542*, 352–356.

Halpern, K.B., Shenhav, R., Massalha, H., Toth, B., Egozi, A., Massasa, E.E., Medgalia, C., David, E., Giladi, A., Moor, A.E., et al. (2018). Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nat. Biotechnol. *36*, 962–970.

Hu, K.H., Eichorst, J.P., McGinnis, C.S., Patterson, D.M., Chow, E.D., Kersten, K., Jameson, S.C., Gartner, Z.J., Rao, A.A., and Krummel, M.F. (2020). ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nat. Methods *17*, 833–843.

Ijsselsteijn, M.E., van der Breggen, R., Farina Sarasqueta, A., Koning, F., and de Miranda, N.F.C.C. (2019). A 40-Marker Panel for High Dimensional Characterization of Cancer Immune Microenvironments by Imaging Mass Cytometry. Front. Immunol. *10*, 2534.

Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., et al. (2020). The single-cell pathology landscape of breast cancer. Nature *578*, 615–620.

Jiang, Y., Zhang, N.R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biol. *18*, 74.

Kann, A.P., and Krauss, R.S. (2019). Multiplexed RNAscope and immunofluorescence on whole-mount skeletal myofibers and their associated stem cells. Development *146*.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. Nat. Methods *10*, 857–860.

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.-R., Kurian, A., Van Valen, D., West, R., et al. (2018). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. Cell *174*, 1373–1387.e19.

Keren, L., Bosse, M., Thompson, S., Risom, T., Vijayaragavan, K., McCaffrey, E., Marquez, D., Angoshtari, R., Greenwald, N.F., Fienberg, H., et al. (2019). MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. Sci Adv *5*, eaax5851.

Kishi, J.Y., Lapan, S.W., Beliveau, B.J., West, E.R., Zhu, A., Sasaki, H.M., Saka, S.K., Wang, Y., Cepko, C.L., and Yin, P. (2019). SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat. Methods *16*, 533–544.

Korem, Y., Szekely, P., Hart, Y., Sheftel, H., Hausser, J., Mayo, A., Rothenberg, M.E., Kalisky, T., and Alon, U. (2015). Geometry of the Gene Expression Space of Individual Cells. PLoS Comput. Biol. *11*, e1004224.

Kwon, S., Chin, K., and Nederlof, M. (2020). Simultaneous Detection of RNAs and Proteins with

Subcellular Resolution. In RNA-Chromatin Interactions: Methods and Protocols, U.A.V. Ørom, ed. (New York, NY: Springer US), pp. 59–73.

Lander, A.D., Lo, W.-C., Nie, Q., and Wan, F.Y.M. (2009). The measure of success: constraints, objectives, and tradeoffs in morphogen-mediated patterning. Cold Spring Harb. Perspect. Biol. *1*, a002022.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., and Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell *159*, 1312–1326.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. Science *343*, 1360–1363.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M., Yang, J.L., Lee, H.S., Aach, J., et al. (2015). Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat. Protoc. *10*, 442–458.

Lin, J.-R., Fallahi-Sichani, M., and Sorger, P.K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. Nat. Commun. *6*, 8390.

Lin, J.-R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P.M., Santagata, S., and Sorger, P.K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. Elife *7*.

Littman, R., Hemminger, Z., Foreman, R., Arneson, D., Zhang, G., Gómez-Pinilla, F., Yang, X., and Wollman, R. (2020). JSTA: joint cell segmentation and cell type annotation for spatial transcriptomics.

Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S., Francis, C., Spaethling, J.M., Wolf, J.A., Grady, M.S., et al. (2014). Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat. Methods *11*, 190–196.

Lundberg, E., and Borner, G.H.H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nat. Rev. Mol. Cell Biol. *20*, 285–302.

Maiser, A., Dillinger, S., Längst, G., Schermelleh, L., Leonhardt, H., and Németh, A. (2020). Super-resolution *in situ* analysis of active ribosomal DNA chromatin organization in the nucleolus. Sci. Rep. *10*, 7462.

Maniatis, S., Äijö, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., Fagegaltier, D., Andrusivová, Ž., Saarenpää, S., Saiz-Castro, G., et al. (2019). Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science *364*, 89–93.

Merritt, C.R., Ong, G.T., Church, S.E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J., Liang, Y., McKay-Fleisch, J., et al. (2020). Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nat. Biotechnol. *38*, 586–599.

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., and Van Valen, D. (2019). Deep learning for cellular image analysis. Nat. Methods.

Moffitt, J.R., Hao, J., Wang, G., Chen, K.H., Babcock, H.P., and Zhuang, X. (2016a). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence *in situ* hybridization. Proc. Natl. Acad. Sci. U. S. A. *113*, 11046–11051.

Moffitt, J.R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C., and Zhuang, X. (2016b). High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. Proc. Natl. Acad. Sci. U. S. A. *113*, 14456–14461.

Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science *362*.

Moor, A.E., Harnik, Y., Ben-Moshe, S., Massasa, E.E., Rozenberg, M., Eilam, R., Bahar Halpern, K., and Itzkovitz, S. (2018). Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis. Cell *175*, 1156–1167.e15.

Motta, P.M. (1998). Marcello Malpighi and the foundations of functional microanatomy. Anat. Rec. *253*, 10–12.

Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. Science *328*, 876–878.

Murray, P.J. (2017). Macrophage Polarization. Annu. Rev. Physiol. *79*, 541–566.

Nichterwitz, S., Chen, G., Aguila Benitez, J., Yilmaz, M., Storvall, H., Cao, M., Sandberg, R., Deng, Q., and Hedlund, E. (2016). Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nat. Commun. *7*, 12139.

Palade, G.E., and Porter, K.R. (1954). Studies on the endoplasmic reticulum. I. Its identification in cells *in situ*. J. Exp. Med. *100*, 641–656.

Perkel, J.M. (2019). Starfish enterprise: finding RNA patterns in single cells. Nature *572*, 549–551.

Ptacek, J., Locke, D., Finck, R., Cvijic, M.-E., Li, Z., Tarolli, J.G., Aksoy, M., Sigal, Y., Zhang, Y., Newgren, M., et al. (2020). Multiplexed ion beam imaging (MIBI) for characterization of the tumor microenvironment across tumor types. Lab. Invest.

Qian, X., Harris, K.D., Hauling, T., Nicoloutsopoulos, D., Muñoz-Manchado, A.B., Skene, N., Hjerling-Leffler, J., and Nilsson, M. (2020). Probabilistic cell typing enables fine mapping of closely related cell types *in situ*. Nat. Methods *17*, 101–106.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.

Rouhanifard, S.H., Mellis, I.A., Dunagin, M., Bayatpour, S., Jiang, C.L., Dardani, I., Symmons,

O., Emert, B., Torre, E., Cote, A., et al. (2018). ClampFISH detects individual nucleic acid molecules using click chemistry-based amplification. Nat. Biotechnol.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell *181*, 236–249.

Saka, S.K., Wang, Y., Kishi, J.Y., Zhu, A., Zeng, Y., Xie, W., Kirli, K., Yapp, C., Cicconet, M., Beliveau, B.J., et al. (2019). Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nat. Biotechnol. *37*, 1080–1090.

Salmén, F., Ståhl, P.L., Mollbrink, A., Navarro, J.F., Vickovic, S., Frisén, J., and Lundeberg, J. (2018). Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. Nat. Protoc. *13*, 2501–2534.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Schulz, D., Zanotelli, V.R.T., Fischer, J.R., Schapiro, D., Engler, S., Lun, X.-K., Jackson, H.W., and Bodenmiller, B. (2018). Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. Cell Syst *6*, 25–36.e5.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron *92*, 342–357.

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

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. *20*, 631–656.

Stark, Z., Schofield, D., Alam, K., Wilson, W., Mupfeki, N., Macciocca, I., Shrestha, R., White, S.M., and Gaff, C. (2017). Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet. Med. *19*, 867–874.

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2020). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nat. Methods *18*, 100–106.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell

Data. Cell *177*, 1888–1902.e21.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Thrane, K., Eriksson, H., Maaskola, J., Hansson, J., and Lundeberg, J. (2018). Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. Cancer Res. *78*, 5970–5979.

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. Cell Syst *6*, 171–179.e5.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

Tsai, T.Y.-C., Sikora, M., Xia, P., Colak-Champollion, T., Knaut, H., Heisenberg, C.-P., and Megason, S.G. (2020). An adhesion code ensures robust pattern formation during tissue morphogenesis. Science *370*, 113–116.

Tutucci, E., and Singer, R.H. (2020). Simultaneous Detection of mRNA and Protein in S. cerevisiae by Single-Molecule FISH and Immunofluorescence. In RNA Tagging: Methods and Protocols, M. Heinlein, ed. (New York, NY: Springer US), pp. 51–69.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for *in situ* tissue profiling. Nat. Methods *16*, 987–990.

Waddington, C.H. (1957a). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology. With an Appendix by H. Kacser.

Waddington, C.H. (1957b). The Strategy of the Genes. Allen.

Wang, C., Lu, T., Emanuel, G., Babcock, H.P., and Zhuang, X. (2019). Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. Proc. Natl. Acad. Sci. U. S. A. *116*, 10842–10851.

Wang, F., Flanagan, J., Su, N., Wang, L.-C., Bui, S., Nielson, A., Wu, X., Vo, H.-T., Ma, X.-J., and Luo, Y. (2012). RNAscope: a novel *in situ* RNA analysis platform for formalin-fixed, paraffin-embedded tissues. J. Mol. Diagn. *14*, 22–29.

Wang, G., Ang, C.-E., Fan, J., Wang, A., Moffitt, J.R., and Zhuang, X. (2020). Spatial organization of the transcriptome in individual neurons.

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science *361*.

Wang, Y., Woehrstein, J.B., Donoghue, N., Dai, M., Avendaño, M.S., Schackmann, R.C.J., Zoeller, J.J., Wang, S.S.H., Tillberg, P.W., Park, D., et al. (2017). Rapid Sequential *in situ* Multiplexing with DNA Exchange Imaging in Neuronal Cells and Tissues. Nano Lett. *17*, 6131–

6139.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*.

Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873–1887.e17.

Xia, C., Babcock, H.P., Moffitt, J.R., and Zhuang, X. (2019a). Multiplexed detection of RNA using MERFISH and branched DNA amplification. Sci. Rep. *9*, 7721.

Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019b). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc. Natl. Acad. Sci. U. S. A.

Young, A.P., Jackson, D.J., and Wyeth, R.C. (2020). A technical review and guide to RNA fluorescence *in situ* hybridization. PeerJ *8*, e8806.

Yuste, R., Hawrylycz, M., Aalling, N., Aguilar-Valles, A., Arendt, D., Arnedillo, R.A., Ascoli, G.A., Bielza, C., Bokharaie, V., Bergmann, T.B., et al. (2020). A community-based transcriptomics classification and nomenclature of neocortical cell types. Nat. Neurosci.

Zhang, M., Eichhorn, S.W., Zingg, B., Yao, Z., Zeng, H., Dong, H., and Zhuang, X. (2020). Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by *in situ* single-cell transcriptomics.

Zhu, Q., Shah, S., Dries, R., Cai, L., and Yuan, G.-C. (2018). Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence *in situ* hybridization data. Nat. Biotechnol.

(2014). Method of the year 2013. Nat. Methods *11*, 1.

CHAPTER 2

Information Transfer in Stimulus-Specific Activation of Macrophages

Abstract

Macrophages control inflammation responses through the NFKB and p38 pathways. Individually, these two pathways control cellular response through temporal patterns of activity that initiate gene expression changes. How these pathways use combinatorial as well as temporal activation to further refine responses to stimuli is not well understood. We utilize a dual-reporter macrophage system to explore the dynamics of p38 and NFKB. We show that both pathways show stimulus-specific response to PAMPS through high-throughput live-cell microscopy as well as MERFISH. Together, these results provide a clearer understanding of the innate immune system.

Introduction

Cells in the body contend with complex environments and are required to respond appropriately to extracellular signals. This fact is especially true for cells in the immune system, whose constant vigilance via the innate immune system is a key component to maintaining health. Signaling that activates appropriate gene expression profiles to accurately and quickly respond to a threat is a critical component in the innate immune system. Two broad hypothesis describe the process by which a cell encodes information about an extracellular threat: the temporal coding hypothesis[1] and the combinatorial encoding hypothesis[2]. The temporal coding hypothesis suggests that information about an extracellular stimulus is encoded in the dynamic of a signaling molecule over time. Combinatorial encoding has been shown to allow cells to perceive information encoded by simultaneous activation of multiple signaling pathways[3,4].

Two major pathways are critical in orchestrating immune response. NFKB controls transcription of DNA, cytokine production, and cell survival. Misregulation of this pathway has been linked to cancer, autoimmune disease, and viral infection[5]. Mitogen-activated protein kinases (MAPK) regulate many cell responses. p38 MAPKs are involved in the production of many inflammatory mediators. p38 plays an essential role in regulating many cellular processes, especially inflammation. The p38 kinase has a well-documented link to the NF-κB pathway[5,6]. Combinatorial signaling through both p38 and NF-κB has been shown to regulate strength of immune response in macrophages[7,8].

Macrophages are among the most fascinating cells in the human body, ranging in function from ingesting and degrading dead cells to wound healing and inflammation response[9]. These myeloid immune cells are found in practically all tissues in the body[10]. Macrophages respond to a wide array of pathogen association molecular patterns (PAMPS) that represent broad classes of cellular threats[11]. Their ability to correctly identify both the identity and the scale of the threat is crucial to mounting an appropriate response. Immune cells have been shown to have gene expression changes that are stimulus- and pathogen-specific[12–14]. These cells make an excellent model system for the study of immune activation. In this chapter, we investigate p38 and NFKB signaling and response in macrophages using a dual-reporter system.

Results

To build upon previous work in establishing the encoding and decoding of signaling dynamics in cell lines, we intended to study the combinatorial and temporal patterns of NF-KB and p38 signaling in primary macrophages in response to a wide range of immune threats (Figure 2.1). Using a previously described mVenus-RelA reporter mouse line[15] transduced with

a mCerulean p38 kinase translocation reporter (KTR)[16], we developed a dual-reporter system for the analysis of p38 and NF-kB response in bone-marrow derived macrophages. Previous work has used RelA fusion proteins ectopically expressed in immortalized cell lines[17,18]. Ectopic expression can lead to artifacts[19] and immortalized cell lines have been shown to have diminished immune responses[20]. Kinase translocation reporters (KTRs) have proven to be effective measurement tools of kinase activity *in vivo* and offer improvements longevity and multiplexing opportunities over the previous most popular kinase reporter, FRET[16,21]. Here, we show the first single-cell study of NF-kB and p38 trajectories in primary macrophages.

Macrophages differentiated from primary bone-marrow cells derived from homogenous $RelA^{v/v}$ mice showed normal levels of nuclear activity of both signaling proteins. Cells were stimulated with the inflammatory molecules TNF, Pam3CSK, and LPS (Figure 2.1). These three ligands represent a diverse array of ligands sources: a native cytokine that stimulates TNFR, a synthetic lipopeptide that stimulates TLR2, and a bacterial TLR4 agonist[22]. The amount of nuclear NFKB and p38 fluorescence in single cells was quantitated with a fully automated imaging pipeline as described in Selimkhanov et al 2014 and Zambrano et al 2016[23,24]. Cells were imaged on 40 mm coverslips treated with poly-l-lysine. We found that using PDMS to create wells was not practical when working with macrophages, as they tended to interact with irregular pieces of PDMS on the walls of the wells. We suspect that using PDMS wells could be a source of artificial activation of some macrophages. Instead, we adapted wells from Ibidi removable chambers to fit on 40 mm coverslips. Macrophages did not interact with these wells and data collected from the 40 mm coverslips correlated highly with experiments performed on 8-well Ibidi SlideTek chambers.

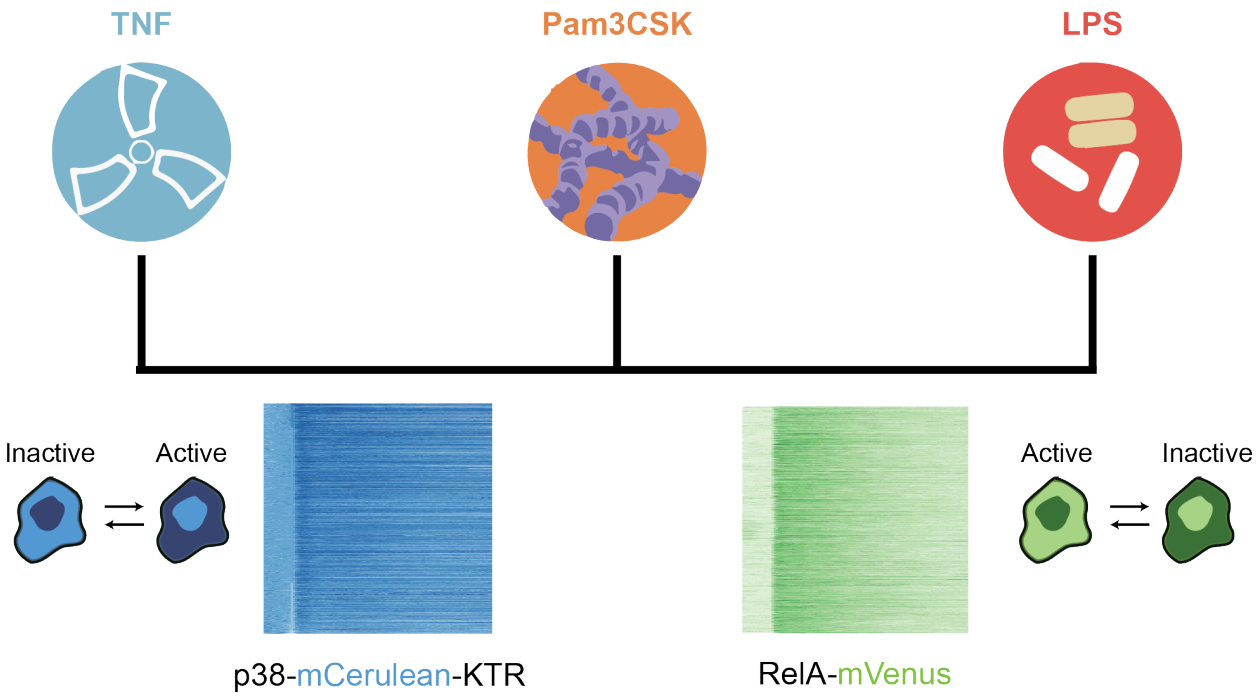


Figure 2.1 Schematic of innate immune signaling network activating p38 and NFκB.

Obvious differences in both NFκB and p38 dynamics are seen at the single-cell level (Figure 2.2). As seen in previous work, LPS shows strong initial and sustained RelA response and TNF induces an oscillatory response in the RelA reporter that desynchronizes over time[1,15]. P3C4 shows a more delayed initial response with a similar overall trajectory to LPS. The p38 dynamics are similar to those of NFκB. The most diverging stimulus is TNF, which shows a fast, relatively weak initial response and a small second peak. This ligand does not seem to produce oscillatory dynamics. In p38 activation, LPS again shows a strong, broad peak of activity and P3C4 is similarly strong. While p38 signaling has previously been shown to oscillate in a FRET-based system in response to the cytokine IL-1β, it does not appear that any of the ligands in the current study induce oscillatory dynamics[25].

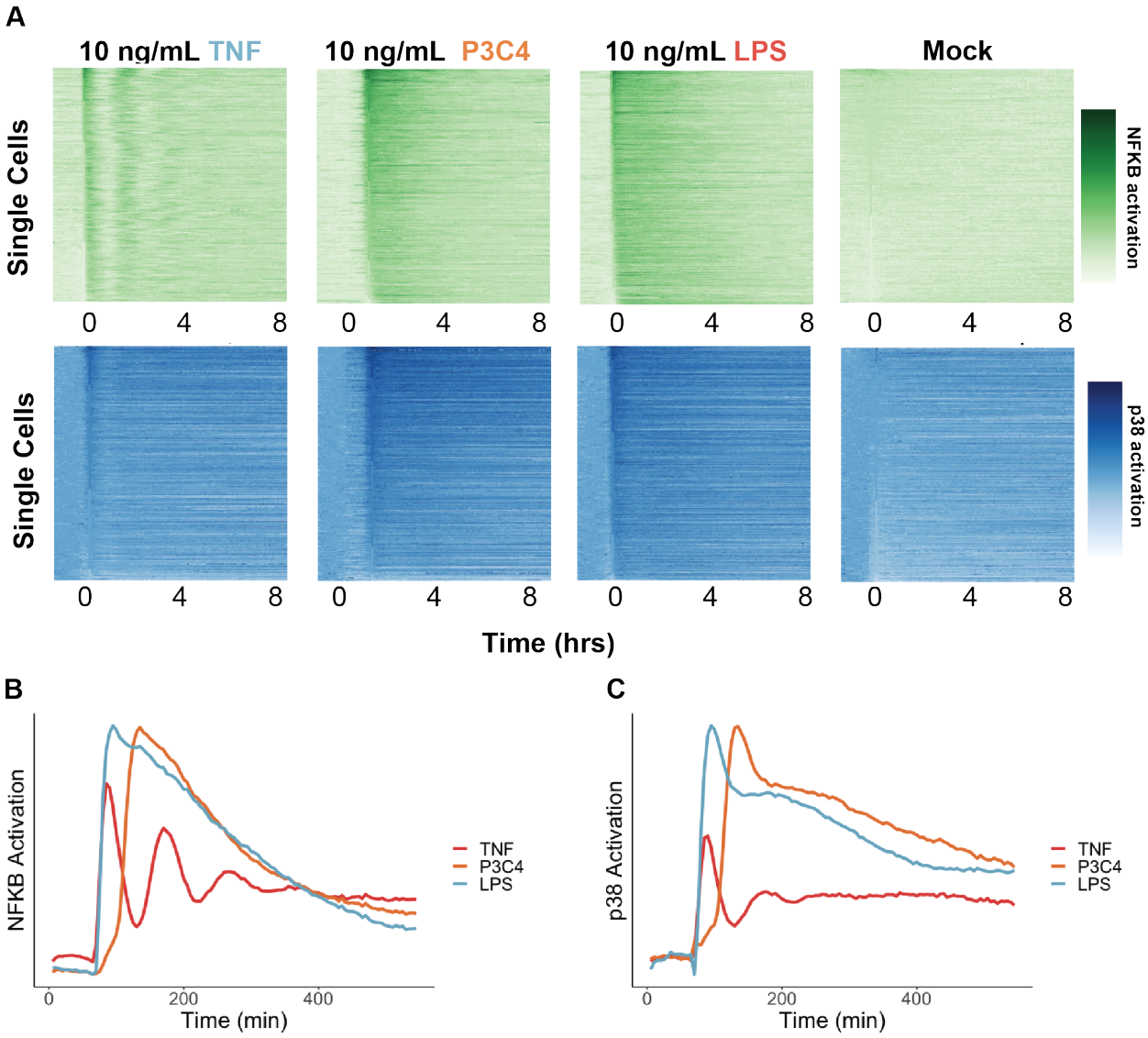


Figure 2.2 Stimulus-specific activation of p38 and NFκB. A) Heatmaps are ordered by maximum initial peak height. Top row shows p38 activation in response to ligands. Bottom row shows RelA activation. B) Average RelA response of cells to ligands. C) Average p38 response of cells to ligands.

In order to further investigate the stimulus-specific activation of p38 and RelA, we employed MERFISH, a method to count single molecules of RNA within cells, to determine the gene expression changes of each cell in response to the stimuli. Custom coverslips containing microgrids as fiduciary markers made the collection of such a multi-omic dataset possible (Figure 2.3). Macrophages were fixed immediately after eight hours of imaging. With this timescale, macrophages show distinct changes in many gene expression groups[14]. We seek to

understand the linkage between signaling activation of NFKB and p38 with gene expression changes. The work of combining these MERFISH datasets with live-cell imaging is ongoing.

In determining the decoding of external stimuli via changes in gene expression, it is critical to study at the single-cell level. Both signaling dynamics and gene expression changes show high degrees of heterogeneity, which means that population-level analyses miss crucial changes that average out across many cells. Coupling live-cell imaging and transcriptional readouts so that the information from the same cell can be linked together in a multi-omic dataset is critical to parsing out the details in the heterogeneity.

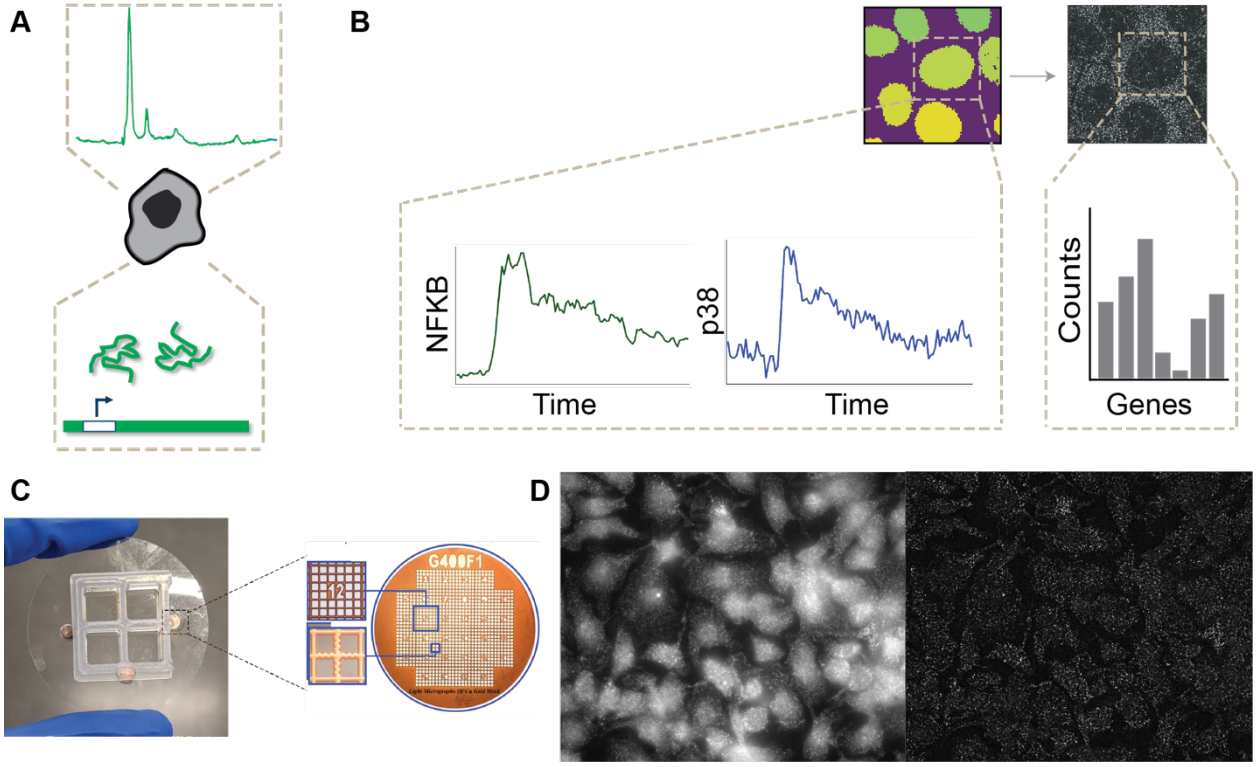


Figure 2.3 Paired single-cell MERFISH and live-cell imaging (A) Measuring live-cell signaling dynamics and gene expression changes in a single cell allow for paired datasets (B) Cells are segmented and signaling dynamics are measured for NFKB and p38. Then the same cells have their gene expression measured by MERFISH (C) Coverslip overview. Setup of coverslip with three microgrids and a modified Ibidi removable chamber. Live macrophage images are paired with MERFISH images via alignment of coordinates via microgrid orientation. (D) Images of macrophage cells during one round of hybridization. Left – raw image. Right – image with background subtraction.

Discussion

In this chapter, we determine the stimulus-specific response of RelA and p38 in macrophages to a variety of ligands. We show that NFκB has oscillatory dynamics to some stimuli, while p38 does not express oscillations in this dual-reporter system. While stimulus-specific temporal responses were obvious, lack of correlation between RelA and p38 responses leave it unclear as to whether combinatorial encoding plays a role in decoding the stimuli's identities.

Our ultimate goal is to generate high-quality multi-omic datasets linking the dynamics of multiple signaling molecules to gene expression changes at the single-cell level. It is well established that NFκB and p38 dynamics are correlated with gene expression[1,26,27]. Future work could also resolve more information in these linked datasets, such as microenvironment effects and neighbor similarities. Extracellular signaling molecules such as cytokines are known to affect the function of immune cells and could alter the encoding and decoding of cellular signaling[13]. Further exploration of the link between heterogeneity in signaling and gene expression are key to understanding the mechanistic underpinnings of the innate immune system.

Acknowledgements

Stefanie Leucke performed the live-cell experiments in this chapter. Zach Hemminger designed the probesets used for MERFISH. Roy Wollman and Alex Hoffmann assisted with experimental design.

Materials and Methods

Cell culture

BMDMs were prepared by culturing bone marrow monocytes from femurs of 8-12 week old in L929 -conditioned medium using standard methods[20]. BMDMs were re-plated in imaging dishes on day 4, then were stimulated on day 7 or day 8.

Live-cell imaging

Bone-marrow macrophages were replated on day 4 at 24,000 or 20,000/cm2 in custom coverslip chambers, for imaging at an appropriate density (approx. 60,000/cm2) on day 7 or day 8. 2 hours prior to stimulation, a solution of 2.5 ng/mL Hoechst 33342 was added to the BMDM culture media. After the start of imaging, additional culture media containing stimulus (TNF, LPS, or P3C4) was injected into the chamber *in situ*. Cells were imaged at 5-minute intervals on a Zeiss Axio Observer platform with live-cell incubation, using epifluorescent excitation from a Sutter Lambda XL light source. Images were recorded on a Hamamatsu Orca Flash 2.0 CCD camera.

Image analysis and processing

Microscopy time-lapse images were exported for single-cell tracking and measurement in MATLAB R2016a. The tracking routines followed those used in earlier work[23]. Briefly, cells were identified using DIC images, then segmented, guided by markers from the Hoechst image. Segmented cells were linked into trajectories across successive images, then nuclear and cytoplasmic boundaries were saved and used to define measurement regions in other fluorescent channels were quantified on a per-cell basis, normalized to image background levels, then were baseline-subtracted. Mitotic cells, as well as cells that drifted out of the field of view, were excluded from analysis.

Coverslip modification

Forty millimeter coverslips (Bioptechs) were allyl silane functionalized according to Moffitt et al (2016), which briefly consists of washing coverslips in 50% methanol and 50% 12M HCl, and then incubating at room temperature in 0.1% (vol/vol) triethylamine (Millipore) and 0.2% (vol/vol) allyltrichlorosilane (Sigma) in chloroform for 30 min. Wash with chloroform and then with 100% ethanol, and air-dry with nitrogen gas. These were stored in a desiccator for less than a month until use. Ibidi removable chambers (Ibidi 80841) were modified to contain only four wells and applied to the glass 40 mm round coverslips. Fiduciary grids in the form of Gilder Finder Grids (VWR) were glued on to the coverslips using small amounts of Gorilla Clear Glue.

Sequential FISH staining

Immediately after imaging cells were fixed for 5 minutes, washed 3x with PBS, then stored with 70% ethanol at -20 degrees C. After storage, the Ibidi wells were removed from the coverslip and the coverslips were washed 3x with PBS and then permeabilized with 0.5% Triton X-100 in PBS for 15 min. Coverslips were washed with 50 mM Tris and 300 mM NaCl (TBS), and then incubated overnight with MelpaX to functionally add acrolyl modificaitons. Coverslips were washed again with TBS and immersed in 30% formamide in TBS for 5 min to equilibrate, all the liquid was aspirated from the petri dishes, and 30 μL of 100 μM encoding probes were added on top of the coverslip. A piece of parafilm was place on top of the coverslip to evenly spread the small volume over the surface and prevent evaporation. The entire petri dish was sealed with parafilm and incubated at 37°C for 36–48 h. The parafilm was removed, and the coverslip was washed 2X with 30% formamide with 30-min incubation at 37°C for both washes. A 4% polyacrylamide hydrogel was then cast to embed the cells before clearing with 2% SDS, 0.5% Triton X-100, and 8 U/ml proteinase k (NEB P8107S), according to previously published methods. Coverslips were incubated in clearing buffer for 12 h and then washed 3x in TBS for 15 min each at room temperature.

MERFISH imaging

smFISH staining was imaged on a custom-modified Zeiss Axio Observer Z1 body with Andor Zyla 4.2 sCMOS camera and 1.4NA 63 Plan-Apo oil immersion objective. Illumination light was provided by LUXEON rebel LEDs (deep red and blue) to excite Cy5, Hoechst, and 200 nm deep blue fiducial markers. The microscope was controlled by micromanager[28] and custom MATLAB software. Automated washing during sequential rounds of hybridization was accomplished by using a previous published setup[29,30]. Briefly, FCS2 bioptech flow chambers were attached to a Gilson Minipuls peristaltic pump pulling liquid from reservoirs attached to Hamilton MVP valves. The pump and valves were controlled with Arduino, and serial commands with Python https://github.com/ZhuangLab/storm-control/tree/master/storm_control/fluidics. This setup was used to automatically wash cells with TBS, then 2 ml of TCEP (Sigma) in TBS incubated for 15 min, then rinse with TBS, then flow in 2 ml of wash buffer [10% ethylene carbonate in TBS with 0.3% polyvinyl sulfonic acid (VWR)], followed by 3 ml 3 nM readout probes in wash buffer incubated for 15 min, then rinsed with 2 ml wash buffer, then 1 ml of TBS, and finally 3 ml of imaging buffer. Imaging buffers 0.15 U/ml rPCO (OYCO), 2 mM PCA (Sigma), 2 mM Trolox (Sigma), 50 mM pH 8.0 Tris–HCl, 300 mM NaCl, and 40 U/ml murine RNase inhibitor (NEB). MERFISH imaging processing and gene calling were performed as described in Foreman et al[31].

References

1. Hoffmann, A., Levchenko, A., Scott, M. L. & Baltimore, D. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* **298**, 1241–1245 (2002).
2. Flores, G. V. *et al.* Combinatorial signaling in the specification of unique cell fates. *Cell* **103**, 75–85 (2000).

3. Klumpe, H. *et al.* The context-dependent, combinatorial logic of BMP signaling. *Cold Spring Harbor Laboratory* 2020.12.08.416503 (2020) doi:10.1101/2020.12.08.416503.
4. Antebi, Y. E. *et al.* Combinatorial Signal Perception in the BMP Pathway. *Cell* **170**, 1184-1196.e24 (2017).
5. Hoesel, B. & Schmid, J. A. The complexity of NF-κB signaling in inflammation and cancer. *Mol. Cancer* **12**, 86 (2013).
6. Vermeulen, L., De Wilde, G., Van Damme, P., Vanden Berghe, W. & Haegeman, G. Transcriptional activation of the NF-kappaB p65 subunit by mitogen- and stress-activated protein kinase-1 (MSK1). *EMBO J.* **22**, 1313–1324 (2003).
7. Miller-Jensen, K. Distinct Signaling Thresholds Distinguish Friend from Foe. *Cell Syst* **2**, 360–361 (2016).
8. Gottschalk, R. A. *et al.* Distinct NF-κB and MAPK Activation Thresholds Uncouple Steady-State Microbe Sensing from Anti-pathogen Inflammatory Responses. *Cell Syst* **2**, 378–390 (2016).
9. Varol, C., Mildner, A. & Jung, S. Macrophages: development and tissue specialization. *Annu. Rev. Immunol.* **33**, 643–675 (2015).
10. Bauer, J. *et al.* A Strikingly Constant Ratio Exists Between Langerhans Cells and Other Epidermal Cells in Human Skin. A Stereologic Study Using the Optical Disector Method and the Confocal Laser Scanning Microscope11Presented in part at the International Investigative Dermatology 1998, Cologne, Germany 1998. *J. Invest. Dermatol.* **116**, 313–318 (2001).
11. Medzhitov, R. & Horng, T. Transcriptional control of the inflammatory response. *Nat. Rev. Immunol.* **9**, 692–703 (2009).
12. Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–263 (2009).

13. Cheng, Q. *et al.* Sequential conditioning-stimulation reveals distinct gene- and stimulus-specific effects of Type I and II IFN on human macrophage functions. *Sci. Rep.* **9**, 5288 (2019).
14. Sheu, K., Luecke, S. & Hoffmann, A. Stimulus-specificity in the Responses of Immune Sentinel Cells. *Curr Opin Syst Biol* **18**, 53–61 (2019).
15. Adelaja, A. *et al.* Six distinct NFκB signaling codons convey discrete information to distinguish stimuli and enable appropriate macrophage responses. *Immunity* **54**, 916-930.e7 (2021).
16. Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S. & Covert, M. W. High-sensitivity measurements of multiple kinase activities in live single cells. *Cell* **157**, 1724–1734 (2014).
17. Ashall, L. *et al.* Pulsatile stimulation determines timing and specificity of NF-kappaB-dependent transcription. *Science* **324**, 242–246 (2009).
18. Tay, S. *et al.* Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267–271 (2010).
19. Mothes, J., Busse, D., Kofahl, B. & Wolf, J. Sources of dynamic variability in NF-κB signal transduction: a mechanistic model. *Bioessays* **37**, 452–462 (2015).
20. Cheng, Z., Taylor, B., Ourthiague, D. R. & Hoffmann, A. Distinct single-cell signaling characteristics are conferred by the MyD88 and TRIF pathways during TLR4 activation. *Sci. Signal.* **8**, ra69 (2015).
21. Kudo, T. *et al.* Live-cell measurements of kinase activity in single cells using translocation reporters. *Nat. Protoc.* **13**, 155–169 (2018).
22. Luecke, S., Sheu, K. M. & Hoffmann, A. Stimulus-specific responses in innate immunity: Multilayered regulatory circuits. *Immunity* **54**, 1915–1932 (2021).
23. Selimkhanov, J. *et al.* Systems biology. Accurate information transmission through dynamic biochemical signaling networks. *Science* **346**, 1370–1373 (2014).

24. Zambrano, S., De Toma, I., Piffer, A., Bianchi, M. E. & Agresti, A. NF-κB oscillations translate into functionally related patterns of gene expression. *Elife* **5**, e09100 (2016).

25. Tomida, T., Takekawa, M. & Saito, H. Oscillation of p38 activity controls efficient pro-inflammatory gene expression. *Nat. Commun.* **6**, 8350 (2015).

26. Gutschow, M. V. *et al.* Combinatorial processing of bacterial and host-derived innate immune stimuli at the single-cell level. *Mol. Biol. Cell* **30**, 282–292 (2019).

27. Whitmarsh, A. J. Regulation of gene transcription by mitogen-activated protein kinase signaling pathways. *Biochim. Biophys. Acta* **1773**, 1285–1298 (2007).

28. Edelstein, A. D. *et al.* Advanced methods of microscope control using μManager software. *J Biol Methods* **1**, (2014).

29. Moffitt, J. R. *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence *in situ* hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11046–11051 (2016).

30. Moffitt, J. R. *et al.* High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14456–14461 (2016).

31. Foreman, R. & Wollman, R. Mammalian gene expression variability is explained by underlying cell state. *Mol. Syst. Biol.* **16**, e9146 (2020).

CHAPTER 3

A method to tag endogenous genes using a barcoded CRISPR tag via intronic regions

Abstract

The localization and abundance of endogenous proteins provides a myriad of useful information. In this chapter we report a high-throughput method to study endogenous protein dynamics. We combine an intronic CRISPR tagging approach with a molecular barcode to identify uniquely tagged cells. We demonstrate this is a scalable endogenous tagging approach for single cell protein tracking. Furthermore, the addition of the *in situ* molecular barcode allows this method to be scaled up to genome-wide proteomic studies.

Introduction

High-throughput screens are playing an increasingly important role in advancing the understanding of biological systems and cell biology[1]. Large-scale screens are greatly facilitated by the ability to pool multiple cell lines of interest together to analyze, instead of having to study each cell's condition independently. The study of the dynamics of endogenous proteins remains an important endeavor to further understand cell signaling pathways, cell cycle progression, differentiation, and more.

Fusing fluorescent or epitope tags to endogenous proteins is a widely used method to study proteins within their natural context[2]. Fluorescent tagging allows for the analysis of protein localization, behavior, and even interactions with other molecules. Tracking cellular proteins *in vivo* with fluorescent tags was enabled by the discovery of GFP and the generation of a suite of fluorescent proteins from this original molecule[3]. Creating a genetic in-frame fusion of GFP to a protein of interest allows the visualization of that protein in time and space in tissues, cells, or

even sub-cellular structures. Their quick folding[4] and fluorescent lifespan[5] makes them a very useful biological tool.

The advent of CRISPR has made modification of the genome, and addition of fluorescent protein tags, even easier[6–8]. However, the traditional method of protein tagging of homology-direct repair (HDR) still involves high amounts of labor and cost for multiplexed tagging efforts. The necessity for fluorescent proteins to be in-frame with their endogenous counterpart requires that CRISPR efforts via HDR are carefully considered and screened. While generic strategies have been attempted via HDR, the creation of indels, which are erroneous insertions or deletions around an entry site, severely curtails its usefulness. Addition of DNA by HDR creates indels at approximately 34% of sites[9]. Any disruption to the exonic portions of genes can disrupt their function and produce unusual results. Individual HDR donors must be constructed for each gene of interest, cloned, and tested.

An alternative to HDR is to capitalize upon a previous approach for tagging endogenous genes through synthetic exons. This method, called “CD-tagging” or “protein trapping”, has used transposons or retrovirus to random tag genes with a synthetic exon[10–12]. This approach has been shown to work in mammalian cells[12], zebrafish[10] and Drosophila[13]. The biggest drawback to this approach has historically been the random nature of its integration into the genome. However, with the use of CRISPR technology this drawback can be overcome. By using CRISPR to integrate the synthetic exon into the intronic regions of a gene, a non-homologous end joining (NHEJ) approach can be used that does not necessitate DNA specific to each gene. A generic template for the synthetic exon tag can be used, which simplifies the tagging and cuts down on cost. The risk of affecting the exonic regions of genes by frame-shift or mutation is greatly diminished. The ability to tag a certain gene of interest also increases, as the choices for sgRNA are greatly widened by the number of introns and their lengths.

In order to truly run large-scale tagging experiments, it is imperative to have an easy and efficient manner of detecting which gene or genes are tagged in each cell. As the number of non-overlapping fluorescent protein emission wavelengths are limited, a limit of one gene per fluorescent protein color can be tagged in each cell. In a large-scale experiment, this would necessitate the isolation and characterization of each type of tagged cell. One method to overcome this issue is to use an RNA barcode. RNA barcodes allow for the determination of a cell's identity without needing to remove or destroy the cell. Barcodes can be read out *in situ*, which vastly increases the application of such a tagging system. Methods such as MERFISH allow for easy readout of *in situ* barcodes[14–16]. These barcodes have previously be demonstrated to be capable of screening 20 million *E. coli* cells[17]. This approach has been attempted by a recent paper using *in situ* sequencing to determine cell identities[18]. A drawback to the ISS approach is the need to image single spots, which is a slow process that requires custom microscopes. By amplifying a barcode many times, lower objectives can be used and cells can be segmented at much lower resolution.

We propose a 24-bit barcode scheme that can be integrated into the genome along with the synthetic exon so as to be able to quickly and easily determine cell identities (Figure 3.1). Each bit of the barcodes is a 20 base pair site that is composed of one of two different sequences, which either represent a value of "1" or "0". Only one of the two sequences, the "1" value, will bind a fluorescently-labeled probe. The other sequence, the "0" value, does not bind to anything at all. To increase the rate of hybridization, the sequences were constructed from a three-letter nucleotide alphabet only containing A, T, and G in order to destabilize potential secondary structures[19]. This design offers 2^24 choices of possible barcodes, or greater than 16.7 million possible barcodes. Having such a large pool of possible barcode options minimizes the probability of two cells having identical or similar barcodes. Only ~21,000 protein-coding

genes are present in the human genome[20], a number more than 500 times smaller than our number of possible barcodes. Therefore, this method is highly error-robust.

Results

Our tagging scheme involves the use of a generic donor plasmid with a Cas9 protein and targeted gRNA to intronically tag endogenous protein-coding genes (Figure 3.1). The generic donor plasmid contains a fluorescent protein, mVenus, surrounded by a splice donor (SD) and a splice acceptor (SA) to create a synthetic exon. The synthetic exon sequence is surrounded by two sgRNA target cut sites comprised of sequences not found in the human genome[21]. The gRNA for these target sites is found on a second plasmid, which contain the Cas9 protein as well. A third plasmid contains the gRNA targeting the protein of interest. This synthetic exon is inserted into the intronic region of the gene via NHEJ. During transcription, the synthetic exon is transcribed, spliced into the mature RNA product, and forms into a protein with an attached fluorescent protein. The use of NHEJ offers many possible insertion sites within intronic regions, which average length is ~3.4 kb and occur approximately 8 times per gene[22]. The only constraint for intronic targeting is to use an intron that allows for the donor plasmid to be in frame in the final mRNA so as not to cause frameshift mutations. We show results from proteins tagged with generic donors in “frame 0” but have also created donors with one or two additional ‘A’s between the splice acceptor and start of the mVenus fluorescent protein for “frame 1” and “frame 2” so that all potential introns can be targeted.

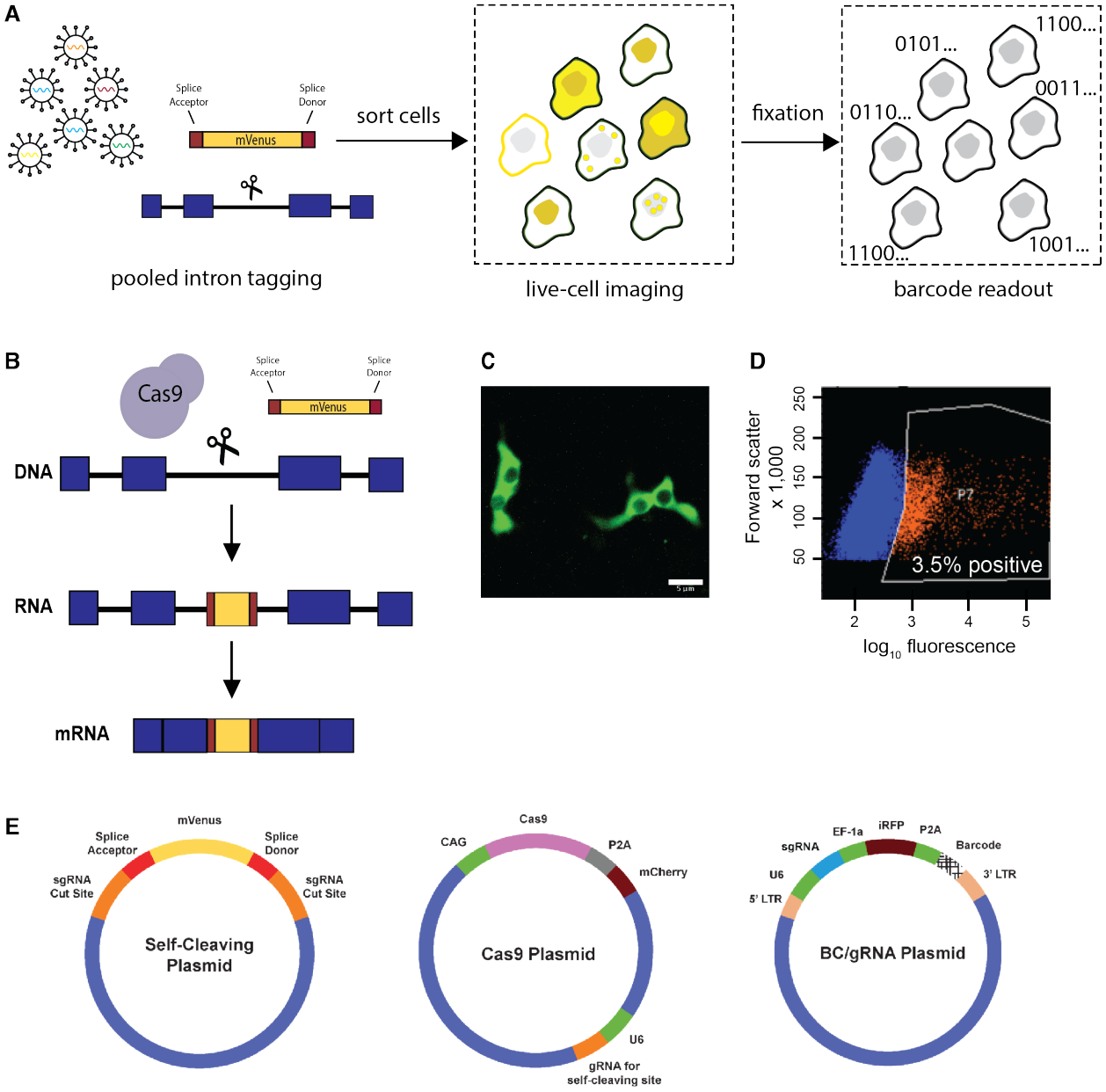


Figure 3.1. Introduction of fluorescent tag into endogenous gene. (A) Overview of scheme to create pooled, intronically tagged cells, then image them using live-cell microscopy, and finally read out their molecular barcode. (B) Using a generic donor DNA that contains a fluorescent protein surrounded by a splice acceptor and splice donor creates a synthetic exon that can be inserted into the intronic region of a gene. Since this way of integration is through NHEJ it does not require the use of lengthy homology arms. Intronic regions also offer many more locations for insertion of synthetic DNA than exons. (C) A549 cells with an intronically tagged ACTB gene at its fourth intron. Scale bar is 5 μm. (D) Flow of A549 cells transfected with the triple plasmid scheme. Untagged cells are represented by blue, tagged cells are represented by orange. (E) Introduction of three plasmids to tag endogenous genes

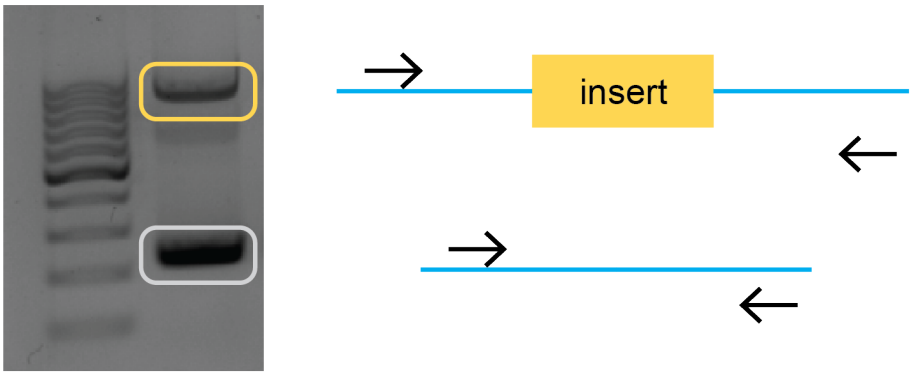
While the fluorescent protein tag can potentially be inserted upside down, the use of flow cytometry to isolate positive fluorescent cells allows for the identification of proteins that have successfully been tagged. Using this approach, we tagged the fourth intron on the actin protein and showed normal localization patterns in A549 cells. We further validated the correct insertion of the mVenus tag through RNA extraction and RT-PCR of the actin RNA. Gel electrophoresis revealed that one allele of the gene was successfully tagged while the other allele remained untagged (Figure 3.2). We see a potential further use for this technique in dual-allele tagging with different colored fluorescent tags for functional analyses and other experiments. Sanger sequencing revealed continuous reads containing the insert at the 5' and 3' junction with the endogenous exonic portion. This result shows that the fluorescent tag can be completely and accurately integrated into the cell's genome and transcriptome.

To allow for this technique to be used in a more high-throughput manner, we explored the use of RNA barcodes for rapid identification of cellular identity (Figure 3.3). The use of an RNA barcode requires that a representation of the population of cells to be analyzed is sequenced before the experiment to determine the sequence of barcode found in each cell along with its tagged gene. Once a table of tagged genes and barcode ID is assembled, the barcodes can be read out *in situ*, allowing for rapid identification of cellular identity as well as the ability to pair live-cell measurements, similar to recent studies[23]. We designed a 24-bit barcode, comprised of sections of 20 nucleotide bits that can either bind a readout probe ("1") or cannot bind a readout probe ("0") (Table 3.1). This barcode has 2^24, or roughly 16 million, possible combinations. Figure 3.3B shows the successful creation of a barcode library which was pooled into lentivirus used to infect cells.

In the convention readout process of the barcode, cells on a coverslip are fixed, permeabilized, and an acrylamide gel loaded with acrydite-modified primers is formed around

the cells. Barcode RNA binds to the primers now embedded in the gel and the rest of the cellular contents are cleared away. The barcode can then be read by MERFISH (Figure 3.3D). In order to increase the speed of imaging by 10x, the barcoded RNA attached to acrydite-modified primers embedded in the acrylamide gel can by reverse transcribed and amplified by PCR (Figure 3.3E). The acrylamide gel limits diffusion of the PCR product, allowing the barcoded RNA to remain in the same area and not diffuse into other cells. The resulting amplified barcode DNA is read out by FISH at a much lower magnification, allowing for greater multiplexing.

A



B

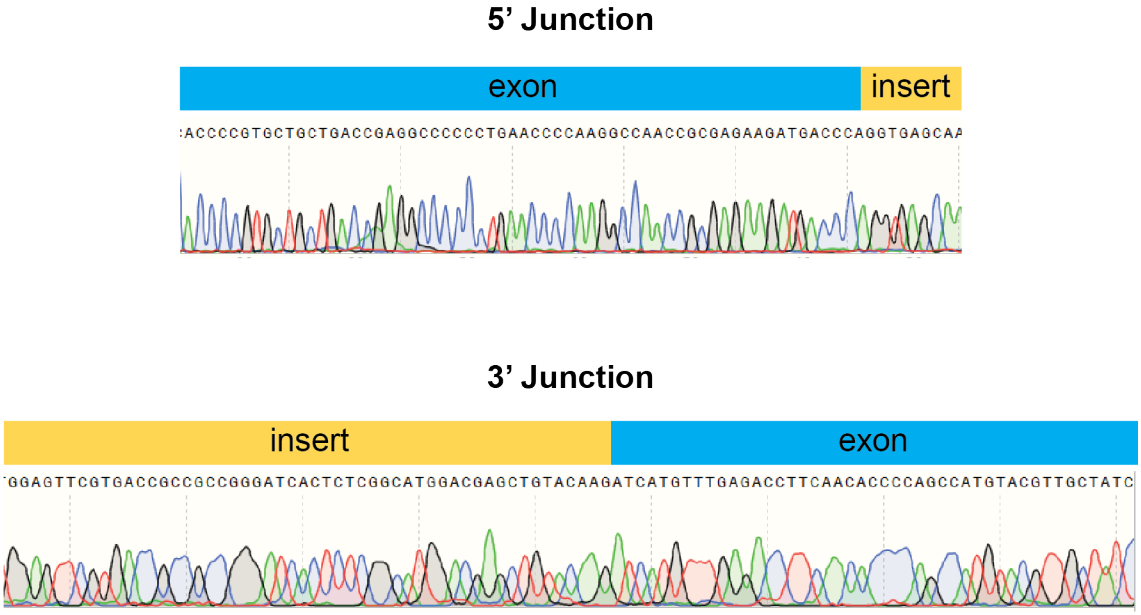


Figure 3.2 Validation of tag insertion (A) PCR result of sequencing of area surrounding the mVenus synthetic exon. Two alleles of the gene are present in this cell line – a tagged allele (top) and untagged allele (bottom) (B) Sanger sequencing results of an RT-PCR amplicon showing continuous reads containing the insert and the exonic region of ACTB. Reads are shown at the 5' junction of the fourth exon and the mVenus insert at the 3' junction of the mVenus insert and the fifth exon.

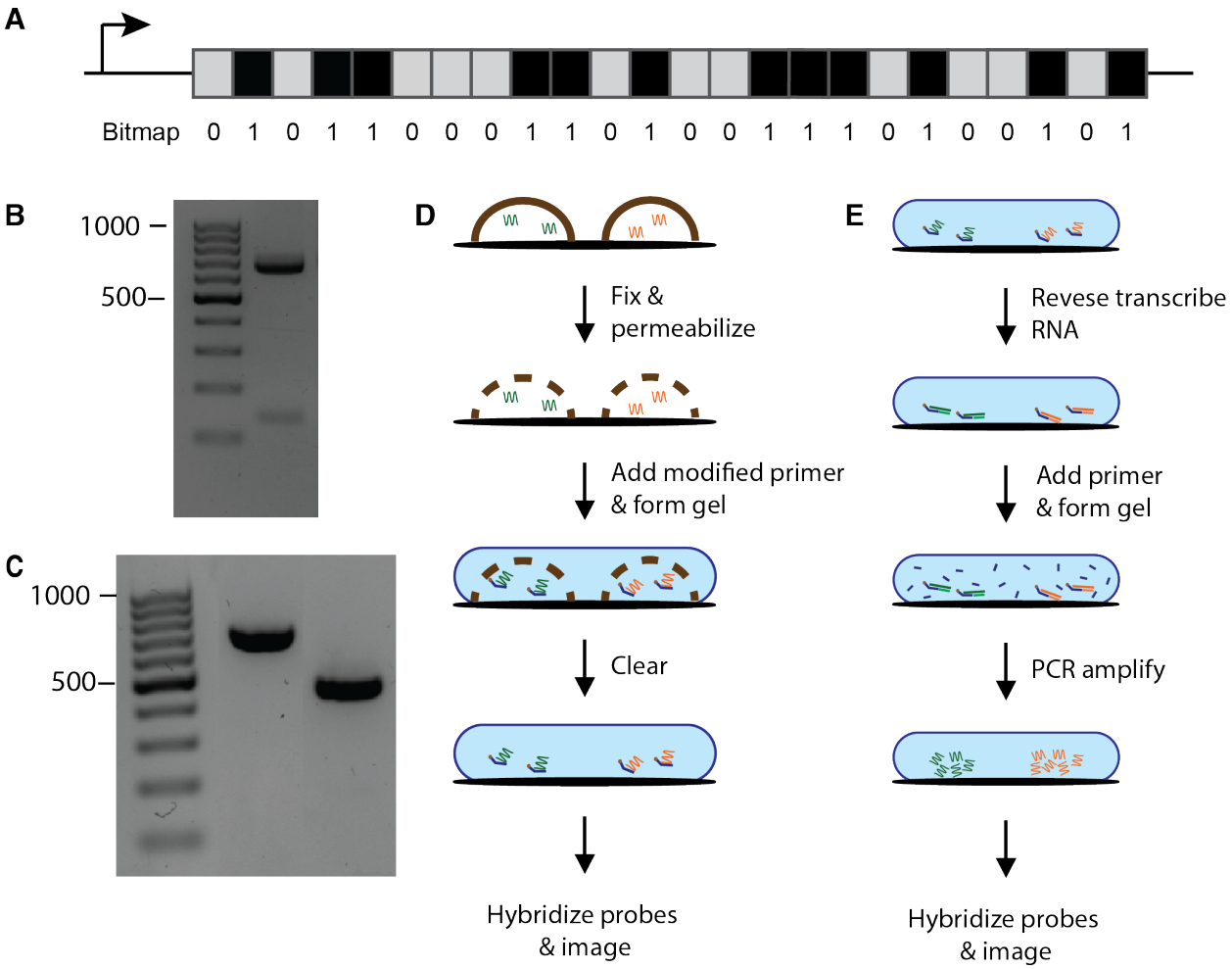


Figure 3.3 RNA barcode scheme (A) A key feature of our approach includes a unique RNA barcode inserted into each cell by a low MOI infection. The barcode can be read out at the end of a live-cell microscopy by FISH and is comprised of 24 bits that can either successfully bind a readout probe or not bind a readout probe (B) Gel image showing the successful creation of a library of 24-bit barcodes (C) Gel image showing the results of a RT-PCR or two different barcoded cell lines. The first lane corresponds to a full-length barcode whereas the second lane represents a truncated barcode (D) Single-molecule readout of the barcodes with clearing (E) Amplification scheme for easier readout of the barcodes. Amplifying the RNA barcode by RT-PCR allows for faster readout using lower magnification microscopy. This will improve speed of barcode analysis by 10X.

Name	Bit 1	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7	Bit 8	Bit 9	Bit 10	Bit 11	Bit 12	Bit 13	Bit 14	Bit 15	Bit 16	Bit 17	Bit 18	Bit 19	Bit 20	Bit 21	Bit 22	Bit 23	Bit 24
Full Barcode	1	1	0	0	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	1	1	1

Table 3.1 Barcode bitmap. Demonstration of the bitmap of a full-length barcode with 13 positive bits and 11 negative bits.

DISCUSSION

In this chapter we have established a scalable protocol for the creation of pooled endogenously tagged library. By demonstrating the usefulness and feasibility of this approach, we hope this information will act as the basis of future studies to address questions of impacts of drugs and drug combinations. Our method improves upon previous methods to fuse fluorescent or epitope flags to proteins to study their functions, localizations, and interactions within living cells. Other methods using plasmids or viral vectors tend to result in overexpression artifacts, lack of endogenous regulatory environments, and scale poorly. The use of a generic DNA donor for large-scale applications makes the use of intronic tagging via CRISPR a highly effective option.

Improvements upon the technique will allow for greater efficiency of the tagging approach. Currently, the random orientation of the tag cuts efficiency in half because half of all fluorophores will be integrated into the DNA upside-down. Methods to ensure the proper integration of the tag will increase efficiency by 2x. The selection of an appropriate intron to integrate a tag into is also crucial to success. Integrating into certain introns could case the proteins to misfold, possibly creating cells that have diffuse fluorescence due to protein mis-localization. A crucial check to the system is comparison to known localization patterns of proteins in wild-type cells. A key metric to ensuring proper localization is comparison to previously published work, such as The Human Protein Atlas which lists the subcellular distributions of proteins encoded by 12813 genes - accounting for 65% of the human protein-

coding genes[24,25]. Future sets of systematic tagging experiments will be needed to understand the parameters by which intronic protein tags are more effective and least disruptive.

Exploring the dynamics of endogenously expressed genes is a useful tool to explore pathways. While previous efforts to explore the dynamics and localization have yielded important insights into the role of non-genetic heterogeneity they have been low-throughput and laborious. By enabling the ability to analyze many hundreds or thousands of genes at a time, we will greatly accelerate the ability to explore cells' reactions to different environments, whether it is in signaling responses or survival patterns in response to a chemotherapy. We believe this technique will be a widely applicable tool that will have a broad range of uses.

We see this method as potentially very useful to address the question of drug resistance in cancer. In particular, this method is highly suited to the study of fractional killing – the phenomenon when tumor cells are exposed to a chemotherapy, some cells are killed while others survive. This fractional killing selects for drug resistance in cancers. Further understanding the processes that lead to drug resistance is crucial in developing new targeted therapies. Recent studies show that heterogenous gene expression plays a role in non-genetic resistance. Cells that stochastically express high levels of certain genes have a higher likelihood of surviving a dose of chemotherapy. Current methods to explore these cell states are time-consuming, costly, and scale poorly. With our method, thousands of cells with hundreds of tagged proteins can be tracked through live-cell imaging to quantify the dynamic changes in expression and localization that leads to differences in drug survivability.

While tagging a single fluorophore here, these methods can be readily adapted to use FRET, complementation, and co-localization analysis with additional organelle markers, this tagging approach readily provides information about variation and heterogeneity in single cells

over time, and thus lineage history. Therefore, this protein tagging strategy will be especially powerful for uncovering latent drug-responsive or -resistant states within single cells. We hope that this method will offer a robust and scalable approach for the study of proteomics at scale.

Acknowledgements

Anna Pilko assisted with barcode library construction. Thanutra Zhang assisted with the CRISPR tagging. Roy Wollman assisted with concept and direction.

Materials and methods

Barcode assembly

The barcode library consists of a set of plasmids, each containing a DNA barcode sequence that encodes and RNA designed to represent a single N-bit binary word. Every barcode in the library has 24 readout sequences, one corresponding to each bit, that are designed to be read out by hybridizing fluorescent probes with the complementary sequence. For each bit position, we assigned one 20-mer sequence to encode a value of “1”. To increase the rate of hybridization, the sequences were constructed from a three-letter nucleotide alphabet only containing A, T, and G in order to destabilize potential secondary structures. Every other bit was separated by a single base pair of “A”. The barcode library was assembled as described in Zhuang et al 2017[17].

Cloning

The mVenus donor tag[26] was amplified out of another plasmid containing the fluorophore. The template was amplified by primers to add the splice donor and acceptor sites. Sequences for the CRISPR cut sites were obtained from Talas et al[21]. The component parts of the plasmid were assembled via Gibson cloning. The sgRNA-expressimg plasmids were generated by

digesting a lentiGuide-Puro plasmid (Addgene 52963) with Esp3I and ligating an annealed sgRNA duplex as described in Ran et al 2013[6].

Cell culture and transfection

Experiments were performed in A549 (ATCC CCL-185), HEK293 (ATCC CRL-1573), and PC9 (ATCC CRL-11350) cells. Each cell line was cultured in DMEM (Thermo Fisher Scientific) +10% fetal bovine serum (FBS; VWR)+ pen-strep (Thermo Fisher Scientific). Cells were transfected at 80% confluency with lipofectamine 3000 with a ratio of 2:1:1 of mVenus generic donor plasmid to Cas9 plasmid to gene-specific sgRNA plasmid.

Flow cytometry and cell sorting

Sorting was accomplished with a BD FACSAria cell sorter. Tracking of expression after sorting was accomplished with a BD LSRII. Cells were initially filtered using forward scatter and side scatter. mVenus signal was measured with FITC-A. TagBFP signal was measured with PacBlue-A. Data were analyzed using custom MATLAB scripts (https://github.com/wollmanlab).

RT-PCR analysis of genomic regions

Roughly 1 million cells were harvested for RNA extraction. SuperScript IV kit was used for the RT-PCR with primers ‘AGGGCTAATTCACTCCCAACG’ and ‘GTTCAGACGTGTGCTCTTCC’. The amplicons were imaged alongside a 1 kb bp DNA ladder (New England Biolabs) and extracted from a 1% agarose gel using the Monarch Gel Extraction kit (New England Biolabs) and analyzed by Sanger sequencing (GENEWIZ) using the primers ‘AGAGAGGCATCCTCACCCTG’ and ‘GATAGCACAGCCTGGATAGCA’.

References

1. Pegoraro, G. & Misteli, T. High-Throughput Imaging for the Discovery of Cellular Mechanisms of Disease. *Trends Genet.* **33**, 604–615 (2017).
2. Crivat, G. & Taraska, J. W. Imaging proteins inside cells with fluorescent tags. *Trends Biotechnol.* **30**, 8–16 (2012).
3. Day, R. N. & Davidson, M. W. The fluorescent protein palette: tools for cellular imaging. *Chem. Soc. Rev.* **38**, 2887–2921 (2009).
4. Balleza, E., Kim, J. M. & Cluzel, P. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat. Methods* **15**, 47–51 (2017).
5. Pliss, A., Zhao, L., Ohulchanskyy, T. Y., Qu, J. & Prasad, P. N. Fluorescence lifetime of fluorescent proteins as an intracellular environment probe sensing the cell cycle progression. *ACS Chem. Biol.* **7**, 1385–1392 (2012).
6. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
7. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
8. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
9. Chakrabarti, A. M. *et al.* Target-Specific Precision of CRISPR-Mediated Genome Editing. *Mol. Cell* **73**, 699-713.e6 (2019).
10. Trinh, L. A. *et al.* A versatile gene trap to visualize and interrogate the function of the vertebrate proteome. *Genes Dev.* **25**, 2306–2320 (2011).
11. Jarvik, J. W., Adler, S. A., Telmer, C. A., Subramaniam, V. & Lopez, A. J. CD-tagging: a new approach to gene and protein discovery and analysis. *Biotechniques* **20**, 896–904 (1996).

12. Jarvik, J. W. *et al.* In vivo functional proteomics: mammalian genome annotation using CD-tagging. *Biotechniques* **33**, 852–4, 856, 858–60 passim (2002).
13. Clyne, P. J., Brotman, J. S., Sweeney, S. T. & Davis, G. Green fluorescent protein tagging Drosophila proteins at their native genomic loci with small P elements. *Genetics* **165**, 1433–1441 (2003).
14. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
15. Moffitt, J. R. *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence *in situ* hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11046–11051 (2016).
16. Moffitt, J. R. *et al.* High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14456–14461 (2016).
17. Emanuel, G., Moffitt, J. R. & Zhuang, X. High-throughput, image-based screening of pooled genetic-variant libraries. *Nat. Methods* **14**, 1159–1162 (2017).
18. Reicher, A., Koren, A. & Kubicek, S. Pooled protein tagging, cellular imaging, and *in situ* sequencing for monitoring drug action in real time. *Genome Res.* (2020) doi:10.1101/gr.261503.120.
19. Zhang, Z., Revyakin, A., Grimm, J. B., Lavis, L. D. & Tjian, R. Single-molecule tracking of the transcription cycle by sub-second RNA detection. *Elife* **3**, e01775 (2014).
20. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* 332825 (2018) doi:10.1101/332825.
21. Tálas, A. *et al.* A convenient method to pre-screen candidate guide RNAs for CRISPR/Cas9 gene editing by NHEJ-mediated integration of a “self-cleaving” GFP-expression plasmid. *DNA Res.* **24**, 609–621 (2017).

22. Hnilicová, J. & Staněk, D. Where splicing joins chromatin. *Nucleus* **2**, 182–188 (2011).
23. Foreman, R. & Wollman, R. Mammalian gene expression variability is explained by underlying cell state. *Mol. Syst. Biol.* **16**, e9146 (2020).
24. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
25. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).
26. Kremers, G.-J., Goedhart, J., van Munster, E. B. & Gadella, T. W. J., Jr. Cyan and yellow super fluorescent proteins with improved brightness, protein folding, and FRET Förster radius. *Biochemistry* **45**, 6570–6580 (2006).

CHAPTER 4

Determining the Underlying Mechanisms of Gene Coexpression in the Human Genome

Abstract

Obtaining a quantitative understanding of how cells regulate gene expression is a central goal of biology. Proper regulation of gene expression is vital to cell development and survival. However, there is a gap in understanding how genes are expressed differentially across the genome. The enormous complexity of the gene regulatory network inhibits the quantitative understanding of gene regulation. Regulation can occur at transcription, RNA processing, and translation; however, the primary point of regulation is transcription. Traditionally, transcriptional regulation has been split into two types: *trans* and *cis*. *Trans* regulation is driven by soluble factors, called transcription factors (TFs), that bind to elements in a gene's local chromatin environment and cause activation or repression of proximate genes. *Cis* regulation refers to DNA elements in a gene's environment that can interact with *trans* factors to affect nearby genes. Previous research has focused efforts on understanding *cis* elements adjacent to the gene of interest, such as promoters. In this chapter, we focus on understanding *cis* effects outside of an open reading frame and their consequences on gene expression.

Introduction

Molecular processes in cells deal with very small numbers of molecules, leading to large fluctuations in output[1]. Gene expression is one of these molecular processes that exhibits large amounts of variability[2]. Fluctuations in gene expression, also known as gene expression noise, are prevalent across multitudes of organisms, ranging from bacteria and yeast to mammalian cells[3,4]. Even within isogenic populations of cells grown under the same conditions, phenotypes can vary significantly[5].

The influence of genomic position on gene expression has been known since the 1930s and is termed the 'position effect'[6]. This effect has important implications in synthetic biology and the study of genetic diseases[7]. The same open reading frame (ORF) placed synthetically in different areas of the human genome will lead to wide differences in expression outcomes. Work by Akhtar et al. has expanded on the position effect by showing a 1,000-fold range of transgene expression in mRNA across 27,000 integration sites in mouse cells[8]. A subsequent report in yeast showed a 20-fold range in average noise of GFP protein expression[9]. However, these studies are single-reporter systems that do not have sufficient controls to account for the effects of cell-wide global factors on gene expression and genomic noise. Both studies only tested histone modifications as possible causative agents of the position effect and have not gone further to propose mechanisms by which the different environments of the transgenes influence their expression.

One of the main causative agents of gene expression noise is transcriptional regulation[5]. A basic model of gene expression is that regulatory proteins called transcription factors (TFs) can act in *trans* to inhibit or promote expression at certain genomic locations by binding to DNA-specific sequences in *cis*-regulatory modules (CRMs). TFs are presumed to be constant across a cell's volume while CRMs change based on location in the chromatin (Figure 4.1). Most previous studies of genes that show covariance across a genome attribute this phenotype to co-regulation by *trans* soluble factors. The effects of cis controllers of expression have focused mainly on promoter architecture as a controller of transcriptional noise[10–12]. There has been a lack of studies exploring how the environment outside of a promoter will affect transcription. Ebisuya et al. attempted to address this question by demonstrating that intensive transcription at one gene can have a "ripple effect" that increases transcription at neighboring genes[13]. However, a more rigorous approach is necessary to understand how the local environment outside an open reading frame affects transcription.

Dual-reporter systems have also been used to study stochastic gene expression[2,14–16]. Raj et al. performed a dual-reporter assay that integrated two reporter genes in one cell at the same genomic location and two reporter cells in another cell at random, distant genomic locations[17]. This experiment showed a high correlation (0.89) for the gene expression at the same location and a very low correlation (0.056) for the genes at distant genomic locations. This result is consistent with a hypothesis of *cis* regulation, as two neighboring loci share the same chromatin environment. However, by not repeating the integrations of distant reporters, the experimenters missed out on sampling distant environments that are under the same kinds of regulatory control. Genes in these environments would also show high levels of co-expression, similar to the two neighboring loci.

The main drawback of current single- and dual-reporter studies is the lack of proper control for global factors. Global factors, such as cell size, cell cycle stage, and number of ribosomes, affect many genes in the cell simultaneously and induce correlated fluctuations between genes[18,19]. In order to separate correlations that are due to similar local chromatin environments and correlations that are due to global factors, a third type of reporter must be introduced. In this chapter, we introduce a triple-reporter system to study the *cis* effects of transcriptional regulation across the chromatin environment in mammalian cells.

Results

Generation of a new triple-reporter assay

We propose a novel approach to study *cis* chromatin regulation through the use of a triple-report system. The key innovation in this approach to improve upon dual-reporter systems is the introduction of a third reporter that has multiple integrations across the genome. This third reporter is under the same promoter as the singly-integrated reporters to control for the

promoter effect. A multiply-integrated reporter will occur in many different local environments across the genome, so the total protein expression will be an average of these conditions. With enough integrations, the third reporter expression will be contingent on only global factors and no local factors. By conditioning the single reporter expression on the multiply-integrated reporter's expression, any correlation between the single reporters that resulted from global factors will be removed. Any remaining correlation will be from local effects.

The human leukemia cell line K562 was chosen for this work because it is well-characterized and widely used[20–23]. K562 is one of three tier-one cell lines of ENCODE and is also most commonly used for large-scale CRISPR/Cas9 screens. We have genetically engineered K562 cells to contain three different types of fluorescent reporter genes: a single integration of a green fluorescent protein gene, a single integration of a red fluorescent protein gene, and multiple integrations of a blue fluorescent protein gene (Figure 4.1C). Lentiviruses were used to randomly insert each gene into the K562 genome. The green and red genes were transduced at a MOI of less than 10,000 viruses per million cells in order to ensure single integrations. The blue genes were integrated at a very high multiplicity of infection to obtain a high number of integrations. All three reporters were inserted with the same open reading frame under the ubiquitin C (UBC) promoter. The UBC promoter was chosen since it has, as the name suggests, ubiquitous and constituent expression in many tissue types and has been shown to be difficult to silence[24–26].

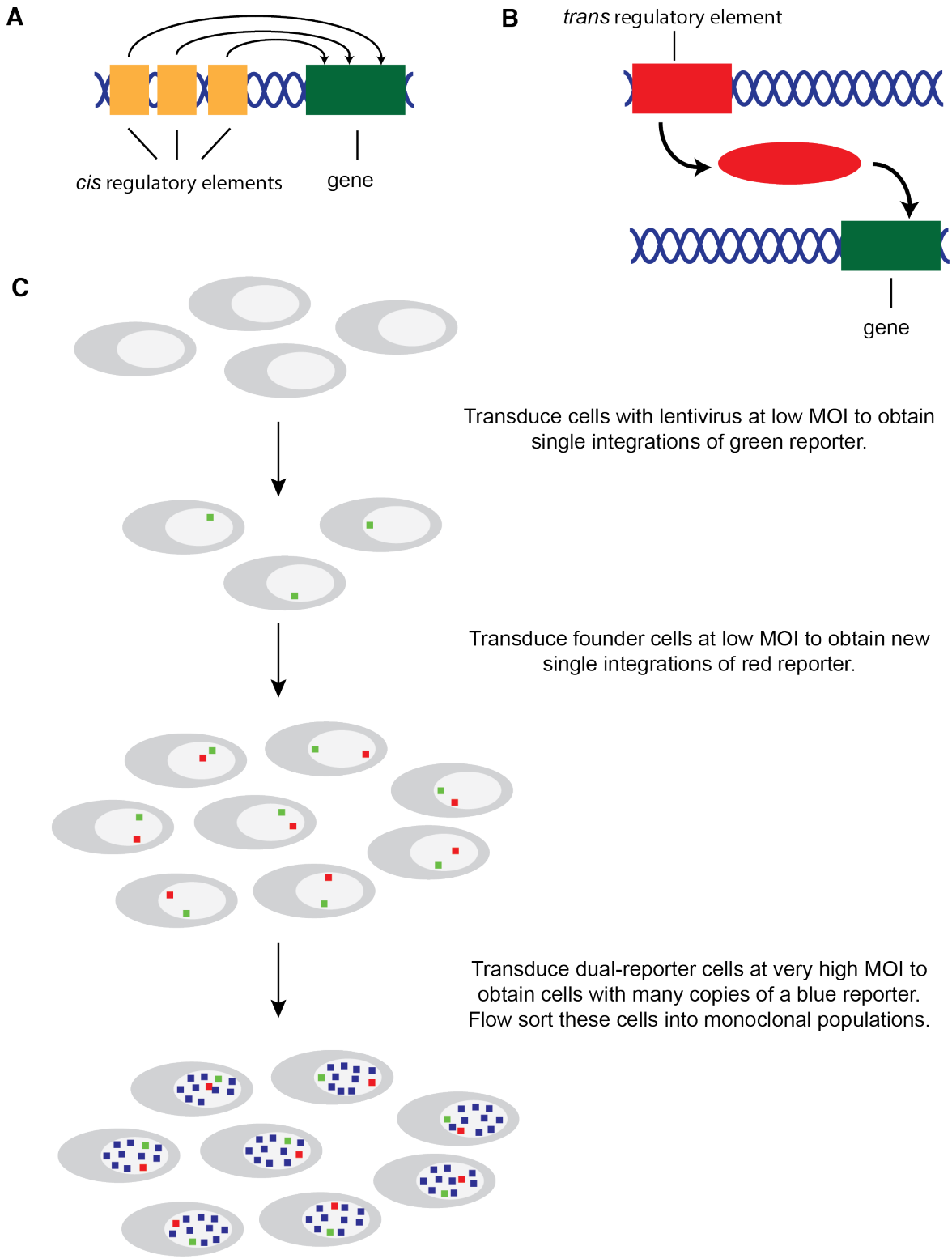


Figure 4.1 Scheme of Triple Reporter Assay (**A**) *Cis* regulatory elements are local elements that regulate gene expression. (**B**) Trans regulatory elements can be distal elements that regulate a gene's expression. **(C)** The process to create a triple-reporter system in mammalian cells via sequential addition of reporters via lentivirus transduction.

Integrations of each gene were done by lentiviral infection, specifically using a modified HIV-1 virus. Lentiviruses were chosen over other gene delivery methods because they infect many cell types, can infect both dividing and non-dividing cells, have a large carrying capacity, and can stably insert their cargo into the host genome[27,28]. Lentiviruses have slight preferences for integrating their cargo into gene-rich areas[28–30]. Although there are some concerns that using lentivirus might change local chromatin structure around the insertion site, Chen et al. found that replacing native yeast genes with a synthetic reporter did not perturb the chromatin landscape[31].

The first challenge to this project was to determine if regions of co-regulation could be found with only a few hundred randomly integrated reporters. Encouragingly, correlations of up to 0.65 were found with only 52 cell lines. After conditioning green and red expression on blue expression, correlations ranging between 0.06 and 0.65 were calculated with an average correlation of 0.31. Figure 4.2 shows how the conditional correlation is calculated using a weighted average of subsamples of cells that share the same blue expression. Using only a small number of integration sites revealed a range of levels of co-regulation.

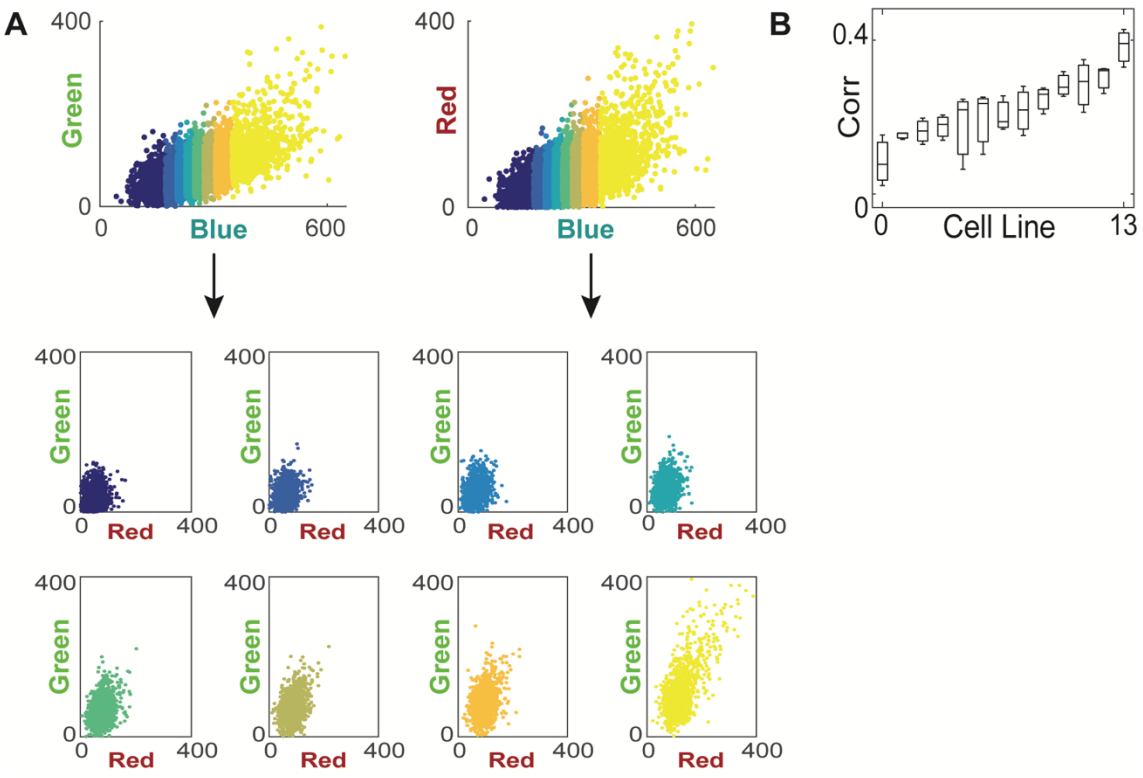


Figure 4.2 Conditional correlation analysis. A) Flow cytometry results outlining the conditional correlation analysis between Green and Red conditioned on Blue. For simplicity, the figure shows 8 discrete bins. The analysis itself is done using a Gaussian moving window B) The measured conditional correlation values calculated based on 13 cell lines.

Next, we sought to identify the sources of variation in green expression across each cell line. We were inspired by the work of Elowitz et al[2], which suggests that expression variance can be decomposed into uncorrelated and correlated variance. Our triple-reporter assay assumes that the correlated component of gene expression with the multiply-integrated blue reporter captures all the regulatory mechanisms that act in *trans* and the uncorrelated component captures *cis* regulation (Figure 4.2). The sources of variation in green expression across each cell line should be able to be decomposed into three types of variation: variance in yellow that is explained by cyan (global variance), variance in yellow that is explained by red and not by cyan (local variance), and variance in yellow that is unexplained by either cyan or red

(unexplained variance). The variance decomposition was based on a modified function of the law of total variance. The law of total variance is as follows:

(Eq. 4.1) $$Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$$

Where E[Var(Y|X)] is the amount of variance in Y that is unexplained by the variance in X and Var(E[Y|X]) is the amount of variance in Y that can be explained by the variance in X. Therefore, to modify this law for a triple-reporter system we can modify the law of total variance to be applicable to our system. Our modified function is as follows:

(Eq. 4.2) $$Var(G) = Var\big(E(G|B)\big) + E(Var\big(E(G|B,R)\big) + E(E\big(Var(G|B,R)\big))$$

In this function, Var(E(G|B)) represents the global variance, which is expressed as the variance of the green expression conditioned on the variance of the blue expression. This can also be thought of as the variance of green expression that can be explained by the blue expression. E(Var(E(G|B,R))) represents the variance in green expression that can be explained by the variance in red expression. This component shows the effects of local chromatin structure on variance. The last variable E(E(Var(G|B,R))) is the variance that is unexplained by either reporter gene. This variance may result from random fluctuations within the cell as well as local effects that are not captured by the red reporter alone.

If all correlated variance acts in *trans* on all three colors, then the component of the variance in green that is explained by red and not blue, i.e. the local variance, should be zero. Any meaningful explanatory power of red reporter after the blue reporter was included indicates that green and red are co-regulated in a way that blue is not. Given that blue, green, and red all

have identical open reading frames it will indicate that two local regulatory mechanisms are dependent on each other.

We generated Gillespie simulations to test our method of conditioning gene expression on a multiply-integrated reporter (Figure 4.3). A Gillespie algorithm generates a trajectory of a stochastic equation[32]. In this case, the Gillespie simulation was used to model the steady-state expression of each reporter gene with changing noise components in the propensity of a gene to be transcribed. The simulation conditions were global noise, biochemical noise, and local noise. Global noise refers to global factors that affect protein expression in the same way for each gene within a cell. Biochemical noise refers to random factors that cause stochasticity in a cell, such as Brownian motion of polymerases. Local effects refer to CRMs that may be similar in the environment of the green and red genes. We showed that, as expected, conditioning on a multiply-integrated reporter decreases the correlation between two genes affected by only biochemical and global noise from 0.64 to 0.01 (Table 4.1). The previous correlation resulted from global factors, but the correlation disappeared when the global factors were removed by conditioning expression on blue expression. The more physiological condition of having biochemical, global, and local noise decreased correlation from 0.78 to 0.65, which matches nicely to our observed maximum correlation in the triple reporter cells of 0.65. These simulation experiments convincingly show that as predicted, conditioning on a multiply integrated blue reporter can remove sources of global noise to allow local effects to become easier to interpret.

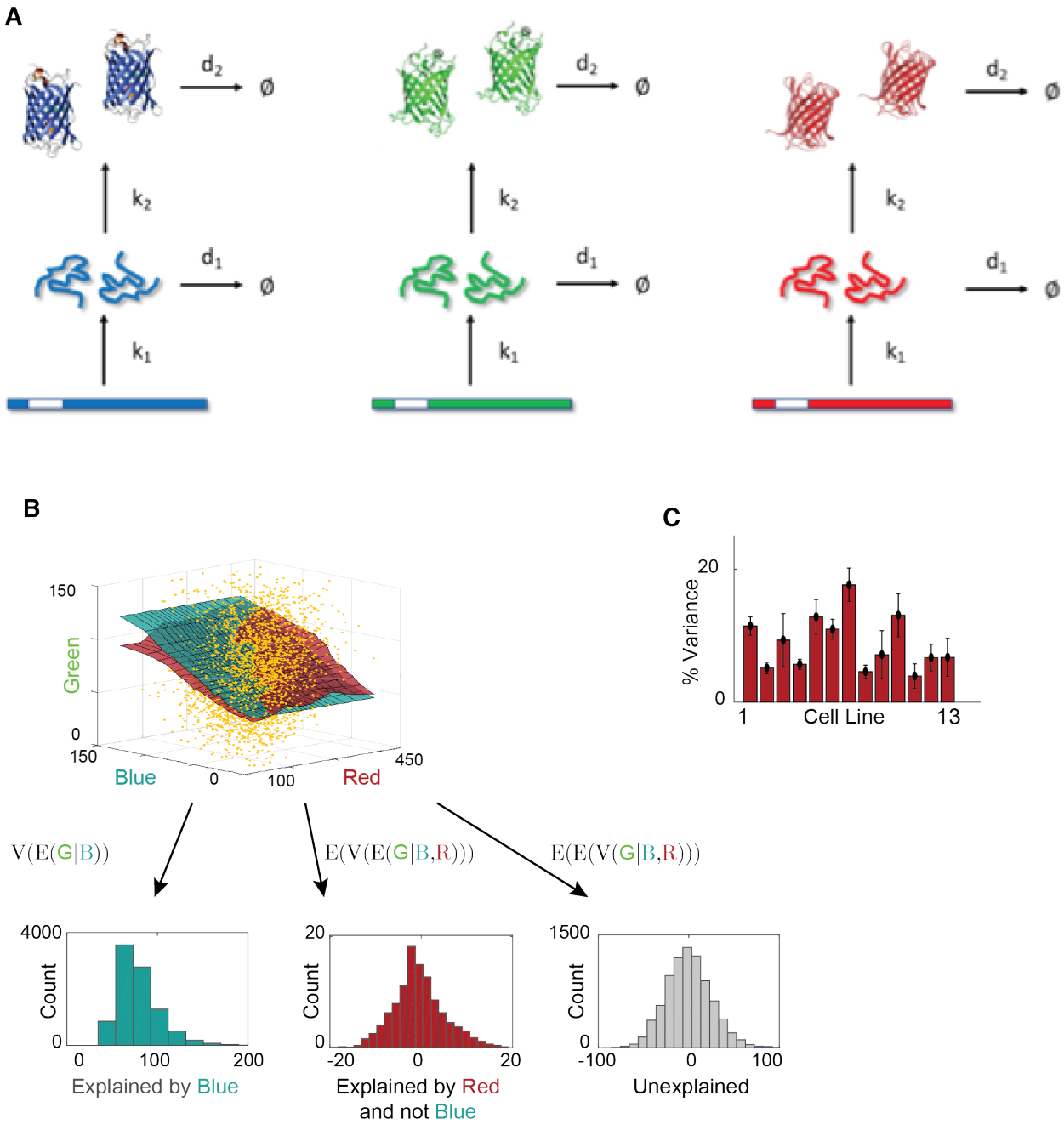


Figure 4.3. Three-way variance decomposition (**A)** Scatter plot of expression of the three fluorescent reporters Red, Green, and Blue. The surface shows the expected value of Green based on Blue alone (cyan) and combined with Red (red). The deviation between these surfaces shows that the addition of Red increases the explained component of the overall variance in Green. This suggests that Red and Green share a co-regulatory mechanism that not shared by Blue. Bottom panels show the decomposition into three components based on successive application of the law of total variance (**B**) The amount of total variance explained by Red and not by Blue in 13 cells lines. Error bars show the standard deviation of duplicate measurement per cell line. The cell lines are significantly different from each other (one-way ANOVA $p<0.05$).

	Correlation coefficient p(V,S) with no conditioning	Correlation coefficient p(V,S) with conditioning on Turquoise
Biochemical, global, and local noise	0.78	0.65
Biochemical and global noise	0.64	0.01
Biochemical and local noise	0.66	0.66

Table 4.1 Gillespie model decomposition of the correlation coefficient. Outcomes of the Gillespie model simulated decomposition of correlation coefficient with and without conditioning on a third reporter.

Lastly, we sought to determine the genomic locations of the green reporters in our assay. Using the Genome Walker method, we extracted DNA from cell lines of interest, used restriction enzymes to cut the surrounding DNA, ligated primers to amplify the region of interest, and used a primer internal to the mVenus sequence to Sanger sequence the surrounding genomic area. We found that our three starting green reporter positions were located in three separate chromosomes (Table 4.2). With this method, the location of all green and red reporter genes could be found and their environments assayed. Scaling future work would benefit in using higher-throughput methods such as the recent work in Zhang et al[33] to more quickly determine many genomic insertion sites in parallel.

Cell Line	Genomic Location	Genomic Sequence
G1	Chromosome 22, 15963910	ATTTTTTTTTTCTTTTCTTTTCTTTTTTTGACAGGTCTTGTTCTGTCATCTAGGCTGGA ATGCAATGGTGTGATCCTAGCTTACTGCGGCCTTGAACCCCTGGGTTCAAGCAGTT CTTTAGCCTCAGCCTCTGACTANGCTAGGATTACAGATACATGCTACTATGCCTGG
G2	Chromosome 1, 25937	TCTGAGAATTCACAACTCAGGCCAGTGTGAGCTCATGCTTGGCAGGGAATTATCTGT AGAATATCTTTGTTGTTCTTCCAGAGATACACTATATTTGATTTTGCTAGCAAAGATTT CCCACAAAGGCCTGCCAAACATAAGCTCACAATTGTGAACACATCAAGAACCAATTAA CCGTGAGCAAACGCAATGCCAATCTAACACAGTTGGATTAGACACTCAGGAACTATA GGCATAGCAATATGAATATTTACCA
G3	Chromosome 12, 129754	GCACACCAAATAATTTGTGGCCTTTTAAAATCTGCCAAAGGTTTAAATGGTGTATATA ATTAATGGCATTCTTATTTTTAAAATCCTGAGAATACCAGCCCGGGCCGTCGACCAC

Table 4.2 Genomic location of the green reporter gene in three cell lines

Discussion and Future Directions

Previous views of gene regulation do not fully consider the *cis* effects outside of a gene's promoter. Here, we present evidence that the local environment of a gene plays a larger role than previously imagined in determining gene expression. In order to show the influence of local chromatin environments, we explored the co-regulation of two distant genes and showed high degrees of correlation and high degrees of variance across genomic positions. Decomposing the variance across positions showed effects of biochemical, global, and local noise. We demonstrate that a triple reporter system is effective at removing global sources of influence on positional correlations. The development of this triple reporter system has implications on position effect and gene expression noise.

Further work in this area could include the incorporation of publicly available data on K562 chromatin environments. For example, The ENCODE database contains information on over 3000 ChIP-seq experiments done so far in K562 cells[22]. The Roadmaps Epigenetics has ChIP-seq data for eleven histone modifications, H2A.Z occurrences, and DNase sensitivity[20]. Using these publicly available data, one could analyze the levels of each possible regulator at the report locations in each cell line within a few kilobases. This data could then create a similarity score between the levels of the regulators at each location. The similarity score could then assess which regulators or combinations of regulators correspond with the observed correlations in gene expression between two locations. We anticipate that no single regulator alone is responsible for the totality of gene expression correlations, but combinations of regulators are more likely culprits. For example, H3K4me has been shown to be predictive of transcription factor binding[34] and H3K4me often co-occurs with H3K79me3 and H3K27ac[35]. Dogan et al. proposed that a handful of main transcription factors and histone modifications are the main predictors of enhancer activity[35].

A better understanding of what drives gene expression correlation in the genome will be the use of perturbation assays. These perturbations would test the effects of removing regulators from the system. By doing a multivariable analysis of multiple gene knockouts at a time with histone perturbations, one could test which factors in combination have the greatest effect on gene regulation. Testing the effect of transcription factors will be more straightforward, as gene knockouts can be performed using CRISPR/Cas9 to specifically remove any TF of interest from the cell lines. Only one or two cell lines showing high correlation will be necessary to test for perturbation assays and gene knockout studies.

The most rigorous test of our understanding of regulatory control of genes across the genome would be the accurate prediction of sites within a genome that should be co-regulated. Inserting reporters into these areas and measuring the outcome in terms of gene expression correlation would demonstrate mastery of understanding the mechanistic underpinnings of gene expression correlation. We hope that this work will further the aim to be able to fully understand the regulatory underpinnings of gene expression control.

Acknowledgments

Thank you to Anna Pilko for her help with cell line construction, Alok Maity for his efforts on modeling work and Roy Wollman for assistance with conception and direction.

Materials and methods

K562 cell culture

A K562 suspension cell line provided by Sigma-Aldrich was grown at 37 °C in RPMI 1640 medium (Gibco) supplemented with 10% FBS (Gibco), 1% penicillin-streptomycin (Gibco) and 1% GlutaMAX (100x) (Gibco) under 5% CO2 atmosphere.

Reporter plasmids

The following elements were included in the base plasmid in order to allow for viral packaging and integration. HIV-1 truncated 5' LTR, HIV-1 packaging signal, HIV-1 Rev response element (RRE), HIV-1 truncated 3' LTR and Central polypurine tract (cPPT). The Ubiquitin promoter (ubi) drove expression of mVenus, mScarlet, and mTurquoise. Lentiviral constructs were constructed through Gibson assembly.

Lentiviral production and cell line generation

Reporter and third generation lentiviral packaging plasmids were transfected into HEK 293T cells to generate reporter vectors. Transfected HEK293T culture supernatant was collected and concentrated by Lenti-X-concentrator (Takara) 48 hours post transfection. K562 cells were transduced with reporter-containing lentivirus in media supplemented with 5 µg/ml polybrene and 20mM HEPES for 2 hours of spinoculation and left to incubate for 24 hours. To generate singly-integrated cell lines, a MOI of 0.01 was used to ensure that the majority of transduced cells integrated with a single reporter copy. Founder cells were then singly sorted by fluorescence-activated cell sorting (FACS) at 72 hours post-transduction to generate unique cell lines.

Gillespie stochastic simulation

Stochastic simulations of the stochastic mRNA and protein model described in Table 4.1 were performed by implementing Gillespie's Direct method[32] in Matlab (The Mathworks).

Fluorescence activated cell sorting and flow cytometry

Sorting was accomplished with a BD FACSAria cell sorter. Tracking of expression after sorting was accomplished with a BD LSRII. Cells were initially filtered using forward scatter (FSC-A)

and side scatter (SSC-A). EGFP signal was measured with FITC-A or GFP-A. TagBFP signal was measured with DAPI-A or Pac-Blue-A. tagRFP signal was measured with PI-A or RFP-A.

Identification of genomic integration sites

Mapping of reporter integration sites was done by nested PCR coupled with Sanger sequencing using the GenomeWalker kit from Takara (638901). Amplicons were sequenced by Sanger Sequencing using primer GCTCCTCTGGTTTCCCTTTCGCTTTCAA. Results were mapped against the human genome using the BLAST algorithm.

References

1. Spudich, J. L. & Koshland, D. E., Jr. Non-genetic individuality: chance in the single cell. *Nature* **262**, 467–471 (1976).
2. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
3. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
4. Balázsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910–925 (2011).
5. Sanchez, A., Choubey, S. & Kondev, J. Regulation of noise in gene expression. *Annu. Rev. Biophys.* **42**, 469–491 (2013).
6. Muller, H. J. Types of visible variations induced by X-rays inDrosophila. *J. Genet.* **22**, 299–334 (1930).
7. Milot, E., Fraser, P. & Grosveld, F. Position effects and genetic disease. *Trends Genet.* **12**, 123–126 (1996).

8. Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914–927 (2013).
9. Chen, X. & Zhang, J. The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. *Cell Syst* **2**, 347–354 (2016).
10. Carey, L. B., van Dijk, D., Sloot, P. M. A., Kaandorp, J. A. & Segal, E. Promoter sequence determines the relationship between expression level and noise. *PLoS Biol.* **11**, e1001528 (2013).
11. Sharon, E. *et al.* Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res.* **24**, 1698–1706 (2014).
12. Zhang, J. & Zhou, T. Promoter-mediated transcriptional dynamics. *Biophys. J.* **106**, 479–488 (2014).
13. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* **10**, 1106–1113 (2008).
14. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
15. Becskei, A., Kaufmann, B. B. & van Oudenaarden, A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.* **37**, 937–944 (2005).
16. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12795–12800 (2002).
17. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
18. Zopf, C. J., Quinn, K., Zeidman, J. & Maheshri, N. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Comput. Biol.* **9**, e1003161 (2013).
19. Edri, S. & Tuller, T. Quantifying the effect of ribosomal density on mRNA stability. *PLoS One* **9**, e102308 (2014).

20. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
21. Koeffler, H. P. & Golde, D. W. Human myeloid leukemia cell lines: a review. *Blood* **56**, 344–350 (1980).
22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
23. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
24. Qin, J. Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* **5**, e10611 (2010).
25. Wiborg, O. *et al.* The human ubiquitin multigene family: some genes contain multiple directly repeated ubiquitin coding sequences. *EMBO J.* **4**, 755–759 (1985).
26. Board, P. G., Coggan, M., Baker, R. T., Vuust, J. & Webb, G. C. Localization of the human UBC polyubiquitin gene to chromosome band 12q24.3. *Genomics* **12**, 639–642 (1992).
27. Durand, S. & Cimarelli, A. The inside out of lentiviral vectors. *Viruses* **3**, 132–159 (2011).
28. Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E. & Engelman, A. Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res.* **42**, 10209–10225 (2014).
29. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, E234 (2004).
30. Schröder, A. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
31. Chen, M., Licon, K., Otsuka, R., Pillus, L. & Ideker, T. Decoupling epigenetic and genetic effects through systematic analysis of gene position. *Cell Rep.* **3**, 128–137 (2013).
32. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).

33. Zhang, T., Foreman, R. & Wollman, R. Identifying chromatin features that regulate gene expression distribution. *Sci. Rep.* **10**, 20566 (2020).
34. Rye, M., Sætrom, P., Håndstad, T. & Drabløs, F. Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. *BMC Biol.* **9**, 80 (2011).
35. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**, 16 (2015).

Conclusion

The studies in this dissertation have offered a glimpse into the complexity of intracellular information transfer. I have broadly discussed the methods by which cells communicate internally and have revealed how much more there is to learn across the fields of biology. In Chapter One, I review the current offering of *in situ* technologies and offer a wide breadth of unanswered questions that these technologies could be key in answering. Chapter Two demonstrated the efficacy of these technologies by discussing how they can be used in conjunction with other techniques, such as live-cell protein imaging, to create multi-omics datasets. This chapter also demonstrated how cells take complex data, such as the incoming signals from two different pathways, to learn more about their extracellular environments. In Chapter Three, I present a method by which many more proteins can be studied by live-cell imaging by intronic CRISPR tagging. This tag coupled with a RNA barcode allows for the high-throughput investigation of many proteins in multiplex. This method allows for the study of intracellular communication through proteins at a very large scale. Finally, in Chapter Four, I investigate the communication of a cell within its own chromatin through *cis* regulatory mechanisms. I show that cells have correlated expression across different regions of the genome, suggesting a greater role of *cis* regulatory mechanisms than previously thought. Altogether, this thesis explores central idea of cellular communication.