UNIVERSITY OF CALIFORNIA SAN DIEGO

Novel Multiomics Method of RNA and Nuclear Protein Characterization in Single Cells

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Daniel Eric Jacobsen

Committee in Charge:

Professor Kun Zhang, Chair
Professor Prashant Mali
Professor Bing Ren
Professor Gene Yeo
Professor Sheng Zhong

2019

The Dissertation of Daniel Eric Jacobsen is approved, and is acceptable in quality and form for publication on microfilm and electronically.

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

B&W-T: Binding and Wash-Tween 20

cDNA: complementary DNA, synthesized DNA from RNA template

DAPI: 4′,6-diamidino-2-phenylindole

DBCO: Dibenzocyclooctyne

DNA: Deoxyribonucleic Acid

dNTP: deoxyribonucleic acid mix of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T)

dsDNA: double stranded deoxyribonucleic acid

DTT: Dithiothreitol

EDTA: Ethylenediaminetetraacetic acid

EGTA: ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid

Fab: Antigen binding fragment

FACS: Flow Assisted Cell Sorting

Fc: Crystallizable fragment

FITC: Fluorescein isothiocyanate

H_2O: Water (dihydrogen monoxide)

HCl: Hydrochloric Acid

HEPES: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HF: High Fidelity

IDT: Integrated DNA Technologies, Coralville IA

Ig: Immunoglobulin

IgG: Immunoglobulin gamma-most common variety of Ig

KOAc: Potassium Acetate

kDa: kiloDaltons

MgOAc: Magnesium Acetate

MW: Molecular weight

NaCl: Sodium Chloride

NaOAc: Sodium Acetate

nfH_2O: nuclease free water

NGS: Next-Generation Sequencing

NP-40: Tergitol-type NP-40/nonyl phenoxypolyethoxylethanol

OAc: Acetate

PBS: Phosphate Buffered Saline

PCR: Polymerase Chain Reaction

PMSF: phenylmethylsulfonyl fluoride

RCA: Rolling Circle Amplification

RI: RNase Inhibitor

RNA: Ribonucleic Acid

RT: Room temperature

RT: Reverse Transcription

SDS: Sodium Dodecyl Sulfate

SDSPAGE: Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis

ssDNA: single stranded deoxyribonucleic acid

TB: Transport Buffer (110 mM KOAc, 5 mM NaOAc, 2 mM MgOAc, 2 mM DTT, 1 mM EGTA in 20 mM HEPES)

U1C: small nuclear protein C

U87-MG: Malignant Gliomal cell line

UMAP: Uniform Manifold Approximation and Projection

LIST OF FIGURES

LIST OF TABLES

ACKNOWLDEGEMENTS

I would like to acknowledge and sincerely thank Dr. Kun Zhang for his support and mentorship. This Ph.D. would not have been possible without his steadfast support or his critical guidance during many experiments of troubleshooting. I would also like to thank my family: my eternally supportive parents, wonderful sisters and brothers in law. Support at home is as critical as support at work, and they have never wavered in it. I would also like to thank the entire Zhang lab for invaluable help on experimental protocols, computational pipelines, as well as general help and friendship.

VITA

2012 Bachelor of Science, University of Minnesota Twin Cities

2019 Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Bioengineering

Professor Kun Zhang

ABSTRACT OF THE DISSERTATION

Novel Multiomics Method of RNA and Nuclear
Protein Characterization in Single Cells

by

Daniel Eric Jacobsen

Doctor of Philosophy Bioengineering

University of California, San Diego 2019

Professor Kun Zhang, Chair

Characterizing single cells has become an important field in understanding human structure and function, both in healthy and diseased individuals. Recently, this characterization has included multi-omics methods, which use complementary data sets to better understand cellular function from several angles. RNA-seq is a staple for characterizing cell phenotypes, but fails to cover the entire picture of the cells phenotype. Protein expression has long been considered as a primary driver in characterizing cells, but inefficiencies in capturing meaningful protein data has long been a problem. Recently, antibodies conjugated to oligonucleotides have helped address these

inefficiencies, bringing the scale of single-cell proteomics assays more on par with single-cell RNA assays. However, these assays have characterized only cell-surface marker proteins, and have not yet shown an ability to capture proteins with more meaningful intracellular data. Of deeper interest are intranuclear transcription factors, which are directly related to RNA expression patterns. This dissertation describes strategies of single-cell protein and RNA capture, including design of probe oligonucleotides and antibody conjugation methods. It also describes use of antibody fragments and the small molecule pitstop 2 as a method of intranuclear protein assaying. Finally, the dissertation describes dual RNA-protein characterization of cell cycle in thousands of cultured cells, and additional conclusions drawn from this combined data.

INTRODUCTION

The human body is composed of around 30 trillion (10^{12}) cells. These cells are organized into tissues, with each tissue performing specialized actions. Further divided, within each tissue individual cells perform roles that, on the whole, contribute to the function of the tissue itself, and therefore of the individual body. Early studies into the cellular makeup of the human body were forced to rely on large populational data, such as millions of cells extracted during a blood harvest. These studies were limited because data obtained from populations of cells contains signatures from each cell type present, which must be deconvolved in order to understand the underlying nature of the cells. Recent advances in technology have allowed for the isolation and characterization of single cells, which has contributed greatly to our understanding of the human body, its structure and function.

All cellular types arise from the same genome, unique to each individual. Within this genome, approximately 20,000 sequences of nucleotides encode genes. The selective use of these genes, typically around 10,000-15,000 per cell, determines the type of cell and therefore its function (Fagerberg et al. 2014). Critically then, determining the representation of these genes in a given cell is one of the foundational ways we have of understanding cellular phenotype. Two important molecular types for investigating this phenotype are messenger RNA (mRNA), the transcripts created by expression of a gene, and protein, the molecules created by translation of mRNA into amino acid structures that make up the architecture of the cell.

mRNA has long been a gold standard among cellular classification. This is because mRNA is nucleotide based, and can be easily processed into a stable double stranded cDNA structure via reverse transcription protocols. Further, the nucleotide system has consistent base pair interactions (Adenosine-Thymine/Uracil and Cytosine-Guanine) that allow for reliable customization of

protocols and molecular probes. Finally, Next-Gen Sequencing (NGS) by Illumina and others relies on nucleotide base-pairing to cheaply produce millions of reads for hundreds of dollars. Since mRNA is a transitional molecule between the stable genome and the functional protein, it is a good metric to determine what actions the cell is taking. This readout can readily be used to identify cell types and functions within mixed populations of single cells (Lake BB et al 2017).

mRNA does not tell the full story however. Ultimately the mRNA is transitional, and therefore can miss out on information from more stable protein expression. An example is immune cells, which are found to have low transcriptional activity and therefore express few mRNAs which do not always reflect the immune cells' characteristics (Ecker *et al* 2017). Protein readouts, however, have long been challenging because protein structure is fundamentally different from nucleotides that make up DNA and RNA. This means that proteins cannot readily be incorporated into NGS-based protocols, which greatly reduces the throughput of protein-based methods. Technologies such as mass spectrometry have greatly expanded protein characterization, but such methods are destructive and often lack the context from mRNA and other molecules.

Recent advances have allowed oligonucleotide conjugation to proteins. By linking oligos to proteins, these methods connect the speed and throughput of NGS-methods, while still using the protein chemistry. The principal proteins used are antibodies. Antibodies are Y shaped molecules created by immune B cells that recognize other proteins, termed antigens, on foreign cells such as bacteria. These antibodies exclusively bind their antigen targets and are used by the B cells to signal the presence of foreign invaders. Antibodies can be raised and purified for a large variety of targets, and have been used extensively to signal the presence of a variety of proteins in human cells. By conjugating an oligonucleotide to the antibodies, the presence of that oligonucleotide is read in NGS technologies, providing a DNA-based readout of protein levels.

This gets around the limitations of many protein detection methods, which are typically limited by fluorescence or other secondary-antibody characteristics. A fluorescence limited study is usually limited to studying around 4 proteins, whereas the limit of NGS based methods is theoretically infinite.

Combining mRNA readouts and protein readouts is a powerful way to better understand the phenotypes of cells present in the body. CITESeq and REAPSeq are two such technologies, which demonstrated combined characterization of proteins and mRNAs in thousands of single cells (Peterson et al 2017; Stoeckius et al 2017). These technologies had some drawbacks, however. Both studies focused on surface markers, proteins present at the surface level of the cell that are in high abundance and are used in identification by the immune system. These are important proteins for the structure of the cell, but do not characterize the activity going on within the cell. Opportunities exist therefore for technologies that address the intracellular and even intranuclear protein content of individual cells, as well as capturing their transcriptional activity.

CHAPTER 1

SYSTEMS FOR COMBINATORIAL-INDEXING BASED METHOD OF RNA AND PROTEIN MEASUREMENT

1.1 Abstract of Chapter 1

The first step towards building single-cell systems for dual omics assays is to work towards compatibility of the two omics methods. For proteomics, which operate on inherently different biological systems than oligonucleotide systems (transcriptomics, genomics, epigenomics), it is key therefore to integrate the protein detection in compatible ways to the RNA detection. For this purpose we conjugated oligonucleotides to antibodies with designed sequences which could be read during RNA detection. We then looked into several systems leveraging the oligo-conjugated antibodies to multiplex with targeted RNA detection systems. Multiplexing systems add in barcoded oligos to the cDNA being read in the sequencer. These barcodes, combined over several rounds of split-pooling, yield combinations of barcodes that allow for much larger amounts of cells. Early experiments in targeted protein systems showed high specificity in capturing both proteins and oligos, but had too low sensitivity to make effective protocols. Experiments with one-step methods that drastically reduced the complexity led to much higher sensitivity, but at an (acceptable) loss of specificity. Still, these one step methods, or a possible combination of them, remain the best solution for targeted capture.

1.2 Introduction to Chapter 1

The first stage in developing new technologies is design. The design concept for this project had several requirements. These requirements will be discussed in further detail but are listed here:

1. Protein must be made to be NGS compatible. This means conjugating an oligonucleotide to a given protein that will be detected and processed during the protocol into sequencing library-ready information.
2. The system must use combinatorial indexing. This means additional oligonucleotides must be added into the process with specified regions of variation (barcoding) that can be mapped back to spatially distinct locations in sequencing results.
3. The system must capture both RNA and protein data, preferably using the same methodology in order to reduce the number of steps involved.

As mentioned in the main introduction, protein data can be immensely valuable in cell classification and understanding. Gathering this data can be difficult, however. Two previous methods used are secondary-antibody visualization methods and protein mass spectrometry. Mass spectrometry will not be discussed, since it is too destructive and has no compatibility available with RNA measurement methods.

Secondary antibody visualization methods rely on a cascade system of antibodies for protein detection and visualization. First, a primary antibody is raised against a specific protein of interest. Usually these primary antibodies are raised in a host animal, typically mouse, rat, goat, or rabbit. Secondary antibodies are also raised which target the primary antibodies of a single species. This can be done because antibodies have constant regions that are reused for every antibody raised in the host, along with variable regions that target specific antigens. These constant domains are

targeted by antibodies raised in a secondary host (e.g. raise anti-rabbit antibodies in a mouse). These secondary antibodies are conjugated to a detection molecule, frequently a fluorophore, which are then imaged via microscopy. This helps to amplify the signal, as many secondary antibodies may bind to a primary, which improves the ability to visualize the protein. Secondary antibody visualization techniques are useful to see proteins and their localization in cells, but are limited in their capacity for testing multiple proteins at once. Since the technique is visual, spectrally distinct fluorophores must be used in order to distinguish one secondary antibody from another. Do to limitations of microscopes as well as the large width of emission spectra, only around 4 spectrally distinct fluorophores (and therefore proteins) can be used in a given experiment.

An NGS method can sidestep the drawbacks inherent in normal antibody studies. The method uses primary antibodies as the original detectors. However, rather than secondary antibodies the primary antibodies are conjugated with oligonucleotides. These oligonucleotides have distinct sequences from one another that, when read by a sequencer, correspond to the individual antibody and therefore the individual protein of interest. Antibody protein detection is less destructive than mass spectrometry, but NGS utilization allows the wider capture rate of MS technologies. Design concepts utilizing this feature must design their oligonucleotides to be compatible with RNA-based capture methods.

Multiplexing is a method of single cell capture that is cheaper and easier than other methods (Cusanovich et al 2015). At its simplest, multiplexing is a split-pooling based strategy that adds additional oligonucleotides to the capture method. Along with constant sequences like PCR handles or adapters, these oligonucleotides contain a barcode (typically 8-12 bp). When read in a sequencer, these barcodes map the cell back to a physical location. By using multiple levels of

split-pooling, cells are given one of a set of oligos (e.g. 96 in a 96 well plate), then pooled back together and split into a new 96 well plate where they are given a second set (e.g. another 96). When read in a sequencer then, the wells have 96x96 or 9216 different combinations of those two oligonucleotides. If the number of cells is much less than the number of barcodes (e.g. 500 cells) the likelihood that two cells will receive the same pair of barcodes is very low. The introduction of these oligonucleotides can be through many factors, including hybridization, PCR extension, reverse transcription, ligation, and others. The second portion of this chapter will discuss multiplexing in further detail.

The final design guideline followed will be compatibility of RNA and protein measurements in process. This step is necessary, but becomes increasingly difficult the more divergent the two methods are. This is because many reactions require specific conditions, be it pH or presence and/or absence of certain molecules (e.g. polymerases require the presence of divalent cations and therefore must contain low levels of the chelating molecule EDTA if any). Though simple in concept the last section of this chapter will include many designs, each of which included adaptations necessary to fulfill this constraint.

1.3 Results and Methods

1.3.1 Antibody-Oligo Conjugation

The first step in design is developing a protocol for antibody-oligonucleotide conjugation. This step has become important enough that at least one company has a dedicated product line of antibody-oligonucleotide conjugates (https://www.biolegend.com/en-us/totalseq). This service was not available at the time the process detailed here was conceived and implemented, and further challenges exist which made it unfavorable to use commercial services.

Custom protocols have previously been developed for production of antibody-oligonucleotide conjugates (Assarsson et al 2014; Darmanis et al 2016). These protocols utilize a reaction setup that targets active groups on protein, as well as a chemical group that is present on the oligonucleotide, typically added by the production company such as Integrated DNA Technologies (IDT, San Diego CA). The two molecules are linked via a crosslinker containing reaction targets for both.

The most common group to target proteins target primary amines ($-NH_2$). Primary amines are present at the N terminues of proteins, as well as being the functional group of lysine amino acids. Many molecules exist that can react with primary amines, but the one chosen for this group is the NHS Ester. The ester group present on this molecule reacts with the primary amine to form a stable amide bond (Fig 1A). This reaction is random, and can occur on any of the primary amine groups present on the antibody.

For oligonucleotide chemistry, many groups again exist. Azide reactions were eventually chosen, due to their more controllable reaction chemistry over amine conjugation (protein and oligo) chemistry. The azide molecule participates with a dibenzocyclooctyne molecule in a reaction that is fast at room temperature, as well as stable in the presence of other functional groups

(Fig 1B; https://www.interchim.fr/ft/D/DQP580.pdf). For antibody-oligo conjugates, the azide group was placed at the 5' end of DNA oligonucleotides. The molecule DBCO-PEG_5-NHSEster was therefore chosen as the chemical crosslinker, which served to react with both functional groups (Fig 1C). Polyethylene glycol (PEG) is a simple, generally nonreactive chemical group often used as a spacer.

The conjugation reaction was performed by first adding the DBCO-PEG_5-NHSEster molecule with the protein at 20X molar excess in 1X PBS. The reaction was left at room temperature for half an hour before being quenched by adding 20 uL of 0.5M Tris-HCl and waiting an additional 5 minutes. Since the tris molecule contains primary amines, this competes with the primary amines on the protein and effectively quenches the reaction. Following quenching, the protein is added to an Amicon Ultra 50 kDa column and washed twice according to directions. This is used as a buffer exchange back to PBS, and helps to get rid of excess DBCO-PEG_5-NHSEster molecules. After buffer exchange and protein retrieval, azide-conjugated oligos were added to the activated protein molecules and incubated overnight at 4C. Afterwards the protein-oligo conjugates were once again purified using an Amicon 50 kDa column. Final results were measured in a Qubit using both the protein module and the ssDNA module. Results were also visualized via gel (Fig 1D). By the gel, it can be seen that the original antibody band (around 150 kDa) is shifted in the conjugated lanes. Additional bands appear above the original band that represent single, double, and larger numbers of conjugations to single antibodies. By tuning the molar ratios used, particularly oligo/protein, we can tune the expected numbers and strength of these bands. The reaction probabilities follow a Poisson distribution, with the lambda being approximately equal to the molar ratio of oligo/protein.

It should be mentioned that the most important region of an antibody is the antigen binding fragment (Fab), which is composed of the variable regions of the antibody. Chemical bonding to this region could affect the antigen/antibody binding, and therefore reduce or eliminate the antibody's function. A total antibody, composed of two heavy chains, each at around 550 residues, and 2 light chains, each around 215 residues (Janeway et al 2001), is around 1430 residues total. The typical lysine content of a protein is around 7.2%, which means that on average in a 1430 residue protein we can expect around 100 lysine residues in a given antibody. The Fab region on the other hand is around 430 residues (although there are two of these per antibody for a total of 860). Given the same average content, we can expect around 30 lysines in a single Fab region, or 60 in the combined Fabs. It is important therefore to keep the number of lysines conjugated by oligonucleotides to a minimum in order to keep the chances of binding in the critical part of the Fab as low as possible.

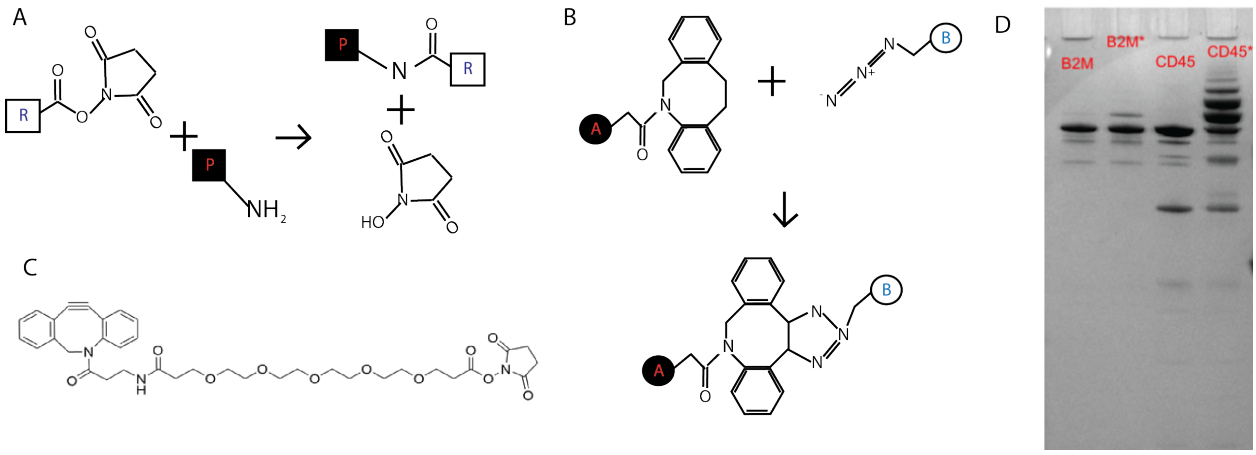

Figure 1: Biochemistry for antibody-oligo conjugates. (A) NHS-Ester reaction with primary amines on protein to form a stable amide bond. Typically this is at the N terminus or on lysine groups. (B) Dibenzocyclooctyne molecule reacts with azide moiety at the 5' end of an oligonucleotide (labeled B) to form a stable ring structure. (C) DBCO-PEG5-NHSEster molecule used for chemical conjugation of antibodies and oligos. (D) SDSPAGE image showing two antibodies conjugated with oligonucleotides. The first and third lane show the protein alone, the second and fourth show the conjugated results. Note the band shift above the unconjugated lanes. CD45 (lanes 3,4) shows more average oligos/protein in conjugation than B2M (lanes 1,2).

1.3.2 Multiplexing

Multiplexing as a strategy for single-cell sequencing that has become popular in the last few years for a variety of applications (Cao *et al* 2017; Sos *et al* 2016; Chen *et al* 2018; Lareau *et. al* 2019). Multiplexing relies on barcodes, designed oligonucleotide sequences usually 8-12 bp long, that are added sequentially during various steps of RNA/DNA capture methods. These barcodes are typically by spreading the cells out into multiple wells, such as the wells of a 96-well plate(See Fig 2A for a 6 well example). Each well has its own unique barcode, as well as adapter sequences, sequences common to all of the barcodes. Adapter sequences allow for all the barcoded oligonucleotides to be treated in the same manner, such as amplification with a PCR primer targeting the adapter sequence or a ligation reaction that hybridizes to the adapter.

The simplest method of multiplexing is split-pooling (Fig 2A). Short oligos are added in each reaction, with each containing a barcode unique to its well of a plate. The number of barcodes is much lower than the number of cells, however. That is because after the first reaction, all cells are collected into a single container and remixed. Afterwards, the cells are distributed into a new 96 well plate, where another reaction is performed. Once again, the number of cells is much larger than the number of barcodes. However, each cell now has two barcodes, one from the first plate and one from the second. This means that the likelihood of any cell having the same two barcodes as another is now the multiplied probabilities of the number of barcodes in each round of the experiment (96x96=9216 for a 2 round, 96 well plate example). This process can be repeated as many times as desired to increase the size of the barcode space. Each new round increases the likelihood of contamination or of losing cells, and the benefits for an additional round must be weighed against the negatives.

While each round of barcoding introduces additional cell loss, it also increases the amount of barcode space (Klein et al 2015). Barcode space is important, because in a split-pooling experiment, the number of barcodes "occupied" by single cells must be much less than the number of barcodes in order to prevent barcode collisions. This problem, how many cells can be processed with N barcodes, is akin to the so called birthday problem. The birthday problem addresses how many people can be in a room before the likelihood is high that 2 people share a birthday. The problem can be modeled as a Poisson distribution where the fraction of barcodes with k cells is given by:

$$p = \frac{\lambda^k e^{-\lambda}}{k!}$$

where lambda is the expected value (number cells/number barcodes). Since collision rate is defined as any multiplet of cells, the collision rate is the sum of all probabilities for k≥2. It is easier therefore to calculate the multiple rate as the inverse of probabilities 0 and 1, written as:

$$p(k \geq 1) = 1 - \big(p(k=0) + p(k=1)\big) = 1 - \left(\frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!}\right) = 1 - e^{-\lambda}(1+\lambda)$$

To keep the number of doublets small, therefore, one must limit the number of cells in the experiment to have lambda approach 0. For a 10% loading, for example, this yields an approximately 0.5% doublet rate. In the 2 plate, 96 well/plate example this means approximately 1000 cells can be used in the experiment.

Advantages of the combinatorial indexing system are removal of expensive, specialized equipment and larger throughputs. One of the earliest, common instruments used in single-cell experiments is Flow-Assisted Cell Sorting (FACS), which can allow for more precise isolation of single cells (Svensson et al 2018). FACS presents challenges, however, as the machines are very

expensive and generally require dedicated personnel for their operation. Combinatorial indexing removes the need for FACS by substituting the active capture technology of FACS with random capture technology. Other sources of individual cell capture technology include passive models, such as Fluidigm's C1 device (Zeisel *et al* 2015). Most passive capture technologies are limited by throughput, however, and often require unique setups/machines similar to FACS. Combinatorial indexing has advantages over these in throughput; a typical C1 experiment contains 10s-100 cells, while a typical combinatorial indexing protocol investigates 1000s-100,000s. Finally, cells can be added through random barcoding, such as Drop-seq and In-drop, microfluidic droplet-based capture technologies (Macosko *et al* 2015; Klein *et al* 2015). While able to process many cells, a microfluidic setup still requires specialized equipment and expertise. Split-pool combinatorial indexing provides easier setup and implementation.

Thus far the design of multiplexing has focused on the throughput-increasing power of split-pool combinatorial indexing, as well as its cost reduction over similar methods. Focusing closer on the method, each split step includes a biochemical reaction which adds a new DNA oligonucleotide to the cDNA read in a DNA sequencer. Several options exist for biochemical addition. The most straightforward is extension (Fig 2B). This can be done by a DNA polymerase in the case of DNA/DNA interactions, or can be done with a reverse transcriptase in the case of RNA/DNA interactions. The primers used for extension target the adapter sequence of the current oligo, making the extension primer universal to all barcoded primers. Another simple modification is ligation (Fig 2C). The new oligo is added to the solution along with another oligo termed the linker. The linker has sequences reverse complementary to adapter sequences on both the original oligo as well as the new oligo being added. This is because DNA ligase needs a double stranded molecule to perform ligation. The linker also serves to anchor both barcode oligos together during

ligation. There are additional ways to add barcoded oligos to the cDNA constructs, but these will be the main methods used here.

Finally, a quick note will be made to the multiplexing design of the antibody oligonucleotides used for protein information capture. Since these oligos are designed, they can be tailored to suit the needs of the protocol. As mentioned previously, the antibodies contain a barcode sequence which identifies them as belonging to a specific antibody and therefore protein, but also contain adapter sequences of their own. These adapter sequences can be specific, as shown in methods in the following section. The adapter sequences can also be more generic, such as a poly dA tail that mimics mRNA, as shown in later chapters.

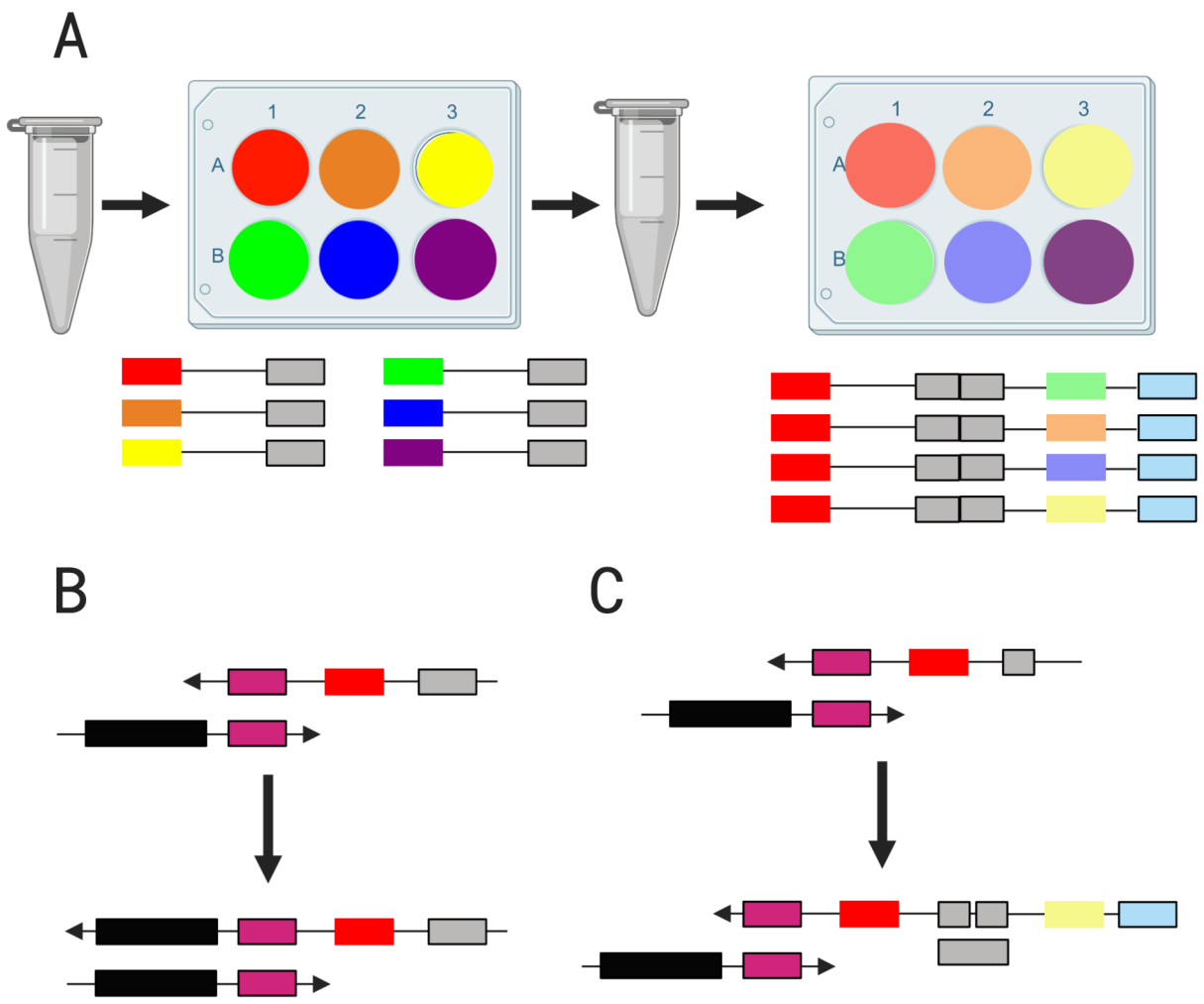

Figure 2: Multiplexing strategies. (A) Overall process of split-pooling. Pooled cells in a microfuge tube are distributed into the wells of a plate, in this example a 6-well plate. Each well has an oligo to be added to the cDNA. The oligo contains a barcode region specific to each well (colored box) and an adapter universal to all wells (grey outlined box). After addition to the cDNA the cells are pooled and redistributed in a new set of wells with different oligos containing new barcodes (pastel colors) and a new adapter (light-blue, outlined box). (B) Extension of cDNA by polymerases. Added oligo (top) anneals to specific region of molecule of interest (magenta; e.g. a dA tail of an mRNA) and is extended by a polymerase across the upstream part of the molecule. (C) Ligation cDNA extension. After initial annealing (magenta region) two more oligos are added, the new barcode primer (right molecule), as well as a linker (large grey box) that spans both the old oligo and new oligo. The two barcoded oligos are fused together by a DNA ligase. Created with biorender.com

1.3.3 Targeted Systems for Protein/RNA Capture

1.3.3.1 Version 1

The first designs of the Combinatorial Padlock (ComboLock) protocol used probes based off the PLAYR protocol (Frei *et al* 2016). The design uses a pair of "C" probes that target two ~20bp sequences on the same RNA/protein oligo molecule, typically about 5 bp apart (Fig 3A). Each of these probes has a region at one end that matches the protein oligo/RNA of interest, followed by a spacer region of around 10 bp, then a region containing a barcode and adapter sequences. The barcode is specific for that RNA/protein oligo, and is used in the Illumina sequencing data to distinguish the RNA or protein of interest being measured. By using two probes instead of one, the specificity of the experiment is greatly increased, since a correct "hit" requires two compatible probes to be near one another. The two probes can be given different barcodes or the same barcode. The same barcode allows for a shorter read during Illumina sequencing, since only one barcode needs to be read, but the second allows for additional quality filtering and specificity.

Once the probes are hybridized to the mRNA/protein oligo, the excess probes are washed away and new probes are added (Fig 3B). These probes are a pair of barcoded probes, termed the latch and padlock. The padlock is the larger of the two, and wraps around the outer adapters present on the two C probes. Padlock probes were initially introduced in 1994 as a method of DNA circularization, with enhanced sensitivity and specificity due to having two nearby hybridization sites on each probe (Nilsson et al 1994). The latch is the smaller of the two and connects the inner adapter sequences. The hybridized form is a circular piece of DNA with two holes in it; these holes are the barcodes that dictate which RNA/protein is being measured. The oligos are used in the first

level of the combinatorial indexing. The padlock barcodes are distributed along the rows of a 96 well plate (8 unique padlocks) and the latches are distributed along the columns (12 unique latches). The combination of the two forms 96 different possibilities. (Fig 3C).

These gaps are filled in by a double procedure of gap fill (by extension) and ligation (Fig 3D). This reaction is performed by Phusion High Fidelity (HF) Polymerase and Amp Ligase. These two enzymes are chosen because they use compatible buffers (Amp Ligase buffer) and temperature ranges (~60C), which allows the enzymes to work in the same thermocycler without adding additional buffer/enzymes. Additionally, Phusion polymerase does not have strand displacement functionality, which prevents the enzyme from displacing either the latch or padlock during extension. This leaves the fully extended oligos as one continuous line, with backbone break points that can be fused by ligase. Once this process is done the whole cDNA molecule is now a single, circular piece of cDNA. Exonucleases are added to remove any material that is not a circular piece of cDNA (i.e. incomplete).

The circular cDNA molecules are then amplified via Rolling Circle Amplification (RCA; Fig 3E). RCA is an isothermal reaction, which eliminates the repeated heating-cooling processes of PCR that can be disruptive to *in vivo* models (Mohsen and Kool 2017). Isothermal reactions can also be more easily paired with additional reactions, since RCA is typically performed between 30-40C. PCR, on the other hand, goes to 95C during the denaturation step, which disrupts the structure of non-thermostable enzymes. Additionally, RCA with a single primer is effective at linearly amplifying material. Linear amplification has been shown at reducing the stochastic nature of amplifying low amounts of starting material (Grisedale and van Daal 2014).

After a short linear amplification, the cDNA can be further amplified with PCR, adding on Illumina sequencing adapters and a second set of adapters (Fig 3F). Combining this set of up to 96

adapters, added during PCR, with the latch/padlock added in previous steps, gives 96x96 or 9,216 different combinations of cell barcodes.

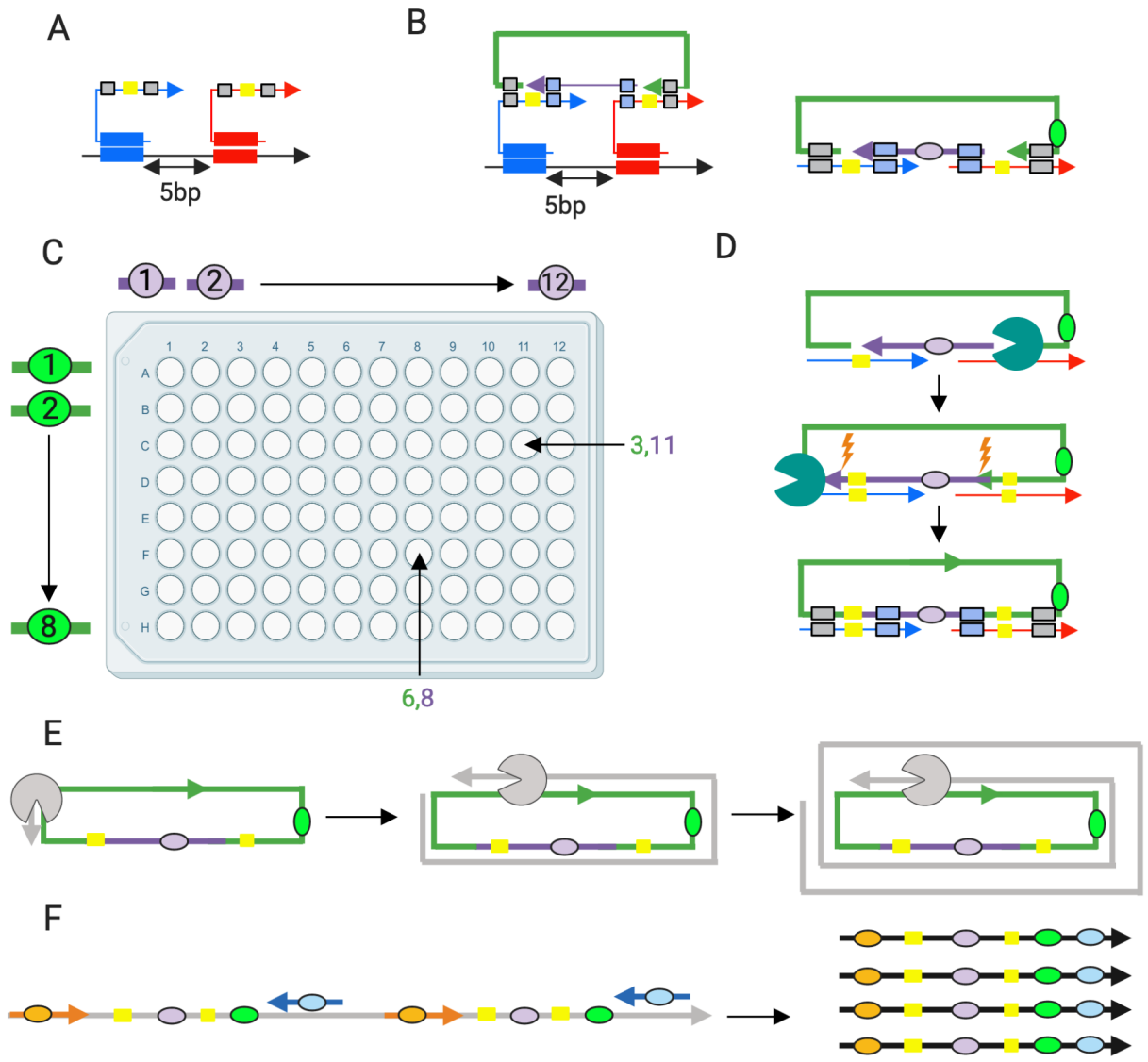

Figure 3: Combolock version 1. A) Two C probes are hybridized to template RNA or antibody-conjugated oligo. C probes are designed with a transcript matching region (red/blue box), spacer region, 20 bp adapter sequences (grey boxes) and a barcode unique to the C probe (yellow) Transcript matching regions are designed to be 5-7 bp apart. B) A latch (purple) and padlock (green) are added. Latches match the interior adapters (light blue) and padlocks match the exterior adapters on each C probe. C) Latches and padlocks contain barcodes (ovals). In a 96 well plate, latches and padlocks are combined so that each well contains a unique pair of latch/padlocks. D) Extension and ligation. A non-strand displacing polymerase (green pacman) extends latches and padlocks to capture C probe barcodes. Following extension Amp ligase fuses the two strands together (lightning bolts). E) Rolling Circle Amplification (RCA) creates a long piece of cDNA that contains many repeats of all adapters and barcodes from C probes, latch and padlock. F) PCR is performed on RCA product, adding in Illumina index adapters using the same barcoding strategy as shown in 3C. Figure made with biorender.com

1.3.3.2 Modifications to Version 1

Several design modifications were made to this system in order to improve it. The original design contained a bottle-neck step in the extension/ligation reaction step (Fig 3D). In order for successful circularization of the product, both the latch and padlock had to be bound to their targets on both adapter positions; 4 adapter hybridizations total. If any of these hybridizations failed, the reactions would either fail or produce an unwanted target. To reduce unwanted interactions it was decided to start by ligating the latch, now termed bridge, to the two C probes before padlock capture (Fig 4Aii). The bridge oligos contained no barcodes, and to compensate padlocks containing 96 different barcodes were designed. Additionally, for this step to work, one of the C probes orientations was switched in order for the ligation to function properly (Fig 4Ai). Post-ligation, padlock probe hybridization was performed and the extension reaction would proceed across a single oligo (Fig 4Aiii).

The bridging process was tested for efficacy. The protocol was tested using a model system that used a biotinylated oligonucleotide template with specially designed C probes against the template. Negative controls used removed either template (No Template), C probes (No C), 5' phosphate that is necessary for ligation (No PO_4), or ligase (No Ligase). Template oligo was added to streptavidin bound beads and excess washed away via magnetic pulldown and supernatant removal. Afterwards, C probes were added in PBS with 0.05% Tween-20 and allowed to hybridize for 15 minutes at 50C. After C probe hybridization beads were magnet pulled down and washed twice with PBS+0.5% Tween-20 before addition of ligation oligos (bridge and complementary bolt oligo). Ligation was performed at 37C for 15 minutes using T4 ligase. Afterwards the beads were washed twice more. Samples were incubated at 95C for 15 minutes to break the biotin-streptavidin bond and release the DNA molecules prior to being added to PCR mixes. Presence of

each C probe was tested, as well as the presence of the final bridged product. Results for each condition were tested against the PCR Non Template Control NTC. NTCs only contain primers, and show the primers tendency to amplify in the absence of template. Anything amplifying in a number of cycles close to the NTC is therefore low or even zero presence of template. Results are shown in number of cycles above the NTC threshold.

Results show that the positive sample had higher quantities of product than any of the controls (Fig 4B). In fact, the sample amplified ~16 cycles before the next-highest control excluding the No PO_4^- control (16 cycle difference is roughly 2^{16} or 60,000 times more material). The No PO_4^- control was later found to be caused by a problem with IDT synthesis, and No PO_4^- controls were greatly reduced in signal when subjected to treatment with recombinant Shrimp Alkaline Phosphotase (rSAP, removes 5' phosphates) before addition to sample. The No Ligation control is important in establishing that, while the C probes were present at similar levels to the sample with ligase, the product was greatly diminished (~1,000,000X).

The full protocol was then run with these conditions, testing out two different enzymes. For this experiment, template binding, C probe hybridization and latch bolt were performed as above, except that for the No PO4 sample the bolt oligo (complementary piece for ligation, green oligo in 4Aii) was incubated with (rSAP) prior to addition in the ligation reaction. After ligation, the mixture was washed and the padlock hybridized for 15 min at 50C. Following a wash step, circularization buffer was added containing 40 nmol NAD+, 600 pmol dNTPs, 15 umol betaine, 10U of Amp Ligase, and 6.4U of Phusion polymerase in 20 uL of 1X Amp Ligase buffer. Circularization (extension/ligation) was performed for 4 hours at 55C. Uncircularized DNA was destroyed by addition of 35U each of exonuclease I and III. Following exonuclease digestion RCA was performed using ThermoFisher Phi29 following standard protocols and using a 3 hour

incubation at 30C. Primers were triple-phosphothiorate bonded for the last three bases to prevent Phi29 exonuclease activity. Following RCA, samples were digested with restriction enzyme BglII for 2 hours at 37C following NEB protocols. BglII cut site was chosen on the padlock backbone (Fig 4Aiii) to create smaller strands and prevent large repeats of product from forming during PCR. Samples were then amplified via PCR to add Illumina adapters for sequencing.

Results from the full protocol are shown in Figure 4C. The final desired product was 221 bp, but due to the poor quality of the gel the bands appear more like smears. Still, the gel shows that the No template and No C probe controls, which should be impossible to produce product from, do not produce any notable product. The No PO4 control still produces some product, likely due to some ligations still occurring without 5' phosphates present on the DNA or incomplete removal of all phosphates by rSAP. Still, the Sample band is considerably darker than the No PO4 band (with same PCR stop time and gel loading volume), suggesting that the No PO4 sample had fewer completed templates than the positive sample.

This difference in quantity of successful circularization events between the sample and the No PO4 condition can be further examined by sequencing. Since the reads are barcode based and therefore contain the same sequence for every read with the same cell barcode (combination of padlock barcode and PCR indexes), Unique Molecular Identifiers (UMIs) were used in order to determine absolute quantity of product in each sample (Kivioja *et al* 2012). UMIs are degenerate sequences of bases, typically 8-12 bp in length (Fig 5A). During normal DNA synthesis, each base (A,C,G,T) is added in appropriate sequence. Only one base is added at a time, and all excess bases are removed prior to proceeding to the next base. To make a degenerate sequence, all bases are added at once in equal numbers. Approximately 25% of the DNA synthesized receives each base. Repeated over multiple cycles, this produces oligonucleotide strings that are unique to individual

molecules while the rest of the oligo is constant. This allows for correction of PCR bias, since PCR clones will have the same UMI, whereas molecules of the same sequence from different origins (e.g. two copies of a transcript) will have different UMIs. The UMI was added to the padlock upstream of the barcode.

For the experiment testing the sensitivity of the Combolock protocol, two different templates were used in varying amounts. Each template had two C probes designed against it as listed above. Each C Probe was tested alone with its correct template, as well as alone against the wrong template. Additionally, samples were tested containing both templates at varying ratios of Template1:Template2; 1:1, 1:9, and 1:9999. After sequencing, each sample was collapsed into total unique UMIs detected count. Results are shown in Figure 5B. As seen in 5B, template 1 or template 2 alone produced roughly 30,000-40,000 unique UMIs. When mixed, the UMI counts for each sample were found to be approximately the same as the ratios present (16,412:10,018 for 1:1 and 5,898:48,174 for 1:9). Due to signal to noise concerns, only T1 was measured for T1:T2 1:9999. This index yielded only 200 UMIs, which was less than the number of UMIs found in the orthogonal case when the same probes were used on the wrong template. It must be concluded then that the signal from the 1:9999 case was too low to be seen above background noise.

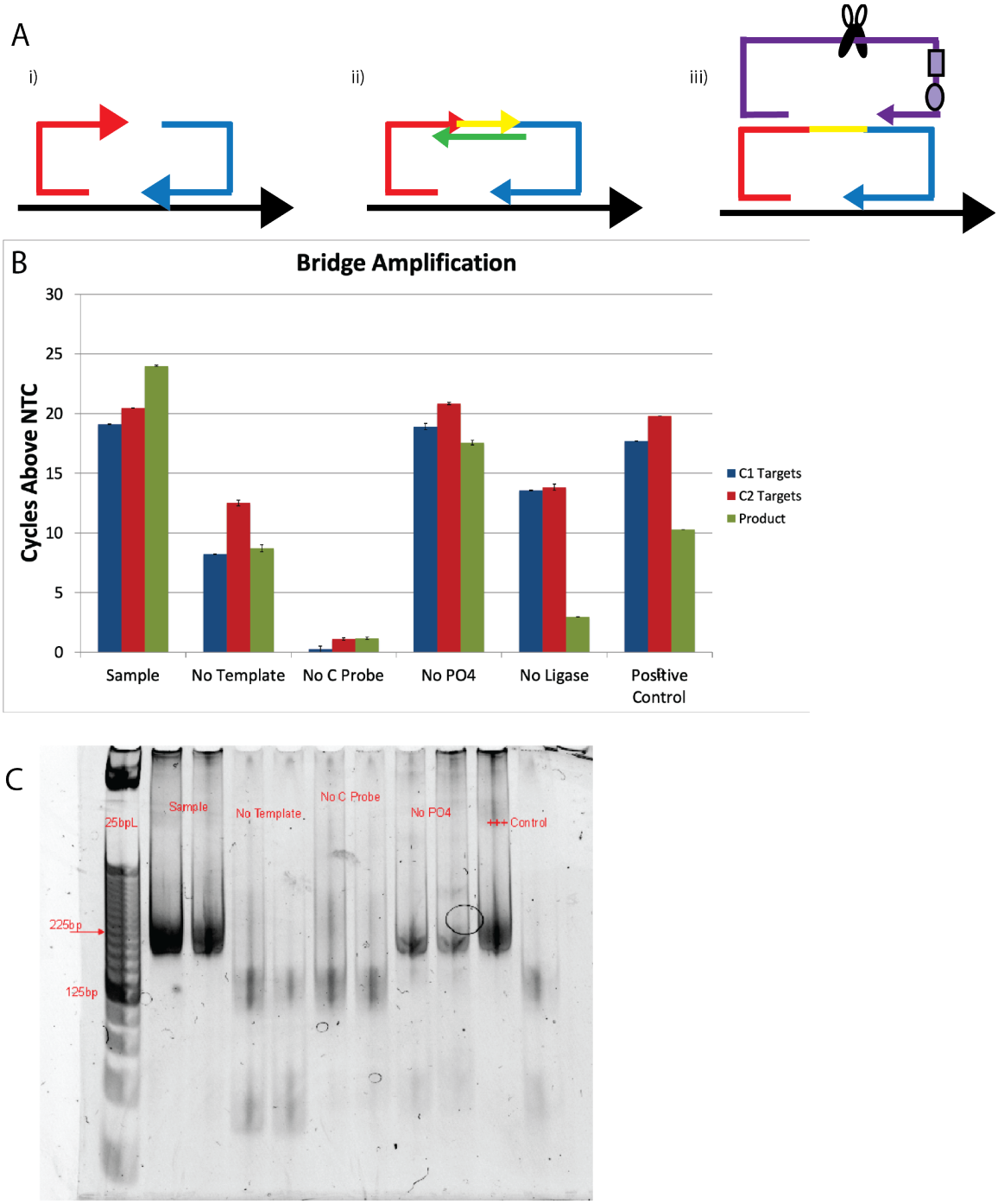

Figure 4: Improvements to Combolock version 1. Ai) Mirror C probe changes orientation of right C probe so that the 3' end matches the template. Aii) Mirror C orientation allows "latch/bolt" ligation to form a single bridge for padlock capture. Aiii) Padlock capture of bridge with barcode (oval), UMI (box) and BglII cut site (scissors). B) C Probe and bridge product amplification for several conditions and controls. Sample shows larger amount present than negative controls. C) Final product bands visible on gel. Final v1 product was 221 bp, shown as smear on poor quality gel. No C Probe and No Template samples show no visible product at correct size, no PO4 control is lighter than sample, indicating less product for same amplification.

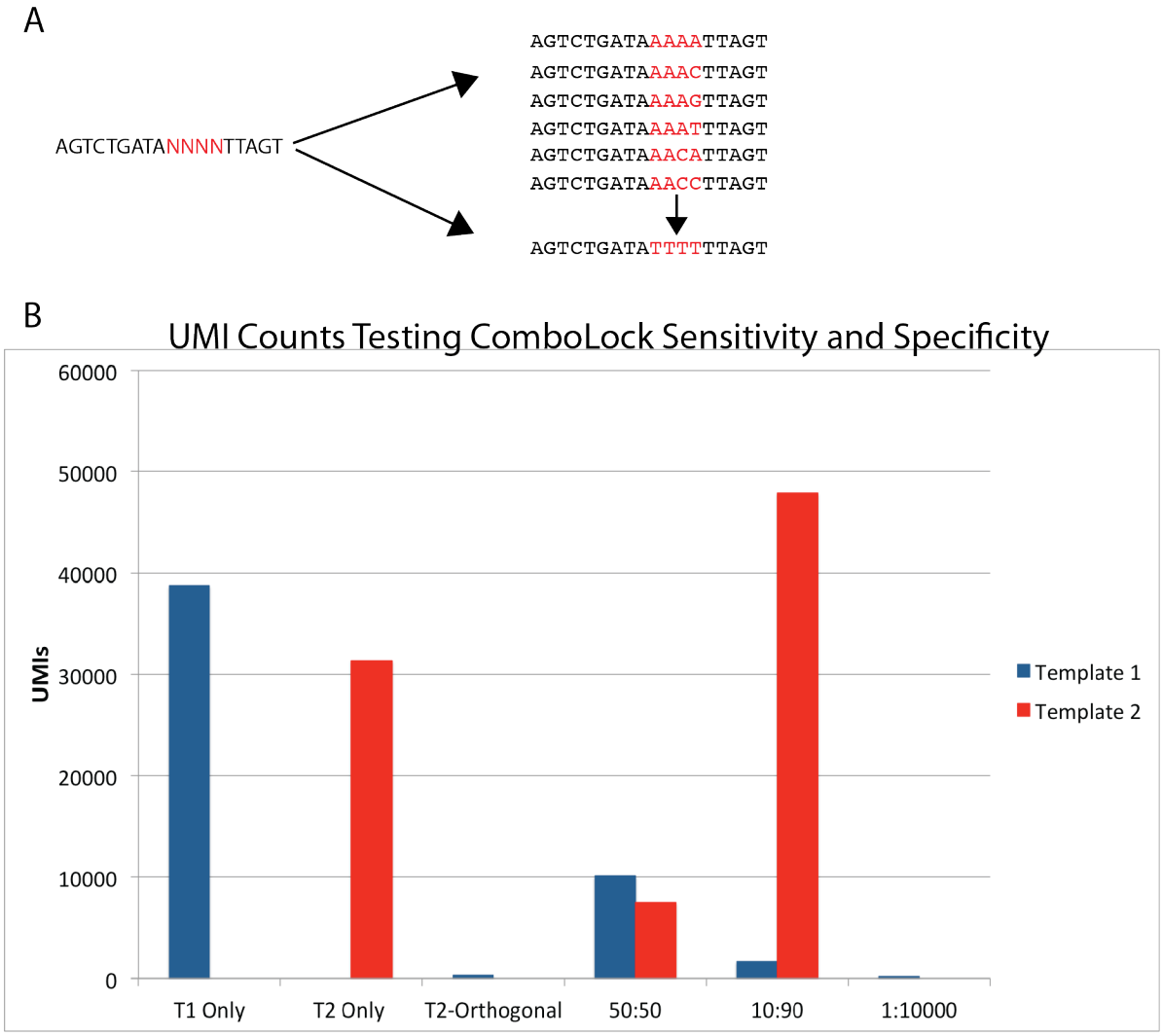

Figure 5: UMI Counting and Combolock sensitivity/specificity. A) Unique molecular indexes are degenerate sequences embedded in normal oligonucleotides. UMI on left is shown as ordered as NNNN, which leads to individual molecules having any combination of bases shown on right. B) UMI counts for sensitivity and specificity test.

1.3.3.3 Protein and Oligo Measurements

Following cDNA testing, multiplexing experiments were performed for protein and oligo templates in the same samples. For this experiment, a protein template was created by conjugating an antibody raised against Bovine Serum Albumin (BSA). Biotin-conjugated streptavidin was used as a template molecule. BSA was first bound to streptavidin beads as described above. For the experiment, 4 samples were used (Table 1). Samples were created using only oligo template, only protein template, a 50/50 molar ratio mix of both templates, and a version with no template molecules at all. C probes corresponding to the oligo template and the protein template were included in all 4 samples. After initial template binding all remaining steps in the protocol were performed as described in the previous section. Upon completion of the protocol up to Illumina Adapter addition, only the No Template sample had no discernable band at 221 bp (Figure 6). After sequencing, samples were demultiplexed and measured for total reads aligning to oligo, protein, and aligning to neither (unaligned). Afterwards reads were collapsed into unique UMIs detected for each template for each sample. Technical replicates were performed (labeled A and B) by performing the entire protocol in two different tubes for each condition.

In the case where the template was alone, the alignments were strongly in favor of the correct template, with negligible aligned reads coming from the wrong template. This suggests that the C probe binding is indeed orthogonal, and produces little false positives when templates are not present. The protein had considerable noise, however, as shown in the relatively large fraction of “unaligned” reads present in the protein sample. Additionally, while the number of protein reads present in the mixed samples is considerably greater than the number of reads aligning to protein template in the oligo-only samples, the actual count of UMIs is quite low. The oligo UMIs in the mixed sample are around 200X greater than their protein counterparts in the

same sample. This suggests that the protein binding model is weaker than the oligo binding, since downstream steps should not be affected in any way. This may be explained by the biotinylated oligos simply binding the streptavidin beads better than the biotinylated BSA molecules, or by the antibody binding being inefficient. Additionally, samples started with approximately 1 pmol (~10^{11} molecules) of template bound to the beads as a theoretical maximum. Given that the UMI counts were, at best, on the order of 10^4, this suggests that the overall process only has a fractional capture rate of 1 in 10 million. Since RNA transcript levels are on the order of 10^3-10^4 per gene per cell for even the highly abundant genes, this suggests the success rate would be too low even for abundant genes (Fagerberg *et al* 2014).

Table 1: Counts for reads and collapsed UMIs in protein/oligo mixture experiment.

Sample	Oligo Aligned Reads	Protein Aligned Reads	Unaligned Reads	Unique UMIs-Oligos	Unique UMIs-Protein
Oligo Only-Replicate A	442,733	21	1,222	15,543	19
Oligo Only-Replicate B	416,724	21	1,255	15,542	13
Protein Only-Replicate A	90	21,837	44,398	77	669
Protein Only-Replicate B	242	30,112	120,826	137	962
50/50 Molar Mix-Replicate A	170,802	463	468	6,014	31
50/50 Molar Mix-Replicate B	259,927	1,227	1,372	8,119	42

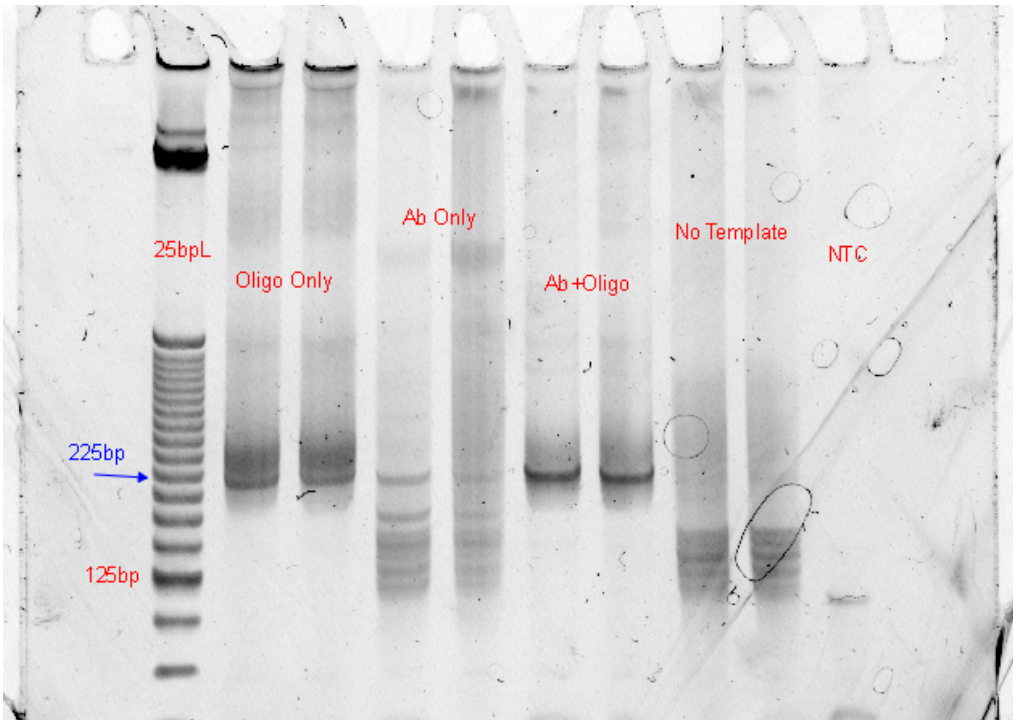

Figure 6: Gel image of oligo/protein mixture experiment. Gel bands at 221bp are strongly present in the oligo only and mixed samples. The antibody only samples have signal but considerably weaker than the ab only samples.

1.3.3.4. One-step Combolock Protocols (Version 4)

Given the low sensitivity of Combolock version 3, it was deemed necessary to adjust the protocol significantly. Each step in the protocol reduces the potential efficiency of the experiment, as no reaction has 100% efficiency, and it was therefore desirable to reduce the protocol to as few steps as possible. Additionally, padlock probes have been shown to have low efficiency, and it was therefore advisable to get rid of the padlock probe capture step. Two possible methods were devised using single oligo addition steps post C probe hybridization. The first used ligation to extend the oligo and add barcodes (Fig 7A-i). This method used 5 total oligos in a single ligation step. The ligated oligos, in green and purple, are added in the same manner as described in Figure 3C, with the green oligos changing along rows and the purple oligos changing along columns. Since the adapters do not overlap, it is possible to do this is in a single ligation reaction or in two separate reactions (data not shown). After the double ligation, the ends of the ligated molecule are amplified via PCR (orange and blue oligos, Fig 7A-ii).

To test the method, differing amounts of starting template were used, ranging from 1 pmol to 1 amol for oligos, and 0.3 pmol to 30 amol for protein. Biotinylated template was bound to streptavidin beads by incubating 2X molar excess with beads for 15 minutes at RT in 1X PBS+0.05% Tween-20. After binding beads were washed twice before addition of C probes. C probes were hybridized in 1X PBS+0.05% Tween-20 for 1 hr at 37C on a thermomixer shaking at 1200 rpm. Beads were washed twice in 1X PBS+0.05% Tween-20 post-hybridization. Following wash, ligation oligos were added from a single mix (20 pmol of each oligo) and incubated at 50C for 20 minutes in T4 ligase buffer to hybridize to C probes. Prior to addition to C probes, the four oligo mixes were incubated at 90C for 2 minutes, then 55C for 15 minutes in order to create annealed structures before addition to the C probe mix. This is expected to improve yield over

having all 5 oligos hybridize in the sample solution. After hybridization at 50C, the temperature was lowered to 20C and 1U T4 Ligase was added. The reaction was incubated at 20C for 30 minutes then 65C for 10 minutes to heat-kill the enzyme. The beads were washed two more times with 1X PBS+0.05% Tween-20 before being readied for Illumina adapter PCR.

This double ligation method was compatible with both the oligo template (Fig 7B-i) and the protein template (Fig 7B-ii), although both showed (lower) amounts of product created in the No Template controls. This is expected, as the methods using two C probes in close proximity are expected to improve specificity at the cost of sensitivity. The double ligation reaction only uses one C probe, and it is therefore expected to have the opposite effect, with increased background noise.

The second one-step method used was an extension motif (Fig 8A). The extension motif only utilizes one barcoded oligo in its first cell barcode (Fig 8A-i; purple). Therefore 96 different extension oligos are required for a 96 well plate to have individual barcodes in each well. This is similar to the changes made with the bridge oligo in 1.3.3.2, which utilized 96 different padlocks. Following extension of the C probe, which adds the first well barcode, the samples can be washed and added to a second PCR mix containing Illumina adapters (Fig 8A-ii). The Illumina adapters again have the same 8 rows by 12 column uniquely paired oligos, creating 96 different PCR combinations. Combined with the first step's 96, this maintains the same 96x96=9,216 combinations described in version 1.

After a 5 minute incubation for the primers to anneal to the C probe, the oligos were extended by a 3 minute reaction at 72C with *Taq* polymerase. Following the extension, a new master mix was created using 0.2 uM primers in 1X Kapa SYBR Fast master mix. 5 uL of extended sample (1:10 dilution) was added to each PCR well for this second reaction. This protocol was the

simplest design yet, and had the highest potential for non-specific interactions. The gel for the oligo test indeed shows that the 158bp product is detectable at as low as 1 amol, but also shows significantly in the No Template lane (Fig 8B). To determine whether or not this background signal (No Template) could be distinguished from the real signal, libraries were prepared for each of the concentrations and sequenced.

Results show that there is a difference between No Template data and positive samples (Fig 8C). To see this difference, reads were first aligned and deindexed to separate them by sample. Within a sample, UMIs were analyzed to remove PCR clones (same UMI for a given set of indices=PCR clone). Once collapsed, UMI counts were normalized by total read count for that cell index. This is because greater read depth in a specific cell barcode will likely lead to more UMIs. UMI counts are therefore normalized by read depth to remove this bias. Looking at the normalized UMI counts, there is a difference between the complexity in the true positive samples against the No Template control samples. This makes sense as PCR amplifications of a few C probes that are not washed out can still lead to signal in qPCR, but the complexity of these samples should be low. This lack of complexity is evident in the sequencing data. It is also interesting to note that the error bars for the three positive samples get larger as the template concentration gets lower. This makes sense also, as dropouts should increase as template concentration decreases, leading to greater variation between samples.

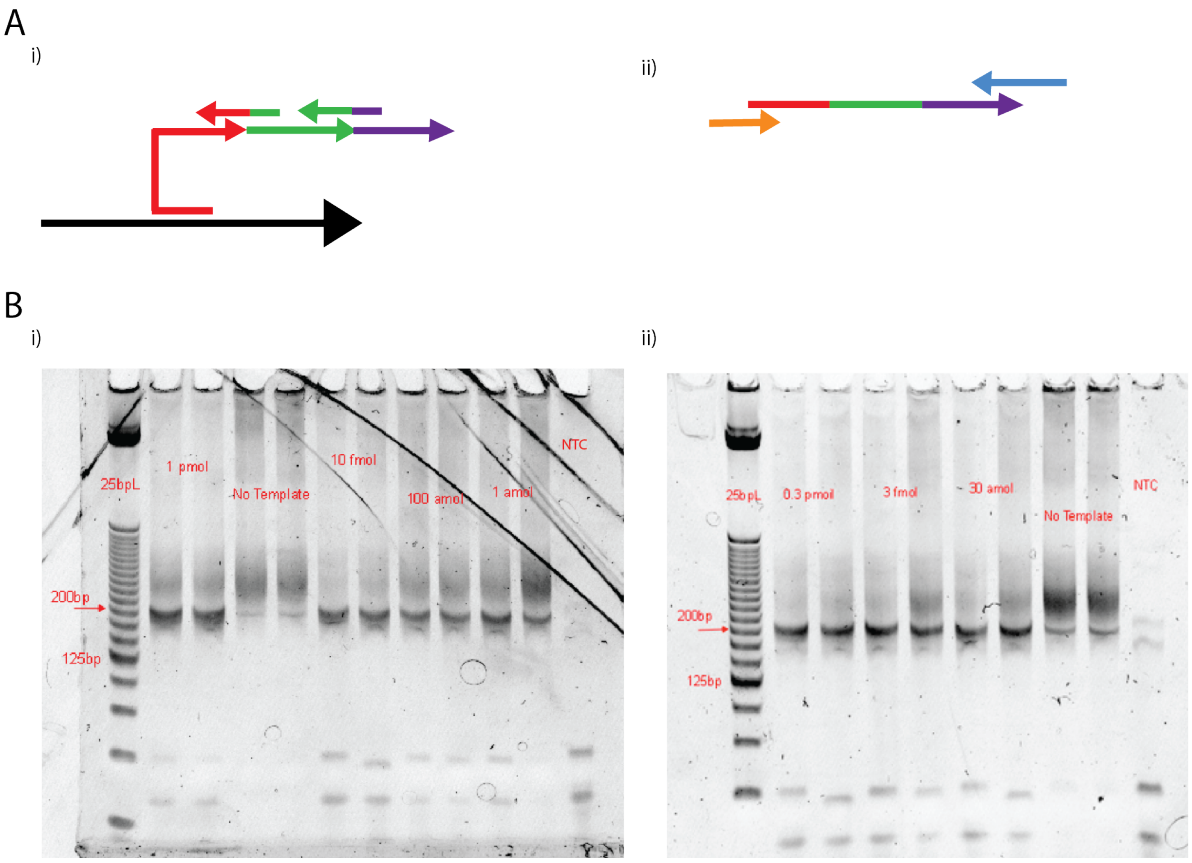

Figure 7: Combolock version 4 ligation. A) Ligation mechanism for v4. (i) Two barcode oligos (green and purple) are ligated onto the original C probe (red). Multi-colored oligos represent linkers used to create double-stranded regions for T4 Ligase. 8 different green oligos and 12 different purple oligos are used across a 96 well plate for 96 combinations. (ii) Illumina indexed-PCR for second cell-barcode addition. 8 orange oligos and 12 blue oligos add on another 96 barcode combinations. Total barcode space is 96x96=9,216. B) Gels showing correct bands at 189bp for oligo templates (i) and protein templates (ii) of differing template amount. No Template controls show product at lower but present quantities to positive libraries.

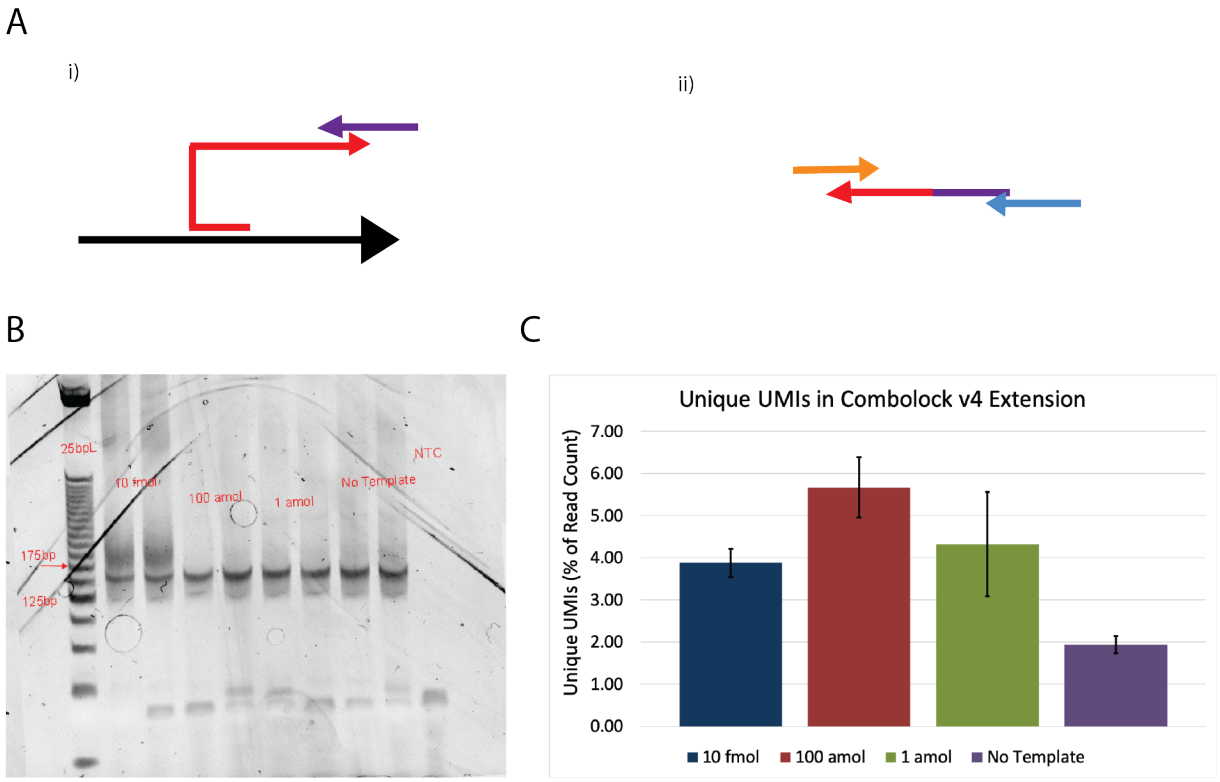

Figure 8: Combolock version 4 extension. A) Scheme for v4 extension. (i) Extension oligo (purple) is annealed to C probe (red) and extended by polymerase. 96 different barcodes are used, one for each well of a 96 well plate. (ii) Samples are washed and pooled before being added to a new 96 well plate for PCR. 2nd PCR adds two Illumina adapters (orange and blue) as described previously. Total combinations are 96 purple, 8 orange, and 12 blue, or 96x96=9,216. B) Gel showing product at 158 bp for samples as low as 1 amol. No template control also shows product. C) UMI complexity for v4 extension samples. UMI counts are expressed as a function of total reads for that index.

1.4 Chapter 1 Conclusions

Original designs for protein and RNA sequencing in single cells relied on targeted systems. These systems used specially designed "C" probes as proxies for the templates themselves, i.e. for RNA the transcript itself is never directly read. Rather, a C probe hybridized to a region of the transcript was then modified by the addition of barcoded oligonucleotides through extensions and ligations. The combination of these barcodes would allow individual cells to be inferred in sequencing data, as cell numbers are designed so that it is unlikely that two cells would get the same set of 3-4 barcodes. For protein, the presence of the target protein is transduced through 2 interactions: the antibody raised against the target and the oligonucleotide bound to the antibody that carries signal for C probe binding.

The targeted system design focused on two principal features: sensitivity and specificity. Paired C probe interactions allowed for greater specificity, as seen in the early versions' lack of product created from No Template scenarios (Figure 4B). However, this increased specificity proved to be at a great cost to the sensitivity of early Combolock protocols. Additionally, the complicated protocols made sensitivity low even in model systems (Figure 5B and Table 1). Version 4 of Combolock discarded the paired C probe approaches and focused on simple strategies for barcoding. Figure 8C shows that, although the noise (byproduct of low specificity) was increased by the simplifications, enough difference still existed in library quality between positive samples and the No Template control that the two could be distinguished.

Chapters 2 and 3 focus on adjustments to these protocols in order to adapt to cellular conditions, as opposed to the simple model systems discussed here. Chapter 2 focuses on the adaptations to antibodies to make them effective intracellular probes, and Chapter 3 discusses a

hybrid extension/ligation strategy used for greater breadth in RNA coverage than the targeted approaches of Chapter 1.

CHAPTER 2

TOOLS AND STRATEGIES FOR INTRACELLULAR PROTEIN DETECTION

2.1 Abstract of Chapter 2

Proteins exist in all locations of the cell from the cell exterior to the cytoplasm to the nucleus. Previous protein/RNA single-cell based methods have focused exclusively on exterior cell proteins, typically looking at cell clustering markers. A gap exists therefore for a method that can look at intracellular proteins. Intranuclear is more challenging even, as two cellular compartments must be breached for measurements. Pitstop 2 is a unique molecule that produces deformations in the nuclear pore complex, creating larger pores than would normally be allowed by the nucleus. These pores are still too large to accommodate full antibodies, but are small enough to allow entry to antibody antigen binding fragments (Fabs). Finally, this chapter contains a protocol for producing antibody Fabs and conjugating them to oligonucleotides. This protocol is the one used for all Fab-conjugates used in Chapter 3.

2.2 Introduction to Chapter 2

Chapter 1 focused on oligonucleotide design strategies for cellular multiplexing and protein/RNA targeting. Chapter 2 focuses on considerations regarding the specifics of protein capture in single cells. When considering protein capture, RNA capture will always be considered, as the two biomolecule types often have conflicting reactions to a given protocol. It is necessary, therefore, to balance the advantages in any protocol against its disadvantages, as something that is good for RNA capture (e.g. proteolysis) can be bad for protein capture. The key concepts to understand for designing split-pool strategies are: compartmentalization and permeabilization. Compartmentalization is the separating of individual cells. Permeabilization is the ability of the compartments to accept new material, namely reaction components such as enzymes and buffers.

Concerning cellular compartmentalization, multiplexed cells must be divided in some way to be able to distinguish them. In the earliest single-cell experiments, this was done literally, such as on the Fluidigm C1 where cells are individually captured into separate microwells. Dropseq continued this trend, physically isolating cells via a microfluidic device that produced oil in water emulsions (Macosko *et al* 2015). Cells are mixed with beads in lysis buffer and their material, specifically RNA, is captured via poly-dT motifs present on oligos bound to the bead. Each of these beads has a unique cellular barcode upstream of the polyT capture site, and during RT this information is merged with the transcript information from the mRNA. This bead and its barcode therefore becomes the new cellular compartment, as the beads are quickly pooled to reduce cost and effort in producing thousands of libraries. As discussed in Chapter 1, split-pooling is an alternative that requires no additional equipment such as FACS machines or microfluidic setups. Split-pooling relies on maintaining the cell as its own compartment. This means keeping all RNA

and protein within the cellular or sometimes nuclear membrane. To restrain cellular components within the membrane, fixation is frequently used.

Fixation can be achieved in several manners. Methanol and other solvents have been shown to retain a large portion of nucleic acids by precipitating them from solution (Srinivasan *et al* 2002). These nucleic acids can be restored with little lasting conformational change by the resuspension in aqueous buffer. However, they have been reported to cause denaturation of proteins even at relatively low concentrations (Fernandez and Sinanoglu 1985; Shao 2014). Formaldehyde is a more commonly used fixative, and has been shown to affect protein structure little (Mason and O'Leary 1991). Formaldehyde does have its drawbacks, however. First, formaldehyde crosslinks with several functional groups on proteins, including the N terminus and lysine, histidine, cysteine, tryptophan and arginine residues (Hoffman *et al* 2015). These, combined with conformational changes that may occur in the protein as a result of crosslinking, can affect the ability of antibodies to bind to their target epitopes. Studies have shown even a 5 minute fixation at 4C can reduce the effect of antibody binding, though only in rare cases is the effect enough to reduce the signal entirely (Otali et al 2009). Even ignoring the slight signal loss in protein, it is important to note that RNA signal is reduced by fixation, although it is recoverable under certain conditions (Russell *et al* 2013). Some methods have been used to chemically bond RNA to gel structures, a feature useful for imaging, but this can interfere with polymerase activity, making gel crosslinking unappealing for sequencing-based methods (Chen et al. Formaldehyde fixation is therefore a viable option, albeit one in which certain guidelines and restrictions should be followed.

Although fixation is useful in preserving cell morphology and integrity, permeabilization must still be used in order to access intracellular antigens or nucleic acids (Amidzadeh *et al* 2014).

This is an especially important consideration in split-pool multiplexing, where the cell membrane or nuclear membrane is the principal physical barrier separating cellular material from other cells (Vitak et al 2017). A membrane must therefore be permeabilized to allow probes and enzymes in while not being over-permeabilized, which can lead to signal bleeding or loss. One common way of permeabilization is through the use of detergents. At low concentrations, detergents are useful for selectively or non-selectively permeabilizing holes in membranes. At higher concentrations they can be disruptive, destroying membranes and even denaturing proteins. For example, NP-40 and Triton X-100 are mild, non-specific detergents that permeabilize at ~0.1% v/v (Amidzadeh et al 2014), but lyse cells at ~1% (Ji 2010). In fact, NP-40 is frequently used in nuclear extraction buffers, as at concentrations that lyse the cellular membrane the nuclear membrane remains relatively intact (Galvis *et al* 2017). Many detergents behave similarly, lysing the cell membrane before adequately permeabilizing the nuclear. Digitonin however has unique properties making it attractive for selective permeabilization. Digitonin binds to cholesterol in order to form pores, which, due to a higher cholesterol content in cellular membranes than intracellular ones, makes digitonin a cellular-membrane specific detergent (Niklas et al 2011). While digitonin still lyses cells at higher concentrations (Holdon and Horton 2009), its specific targeting of cholesterol makes it more controllable than non-ionic detergents such as Triton X-100 and NP-40.

With digitonin providing cellular membrane specific permeabilization, the next step is to target nuclear membrane permeabilization. Liashkovich *et al* reported a novel method of nuclear permeabilization using a molecule called pitstop 2 (Liashkovich et al 2015). Pitstop 2 is a small molecule initially designed to inhibit clathrin-independent endocytosis (Dutta et al 2012; von Kleist et al 2011). In the Liashkovich paper, the researchers showed that pitstop 2 allowed for permeabilization of the nuclear membrane by 70 kDa dextran (FITC labeled). This

permeabilization did not extend to larger molecules however, as a 250 kDa dextran molecule was found to be membrane impermeable even with pitstop 2. This makes pitstop an attractive target, as it creates membrane permeability without making large holes by which cellular material could be lost.

Chapter 2 will discuss experiments around fixation and permeabilization conditions that were used, along with molecular probing tools, in order to probe intranuclear protein concentrations.

2.3 Results and Methods

2.3.1. Pitstop Allows Entry of Small Molecules into the Nucleus

The first step in testing antibodies as intracellular and indeed, intranuclear molecular probes was to recapitulate the findings of Liashkovich *et al*. In their 2015 paper, they used the small molecule (473.36 Da MW) pitstop 2 to permeabilize the nucleus. This had the effect of making the cell nucleus permeable to 70 kDa Dextran-FITC while still being impermeable to 250 kDa Dextran-FITC. Antibodies, the principal probe used in NGS based protein sequencing methods, have a standard molecular weight of around 150 kDa. Since this weight is between the recorded permissible and impermissible sizes, it was important to gauge whether the antibodies would be selectively permeabilized by pitstop 2.

For this experiment, U87-MG cells were selected as a test case. U87-MG is a glioblastomal line that has been well characterized, including being similar to a glioblastomal line that has been used previously for RNA and protein measurements (Darmanis *et al* 2015). For the experiment, a trial antibody was chosen against Activating Transcription Factor 2 (ATF2). ATF2 was chosen since it is a ubiquitous nuclear protein that is prevalent in most cell types, including U87-MG. First, cells were grown to near confluency in a 6 well plate. Then, after removing media cells were washed once with 1X PBS. After washing, cells were incubated with 1 mL 20 ug/mL digitonin (0.002% w/v) in TB for 10 minutes at 37 (TB: Transport Buffer: 110 mM KOAc, 5 mM NaOAc, 2 mM MgOAc, 2 mM DTT, 1 mM EGTA in 20 mM HEPES). After permeabilization, digitonin containing buffer was removed and replaced by 500 uL of TB with 50 uM Pitstop 2 in 0.1% DMSO + 20 ug/mL 70 kDa Dextran-FITC (Fig 9A); 500 uL of TB with 0.1% DMSO + 20 ug/mL 70 kDa Dextran-FITC (Fig 9B); 500 uL of TB with 50 uM Pitstop 2 in 0.1% DMSO + 1:100 diluted anti-ATF2 (Fig 9C); 500 uL of TB with 0.1% DMSO + 1:100 diluted anti-ATF2 (Fig 9D). Cells were

incubated for 15 minutes at 37C in the dark in these pitstop 2 probe-containing buffers. After 15 minutes, a 1:200 dilution of FITC-labeled secondary antibody was added to the anti-ATF2 containing wells following manufacturer recommendations. The samples were incubated another 15 minutes at 37C in the dark before washing twice with 2 mL PBS. For each wash the cells were incubated at 37C for 5 minutes prior to removing the wash buffer. This is because with nuclear staining, it is important to give time for diffusion to occur. After the last wash, cells were suspended in 500 uL PBS with 0.5 ug/mL DAPI. Samples were incubated for 5 minutes at RT in this buffer before imaging.

Results are shown in Figure 9. Dextran samples show pitstop 2-dependent signal. In the pitstop positive sample, nuclear stain DAPI outlines correlate well with large amounts of FITC signal (Figure 9A; outline of white DAPI on left shown in red on right image; FITC in green on right image). In the DMSO only sample, this signal is removed by the washing (Figure 9B). This result is the same as that found in the Liashkovich paper and shows that the pitstop 2 is working as previously reported. When antibodies are used, however, signal does not correlate at all with nuclear stains whether pitstop 2 is used (Figure 9C) or not (Figure 9D). This results suggests that the 150 kDa antibodies are indeed too large to be allowed entry into the nucleus with pitstop 2.

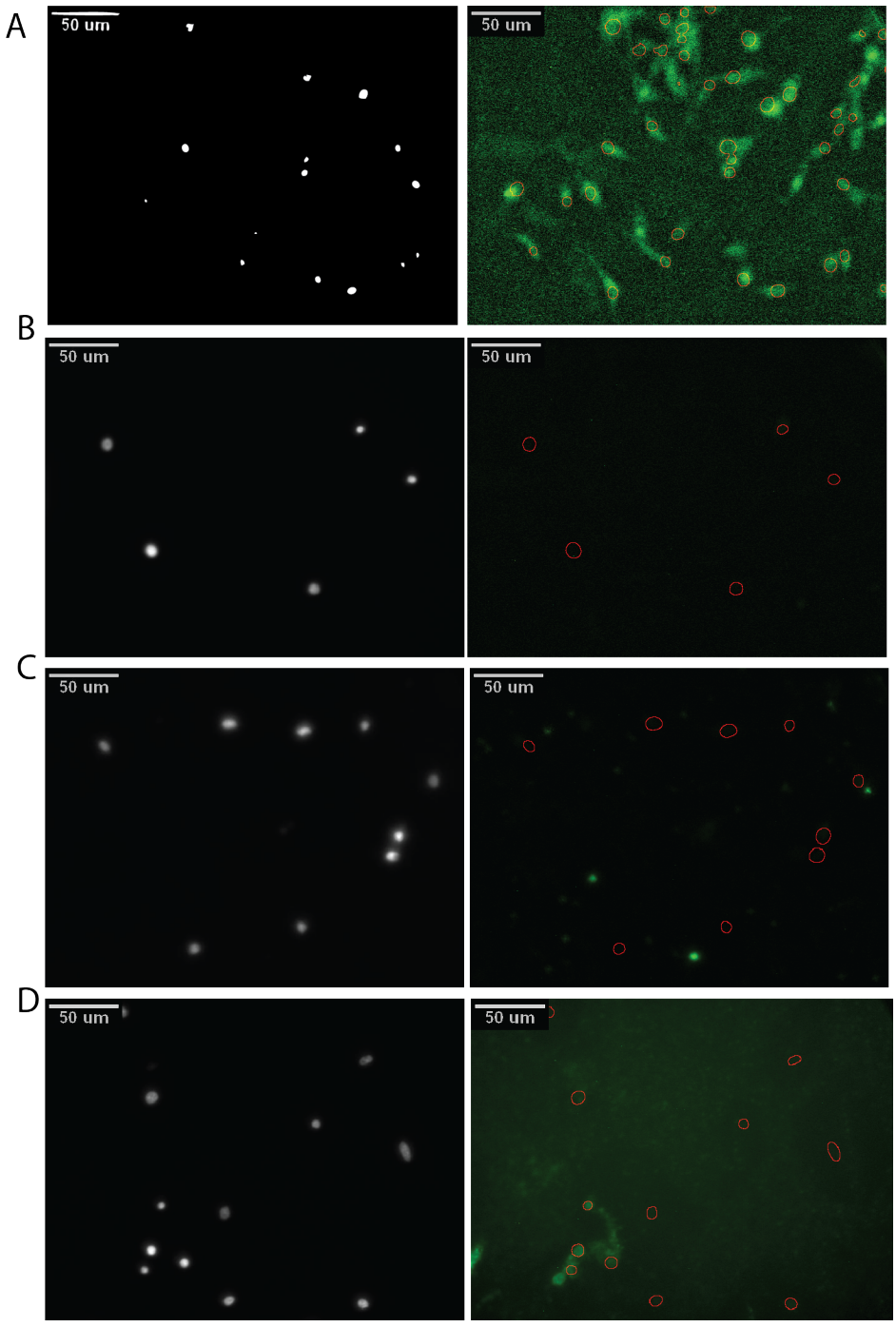

Figure 9: Pitstop 2 selectively permeabilizes the nuclear membrane to small molecules. Images show DAPI channel in white on left image and FITC channel in green on right image. DAPI outline is included in red on FITC channel images. A) 70 kDa Dextran-FITC imaging in presence of 50 uM pitstop 2. B) 70 kDa Dextran-FITC imaging in 0.1% DMSO. C) anti-ATF2 staining in presence of 50 uM pitstop 2. Secondary mouse antibody used for FITC. D) anti-ATF2 staining with secondary mouse FITC imaging in 0.1% DMSO.

2.3.2 Antibody Fabs are Selectively Permeable to the Nucleus with Pitstop 2

Section 2.3.1 showed that full-size antibodies, termed immunoglobulins (Ig) or IgGs (immunoglobulin gamma, the most common variety), are too large to pass through the pitstop 2-permeabilized nuclear membrane. Antibodies do not need to be full size, however, to be functional. Antibodies are constructed by B cells as a response to foreign invaders, typically bacteria and viruses. The antibodies are raised against specific proteins, termed antigens, present in the invaders. All antibodies follow the same basic structure, being composed of two heavy chains and two light chains (Woof and Burton 2004; Janeway et al 2001). These four chains interact in a manner to form a "Y" shape (Fig 10A). The two heavy chains form the backbone (blue), and each contain approximately 450 amino acids. On these heavy chains, approximately ¾ of the amino acids are constant amongst all antibodies from the host. This region is termed the constant region (dark blue). The last 25% is specific to each antigen detected, and is termed the variable region (light blue). Similarly, the light chains (red) are constructed with a constant region (dark red) and variable region (light red). Disulfide bonds link the fragments together (black lines), with two disulfide bonds linking the heavy chain regions together and one disulfide bond linking the light chain constant region to the heavy chain constant region. It is important to note in this structure, that it is the lightly shaded variable regions that contain the antigen binding site, or paratope. For full antigen binding functionality, only this region (both heavy chain part and light chain) must be maintained (Fig 10B).

This structure has not gone unnoticed. Antigen binding fragments (Fabs) have long been used as potential tools in immunobiology (Lewis Carl *et al* 1993; Negoescu *et al* 1994). Antibody Fabs are useful for several reasons. First, Fabs are smaller and are known to have better tissue penetration as a result. Second, the constant region that is removed, termed the crystallizable

fragment or Fc domain, is responsible for much non-specific binding, particularly on immune cells which have specific Fc receptors. Fabs are typically created by the addition of papain (Zhao *et al* 2009), a protease whose cut site is above the disulfide bonds holding the two heavy chains together but without interfering in the Fab structure (Fig 10C). The resulting fragment is 50 kDa, which is smaller than the dextran-FITC used as a positive control.

To test the viability of Fabs as an intranuclear target, the experiment from 2.3.1 was repeated with a few changes. The major change was that only a secondary Fab was used for the antibody. This Fab was purchased premade from a distributor, and distributors only produce secondary Fabs typically. This Fab, against mouse, was labeled with FITC. Since primary antibodies weren't used (as per 2.3.1 they don't enter the nucleus even with pitstop 2), there was no primary antibody target for the Fabs. However, a secondary Fab alone, with no primary antibody target, behaves much like the dextran positive control. Since the main purpose of the experiment was to test Fab entry, the binding ability of the Fab was not necessary. When the experiment was repeated using Fabs instead of primary/secondary antibodies, the dextran controls showed nuclear signal in the presence of pitstop 2 (Figure 10D) and absence of nuclear signal in the DMSO only control (Figure 10E). The Fabs showed the same behavior, with FITC signal overlapping nuclear boundaries when pitstop was present (Figure 10F) but showing no signal in DMSO only controls (Figure 10G). This experiment showed therefore that antibody Fabs are potential molecular probes for intranuclear proteins.

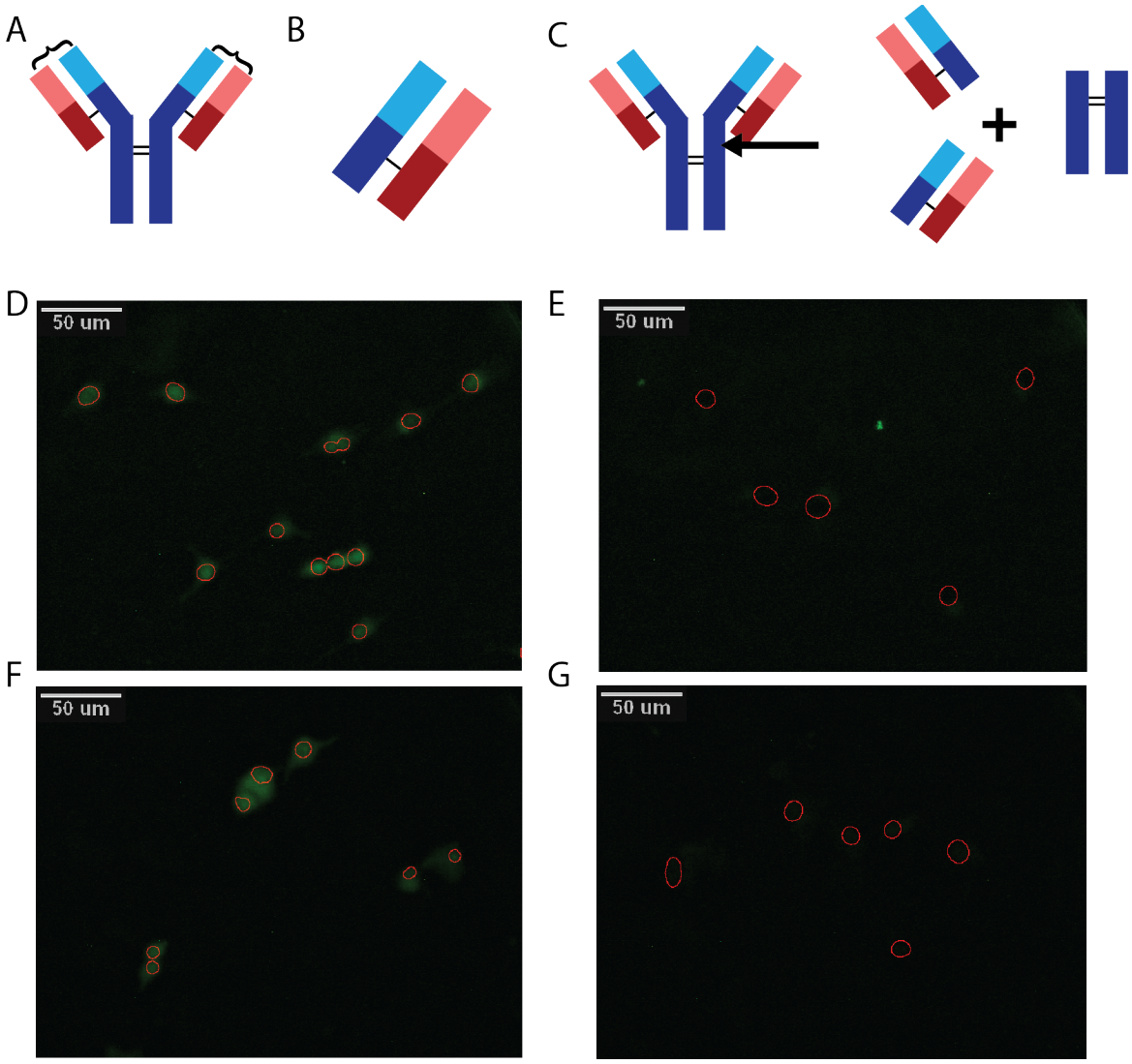

Figure 10: Antibody Fabs are viable molecular probes with pitstop 2 permeabilization of the nucleus. A) Structure of antibody immunoglobulin with two heavy chains (blue) and two light chains (red). Each chain contains a constant region (dark) and variable region (light). Disulfide bonds (black lines) connect the two heavy chains together, as well as the constant regions of the heavy chains and light chains. Brackets show antibody binding region, called the paratope. B) Antigen binding fragment (Fab) close up. C) Papain digestion cuts (black arrow) just above hinge on the heavy chains to produce two Fab fragments and one Fc (crystallizable) fragment. D-G) Images showing pitstop 2 dependent entry of small proteins into the nucleus. Nuclear outlines in red and FITC-dextran (D-E) or FITC labeled Fabs (F-G) signal in green. 70 kDa dextran enters the nucleus in the presence of pitstop (D) and not in DMSO control (E). 50 kDa Fab fragments enter the nucleus in the presence of pitstop (F) and not in DMSO control (G).

2.3.3. Production and Functional Testing of Antibody Fabs

With antibody Fabs as a viable source for molecular probes, it remained to produce antibody Fabs and test their efficacy once conjugated with oligonucleotides. For production of Fabs, there are few companies that produce Fabs of primary antibodies, and fewer that do it in small quantities (~100 ug per antibody) or cheaply. However, there do exist plenty of kits for the production of Fabs. The kit used in this work was the Pierce Fab Micro Prep Kit from ThermoFisher (product #44685). The basic workflow of the kit is outlined in Figure 11A (red outlines).

The protocol starts with a buffer exchange. Papain requires L-cysteine at a concentration of around 20 mM for proper functioning, and the protocol also removes storage buffers from the papain and antibodies. Following buffer exchange the antibodies are incubated with papain, attached to beaded agarose superstructures, at 37C for 5-6 hours. As said earlier, this fractures the full IgG proteins into 2 Fabs and 1 Fc per IgG. The Fcs and Fabs are both 50 kDa, so size exclusion is infeasible. Affinity exclusion works however, as Protein A, a surface protein found in *Staphylococcus aureus,* has been found to bind the Fc domain of most antibodies (Duhamel et al 1979; Hober et al 2007). This is a negative selection, and the Fab fragments are collected in the flow through and 1st wash step. Following collection, Fabs are purified using Amicon Ultra Centrifugal Filter Units with a 30 kDa cutoff. Choice of cutoff is in Appendix to Chapter 2, Figure 1. The filter units are size exclusion based, and are run twice with washes of PBS in order to buffer exchange for the next reaction. After buffer exchange, Fabs are measured in Qubit for a quantification. After this, DBCO-PEG5-NHSEster is added to the Fabs in 20X molar ratio with PBS added to 100 uL. Fabs are incubated for half an hour at room temperature for the NHS Ester reaction. Afterwards, 5 uL 500 mM Tris-HCl is added to quench the reaction, as Tris contains

primary amines that react with unreacted NHS Ester in the buffer. After a 5 minute quench at RT, the Fabs are once again loaded onto a 30 kDa filter unit and buffer exchanged with PBS. This removes unbound DBCO molecules from the solution, which would normally compete with bound DBCO groups for azide oligo addition. Following buffer exchange, azide-conjugated oligos are added at a 3:1 molar ratio to the protein with PBS to 100 uL. Reaction is incubated overnight before a final buffer exchange to PBS and quantification (ssDNA and protein) with Qubit. Additionally, conjugated Fabs are visualized on a gel (Figure 11B).

Finally, it is important to ensure that the Fabs, once conjugated, are still biologically active. To test this conjugated Fabs were produced against BSA. Biotinylated BSA was then bound to superparamagnetic C1 beads coated with streptavidin (ThermoFisher #65001). A second set of samples were prepared with biotinylated Protein A and a third with biotinylated oligos as templates, used as negative controls. After binding, beads were washed with PBS+0.02% Tween-20 to remove unbound BSA. Then 12 pmol of conjugated Fabs were added to each sample at incubated at 37C for 30 minutes to bind. Afterwards beads were washed with PBS+0.02% Tween-20 3 times before being diluted and added to a PCR mix. The PCR tested against the conjugated oligo, and included a positive control of pure conjugated Fabs and a negative control with no template. For both polyclonal and monoclonal antibodies, the samples (n=3 for all conditions) with BSA template showed faster amplification than their respective controls (Figure 11C). As expected, the polyclonal Fabs showed less distance between positive samples and negative controls. This makes sense, as polyclonal mixes contain many isotypes and have generally higher non-specific binding. Still, either type of antibody presents as a viable target for probing protein targets in cells.

The fragmentation and conjugation process has been performed on many types and subtypes of antibodies, including both polyclonal and monoclonal antibodies. Monoclonal antibodies are raised by B lymphocyte clones, and recognize only a single epitope on an antigen. Polyclonal antibodies are typically raised in animal models, where the heterogeneous response from many B lymphocyte lines produces a plurality of antibodies, often corresponding to many epitopes on a single antigen. Monoclonal antibodies are of a specific subclass, such as IgG1, IgG2a, IgG2b, etc., which contain small variations in the constant region and backbone structure of the antibodies (Vidarsson *et al* 2014; Irani *et al* 2015). Polyclonal antibodies typically are mixtures of available types in the host species. Results for several purifications are shown in Figure 11D. Results are given by subtype, as it has been shown each subtype responds differently to papain digestion (Adamczyk *et al* 2000). Although papain digests all fragments, IgG1 has been shown to be the most resistant to digestion and IgG2a the most susceptible. This makes sense as the one IgG2a digestion shows considerably higher digestion than any other subclass. IgG1 conversely shows low yields with relatively low variance, indicating that the class is simply resistant to papain digestion. Polyclonal antibodies appear to have the most variance, which makes sense as their susceptibility to papain digestion will depend on the fractional makeup of different IgG subclasses in a given polyclonal pool. This indicates for future experiments that parameters should likely be tuned for a given antibody subclass.

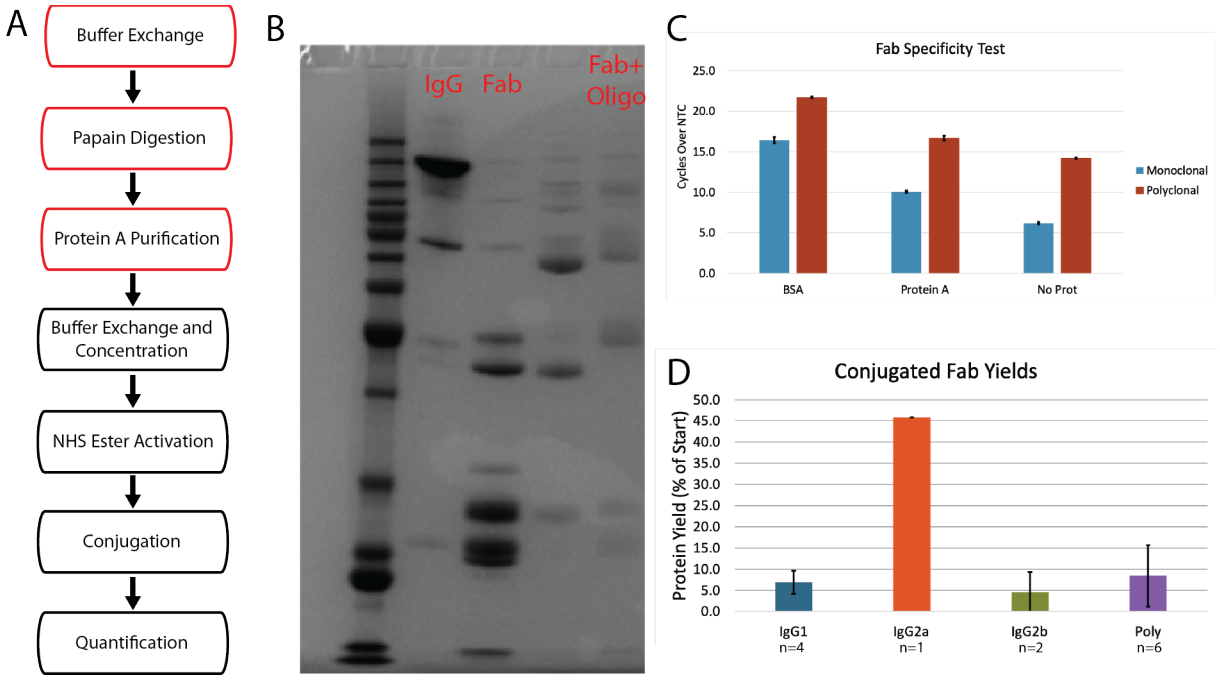

Figure 11: Antibody Fab probe production via papain digestion and conjugation. A) Workflow for Fab production and conjugation. B) SDSPAGE image showing, from left to right, Benchmark Protein Ladder, Whole IgG for anti-U1C antibody, unconjugated aU1C Fab, conjugated Fab for a different protein, conjugated Fab U1C. Note band shifts seen as smears from the unconjugated to conjugated, indicating successful conjugation process. C) Fab activity test using antiBSA model for both monoclonal (blue) and polyclonal (red) antibodies. Results are given as cycles over NTC, with higher numbers indicating greater amount of product. Polyclonal shows higher amounts than monoclonal, but both show significantly higher signal than their control counterparts. D) Yields from Fab production and conjugation process, distinguished by IgG subtype. Error bars indicate sample standard deviations, with number of samples listed.

2.4 Discussion and Conclusions on Chapter 2

Chapter 2 focused on tools for protein detection inside cells. Using imaging experiments, it was shown that the molecule pitstop 2 can be used as a selective permeabilizing agent of the nuclear membrane (Figure 9). Pitstop 2 only permeabilizes the nuclear membrane to small molecules, however, and it was shown in Figure 9E-H that IgG antibodies (150 kDa) are too large to be allowed entry.

Using papain, antibodies can be fragmented into smaller fragments. One of these pieces, the antigen binding fragment (Fab), retains the antigen binding region, or paratope, that allows the antibody to function. By isolating these fragments from the papain reaction and conjugating them to oligonucleotides discussed in Chapter 1, the Fabs provide protein-sensing molecular probes of around 50 kDa. Figure 11 details the process of producing these Fabs. Figure 10 D-G shows that Fabs are small enough to enter the nucleus with pitstop 2 permeabilization. Chapter 2 therefore details antibody Fabs as molecular probes for intranuclear protein detection. Chapter 3 will focus on combining these probes, along with established RNA detection methods, into single cell sequencing of RNA and protein.

2.5 Appendix to Chapter 2

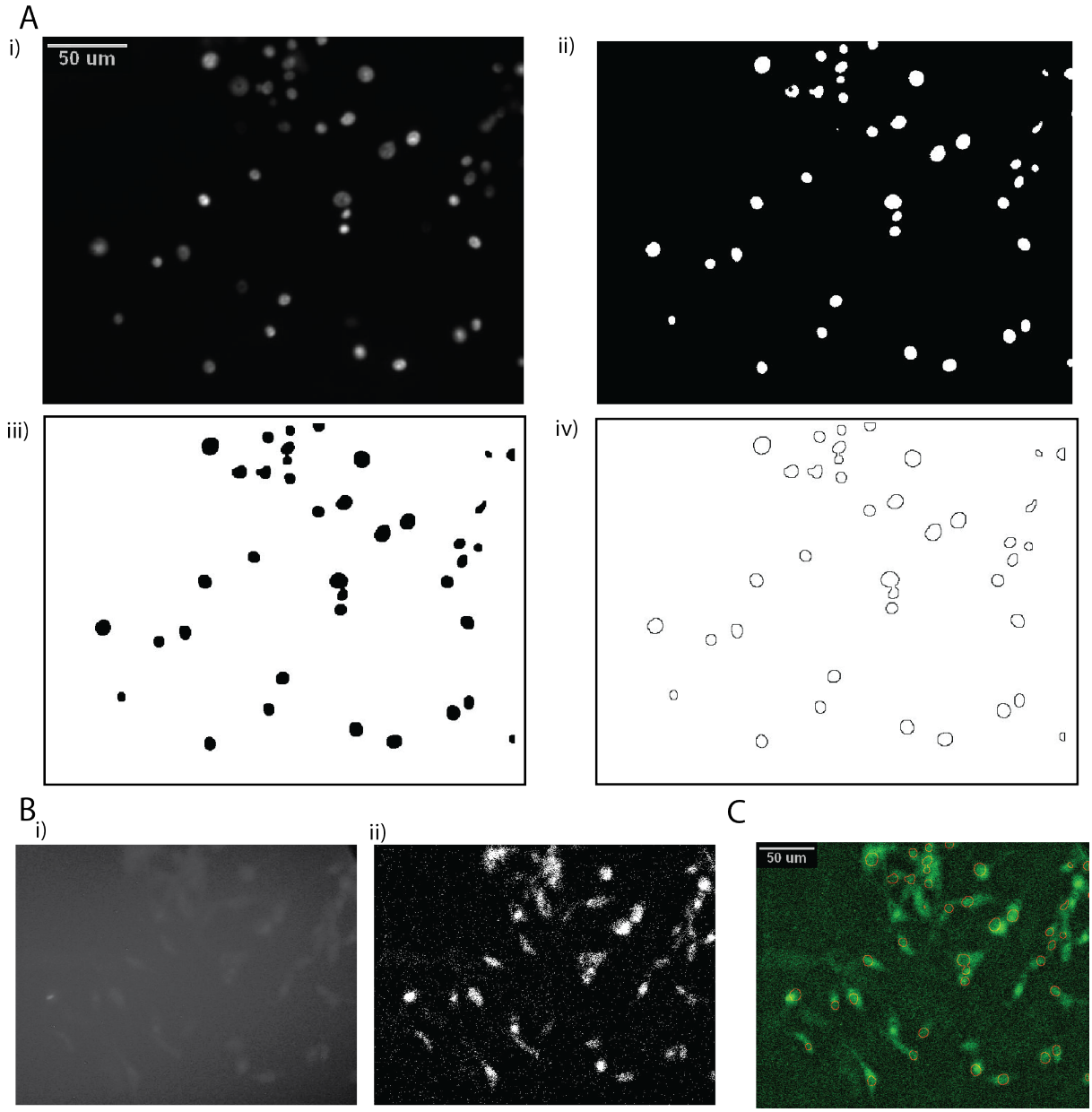

Figure 12: **Image processing in Chapter 2**. A) DAPI image processing. Original DAPI image (i) is first thresholded to make all pixels same intensity (ii). The image is inverted and, when necessary modified to close holes or remove unwanted nuclei connections by erosion/dilation (iii). Finally, an outline is created of each object (iv) for merging with GFP image. B) GFP image processing. Original GFP channel image (i) is background subtracted (ii) then optionally thresholded for clarity. C) Merging of DAPI outline channel (A-iv) and GFP background-subtracted image (B-ii). Colors are added for clarity. Final figure shown is image 9A.

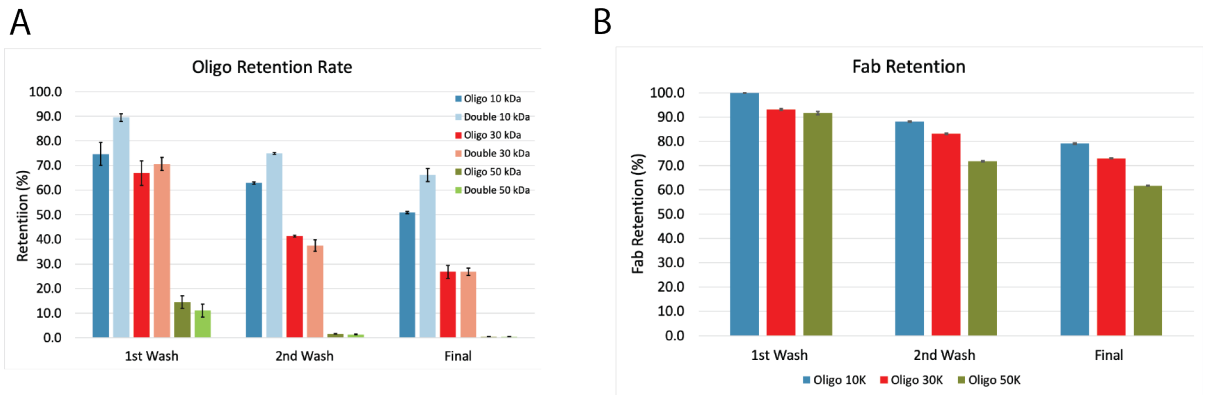

Figure 13: Amicon column retention. Retention percentages for A) oligonucleotides and B) Fab proteins in various Amicon Ultra centrifugal filter columns. Retention experiment was performed for 3 sizes of filter: 10 kDa, 30 kDa, and 50 kDa. Oligo length was 91 bp, same oligos used for antibody conjugation. In A) lower retention percentage is desired, as the experiment simulates free oligonucleotides alone ("oligo" samples) and in the presence of conjugated protein ("double"). In B), greater retention of Fabs are desired. 30 kDa columns and 2 washes condition was selected as having best mix of both qualities.

CHAPTER 3

MEASUREMENTS OF PROTEIN AND RNA IN THOUSANDS OF SINGLE CELLS

3.1 Abstract of Chapter 3

Having developed tools to probe intracellular protein systems, the next step is to produce single cell data sets measuring RNA and protein levels. SPLiTSeq, a 2018 method of RNA Seq libraries produced by four rounds of split-pooling, provides a cost-effective method for producing these libraries. By slight modifications to antibody oligos and the SPLiTSeq protocol, it is possible to convert the experiment into a dual assay. For proof of concept, examining the cell cycle in cultured U87-MG cells was selected. Cultured cells are constantly growing and dividing, and a collection of these cells from any given time point yields cells from every phase of the cell cycle. After several experiments to optimize certain conditions, a proof of concept experiment was run using 8 cell cycle antibodies and several controls. Results from the experiment showed strong correlations between quality of RNA and protein reads, with both data types indicating similar levels of occupied barcodes. These data sets were then used for single cell Seurat analysis, which showed poor individual correlation on the biological level between protein and RNA data. However, an analysis of the RNA differences in two dissimilar protein groups showed a strong correlation with mitotic genes. This suggests that due to intracellular differences in RNA and protein levels the two assay types can act as complementary to one another to yield new biological insights.

3.2 Introduction to Chapter 3

Chapter 1 covered the design of oligonucleotide multiplexing strategies for RNA and protein sequencing in single cells. Chapter 2 discussed the design of antibody Fabs conjugated to oligos as potential targets for intranuclear protein detection. Chapter 3 focuses on RNA and protein single cell sequencing experiments.

With tools for protein detection and strategies for combining RNA and protein methods, it remained to choose a method of RNA probing. Protein oligonucleotides could be tailored to different RNA capture methods, and so it was deemed optimal to adopt previously used RNA methods and alter protein capture to sync with them. Several RNA capture methods for single cells were considered. First, Dropseq (Macosko et al 2015) and Indrop (Klein et al 2015) provide thousands of single cell datasets per experiment using droplet isolation and bead-based barcoding. However, both of these use microfluidic devices that are difficult and costly to setup, and can have large variations within runs. Another method, sci-RNA-seq (Cao et al 2017), uses combinatorial indexing in split-pool reactions but relies on Tn5 transposase for part of its indexing. Tn5 inserts fragments into dsDNA, and upon SDS addition the Tn5 leaves causing double stranded breaks. Tn5 transposase has been shown to have reduced effectiveness at fragments below 200bp (Adey and Shendure 2010; Picelli et al 2014). Since antibody oligos are typically less than 200bp this is problematic. Additionally the antibody oligos are heavily structured and random insertions of barcodes can lead to catastrophic loss of information within the antibody oligo barcodes.

As mentioned in the introduction, two recent methods combined RNA and protein in single cell sequencing; REAPSeq (Peterson *et al* 2017) and CITESeq (Stoeckius *et al* 2017). CITESeq used Dropseq for some of its capture, and both papers used the 10X Genomics platform. Dropseq has been discussed before, and while the 10X Genomics corporation does have a module for RNA

and protein sequencing now, the protocol requires specific antibody barcoding strategies and the use of a 10X Genomics device for library prep. The 10X Genomics device is costly, and its purchase is roughly the equivalent of needing to buy a FACS machine. Alternative strategies were sought that would allow more flexibility in the protocol, as well as eliminate the need to purchase any large equipment or setup besides an Illumina sequencer.

In 2018, a method for RNA sequencing was introduced that used only split-pooling for sequencing tens of thousands of cells (Rosenberg et al 2018). The method relies on barcode introduction from reverse transcription (RT), followed by 2 successive rounds of split-pool ligation. A fourth and final barcode is added during library prep via Illumina indexing. This method is cost effective, as millions of cells can be sequenced from a single experiment. There is no special equipment involved, and since the reactions are based on extension and ligation, there are no special requirements for template length as with sci-RNA-seq. The methods are also similar to the Combolock v4 extension and ligation modalities, building on the same principals and reactions. The major difference is in the extent of the barcoding. Most Combolock protocols used two rounds of barcoding providing 96 unique barcodes each round for a total of 9216 combinations. SPLiTSeq uses 4 rounds of barcoding with a 48x96x96x**X** format, where **X** is the number of Illumina indexes used. The first three rounds provide 442,368 unique barcodes, which allows for much larger inputs of cells than the Combolock protocols intended. This is helpful, as the antibody staining and washing protocols require large amounts of cells to accommodate the many centrifugal steps.

Finally, the work will address the biological significance of working with protein and RNA data by evaluating single cell data sets of cell-cycle protein and RNA markers. The cell cycle is one of the fundamental components of cellular function. Cells begin in Gap 1 (G1) phase, where they take in nutrients and grow. Once the cell grows to large enough size, it enters the Synthesis

(S) phase, where it replicates its DNA. Now containing two copies of every chromosome, the cell proceeds to a second growth phase (G2). When a number of checkpoints are achieved in G2 phase, the cell proceeds through mitosis (M), where it splits itself in half, with each half retaining a copy of every chromosome. After mitosis the cell returns to G1 phase and repeats the cycle (or enters G0 phase, where it stops dividing altogether). Most cultured cells repeat this cycle over and over until they fill up their container, also known as reaching confluency. Depending on the cell line, this cycle can be short (~20 minutes for *E. coli*) or much longer (~1 day for U87-MG) cells. Since the cells do not grow and divide synchronously, by taking a culture of actively growing cells it is expected that a fraction of them will be in each state of the cell cycle. The fraction of cells in each phase roughly corresponds to the relative times that the cell spends in that phase.

Both RNA and protein levels change during the course of the cell cycle (Kowalczyk *et al* 2015; Gookin *et al* 2017). Several methods have already been established in identifying marker genes to assign cells to one phase or another (Scialdone *et al* 2015). In fact several studies have shown underlying expression patterns affected by cell cycle beyond simply the existence of marker genes (Buettner *et al* 2015; Barron *et al* 2016). Cell cycle proteins and RNA were chosen as a model biological system in order to evaluate the potential of the combined data at providing biological insights.

3.3 Methods and Results

3.3.1 Split-pool library preparation with SPLiTSeq

SPLiTSeq barcodes are derived from 4 rounds of successive split-pooling (Figure 14A). Library prep with SPLiTSeq starts with a reverse transcription reaction (Figure 14B-i). In this first round, primers (solid line, top) with poly-dT tails are annealed to mRNA templates' (dashed line) poly-dA tails. Cells fixed with 1.6% formaldehyde are spread out into 48 wells of a 96 well plate, each containing a uniquely barcoded oligo (red section) at final concentration of 2.5 uM. A master mix is prepared and added to the cell/barcode mixture. The final concentration of the reagents is 20 U/uL Maxima H Minus Reverse Transcriptase, 500 uM dNTPs, and 0.25 U/uL each of RNase Inhibitors (RIs) from Enzymatics and Superase from Thermofisher in 1X Reverse Transcriptase buffer. This plate is incubated at 50C for 10 minutes, followed by 3X cycles of 8C for 12s; 15C for 45s; 20C for 45s; 30C for 30s; 42C for 2 min; 50 C for 3 min. Finally the reaction is incubated at 50C for 5 minutes before being held at 4C for the next step.

Following RT, cells are washed once before the first round of ligation. Cells are pooled together into a 15 mL falcon tube and centrifuged for 5 minutes at 500xg at 4C. Afterwards the supernatant is removed and the cells are resuspended in 2 mL of NEB Buffer 3.1 with 20 uL of Enzymatics RNase Inhibitor.

The cells are then ready for the first ligation reaction (Figure 14B-ii). This reaction uses a barcoded oligo (blue oligo, top) that is added to the growing probe. Each barcode oligo has a barcode unique to each well (blue region), as well as adapter sequences on either side (black boxes w/blue outlines) common to all wells. Additionally, a linker oligo is added that bridges adapter regions of the first and second barcode probes (bottom oligo, left). The right depiction shows how these two new oligos fit in with the previously added oligos. For this first ligation, cells are spread

out into a new 96 well plate containing the 96 differently barcoded ligation oligos. The ligation reaction is performed in a final volume per well of 50 uL with 8 U/uL T4 Ligase (NEB), 0.2 mg/mL BSA, 2.4 uM barcode oligo, 2.2 uM linker oligo, and Enzymatics/Superase RIs at 0.32U/uL and 0.05 U/uL respectively. The reaction is performed in 1X T4 ligase buffer. The reaction is incubated at 37C with light shaking on a thermomixer for 30 minutes. Following ligation, a blocking oligo is added to final concentration of 4.4 uM. The blocking oligo is the reverse complement of the round 2 linker oligo in order to remove it from future reactions (buffer is not exchanged before the second ligation). The ligation reaction with blocking solution is incubated another 30 minutes at 37C with light shaking.

Following the first ligation and blocking, the cells are re-pooled in a basin then added to a new plate with the second round of ligation oligos (Figure 14B-iii). The second round ligation oligo (orange, left-top) is set up in the same manner as the round one ligation oligo, with unique barcode region (solid orange) and flanking adapters (orange outline). The second round ligation oligos differ from the first round oligos as they contain a 5' biotin and a UMI that is just upstream (towards 5' end) of the barcode. A linker oligo is again added that bridges the first and second ligation oligos (left, bottom). The right depiction again shows how these oligos fit together with the previous structure. The reaction is performed in the same conditions as before, with another 100 uL of T4 ligase added to the pooled basin before adding to the round 3 plate. Final round 3 oligo concentrations are 2.3 uM for barcode oligo and 2.16 uM for linker oligo. Reaction is again incubated at 37C for 30 minutes with light shaking. A blocking oligo is added, although there is no second incubation as the reactions are immediately pooled and washed.

For the wash, cells are once again pooled into a 15 mL falcon tube and centrifuged for 5 minutes at 1000xg at 4C. The supernatant is removed and cells washed again with 4 mL of 0.1%

Triton X-100 and 0.05 U/uL Superase RI in 1X PBS. Following the wash, cells are resuspended in 50 uL 1X PBS+0.05U/uL Superase RI and counted in a cell counter. At this point, libraries are separated into as many pools as desired, tuning the number of cells to desired number per library. For the experiments in chapter 3, this was most commonly 5000 cells/library. Following cell splitting, 2X lysis buffer is mixed in 50/50 with final cell volume. 1X concentration of the lysis buffer is 200 mM NaCl, 50 mM EDTA, 2.2% w/v SDS in 10 mM Tris-HCl. Proteinase K is then added to 1.8 mg/mL and the whole reaction is incubated at 55C for 2 hours with shaking. After this incubation, libraries are frozen at -80C for storage.

On the second day of the protocol, libraries are removed from -80C and 5 uL of 100 uM phenylmethylsulfonyl fluoride (PMSF) to deactivate the Proteinase K. Libraries are then incubated with 440 ug of My One C1 Streptavidin-Coated magnetic beads (ThermoFisher) for 1 hour at RT. This allows the 5' biotinylated probes to bind to the streptavidin beads. Following binding, streptavidin beads are washed twice with B&W-T buffer (1M NaCl, 500 uM EDTA,0.35 U/uL Superase RI, and 0.05% Tween-20 in 5 mM Tris-HCl) and once with 10 mM Tris+0.05% Tween 20+0.05U/uL Superase RI. Following a nfH2O rinse, beads are incubated in an RT reaction mix with template switching oligo (TSO; Zhu *et al* 2001). The TSO is a special oligo with its 3' end consisting of 3 ribonucleotide guanine residues (Figure 14C). When Reverse Transcriptase performs the initial reaction (first RT) it leaves 3 cytosine residues at the 3' end of the newly created cDNA strand. The TSO oligo contains 3 guanine residues that make the reverse complement to this, and because these residues are ribonucleotides not deoxyribonucleotides, Reverse Transcriptase does not extend the TSO strand but instead extends the template strand. The final reaction mix for the reverse transcription is: 10 U/uL Maxima H Reverse Transcriptase, 2.5 uM TSO oligo, 1 mM dNTPs, 4% Ficoll-PM 400, and 0.45 U/uL Superase RI in 1X Maxima H

RT Buffer. The reaction is incubated for 30 minutes at RT with light agitation, followed by 90 minutes at 42C with light agitation.

After the second RT, the RNA reads now have PCR adapters on both sides, making it possible to do standard PCR amplification. Using Kapa HiFi with 1X concentration and 0.4 uM primers, cDNA is amplified first on the beads for 5 cycles. Afterwards the beads are magnetically pulled down and the PCR reactions added to optical tubes for qPCR. 1X evagreen (ThermoFisher) is used as dye. The cDNA in supernatant is now amplified for around 5-15 cycles, until the qPCR reaction plateaus. Following cDNA amplification the libraries are purified via 0.8X Kapa bead size selection (Kapa product # KK8000). After Kapa purification the cDNA is tagmented and amplified with Illumina Nextera Tagmentation kits via standard protocol. This is where the final (4^{th}) barcodes are added before Illumina sequencing.

The final library is a paired end read (Figure 14D). The first read begins at the 5' end of the cDNA region, which for RNA is dictated by the site of the tagmentation cut. This read is 66 bp long. The second read contains all of the barcodes. The first base read is the first base of the UMI, going through the barcodes in reverse order. The read is 94 bp, finishing on the last base of the round 1 barcode.

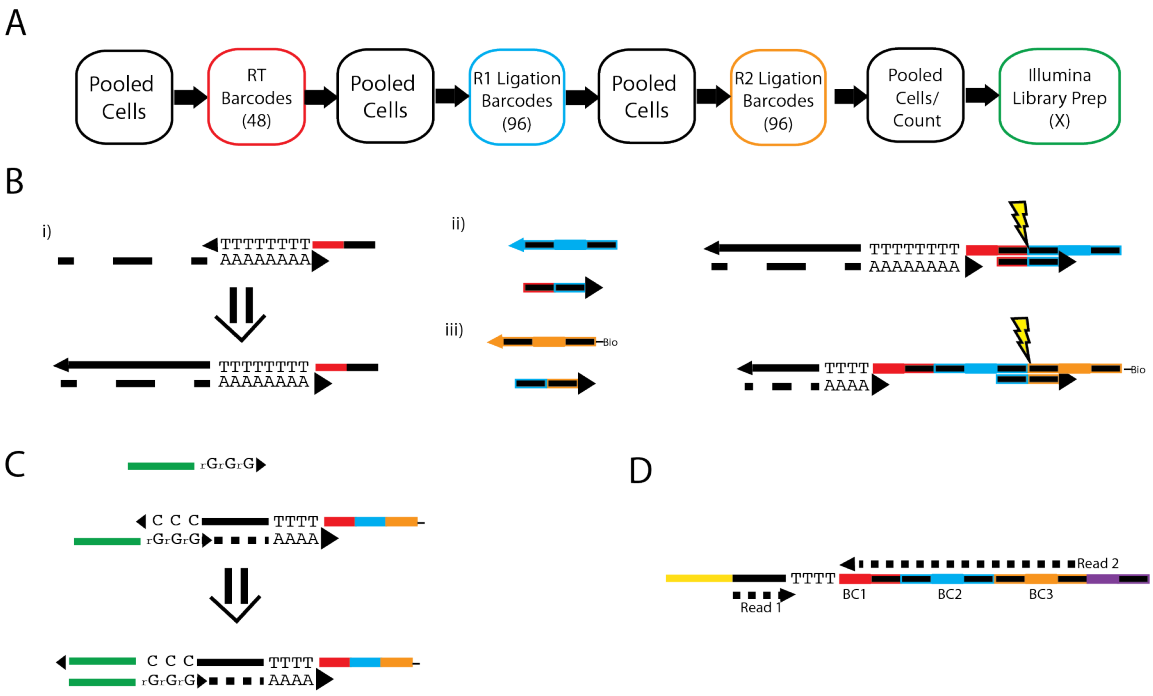

Figure 14: SPLiTSeq protocol overview. A) The SPLiTSeq protocol workflow diagram. The protocol uses four rounds of split-pooling, starting with reverse transcription, then two rounds of ligation before a final pool and split into libraries for Illumina sequencing. B) SPLiTSeq protocol as view at the individual oligo level. (i) Round 1 barcode (red) poly dT region anneals to the polyA tail of an mRNA template (dashed line). Reverse transcriptase produces cDNA of the mRNA template. Arrow indicates 5'->3' direction. (ii) Ligation reaction in the first round. Ligated barcode oligo (blue, top) and linker oligo (red/blue outline, bottom) are added to reaction mix and form a complex with round one oligo adapters. T4 ligase operates at the red/blue junction (lightning bolt) (iii) Second round ligation adds barcoded oligo (orange, top) and linker oligo (blue/orange outline, bottom) form a complex with the 5' adapter from the round 2 oligo (right). T4 ligase operates at the red/blue junction (lightning bolt). C) Template switch reaction. Template Switch Oligo (TSO) contains 3 rG residues on 3' end. When bound to the 3' end of the cDNA, the ribonucleotides cause Reverse Transcriptase to extend along the cDNA. D) SPLiTSeq sequencing read. After Illumina adapters (yellow, purple) have been added by tagmentation and PCR, read is sequenced on an Illumina sequencing platform. Read 1 of paired-end read is 66bp going from the 5' end of the cDNA towards the 3' end. Read 2, 94 bp, starts at the UMI immediately upstream of the round 3 (ligation 2) barcode and proceeds until picking up the round 1 (reverse transcription) barcode.

3.3.2 Protein Oligonucleotide Design for SPLiTSeq Compatibility

Design of the antibody oligonucleotides was modified in order to suit the needs of SPLiTSeq (Figure 15A). The final oligo is constructed by the ligation of two different oligos. The first oligo contains the 5' azide chemical group necessary for protein conjugation (black and grey oligo, top). The second oligo contains the barcode used to identify the antibody used (white box) and a 3' polyA tail. The dual oligo system is useful in allowing easy ordering from IDT, as azide-modified oligos may take several weeks but phosphorylated oligos take only a few days, or even one. Therefore switching out barcodes or ordering new barcodes can be done quickly as long as a stock of the constant azide-modified oligo is kept on hand.

The 3' polyA tail on the antibody molecule causes the protein oligos to be picked up in the same capture method as the mRNA. The antibodies undergo the same procedures as discussed in the previous section until cDNA amplification. There a primer is added at 0.04 uM that targets the region just upstream of the barcode (grey box in Figure 15A). After cDNA amplification, the reaction is purified by a 1.8X Kapa pure bead process, and the sample split in order to amplify the cDNA and antibody oligos separately. The cDNA is processed with 0.8X Kapa purification as the antibody oligos are amplified on their own with primers targeting the round 3 oligo adapter and adapter region mentioned earlier (Figure 15B). Therefore the protein capture is the same as the RNA, but its amplification is treated separately. Since capture is the same manner for both molecules, it makes sense for the protein tagging to be done before entering in to the SPLiTSeq protocol.

For antibody tagging, U87-MG cells and 3T3 cells were grown in culture under standard conditions (MEM and DMEM high glucose, respectively). Once cells reached confluency, they

were incubated with Trypsin LE Express until detached. Cells were pelleted, washed once with PBS, then stored in RNALater storage buffer (ThermoFisher).

The antibody tagging workflow starting on the day of the experiment is covered in Figure 15C. First the cells were counted in a cell counter and mixed in a 1:1 ratio in a 15 mL falcon tube, 1 million cells each. 1X PBS + 0.05 U/uL Superase RI was added in a 1:1 volume with the RNALater to reduce viscosity. Cells were then pelleted by centrifugation at 1000xg for 10 minutes at 4C. Following centrifugation, the supernatant was aspirated out and replaced with 2 mL permeabilization buffer, which is Transport Buffer (TB) with 0.005% digitonin and 30 uM pitstop 2 in 0.5% DMSO. Cells were incubated in permeabilization buffer for 10 minutes at RT before being centrifuged at 500xg for 5 minutes at 4C. Permeabilization buffer was removed and replaced with 3 mL chilled 1.6% formaldehyde in 1X PBS. Cells were incubated in fixation buffer for 10 minutes on ice.

Following fixation, cells were pelleted by centrifugation at 500xg for 5 minutes at 4C before removal of supernatant. Cells were washed with 4 mL 1X PBS+0.05 U/uL Superase RI+0.05% Tween 20 before resuspending in 200 uL PBS+RI. 10 uL Fc Block (Biolegend) was added to solution and cells were incubated in blocking solution for 10 minutes on ice. Afterwards, antibodies, premixed together in 600 uL 1X PBS were added to the final solution. Hybridization was performed at RT for 30 minutes, then transferred to ice for 30 minutes.

After hybridization cells were pelleted by centrifugation at 500xg for 5 minutes at 4C and washed with 4 mL Wash Buffer (WB: 0.5% BSA and 0.05% Tween-20 in 1X PBS) before being replaced with 1.6% formaldehyde in 1X PBS. After a 10 minute fixation on ice, cells were washed twice more with WB before being counted. Counted cells were resuspended in appropriate volume

of PBS+0.1 U/uL Enzymatics RI before being added to first step of SPLiTSeq (See Figure 14A for SPLiTSeq protocol).

One more factor must be addressed for adaptation of the protein oligos to the SPLiTSeq protocol. Reverse Transcriptase acts as a 5'->3' polymerase that can generate DNA from an RNA template as its main function, but it has also been noted to generate DNA from a DNA template (Gerard *et al* 2002). This DNA-directed DNA synthesis is the function being used for the antibody oligo extension in the RT step of SPLiTSeq. However, while the RNA bases on the mRNA template will prevent Reverse Transcriptase from extending the mRNA template past the polyA tail (Figure 16D-i), there is no such inhibition for DNA templates. If this extension proceeds, the single-stranded overhang left after extension will not be present (Figure 16 D-ii). If there is no overhang, there will be no ssDNA available for the first round of ligation. Therefore, the 3' end of the antibody oligos was given a phosphate group. 5' phosphates are necessary for ligation reactions, but 3' phosphates on oligos prevent extension by polymerases.

To test this, a template oligo was produced with a 5' biotin instead of the 5' azide. Template oligos were bound to streptavidin-coated magnetic beads and then went through all the steps of SPLiTSeq, with washes performed using 1X PBS+0.05%Tween-20 instead of centrifugation steps. After the final oligo was ligated and the solutions washed, the oligos underwent PCR with primers targeting the same regions used for Illumina adapter addition (Figure 15B). For the experiment, 3 conditions were used: the first using 3' phosphorylated templates, the second using unphosphorylated templates, and the third with no template at all. Results of the PCR amplification are shown in Figure 15E. The 3' phosphorylated samples amplified several cycles faster than the unphosphorylated samples, which amplified at the same time as the no template control. The gel image shows that while the unphosphorylated oligos do have some product (285bp), the signal is

considerably diminished compared to the 3’ phosphorylated templates. Thus it can be concluded that 3’ phosphorylated antibody oligos should be used in order to better mimic the mRNA templates expected in the SPLiTSeq design.

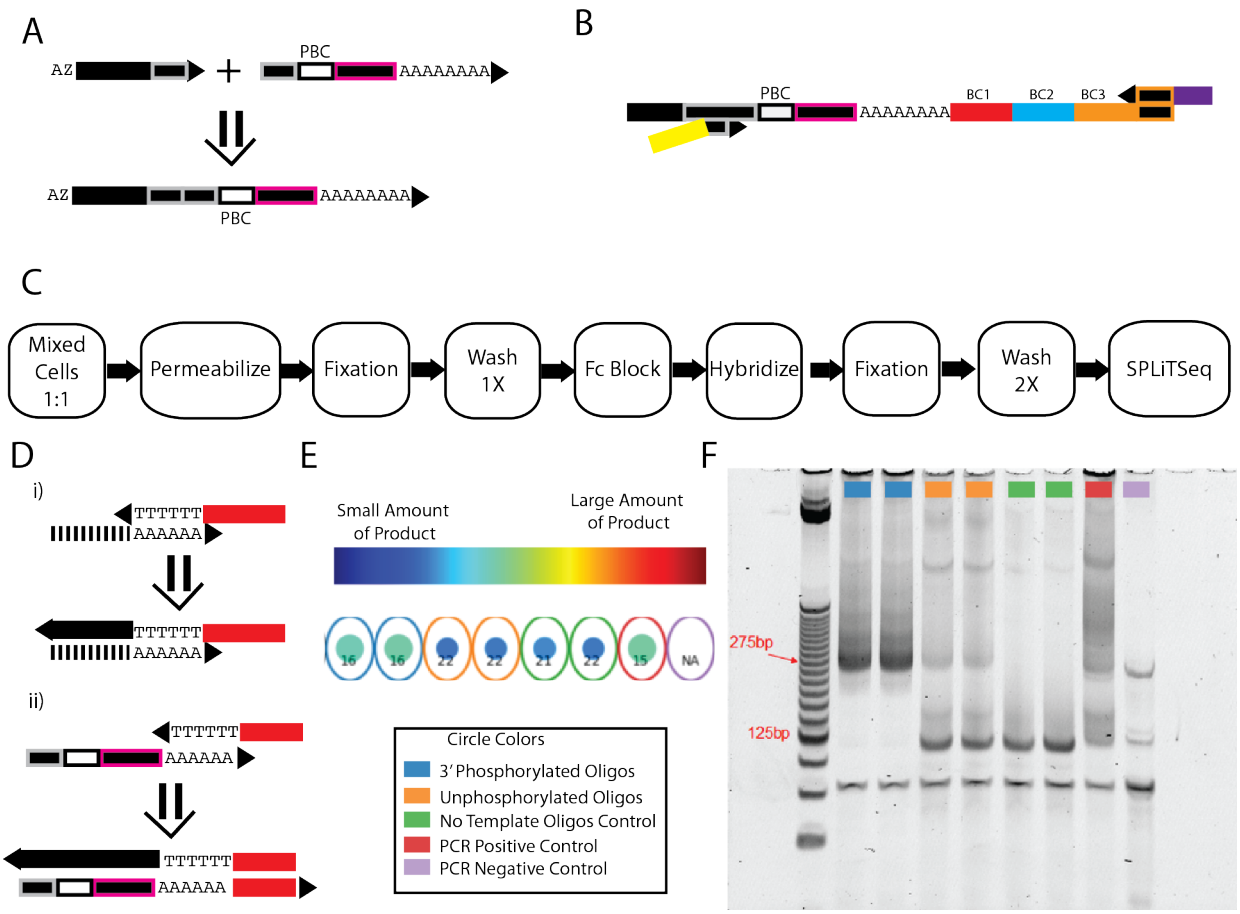

Figure 15: Protein oligo adaptations for SPLiTSeq. A) Construction of the protein oligos using a two oligo ligation system prior to conjugation. Azide conjugated oligos This reduces lag time in adding new barcodes. Upstream oligo (left) contains azide moiety and downstream oligo (right) contains protein barcode (PBC) and polyA tail. B) Final oligo PCR for Illumina adapter addition after SPLiTSeq. The protein oligo undergoes the same reactions as mRNA until the cDNA amplification step, where PCR primers target the adapter just upstream of (grey box oligo) the protein barcode to add on the P5 adapter (yellow). The P7 adapter (purple) is added with the same barcode used for P7 addition to cDNA. C) Antibody tagging protocol used before SPLiTSeq. Last step of this workflow "SPLiTSeq" corresponds to the first step "Pooled Cells" in Figure 14A. D) mRNA extension differs from extension on DNA template antibody oligos. The inability of Reverse Transcriptase to extend mRNA (i) is not a problem with DNA templates, leading to double-stranded extension (ii). This reaction can interfere with the SPLiTSeq protocol. E) PCR amplification (same primers as B) for antibody oligos with and without 3' phosphorylation. Size and color of interior circles corresponds to quickness of amplification; CT value listed in the circle. Exterior circle color indicates sample as given in the legend. F) Gel image of products from (E). Samples use the same legend. 3' phosphorylated samples (blue) show large amounts of correct product at 285bp over the unphosphorylated (orange) and No template samples (green).

3.3.3 Optimizing Conditions in Dual Omics SPLiTSeq

The previous section discussed modifications made to the antibody oligos for adaptation into the SPLiTSeq protocol. This section details different conditions tested and their effects on the yields of protein and RNA in SPLiTSeq. Since SPLiTSeq has previously been tested for RNA performance, most experiments were designed with the intention of ascertaining whether the changes improved protein capture quality, without sacrificing RNA quality.

The first test involved the fixation and wash conditions in the antibody protocol (Figure 15C). Two variables were tested: the number of fixations and the number of washes. To test the number of fixations, samples were either fixed after permeabilization and after hybridization, or only after hybridization. Fixation is useful in that it stops many intracellular processes, importantly the ones that contribute to RNA degradation. However, as the name implies, fixation also freezes cells, and may affect protein binding (Vani *et al* 2006). Additionally, Liashkovich *et. al* hypothesized that the pitstop 2 molecule creates nuclear permeability by deforming nuclear pore proteins. If the proteins are fixed, this may result in pitstop 2 not affecting them. This is why the permeabilization step was placed before the first fixation. The second fixation is used to bind antibodies to their epitopes in a more permanent fashion. This is a common technique used with other methods (e.g. Chromatin Immunoprecipitation), as formaldehyde crosslinks are only created between molecules in close proximity to one another, such as epitope-paratope binding (Hoffman *et al* 2015). Additional washes provide more stringency but at the cost of potentially removing some correctly bound antibodies. Therefore a test was performed to inform on the characteristic losses and gains associated with additional wash steps.

For the experiment, four samples were prepared, each starting with 1 M U87-MG cells and 1M 3T3 cells. The samples were run through the protocol in Figure 15C, with some steps removed

for each sample. The overall tally was as follows: Sample 1 received one wash after hybridization/fixation and was only fixed after hybridization; Sample 2 received two washes after hybridization/fixation and was only fixed after hybridization; Sample 3 was fixed before and after hybridization and received one wash after the second fixation; Sample 4 was fixed before and after hybridization and received two washes after the second fixation. After all samples were washed and counted at the end of the antibody protocol, cells were pooled into the RT step such that each sample was only in one row (i.e. Sample 1 in row "A", Sample 2 in row "B", etc). Each sample was normalized so that approximately the same number of cells was loaded in each row. After library prep and sequencing, the samples could be traced back to their initial row for comparison. Results are shown in Figure 16A for RNA (i) and protein (ii). Results are expressed as a heat map, with colors corresponding to the log2 fold expression over the expected value of fraction of barcodes for that round. The expected value for a given round is 1(/# barcodes), or the fraction of reads that would come from that well if all barcodes were distributed evenly. That is:

$$f = \frac{\#\ reads\ from\ well\ i}{\sum_i reads\ from\ well\ i}$$

$$fold\ expression = \log_2\left(\frac{f}{Expected\ value\ of\ f = \frac{1}{\#\ barcodes}}\right)$$

To compare conditions, it is best to examine groups of rows against one another. To compare fixations (1 fixation vs. 2), we look at rows A & B (1 fixation) against rows C & D (2 fixations). For the RNA, there is a clear favorability for the 2 fixation case (student's t-test p-value = 1.5E-9). In proteins unfortunately, the opposite is true, although to a weaker extent (student's t-test p-value=0.021). To examine washing conditions, compare rows A & C (1 wash) against rows B & D (2 washes). In both cases, the condition is inconclusive (p-value 0.45 for RNA and 0.55 for protein). In that case, it is not a significant loss to include more washing.

It should also be expanded upon that different formaldehyde concentrations were tested. 1.6% formaldehyde is a common percentage used for cell analysis. Lower formaldehyde concentrations will reduce the number of crosslinks in a given cell, potentially creating larger pores and interfering less with protein structure. However, weakly fixed cells may lead to loss of intracellular material and inability to withstand the repeated wash steps in the protein hybridization/SPLiTSeq combined protocol. Figure 16B shows the amount of cell retention across the SPLiTSeq experiments with good libraries at the end (1, 2, and 4 did not have viable libraries).

To test the effect of formaldehyde concentration more directly, cells were grown in a 6-well plate until near confluency, then subjected, while still adhered to the dish) to the antibody protocol as shown in Figure 15C. Once again, the number of fixations were tested, as well as switching the permeabilization and 1st fixation step. Anti-CycD1, a common transcription marker especially prevalent in G1 cells was used. Before antibodies were hybridized, a complementary oligo with a Cy3 dye on the 5' end was annealed to the antibody-conjugated oligo. This allowed for visualization of the hybridized antibody. In addition to antibody staining, cells were counted via looking at several windows for each plate and counting the number of cells present in each window. The results are shown in Figure 16C. Along the X-Axis is normalized cell count, expressed as a fraction of the maximum number of cells counted in one sample (57). Along the Y-axis is the fraction of cells in that sample that had yellow stain, indicating a presence of CycD1. The results show that the double fixation with fixation after permeabilization (red and light red) have the best mix of cell viability and CycD1 shading. Although the most cells are viable when fixed before permeabilization (green and light green), neither sample had any staining visible from the antibodies. The single fixation samples (blue and light blue), in keeping with the results of 14A, show that 1 fixation has reduced number of cells and antibody signal when compared to their

double fixation counterparts. Finally, the 1.6% formaldehyde samples (dark) appear to produce more viable cells than their 0.5% counterparts (light), except in the single fixation case, where they are similar.

Another important improvement was in the use of protein-specific barcodes. The oligonucleotides conjugated to the antibodies specifically contain a poly-dA tail in order to be captured by the SPLiTSeq round 1 barcode oligos. However, this means that the two capture methods are competing. Instead, new oligos were developed that targeted an adapter upstream of the poly-dA tail (Figure 15A; pink box). This adapter is still downstream of the protein barcode and can therefore capture all the information the poly-dA capture does. A plate of 48 different primers were created that were identical to the SPLiTSeq round 1 primers in every way, except for substituting the dT region with the reverse complement of the adapter sequence. To test the efficacy of the new barcodes, a SPLiTSeq experiment was performed where the first two rows of the RT plate contained both poly-dT primers and the new adapter-targeting primers, each 12.5 uM. In the last two rows, 25 uM poly-dT primers were used. Note that the original SPLiTSeq protocol uses 12.5 uM each for poly-dT primers and N6 primers, or random hexamers. Random hexamers are potentially a way to gain greater diversity in capture, but because of the structured nature of the antibody oligos, these were deemed to inaccurate for inclusion.

Results are shown in Figure 17A for both RNA (i) and protein (ii). As expected, the protein primers show significant increase between inclusion (rows A & B) and absence (rows C & D) of protein-targeting primers (students t-test p-value 0.002). Interestingly, the RNA showed a similar increase in expression, though less pronounced (students t-test p-value 0.012).

One final implementation was used, this time adding 25X the antibodies (5 ug) per antibody. This experiment will be talked about in depth in the later sections, but it is noted here as

an improvement in the protocol. The increase was added in order to improve the probe/target ratio. Transcription factor (TF) estimates vary widely by cell type and by transcription factor (Biggin 2011), reasonably ranging from thousands to hundreds of thousands of cells. Taking a middle range of tens of thousands (10^4), we multiply that by a million (human) cells per experiment, or (10^{10}) potential targets. Previous SPLiTSeq experiments used ~300 ng of protein per antibody (6000 pmol or ~10^{15} molecules), giving a probe to target ratio of 10^4 for high copy TFs and 10^6 for low copy. The SPLiT9 experiment used 5 ug per protein (~10^{16} molecules), which increases this range to 10^5-10^7.

With all these trends in mind, it can be seen that the scale of the SPLiTSeq protein experiments has improved over the course of the proof-of-concept runs (Figure 17B). These data points show the number of human cells (as defined by RNA data, see section 3.3.4), Early experiments show <10 cells with unique protein reads totaling <100, which is greatly expanded on in subsequent runs. Most recently, the increase in ~25X antibodies added during the SPLiT9 experiment added outsized gains, resulting in a ~25X fold increase in number of human cells with protein and a 100-fold increase in the total unique proteins observed.

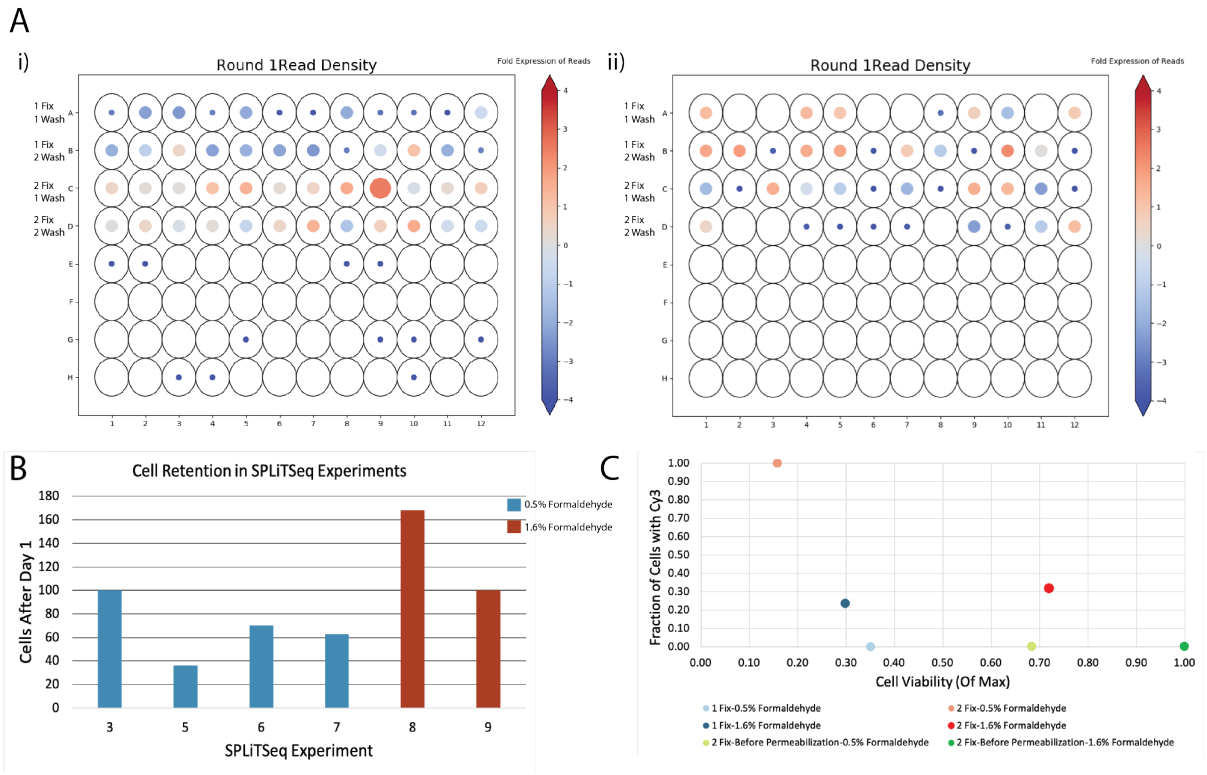

Figure 16: Washing and fixation conditions for combined protein/RNA SPLiTSeq. A) Barcode density plots from round 1 of SPLiTseq showing RNA (i) and protein (ii) read densities for the R1 plate. Heatmap shows log2 values of reads containing that barcode divided by the expected value of a uniform distribution (e.g. 1/48~2.08% for round 1). Each row corresponds to a different condition as labeled on the left. Barcode densities show greater RNA in 2 fix over 1 fix (t-test p-value=2E-9) vs. a slight dropoff for protein (t-test p-value=0.028). Note: small circles in lower half of the plate indicate misalignments, since these reads should not be possible. This was later fixed by barcode correction as described in Section 3.3.4. B) Formaldehyde concentration effects on cell retention. Estimated number of cells (in thousands) at the end of day 1 of the protocol. All experiments started with 2M cells except for SPLiT5, although it was normalized to 1M cells after antibody steps, a common input from the other experiments. C) Cell counts (as fraction of max) and Cy3+ cell counts from formaldehyde experiment. 2 fixations seems to provide best compromise of Cy3 signal and cell viability (number of cells). 1.6% formaldehyde samples show higher or equal viability with their 0.5% counterparts.

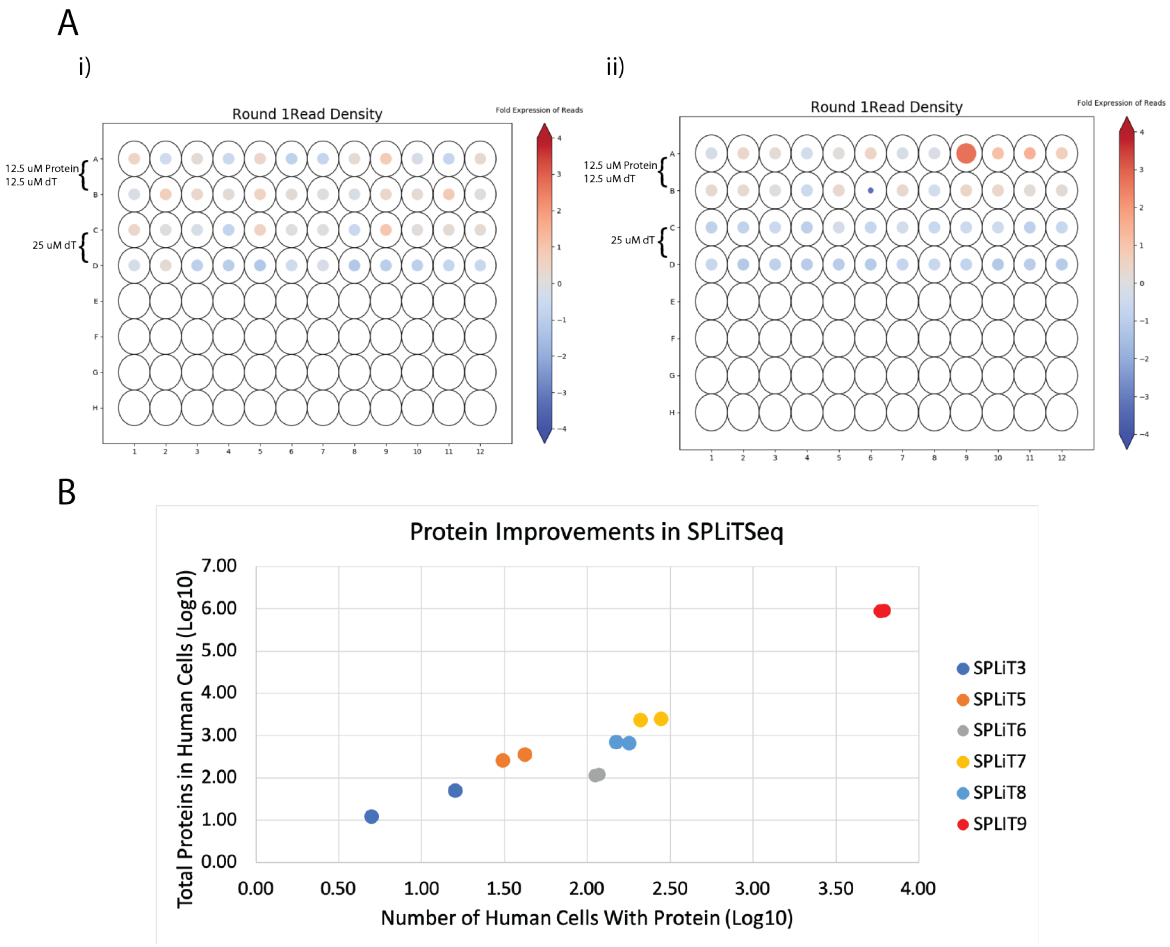

Figure 17: Improvements in protein capture over proof of concept experiments. A) Barcode density plots from round 1 of SPLiTseq showing RNA (i) and protein (ii) read densities for the R1 plate. Heatmap shows log2 values of reads containing that barcode divided by the expected value of a uniform distribution (e.g. 1/48~2.08% for round 1). Rows A & B correspond to mixed protein/dT primers, whereas rows C & D correspond to rows to dT primers only. Both RNA and protein show statistically significant raises in the expression in the mixed primers case (t-test p-value 0.02 for RNA and 0.002 for protein). B) Improvements in protein capture over SPLiTSeq experiments. On a log10 scale, number of human barcodes with protein reads against number of total protein reads found. Number of cells has improved 3 orders of magnitude and number of total unique protein reads has improved 6 orders of magnitude.

3.3.4 Computational Analysis

The overall data analysis pipeline is illustrated in Figure 18A. Data analysis begins with cell demultiplexing using bcl2fastq (Illumina). This program reads Illumina image files and converts them to the fastq format. Recall that the Illumina sequencing of SPLiTSeq reads is a paired end 66+6+94 (66bp in read 1, 6 bp index, and 94 bp read 2). Read 1 is composed of the cDNA for mRNA or the protein oligo and read 2 contains 3 cell barcodes and a UMI with interspersed adapter regions.

Once in fastq, reads are fed into a custom python script for read-splitting if the cDNA reads and antibody reads are in the same indexes. This script looks for adapter sequences present in the antibody read 1 file and separates those into a separate fastq than those without the adapter sequence, which are by default deemed RNA reads. The program also separates the cell barcodes and UMIs out from the read 2 file, adding them to the read name of the read 1 file for later use. All barcodes are also diverted to new fastq files that only have the 8 bp barcode (Figure 18B). This makes 3 different barcode reads for each RNA fastq read and 4 different barcode files for each protein read, with the additional file for the protein barcode. These 8 bp read fastq files are analyzed by deindexer (https://github.com/ws6/deindexer/), a custom script for aligning index sequences as an alternative to bcl2fastq. This script has advantages over standard aligner programs, in that it is designed for short sequence alignment. The output of deindexer is new fastq files, one for each index (i.e. 48 files for round 1 each corresponding to a well, 96 for rounds 2 and 3). These fastq files are then read and the read names stitched together across the 3 cell barcode files. The final step is to compare these new corrected barcodes to the original fastq barcodes (in the readname), and replace the originals with corrected where necessary. For protein this process also adjusts the protein barcodes, and outputs the result not as a fastq but a SAM file (Li et. al 2009).

The effects of barcode correction can be seen in Figure 18C. For both protein and RNA, it is not uncommon for at least one barcode in the 3 8bp regions found in read 2 to need correction (around 8%). RNA, however, has a higher amount of "unaligned" reads, which are reads where one or more of the cell barcode components cannot align to any of the indexes. This may be due to the larger amount of processing that goes on in RNA reads. Overall these reads are not especially important, however, as they are typically unaligned during STAR alignment if uncorrected (data not shown). Note also that the modified barcodes are still aligned reads, they only need barcodes corrected to

Following barcode correction, the RNA is now aligned using STAR (Dobin *et al* 2013). Since all experiments mentioned in this work contained both mouse and human cells, cells were aligned to a combined genome of mouse and human reads. After alignment, barcodes and UMIs are added (via custom script) to the end fields of the SAM format using "XC:Z:CellBarcode" and "XM:Z:UMI" notation, as described in SAM formatting. Once this is done, the bam files are then input through the Drop-seq analysis pipeline (Macosko et al 2015) which is a subset of Picardtools from the Broad Institute (https://broadinstitute.github.io/picard/), using the functions "TagReadWithGene", "DigitalExpression", and "BAMTagHistogram". These scripts (in order) tag the alignments to specific genes, counts #UMIs (in different cell barcodes) present for each gene, and finally gets a read summary by cell barcode. The main output of "DigitalExpression" is a large tab separated matrix of n genes x m cells. This table can be used as an input to Seurat Analysis (Butler *et al* 2018).

Seurat is a data analysis platform in R that specializes in single cell data sets. Seurat can perform normalization and scaling on data sets and then perform differential analysis, and is useful in data visualization. For these data sets, Seurat first normalizes the data by dividing individual

gene counts by the total expression and multiplying it by a scale factor before taking the log of this value. After this normalization and scaling, data can be subjected to differential analysis, which highlights differentially expressed genes among populations of cells. More Seurat applications will be discussed in the next sections dealing with data.

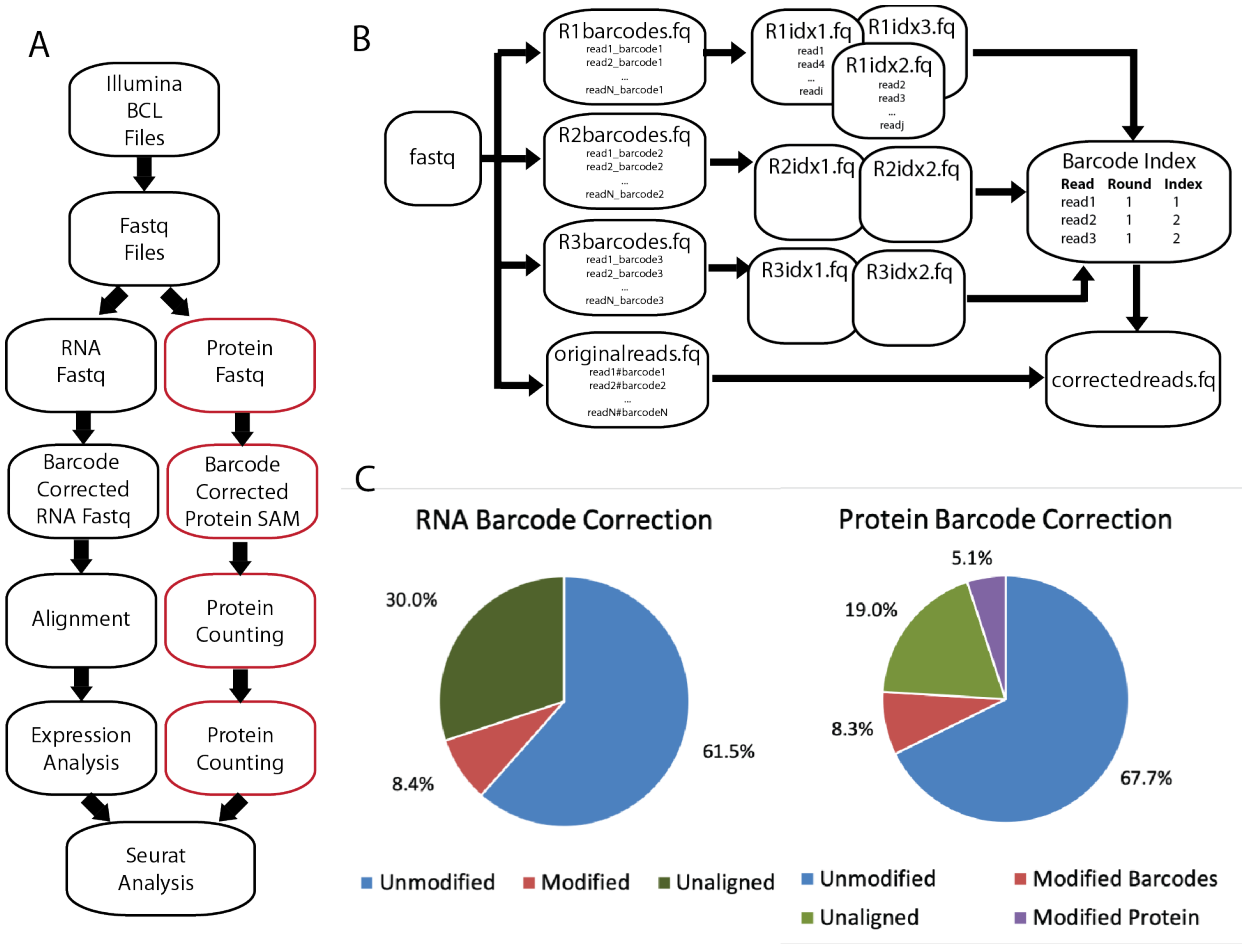

Figure 18: Data analysis pipeline in combined RNA/protein data. A) Pipeline for sequencing read analysis. Illumina bcl captures are converted to fastq format then split into RNA/protein reads. Both types of read go through barcode correction before alignment and counting/expression analysis. Afterwards cell count matrices are input to Seurat for normalization, variable expression, and visualization. B) Steps of barcode direction. Fastq read containing cell barcodes (read2) are extracted and added to names of reads in originalreads.fq. Meanwhile, 3 separate barcode files are created each listing the barcode sequence for a given round for that read. Those are aligned using deindexer, which creates fastq files for every index available (i.e. 48,96,96 for SPLiTSeq). These files contain a list of all reads that aligned to that index. That information is pooled in a table, which is reassembled and used to correct original barcodes. C) Statistics for barcode correction in RNA and protein reads. Modified reads contained at least one modified barcode, protein modified reads were modified in their protein barcode region. Unaligned reads refers to a condition where at least 1 of the 3 round barcodes failed to align to any index.

3.3.5 Data Quality Metrics in RNA and Protein Single Cell Data

In single cell data sets that use barcoding schemes it is important to show that the data in individual barcodes are indeed coming from single cells. Species mixing, first introduced by Macosko *et al*, has become a common method for showing quality of single-cell data. Reads from each cell barcode can be separated into mouse and human reads based on the results from the dual alignment step. This can best be viewed in a species mixing plot, where each dot represents a cell barcode and the x and y coordinates are given by the number of human and mouse reads in that barcode, respectively (Figure 19A). Cell barcodes that fall along either axis contain predominately (or exclusively) reads from one species and can reasonably be labeled as one species or another. Cell barcodes with large numbers of reads in both species are considered "collisions", which occur when two cells are stuck together or simply caused by too many cells in the barcode space, creating overlap. The collision rate, or estimated doublet rate, is two times the fraction of cells exhibiting this phenomena. This is because only collisions of human-mouse can be readily seen in the data, collisions of human-human or mouse-mouse simply appear as barcodes with many counts, or are hard to see.

The next important plot is the kneeplot (Figure 19B). For a kneeplot, cell barcodes are ordered by read count from highest to lowest, and plotted as a cumulative fraction of all reads. The plot includes many more barcodes than are expected to contain reads and so contains a "knee" portion. In the early parts of the graph each of the barcodes adds a considerable amount of reads to the cumulative total, leading to a near linear growth pattern. Eventually however the reads per barcode tapers off and the growth is nearly flat (some reads are typically found even in "empty" barcodes). The linear portion indicates the barcodes that are counted as cells and the flat portion is empty barcodes.

Next it is important to check the protein reads. The kneeplot for protein shows a similar sharp bend as the RNA (Figure 19C). Additionally, this plot is shaded by the average RNA UMI content. Here, each dot represents 100 cell barcodes, ordered from most protein reads to least, and the shading is the average RNA UMI count for those 100 barcodes. From the plot it is apparent that overall the protein content and RNA content correlate well with one another. Protein barcodes with the highest read count (left side) also show the highest RNA count (red). This correlation can be more directly examined with a correlation plot (Figure 19D), where each dot represents a cell barcode whose rna content and protein content are plotted on the x and y axes, respectively. While the correlation for the top 50,000 barcodes is respectable (ρ=0.508), the correlation for the top 12,500 barcodes, or the barcodes corresponding to occupied barcodes, is better (ρ=0.686). This makes sense as the junk reads coming from empty barcodes should be random, whereas the reads coming from occupied barcodes should be related. Overall these plots paint a picture of a high quality library where the RNA and protein reads both contain ~12,500 occupied barcodes, and correlate well with each other as to which barcodes those are.

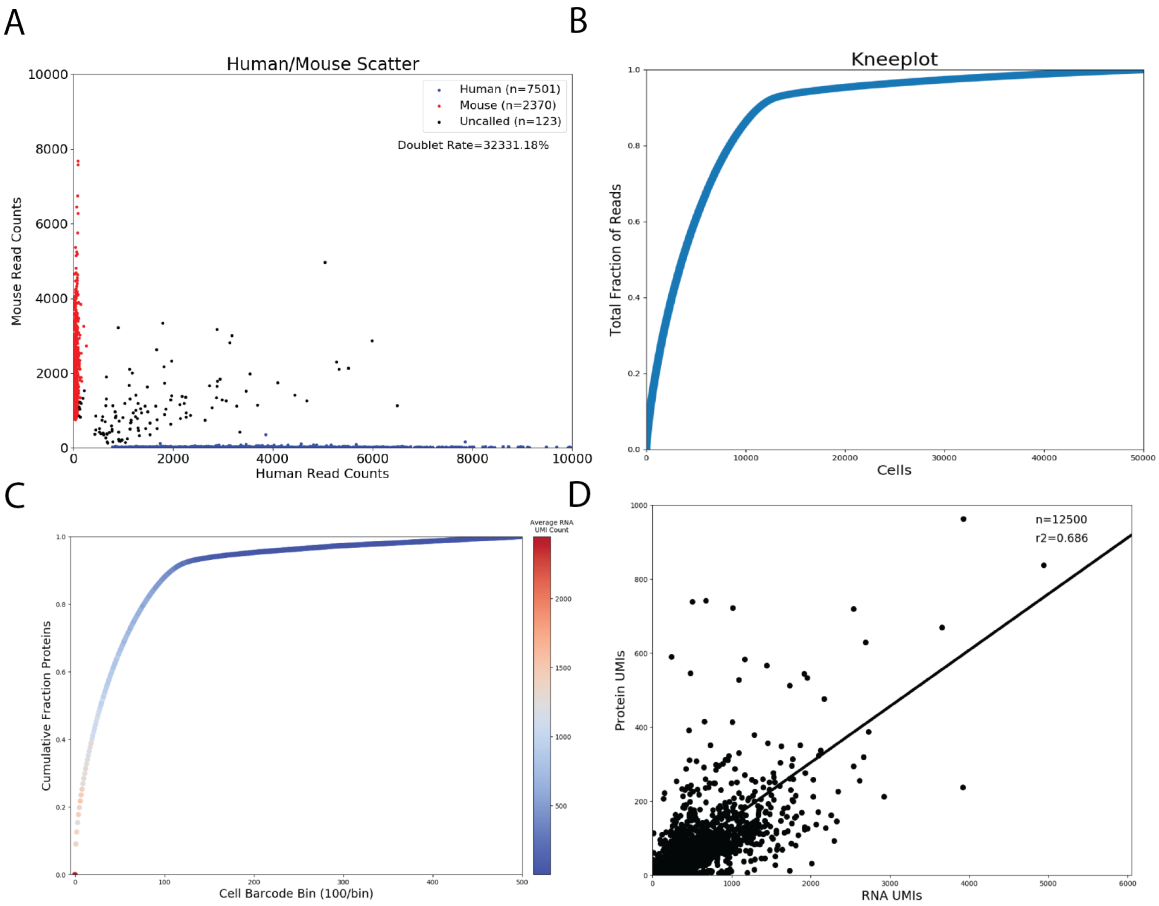

Figure 19: Quality metrics for SPLiT9 experiment. A) Human mouse scatter. Each dot corresponds to a single cell barcode plotted as human read counts and mouse read counts (x and y respectively). Blue dots were judged to be human cells, red dots as mouse cells, black dots are labeled as collisions. Note: any cell barcode containing fewer than 500 reads was considered as "empty". B) Kneeplot for RNA data. Cell barcodes are listed in descending order of read count. Y axis is cumulative read count. Discontinuous "knee" at ~12,500 barcodes indicates switch between "occupied" and "empty" cell barcodes, that is barcodes containing real cell data and barcodes made of noise. C) Kneeplot for protein data, binned into sets of 100. Color corresponds to average RNA UMI count for the 100 barcodes. RNA content and protein content correlate well. D) Correlation plot of top 12,500 cell barcodes (occupied barcodes). Correlation for top 12,500 is 0.686).

3.3.6. Assessing Cell Cycle Using RNA and Protein Data

So far this work has focused principally on the mechanics of achieving RNA and protein reads in single cells. This section focuses on using the combined data to assess biological variability in actively dividing cells. For the data set, 11 different proteins were used, 8 of which have been used as markers for phases of the cell cycle (Table 2). 2 proteins (FUS and U1C) are more ubiquitous transcription factors, and shouldn't vary much over cell cycle. The 11th protein examined was anti beta-2 microglobuin (aB2M), a generic marker that stains human cells generically (Stoeckius *et al* 2017).

After normalizing and scaling the data in Seurat, variable genes can be computed as mentioned in section 3.3.4. This data will contain a multivariate principal component set which can be difficult to visualize. These principal components are not all made equal, however. Indeed, by computing the eigenvalues it is possible to determine the amount of the variance each of the principal components contributes to the data. This can be shown in an elbow plot (Figure 20A). The elbow plot shows that the first three principal components contain a high fraction of the variance (15.1%, 7.9%, and 6.8% respectively). Plotting the first two principal components on the x and y axes, along with cell cycle shading via Seurat's cell cycle vignette package, it becomes clear that the first principal component, and to a lesser extent the second, tracks well with the phases of the cell cycle (Figure 20B). This can be shown even better in a Uniform Manifold Approximation and Projection plot (UMAP; Becht *et al* 2018). UMAP is a visualization package that projects multivariate data sets onto a 2 dimensional space, preserving the clustering and distances from the multivariate space. In this feature, the principal distinction of the data set, UMAP dimension 1, is based principally on the cell cycle phase (Figure 20C). The distinction is not related to read depth (Figure 20D).

The variable gene analysis on RNA data showed that cell cycle is a driving factor in the differential RNA expression in this data set. The same variable analysis can be performed, this time on the protein data. First, looking at the component analysis, it is clear that the protein data set is considerably less complicated than the RNA set (Figure 21A). In the protein data set, 53.1% of the variance is contained in the first principal component, with the next 6 taking up around 5% each (resulting in >80% total). This is expected as the protein data set is considerably less complicated than the RNA set, containing only 11 inputs. The resulting UMAP plot is shown in Figure 21B. It is clear that while there is strong differentiation along UMAP dimensions 1 and 2, the differentiation is not related to cell cycle as determined by RNA phase. This lack of direct correlation between RNA and protein markers of cell cycle makes some sense. Low correlation has been reported in single cell RNA and protein experiments (Darmanis et al 2016). Variation between the two can be explained by many factors, including transcriptional bursting, signal on demand (waiting to translate pre-existing mRNA), and latency between mRNA synthesis and protein synthesis, especially in mammalian cells (Liu et al 2016).

Instead, it makes more sense to define the groups by the protein (Figure 21C). It is clear from the protein UMAP that there are two fairly distinctive groups, with potentially a third, very small group, straddling between them. These groups actually correlate well with the overall protein content of the cells (Figure 21D). Although group 2 does connect them, the majority of cells actually fall into either group 1 or group 3. In fact, 50% of the cells fall into group 1, the low protein group, alone.

These protein groups were then implemented for differential analysis in RNA expression. First, plotting the protein groups on the RNA UMAP shows little clear pattern (Figure 22A). It is notable however that group 3 appears to be along the periphery of the UMAP. This becomes more

interesting when looking back at Figure 20D, where most cells containing large amounts of RNA were likewise along the periphery. Plotting the integration of the 2 protein UMAP dimensions on one axis and the RNA read count on the other shows that the RNA content of cells does seem to be correlating with the protein groups (Figure 22B). In group 1, most cells (92.5%) are below 3000 total reads. In group 3, this fraction is 42.7%, with 12.4% over 5000 reads.

Next, differential expression analysis was performed on the RNA expression in the two protein groups. The results are in Table 3. 38 total genes were found to be significant after Bonferroni correction. 21 of them were included in the Table 3, selected principally for relevance to cell cycle. DAVID analysis (Huang *et al* 2008) of the 38 genes yielded two significant GO terms, for cytokinesis (GO enrichment score of 7.87) and ribosome production (GO enrichment score of 20.4). Ribosome production has been showing to take place around cytokinesis, as ribosomes are critical machinery for cell function (Hernandez-Verdun 2011; Carron *et al* 2012). Violin plots showing the named differentially expressed genes for cytokinesis (Figure 22C) and ribosomal production (Figure 22D) show that group 3 has these genes upregulated. It is therefore surmised that group 3 consists principally of cells near the end of G2M, and that group 1 consists of cells that have just finished cytokinesis. This is consistent with previous findings (Tanenbaum *et al* 2015). It is important to note that this distinction is visible only in the protein, and that the cell cycle-related difference between the two groups is not immediately evident from RNA data alone.

Table 2: Antibodies used in protein analysis.

Antigen	Cell Cycle Marker	Manufacturer Product #	Barcode	Barcode Seq	Oligos/ Protein
FUS		Biolegend 855002	SS0517	AGTACGGC	4.6
U1C		Millipore-Sigma SAB4200188	SS0518 SS0519	ATACCTCT CCAACTGG	1
CycB1	G2M	R&D Systems AF6000	SS0514	AACTTGAT	1.9
CDT1	G1/S	Sigma SAB2701049	SS0515	GAGCGCTA	3.8
Geminin	S/G2M	Proteintech 66566-1-Ig	SS0513 SS0516	TTCGAGTA GATATAGT	3.2
CycA1	G2	R&D Systems MAB7046	SS0512	GCACGAAG	2.6
CDC2	G2M	R&D Systems AF888	SS0511	TCGCGCCA	3.2
CDK2	G1/S	R&D Systems AF4654	SS0510	ACGAAGCT	1.9
HUMAN (aB2M)		Biolegend 316302	SS0508/SS0509	GTAACTTC	3.1
CycD1	G1	R&D Systems MAB4314	SS0507	CTGCCAAC	2
CDC25C	G2M	R&D Systems MAB4459	SS0506	TCATCACT	2.3

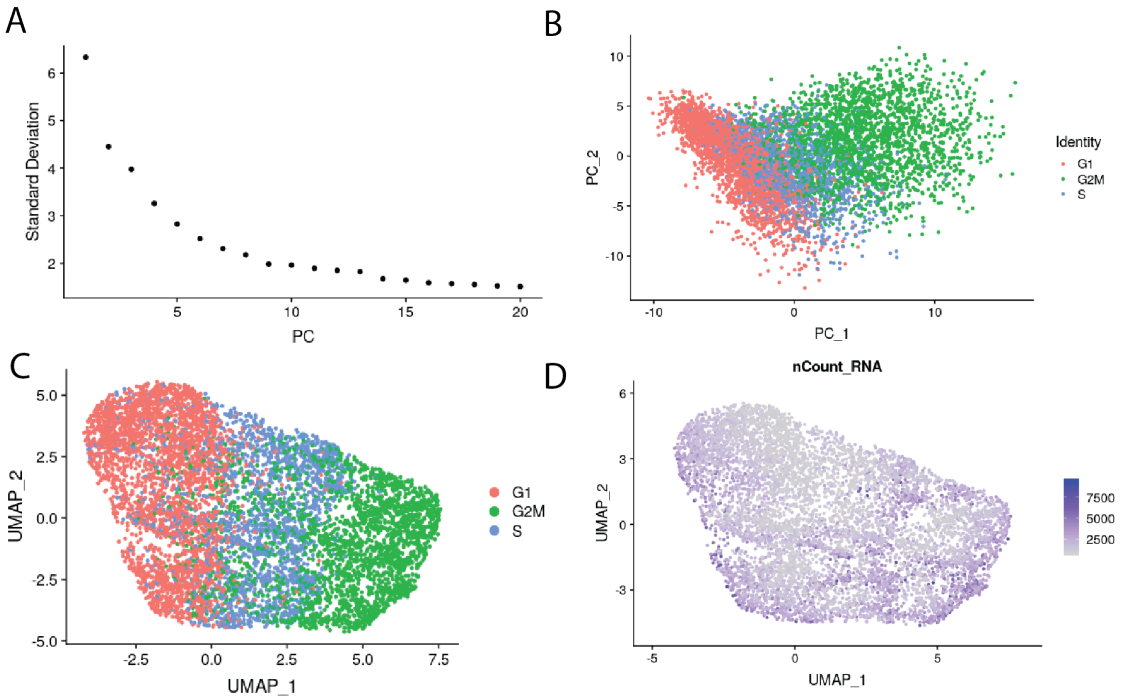

Figure 20: Differential RNA expression analysis using Seurat. A) Elbow plot showing contributions of each principal component (PC) to the overall variance of the RNA data. B) PC plot showing first two PCs with cells labeled by cell cycle phase. C) Uniform Manifold Approximation and Projection (UMAP) analysis showing multivariate data projected onto 2D plot. Principal variation (UMAP_1) is along cell cycle. D) Same UMAP plot as (C), but shaded by unique read count.

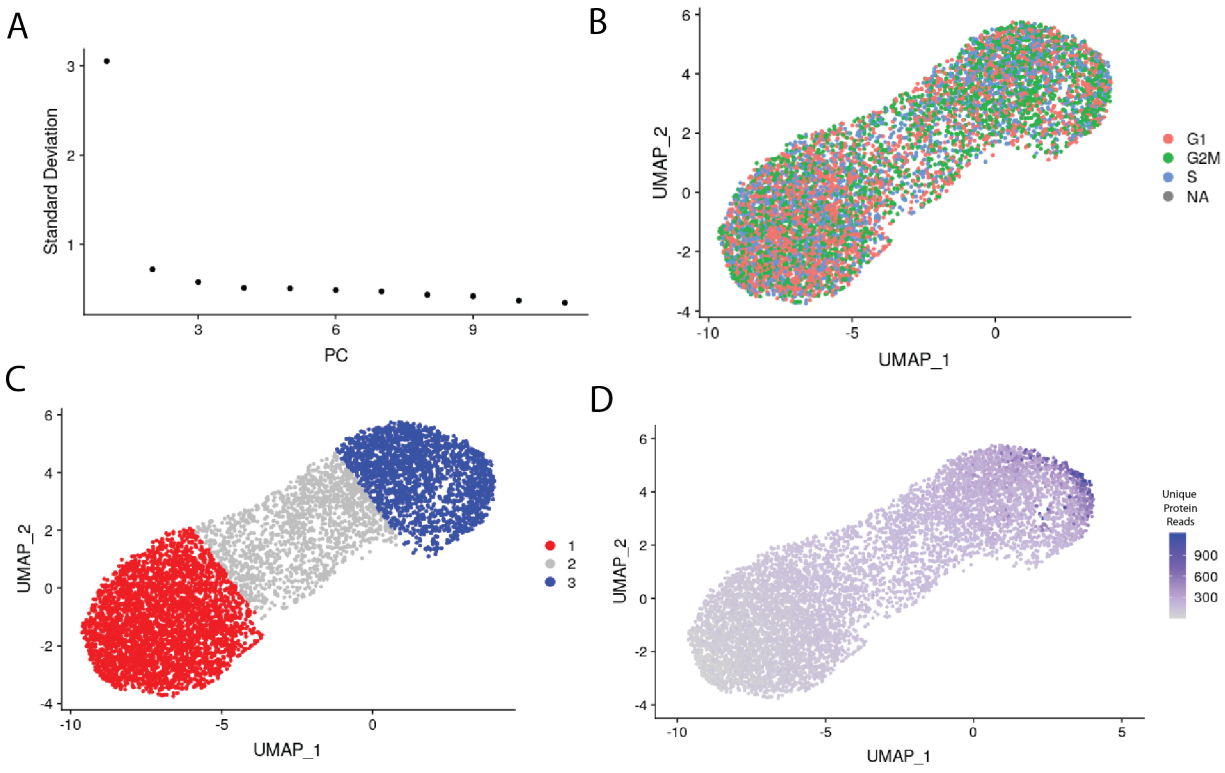

Figure 21: Differential protein expression analysis and integrated analysis using Seurat. A) Elbow plot for protein data showing most variation (53.1%) is explained in the first PC. B) UMAP plot for protein data, colored by cell cycle. Protein variation seems to have no correlation with cell cycle as determined by RNA markers. C) Protein UMAP shaded by UMAP dimension 1 values showing 3 groups, two polar groups (1 and 3) with a region in the middle showing cells not belonging strongly to either group. D) Same groups now applied to RNA UMAP from 18C. Though still not absolute, some polarity seems to exist.

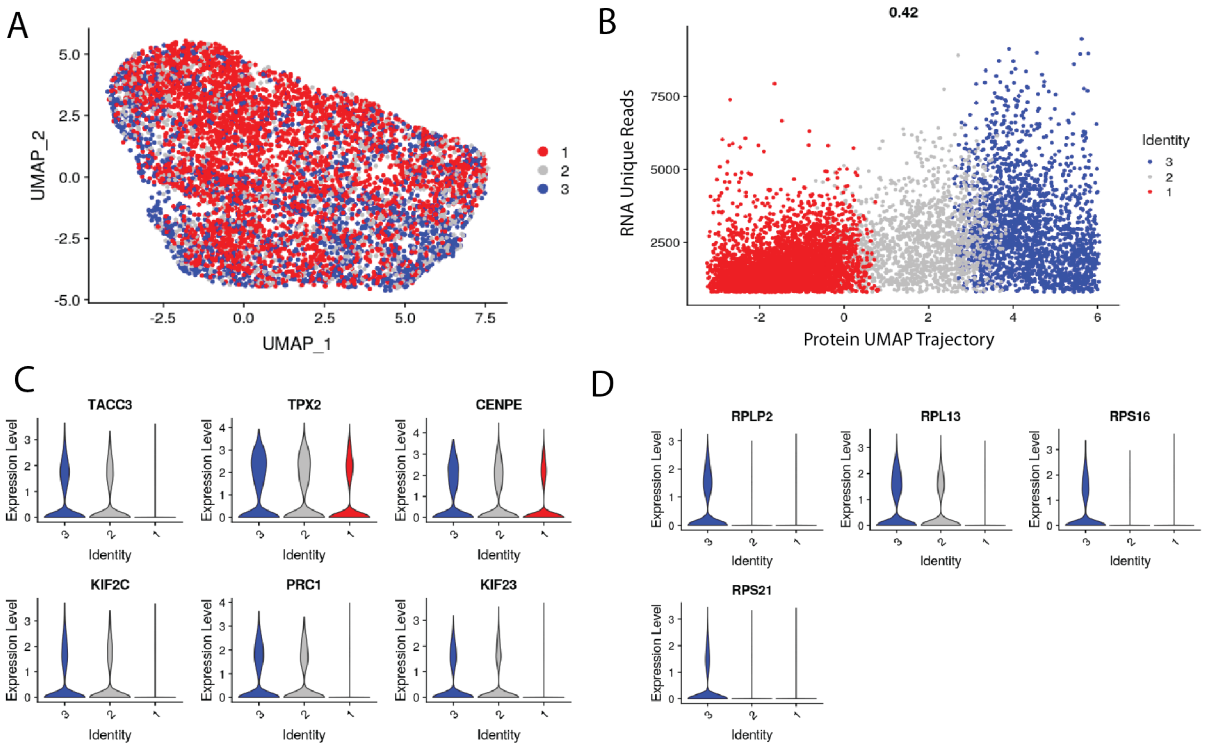

Figure 22: RNA expression between protein groups. A) RNA UMAP (first in 19C) shown with protein colors from Figure 20C. Note that the only clear pattern is that most cells in group 3 cluster near the outside of the graph. B) Correlation of protein UMAP trajectory (UMAP 1 and 2 merged to a single axis) and RNA Read count. Note that protein group 1 has few reads above 2500 RNA UMI, whereas protein group 3 has a large population above 5000 UMI. C) RNA count violin plots for 6 differentially expressed genes related to cytokinesis between protein groups 1 and 3 (GO enrichment score 2.56). C) RNA count violin plots for 6 differentially expressed genes related to ribosome production between protein groups 1 and 3 (GO enrichment score 2.43).

Table 3: Significant RNA markers between protein groups 1 and 3.

Gene	Adj. P (Bonferroni)	Function
UBC	4.47E-57	G2M related signaling (Gilberto et al 2017; PMID: 17491588)
TACC3	5.01E-48	May play a role in stabilization of the mitotic spindle
RPLP2	2.35E-47	Component of 60S ribosomal subunit
RPS21	8.12E-45	Component of 40S ribosomal subunit
MALAT1	2.07E-44	May regulate genes involved in cancer metastasis and cell migration, and it is involved in cell cycle regulation.
TOP2A	5.97E-44	Mitotic
KIF23	5.30E-43	Move chromosomes during cell division
FTH1	1.12E-42	Ferritin heavy chain 1 unit
S100A6	3.75E-41	Involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation
RPL13	1.08E-38	Component of 60S ribosomal subunit
HIST1H4C	2.04E-37	Histone protein
KIF2C	3.08E-37	Mitotic
AURKA	3.35E-37	Microtubule formation and/or stabilization at the spindle pole during chromosome segregation
RPS16	9.25E-37	Component of 40S ribosomal subunit
PRC1	1.33E-35	Key regulator of cytokinesis
TPX2	5.70E-35	Spindle assembly factor required for normal assembly of mitotic spindles
CENPE	1.12E-28	Centromere protein E
CCNB1	2.63E-26	Necessary for proper control of the G2/M transition phase of the cell cycle
SERPINE1	1.62E-25	Inhibitor of fibrinolysis
ITGA2	1.03E-24	Adhesion to the extracellular matrix
MEIS2	1.21E-06	Highly conserved transcription regulators shown to be essential contributors to developmental programs

3.4 Conclusions

This chapter discuss single cell measurements of RNA and protein. Starting with the SPLiTSeq protocol from Rosenberg *et al*, the protocol was modified to accommodate protein measurements. The adjustments to include protein were principally before the beginning of the SPLiTSeq protocol, staining with antibodies and washing prior to the first reverse transcription step. Several experiments were performed aimed at increasing the protein signal, including adjusting fixation and washing conditions, as well as making barcoded oligos specific for protein capture. After optimizing several conditions, a trial was conducted using a fairly large amount of antibodies that serves as the benchmark.

Results from the featured experiment indicate strong correlation between occupied barcodes in both RNA and protein data sets. Both sets indicate approximately 12,500 cells in the experiment as indicated by the kneeplots (Figure 19B-C). Moreover, the correlation between those barcodes is high (ρ=0.686). We can therefore concluded that both data sets are measuring from cells, and the human/mouse scatter suggests both data sets contain predominantly single cells. In looking at the biology, it becomes clear that protein and RNA are not closely correlated within individual cells, which corresponds to previously published findings (Darmanis *et al* 2015; Liu *et al* 2016). Moreover, it is difficult to use the RNA as markers for protein (21-B), although the converse does hold some promise (Figure 22 and Table 3).

The markers seen most closely defining the difference between the two polar protein groups (red and blue in Figure 21C), are few but are consistently related to cell polarity and mitosis. This makes sense then, seeing them at the exterior of the RNA UMAP (Figure 22A), most closely tracking to the outer edges of their respective groups. Their position, in both cases, indicates that the cells are within the G1 or G2M phase groups, but are further from the S phase. For G1 phase,

this is cells immediately leaving mitosis and re-entering G1, for G2M cells this would be cells actively engaged in mitosis. These genes are well correlated with that theory, and the fact that these two protein groups create an altogether different divide than the RNA groups is an example of why intracellular protein/RNA analysis can be a useful feature. Even in the limited cell cycle model, the protein is offering quite different insights from RNA data alone. Complementary protein and RNA single-cell analysis could therefore be a powerful tool in identifying cell function and tissue organization in the future.

Finally, a word on comparison between current methods and the SPLiTSeq+Protein option described in this paper. The closest comparison is with the commercial 10X Genomics system with added on Cell Surface Profiling (https://www.10xgenomics.com/solutions/vdj/), based on work described in Stoeckius *et* al 2017. Both the 10X method and the one described in this work are capable of targeting many proteins, with limitations mostly based on the cost of producing new antibodies. However, the 10X system is explicitly restricted to surface proteins, which often lack strong biological meaning outside of basic classification. The SPLiTSeq+Protein method described here uses Fabs for intracellular and intranuclear targeting, greatly expanding the breadth and depth of targets available. Also, being based on SPLiTSeq, the number of cells achieved during a single run can be in excess of 100,000 (Figure 16B). At comparable doublet rates to the most recent experiment (~3.4%), a single V2 chip can sequence around 4500 cells with costs around $400 per thousand cells (Wang *et al* 2019) . Whereas SPLiTSeq costs $5.45 per thousand cells, plus about $2 per antibody for the experiment. Even with 10 antibodies or more, this still makes a drastic improvement over the cost of 10X. Finally, the method requires no specialized equipment and is therefore easily scalable; users can adjust how many antibodies they desire per experiment without the need for different kits.

Table 4: Comparison of SPLiTSeq+Protein and 10X Protein Analysis.

	10X Protein Analysis	SPLiTSeq + Protein
# Proteins (max)	100s	100s
Protein Targets	Immune Surface Markers	**Intracellular/ Intranuclear**
Usable Cells/Experiment	40,000	**100,000+**
Cost per 1000 cells	$400	**$5.45 + $2 /Ab**

3.5 Appendix to Chapter 3

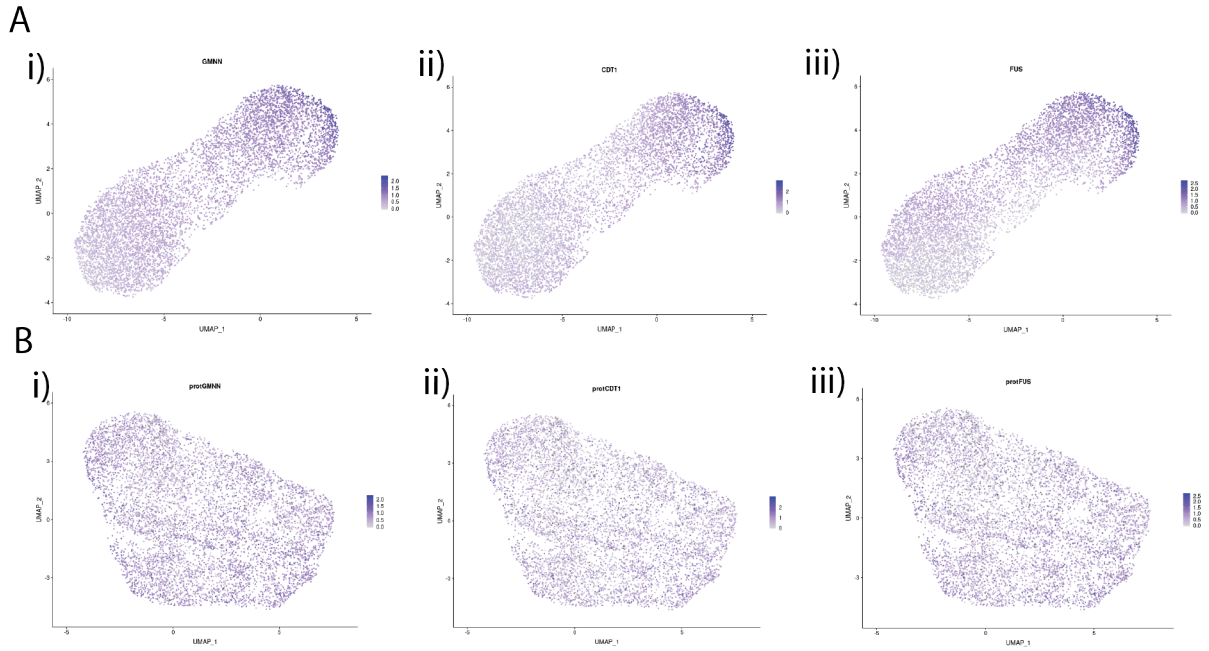

Figure 23: Individual gene concentrations for select proteins. A) Individual protein expression for Geminin (GMNN, i), Chromatin Licensing And DNA Replication Factor 1 (CDT1, ii), and Fused in Sarcoma (FUS, iii). GMNN is positively upregulated from G1->G2M, starting with S, making its expression pattern consistent with expected results. CDT1 is supposed to have the opposite pattern, decreasing between G1->G2M. FUS should be cell-cycle independent. B) Individual protein expression for the same three genes over the RNA UMAP from Figure 20C. No clear expression patterns are visible.

REFERENCES

Adamczyk, Maciej, John C. Gebler, Jiang Wu. 2000. “Papain Digestion of different mouse IgG subclasses as studied by electrospray mass spectrometry”. *Journal of Immunological Methods* 237: 95-104. PMID: 10725455

Adey, Andrew and Jay Shendure. 2010. “Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing”. *Genome Research* 22:1139–1143. PMID: 22466172

Amidzadeh, Zahra, Abbas Behzad Behbahani, Nasrollah Erfani, Sedigheh Sharifzadeh, Reza Ranjbaran, Leili Moezi, Farzaneh Aboualizadeh, Mohammad Ali Okhovat, Parniyan Alavi, and Negar Azarpira. 2014. “Assessment of Different Permeabilization Methods of Minimizing Damage to the Adherent Cells for Detection of Intracellular RNA by Flow Cytometry”. PMID: 24523954

Assarsson, Erika, Martin Lundberg, Goran Holmquist, Johan Bjorkesten, Stine Bucht Throsen, Daniel Ekman, Anna Eriksson, Emma Rennel Dickens, Sandra Ohlsoon, Gabriella Edfeldt, Ann-Catrin Andersson, Patrik Lindstedt, Jan Stenvang, Mats Gullberg, Simon Fredriksson. 2014. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One*. PMID: 24755770

Barron, Martin & Jun Li. 2016. “Identifying and removing the cell-cycle effect from single-cellRNA-Sequencing data”. *Scientific Reports* 6,33892. PMID: 27670849

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoione Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2018. “Dimensionality reduction for visualizing single-cell data using UMAP”. *Nature Biotechnology*. PMID: 30531897

Biggin, Mark D. 2011. “Animal Transcription Networks as Highly Connected, Quantitative Continua”. *Developmental Cell* 21(4):611-26. PMID: 22014521

Buettner, Florian, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. 2015. “Computational analysis of cell-to-cell heterogeneity in single-cellRNA-sequencing data reveals hidden subpopulations of cells”. *Nature Biotechnology* 33,155–160. PMID: 25599176

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi and Rahul Satija. 2018. “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. *Nature Biotechnology* 36(5):411-420. PMID: 29608179

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston , Cole Trapnell , Jay Shendure. 2017. “Comprehensive single-cell transcriptional profiling of a multicellular organism.” *Science.* PMID: 28818938

Carron, Coralie, Stephanie Balor, Franck Delavoie, Celia Plisson-Chastang, Marlene Faubladier, Pierre-Emmanuel Gleizes, and Marie-Francoise O'Donohue. 2012. "Post-mitotic dynamics of pre-nucleolar bodies is driven by pre-rRNA processing". *Journal of Cell Science* 25(Pt 19):4532-42. PMID: 22767511

Chen, Fei, Asmamaw T Wassie1, Allison J Cote, Anubhav Sinha, Shahar Alon, Shoh Asano, Evan R Daugharthy, Jae-Byum Chang, Adam Marblestone, George M Church, Arjun Raj, and Edward S Boyden. 2016. "Nanoscale imaging of RNA with expansion microscopy". *Nature Methods* 2016.

Chen, Xi, Ricardo J. Miragaia, Kedar Nath Natarajan, & Sarah A Teichmann. 2018. "A rapid and robust method for single cell chromatin accessibility profiling". *Nature Communications.* PMID: 30559361

Cusanovich, Darren A, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, Jay Shendure. 2015. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing." *Science.* PMID: 25953818

Darmanis, Spyro, Caroline Julie Gallant, Voichita Dana Marinescu, Mia Niklasson, Anna Segerman, Georgios Flamourakis, Simon Fredriksson, Erika Assarsson, Martin Lundberg, Sven Nelander, Bengt Westermark, and Ulf Landegren. 2015. "Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells". *Cell Reports*. PMID: 26748716

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Pilippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: ultrafast universal RNA-seq aligner". *Bioinformatics* 29(1):15-21. PMID: 23104886

Duhamel, Raymond C., Peter H. Schur, Klaus Brendel, and Elias Meezan. 1979. "pH Gradient Elution of Human IgG1, IgG2, and IgG4 from Protein A-Sepharose". *Journal of Immunological Methods* 31: 211-217. PMID: 42659

Dutta, Dipannita, Chad D. Williamson, Nelson B. Cole, and Julie G. Donaldson. 2012. "Pitstop 2 Is a Potent Inhibitor of Clathrin-Independent Endocytosis". *PLoS One.* PMID: 23029248

Ecker, Simone, Lu Chen, Vera Pancaldi, Frederik O. Bagger, José María Fernández,Enrique Carrillo de Santa Pau1, David Juan, Alice L. Mann, Stephen Watt, Francesco Paolo Casale, Nikos Sidiropoulos, Nicolas Rapin, Angelika Merkel, BLUEPRINT Consortium, Hendrik G. Stunnenberg, Oliver Stegle, Mattia Frontini , Kate Downes, Tomi Pastinen, Taco W. Kuijpers, Daniel Rico, Alfonso Valencias, Stephan Beck2, Nicole Soranzo and Dirk S. Paul. 2017. "Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types". *Genome Biology* 18(1):18. PMID: 28126036

Fagerberg, Linn, Bjorn M. Hallstrom, Per Oksvold, Caroline Kampf, Dijana Djureinovic, Jacob Odeberg, Masato Habuka, Simin Tahmasebpoor, Angelika Danielsson, Karolina Edlund, Anna Asplund, Evelina Sjostedt, Emma Lundberg, Cristina Al-Khalili Szigyarto, Marie Skogs, Jenny Ottosson Takanen, Holger Berling, Hanna Tegel, Jan Mulder, Peter Nilsson, Jochen M. Schwenk, Cecilia Lindskog, Frida Danielsson, Adil Mardinoglu, Åsa Sivertsson, Kalle von Feilitzen, Mattias Forsberg, Martin Zwahlen, IngMarie Olsson, Sanjay Navani, Mikael Huss, Jens Nielsen, Fredrik Ponten, and Mathias Uhlen. 2014. "Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics". Molecular and *Cellular Proteomics.* PMID: 24309898

Fernandez, Ariel and Oktay Sinanoglu. 1985. "Denaturation of proteins in methanol/water mixtures." *Biophysical Chemistry.* PMID: 17007768

Frei, Andreas P, Felice-Alessio Bava, Eli R Zunder, Elena W Y Hsieh, Shih-Yu Chen, Garry P Nolan, Pier Federico Gherardini. 2016. "Highly multiplexed simultaneous detection of RNAs and proteins in single cells". *Nature Methods*. PMID: 26808670

Hernandez-Verdun, Daniele. 2011. Assembly and disassembly of the nucleolus during the cell cycle. *Nucleus* 2:3, 189-194. PMID: 21818412

Galvis, A., Fisher, H. E. and Camerini, D. 2017. "NP-40 Fractionation and Nucleic Acid Extraction in Mammalian Cells". Bio-protocol 7(20): e2584. DOI: 10.21769/BioProtoc.2584.

Gerard, Gary F, R. Jason Potter, Michael D. Smith, Kim Rosenthal, Gulshan Dhariwal, Jun Lee, and Deb. K. Chatterjee. 2002. "The role of template-primer in protection of reverse transcriptase from thermal inactivation". *Nucleic Acids Research* 30(14): 3118-3129. PMID: 12136094

Gilberto, Samuel, and Matthias Peter. 2017. "Dynamic ubiquitin signaling in cell cycle regulation". *Journal of Cell Biology* 216(8):2259-2271. PMID: 28684425

Gookin, Sara, Mingwei Min, Harsha Phadke, Mingyu Chung, Justin Moser, Iain Miller, Dylan Carter, Sabrina L. Spencer. 2017. "A map of protein dynamics during cell-cycle progression and cell-cycle exit". *Public Library of online Sciences Biology* 15(9):e2003268. PMID: 28892491

Grisedale, Kelly and van Daal, Angela. 2014. "Linear amplification of target prior to PCR for improved low template DNA results". *Biotechniques.* PMID: 24641479

Hober, Sofia, Karin Nord, and Martin Linhult. 2007. "Protein A chromatography for antibody purification". *Journal of Chromatography B* 848: 40-47. PMID: 17030158

Holden, Paul, and William A. Horton. 2009. "Crude subcellular fractionation of cultured mammalian cell lines". *Biomed Central Research Notes*. PMID: 20003239

Hoffman, Elizabeth, Brian L Frey, Lloyd M Smith, and David T. Auble. 2015. “Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes”. *Journal of Biological Chemistry* vol 290:44. 26404-26411. PMID: 26354429

Huang, Da Wei, Brad T. Sherman, and Richard Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. *Nature Protocols* 4(1): 44-57. PMID: 19131956

Irani, Vashti, Andrew J. Guy, Dean Andrew, James G. Beeson, Paul A. Ramsland, and Jack S. Richards. 2015. “Molecular properties of human IgG subclasses and their implications for designing therapeutic monoclonal antibodies against infectious diseases”. *Molecular Immunology* 67: 171-182. PMID: 25900877

Janeway CA Jr, Travers P, Walport M, et al. “Immunobiology: The Immune System in Health and Disease”. 5th edition. New York: Garland Science; 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10757/?depth=10

Ji, Hong. 2010. “Lysis of Cultured Cells for Immunoprecipitation”. *Cold Spring Harbor Protocols.* PMID: 20679375

Kivioja, Teemu, Anna Vaharautio, Kasper Karlsson, Martin Bonk, Martin Enge, Sten Linnarsson & Jussi Taipale. 2011. “Counting absolute numbers of molecules using unique molecular identifiers.” *Nature Methods.* PMID: 22101854

Klein, Allon M, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W. Kirschner. 2015. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells”. *Cell*. PMID: 26000487

Kowalczyk, Monica S, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J. Haas,1 Rebekka K. Schneider, Amy J. Wagers, Benjamin L. Ebert, and Aviv Regev. “Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells”. *Genome Research* 25:1860-1872. PMID: 26430063

Lareau, Caleb A, Fabiana M Duarte, Jennifer G. Chew, Vinay K Kartha, Zach D Burkett, Andrew S Kohlway, Dmitry Pokholok, Martin J. Aryee, Frank J. Steemers, Ronald Lebofsky, Jason D. Buenrostro. 2019. “Droplet-based combinatorial indexing for massive single-cell epigenomics”. *BioRxiv*. doi: https://doi.org/10.1101/612713

Lewis Carl, Stephanie A, Illona Gillete-Ferguson, and Donald G Ferguson. 1993. “An Indirect Immunofluorescence Procedure for Staining the Same Cryosection with Two Mouse Monoclonal Antibodies”. *Journal of Histochemistry and Cytochemistry* 41(8):1273-8. PMID: 7687266

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." Bioinformatics 25 (16): 2078–79. PMID: 19505943

Liashkovich, Ivan, Dzmitry Pasrednik, Valeria Prystopiuk, Gonzalo Rosso, Hans Oberleithner and Victor Shahin. 2015. "Clathrin inhibitor Pitstop 2 disrupts the nuclear pore complex permeability barrier". *Scientific Reports*. PMID: 25944393

Liu, Yansheng, Andreas Beye, and Reudi Aebersold. 2016. "On the Dependency of Cellular Protein Levels on mRNA Abundance". *Cell* 165(3):535-50. PMID: 27104977

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, aviv Regev, and Stene A. McCarroll. 2015. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". *Cell.* PMID: 26000488

Mohsen, Michael G and Kool, Eric T. 2016. "The Discovery of Rolling Circle Amplification and Rolling Circle Transcription". *Accounts of Chemical Research.* PMID: 27797171

Niklas, Jens, Armin Melnyk, Yongbo Yuan, and Elmar Heinzle. 2011. "Selective permeabilization for the high-throughput measurement of compartmented enzyme activities in mammalian cells". *Anals of Biochemistry.* PMID: 21683676

Negoescu, Adrien, Francoise Labat-Moleur, Philippe Lorimier, Laurence Lamarcq, Christiane Guillermet, Edmond Chambaz, and Elisabeth Brambilla. "F(ab) Secondary Antibodies: A General Method for Double Immunolabeling with Primary Antisera from the Same Species. Efficiency Control by Chemiluminescence". *Journal of Histochemistry and Cytochemistry* 42(3):433-437. PMID: 7508473

Nilsson, Mats, Helena Malmgren, Martina Samiotaki, Marek Kwiatowski, B.P Chowdhary and Ulf Landegren. 1994. "Padlock probes: circularizing oligonucleotides for localized DNA detection". *Science*. PMID: 7522346

Otali, Dennis, Cecil R Stockard, Denise K oelschlager, Wen Wan, Upender Manne, Stephen A Watts, and William E Grizzle. 2009. "Combined Effects of Formalin Fixation and Individual Steps in Tissue Processing on Immunorecognition". *Biotechnology Histochemistry.* PMID: 19886759

Peterson, Vanessa M, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova and Joel Klappenbach. 2017. "Multiplexed quantification of proteins and transcripts in single cells". *Nature Biotechnology* 35: 936-939. PMID: 28854175

Picelli, Simone, Asa K. Bjorklund, Bjorn Reinius, Sven Sagasser, Gosta Winberg, and Rickard Sandberg. 2014. "Tn5 transposase and tagmentation procedures for massively scaled sequencing projects". *Genome Research* 24:2033–2040. PMID: 25079858

Rosenberg, Alexander B, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas Gray, David J. Peter, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, Georg Seelig. 2018. "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding". *Science* 360:176-182. PMID: 29545511

Russell, Julia N, Janice E Clements, and Lucio Gama. "Quantitation of Gene Expression in Formaldehyde-Fixed and Fluorescence-Activated Sorted Cells". *PLoS One.* PMID: 24023909

Scialdone, Antonio, Kedar N. Natarajan, Luis R. Saraiva, Valentine Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. 2015. "Computational assignment of cell-cycle stagge from single-cell transcriptome data". *Methods* 85:54-61. PMID: 26142758

Shao, Qiang. 2014. "Methanol Concentration Dependent Protein Denaturation Ability of Guanidium/Methanol Mixed Solution". *Journal of Physical Chemistry.* PMID: 24846320

Sos, Brandon Chin, Ho-Lim Fung, Derek Rui Gao, Trina Faye Osothprarop, Amirali Kia, Molly Min He, and Kun Zhang. 2016. "Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-Seq) assay". *Genome Biology.* PMID: 26846207

Srinivasan, Mythily, Daniel Sedmak, and Scott Jewell. 2002. "Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids". *American Journal of Pathology.* Vol 161: 6. PMID: 12466110

Stave, James W, and Lindpaintner, Klaus. 2013. "Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure". *Journal of Immunology.* PMID: 23797669

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. "Large-scale simultaneous measurement of epitopes and transcriptomes in single cells". *Nat Methods*. 14(9): 865–868. PMID: 28759029

Svensson, Valentine, Roser Vento-Tormo, Sarah A. Teichmann. 2018. "Exponential scaling of single-cell RNA-Seq in the last decade". *Nature Protocols*. PMID: 29494575

Tanenbaum, Marvin E, Noam Stern-Ginossar, Jonathan S Weissman, and Ronald D Vale. 2015. "Regulation of mRNA translation during mitosis". *eLife*. PMID: 26305499

Vani, Kodela, Steven A. Bogen, and Seshi R. Sompuram. 2006. “A High Throughput Combinatorial Library Technique for Identifying Formalin-Sensitive Epitopes”. *Journal of Immunological Methods* 317(1-2): 80-89. PMID: 17056057

Vidarsson, Gestur, Gillian Dekkers, and Theo Rispens. 2014. “IgG subclasses and allotypes: from structure to effector functions”. *Frontiers in Immunology* 5:520. PMID: 25368619

Vitak, Sarah A., Kristof A. Torkenczy, Jimi L. Rosenkrantz, Andrew J. Fields, Lena Christiansen, Melissa H. Wong, Lucia Carbone, Frank J. Steemers, and Andrew Adey. 2017. “Sequencing thousands of single-cell genomes with combinatorial indexing”. *Nature Methods*. PMID: 28135258

vonKleist, Lisa, Wiebke Stahlschmidt, Haydar Bulut, Kira Gromova, Dmytro Puchkov, Mark J. Robertson, Kylie A. MacGregor, Nikolay Tomilin, Arndt Pechstein, Ngoc Chau, Megan Chircop, Jennette Sakoff, Jens Peter von Kries, Wolfram Saenger, Hans-Georg Krausslich, Oleg Shupliakov, Phillip J. Robinson, Adam McCluskey, and Volker Haucke. 2011. “Role of the Clathrin Terminal Domain in Regulating Coated Pit Dynamics Revealed by Small Molecule Inhibition”. *Cell*. PMID: 21816279

Wang, Yue J., Jonathan Schug, Jerome Lin, Zhiping Wang, Andrew Kossenkov, the HPAP Consortium,5 Klaus H. Kaestner. 2019. “Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues”. *bioRxiv*. https://doi.org/10.1101/541433

Woof, Jenny M, and Dennis R Burton. 2004. “Human Antibody-Fc Receptor Interactions Illuminated by Crystal Structures”. *Nature Reviews Immunology*. PMID: 15040582

Zeisel, Amit, Ana B Munoz-Manchado, Simone Codeluppi, Peter Lonnerberg, Gioele La Manno, Anna Jureus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Goncalo Castelo Branco, Jens Hjerling-Leffler, Sten Linnarsson. 2015. “Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq”. *Science.* PMID: 25700174

Zhao, Yonghong, Lester Gutshall, Haiyan Jiang, Audrey Baker, Eric Beil, Galina Obmolova, Jill Carton, Susann Taudte, Bernard Amegadzie. 2009. “Two routes for production and purification of Fab fragments in biopharmaceutical discovery research: Papain digestion of mAb and transient expression in mammalian cells”. *Protein Expression and Purification* 67: 182-189. PMID: 19442740

Zhu, YY, E.M. Machleder, A. Chenchik, R. Li, and P.D. Siebert. 2001. “Reverse Transcription Template Switching: A SMART™ Approach for Full-Length cDNA Library Construction”. *BioTechniques* 30: 892-897. PMID: 11314272