Expanding Single-Cell RNA-Sequencing in Scale and Dimension

by

Jase Gehring

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lior Pachter, Chair
Assistant Adjunct Professor Jacob Corn
Assistant Professor Nick Ingolia
Assistant Professor Aaron Streets

Spring 2018

Expanding Single-Cell RNA-Sequencing in Scale and Dimension

Abstract

Expanding Single-Cell RNA-Sequencing in Scale and Dimension

by

Jase Gehring

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Lior Pachter, Chair

Multicellular organisms rely on diverse cell types to carry out the multitude of tasks necessary for complex life. Understanding the interplay between cell populations within a tissue or organism is a major goal of biological and medical research. In pursuit of this aim, recent advances in microfluidics and molecular biology have propelled single-cell RNA-sequencing (scRNA-seq) to the forefront of cell population analysis. Routine scRNA-seq experiments can profile tens of thousands of genes from tens of thousands of cells in parallel, offering a platform ripe for technological development, a sandbox in which a clever molecular biologist may develop more varied experiments at unprecedented scale and depth. Accordingly, we have made significant inroads toward the goals of simultaneous RNA/epitope quantification and ultra-low-cost library preparation, and we have expanded the capacity of scRNA-seq with a novel sample multiplexing method. We demonstrate the power of parallel cell population analysis with a high-resolution screen of experimental perturbations, introducing a new paradigm in which scRNA-seq is used to understand a cell population at multiple scales with unprecedented depth of information.

To Kate

Contents

List of Figures

List of Tables

Acknowledgments

My path through graduate school has been neither conventional nor ideal. Many times I considered leaving Berkeley without a PhD, sometimes with good reason. The following people, at different times and in different ways, gave me support and honest advise when I needed it most. Any accomplishments I've made are due to them.

First and foremost I thank my advisor, Lior Pachter, who is responsible for reigniting the passion for science that I lost somewhere in grad school. His fearlessness, curiosity, and attention to ethics have capped my scientific education. Our disparate backgrounds, mixed with our overlapping interests, have resulted in exciting advances and a dynamic new laboratory wedged between experimental and analytical science. Together we've built on the pioneering efforts of his first experimental biologists and established a foundation for innovation and discovery that cannot be confined to any field of study. Shannon Hateley, Lorian Schaeffer, Lynn Yi, Shannon McCurdy, Vasilis Ntranos, Valentine Svensson, Akshay Tambe, Robert Tunney, Ali Booeshagi, Eduardo Beltrame - thank you for welcoming to a new lab and a new institution.

For three years I worked in the laboratory of Michelle Chang. Although she and I have little to show for our efforts, looking back I realize just how much I learned while struggling and failing in Michelle's lab, the successes and shortcomings of which have influenced my personal views immensely. Two of her students, Mike Blaisse and Jon McMurry, are responsible for most of my hands-on scientific training. I try to emulate their drive, brilliance, and creativity. Vivian Yu and Joe Gallagher supported me throughout the trials of graduate school. This thesis would not exist without their friendship. Amy Weeks, Brooks Bond-Watts, Ben Thuronyi, Mark Walker, Maggie Brown, Miao Wen, Jeff Hanson, Stephanie Jones, Omer Ad, Ningkun Wang, Jen Wang, Heng Song, Hongjun Dong, Chia-I Lin, Sasilada Sirirungruang, Kersh Theva, Jorge Marchand, and Monica Neugebauer - thank you for all the conversations, advise, and companionship.

I've had the unusual privilege of training at two of the world's top research institutes, and while these universities could scarsely be more different, they share in commmon communities of scientists eager to help others. I never had the resources or knowledge to complete my research projects in a single lab. To all the lab members and PIs who have contributed so much to my education and research, in particular Djem Kissiov, Matt Thomson, Jeff Park, Sisi Chen, Can Li, John Thompson, and Brady Weissbound, thank you. May your inclusiveness spread to all corners of the scientific endeavour.

I must remember the people who helped me choose the life of science. Brian and Kelly Hogan were kind and passionate advisors during my time at UNC. My undergraduate research advisor, Jeff Dangl, exemplified logic and perseverance, and my mentors Sarah Lebeis, Derek Lundberg, Sur Herrera-Paredes, and Scott Yourstone showed me how to have fun while rolling with the punches science dishes out.

Finally, I thank my friends and family who have been with me along this journey. If we must struggle, best to do so in a beautiful place filled with beautiful people. My mind has been opened and my heart filled. Thank you.

Chapter 1

Introduction

1.1 Life as Letters

Life on Earth, in all its wonder, creativity, and brutality, can be accurately described by sequences of letters. The primary sequences of three linear polymers - DNA, RNA, and protein - determine the functions of these molecules, the conductors of the chemical orchestra of life. For the nucleic acids, DNA and RNA, sequence and function are (to a first approximation) one and the same, while proteins are remarkable in their ability to precisely position chemical functional groups in space based on their sequences. Sixty years ago, Francis Crick proposed a model of molecular biology in which a DNA genome generates a dynamic population of RNA molecules (the transcriptome), which codes for corresponding protein molecules responsible for carrying out the chemistry of life (Crick 1958). In the years since Crick's famous proposal, many additions have been made to this "Central Dogma", but the core idea that the polymer sequences of a relatively small set of molecules uniquely defines a cell has proven incredibly powerful. In a sense, Cricks model reframed biology from an impossibly complex chemical problem to a potentially tractable informatics problem. But tractable is a relative term. Biological genomes are enormous (commonly 10^6-10^9 DNA bases), and the four DNA bases conspire to produce a combinatorial sequence space of cosmic proportions. Crick's simple rules mask the staggering, incomprehensible complexity of earthly life. Even if we make the tentative assumption that the immense chemical networks underlying all of metabolism, signal processing, development, etc. can be thought of as an emergent property arising from a fundamental information structure, our understanding of and control over biology is at best limited by our ability to code and decode the primary sequences of the molecules of life. Reflecting on the vast ocean of sequence space, our efforts to explore it are perhaps akin to a beachgoer splashing in the surf.

In molecular biology, sequence determines structure, and structure determines function. Cells long ago realized the utility of this paradigm. Function is complicated, but sequence is not. Sequence gives cells (and biologists) a handle by which they can manipulate the chemical world. It is a language of abstraction that brings all the power of biomolecules within reach. While a complete understanding of cell biology is well beyond our current capability, sequence abstraction allows us to perform specific, controlled experiments and begin to tease apart the mechanisms by which biological systems operate. The Sequence Hypothesis transformed molecular biology from a descriptive, mysterious field of study to a mechanistic, hypothesis-driven hard science in which researchers can use experiments to ask and answer specific questions. It was the beginning of the biotechnology revolution, a framework in which we ignorant scientists can harness the results of billions of years of biological evolution and direct the evolution of biological function to suit our ends.

1.2 History of Biomolecule Sequencing

With the realization that biology can be thought of as a problem in sequence informatics, it is unsurprising that progress in molecular biology closely follows the ability to read and write defined sequences of DNA, RNA, and protein. Sequencers and synthesizers are the tools by which we have begun to wade into the vastness of sequence space. Things started slowly, with the determination of individual sequences earning at first Nobel Prizes and then entire doctoral theses. In the early 1950s, Fred Sanger and Hans Tuppy reported the primary structure of of a biomacromolecule for the first time (Sanger and Tuppy 1951). Employing a combination of chemical modifications and chromatography, they determined the amino acid sequence of insulin and hypothesized that all proteins, by extension, were defined by specific amino acid sequences. The impact of this result simply cannot be overstated. The exquisite properties of a highly evolved biomolecule could be now be explained, in full, with nothing more than a sequence of letters. The underlying chemistry of protein folding and reactive group position and function is still far from understood, yet any molecular biologist can describe and even reproduce insulin's complex functionality from its sequence alone. Sanger's work proved highly influential in the development of Crick's Sequence Hypothesis, which linked protein sequences with those of DNA, the genetic material.

At first, nucleic acid sequencing lagged behind protein sequencing. Most methods for DNA sequencing rely on synthetic DNA primers and DNA polymerases, two reagents that were not widely available until the 1970s. But the relative uniformity of the DNA double-helix along with the abundance of natural enzymes evolved to manipulate nucleic acids eventually propelled DNA sequencing to the forefront of molecular biology research. Sanger sequencing, the most successful DNA sequencing approach, was introduced by Sanger and colleagues in 1977 (Sanger, Nicklen, and Coulson 1977). It involves enzymatic incorporation of modified chain-terminating dideoxynucleotides into a growing DNA strand complementary to a template of interest. This reaction produces new DNA strands terminated at all positions along the template molecule. The identity of the last, terminating base of every truncated sequence can be determined by a number of methods, the most popular being fluorescence detection in a polyacrylamide gel.

For 40 years, Sanger sequencing has been a mainstay in molecular biology, culminating in the determination of the human genome in the early 2000s. It was during the push to unlock the secrets of the human genome that a radically different approach to DNA sequencing was brought to prominence. High-throughput, or ”shotgun”, DNA sequencing involves sequencing many (currently 10^6-10^9) DNAs in parallel (Staden 1979, Anderson 1981). The information from each sequenced DNA molecule, or ”read”, is typically 20-500bp of sequence along with corresponding quality metrics for each base. Illumina, Inc. sells the most commercially successful platforms. On these machines, a glass slide coated in DNA, termed a flow-cell, is used to capture DNA molecules to be sequenced. The molecules are amplified by surface-bound primers to create clonal colonies spread across the flow-cell, and fluorescent reversible chain terminators are used to read out the DNA bases one-by-one for the entire flow-cell. The result is millions of short sequencing reads produced in a short timespan (days) at very low cost (Reuter, Spacek, and M. P. Snyder 2015).

1.3 Single-Cell RNA-Sequencing

In parallel with the emergence of the first successful strategies for high-throughput DNA sequencing, microarray technology was the subject of intensive development. Microarrays are capable of monitoring gene expression patterns for thousands of genes in parallel (Schena et al. 1995), and have been adapted to a wide range of functional assays (Hoheisel 2006). Biologists were quick to realize that highly parallel assays were key to understanding biological phenomena that operate at a scale and complexity beyond the reach of individual experiments. The field of systems biology emerged, revealing gene networks that govern cellular functions. In the past decade, the availability of cheap, massively-parallel DNA sequencing has sparked an explosion of technological advances aiming to convert traditional, singlet biochemical experiments into highly multiplexed sequencing-based assays. One of the most powerful such applications is RNA-sequencing (RNA-seq), in which an RNA sample is enzymatically converted into a corresponding complementary DNA (cDNA) library to be analyzed by sequencing. In RNA-seq experiments, the sequencer is not used for sequence analysis or assembly but rather as a way to count RNA molecules according to their sequences and relative abundances. Originally developed in 2008 (Nagalakshmi et al. 2008, Mortazavi et al. 2008, Lister et al. 2008), RNA-seq provides a snapshot of the RNA content of a sample. Thousands of genes can be detected and quantified for each sample, and much research has focused on problems associated with generating cDNA libraries representative of the input RNA as well as computational tools to quickly and accurately analyze individual samples. Success in these areas has prompted adoption of RNA-seq as an experimental platform (Wang, Gerstein, and M. Snyder 2009. In the context of controlled experiments, RNA-seq gives researches a quantitative, genome-wide view of the cellular response to experimental variables. One aim of systems biology is the integration of high-throughput experimentation and analysis into more realistic models for cellular behavior that incorporate the complex interactions between various pathways and functions in the cell.

While some researchers were busy perfecting RNA-seq for experimental use, others were thinking small and working to reduce the amount of RNA input required to produce a viable cDNA library. Just a year after the initial demonstrations of RNA-seq, the first report of single-cell RNA-seq emerged (Tang et al. 2009). Scaling the quantitative, transcriptome-wide information produced by RNA-seq down to the level of individual cells represented a major milestone in biological research. This approach yielded truly high-dimensional, meaningful biological information at the level of the cell, the basic functional biological unit, and bridged the gap between cytometry (single-cell characterization) and genomics (genome-wide functional profiling). What has followed is nothing short of an explosion of interest in single-cell RNA-sequencing (scRNA-seq), one of the most rapidly advancing areas in biology. Since 2009, scRNA-seq has incorporated rare cell isolation (Torre et al. 2018), spatial information (FISH) (Karaiskos et al. 2017), protein quantification (Stoeckius, Hafemeister, et al. 2017), Peterson et al. 2017), and electrophysiology (Cadwell et al. 2016) while being miniaturized, optimized, and parallelized. Correspondingly, analysis of scRNA-seq data is now a major topic of interest in computational biology, with mathematicians and computer scientists alike racing to keep up with the newest datasets and prompting still further advances in experimental approaches.

Initially, scRNA-seq was carried out on individual cells one-by-one. A major step forward was the exploitation of a curious activity of certain reverse transcriptase enzymes known as template switching (Picelli, Björklund, et al. 2013, Picelli, Faridani, et al. 2014). The first step of template-switching involves the addition of non-templated bases at the end of an RNA-DNA or DNA-DNA duplex. For biotechnology applications, the MMLV reverse transcriptase displays a strong preference for the addition of non-templated deoxycytosine bases to the 3'-end of the first-strand cDNA synthesis product. A template switching oligo or TSO with several deoxyguanosine bases at its 3' end can base-pair with non-templated bases added by the reverse transcriptase, positioning the TSO to serve as a template for further extension of the first-strand cDNA by the reverse transcriptase. Template switching provides an efficient means to append specific nucleic acid sequences to the 3' end of cDNAs, addressing a major problem in low-input reverse transcription reactions. With a primer binding site added to the 5' end of first-strand cDNAs via a poly(dT) capture primer and a 3' primer binding site attached via template switching, a poly(A)-selected cDNA library can be easily and efficiently amplified from the RNA content of a single cell. Template switching, under the moniker SMART-Seq, was first used for scRNA-seq in 2012 (Picelli, Björklund, et al. 2013).

Concurrent with advances in molecular biology that enabled cDNA library preparation from limiting inputs, great strides were being made to miniaturize and parallelize biochemical reactions. The field of microfluidics, once strictly the realm of physicists, has come to play a central role in molecular biology. Microfluidics is a broad term applied to manipulation of fluids at small (nanometer to millimeter) scale. The devices used in microfluidics experiments vary greatly in design and purpose, and various biotechnologies, from electrophoresis to PCR, have been adapted to the microfluidic scale, resulting in drastic reductions in reagent cost, sample requirements, and time for many applications (Sackmann, Fulton, and Beebe 2014). Three approaches have played prominent roles in the recent history of scRNA-seq. The first was a commercial platform offered by Fluidigm, the C1 chip, capable of generating scRNA-seq libraries for dozens of cells in parallel. The C1 chip uses an extensively engineered series of channels that sequentially add reagents to an increasing reaction volume. Individual cells are captured by size in small traps, then flowed into one of 96 individual reaction chambers. For scRNA-seq, cells are lysed, reverse-transcribed, amplified, and fragmented in a series of discrete steps, resulting in full-length cDNA libraries with sequencing adapters specific for each channel on the microfluidic device (Xin et al. 2016). Compared to alternative scRNA-seq platforms, Fluidigm libraries show very high sequence complexity and are often sequenced very deeply (millions of reads per cell). The Fluidigm C1 chip gave scientists ready access to scRNA-seq, and some progress has been made in adapting this chip design for other molecular assays, most notably single-cell ATAC-seq (Buenrostro et al. 2015), which uses integration events by Tn5 transposase as a proxy for chromatin accessibility. But materials costs to operate Fluidigm chips prevented massively parallel experimentation, and most studies were limited to fewer than 10,000 cells.

A remarkable breakthrough was made in 2015 when two groups succeeded in manufacturing barcoded microparticles for scRNA-seq, abandoning traditional well-based library preparation entirely. Rather than miniaturizing and parallelizing known cDNA synthesis reactions, Drop-Seq (Macosko et al. 2015) and inDrops (Klein et al. 2015) were designed from the ground up to utilize advanced microfluidic technologies and generate truly massive single-cell libraries. The challenge was to generate large numbers of cell barcodes for parallel library preparation, and both groups realized that barcoded, microparticle-immobilized primers could be generated by split-pool synthesis, resulting in clonal particles, or beads, in which each bead contains many copies of the same DNA sequence but is also unique from the other clonal beads in the population. By isolating individual cells with individual beads containing barcoded cDNA capture probes, the RNA content of a cell can be uniquely tagged with a single, cell-specific barcode sequence. After cDNA synthesis, the cDNAs from many cells can be pooled and amplified in a single batch, eliminating the need for extremely sensitive or specialized molecular biology procedures.

In the Drop-Seq procedure (Macosko et al. 2015), barcoded beads are synthesized using the same phosphoramidite chemistry used to make standard DNA oligos. First, a bead-bound DNA primer is synthesized using a non-cleavable linker, giving a homogeneous population of beads, each coated with millions of copies of the same DNA oligo. Next, a series of split-pool steps are performed in which the beads are split into four groups, each of which receives just one of the four standard DNA bases. After a round of single-base extension, the beads are pooled, mixed, and redistributed into four new groups. The split-pool synthesis is repeated a total of 12 times to give over 16 million possible bead barcodes, with each bead coated in many clones of a single barcode from the theoretical set. The beads are then subjected to 8 rounds of random single-base extensions, giving each oligo its own 8 bp unique molecular identifier (UMI). Finally, 25 deoxythymidine bases are added to serve as a poly(dT) capture sequence to enrich for polyadenylated mRNA molecules. The result is a population of barcoded beads with the following structure: PCR primer binding site, cell barcode, UMI, poly(dT) capture. These beads are sufficient for scRNA-seq using a droplet generating microfluidic device. Such chips can generate nanoliter droplets at incredible rates (>10kHz).

In a Drop-Seq experiment, a co-flow droplet generator is used in which two aqueous streams are met with a fluorous organic flow at a T-junction, producing highly monodisperse water-in-oil emulsions. The two aqueous streams contain a cell suspension and a bead/lysis buffer suspension, respectively. Upon droplet formation, the aqueous flows are combined in equal proportion and rapidly mixed, lysing any captured cells with a mild detergent. Cells and beads are encapsulated according to Poisson statistics, and in cases where a cell and a bead are co-encapsulated, the cellular mRNAs are released into the aqueous droplet volume where they are captured by annealing to the poly(dT) oligos on the bead. An on-bead reverse transcription step produces bead-bound, cell-specific cDNA libraries which can be amplified, fragmented, and size selected for sequencing. The Drop-seq approach proved enormously successful thanks in large part to an outstanding effort by the McCarroll laboratory to provide detailed instructions as well as the lack of a viable commercial alternative. Today, Drop-seq has fallen out of favor, suffering from double-Poisson capture statistics (scRNA-seq libraries are produced for only 5% of cells) and highly error-prone bead synthesis beyond the capabilities of most molecular biology laboratories.

A contemporaneous, more sophisticated approach to barcoded bead production was also reported (Klein et al. 2015). The inDrops method is based on the same principles as Drop-seq but offers greatly improved cell capture rates and an accessible synthetic route to barcoded microparticles. The plastic microparticles of Drop-seq are replaced with monodisperse hydrogel beads produced on a microfluidic droplet generator. A common, acrydite-modified DNA oligo with a photocleavable moiety is covalently incorporated into a polyacrylamide gel upon droplet formation. The resulting gel beads are modified throughout with a solvent-accessible DNA oligo, and split-pool synthesis is achieved using enzymatic, rather than organic chemical reactions. The bead oligo is extended in a series of split-pool rounds with a panel of 96 different oligos. After three rounds of split-pool, over 884,000 possible barcode sequences are produced. The advantages of enzymatic synthesis will be further discussed in detail in Chapter 4. A second key advantage of hydrogel bead-based approaches is the possibility of super-Poisson loading in a droplet microfluidic device. Hydrogel beads are mechanically deformable, a property which has been exploited to squeeze the beads in single-file through a narrow channel, resulting in a controlled flow rate that can be tuned with the droplet generation rate of the microfluidic device. In this way almost every droplet produced will contain a single barcoded bead, and correspondingly almost every cell will generate a scRNA-seq library. Cell capture rates of 60-70% have been achieved by inDrops and the commercial 10x Genomics platform.

The most recently developed approaches for massively parallel scRNA-seq involve barcoded microparticles but isolate cells and beads in a patterned array of microwells instead of emulsion droplets (Gierahn et al. 2017, Han et al. 2018). In these approaches, beads can be loaded at super-Poisson rates by size, and cells are captured by gravity, settling into the bead-containing wells. Micro-well arrays offer many distinct advantages, including straightforward chip design and fabrication, arbitrary scaling, and ease of use. These features are obvious in a recently published ”Mouse Cell Atlas” (Han et al. 2018) which surveyed all the major mouse organs across a total of 400,000 single-cell transcriptome profiles. As scRNA-seq technology becomes more accessible and library costs continue to drop, experiments of this scale will soon become the norm.

With the advent of barcoded beads, scRNA-seq has exploded in scale. The first Drop-seq protocols generated libraries of about 10,000 cells in a single batch, approximately 100-fold larger than typical scRNA-seq libraries prepared in microtiter plates or on the Fluidigm C1. Today, multiplexing methods enable much greater cell loadings, and libraries of 20,000-50,000 cells can be prepared in a single batch. Combined with plummeting library prep costs, the increased batch size has resulted in the first studies analyzing hundreds of thousands and even millions of cells. In the coming years, the key limitation in scRNA-seq will be sequencing costs, and methods to either selectively sequence cells from large populations or glean key information of fewer reads per cell will be imperative to continue to scale scRNA-seq libraries.

Chapter 2

Highly Multiplexed Single-Cell RNA-Seq for Defining Cell Population and Transcriptomic States

2.1 Chapter Summary

We describe a universal sample multiplexing method for single-cell RNA-seq in which cells are chemically labeled with identifying DNA oligonucleotides. Analysis of a 96-plex perturbation experiment revealed changes in cell population structure and transcriptional states that cannot be discerned from bulk measurements, establishing a cost effective means to survey cell populations from large experiments and clinical samples with the depth and resolution of single-cell RNA-seq.

2.2 Methods

Overview of Cell Tagging Procedure

Barcoded DNA oligonucleotides (tags) are attached to exposed NHS-reactive amines on the cells of interest. Sample tagging is achieved in a one-pot, two-step reaction by exposing cell samples to methyltetrazine-activated DNA (MTZ-DNA) oligos and the amine-reactive cross-linker NHS-trans-cyclooctene (NHS-TCO) (Figure 2.1b). NHS-functionalized oligos are formed in situ via inverse-electron demand Diels-Alder (IEDDA) chemistry, and nucleophilic attack by accessible cellular amines chemoprecipitates the oligos directly onto the cells. Our one-pot reaction based on the IEDDA reaction improves on a previous cell surface modification scheme (Hsiao et al. 2009) that requires far higher DNA concentrations and isolation of unstable activated DNAs immediately before use. A library of methyltetrazine-modified sample tags can be prepared in advance, stored frozen for long periods, and applied to many cell samples in parallel. Sequencing library preparation is derived from recently published methods for multi-modal scRNA-seq (Peterson et al. 2017; Stoeckius, Hafemeister, et al. 2017).

Oligo Activation

Sample tags were prepared with either 5'- or 3'-amine modified oligonucleotides (100-250 nmol scale, Integrated DNA Technologies, Table A.1). HPLC purification was critical to obtain highly reactive preparations of 5'-modified oligos, while 3'-modified oligos can be purchased without HPLC purification (data not shown). In either case, oligos were resuspended to a concentration of 500 µM in 50 mM sodium borate buffer pH 8.5 (Thermo). Activation reactions were performed by combining 25 µL oligo solution with 41.8 µML DMSO (Sigma) and 8.2 µL of 10 mM NHS-methyltetrazine (Click Chemistry Tools). The reaction was allowed to proceed for 30 minutes at room temperature on a rotating platform. After 30 and 60 minutes, additional 8.2 µL aliquots of 10 mM NHS-methyltetrazine were added. After 90 minutes total reaction time, ethanol precipitation was performed by addition of 180 µL 50 mM sodium borate buffer and 30 µL 3 M NaCl. After mixing, 750 µL ice-cold ethanol was added and the mixture precipitated at −80 °C overnight. The precipitate was pelleted at 20,000 x *g* for 30 minutes, washed twice with 1 mL ice-cold 70% ethanol, then resuspended in 100 µL 10 mM HEPES pH 7.2. Yield was determined by absorbance at 260 nm. Typical final concentrations ranged between 40 and 80 µM.

Relative oligo activity was determined by electrophoretic mobility shift assay using Cy5-*trans*-cyclooctene (Click Chemistry Tools). Methyltetrazine-derivatized oligos were diluted 100-fold in 10 mM HEPES pH 7.2, then 4 µL of this solution was added to 1 µL of a 500 nM solution of TCO-Cy5 in DMSO. All tetrazine reactions in this work were protected from light to reduce degradation of *trans*-cyclooctene. The reaction was allowed to proceed at room temperature for 20-120 minutes and analyzed on a 12% SDS-PAGE gel. Oligo activity varied within a 2-fold range across preparations. Oligos were stored at −20 °C and used without further normalization.

Cell Culture and Fixation

Neural stem cells were cultured according to the following protocol: Cryopreserved mouse neural stem cells (NSCs) were thawed for 2 minutes at 37 °C then transferred to a 15 mL conical tube. Pre-warmed Neural Stem Cell Basal Medium (SCM003, Millipore) was slowly added to a total volume of 10 mL, and the resulting cell suspension centrifuged at room temperature for 2.5 minutes at 200 x g. The supernatant was removed and the cell pellet was resuspended in 2 mL pre-warmed Neural Stem Cell Basal Medium and counted on a Countess II Automated Cell Counter (Thermo). Cells were seeded on poly-L-ornithine (Millipore) and laminin (Thermo) coated 100mm culture plates at 700,000 cells per plate in 10 mL of pre-warmed Neural Stem Cell Basal Medium supplemented with EGF (Millipore) and bFGF (Millipore) at 20ng/mL each, heparin (Sigma) at 2 µg/mL, and 1% Penicillin-

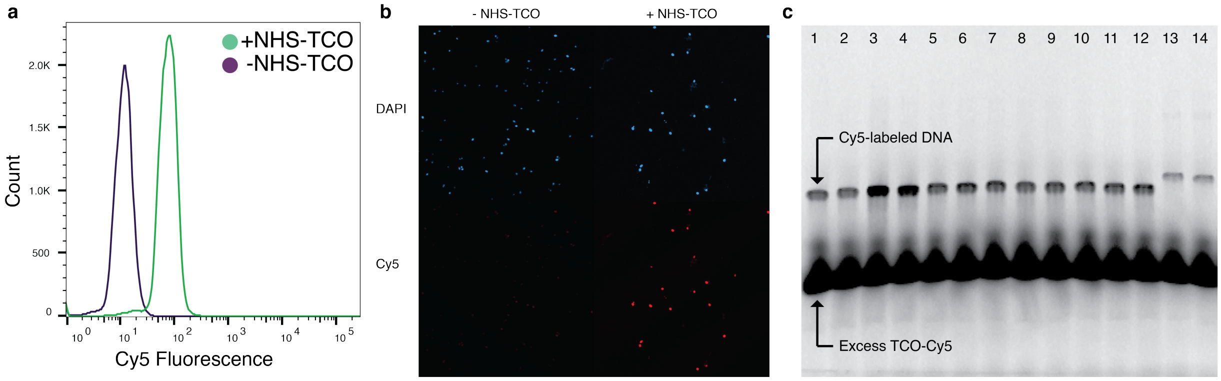


Figure 2.1: Direct cell labeling with Inverse Electron-Demand Diels-Alder (IEDDA) chemistry (a) Yeast cells were fluorescently labeled in a one-pot, two-step reaction with NHS-TCO and MTZ-Cy5. Control reactions omitted NHS-TCO. (b) Fluorescence microscopy of yeast cells labeled with NHS-TCO and MTZ-Cy5 shows labeling only in the presence of NHS-TCO cross-linker. (c) Activity assay for panels of methyltetrazine-activated DNA sample tags. MTZ-DNAs were reacted with TCO-Cy5 and the products separated by polyacrylamide gel electrophoresis. Lanes 1-12 are 3'-amine modified, while lanes 13 and 14 are 5'-amine modified.

Streptomycin (Thermo). Supplemented medium was changed the next day and every other day thereafter until confluent.

Neural stem cells for 96-sample growth factor screen were cultured according to the following protocol after previously described cell culture plate reached 80% confluence: Stock solutions (10x) were prepared in Neural Stem Cell Basal Medium for every factor and at every concentration used: EGF+bFGF at 200 ng/mL, 40 ng/mL, 8 ng/mL, 0 ng/mL; BMP4 (Peprotech) at 200 ng/mL, 40 ng/mL, 8 ng/mL, 0 ng/mL; Retinoic Acid (Sigma) at 10 µM, 2 µM, 0 µM; Scriptaid (Selleckchem)/Decitabine (Selleckchem) at 1 µM/5 µM, 0.2 µM/1 µM, 0 µM/0 µM; heparin at 20 µg/mL + Penicillin-Streptomycin at 10%. 20 µL each of EGF/bFGF, BMP4, Retinoic acid or Scriptaid/Decitabine, heparin, and Penicillin-Streptomycin were added to each well of a poly-L-ornithine and laminin coated 96-well plate for a total of 80 µL.

NSCs previously plated on 100mm culture plates until 80% confluent were dissociated by incubation in 4 mL of ESGRO Complete Accutase (Millipore) for 2 minutes at 37 °C. After incubation, the Accutase and NSCs were transferred to a 15 mL conical tube and centrifuged at room temperature for 2.5 minutes at 200 x *g*. Supernatant was removed and the cell pellet was resuspended in 2 mL Neural Stem Cell Basal Medium. Centrifugation and medium replacement were repeated one more time and cells were counted on a Countess II Automated Cell Counter. The cell suspension was then diluted with additional Neural Stem Cell Basal Medium to a concentration of 18.3 cells/µL. From this stock 120 µL was added to each well of the 96-well plate for a total of 2,200 cells/well. Supplemented media for every well in the 96-well plate was replaced every other day during the 5-day incubation.

Before NSC dissociation and fixation, 80 µL of ice-cold methanol was added to each well of twelve 8-well PCR strips on an ice block. After 5 days in culture, all media in the 96-well plate were removed and the cells washed three times with 150 µL of Neural Stem Cell Basal Medium. Any remaining media were removed and replaced with 20 µL of Accutase and incubated at 37 °C for 2 minutes with gentle pipetting to help break cell clumps. Next, 20 µL of dissociated NSCs in Accutase were transferred to the 8-well strip tubes containing 80 µL of 100% methanol, and the entire volume was pipetted to mix. After fixation, the NSCs were stored at −20 °C until sample labeling.

For 4-sample NSC labeling and species-mixing experiments (below), NSCs were cultured on a 100mm poly-L-ornithine and laminin coated culture plate according to the protocol previously described until 80% confluent. NSCs were dissociated by removing culture medium followed by incubation with 4 mL Accutase for 2 minutes. NSCs in Accutase were transferred to a 15 mL conical tube and centrifuged at room temperature for 2.5 minutes at 200 x *g*. The supernatant was removed and the cell pellet was resuspended in 2 mL Hanks Balanced Salt Solution (HBSS, Thermo) with 0.04% BSA (Sigma). Centrifugation and medium replacement were repeated once and cell concentration was determined on a Countess II Automated Cell Counter. Cells were then fixed by addition of 4 volumes ice-cold methanol added slowly with constant mixing. Fixed cells were stored at −20 °C until sample labeling and scRNA-seq.

Frozen stocks of HEK293T cells (ATCC) were thawed for 2 minutes at 37 °C with gentle agitation. Thawed cells (500 µL) were added to 5 mL pre-warmed media (DMEM (Corning) + 10% FBS (Gemini Bio-Products) + 1% Penicillin-Streptomycin (Corning) and centrifuged at 1,500 x *g* for 5 minutes. The cells were resuspended in 5 mL media and transferred to a T-25 cell culture flask. Cells were grown at 37 °C with 5% CO2 following standard practices. HEK293T cells were dissociated by incubation with TrypLE Select (Thermo) for 5 minutes at 37 °C, washed twice with HBSS, and resuspended in 1 mL at a concentration of 6x10^6 cells/mL. Cell number and viability were measured using a Countess II Automated Cell Counter (ThermoFisher). Four mL ice-cold methanol was added slowly with constant mixing, and the resulting cell suspension incubated at −20 °C for at least 20 minutes. Cells were stored at −20 °C until sample labeling and scRNA-seq.

Flow Cytometry and Fluorescence Microscopy

Yeast cells (Fleischmann's Rapid Rise) were used as an abundant cellular substrate to test cell labeling reactions. Approximately 5 g of dehydrated cells were rehydrated in 4 mL PBS + 0.1% Tween-20 (Sigma) for 10 minutes at room temperature with rotation. One mL of the resulting cell suspension was diluted with 7 mL PBS-Tween and fixed by slow addition of 32 mL ice-cold methanol with constant mixing. Cells were incubated at −20 °C for at least 20 minutes before further use.

Methanol-fixed cells were rehydrated by combining 700 µL HBSS with 500 µL fixed cells in 80% methanol. This suspension was centrifuged at 3,000 x *g* for 5 minutes, then washed twice more with HBSS. Cells were resuspended in 1 mL HBSS, and 50 µL of this cell suspension was used for cell labeling. Methyltetrazine-Cy5 (Click Chemistry Tools) was added to 2 µM final concentration, NHS-TCO to 5 µM, and DAPI to 1 µg/mL. Cell labeling reactions were incubated for 30 minutes at room temperature with rotation then quenched by addition of Tris-HCl to 10 mM and methyltetrazine-DBCO (Click Chemistry Tools) to 50 µM. Samples were diluted 20-fold in HBSS and analyzed on a MACSQuant VYB flow cytometer.

Fluorescence microscopy samples were prepared as above except NHS-TCO was used at 1 µM and MTZ-Cy5 was used at 62.5 µM. Samples were imaged on a Zeiss LSM 800 laser scanning confocal microscope.

Sample Labeling Proof of Concept

Fixed NSCs were split into four aliquots with 400,000 cells in 100 µL 80% methanol. Live NSCs were prepared as described above, washed into HBSS, and similarly aliquoted to four samples with 400,000 cells in 100 µL. Prior to cell labeling, 8 labeling combinations were made by combing 6 µL each of two different sample tags. A 5-minute pre-incubation reaction was performed in the dark at room temperature by addition of 4 µL 1 mM NHS-TCO. After pre-incubation, cell suspensions were thoroughly mixed with the entire volume of a single sample label mix. Cell labeling proceeded for 30 minutes at room temperature on a rotating platform. Reactions were quenched by addition of Tris-HCl to 10 mM final concentration and methyltetrazine-DBCO (Click Chemistry Tools) to 50 µM final concentration. After quenching for 5 minutes, cells were pooled to create a single sample for fixed cells and a single sample for live cells. The two samples were combined with two volumes PBS-BSA and pelleted by centrifugation at 500 x *g* for 5 minutes. Cells were washed three times with PBS-BSA and vigorously resuspended in a final volume of 150 µL. Cells were analyzed and counted, then fixed and live samples were combined at equal concentration and loaded onto a single lane of the Chromium Controller (10x Genomics, Inc.) targeting 10,000 cells. Library preparation was adapted from the REAP-Seq protocol (Peterson et al. 2017). The 10x Genomics v2 Single Cell 3' Seq Reagent kit protocol (10x Genomics) was used to process samples according to the manufacturer's procedure with modifications as follows. After initial ampilifcation of cDNA and sample tags, the two libraries were separated during SPRI size-selection. The manufacturers instructions were used to complete cDNA library preparation. For sample tags, rather than discarding 80 µL SPRI supernatant, this fraction was added to 45 µL SPRI beads and incubated at room temperature for 5 min. The SPRI beads were washed twice with 80% EtOH and sample tags eluted in 20 µL nuclease-free water. Sample tags were quantified by Qubit High-Sensitivity DNA Assay (Invitrogen) and amplified using primer R1-P5 and indexed reverse primers as appropriate (Table A.1). PCR was performed in a 25 µL volume including 2.5 µL sample tag library, 1.5 µL of 10 µM forward and reverse primer, 7 µL nuclease-free water, and 12.5 µL KAPA 2x HIFI PCR master mix (Kapa Biosystems). The samples were cycled as follows: 98 °C 3 min, 16 cycles of: 98 °C 20 sec, 58 °C 30 sec, and 72 °C 20 sec; and then a final extension step of 72 °C for 4 min. Final sample tag libraries were obtained using a PippinPrep automated size selection system with

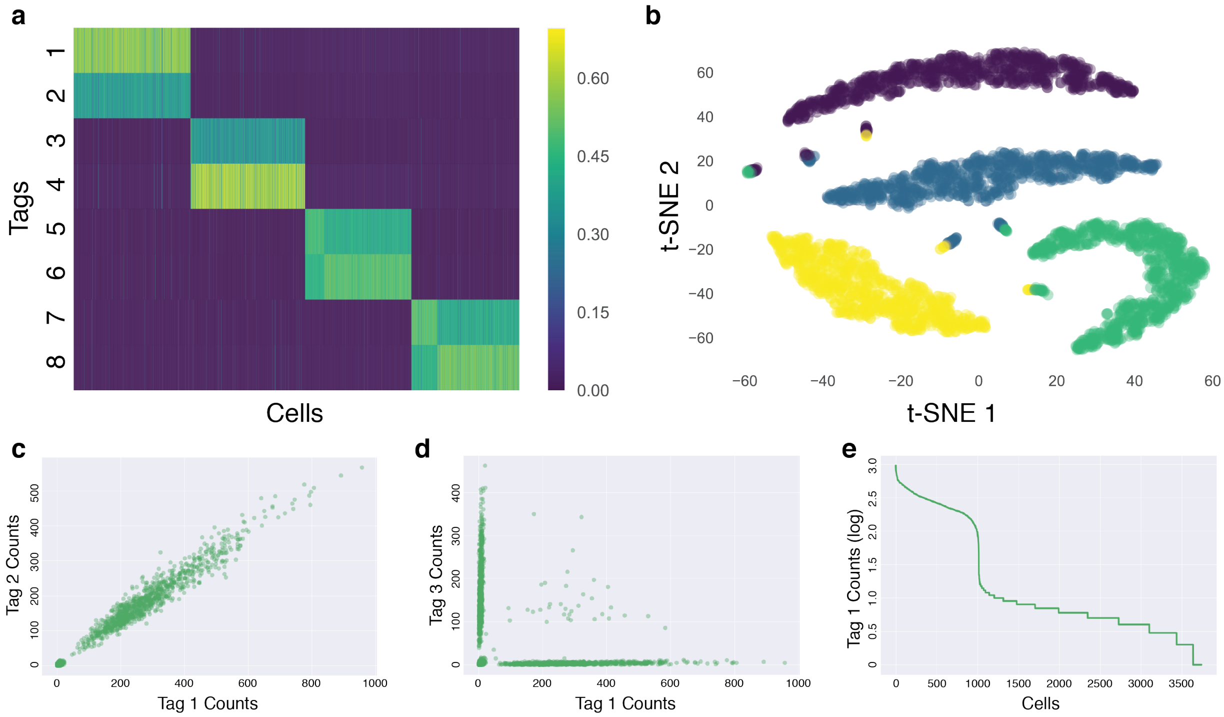


Figure 2.2: Proof-of-concept sample tagging experiment (a) Heatmap showing 3,768 detected cells originating from four methanol-fixed samples each labeled by a pair of sample-specific tags. (b) t-SNE visualization of sample tag data colored by k-means clustering (k=4). Four main clusters are observed, corresponding to the four individual samples, as well as 6 = (42) small clusters corresponding to each possible combination of cell doublet originating from two different samples. (c) Scatter plot of counts for tags 1 and 2, which were used to label the same sample. The low-count population (bottom-left) is background from droplets not containing cells from the sample, while the high-count population corresponds to positive cells from the sample and shows a striking correlation between the two tag counts (Pearsons correlation coefficient r = 0.96). (d) Barnyard plot showing two tags from separate samples. Tags are clearly orthogonal, with doublets easily identified. (e) Counts for tag 1 from each cell in the experiment, ordered from highest to lowest and showing a clear inflection point between Tag 1 (+) and Tag 1 (-) cells.

a 3% agarose gel set for a broad purification range from 200-250 bp (target library size is 225 bp). A Qubit assay was again used to determine library concentration for sequencing. Sample tag and cDNA libraries were analyzed on a BioAnalyzer High Sensitivity DNA kit (Agilent). Example traces are provided for reference (Figure 2.5). Sample tag libraries were sequenced on an Illumina MiSeq using a MiSeq V3 150 cycle kit (26x98bp reads), and cDNA libraries were sequenced on an Illumina HiSeq 4000 using a HiSeq SBS 3000/4000 SBS 300 cycle kit (2x150bp reads).

Species Mixing and Sample Label Multiplexing

Methanol-fixed human HEK293T and mouse NSCs were prepared as described above. Samples were labeled with non-overlapping tags sets of increasing size (Table A). Suspensions of both cell types were prepared at 700,000 cells/mL in 80% methanol. Samples of 100 µL were prepared for each condition, with species mixing conditions comprising 50 µL of cell suspension from each species. For this experiment, 3'-modified oligos isolated by standard desalting were used as opposed to the 5'-modified, HPLC-purified oligos used in all other experiments presented. Tag sets were prepared by reacting 6 µL of each oligo along with 2 µL of 1 mM NHS-TCO per oligo at room temperature. After 5 minutes, the entire volume of each tag set was added to the appropriate cell suspension. Cell labeling was performed for 30 minutes at room temperature on a rotating platform. Reactions were quenched as above, pooled, and added to 2 mL PBS + 1%BSA. Samples were split across two Eppendorf tubes and centrifuged at 500 x *g* for 5 minutes. Cell pellets were resuspended in 500 µL PBS-BSA, combined, and centrifuged once more. The cell pellet was washed twice more with 1 mL PBS-BSA. Finally, the cells were resuspended in 150 µL PBS-BSA, counted, and diluted to 1x10^6 cells/mL and loaded on a single lane of the Chromium Controller targeting 12,000 cells. Sample tag and cDNA libraries were prepared as described. Libraries were submitted as part of an Illumina NovaSeq library, targeting 500 M reads total (2x150bp reads), with sample tags submitted at 10% of the total library concentration.

96-Sample Growth Factor Screen

Cells for the 96-sample perturbation experiment were prepared as described above. For each sample, two sample tags (6 µL each) were combined with 4 µL 1 mM NHS-TCO according an 8x12 matrix. Columns 1-12 of the 96-well plate correspond to tags BC21-BC32, while rows A-H correspond to tags BC33-BC40 (Table A.1). Fixed cells from each experimental condition (100 µL) were labeled with the entire volume of the corresponding sample tag mix for 30 minutes at room temperature on a rotating platform. Samples were quenched as described above, pooled, and combined with 15 mL PBS-BSA. Samples were split across two 15-mL conical tubes and spun at 500 x *g* for 5 minutes. Cell pellets were resuspended in 3 mL PBS-BSA each and centrifuged again. The pellets were washed twice with one mL PBS-BSA and resuspended in a final combined volume of 200 µL. Cells were loaded on two lanes of the 10x Chromium Controller targeting 10,000 cells per lane. Sequencing libraries were prepared as described, with sample tag libraries sequenced on two lanes of Illumina MiSeq using MiSeq v3 150 cycle kits (26x98bp reads), and cDNA libraries pooled and sequenced on Illumina HiSeq 4000 using two HiSeq 3000/4000 SBS 300 cycle kits (2x150bp reads).

cDNA Data Processing

Standard bioinformatics tools were used to process and analyze DNA sequencing information. Raw sequencing data were processed using the 10x Genomics Cell Ranger pipeline

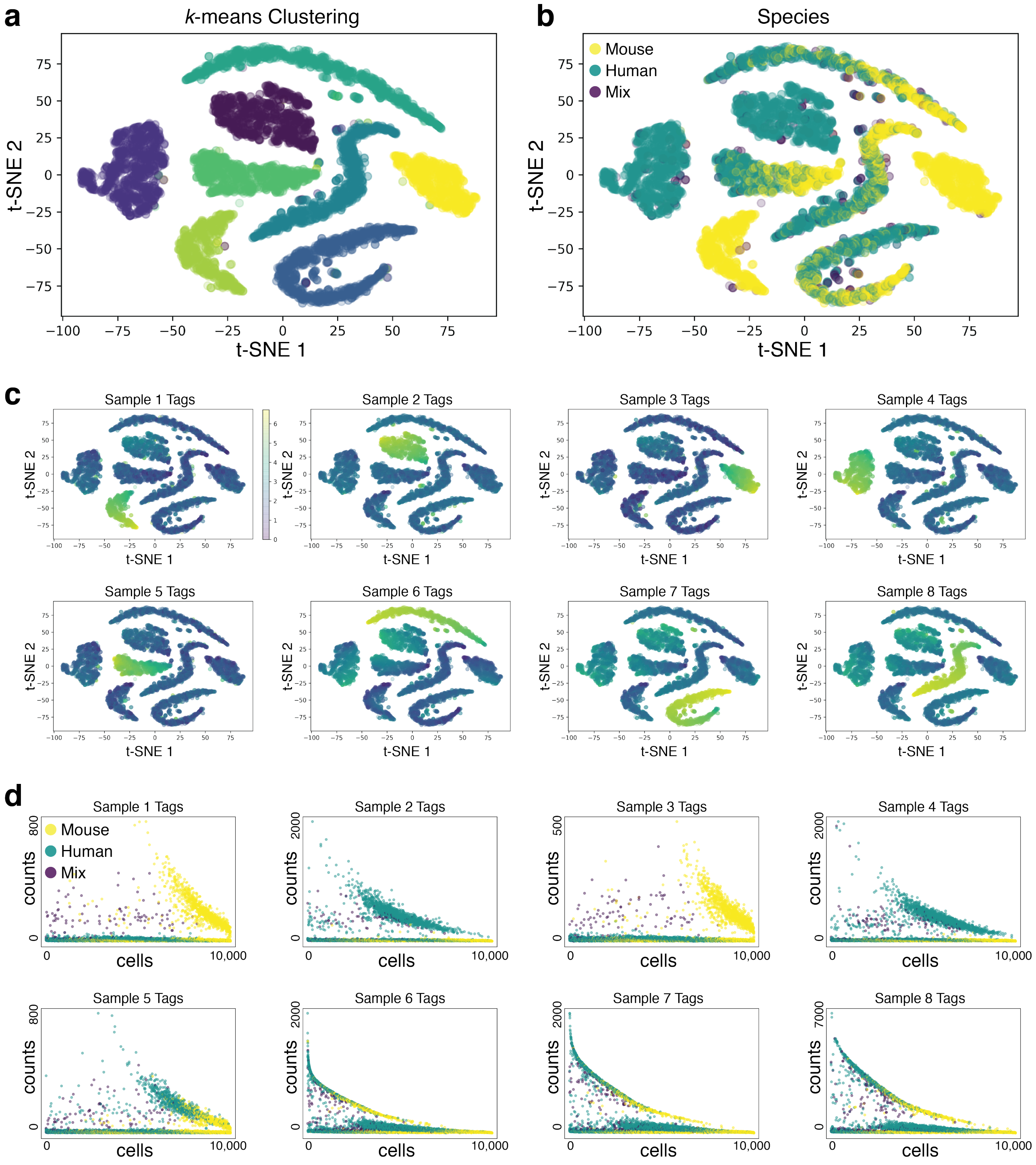


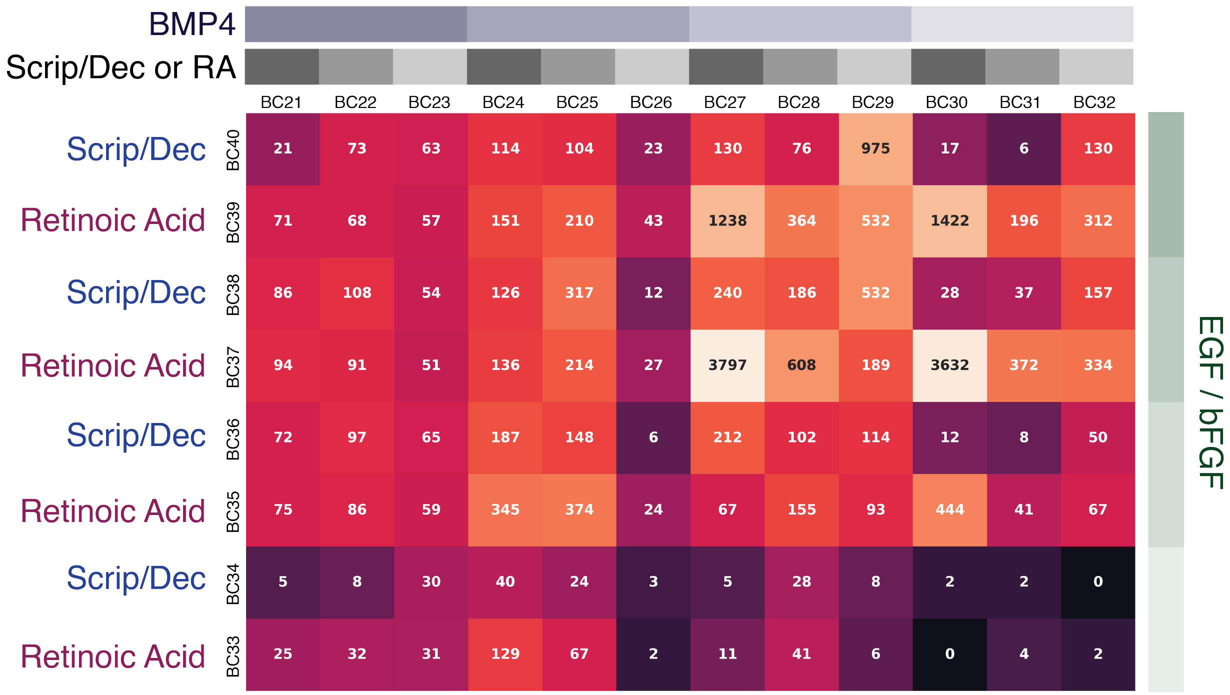
Figure 2.3: Species Mixing and Tag Multiplexing. (Continued on the following page.)

Figure 2.3: Species Mixing and Tag Multiplexing. Mouse neural stem cells and human HEK293T cells were labeled as follows: Sample 1: mouse, one label; Sample 2: human, one label; Sample 3: mouse, two labels; Sample 4: human, two labels; Sample 5: mix, two labels; Sample 6: mix, three labels; Sample 7: mix, 4 labels; Sample 8: mix, 5 labels (a) 10,054 cells were detected. t-SNE of sample tags x cells count matrix, colored by sample assignment from k-means clustering performed on a matrix normalized for tag numbers and counts per cell. Eight major clusters are clearly identified. (b) t-SNE colored according to species assignment based on cDNA content. Four clusters represent a single species, and the remaining four are mixed, concordant with the experimental design. Cells identified as a mix of human and mouse are explained as cell doublets, and as expected fall outside of the major clusters, indicating a mix of sample tag signals as well. (c) t-SNE representation with detected cells colored as the logarithm of the sum of sample tags used in each of the eight experimental samples. (d) Sum of sample tag counts for each sample across all detected cells. Smaller NSCs present fewer sample tags than HEK293Ts from the same samples, indicating a correlation between cell size and the extent of labeling.

(version 2.0). Cellranger mkfastq was used to demultiplex libraries based on sample indices and convert the barcode and read data to FASTQ files. Cellranger count was used to identify cell barcodes and align reads to mouse or human transcriptomes (mm10 and hg19) as appropriate. For the 96-sample perturbation experiment, cellranger aggr was used to combine and normalize sequencing data from the two 10x lanes split across two HiSeq lanes. Cells were selected by cellranger using the inflection point of detected cell numbers as a function of ordered read counts as a cutoff. For the sample labeling proof of concept and species mixing experiments, no further analysis of the cDNA data was performed.

Sample Tag Data Processing and Assignment

Sequencing reads from sample tag libraries were processed using cellranger and synthetic transcriptomes corresponding to the sequences of the tags used in a given experiment. Cellranger count outputs a post-sorted genome BAM file containing error-corrected cell barcodes and UMIs as well as read2 sequence containing sample tag information. The post-sorted genome BAM file was used to generate a digital count matrix for the sample tags corresponding to each cell barcode. A modified version of CITE-Seq Count (Stoeckius, Hafemeister, et al. 2017) was used to count sample tag data. Briefly, a fuzzy matching package, fuzzywuzzy (https://github.com/seatgeek/fuzzywuzzy), was implemented to find the sample barcode region in staggered sample tag libraries that were synthesized to improve sequencing quality. Tag reads were summed according to the combinations used in a given experiment, and sample calling was based simply on the sample with the highest number of reads. Sample assignment was performed by querying the sample tag matrix with cell barcodes identified from cDNA data, generating a vector of sample assignments that can be input into standard scRNA-seq analysis packages. For the species mixing experiment (Table A), in which up



		BC21	BC22	BC23	BC24	BC25	BC26	BC27	BC28	BC29	BC30	BC31	BC32
Scrip/Dec	BC40	21	73	63	114	104	23	130	76	975	17	6	130
Retinoic Acid	BC39	71	68	57	151	210	43	1238	364	532	1422	196	312
Scrip/Dec	BC38	86	108	54	126	317	12	240	186	532	28	37	157
Retinoic Acid	BC37	94	91	51	136	214	27	3797	608	189	3632	372	334
Scrip/Dec	BC36	72	97	65	187	148	6	212	102	114	12	8	50
Retinoic Acid	BC35	75	86	59	345	374	24	67	155	93	444	41	67
Scrip/Dec	BC34	5	8	30	40	24	3	5	28	8	2	2	0
Retinoic Acid	BC33	25	32	31	129	67	2	11	41	6	0	4	2

Figure 2.4: Organization of 96-plex perturbation experiment. Matrix entries correspond to the number of cells recovered from each sample.

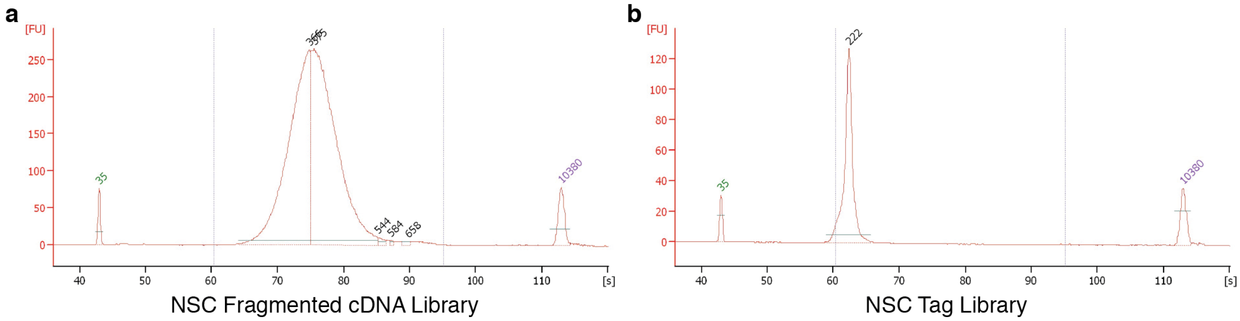


Figure 2.5: Representative BioAnalyzer traces for (a) fragmented cDNA libraries and (b) sample tag libraries.

to five tags were used for each cell, t-SNE was performed on the sample tags x cells count matrix while k-means clustering was performed on a normalized count matrix in which the counts corresponding to each cell were first (1) collapsed and normalized according to the tag sets used by adding the tag counts corresponding to each sample and dividing by the size of the tag set then (2) dividing each normalized sample count by the sum of all normalized samples for that cell.

Data Analysis

For the 96-sample perturbation experiment, the ScanPy Python package (version 1.0.4, Wolf, Angerer, and Theis 2018) was used to process the filtered genes x cells matrix produced by cellranger. The data was log transformed, normalized per cell, and highly variable genes were selected as those with mean normalized counts > 0.0125 and < 3 and with dispersion > 0.5, giving 1,221 highly variable genes. The per-cell read counts were regressed out and the data scaled to unit variance. PCA was performed on this matrix, followed by t-SNE visualization based on the top 20 principal components. Clustering was performed using the neighbors and louvain tools in ScanPy with the size of the local neighborhood set to 30. For clustering based on Louvain community detection, the resolution parameter was adjusted to agree well with subpopulations produced by the perturbation experiment. We reasoned that these natural groupings represent reproducible, quantitatively distinct biological states under the conditions of our experiment and would thus hold the most information relevant to the changing experimental parameters. In practice, a resolution setting of 2 yielded clusters that agreed quite well with the sample-specific subpopulations produced by the perturbation experiment. Sample assignments were combined with cluster assignments from each cell to produce a matrix of cluster occupancy x experimental condition as well as a normalized version of the same matrix showing cluster relative abundance for each sample (Figure 2.7a). Principal component analysis was performed on the cluster relative abundance matrix to visualize relationships between the experimental conditions used in our perturbation (Figure 2.7b). Differential expression analysis was performed with the rank_genes_groups function in ScanPy. The top differential genes between the cluster(s) of interest and the rest of the dataset are shown (Figure 2.7c,d).

2.3 Results

Massively parallelized single-cell RNA-sequencing (scRNA-seq) is transforming our view of complex tissues and yielding new insights into functional states of heterogeneous cell populations. Currently, individual scRNA-seq experiments can routinely probe the transcriptomes of more than ten thousand cells (G. X. Y. Zheng et al. 2017, Svensson, Vento-Tormo, and Teichmann 2018), and in the past year the first datasets approaching and exceeding one million cells have been reported (*Datasets - Single Cell Gene Expression - Official 10x Genomics Support* 2018, Han et al. 2018). However, despite numerous technical breakthroughs that have increased cell capacity of many scRNA-seq platforms, researchers are at present limited in the number of samples that can be assayed. Many biological and therapeutic problems rely on finding genes or signals responsible for a phenotype of interest, but the enormous space of possible variables calls for screening hundreds, or even thousands, of conditions. At present, analyzing genetic, signaling, and drug perturbations (and their combinations) at scale with scRNA-seq is impeded by microfluidic device operation, high reagent costs, and batch effect. While a multiplexing method based on epitope expression has been developed

(Stoeckius, S. Zheng, et al. 2017, Peterson et al. 2017), it can only be practically applied to about a dozen samples. The in silico demuxlet algorithm (Kang et al. 2018) is more scalable but requires samples from distinct genetic backgrounds.

The scRNA-seq sample multiplexing method presented here allows for cells from individual samples to be rapidly chemically labeled with identifying DNA oligonucleotides, or sample tags (Figure 2.6). This universal approach can be applied to cells from any organism without the need for specific epitopes, sequence markers, or genetic manipulation, and is compatible with any scRNA-seq protocol based on poly(A) capture. We demonstrate the utility and versatility of our technique in the context of a multifaceted experimental perturbation in which neural stem cells (NSCs) were exposed to 96 unique combinations of growth factors, with the perturbed cell populations profiled as a single pooled library (Figure 2.6a). Despite the cell capacity of scRNA-seq platforms, single-cell transcriptome-wide analysis of such an experiment, which produces a unique cell population in each condition, has been technically and financially inaccessible in the absence of a suitable means of sample pooling. This experiment introduces a powerful new experimental and analytical paradigm, underpinned by our flexible, scalable cell tagging procedure, in which the massive cell capacity of scRNA-seq is effectively leveraged to analyze and compare large numbers of cell populations.

Neural stem cells (NSCs) are known to differentiate into many unique cell types in vivo, primarily neurons, astrocytes, and oligodendrocytes (Bond, Ming, and Song 2015). In vitro, NSCs can be forced into different differentiation trajectories by exposing the cells to a variety of synthetic chemicals, hormones, and growth factors. We investigated the response of NSCs to varying concentrations of Scriptaid/Decitabine, epidermal growth factor (EGF)/basic fibroblast growth factor (bFGF), retinoic acid, and bone morphogenic protein 4 (BMP4), producing a 4x4x6 perturbation array representing a large space of experimental conditions (Figure 2.6a). NSCs were cultured in a single 96-well plate with each sample corresponding to a unique combination of factors (Figures 2.4, 2.6c). After chemical DNA labeling (Figure 2.6b), the samples were pooled and subjected to a modified version of the 10x Genomics Single-Cell Expression protocol. A total of 21,232 cells were detected based on cDNA counts, and sample assignment was performed for the detected cells based on the sample tags with the highest UMI counts.

Visualization of the cell populations produced by each experimental condition revealed a complex interplay between the perturbations used in this 96-plex experimental space (Figure 2.6e). On a global level, cell proliferation varied widely across the experiment, revealing growth rates specific not just to each condition but also to each cell state across the experiment. Highly proliferative states (clusters 1, 2, 3, 6, 7, 8, 9, 12, and 16), which account for large regions of the cell state space when plotted according to t-SNE, differentially express various genes associated with cell growth and the cell cycle, including ribosomal, cytoskeletal, and cyclin-dependent proteins. Conversely, samples deprived of EGF/bFGF exhibited apoptotic phenotypes including low cell counts and expression of stress response genes such as Cryab, Mt1, and Gpx4. We sought to define the cell states produced by the array of experimental conditions, a frequently challenging procedure in scRNA-seq analysis and a potential roadblock to perturbation experiments where the presence of classical marker genes

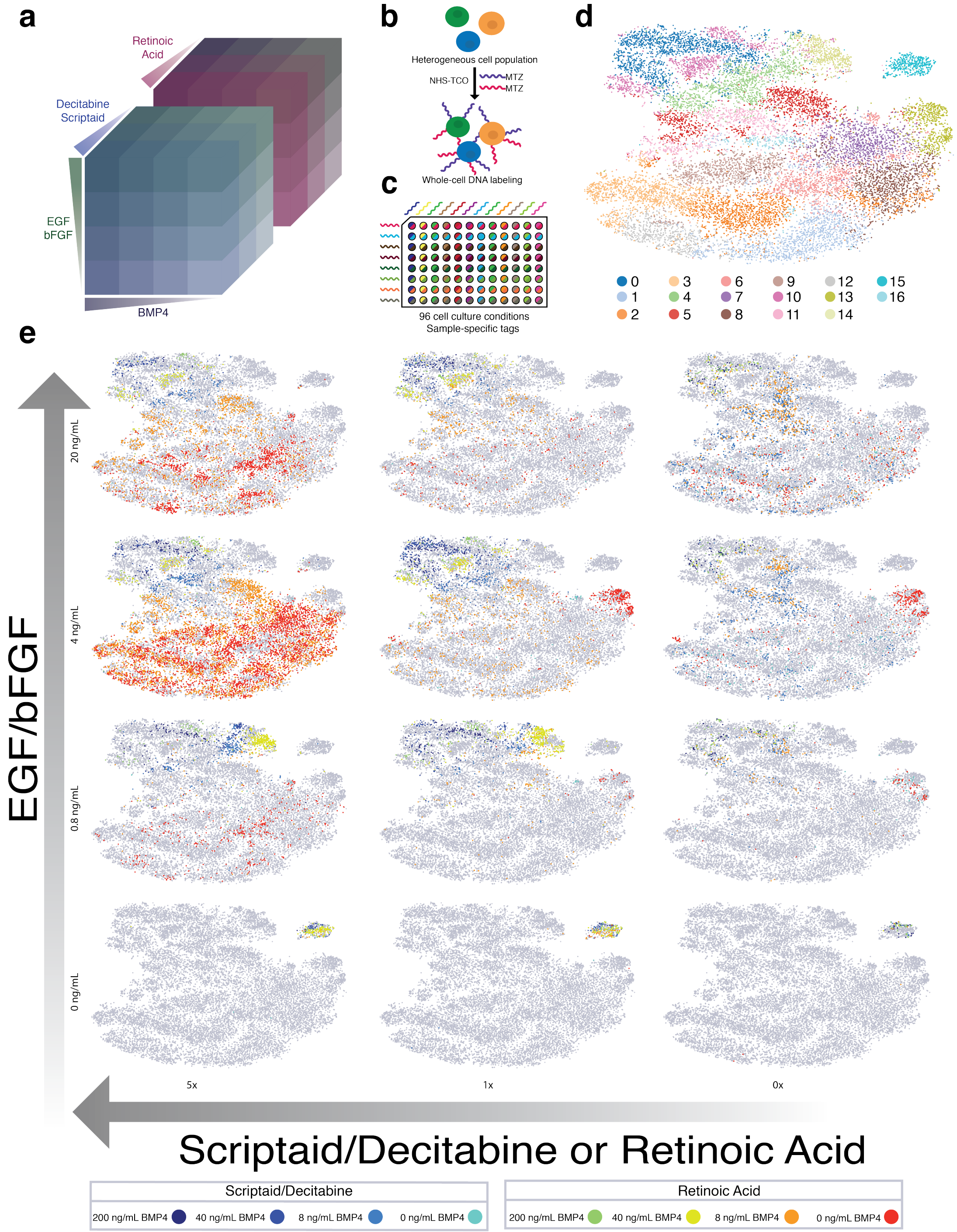


Figure 2.6: 96-Plex scRNA-seq experiment. (Continued on the following page.)

Figure 2.6: 96-Plex scRNA-seq experiment (a) Four experimental factors (EGF/bFGF, BMP-4, Decitabine/Scriptaid, and Retinoic acid) were titrated against one another to produce an array of 96 unique perturbations. (b) Prior to scRNA-seq, a one-pot, two-step reaction with MTZ-DNA and NHS-TCO labeled cells with sample-specific tags (c) Neural stem cells subjected to a 96-plex array of growth conditions were dual-labeled with a unique pair of sample tags (d) t-SNE of 21,232 cells from 96-plex perturbation. Cluster assignments closely match population behavior driven by experimental parameters (e) Visualization of cell populations produced by each experimental condition. Each t-SNE corresponds to a given EGF/bFGF concentration against a series of retinoic acid or Scriptaid/Decitabine concentrations and displays eight samples colored by BMP-4 concentration.

may depend on experimental conditions. Identification of functional cell states was greatly aided by the large number of samples in our experimental perturbation. Various distinct regions of transcriptome space were repeatedly populated by cells originating from multiple samples in localized regions of perturbation space, forming natural groupings of cells that were validated and assigned by clustering using Louvain community detection (Figure 2.6d). Plotting the cluster occupancy of each sample revealed the structure of the cell populations produced across the experiment (Figure 2.7a). Overall trends, such as high proliferation under low BMP4 conditions and high cluster specificity under high BMP4 conditions, are readily observed. Principal component analysis of the relative cluster abundance x sample matrix was used to identify relationships between the experimental inputs (Figure 2.7b). The experimental perturbations associate directly with the cell populations observed in the scRNA-seq samples. The absence of EGF/bFGF has a drastic effect, yielding an isolated group of samples, while BMP4 concentration has a graded effect and a strong interaction with either Scriptaid/Decitabine or retinoic acid, each of which produces a separate branch of samples when combined with the two highest BMP4 concentrations. This analysis demonstrates that multiplexed scRNA-seq can be used to classify cell populations and interpret the conditions that produced them. In the context of a perturbation experiment, relevant features of the experimental space can be learned, e.g. the strong effect of BMP4 concentration shown here. Of perhaps greater interest would be to extend this proof-of-principle to biomedical diagnostics: by applying Bayes Rule to the relative cluster abundance x samples matrix, it should be possible to infer disease conditions from high-resolution cell population observations.

After evaluating the high-level information that can be gleaned from a large perturbation array, we closely examined two regions of our experimental space to illustrate the depth of analysis afforded by multiplexed scRNA-seq. First, we explored an isolated portion of cell state space, cluster 13, which was populated under a strict range of conditions with intermediate EGF/bFGF concentrations, no BMP4, and moderate to no retinoic acid. Cells from just five samples accounted for practically all the cells in cluster 13 and little across the rest of cell state space, exhibiting strong condition dependence (Figure 2.7c). Differential expression analysis showed that this cluster is strongly enriched for Hes5, a gene with important roles in cell fate determination (Imayoshi et al. 2013).

A more complex cellular response was observed under high BMP4 conditions, where numerous cell states were identified, many populated only within a small region of experimental space. Cells from conditions with ≥ 0.8 ng/mL EGF/bFGF and BMP4 ≥ 4 ng/mL, 36 samples in total, mapped to just three clusters (0, 10, and 14) which were further subdivided by orthogonal experimental factors (Figure 2.7d). The cell state defined by cluster 14 was not observed in samples with high EGF/bFGF, high BMP4, or high Scriptaid/Decitabine or retinoic acid concentrations. Instead, cells from those conditions were found in clusters 0 and 10, with cells treated with Scriptaid/Decitabine appearing almost exclusively in cluster 0, while those treated with retinoic acid mapped strongly to cluster 10 with secondary populations mapping to cluster 0. Such a dissection of cellular response to perturbations has been a long-standing goal in cell biology (Janes 2005, Nelander et al. 2008, Sims et al. 2011, Lamb et al. 2018, Datlinger et al. 2017). It has been hypothesized that cells occupy a relatively limited number of transcriptional states in response to disease or experimental perturbation, and elucidating the connections between various perturbations will help in understanding cellular behavior. One such endeavor, the Connectivity Map (CMap) project (Lamb et al. 2018), is a large-scale effort to measure gene expression response to molecular perturbations. While impressive in scope CMap has been used to profile more than a million perturbation experiments major challenges have included batch effects, averaging across cell populations, and difficulty in examining conditions that yield very few cells. The multiplexing method presented here overcomes these obstacles and provides single-cell whole-transciptome resolution at very low cost. To further validate sample multiplexing and explore its limits, we performed a multiplexing experiment in which four samples of live mouse neural stem cells (NSCs) and four samples of methanol-fixed NSCs were each labeled with unique sets of two methyltetrazine-modified sample tags. The samples were then quenched, pooled, and processed with the 10x Genomics Single-Cell Gene Expression Kit. Analysis of sample tag profiles from methanol-fixed cells recapitulated matched pairs of sample tags, indicating efficient single-cell labeling, and permitting facile sample demultiplexing (Figure 2.2)). Cell doublet events were unambiguously detected as collisions of four pairs of tags corresponding to two separate samples. In methanol-labeled samples, we noted a strong correlation between UMI counts for pairs of tags applied to the same samples (Figure 2.2c), suggesting that the extent of chemical tagging may be correlated with cell size. To test this hypothesis, we devised a species-mixing experiment in which large, human HEK293T cells and small, mouse NSCs were reacted individually and in combination with a series of non-overlapping sample tag pools of increasing size (Figure 2.3). We found that up to five cell tags could be deposited on a single cell without loss of tag recovery, implying that 15,504 experiments could be multiplexed with a panel of just 20 tags. In addition, a strong correlation was observed between species of origin and sample tag counts, indicating our chemical tagging method is indeed sensitive to cell size, a relatively unexplored biological phenotype with intriguing implications for future work. Live cell labeling in aqueous solution resulted in diminished signal-to-noise (data not shown), likely a result of the high rate of NHS-ester hydrolysis in aqueous solution, along with the reduced rate of IEDDA reactions in water

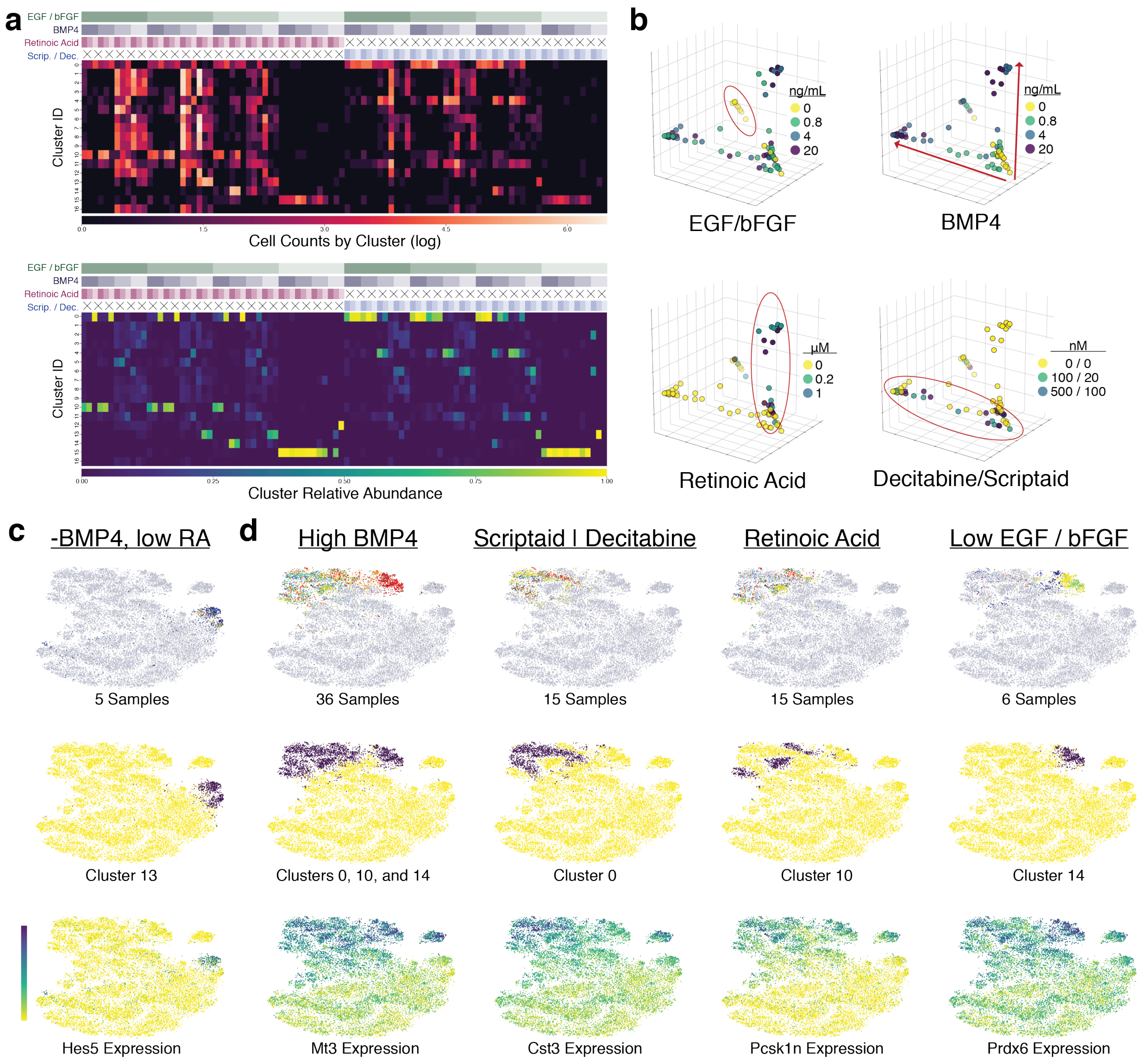


Figure 2.7: Cellular response to perturbation (a) Cluster occupancy versus experimental condition (above), shown as number of cells detected (top) or the relative abundance of cells assigned to each cluster in each sample (bottom). (b) PCA of relative cluster abundance matrix from Fig 2a. Each point represents a cell population from one of 96 experimental conditions, revealing patterns of influence for each experimental factor (highlighted). (c) Five conditions map specifically to cluster 15, characterized by Hes5 expression. (d) 36 samples from high BMP4 conditions map to clusters 0, 10, and 14, each specific to localized regions of the perturbation space. Expression of top differentially expressed genes for each selected group is shown.

compared to methanol. Under methanol fixation conditions, cell tagging is a robust and flexible method for multiplex scRNA-seq with high capacity for tag multiplexing on individual cells. Compared to labeling strategies based on antibody-oligo conjugates (Stoeckius, S. Zheng, et al. 2017, Stoeckius, Hafemeister, et al. 2017, Peterson et al. 2017), our chemical tagging procedures are cheaper, not reliant on epitope markers, compatible with fixed cells, and, most notably, subject to chemical quenching, permitting high-throughput scRNA-seq analysis of low-input samples by pooling many cell populations before washing. While we have demonstrated multiplexing on the 10x Chromium system, our method is compatible with other similar platforms (e.g. Drop-Seq (Macosko et al. 2015), inDrops (Klein et al. 2015), sci-RNA-seq (Cao et al. 2017), Bio-Rads ddSEQ), and should be readily extendible to full-length scRNA-seq (Picelli, Faridani, et al. 2014) and other single-cell genomic assays.

We envision our chemical multiplexing strategy playing a central role as sequencing-based single-cell profiling continues its phenomenal increase in scale. As multiplexing of DNA libraries has vastly improved the utility and adoption of high-throughput DNA sequencing, our solution for scRNA-seq will similarly reduce costs, drive increases in cell capacity, and extend the scope of scRNA-seq beyond bulk tissue profiling. Furthermore, the increasing throughput of scRNA-seq will facilitate even higher multiplexing, and our method can be readily applied to thousands of samples. For diagnostic purposes, the cost savings associated with multiplex scRNA-seq also have the potential to accelerate the adoption of single-cell genomics in the clinic.

2.4 Future Directions

SUGAR-seq

We have demonstrated a powerful application for labeling cells with identifying nucleic acids - sample multiplexing - but the concept of cell tagging can be expanded to many other important areas. For example, the collection of sugars that decorate the cell surface, known as the glycome, is dynamically regulated and plays important roles in aging and disease. The glycome, unlike the transcriptome or proteome, is not directly encoded by the genome, and thus impervious to standard genomic analysis. In fact, the challenges associated with study of the glycome are so great that it is has been synthetic chemists, rather than molecular biologists, who have made the greatest strides in understanding the role of glycan-modified proteins in biology.

Glycosylation is a form of post-translational protein modification involving the attachment of one or many sugars to a protein, often resulting in a complex chain consisting of a handful of specific sugar residues (Moremen, Tiemeyer, and Nairn 2012). Glycan chains are synthesized stepwise by glycosyltransferases whose substrates are most often uridine diphosphate (UDP)-activated sugars. A merger of synthetic chemistry with biochemistry has succeeded in chemically labeling specific glycans (Aguilar et al. 2017). First, a small, bioorthoganol functional group is installed on a sugar of interest. Then, the modified sugar

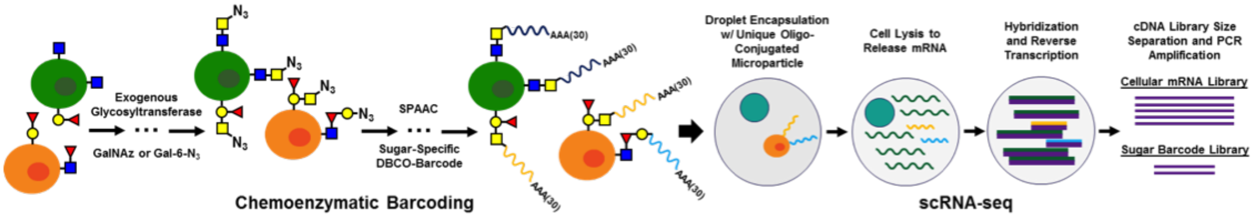

N$_3$
Exogenous
Glycosyltransferase
GalNAz or Gal-6-N$_3$
SPAAC
Sugar-Specific
DBCO-Barcode
AAA(30)
Chemoenzymatic Barcoding
Droplet Encapsulation w/ Unique Oligo-Conjugated Microparticle
Cell Lysis to Release mRNA
Hybridization and Reverse Transcription
cDNA Library Size Separation and PCR Amplification
Cellular mRNA Library
Sugar Barcode Library
scRNA-seq

Figure 2.8: Overview of SUGAR-seq. Cultured cells or tissue samples are dissociated into a single-cell suspension before three sequential rounds of chemoenzymatic labeling to affix a unique, sugar-specific DNA-oligo barcode to each glycan. These labeled cells can then be submitted for scRNA-seq where they are encapsulated in droplets along with oligo-conjugated microparticles and lysed. Because the sugar specific barcodes have a poly(dA) tail, they are captured and reverse transcribed along with the poly-adenylated cellular mRNAs to form cDNA libraries that can be separated by size. These libraries can then be PCR amplified and sequenced using high-throughput sequencing and the transcript and glycan counts for each individual cell determined based a unique identifier derived the droplet-specific oligos on each microparticle.

is ligated to uridine diphosphate, forming the modified UDP-sugar. Finally, a glycosyltransferase is engineered to accept this modified sugar and install it onto protein substrates. So far, a handful of enzymes has been developed that can be used in this way. To selectively label terminal GlcNAc sugars, bovine β-1,4-galactosyltransferase 1 engineered with a Y289L mutation in its active site (Y289L GalT) has been successfully used to append azidoacetylgalactosamine (GalNAz) onto the C-4 hydroxyl group of N-acetylglucosamine (GlcNAc) (Boeggeman et al. 2009). To selectively label fucose-α(1-2)galactose sugars, the bacterial homolog of the blood human blood group A antigen glycosyltransferase (BgtA) has been employed to transfer a GalNAz sugar to the C-3 position of galactose in fucose-α(1-2)galactose (Chaubard et al. 2012). Finally, core fucose residues can be selectively labeled using a -1,4-galactosyltransferase from C. *elegans* (GALT-1), which transfers Gal-6-N3 onto the C-4 position of core fucose (Titz et al. 2009).

In the past, chemoenzymatic labeling has been used for microscopy (attaching clickable fluorophores) or proteomics (attaching mass tags) of specific sugars. While powerful, these approaches fail to capture the transcriptional state of the cell. We will combine the cell tagging methods developed in this work to sequentially label each of these three classes of glycan with a unique barcoded oligo (Figure 2.8). The oligos will be captured, along with the cellular transcriptome, by scRNA-seq, as described. This experiment will reveal for the first time the relationship between the transcriptome and the glycome at single-cell, transcriptome-wide resolution for multiple glycans simultaneously. We have termed this new experiments SUGAR-seq, for Single-cell Unified Glycan And RNA sequencing.

Taking this approach a step further, we will combine the idea of sample multiplexing with SUGAR-seq. Cultured neurons will be exposed to a perturbation screen and their combined glycomes and transcriptomes profiled. Such an experiment will not only describe the contents of each sugar on a cell and associate such a measurement with the transcriptome, but also explore the potential space of transcriptional and glycomic response to the environment, providing an unprecedented connection between glycobiology and genomics. Finally, SUGAR-seq will provide a new window onto Alzheimer's disease as we sample primary tissue samples from both mice disease models and recently deceased humans. We hope to shed new light onto the role of glycome disregulation in Alzheimer's disease and to identify transcriptional states and cell types most highly associated with abnormal glycan profiles.

A Molecular Roadmap for Disease

Biology is unique in the daunting scale of potential experimental space. As stated previously, biological sequence space is undeniably huge, but even the vastness of sequence space pales in comparison to that of functional space. The range of functional states in which cells may exist is almost completely unknown, as evidenced by the fact that we have only just begun to appreciate the functional states in which cells typically occur in healthy individuals, and in those cases only for a select few model organisms.

Cells are thought to populate a high-dimensional functional landscape with each cell's position in such a landscape determined by its genotype, life history, and physical environment. The extent to which the potential landscape is actually explored in vivo is at present entirely unknown, but we have begun to model many diseases as perturbations in functional space away from "healthy" regions. Ailments from cancer to Alzheimer's to pathogenic infections can all be modeled as stable cell states far from normal. From this viewpoint, our understanding of disease can only be as good as our understanding of cell state space.

The method for high-throughput scRNA-seq presented here represents a breakthrough in defining the landscapes of cell population and transcriptional states. The ability to observe many (hundreds to thousands) of samples each with many (hundreds to thousands) of cells in a single experiment will enable researchers to explore functional space through experimentation. By pushing cells out of frequently populated "healthy" regions of cell space, we can begin to build an understanding of disease at the fundamental level of biology - that of the individual cell.

Such an endeavour will first necessitate a broad survey of disease states across individuals, immediately highlighting the need for sample multiplexing. First, samples from many individuals across many disease states will require hundreds of samples. From there, perturbations could be used to drive healthy cell populations into disease states or to drive diseased cell populations into healthy states. Finally, time-course experiments can be utilized to build a model of disease progression over time both in model systems and clinical patients.

Chapter 3

Antibody-Oligo Conjugation for Simultaneous Single-Cell Protein and RNA Quantification

3.1 Chapter Summary

Cell populations have long been described by the presence of certain marker genes expressed on the cell surface. Antibodies raised against such markers can be used to isolate cells of interest from large populations of undesired cells. For this reason flow cytometry has become the backbone of immunology, providing researchers with tools to characterize and obtain target cell populations. Single-cell RNA-seq, essentially a transcriptome-wide RNA cytometry assay, follows many of the same principles. Indeed, the primary results from scRNA-seq are the transcriptomes of various cell populations defined by clustering algorithms based on pairwise similarities. In practice, biologists interpret these clusters by the presence of marker genes first identified by flow cytometry of immunohistochemistry. However challenges can arise when the genes of interest do not represent a significant fraction of the mRNA pool or when splice variants of biological importance cannot be readily distinguished by sequencing. In these cases it is desirable to specifically quantify proteins of interest in the context of a scRNA-seq dataset, generating a ”multi-modal” dataset that incorporates a transcriptome profile as well as supplementary information mapping to the same cells.

We envisioned an approach to convert epitope expression into sequencing data using antibody-oligo conjugates (Figure 3.1). In such an experiment, synthetic oligonucleotides are designed with the following components: 1) Poly(dA) tail for capture by barcoded poly(dT) primers in scRNA-seq 2) unique barcode for identification during sequencing 3) PCR primer binding site 4) chemical modification for antibody conjugation. Antibody-oligo conjugates are prepared by covalently attaching the modified DNA oligos to antibodies using NHS-ester chemistry, which modifies primary amines, and the tetrazine ligation to link specific antibodies with unique barcodes. Large panels of such antibody-oligo conjugates can be used in

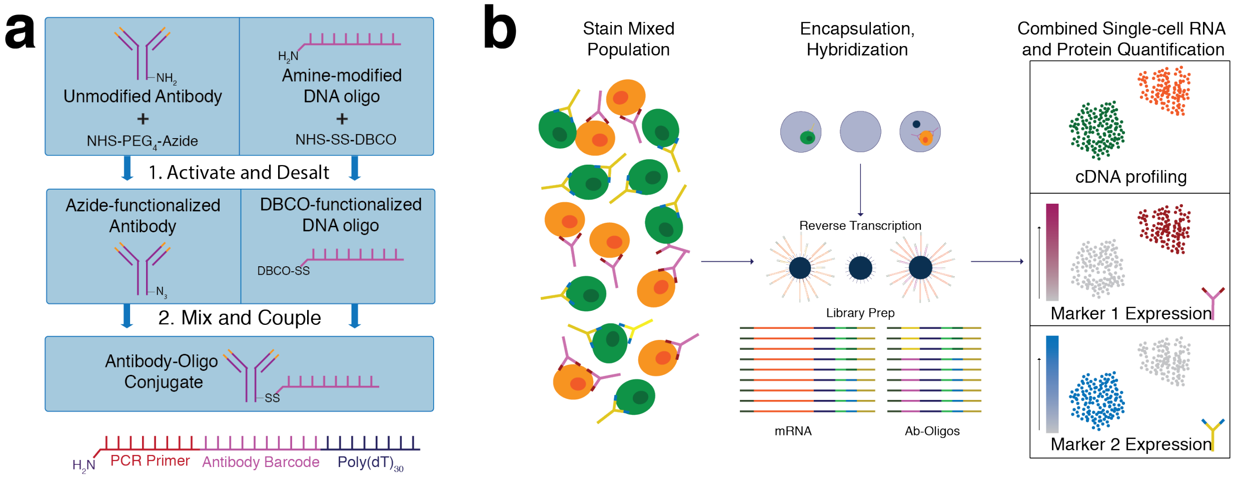


Figure 3.1: Antibody-Oligo Sequencing Overview (a) Antibody-Oligo conjugates are formed using either strain-promoted azide-alkyne cycloaddition (SPAAC, shown) or the tetrazine ligation. A cleavage linker, such as a disulfide can be included. The oligo has components required for capture during scRNA-seq (b) Cell populations are stained with antibodies against multiple markers, each with unique oligo tags. After sequencing cDNA and antibody-oligo libraries, epitope expression can be associated with transcriptional states.

parallel to generate high-dimensional epitope quantification alongside transcriptome profiles from the same single cells (Peterson et al. 2017). Immediately prior to scRNA-seq, the cell population is stained with a panel of antibody-oligo conjugates and thoroughly washed. Once the cells are encapsulated with barcoded microparticles, the synthetic, antibody-conjugated oligos serve as a template for DNA-dependent DNA polymerization by reverse transcriptase using the poly(dT) cell barcode primers. In this way, the epitope profile of the cell is encoded in the same cDNA library as the mRNA from the same cell. After amplification of the pooled cDNA and antibody-tag library, the antibody-tag library can be separated from the cDNA library by DNA size selection using solid-phase reversible immobilization (SPRI) beads. From here, the antibody-tag library is amplified with primers containing sample indexes and Illumina sequencing adapters, and the cDNA library is processed as normal, typically involving fragmentation, adapter ligation, and final amplification. This strategy is applicable to practically any epitope for which a corresponding antibody is available. Because the antibody-tags are captured via a poly(dA) tail, no additional modifications to the library prep need to be made, and the approach is compatible with any scRNA-seq platform that relies on poly(A) capture and a reverse transcriptase possessing DNA-dependent DNA polymerase activity.

3.2 Antibody-Oligo Conjugation Procedures

During development of this procedure, two other groups reported success in applying antibody-oligo conjugates to scRNA-seq (Peterson et al. 2017, Stoeckius, Hafemeister, et al. 2017). The Satija group performed a proof-of-concept experiment (dubbed CITE-Seq) in which cord blood mononuclear cells (CBMCs) were stained simultaneously with 13 antibody-oligo conjugates, 10 of which (77%) were found to give sufficient signal:noise to determine positive and negative cell populations. This work was greatly expanded in the form of REAP-Seq, an antibody-oligo conjugate approach based on the same principles. The developers of REAP-Seq succeeded in staining peripheral blood mononuclear cells PBMCs with 82 unique antibody-oligo conjugates, and impressive improvement in scale. REAP-Seq surpasses even commercial mass-spectrometry-based approaches (CyTOF, Fluidigm) for high-dimensional single-cell epitope profiling. Such high-dimensional analysis alone would be a significant technological advance, but the additional same-cell transcriptome profiles make REAP-Seq a singularly powerful approach to single-cell analysis.

Compared to CITE-seq and REAP-seq, our antibody-tagging approach is significantly cheaper in large part due to the development of custom labeling procedures. While significant time was invested in developing strategies for antibody-oligo conjugation, our optimized procedures are simple, rapid, and far less expensive than commercial alternatives. We were able to prepare six unique antibody-oligo conjugates in a single day using the PK136 universal anti-NK cell antibody, saving around $3,000 by avoiding use of proprietary reagents. Our initial antibody-oligo conjugation strategy involved a RedOx-cleavable crosslinker DBCO-SS-NHS ester which contains a disulfide bond for controllable cleavage, an NHS-ester moiety which is susceptible to nucleophilic attack by primary amines, and a dibenzocyclooctyne (DBCO) functional group, which reacts specifically with azides in one of the most popular click chemistry reactions known as strain-promoted azide-alkyne cycloaddition, or SPAAC. Over the course of this work, it was reported that non-cleavable linkers suffice for scRNA-seq detection, and we turned once again to the incredibly fast tetrazine ligation. Antibody-oligo conjugation was achieved according to the following procedures:

First, antibodies (500 µg) are buffer-changed into borate-buffered saline buffer, pH 8.5, using a 0.5 mL Zeba spin desalting column with 40 kDa molecular weight cutoff according to the manufacturer's incstructions. To the resulting antibody solution, NHS-TCO is added to 250 µM concentration and the reaction is allowed to proceed for 30 minutes at room temperature protected from light. The reaction is then diluted with PBS pH 7.4 to a volume of 1 mL and desalted using a 2 mL Zeba spin desalting column with 40 kDa molecular weight cutoff equilibrated with the same PBS solution. Antibody concentration is determined by NanoDrop spectrophotometer using a molar extinction coefficient of 210,000 M^-1cm^{-1}. At this point, methyltetrazine activated DNA oligos (prepared as in chapter 2), are added at 1.5x molar excess over the antibody concentration. The reactions are allowed to proceed for 30 minutes, and the products can be analyzed by polyacrylamide gel electrophoresis (PAGE) using a 12% denaturing PAGE gel (Figure 3.2).

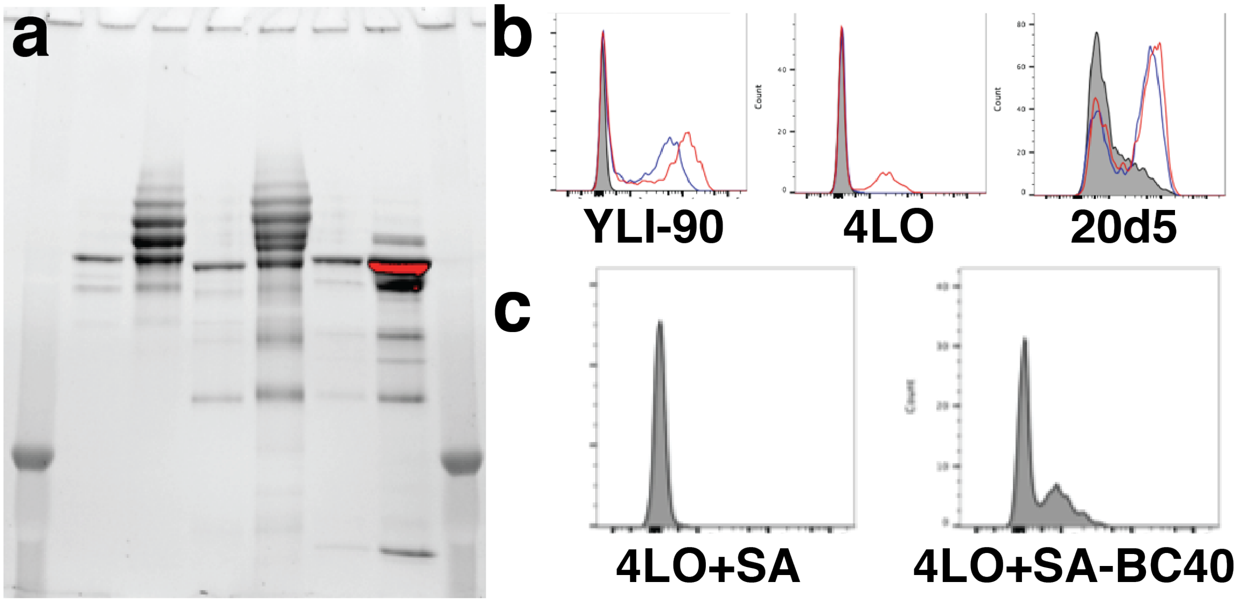


Figure 3.2: Antibody-Oligo Conjugation and Activity Assay (a) SDS-PAGE depicting antibody oligo conjugation. Lanes: 1) ladder 2) 4LO unmodified 3) 4LO + BC40 4) 20d5 unmodified 5) 20d5 + BC38 6) YLI-90 unmodified 7) YLI-90 + BC39 8) ladder. Antibody-oligo conjugation is seen as laddering child bands above the unmodified parent band. Each rung of the ladder corresponds to an additional oligo. 4LO and 20d5 have an average of 2-3 oligos per antibody, while YLI-90, which was conjugated at lower pH, is poorly modified and was conjugated again in a separate reaction under optimal pH conditions. (b) Flow cytometry demonstrating activity of antibody oligo conjugates. Antibody-oligo conjugates and their unmodified versions were stained with appropriate secondary antibodies. YLI-90 and 20d5 retained activity, while 4LO displayed no activity after oligo modification. (c) To recover 4LO activity, a different strategy was employed in which streptavidin (SA) was oligo-conjugated and targeted against biotin-4LO. This strategy proved effective, recovering cell labeling by an anti-DNA fluorescent secondary antibody only when oligo-conjugated, and not unmodified, streptavidin was used.

3.3 Application to Inhibitory Receptor Variegation in Natural Killer Cells

With antibody-oligo conjugation protocols in hand, we looked for intriguing biological processes that would be uniquely suited to simultaneous protein and RNA quantification. We have noticed a strong tendency for proof-of-principle studies to dominate the scRNA-seq literature. In the cases of REAP-Seq and CITE-Seq, few experiments were performed in the initial publications. Known antibodies were used to stain cell populations that had long been studied by flow cytometry and previously analyzed by scRNA-seq. Relatively little computational analysis was performed to leverage the unprecedented datasets being produced. The resulting publications, although technically impressive, offered relatively few biological insights. When developing new technologies, we believe that the applications for a new technology are as consequential as the technology itself. In the case of antibody-oligo conjugates for scRNA-seq, we sought a biological system in which complex cell types traditionally characterized by flow cytometry could be understood with previously unattainable resolution by scRNA-seq. More specifically we wanted a system in which experimental perturbations could be used to tease apart subtle differences in the cell populations.

We turned to the phenomenon of receptor variegation in natural killer cells. Natural killer (NK) cells surveil host cells in the body for signs of damage or infection. NK activity is a balance between inhibitory and activating signals transduced by a series of receptors on the cell surface (Joncker et al. 2009). Inhibitory receptors recognize MHC Class I molecules on the target cell, while stimulatory receptors recognize cellular stress ligands. Mice possess three unique inhibitory receptors that are expressed in an overlapping, variegated pattern in which the expression of a any given receptor is random and independent of expression of either of the other two receptors. Such behavior presents a problem, however, because individual NK cells need to appropriately tune their response to damaged host cells. An NK cell lacking any inhibitory receptors must respond to host cell stress with equivalent potency as an NK cell expressing three inhibitory receptors. This dilemma led our collaborators to propose a rheostat model for NK cell activation in the face of inhibitory receptor variegation. This model suggests that individual NK cells tune their responsiveness based on their inhibitory receptor expression profile.

This system is an excellent test-bed for simultaneous, single-cell protein and RNA detection. Variegated expression of three inhibitory receptors, Ly49C, Ly49I, and NKG2A, gives rise to 8 unique and rare cell populations, each representing less than 1% of the total lymphocyte population, presenting a significant barrier to performing bulk RNA-seq on sorted populations. Conversely, scRNA-seq alone is insufficient to address this biological question because the transcriptome generated for each cell is too sparse to accurately quantify individual genes, in this case the inhibitory receptors. To describe transcriptome profiles corresponding to the eight unique receptor expression profiles in NK cells, we need to capture accurate receptor expression information as well as a full transcriptome for each cell. We created antibody-oligo conjugates specific for each of the three inhibitory receptors expressed by NK cells using corresponding antibodies 4LO (anti-Ly49C), YLI-90 (anti-Ly49I), and (anti-NKG2A). Antibody-oligo conjugates were tested for activity via flow cytometry using either fluorescent secondary antibodies or a fluorescent anti-DNA antibody (AE-2, EMD Millipore). In the case of 4LO, it was found that oligo conjugation ablated avidity, so an alternative approach was devised in which biotinylated 4LO, which retains biological activity, was stained with a streptavidin oligo conjugate. This circumvented the loss of activity seen in 4LO-oligo conjugates and completed the set of antibody-oligo conjugates required for receptor variegation studies in B6 mice. Finally, we prepared 6 unique antibody-oligo conjugates using the PK136 antibody which stains NK1.1. This general marker of NK cells was used to pool cells from multiple animals without losing animal-of-origin information.

As a negative control, the β2m knockout mouse strain was compared against wild-type B6 mice. β2m mice lack expression of MHC Class 1 protein, the main ligand for inhibitory receptors expressed by NK cells. Correspondingly, NK cells from β2m mice display severely depressed NK cell activation phenotypes. NK cells were harvested from 3 female B6 and 3 female β2m mice, and, in a biological replicate, 3 male mice of each genotype. To reduce cost and batch effects, a multiplexing strategy was employed in which white blood cells from each mouse were separately stained with a unique PK136 antibody-oligo conjugate as well as 4LO biotin. The cells were washed in parallel, then pooled and stained with oligo-conjugated antibodies against NKG2A and Ly49I as well as a streptavidin-oligo conjugate as a secondary stain against 4LO biotin. The cells were thoroughly washed then subjected to flow cytometry using a gating scheme designed to isolate all NK cells. The cells, purified and stained with oligo-conjugates for each of the three inhibitory receptors, were processed using the 10x Genomics Single-Cell Gene Expression Kit according to a modified workflow (Figure 3.3). This experiment did not yield the expected results. While library preparation and cDNA sequencing appeared normal (Figure 3.3a and 3.3b), antibody-oligo tags were poorly distributed among the cell population (Figure 3.3c) and failed to generate positive and negative populations (Figure 3.3d, 3.3e, and 3.3f). Dissection of this result and optimization of antibody labeling methods will be the subject of future work.

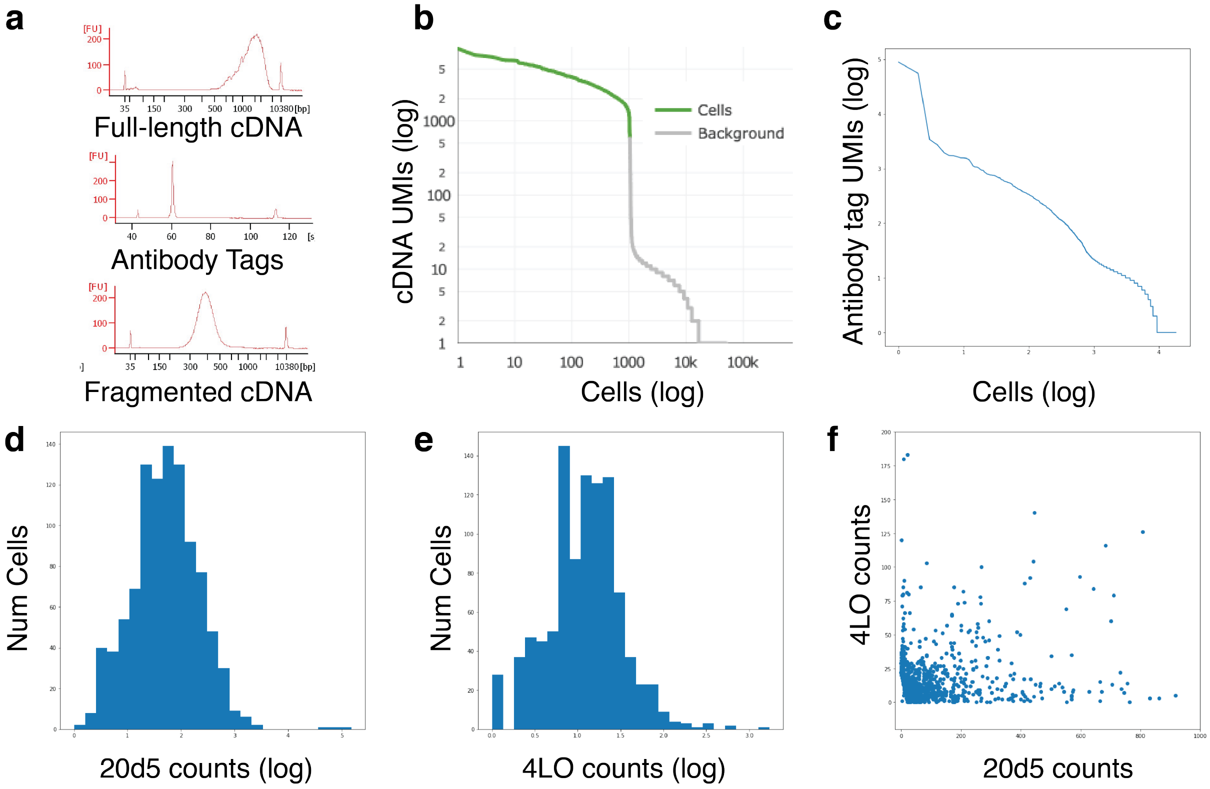


Figure 3.3: Antibody-Oligo Sequencing Results (a) BioAnalyzer traces for a full-length natural killer cell cDNA library, and antibody tag library, and a fragmented cDNA library (b) cDNA sequencing results for 1,039 natural killer cells. An average of 16,558 reads and 1,071 genes were detected across the population (c) Antibody tag reads for the same 1,039 cells. A clear distinction between positive and negative droplets cannot be made (d) Histogram showing antibody tag reads for the 20d5 antibody. A single population is observed, rather than positive and negative populations (e) Histogram showing antibody tag reads for the 4LO antibody. Again, a single population is observed, failing to distinguish positive and negative populations (f) Scatter plot of 20d5 and 4LO antibody tag reads plotted against one another for each cell. No clear populations are observed.

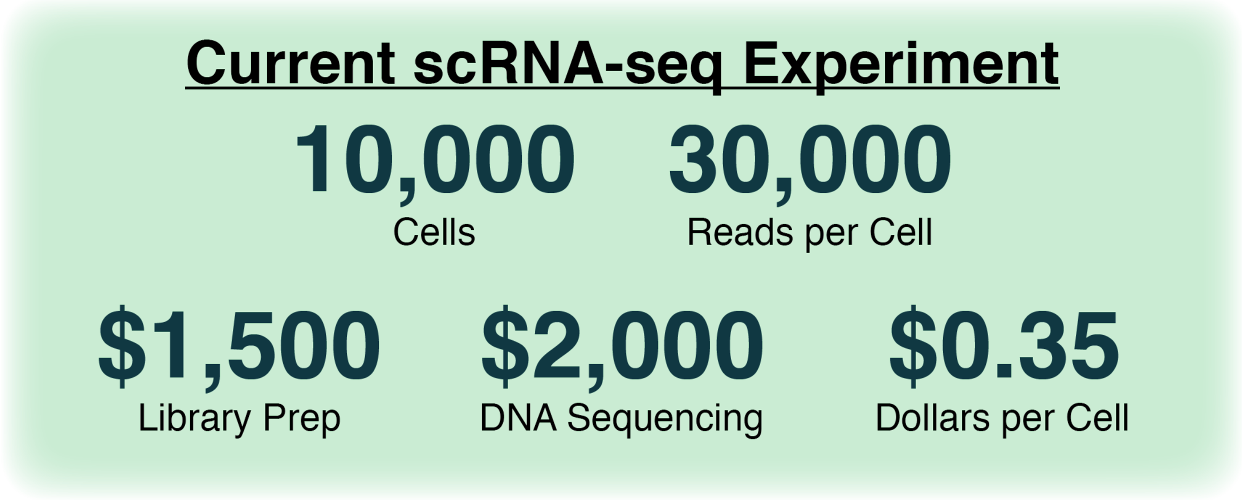
Chapter 4

Split-Pool DNA Barcoding: Targeted Single-Cell RNA-Sequencing

4.1 Chapter Summary

In the previous chapters, I have addressed two limitations in single-cell genomics, sample throughput and multi-modal analysis. In both cases, barcoded oligonucleotides were attached to cells, either chemically or through an antibody mediator, to add identifying information to cells prior to scRNA-seq. These advances represent two ways in which the scRNA-seq library prep can be leveraged for more complex or larger scale analysis. So what, if any, are the limitations of scRNA-seq? The current challenges fall into three general classes: sample preparation, cDNA capture efficiency, and cost. Sample preparation, which involves dissociating cultured cells or tissue samples into single cells, is a major concern because every tissue type is unique, potentially requiring individual optimization, and because it is known that tissue dissociation causes transcriptional changes that can be difficult to control (Wu et al. 2017). For now, orthogonal approaches, such as high-dimensional, quantitative FISH probing, may be required to ensure the validity of biological results based on scRNA-seq of tissues requiring extensive dissociation procedures (Shah et al. 2016, G. Wang, Moffitt, and Zhuang 2018). Capture efficiency, referring to the conversion rate of transcript molecules into an amplified cDNA library, is a more general problem where any improvements will broadly impact single-cell genomics. Low capture efficiency increases false negative rates for transcript detection, especially for low-abundance transcripts typically associated with cell state, such as transcription factors.

Capture efficiency is also directly related to cost, perhaps the biggest barrier in scRNA-seq. The economics of current scRNA-seq workflows break down as follows: 10,000 cells captured with 30,000 reads each, with library prep costs at $1,000-2,000 and sequencing costs at $2,000 (Figure 4.1). Improvements in microfluidic platforms and sample multiplexing are rapidly driving down the cost of library preparation, meaning sequencing cost is the only barrier to performing scRNA-seq at scales orders of magnitudes larger than today's



Current sccRNA-seq Experiment

10,000 Cells

30,000 Reads per Cell

$1,500 Library Prep

$2,000 DNA Sequencing

$0.35 Dollars per Cell

Figure 4.1: A typical scRNA-seq experiment profiles 10,000 cells at 30,000 reads each. Costs for library preparation and sequencing are roughly equivalent, but library preparation costs are plummeting due to sample multiplexing and improvements in microfluidic technologies. Sequencing costs are stagnant in comparison, prompting an urgent need to decrease the number of reads per cell without sacrificing biological information.

experiments. Single-cell genomics is putting more pressure on the high-throughput sequencing industry than at any time since the human genome project. For the first time in many years, DNA sequencing is preventing, rather than enabling, massive increases in experimental scope. The price of sequencing is sure to continue to drop, although one cannot reasonably expect capacity per dollar to increase more than around two-fold year over year, at best. The only conclusion that can be drawn is that sequencing cost per cell must be dropped, and that cost reductions will come from molecular biologists, not Illumina, Inc. To this end, we are developing a platform for targeted scRNA-seq that will improve capture efficiency, reduce the cost of library preparation, and, most importantly, directly reduce the requisite sequencing depth for scRNA-seq.

4.2 Targeted scRNA-seq

We reasoned that targeted RNA capture could be transformative for scRNA-seq by enriching for biologically relevant genes while reducing the number of reads required for each cell. Advances in high-throughput oligo array synthesis could be leveraged to generate panels of capture probes suitable for a given biological system (i.e. the immune system, cancer, neuroscience). Hundreds to thousands of individual targets could be selected, including non-coding RNAs that are currently undetectable with poly(A) capture protocols. Single-cell sequencing information could be used to identify the most informative genes for a given tissue, for example by selecting the top 1,000 genes contributing to the overall variance in a standard scRNA-seq experiment. The cost reduction provided by targeted scRNA-seq is potentially enormous. Capture strategies based on poly-adenylated RNAs successfully avoid abundant, uninformative ribosomal RNAs, but targeted capture takes this concept a step further and avoids abundant, uninformative mRNAs as well. Most cDNAs sequenced in a given scRNA-seq experiment map to a small subset of genes. A preliminary analysis found that over 50% of the reads captured in scRNA-seq map to just 20 genes. Our targeted approach will turn this dynamic on its head. We will design capture panels that correspond to the most variant genes in a population, and using array-based oligo synthesis we can control the relative ratios of the capture sequences, effectively normalizing for average gene expression and distributing sequencing depth more evenly across the most informative subset of the transcriptome.

4.3 Split-Pool Synthesis via the Primer Exchange Reaction

The key advance responsible for the recent dramatic increases in scRNA-seq capacity is the DNA-barcoded microparticle. The beads used in Drop-Seq are polystyrene microspheres coated in DNA oligos produced by split-pool phospharmidite synthesis, while those used by the competing inDrops protocol are polyacrylamide hydrogel beads synthesized enzymatically by primer extension. Both of these methods for barcoded bead synthesis are extremely labor intensive and relatively inefficient. Drop-Seq beads are infamous for their high error rates and low number of oligos per bead, while adoption of the inDrops protocol has been slow due to high up-front costs and labor demands. We sought a flexible split-pool barcoding procedure that could reduce the high costs, labor, and inefficiencies associated with barcoded bead synthesis while providing an avenue to generate beads with custom capture panels for targeted scRNA-seq. We turned to a recently demonstrated method for single-stranded nucleic acid synthesis dubbed the primer exchange reaction (PER) (Kishi et al. 2018).

PER is a primer extension reaction that relies on the phenomenon of branch migration to recycle a catalytic template molecule. The PER catalytic cycle begins with a toe-hold primer extension reaction in which a catalytic template hairpin anneals to a short (9bp) complementary region at the 3' end of a DNA primer (to be extended). A DNA polymerase with strand-displacing activity (such as Bst DNA polymerase), extends the primer while displacing the upstream hairpin sequence (Figure 4.2). Primer extension is stopped when the polymerase reaches a blocking sequence, such as a modified template base. Having stalled, the DNA polymerase will dissociate, leaving a branched structure that can explore base-pairing combinations in a process known as branch migration. When the branch migrates sufficiently in favor of hairpin formation, the melting temperature of the partially displaced primer will fall below the reaction temperature, and the primer dissociates. The product

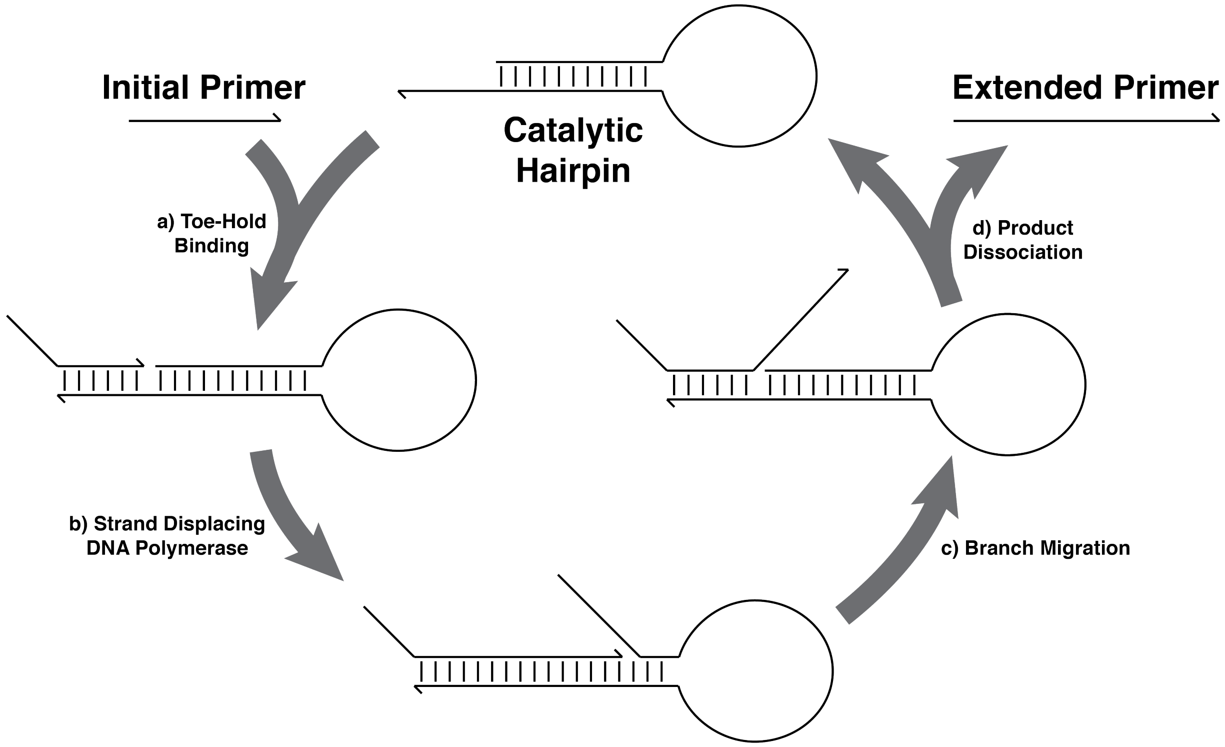


Figure 4.2: The primer exchange reaction (PER) consists of a four-step catalytic cycle. (a) First, a primer to be extended binds to a complementary toe-hold region on a catalytic hairpin. (b) A DNA polymerase with strand displacement activity extends the primer and displaces the hairpin top strand until reaching a lesion through which it cannot continue polymerization. (c) The polymerase dissociates, and branch migration proceeds as the primer and hairpin top strand compete for binding to the hairpin bottom strand. (d) With the branch migrated sufficiently in favor of hairpin formation, the primer-template interaction is sufficiently destabilized, the primer dissociates, producing an extended primer and recycling the catalytic hairpin for another round of primer extension.

of the reaction is a primer extended with a specific sequence, dependent on its original 3'-terminal sequence, and a recycled PER hairpin that can go on to catalyze primer extension on a new primer. Repurposing the template as a catalyst, rather than a substrate to be consumed in the reaction, dramatically reduces the concentration of template oligo required while making PER highly efficient at moderate temperatures (typically 37 °C) and capable of high-fidelity, one-pot multiple extension reactions. These advantages make PER extremely attractive as the basis for split-pool DNA barcoding on primer-coated microspheres.

4.4 Results

We began by synthesizing hydrogel beads according to the inDrops protocol (Klein et al. 2015). Briefly, oligos are covalently incorporated into a polyacrylamide gel matrix via a 5'-acrydite modification. Beads are formed using a microfluidic emulsion generator operated by flowing a fluorous oil continuous phase and an aqueous discontinuous phase through a T-junction. The radical catalyst TEMED is dissolved in the continuous phase. As a monodisperse emulsion is formed at the T-junction, TEMED diffuses into the aqueous droplets and initiates polyacrylamide polymerization. Each droplet produces a polyacrylamide gel bead decorated with up to 10^9 oligos per bead. A suspension of clear gel beads is depicted in Figure 4.3a. We confirmed the presence of covalently incorporated DNA oligos using a fluorescence in situ hybridization (FISH) experiment in which an fluorescently labeled complementary oligo was annealed to gel beads synthesized in the presence and absence of an acrydite-modified DNA oligo (Figure 4.3b). Strong fluorescence signal was observed only for beads synthesized with acrydite-modified DNA oligos, indicating efficient incorporation of DNA during polymerization as well as accessibility to non-immobilized macromolecules.

Bead barcodes are generated by split-pool barcoding using PER. For scRNA-seq, it is advantageous for capture oligos to be released upon droplet encapsulation with target cells. We used a modified polyacrylamide cross-linker, *N,N'*-Bis(acryloyl)cystamine (BAC), to serve as a reversible cross-linker. The disulfide bond in BAC allows cleavage via a reducing agent such as DTT, resulting in the dissolution of the gel matrix (Figure 4.3a). After BAC cleavage, bead oligos are still covalently attached to a linear polyacrylamide tail. We discovered that the presence of this tail led to decreased binding to silica matrix, preventing standard nucleic acid purification. To circumvent this problem, we installed a deoxyuridine residue on the bead oligo, permitting facile cleavage from the polyacrylamide support by USER enzyme, a cocktail of uracil-DNA glycosylase, which leaves abasic sites at deoxyuracil residues, and Endonuclease VIII, which cleaves single- or double-stranded DNA at abasic sites. Both DTT and USER enzyme are compatible with reverse transcription, meaning we can dissolve the polyacrylamide gel matrix, cleave bead oligos from the support, and perform reverse transcription all within a microfluidic droplet upon cell capture. In sum, this protocol serves as a drop-in, non-proprietary replacement for commercial scRNA-seq procedures. Coupled with an improved split-pool barcoding strategy, these tools will open up new avenues of research while applying additional pressure to reduce costs and add

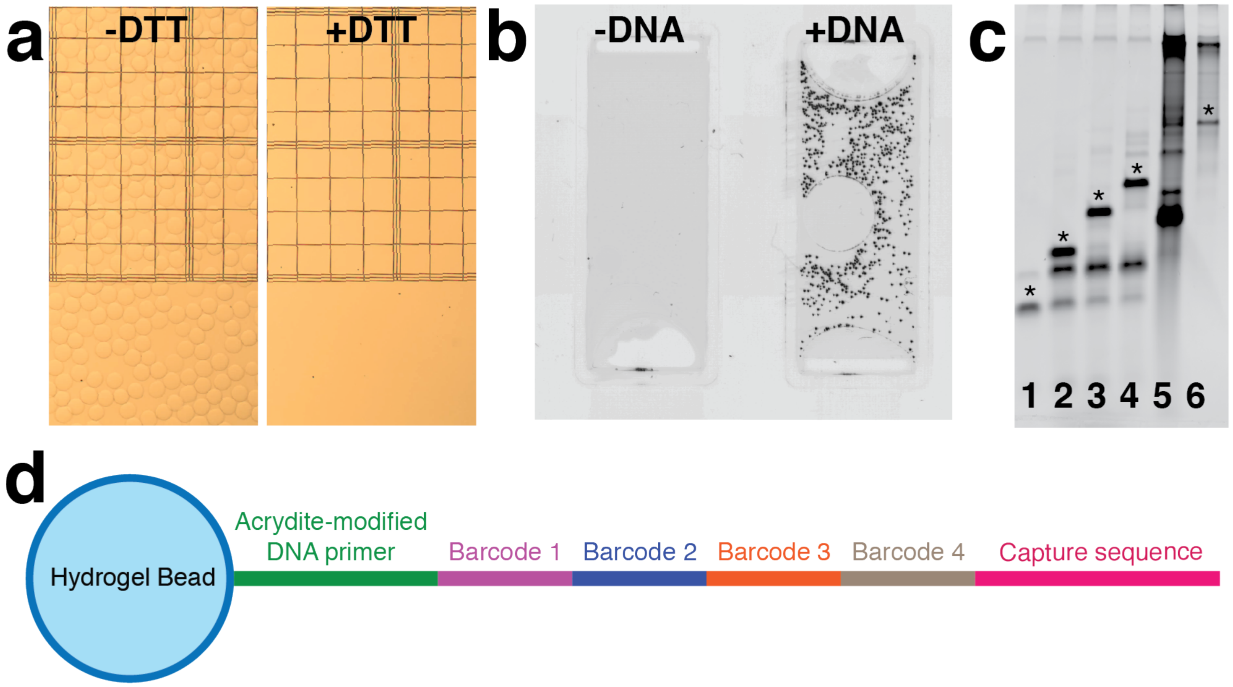


Figure 4.3: (a) Clear hydrogel beads can be visualized by light microscopy. Beads (60 µm)are synthesized in a microfluidic emulsion. DNA oligos are covalently incorporated via a 5'-acrydite modification. Treatment of BAC-cross-linked beads with reducing agent (DTT) dissolves the gel matrix in minutes. (b) FISH probe for bead-bound DNAs. A fluorescein-labeled DNA oligo complementary to the bead oligo was hybridized to either control beads synthesized without DNA (left) or positive beads synthesized with an acrydite-modified DNA oligo (right). Beads were imaged on a Typhoon laser scanning imager, and fluorescent microparticles are cleary seen when beads are synthesized in the presence of acrydite-modified DNA. (c) Synthesis of mini-split-pool barcoded bead library. Bead-bound oligos are sequentially extended by PER. Shown are one (1), two (2), three (3), and four (4) rounds of primer extension. Lane 5 is the product of a primer extension reaction that caps the bead oligos with a poly(dT) capture sequence, and lane 6 shows this product after exonuclease treatment and denaturing wash. The final product is a single species of correct molecular weight. (d) Schematic of the hydrogel bead barcode. Beads oligos are covalently attached to a hydrogel support and contain a series of four split-pool barcodes appended with a variable capture sequence.

features in commercial scRNA-seq kits.

With a viable chemistry strategy to prepare oligo-functionalized beads compatible with scRNA-seq, we began serial primer extension reactions to prepare barcoded bead libraries. We designed a mini-split-pool library consisting of four rounds of split-pool synthesis with 10 unique barcodes at each round, generating a final library of 10^4 possible bead barcodes. PER was used to perform serial bead oligo extension reactions. A crucial, unique aspect of PER is that the product of the reaction is a single-stranded DNA oligo. This feature, combined with the sequence specificity of primer extension, makes PER an ideal reaction for split-pool synthesis of barcoded oligos. The reactions, which require very little catalytic hairpin, can be performed sequentially without denaturation or washing steps, greatly reducing the labor, costs, and sample loss typically associated with washing-intensive split-pooling approaches. Figure 4.3 shows the results of a mini-split pool barcoding experiment. Each sequential extension reaction converts almost all of the substrate into a singly-extended product. After four rounds of split-pooling, the oligos are capped with a poly(dT) capture sequence. A single denaturing wash and exonuclease treatment yields a single species (lane 6) with the structure shown in Figure 4.3d. In all, these improvements represent a major advance in synthesis of barcoded beads for scRNA-seq, a process that until now has proven so prohibitively challenging and expensive that only a handful of laboratories in the world have successfully produced barcoded beads in-house. In demonstrating a user-friendly, cost-effective synthesis, we hope to democratize scRNA-seq technology, allowing individual laboratories to prepare custom barcoded beads for applications beyond our current imagination.

4.5 Future Directions

In parallel with our goal of demonstrating user-friendly bead synthesis, we plan to use the first PER-barcoded hydrogel beads for previously intractable experiments. First, we need to perform a quality control experiment in which oligo populations from single beads are amplified and appended with bead-specific Illumina indexes. By sequencing ten to twenty beads, we will determine the barcode error rate. For a full-scale library, four rounds of split-pool with 48 barcodes at each round will yield a total library size of over five million unique bead barcodes, sufficient for scRNA-seq of more than 10,000 cells. We will first perform a species mixing experiment using standard mouse and human cell lines (3T3 and HEK293T, respectively). Library preparation will combine elements from the 10x Single Cell 3' Gene Expression Kit and the inDrops protocol. Once we have demonstrated scRNA-seq using our homemade beads, we will focus on untouched experimental territory. A clear implication of making is our beads is that we can design any sequences we want. We will start by creating a targeted scRNA-seq panel for the mouse brain by identifying the most variable and informative genes capture by previous scRNA-seq of brain tissue. We will cap our bead oligos with a targeted panel designed against 100-1,000 genes of interest. We hypothesize that targeted capture sequences will enrich for informative genes, thus greatly reducing the read depth required for effective cell clustering and analysis. Such an experiment is the first step toward a leap in the capacity of scRNA-seq. By attacking the problem of sequencing cost, we plan to perform scRNA-seq experiments on significantly larger numbers of cells with no increase in cost per cell.

Bibliography

Aguilar, Aime Lopez et al. (2017). "Tools for Studying Glycans: Recent Advances in Chemoenzymatic Glycan Labeling". In: *ACS chemical biology* 12.3, pp. 611–621. ISSN: 1554-8929. DOI: 10.1021/acschembio.6b01089. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5469623/ (visited on 05/10/2018).

Anderson, S (1981). "Shotgun DNA sequencing using cloned DNase I-generated fragments." In: *Nucleic Acids Research* 9.13, pp. 3015–3027. ISSN: 0305-1048. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC327328/ (visited on 05/10/2018).

Boeggeman, Elizabeth et al. (2009). "Site specific conjugation of fluoroprobes to the remodeled Fc N-glycans of monoclonal antibodies using mutant glycosyltransferases: application for cell surface antigen detection". eng. In: *Bioconjugate Chemistry* 20.6, pp. 1228–1236. ISSN: 1520-4812. DOI: 10.1021/bc900103p.

Bond, Allison M., Guo-li Ming, and Hongjun Song (2015). "Adult Mammalian Neural Stem Cells and Neurogenesis: Five Decades Later". In: *Cell Stem Cell* 17.4, pp. 385–395. ISSN: 1934-5909. DOI: 10.1016/j.stem.2015.09.003. URL: http://www.sciencedirect.com/science/article/pii/S1934590915004105 (visited on 04/25/2018).

Buenrostro, Jason D. et al. (2015). "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561, pp. 486–490. ISSN: 1476-4687. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590 (visited on 05/10/2018).

Cadwell, Cathryn R. et al. (2016). "Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq". en. In: *Nature Biotechnology* 34.2, pp. 199–203. ISSN: 1546-1696. DOI: 10.1038/nbt.3445. URL: https://www.nature.com/articles/nbt.3445 (visited on 05/10/2018).

Cao, Junyue et al. (2017). "Comprehensive single-cell transcriptional profiling of a multicellular organism". en. In: *Science* 357.6352, pp. 661–667. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam8940. URL: http://science.sciencemag.org/content/357/6352/661 (visited on 05/05/2018).

Chaubard, Jean-Luc et al. (2012). "Chemoenzymatic Probes for Detecting and Imaging Fucose-fffdfffd(1-2)-galactose Glycan Biomarkers". In: *Journal of the American Chemical Society* 134.10, pp. 4489–4492. ISSN: 0002-7863. DOI: 10.1021/ja211312u. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3303202/ (visited on 05/10/2018).

Crick, F. H. (1958). "On protein synthesis". eng. In: *Symposia of the Society for Experimental Biology* 12, pp. 138–163. ISSN: 0081-1386.

Datasets - Single Cell Gene Expression - Official 10x Genomics Support (2018). URL: https://support.10xgenomics.com/single-cell-gene-expression/datasets (visited on 04/25/2018).

Datlinger, Paul et al. (2017). "Pooled CRISPR screening with single-cell transcriptome readout". en. In: *Nature Methods* 14.3, pp. 297–301. ISSN: 1548-7105. DOI: 10.1038/nmeth.4177. URL: https://www.nature.com/articles/nmeth.4177 (visited on 04/25/2018).

Gierahn, Todd M. et al. (2017). "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput". en. In: *Nature Methods* 14.4, pp. 395–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.4179. URL: https://www.nature.com/articles/nmeth.4179 (visited on 05/10/2018).

Han, Xiaoping et al. (2018). "Mapping the Mouse Cell Atlas by Microwell-Seq". In: *Cell* 172.5, 1091–1107.e17. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.02.001. URL: http://www.sciencedirect.com/science/article/pii/S0092867418301168 (visited on 04/25/2018).

Hoheisel, Jörg D. (2006). "Microarray technology: beyond transcript profiling and genotype analysis". en. In: *Nature Reviews Genetics* 7.3, pp. 200–210. ISSN: 1471-0064. DOI: 10.1038/nrg1809. URL: https://www.nature.com/articles/nrg1809 (visited on 05/10/2018).

Hsiao, Sonny C. et al. (2009). "Direct Cell Surface Modification with DNA for the Capture of Primary Cells and the Investigation of Myotube Formation on Defined Patterns". In: *Langmuir : the ACS journal of surfaces and colloids* 25.12, pp. 6985–6991. ISSN: 0743-7463. DOI: 10.1021/la900150n. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2812030/ (visited on 04/25/2018).

Imayoshi, I. et al. (2013). "Oscillatory Control of Factors Determining Multipotency and Fate in Mouse Neural Progenitors". en. In: *Science* 342.6163, pp. 1203–1208. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1242366. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1242366 (visited on 05/04/2018).

Janes, K. A. (2005). "A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis". en. In: *Science* 310.5754, pp. 1646–1653. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1116598. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1116598 (visited on 04/25/2018).

Joncker, N. T. et al. (2009). "NK Cell Responsiveness Is Tuned Commensurate with the Number of Inhibitory Receptors for Self-MHC Class I: The Rheostat Model". en. In: *The Journal of Immunology* 182.8, pp. 4572–4580. ISSN: 0022-1767, 1550-6606. DOI: 10.4049/jimmunol.0803900. URL: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.0803900 (visited on 05/09/2018).

Kang, Hyun Min et al. (2018). "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation". en. In: *Nature Biotechnology* 36.1, pp. 89–94. ISSN: 1546-1696. DOI: 10.1038/nbt.4042. URL: https://www.nature.com/articles/nbt.4042 (visited on 04/25/2018).

Karaiskos, Nikos et al. (2017). “The Drosophila embryo at single-cell transcriptome resolution”. en. In: *Science*, eaan3235. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aan3235. URL: http://science.sciencemag.org/content/early/2017/08/30/science.aan3235 (visited on 05/10/2018).

Kishi, Jocelyn Y. et al. (2018). “Programmable autonomous synthesis of single-stranded DNA”. In: *Nature chemistry* 10.2, pp. 155–164. ISSN: 1755-4330. DOI: 10.1038/nchem.2872. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784857/ (visited on 05/10/2018).

Klein, Allon M. et al. (2015). “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells”. English. In: *Cell* 161.5, pp. 1187–1201. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.04.044. URL: https://www.cell.com/cell/abstract/S0092-8674(15)00500-0 (visited on 05/05/2018).

Lamb, Justin et al. (2018). “The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease”. en. In: 313, p. 8.

Lister, Ryan et al. (2008). “Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis”. In: *Cell* 133.3, pp. 523–536. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.03.029. URL: http://www.sciencedirect.com/science/article/pii/S0092867408004480 (visited on 05/10/2018).

Macosko, Evan Z. et al. (2015). “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. English. In: *Cell* 161.5, pp. 1202–1214. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.05.002. URL: https://www.cell.com/cell/abstract/S0092-8674(15)00549-8 (visited on 05/05/2018).

Moremen, Kelley W., Michael Tiemeyer, and Alison V. Nairn (2012). “Vertebrate protein glycosylation: diversity, synthesis and function”. en. In: *Nature Reviews Molecular Cell Biology* 13.7, pp. 448–462. ISSN: 1471-0080. DOI: 10.1038/nrm3383. URL: https://www.nature.com/articles/nrm3383 (visited on 05/10/2018).

Mortazavi, Ali et al. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. en. In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7105. DOI: 10.1038/nmeth.1226. URL: https://www.nature.com/articles/nmeth.1226 (visited on 05/10/2018).

Nagalakshmi, Ugrappa et al. (2008). “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing”. en. In: *Science* 320.5881, pp. 1344–1349. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1158441. URL: http://science.sciencemag.org/content/320/5881/1344 (visited on 05/10/2018).

Nelander, Sven et al. (2008). “Models from experiments: combinatorial drug perturbations of cancer cells”. en. In: *Molecular Systems Biology* 4.1, p. 216. ISSN: 1744-4292, 1744-4292. DOI: 10.1038/msb.2008.53. URL: http://msb.embopress.org/content/4/1/216 (visited on 04/25/2018).

Peterson, Vanessa M. et al. (2017). “Multiplexed quantification of proteins and transcripts in single cells”. en. In: *Nature Biotechnology* 35.10, pp. 936–939. ISSN: 1546-1696. DOI: 10.1038/nbt.3973. URL: https://www.nature.com/articles/nbt.3973 (visited on 04/25/2018).

Picelli, Simone, fffdfffdsa K. Björklund, et al. (2013). “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”. eng. In: *Nature Methods* 10.11, pp. 1096–1098. ISSN: 1548-7105. DOI: 10.1038/nmeth.2639.

Picelli, Simone, Omid R. Faridani, et al. (2014). “Full-length RNA-seq from single cells using Smart-seq2”. en. In: *Nature Protocols* 9.1, pp. 171–181. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.006. URL: https://www.nature.com/articles/nprot.2014.006 (visited on 05/05/2018).

Reuter, Jason A., Damek Spacek, and Michael P. Snyder (2015). “High-Throughput Sequencing Technologies”. In: *Molecular cell* 58.4, pp. 586–597. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.05.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494749/ (visited on 05/10/2018).

Sackmann, Eric K., Anna L. Fulton, and David J. Beebe (2014). “The present and future role of microfluidics in biomedical research”. en. In: *Nature* 507.7491, pp. 181–189. ISSN: 1476-4687. DOI: 10.1038/nature13118. URL: https://www.nature.com/articles/nature13118 (visited on 05/10/2018).

Sanger, F., S. Nicklen, and A. R. Coulson (1977). “DNA sequencing with chain-terminating inhibitors”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12, pp. 5463–5467. ISSN: 0027-8424.

Sanger, F. and H. Tuppy (1951). “The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates”. In: *Biochemical Journal* 49.4, pp. 463–481. ISSN: 0264-6021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1197535/ (visited on 05/10/2018).

Schena, M. et al. (1995). “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”. eng. In: *Science (New York, N.Y.)* 270.5235, pp. 467–470. ISSN: 0036-8075.

Shah, Sheel et al. (2016). “In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus”. English. In: *Neuron* 92.2, pp. 342–357. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2016.10.001. URL: https://www.cell.com/neuron/abstract/S0896-6273(16)30702-4 (visited on 05/10/2018).

Sims, David et al. (2011). “High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing”. In: *Genome Biology* 12, R104. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-10-r104. URL: https://doi.org/10.1186/gb-2011-12-10-r104 (visited on 04/25/2018).

Staden, R (1979). “A strategy of DNA sequencing employing computer programs.” In: *Nucleic Acids Research* 6.7, pp. 2601–2610. ISSN: 0305-1048. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC327874/ (visited on 05/10/2018).

Stoeckius, Marlon, Christoph Hafemeister, et al. (2017). “Simultaneous epitope and transcriptome measurement in single cells”. en. In: *Nature Methods* 14.9, pp. 865–868. ISSN: 1548-7105. DOI: 10.1038/nmeth.4380. URL: https://www.nature.com/articles/nmeth.4380 (visited on 04/25/2018).

Stoeckius, Marlon, Shiwei Zheng, et al. (2017). “Cell ”hashing” with barcoded antibodies enables multiplexing and doublet detection for single cell genomics”. en. In: *bioRxiv*,

p. 237693. DOI: 10.1101/237693. URL: https://www.biorxiv.org/content/early/2017/12/21/237693 (visited on 04/25/2018).

Svensson, Valentine, Roser Vento-Tormo, and Sarah A. Teichmann (2018). “Exponential scaling of single-cell RNA-seq in the past decade”. en. In: *Nature Protocols* 13.4, pp. 599–604. ISSN: 1750-2799. DOI: 10.1038/nprot.2017.149. URL: https://www.nature.com/articles/nprot.2017.149 (visited on 04/25/2018).

Tang, Fuchou et al. (2009). “mRNA-Seq whole-transcriptome analysis of a single cell”. en. In: *Nature Methods* 6.5, pp. 377–382. ISSN: 1548-7105. DOI: 10.1038/nmeth.1315. URL: https://www.nature.com/articles/nmeth.1315 (visited on 05/10/2018).

Titz, Alexander et al. (2009). “Molecular Basis for Galactosylation of Core Fucose Residues in Invertebrates”. In: *The Journal of Biological Chemistry* 284.52, pp. 36223–36233. ISSN: 0021-9258. DOI: 10.1074/jbc.M109.058354. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2794738/ (visited on 05/10/2018).

Torre, Eduardo et al. (2018). “Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH”. en. In: *Cell Systems* 6.2, 171–179.e5. ISSN: 24054712. DOI: 10.1016/j.cels.2018.01.014. URL: http://linkinghub.elsevier.com/retrieve/pii/S2405471218300516 (visited on 05/10/2018).

Wang, Guiping, Jeffrey R. Moffitt, and Xiaowei Zhuang (2018). “Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy”. en. In: *Scientific Reports* 8.1, p. 4847. ISSN: 2045-2322. DOI: 10.1038/s41598-018-22297-7. URL: https://www.nature.com/articles/s41598-018-22297-7 (visited on 05/10/2018).

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). “RNA-Seq: a revolutionary tool for transcriptomics”. en. In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0064. DOI: 10.1038/nrg2484. URL: https://www.nature.com/articles/nrg2484 (visited on 05/10/2018).

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis (2018). “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biology* 19, p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: https://doi.org/10.1186/s13059-017-1382-0 (visited on 05/10/2018).

Wu, Ye Emily et al. (2017). “Detecting Activated Cell Populations Using Single-Cell RNA-Seq”. en. In: *Neuron* 96.2, 313–329.e6. ISSN: 08966273. DOI: 10.1016/j.neuron.2017.09.026. URL: http://linkinghub.elsevier.com/retrieve/pii/S0896627317308681 (visited on 05/10/2018).

Xin, Yurong et al. (2016). “Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells”. en. In: *Proceedings of the National Academy of Sciences* 113.12, pp. 3293–3298. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1602306113. URL: http://www.pnas.org/content/113/12/3293 (visited on 05/10/2018).

Zheng, Grace X. Y. et al. (2017). “Massively parallel digital transcriptional profiling of single cells”. en. In: *Nature Communications* 8, p. 14049. ISSN: 2041-1723. DOI: 10.1038/ncomms14049. URL: https://www.nature.com/articles/ncomms14049 (visited on 04/25/2018).

Appendix A

Name	**Sequence**	**Scale**	**Purification**
BC21	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAAGCAGTTACAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC22	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTACTTGTACCCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC23	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAGAACCCGGCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC24	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTATCGTAGATCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC25	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAACGCGGAACAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC26	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTACGCTATCCCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC27	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAGTTGCATGCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC28	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTATAAATCGTCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC
BC29	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAATCGCCATCAGBAAAAAAAAAA AAAAAAAAAAAAAAA	250nm	HPLC

BC30	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTACATAAAGGCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC31	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTATCACGGTACAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC32	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTACACTCAACCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC33	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAGCTGTGTACAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC34	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTATTGCGTCGCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC35	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAATATGAGACAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC36	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTACACCTCAGCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC37	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAGCTACTTCCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC38	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTATGGGAGCTCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC39	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTAATCCGGCACAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
BC40	/5AmMC6/TCGTCGGCAGCGTCAGATG TGTACCGTTATGCAGBAAAAAAAAAA AAAAAAAAAAAAAAAA	250nm	HPLC
REAP_BC41_v4	TCGTCGGCAGCGTCAGATGTGTAGGT AATGTCAGBAAAAAAAAAAAAAAAAA AAAAAAAA/3AmMO/	100nm	
REAP_BC42_v4	TCGTCGGCAGCGTCAGATGTGTATAA GCCACCAGBAAAAAAAAAAAAAAAAA AAAAAAAA/3AmMO/	100nm	

REAP_BC43_v4	TCGTCGGCAGCGTCAGATGTGTAACCGAACACAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC44_v4	TCGTCGGCAGCGTCAGATGTGTACGACTCTTCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC45_v4	TCGTCGGCAGCGTCAGATGTGTAGTTTGTGGCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC46_v4	TCGTCGGCAGCGTCAGATGTGTATAGACGACCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC47_v4	TCGTCGGCAGCGTCAGATGTGTAACGCTTGGCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC48_v4	TCGTCGGCAGCGTCAGATGTGTACGCTACATCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC49_v4	TCGTCGGCAGCGTCAGATGTGTAGAAAGACACAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC50_v4	TCGTCGGCAGCGTCAGATGTGTATTTGCGTCCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC51_v4	TCGTCGGCAGCGTCAGATGTGTAATGGTCGCCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC52_v4	TCGTCGGCAGCGTCAGATGTGTACGACATAGCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC53_v4	TCGTCGGCAGCGTCAGATGTGTAGATTCGCTCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC54_v4	TCGTCGGCAGCGTCAGATGTGTATCCAGATACAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC55_v4	TCGTCGGCAGCGTCAGATGTGTAACTACTGTCAGBAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	

REAP_BC56_v4	TCGTCGGCAGCGTCAGATGTGTACGGGAACGCAGBAAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC57_v4	TCGTCGGCAGCGTCAGATGTGTAGACCTCTCCAGBAAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC58_v4	TCGTCGGCAGCGTCAGATGTGTATTATGGAACAGBAAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC59_v4	TCGTCGGCAGCGTCAGATGTGTAACAGCAACCAGBAAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
REAP_BC60_v4	TCGTCGGCAGCGTCAGATGTGTACGCAATTTCAGBAAAAAAAAAAAAAAAAAAAAAAAAAA/3AmMO/	100nm	
P7+A1.3_11bpS3v4	CAAGCAGAAGACGGCATACGAGATTCGGCGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGGGATGTCGTCGGCAGC	100nm	
P7+A1.2_11bpS2v4	CAAGCAGAAGACGGCATACGAGATCTAAACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGGGATTCGTCGGCAGC	100nm	
P7+A1.1_11bpS1v4	CAAGCAGAAGACGGCATACGAGATGGTTTACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGGGATCGTCGGCAGC	100nm	
P7+A1.4_11bpS0v4	CAAGCAGAAGACGGCATACGAGATAACCGTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGGGTCGTCGGCAGC	100nm	
R1-P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT		

Table A.1: Primers used in this study

Sample Number	Species	Tag(s)
1	Mouse 1 tag	BC41
2	Human 1 tag	BC42
3	Mouse 2 tags	BC43, BC44
4	Human 2 tags	BC45, BC46
5	Mouse and Human Mix 2 tags	BC47, BC48
6	Mouse and Human Mix 3 tags	BC49, BC50, BC51
7	Mouse and Human Mix 4 tags	BC52, BC53, BC54, BC55
8	Mouse and Human Mix 5 tags	BC56, BC57, BC58, BC59, BC60

Table A.2: Sample Labeling Scheme for Limits of Multiplexing Experiment