UNIVERSITY OF CALIFORNIA SAN DIEGO

Ultra-High Throughput Single Cell Co-Sequencing of DNA Methylation and RNA using 3-Level Combinatorial Indexing

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

in

Bioengineering

by

Huy Lam

Committee in charge:

Professor Kun Zhang, Chair
Professor Xiaohua Huang
Professor Eran Mukamel
Professor Bing Ren
Professor Sheng Zhong

2022

The Dissertation of Huy Lam is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

This work would not be possible without the support and mentorship of Dr. Kun Zhang and the extended Zhang lab members and alumni. In addition, I would like to thank my family, friends, and my partner Tova.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Zhang and members of the lab for their invaluable advice and mentorship. This work would not be possible without the support and mentorship of them and the extended Zhang lab members and alumni. In addition, I would like to thank my family, friends, and my partner Tova.

Portions of Chapters 1, 2, and 3 in part are a reprint of material in submission as it appears in “Ultra-High Throughput Single Cell Co-Sequencing of DNA Methylation and RNA using 3-Level Combinatorial Indexing” The dissertation author was the primary author of this paper along with Andrew Richards and Kun Zhang.

VITA

2016 Bachelor of Science, Washington University in St. Louis

2022 Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Bioengineering

Professor Kun Zhang

ABSTRACT OF THE DISSERTATION

Ultra-High Throughput Single Cell Co-Sequencing of DNA Methylation and RNA using 3-Level Combinatorial Indexing

by

Huy Lam

Doctor of Philosophy in Bioengineering

University of California San Diego, 2022

Professor Kun Zhang, Chair

DNA methylation at cytosines has long been associated with early development, maturation, and aging of human tissues. Traditionally, DNA methylation is associated with gene silencing. However, recent single cell multi-omic DNA methylation and RNA sequencing methods have shown that the role of DNA methylation on the expression of nearby genes could silence or activate them depending on the gene and cell type. The recent developments these

assays have detected cell type specific DNA methylation and RNA coupling in stem cell rich and human brain tissues. This specificity underscores the need for future growth in DNA methylation and RNA co-sequencing technologies and analysis tools. Presently, about 100,000 single cell profiles are required to adequately map tissues. DNA methylation and RNA co-sequencing methods require the physical isolation of single cells in individual wells. There is no method that can assay 100,000 cells without utilizing extensive liquid handling systems.

We address this challenge by developing a novel ultra-high throughput DNA methylation and RNA co-sequencing platform, sci-Gel, that utilizes three levels of combinatorial indexing to increase the throughput of existing technologies to 50,000-100,000 cells per experiment with just three 96 well plates. In this dissertation, we first push the boundaries of present combinatorial indexing techniques where the DNA and RNA of single cells are simultaneously extracted and immobilized within polyacrylamide gel beads that are used for indexing. This resulted in the development of a 2-level combinatorial indexing platform that could be used to co-sequence DNA copy-number variations, relevant in cancers, and RNA at the scale of thousands of cells. We then describe the adaptations made from existing bisulfite conversion chemistries to our gel beads to incorporate the DNA methylation feature. We then describe the development of a 3-level combinatorial indexing platform to increase the cell throughput of our technology to 50,000-100,000 cells per experiment. Finally, we discuss future efforts to utilize sci-Gel to create the first single cell DNA methylation and RNA co-sequencing map of peripheral blood mononuclear cells.

INTRODUCTION

Cytosine-guanine dinucleotide (CpG) and non-CG DNA methylation have been associated with a variety of mammalian processes such as development, aging, and are disrupted in diseases such as cancer. Recent studies have shown that these methylation marks are cell-type specific and positively or negatively affect transcription factor binding affinity at regulatory elements such as enhancers and promoters(Mulqueen et al. 2018; Callaway et al. 2021). Single cell bisulfite sequencing opens the door for cell type specific methylome profiling for human cell atlas initiatives, for identifying cell-specific methylation markers associated with disease states, and can provide additional epigenetic context to single cell RNA sequencing datasets.

Review of Existing Single Cell DNA Methylation Sequencing Techniques

Single cell DNA methylation can be assayed using whole genome-bisulfite sequencing (WGBS) or reduced representation bisulfite sequencing (RRBS). WGBS interrogates the DNA methylation status of the whole genome. Most single cell WGBS studies have focused on mammalian brain or stem cell tissues (Argelaguet et al. 2019; Angermueller et al. 2016; Luo et al. 2018). Compared to other tissues, these tissues exhibit elevated non-CG methylation which greatly assists in the clustering of single cells. In contrast, the low level of non-CG methylation in other tissues generally requires the use of CG methylation to cluster single cells. The number of non-CG cytosine positions vastly outnumbers the number of CG cytosine positions. For example, our single cell WGBS analyses of kidney tissue, showed to us that the number of non-CG cytosine positions can outnumber the number of CG cytosine almost 10-fold as shown in Figure 1. The lesser number of potentially differentially methylated cytosine positions lowers the

ability to cluster single cells in these tissues.

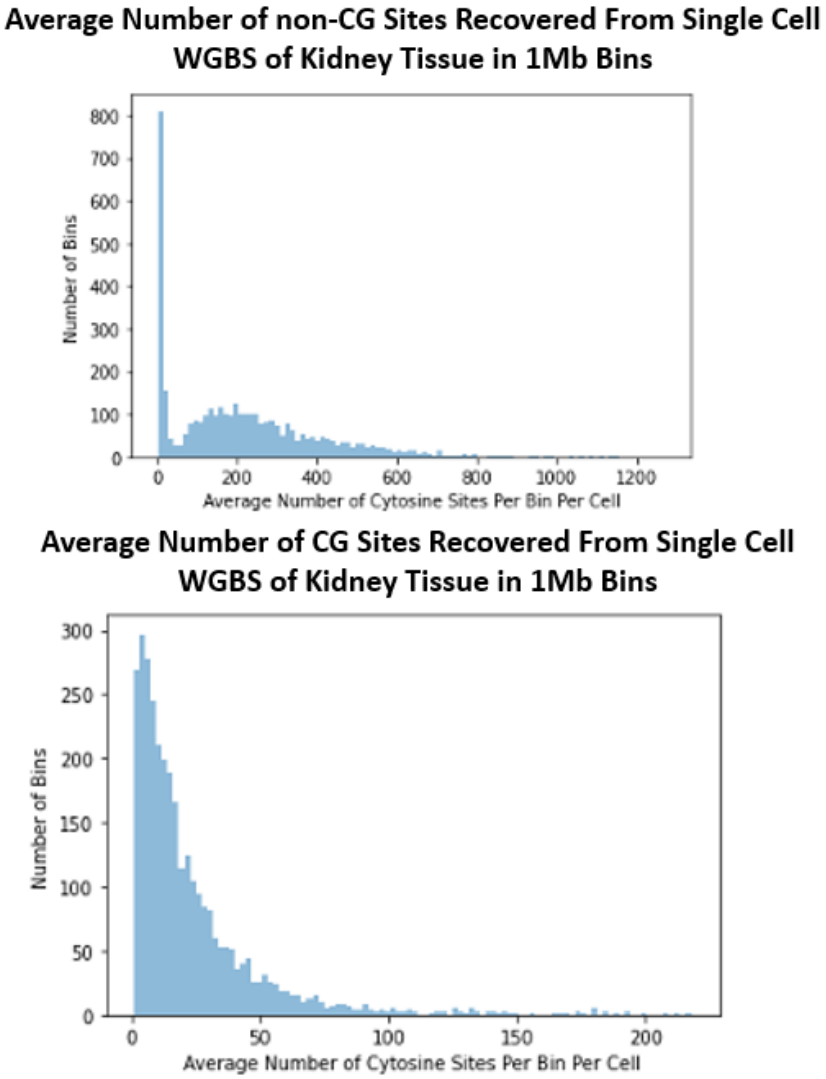


Figure 1: Comparison of the average number of CH sites in 1Mb bins per cell versus the number of CG sites.

To cluster cells, WGBS typically requires a high sequencing depth of at least 1 million unique reads per cell. RRBS aims to lower the sequencing costs by enriching for CG sites by using a restriction enzyme, MspI, that cuts at high density CG islands. However, RRBS doesn't recover biologically relevant non-CpG methylation and misses low density CG sites. Thus, single cell RRBS technologies still require sequencing depths in the millions to reads like WGBS

to perform downstream analyses. (Gu et al. 2021; Hu et al. 2016)In addition, RRBS doesn't recover variable cell type specific non-CG methylation as found in the context of brain and stem cell tissues which limits its use as a platform technique.

Typically, thousands of cell libraries are needed to characterize heterogenous human tissues. snmC-seq is by a large margin the most prolific single cell WGBS method and has been used as the backbone to methylome cell atlas studies with the ability to generate thousands of single cell methylomes per study, 10-fold higher than most other techniques. (Callaway et al. 2021) Briefly, extracted nuclei are sorted into individual reaction vessels which are given a well-specific DNA barcode during library construction.(Callaway et al. 2021; Mulqueen et al. 2018) Using a liquid handling system, this protocol can reportedly generate an astonishing 10,000 single cell methylomes per week by automating the thousands of reactions in parallel.(Luo et al. 2022) The optimized adaptation of this protocol in 384 well plates to liquid handlers is key to the high throughput of this method. However, the use of liquid handlers prevents the practical widespread adoption of this method and its ability to practically scale to millions of cells like other single cell technologies. (Cao et al. 2020; Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnel, et al. 2020)

Recent combinatorial indexing methods offer a potential solution to exponentially scale the cell throughput of single cell sequencing technologies without the extensive use of liquid handlers. Briefly, these technologies leverage a split-pool barcoding scheme that virtually creates an exponentially scaled barcode space. For example, a barcode space of 56 million barcodes can be created with 3-levels of combinatorial barcoding using 3x384 well plates. The single cell input into this barcode space is typically restricted to 10% of this barcode space to minimize the chance that two cells have the same barcode. This technique can potentially sequence millions of

cells and has been demonstrated to perform single cell RNA and chromatin accessibility sequencing of organ systems.(Cao et al. 2020; Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnell, et al. 2020) sci-MET is a recently published single cell WGBS technique that uses a 2-level combinatorial indexing approach. Isolated nuclei are first fixed with formaldehyde and then nucleosome depleted whereby a careful balance is struck between the denaturation of chromatin organization proteins for whole genome coverage and structural integrity of the nucleus. Next, thousands of nuclei per well are flow sorted into a 96 well plate, and a well specific DNA barcode is inserted using Tn5 transposase into all genomic fragments. The nuclei are then mixed and then roughly 10 nuclei are flow sorted into a second 96 well plate where bisulfite conversion takes place. Post bisulfite conversion, a second well-specific barcode is added during the final PCR. Using 2x96 well plates, this protocol demonstrated the ability to generate roughly 1000 single cells per experiment at a mean sequencing depth of 200,000 reads per cell. As indicated in this study, this method has at least 5 fold lower library complexity compared to snmC-seq.(Mulqueen et al. 2018) Because the extent of DNA accessibility to Tn5 barcoding is in tension with the structural integrity of the nucleus, the low coverage may be due to continued existence of DNA binding proteins after nucleosome depletion. SnmC-seq still comfortably scales to the throughput of sci-MET. However, continued innovation in combinatorial indexing schemes such as the addition of a third level of indexing would out scale snmC-seq as barcode layer exponentially scales the cell throughput. In addition to technological challenges to generate single cell WGBS datasets, the biological interpretation of WGBS data is also an immense challenge.

Challenges in Biological Interpretation of Single Cell DNA Methylation Datasets

CG methylation is typically associated with gene repression. For example, X-chromosome inactivation is a critical feature of female mammalian embryonic development which is established and maintained by CG methylation gene repression.(Heard, Clerc, and Avner 1997) SnmCAT-seq, derived from snmC-seq, was recently developed to profile the transcriptome, DNA cytosine methylation, and chromatin accessibility in postmortem human frontal cortex tissue.(Luo et al. 2022) Currently, this is the only study that has generated thousands of single cell coupled WGBS and RNA datasets as single cell per well methods can only reasonably generate low hundreds of cells without liquid handler robotics. This study showed that CH methylation within gene bodies of neuronal cells can have different effects in different contexts. For example, the expression of KCNIP4 has a strong negative correlation between RNA expression and gene body methylation in excitatory neurons but a slight positive correlation in inhibitory neurons. In contrast, the expression of ADARB2 shows a strong negative correlation with gene body methylation in inhibitory neurons but a slight positive correlation in excitatory neurons. Interestingly, the expression of GPC5 is positively associated with gene body methylation for both inhibitory and excitatory neurons as shown in figure 2, taken from snmCAT-seq. (Luo et al. 2022)

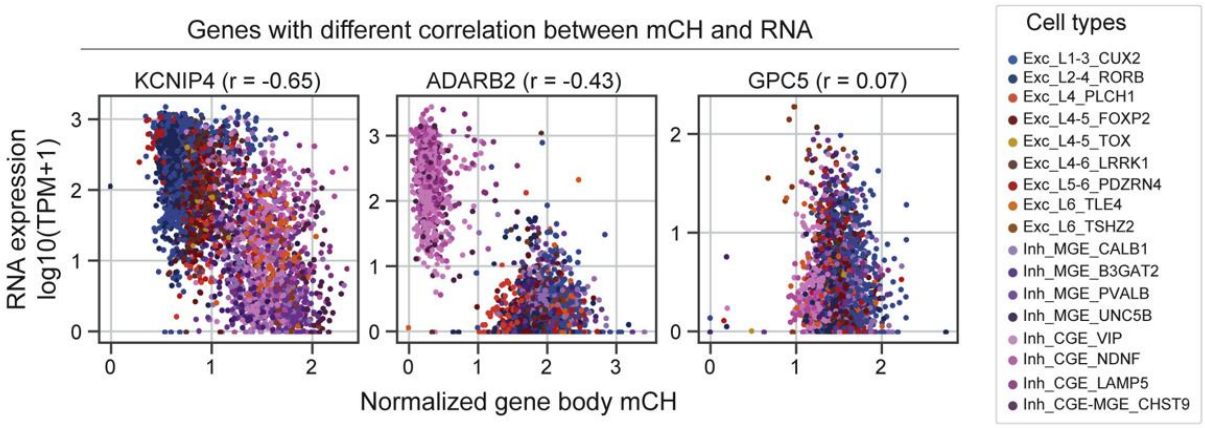


Figure 2: Variable correlations between mCH and RNA in different cell types and genes.

Another noteworthy co-sequencing method called scNMT-seq has been used to profile the transcriptome and methylome of differentiating stem cells. (Clark et al. 2018; Argelaguet et al. 2019; Angermueller et al. 2016) An RNA expression predictive model using WGBS based on these scNMT-seq studies found positive correlations between DNA methylation at promoters and gene expression for those genes. This correlation is opposite from most bulk DNA methylation studies Because the data used for training this model is from stem cell rich tissue, this opposite correlation could be a distinguishing feature of stem cells.(Uzun, Wu, and Tan 2021) Therefore, the modulation of gene activity of a nearby methylated feature is extensively cell type dependent.

The high context dependency of a methylation mark's effect on nearby gene expression presents a formidable challenge in integrating single cell methylation with the more broadly used single cell RNA modality. In contrast, single cell DNA accessibility shows relatively consistent positive correlations with nearby RNA gene expression. The Cicero computational method demonstrated that this general positive correlation can be used to quantify the relative predicted gene expression of a particular gene using only the number of DNA accessible sites and their

distances from that gene (Pliner et al. 2018). Since DNA methylation can have an ambivalent effect or no effect on gene expression depending on the context, computational methods have limited ability to couple proximal methylated features to a gene without additional RNA information. Nevertheless, single cell WGBS in the form of snmC-seq and snmCAT-seq has demonstrated cell-type clustering of brain cells with similar resolution to RNA.(Callaway et al. 2021; Luo et al. 2022) In contrast, single cell DNA accessibility clustering of human brain cells have been shown to be lowest in resolution. (Chen, Lake, and Zhang 2019; Lake et al. 2018) Thus, the integration of the methylome and transcriptome could potentially reveal how DNA methylation, at loci resolution, establishes and maintains specific cell type identity in the broader context of DNA methylation associated phenomena such as cancer and aging.

Multi-omic methods such as snmCAT-seq and scNMT-seq are therefore critical to elucidate the epigenetic context of DNA methylation for a specific cell type. These methods integrate the RNA expression and the whole genome DNA cytosine methylation of the same single cell. Nuclei are first isolated from brain tissue followed by the methylation of cytosines in the GC context of DNA accessible cytosines with GpC methyltransferase. DNA binding proteins such as nucleosomes block the inaccessible GC positions from receiving the methyl groups. During bisulfite sequencing, the unmethylated cytosines convert to thymines. As a result, cytosine conversions in the GC context are interpreted as inaccessible and vice versa. The nuclei are then flow sorted into individual reaction wells where reverse transcription and cDNA amplification with methylated cytosine takes place using the SMART-Seq protocol. The reaction then undergoes bisulfite conversion follow by post bisulfite adapter ligation using the adaptase enzyme. DNA and cDNA libraries are then co-sequenced and bioinformatically split based on highly methylated and lowly methylated reads in the CH sequence motif. Highly methylated

reads are presumed to be cDNA reads which were amplified with methylated cytosine prior to bisulfite conversion while DNA reads are lowly methylated, as expected for human cells. This crucially allows for the hypothesized biological relevance of a particular methylated locus to be cross-validated with the RNA expression of nearby genes. Like snmC-seq, this method achieves high cell throughput by flow sorting nuclei into individual wells in a 384 well plate and using optimized liquid handlers. Without one, team would have to run the snmCAT-seq protocol in at least 5,000 individual wells to generate the roughly 4,358 single nuclei datasets reported.

Challenges in Single Cell Clustering of Terminally Differentiated Cell-Types

All single cell methylation technologies suffer from low coverage of the genome. In the mammalian diploid genome, there are only two copies of a cytosine site that can be detected. In contrast, the RNA transcript drop-outs during library preparation is mitigated by the naturally high copy numbers of gene transcripts. In addition, the harsh bisulfite chemistry required for DNA methylome sequencing causes extensive DNA loss due to double stranded breaks. Thus, most genomic fragments are lost during WGBS library preparation. For example, the maximum genome coverage possible projected library complexity of scnmC-seq is 30% per cell. This poses a significant challenge to computationally cluster single cells into cell-types. In summary, the methylated cytosine information is binned across vast genomic windows (typically 100kb in size) by cell. Only bins with high coverage across all cells are considered. Single cells of the same cell type can be clustered based on similar methylation levels across these bins. Generally, millions of reads per cell are minimally required to capture enough shared methylated cytosine sites across the bins for clustering. For example, the average sequencing depth of scnmC-Seq is 5 million reads per cell to cover approximately 10% of the genome per cell to cluster brain cells.(Callaway et al. 2021) Notably, most single cell methylation or methylation and RNA co-

sequencing studies focus on mammalian neuronal or stem cell populations where cell type specific CH methylation is elevated. For example, neurons can have 5 fold more CH methylation compared to glia. In addition, both tissue types contain cell-type specific elevated 5-hydroxymethylcytosine which is captured, but not distinguished from, methylcytosine during WGBS. For example, it's estimated that 5hmC methylation is 40% as abundant as CG methylation in Purkinje cells.(Kriaucionis and Heintz, n.d.) Therefore, informative methylation features are bolstered by the additional cell type specific 5hmC sites. In contrast, terminally differentiated tissues demonstrate low levels of CH methylation. Thus, CG methylation would be used to cluster single cells. We have found that the number of CH sites can be over 5-10 fold more abundant than CG sites based on our WGBS study on kidney tissue. Therefore, it's plausible that the required sequencing depth to cluster terminally differentiated cell types will require vastly more than 10% genome coverage, possibly beyond the snmC-seq projected maximum library complexity of 30%.(Luo et al. 2018) Unsurprisingly, single cell methylation of terminally differentiated tissue remains vastly understudied because of these complications.

Multi-omic technologies such as snmCAT-seq offer part of the solution to studying the methylome of terminally differentiated tissues. With multi-omic RNA and WGBS co-sequencing, single cells can be clustered and grouped into a pseudo-bulk with as little as 50,000 unique RNA reads per cell. These cell type group labels can be then transferred to the WGBS library where these same cells can be pooled into a pseudo-bulk. Differential methylation analysis can then be performed between these pseudo-bulk profiles defined by the RNA cell type label. This framework leverages the powerful ability of single cell RNA-seq to discriminate most cell types as demonstrated by numerous cell atlas studies of human organs using the transcriptome. (Quake 2022) For example, if the single cell methylome library is sequenced to

1,000,000 reads per cell, roughly 500 cells within a cell type pseudo-bulk would be needed to have 30X coverage of that cell type as shown in Figure 3. This high coverage could plausibly contain enough CG methylation information to identify novel cell-type specific CG methylation features, currently understudied in terminally differentiated tissue. Furthermore, the methylome of rare cell types that can only be observed in high throughput single cell RNA-seq experiments could also be profiled. This analysis framework requires an ultra-high throughput method on the order of tens of thousands of cells. Basically, a higher throughput co-sequencing assay results in higher methylome coverage of a particular cell type as more cells constitute the corresponding methylome pseudo-bulk. All DNA methylation and RNA co-sequencing platforms currently lack the cell throughput required for this analysis.

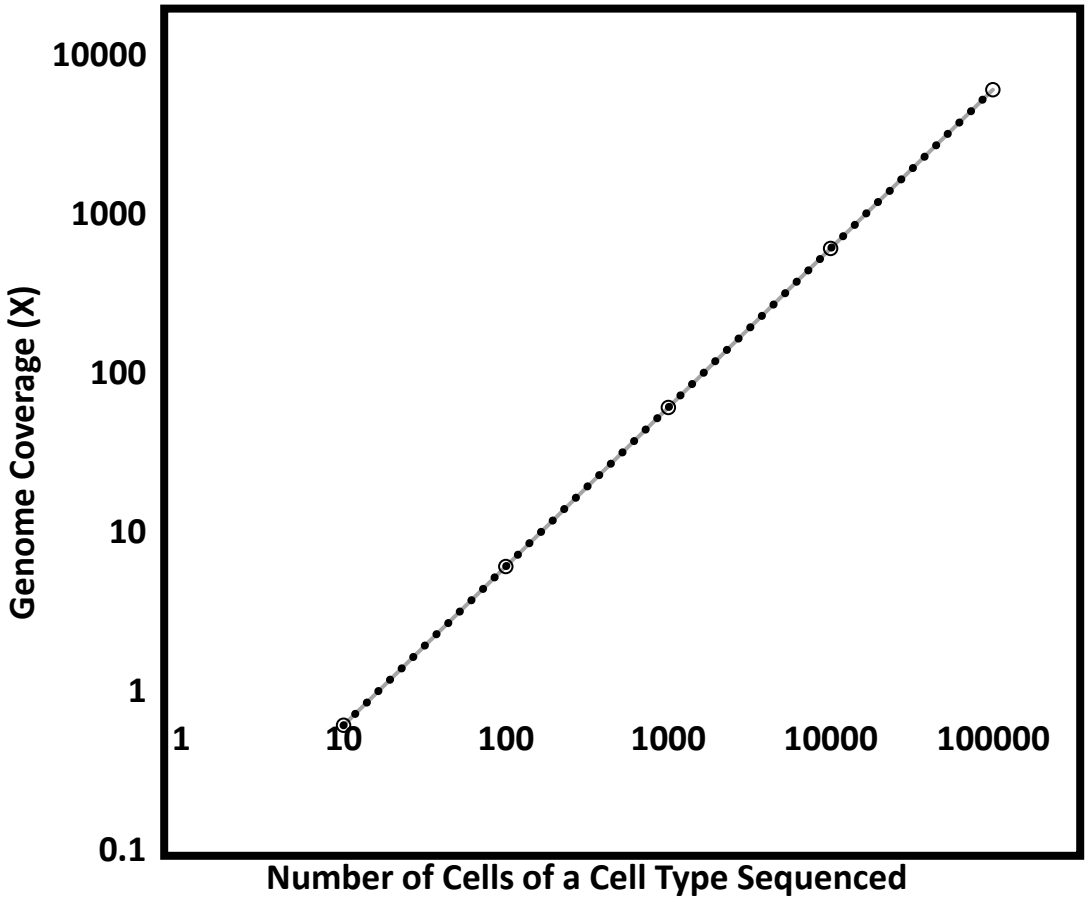


Figure 3: Genome coverage of a pseudo-bulk cell type based on the number of cells sequenced at 1M reads/cell within the pseudo-bulk.

Motivations for the Development of the Method

We seek to build upon existing multi-omic DNA methylation and RNA co-sequencing technologies by expanding the throughput from hundreds of cells to tens of thousands of cells per experiment. Here, we propose an ultra-high cell throughput multi-omic DNA methylation and RNA co-sequencing platform as the basis for the pseudo-bulk analysis framework previously mentioned. This method utilizes a combinatorial indexing approach inspired by sci-MET, but crucially increases the throughput of this scheme 100-fold to allow sequencing of tens of thousands of cells using 3x96 well plates by adding a third round of barcoding in one experiment. We demonstrate how the nucleosome depletion process as described in sci-MET

severely reduces the structural integrity of the nucleus, preventing the additional reverse transcription and barcoding reactions required for 3-level co-sequencing of DNA methylation and RNA. We show case a solution that involves the simultaneous encapsulation and lysis of single cells or nuclei within polyacrylamide hydrogel beads. This combinatorial indexing vessel, in contrast to nucleosome depleted nuclei, displays drastically higher vessel stability, allowing for the robust addition of reverse transcription and additional barcoding reactions beyond 3-levels. For example, the polyacrylamide remains intact after exposure to high concentrations of SDS and protease K which is crucial to robustly denature DNA binding proteins. We describe a 3x96 well plate that can sequence 50,000-100,000 single cells per experiment. However, a 3x384 well plate adaptation could sequence 3-5 million single cells per experiment. Ultimately, this technology aims to be the next step in single cell WGBS and RNA co-sequencing technology development by unlocking the possibility to profile the methylomes of terminally differentiated tissues using an ultra-high throughput approach.

Scope of the Dissertation

Here, we describe the development of a novel combinatorial indexing method where single cells or nuclei are simultaneously encapsulated and lysed within polyacrylamide gel beads. These gel beads act as the vessel that compartmentalizes both the DNA and RNA during the barcoding steps. The motivation of this gel bead encapsulation method stems from our difficulties in adding additional reactions to reverse transcribe RNA and perform additional barcoding using nucleosome depleted nuclei. The design of this novel gel bead platform was the first aim of this work, resulting in the development of a gDNA and RNA co-sequencing platform. This platform could be published by applying it to the profiling of DNA copy number variations in various cancers and their effects on cancer cell RNA expression. The second aim of

this work was the design of a novel bisulfite sequencing method to solve the unique challenges involving the bisulfite conversion of these polyacrylamide gel beads. This novel approach could solve some existing problems in bulk WGBS. Upon slight optimization, this new method could offer a better bulk bisulfite sequencing method. Finally, the third aim of this work was to couple both the first and second aims to develop a whole methylome and transcriptome co-sequencing method. We showcase optimizations performed to build upon this third aim and future work required to publish this method.

CHAPTER 1: DESIGNING A SYNTHETIC COMBINATORIAL INDEXING VESSEL

1.1 Abstract

Our 2-level combinatorial indexing design incorporates an initial hyperactive Tn5 tagmentation step to insert the first barcode followed by a PCR step to add the second barcode. Whole genome sequencing of DNA methylation using this approach is challenging because the complete lysis of histone proteins is required. The DNA is tightly wound around these proteins which block the accessibility of the DNA to hyperactive Tn5. Simply, without the denaturation of these histone proteins, only the small portion of originally accessible DNA will be sequenced. However, the methods to denature these proteins also compromise the structural integrity of the nucleus which is required through the first tagmentation step. Here, we explore several strategies to denature histone proteins while immobilizing the DNA and RNA. We identified that single cell encapsulation and lysis within polyacrylamide gel beads was the best solution. The gel beads provide the structural integrity to withstand high concentrations of SDS and protease K to denature histone proteins. In addition, the DNA is intertwined in the mesh of the gel bead and the RNA is anchored to the gel bead matrix with the use of an acrydite modified primer. Although our initial tests showed high barcode doublet rates ~30-50%, we identified a novel strategy to perform DNA and RNA co-sequencing.

1.2 Introduction

Single cell methods require the compartmentalization of either DNA or RNA during the single cell barcoding steps. In the case of single cell per well methods, the reaction well physically provides this compartmentalization where the nucleic acids of each single cell is given a well specific barcode. For example, in snmCAT-seq the well specific barcode is added to both

the DNA and RNA during the post PCR bisulfite conversion(Luo et al. 2022). In the case of combinatorial indexing methods, the cell nucleus provides the compartmentalization during the combinatorial barcoding steps(Mulqueen et al. 2018). Therefore, the success of this technology depends on the single cell compartmentalization of both the DNA and RNA through the combinatorial barcoding steps.

The cell or nucleus must be completely lysed because DNA binding proteins such as nucleosomes only allow the accessible DNA to be barcoded. This blocking of barcoding enzymes by nucleosomes is the basis of existing DNA accessibility combinatorial indexing technologies like sci-ATAC seq. In contrast, whole genome sequencing methods require the inaccessible DNA to also be barcoded. Therefore, these DNA binding proteins must be adequately denatured. For single cell per well methods, single cells or nuclei are fully lysed in the well. In the case of snmCAT-seq, the nuclei are sorted into a reverse transcription buffer that also permeabilizes the nuclei allowing reverse transcriptase to access the nuclear RNA. The thermocycling that accompanies amplification of full-length cDNA and subsequent bisulfite conversion denatures the nucleus and chromatin organization proteins. This process allows for both the DNA and cDNA to be fully accessible to the post bisulfite adapter tagging enzyme, adaptase, theoretically barcoding the full methylome and transcriptome. The challenge for whole genome combinatorial indexing is that the full lysis of DNA binding proteins often results in the lysis of the nucleus. However, the structural integrity of the nucleus is required to compartmentalize the DNA and RNA during combinatorial indexing. In the case of sci-MET, this problem is mitigated by first fixing the cells or nuclei with formaldehyde followed by SDS treatment. This careful balancing of fortifying the nuclear structure yet denaturing some of the DNA binding proteins allows for both increase genome coverage and nucleus integrity. As

published, this balanced technique is called nucleosome depletion. Although increased genome coverage is demonstrated, the genome coverage is 5-10 fold lower than single cell per well methods as not all the DNA binding proteins are denatured. (Mulqueen et al. 2018)

We explored different DNA and RNA immobilization and whole methylome accessibility techniques. We first tried to adapt the sci-MET nucleosome depletion technique to reverse transcription, required for transcriptome sequencing. This approach failed because the nucleosome depletion severely compromises the nuclear integrity causing 90% of the nuclei to be destroyed after reverse transcription. This motivated us to identify a more robust vessel with much higher structural integrity compared to the nucleosome depleted nuclei. This led to the development of a simultaneous cell encapsulation and lysis within hydrogel beads method. With inspiration from previously published combinatorial indexing RNA sequencing methods, in-nuclei reverse transcription was performed followed by nuclei encapsulation and lysis by high concentrations of SDS and proteinase K. (Rosenberg et al., n.d.; Plongthongkum et al. 2021; C. Zhu et al. 2019) The microfluidic hydrogel encapsulation approach offers the added advantage of using strong protein denaturation buffers to ensure the complete denaturation of DNA binding proteins, and the robust compartmentalization of nucleic acids. This high stability allows for the easy incorporation of reverse transcription and additional barcoding enzymes to allow for the development of a 3-level WGBS and RNA co-sequencing platform.

We will describe below how we use DNA staining and imaging, to confirm the adequate lysis of DNA binding proteins. However, the immobilization of RNA required the screening of different hydrogel structures. Simply, the RNA is over 50,000X shorter in length than DNA which allows the RNA to easily diffuse out of the hydrogels. Here, we describe three hydrogel structures: agarose gel beads, polyethylene glycol (PEG) gel beads, and finally polyacrylamide

gel beads. The polyacrylamide gel beads offered the best solution as reverse transcription primers could be modified with an acrydite group. During gel polymerization, this acrydite modified primer covalently anchors the cDNA to the polyacrylamide matrix. The long DNA is intertwined in the polyacrylamide gel matrix. Thus, this structure successfully immobilizes both the fully accessible DNA and RNA which enables whole genome and transcriptome combinatorial indexing. We demonstrate the success of this approach by performing single cell whole genome and transcriptome sequencing on a mixture of human and mouse cells. After sequencing, we observed cell barcodes that contained only human or mouse reads.

1.3 Methods & Results

1.3.1 Nucleosome Depletion Adaptation

My initial approach was inspired by two combinatorial indexing techniques: sci-RNA seq and sci-MET seq. The goal was to combine the in-nuclei reverse transcription technique to process the RNA and the nucleosome depletion technique for whole genome barcoding. I first tried to perform nucleosome depletion followed by reverse transcription to generate a nuclear structure containing nucleosome depleted DNA and cDNA. Using a two-level indexing scheme, 1000-2000 nuclei would first be FACS sorted into a 96 well plate where Tn5 would be used to add the first cell barcode. The nuclei would then be pooled and then 10-20 nuclei per well were FACS sorted into as second 96 well plate where PCR indexed adapters reverse complement to the Tn5 adapter sequences would be used to add the second bell barcode, completing the combinatorial indexing process.

The primary issue with nucleosome depletion was the integrity of the nuclei following depletion. This was assessed by first staining the nuclei with a standard DNA stain, DAPI. Intact nuclei contain higher levels of DAPI compared to nuclear/chromosome debris. The number of

intact nuclei and nuclear debris can be measured using FACS and plotting the DAPI fluorescent intensity. Figure 4 illustrates the large difference of intact nuclei after nucleosome depletion compared to freshly isolated nuclei. Briefly, the FACS machine measures the forward and side light scattering and DAPI fluorescent intensity of the nuclei or debris. A gate is manually drawn to distinguish nuclei from debris. Particles with sufficient DAPI fluorescence are collected as nuclei whereas all other particles of lower fluorescence are assumed to be debris. For clarity, the DAPI gate is labeled in each plot. Freshly isolate nuclei are first sorted to identify a baseline DAPI fluorescent intensity. Examining the DAPI signal plot, most particles have high DAPI signal and a threshold of 1000 460/50[405] is used to differentiate intact nuclei and debris. Next, nucleosome depleted nuclei are sorted using the same DAPI fluorescent threshold. Clearly, the nucleosome depletion process generates large amounts of nuclear debris as a large population of particles have low DAPI fluorescence.

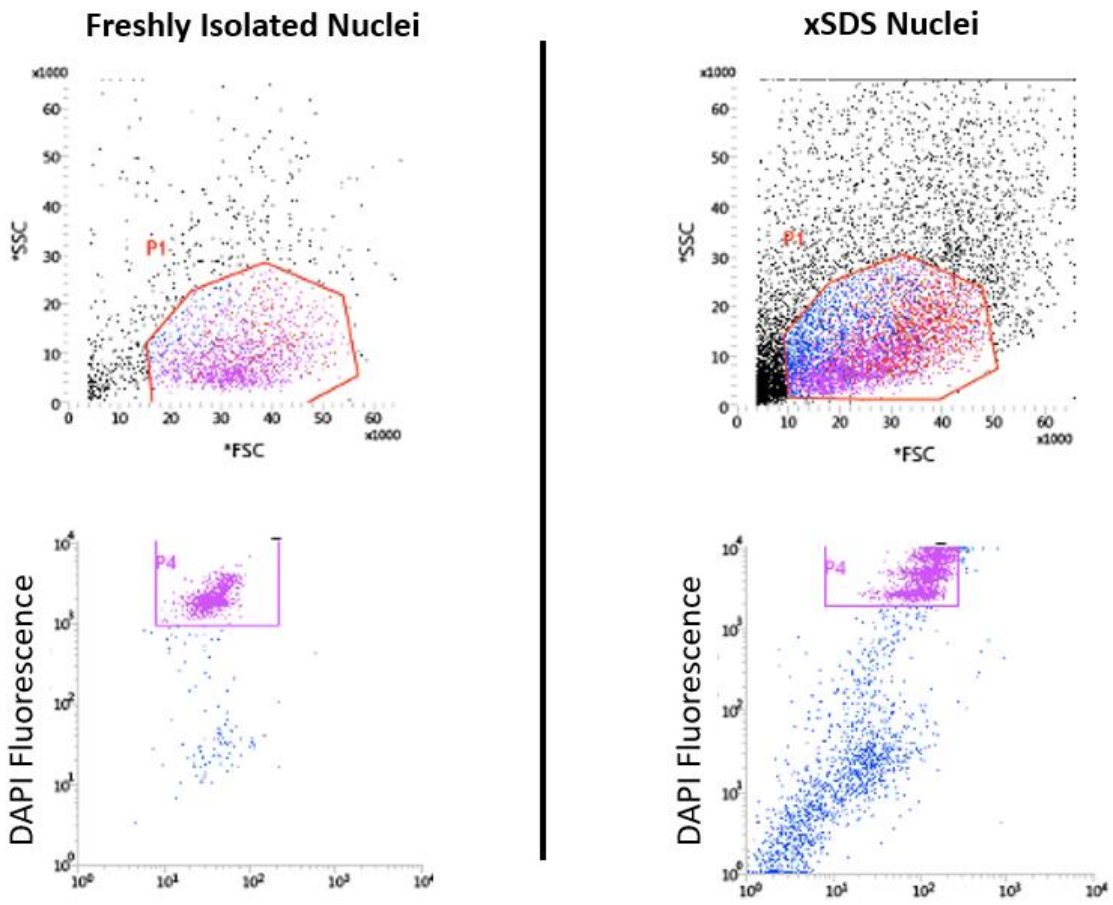


Figure 4: FACS plots comparing the proportion of DAPI positive events before and after nucleosome depletion.

This FACS study encapsulates the immense difficulty in recovering intact nuclei from each required reaction. After nucleosome depletion, the SDS and formaldehyde are removed by pelting the nuclei and removing the supernatant. The nuclei were then resuspended in reverse transcription buffer and incubated at 55C for 10 minutes, following the sci-RNA seq reverse transcription protocol. Afterwards, the nuclei are pelleted and resuspended in PBS containing DAPI for FACS. In our experiments, we typically recover only 1% of nuclei after nucleosome depletion and reverse transcription from FACS. Thus, this nucleosome depletion approach was deemed unfeasible.

1.3.3 In-Nuclei cDNA Generation and Agarose Gel Encapsulation

The immense difficulty in handling nucleosome depleted nuclei motivated a new approach inspired by SiC-seq where single microbes were encapsulated in agarose micro-sized gel beads, lysed with SDS and proteinase K, and finally individually barcoded using a system of microfluidic devices.(Lan et al. 2017) We sought to adapt the agarose gel bead encapsulation and lysis approach to immobilize the DNA and cDNA of nuclei after in-nuclei reverse transcription. The microfluidic device used to achieve this encapsulation was custom designed by a previous PhD student, Andrew Richards. The specific microfluidic device engineering and encapsulation protocol is detailed in the supplemental methods. With inspiration from InDrops and Drop-seq, the microfluidic device encapsulates single cell or nuclei within oil droplets. In our adaptation, we create a suspension of single nuclei in low melting temperature agarose kept at 37C. We input this mixture through the encapsulation device along with 0.5% SDS and 0.016U/μL proteinase K. A space heater is used to warm the encapsulation device and fluid reservoirs to 37C to prevent gelling of the agarose prior to encapsulation. Figure 5 illustrates this process as

described.

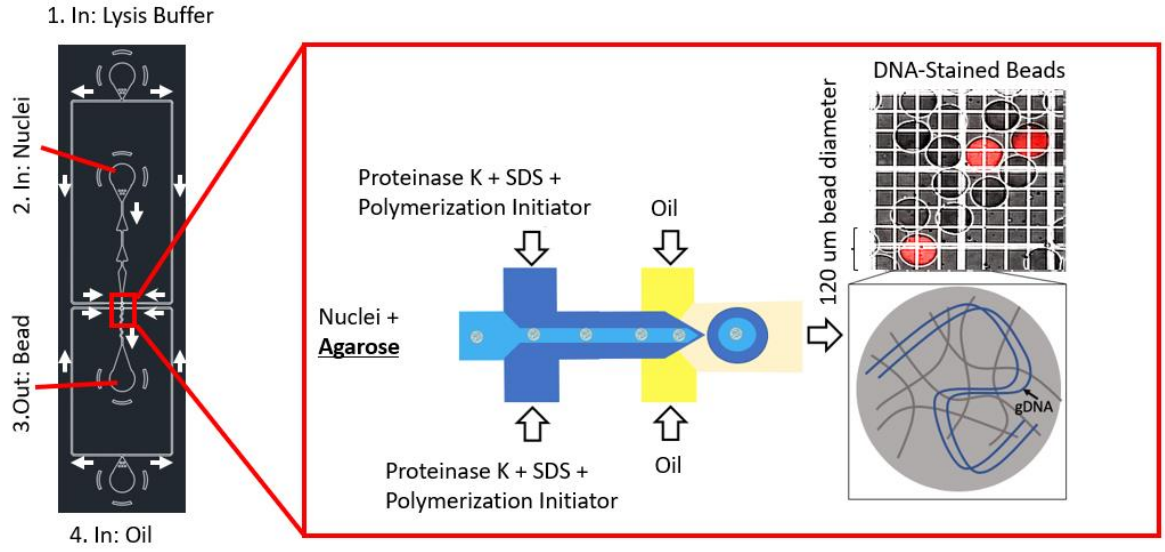


Figure 5: Microfluidic encapsulation scheme of nuclei in low melting temperature agarose.

Agarose demonstrates robust structural integrity when exposed to high concentrations of SDS and proteinase K. The size of a typical nucleus is roughly 1-5 microns while the gel bead is roughly 120 microns in diameter. The DNA content of gel beads can be visualized by staining them with DAPI. The robust denaturation of DNA binding proteins can also be confirmed by observing the diffusion of DNA throughout the hydrogel matrix as shown in figure 6.

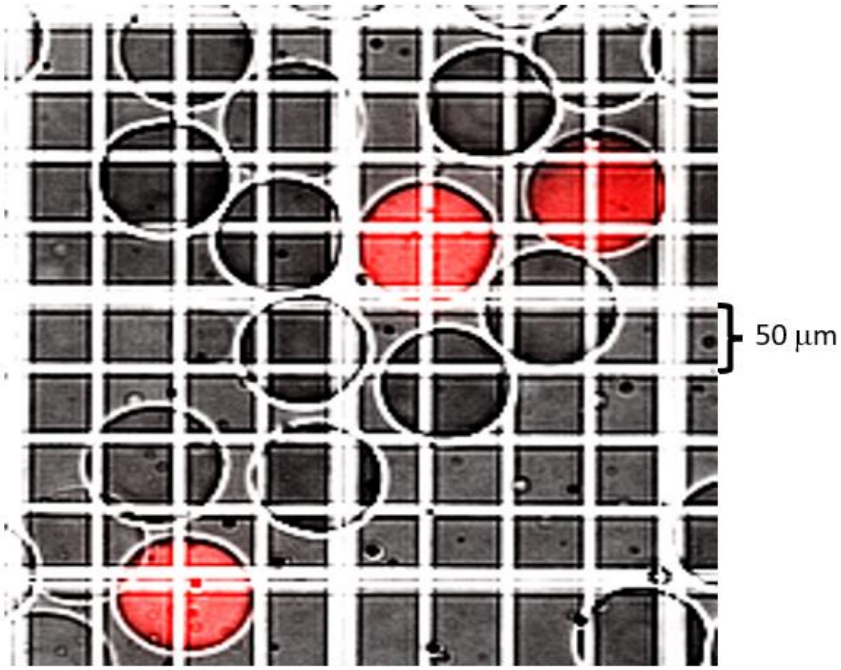


Figure 6: Agarose microbeads containing encapsulated and lysed nuclei. DNA content of the beads are stained with the DNA stain, DAPI. The microbeads were placed on a hemocytometer to quantify the bead size and imaged using a fluorescent microscope at the excitation wavelength for DAPI. About 10% of beads are occupied by nuclei.

The encapsulation of single cells or nuclei can be described by a Poisson probability distribution as described in previous cell encapsulation methods such as InDrops and Drop-Seq.(Klein et al. 2015; Macosko et al. 2015) Using the volume of the gel bead and a goal of roughly 10% of beads occupied by single nuclei, 90% of beads empty, and negligible numbers of beads containing multiple nuclei, we used the Poisson distribution to predict the required concentration of nuclei prior to encapsulation as 3000 nuclei/μL. After encapsulation, the occupancy of the beads is visually calculated by counting the number of empty beads and stained beads. With 10% of the beads DAPI positive, we verified that our encapsulation method follows a Poisson distribution as described previously. (Klein et al. 2015; Macosko et al. 2015)

In our adaptation, nuclei are first freshly isolated from cultured cells and then undergo the reverse transcription and second strand synthesis reactions previously described in sci-RNA seq. Afterwards, the nuclei are washed once with nuclei isolation buffer without NP-40 and filtered through a 30-micron filter to remove nuclei aggregates. The nuclei were then resuspended in a low melting temperature 1.5% agarose PBS mixture pre-warmed to 37C to prevent gelling. Encapsulation was then performed using a microfluidic device illustrated in Figure 5. To keep the agarose from polymerizing, the encapsulation was performed with a space heater to keep the agarose on the device and in the fluid reservoirs at roughly 37C. Figure 7 illustrates the general steps prior to gel bead formation. Post encapsulation, the agarose gel beads were removed from the emulsion using previously described methods. (Klein et al. 2015) Briefly, the emulsion oil was carefully removed, and the emulsion was broken using 20% 1H,1H,2H,2H-Perfluorooctan-1-ol (PFO) v/v in HFE7500. The beads were then washed with 1% Span80 in hexane followed by 0.1% tween 20 in Tris HCl ph=7.5. The agarose beads were then wash twice in H2O and then

stained with DAPI to calculate occupancy and establish input amounts for DNA and cDNA library generation.

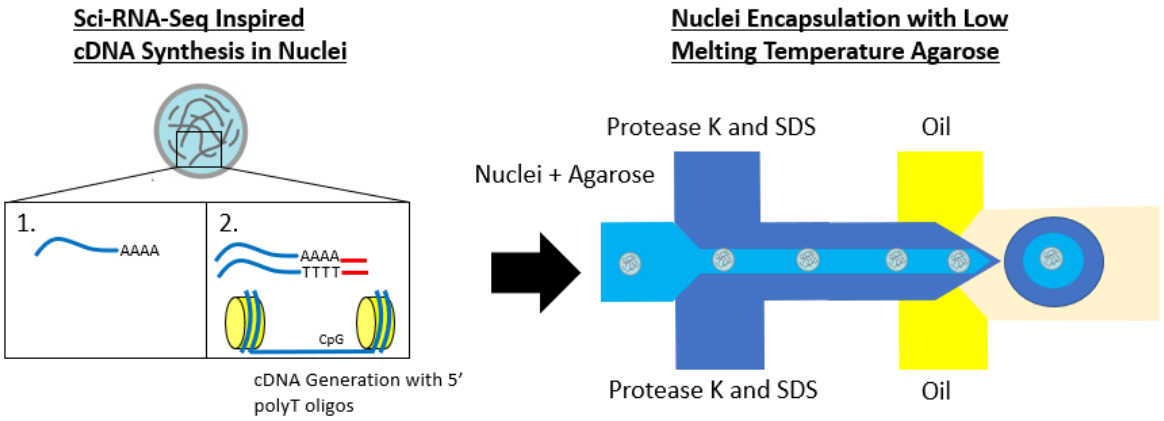


Figure 7: cDNA synthesis and nuclei encapsulation with molten low temperature agarose.

Due to its length the DNA could safely be assumed to be immobilized in the gel matrix. However, the cDNA could freely diffuse out of the gel bead. To assess this possibility, we first used Tn5 to tagment the cDNA and then amplified the cDNA using PCR primers reverse complement to the reverse transcription primer and the Tn5 adapter. We then use qPCR to quantify the amount of encapsulated cDNA compared to a positive control: cDNA in a tube and a negative control: no cDNA. From this experiment, we realized that cDNA was not retained inside of the gel bead as the amplification dynamics of the agarose gel bead samples matched that of the negative control.

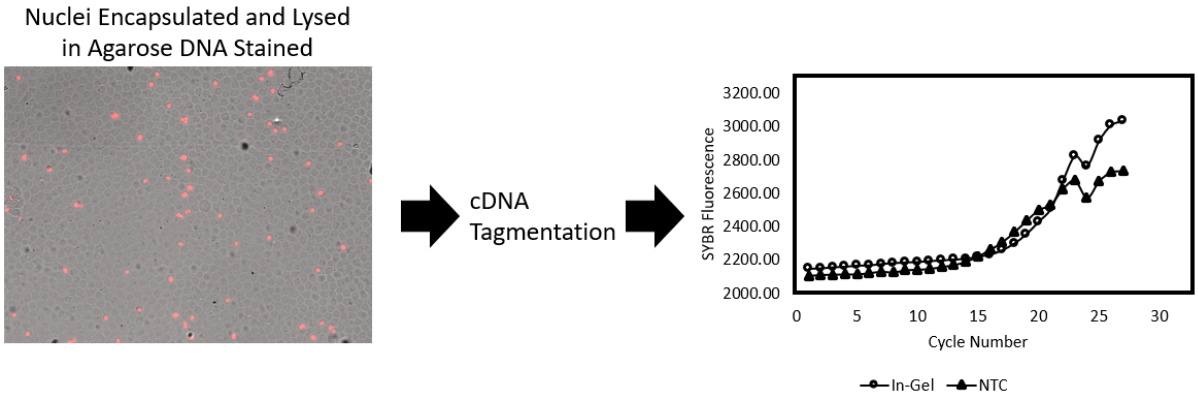


Figure 8: cDNA retention experiment showing loss of cDNA due to diffusion out of the gel bead matrix

Figure 8 illustrates the workflow of this experiment. Although the agarose gel bead structure was relatively simple to work with due to the ease of nucleic acid extraction under heat, the large pore sizes (estimated to be between 100-200 nanometers) resulted in loss of the cDNA.

1.3.4 PEG Acrylate Gel Formation

Our next approach was to lower the pore size of the hydrogel. We considered a polyethylene glycol (PEG) hydrogel inspired by a virtual microfluidic method by the Paul Blainey group.(Xu et al. 2016) The pore size diameter of this hydrogel was reported to be 25 nanometers, drastically lower than the pore size of the agarose gel. Using a flexible chain molecular model, we estimated that DNA above 60 bases could potentially be immobilized by this gel.(Pluen et al. 1999) With a protocol like the agarose one described previously, we planned to encapsulate nuclei with this polymer. In our adaptation, we planned on resuspending the nuclei in 8-arm PEG and co-encapsulated the nuclei with a thiol-PEG crosslinker dissolved in 0.5% SDS. Proteinase K was removed since proteinase K destroys the ester bonds formed during gel polymerization. This process was described in Figure 9.

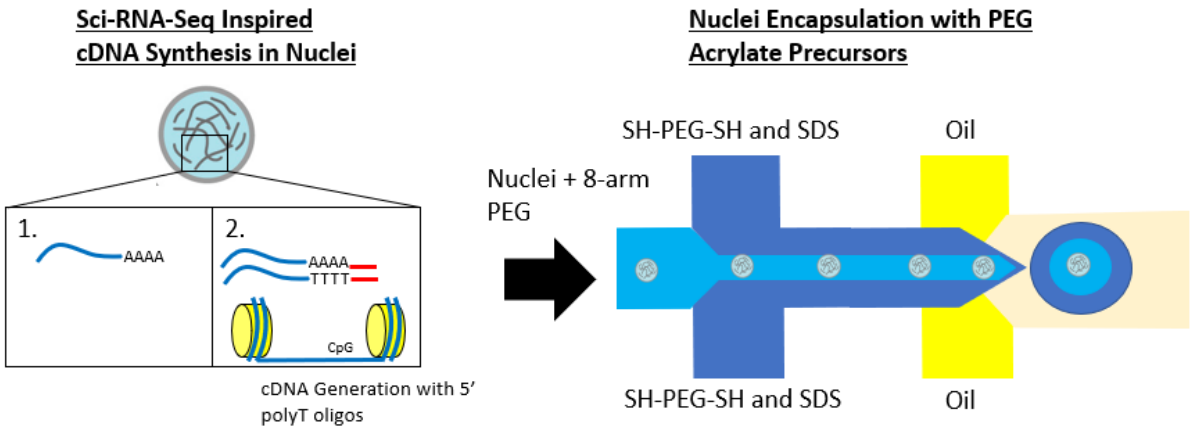


Figure 9: The encapsulation scheme using PEG based hydrogel formula published previously

To test the retention of DNA within these gel beads, we first tried to encapsulate a DNA ladder in the size range of cDNA and washed these gel beads. We then dissolved the gel beads with proteinase K and then ran a polyacrylamide gel electrophoresis experiment with the unencapsulated ladder. After DNA staining and imaging the gel, we used the ratio of fluorescent intensity of the DNA unencapsulated ladder and encapsulated ladder to estimate the loss of encapsulated DNA based on size. Unfortunately, we noticed clear loss of DNA within the typical size range of cDNA (800-2000 bases in length) as shown in Figure 10.

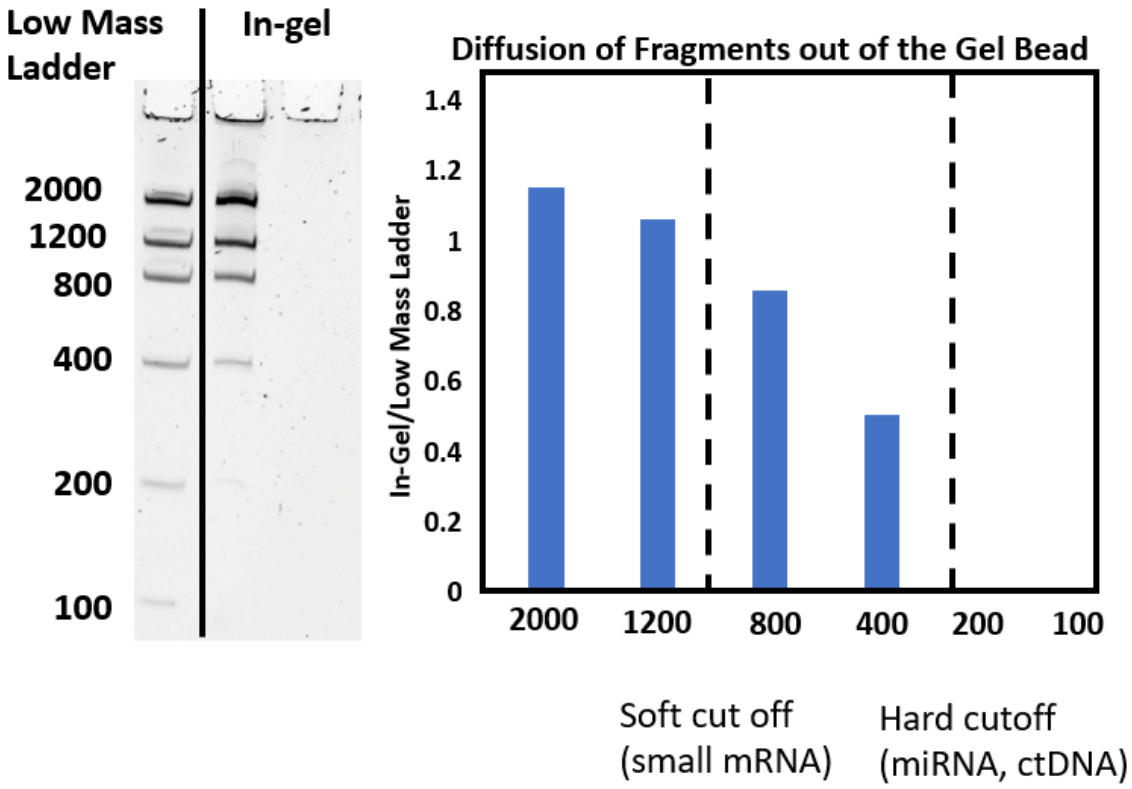


Figure 10: Loss of DNA ladder after encapsulation of the ladder with PEG microbeads

From these experiments, we realized that the amount of bead pelleting and washing post encapsulation to break the emulsion adds a fluidic and mechanical force that will cause the cDNA to diffuse out of the gel bead. Thus, we concluded that cDNA can only be immobilized chemically, ideally with a covalent bond.

1.3.5 Polyacrylamide Gel Formation

Our next approach took inspirations from a version of polony sequencing developed by the George Church group.(Mitra and Church 1999) In this iteration, we designed a polyacrylamide hydrogel approach where the reverse transcription primer contains an acrydite modification allowing it to be covalently anchored to the gel bead matrix during gel bead

polymerization. Like previous approaches, cDNA is first synthesized in-nuclei using an adapted version of the sciRNA-seq protocol with acrydite modified reverse transcription primers. The nuclei are then pelleted and resuspended in an encapsulation buffer containing acrylamide monomers. The nuclei are then encapsulated with protease K, SDS, and a bisacrylamide crosslinker as illustrated in Figure 8. This solution proved to be the correct approach to immobilize both the DNA and cDNA. The polyacrylamide hydrogel is also structurally resistant to SDS and proteinase K. Through the acrydite modification, the synthesized cDNA using the reverse transcription primer is covalently anchored to the polyacrylamide matrix.

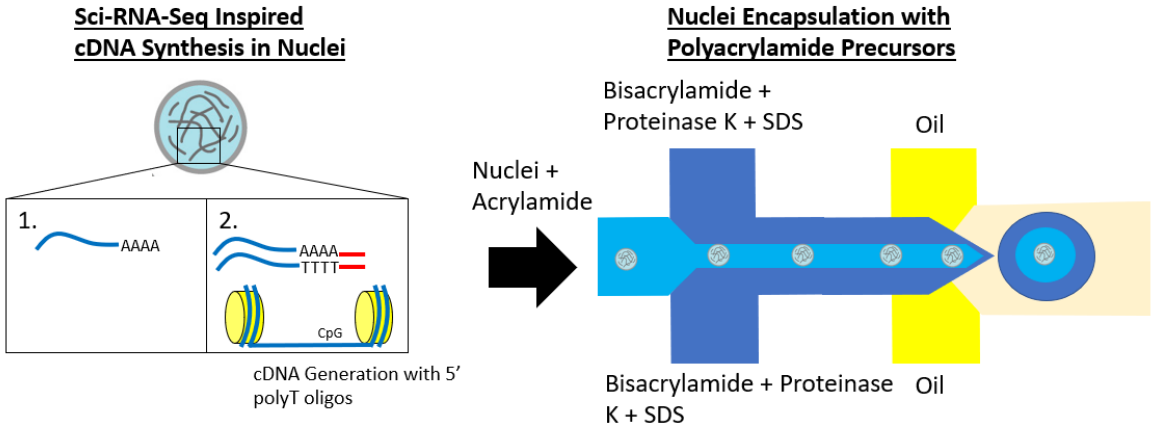


Figure 11: Nuclei encapsulation with polyacrylamide precursors scheme

Figure 12 illustrates the gel bead structure where the DNA is intertwined in the gel bead matrix and cDNA is covalently anchored into it.

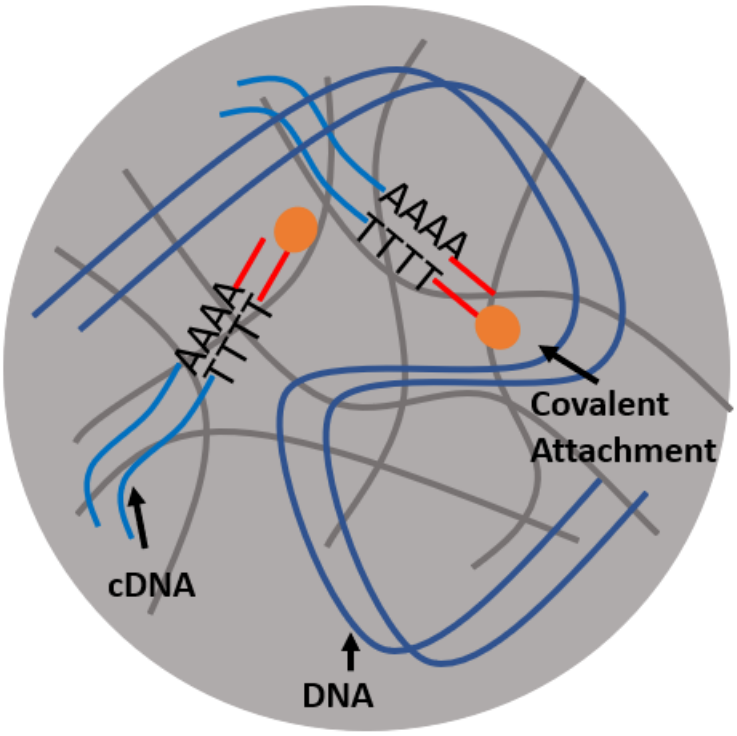


Figure 12: The encapsulated and lysed nucleus within the polyacrylamide gel bead.

To assess the efficiency of acrydite incorporation, we performed a polyacrylamide electrophoresis experiment where the polyacrylamide gel beads were directly added to the wells of the gel during electrophoresis. In parallel, we ran a denaturing polyacrylamide electrophoresis experiment where the cDNA within the polyacrylamide beads was first denatured in urea at 98C for 5 minutes and then placed on ice for 2 minutes. These gel beads were then directly added to the wells of a polyacrylamide gel infused with urea to keep the cDNA denatured. Because only one strand of the cDNA is anchored to the gel bead, the complement strand will migrate through the polyacrylamide gel infused with urea after urea denaturation of the cDNA. In contrast, the undenatured cDNA will not migrate through the gel during electrophoresis. Figure 13 illustrates this concept and shows how we validate the robust covalent anchoring of cDNA in the gel bead.

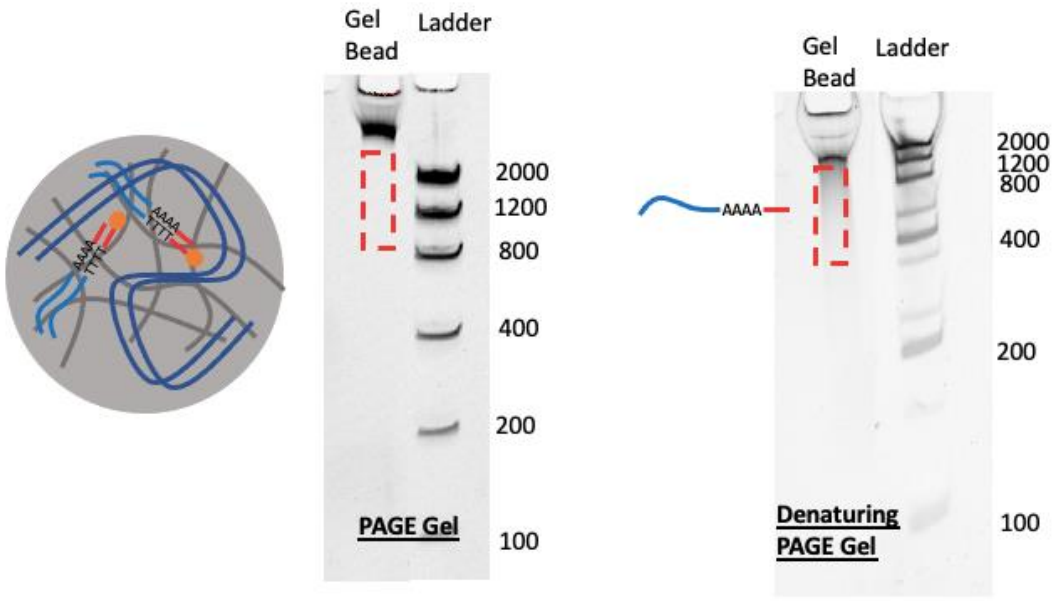


Figure 13: Left, PAGE, and right, Urea-PAGE, demonstrating the robust anchoring of cDNA to the gel bead.

1.3.6 sciGel Version 1: gDNA and RNA Sequencing Library Formation Protocol

We next tested our combinatorial indexing scheme on this gel bead structure by processing an even mixture of mouse and human nuclei. In this way, the success of our combinatorial indexing scheme can be revealed after sequencing the library. Nuclei barcode combinations that contain only human reads or mouse reads suggest single cell resolution. Barcodes that contain both human and mouse reads are considered mixed. In an even mouse and human mixture, the barcode collision rate is estimated as two times the mixed rate as mouse-mouse and human-human doublets cannot be measured. The detailed nuclei isolation protocol can be found in the supplementary methods section. Briefly, cDNA is synthesized in-nuclei like in previous designs. The nuclei are then simultaneously encapsulated and lysed using the same microfluidic device. After an overnight polymerization, the emulsion is broken to extract the gel beads. The beads are then stained with DAPI and the occupancy and concentration of nuclei are calculated. 100-200 nuclei/well are added to a 96 well plate and then tagmented with Tn5

mixture loaded with two different transposon sequences now referred to as Tn5 A and Tn5 B. This Tn5 A is well specific and contains the first nuclei barcode to the DNA and cDNA while Tn5 B is simply a PCR handle. The beads are then pooled and washed twice with 0.1% tween20 in Tris-HCl pH=8 followed by two washes in H2O by pelleting the beads 300xg for 2 minutes. The beads are then counted with a hemocytometer and then 10-20 nuclei are split into a second 96 well plate. The Tn5 was denatured with 0.1% SDS and then quenched with 2% Triton-X. PCR master mix was then added to each well with a PCR primer reverse complement to the cDNA capture primer. Because polyacrylamide is extremely stable, we discovered that in-gel PCR had to be used to extract both the gDNA and cDNA.

The cDNA was then linearly amplified for 10 cycles. Then, a well specific PCR primer reverse complement to Tn5 A and a PCR primer reverse complement to Tn5 B was added. Both the cDNA and gDNA was then exponentially amplified together for 6 cycles. Each reaction was then individually bead purified with SPRI beads at a 0.8X ratio. The eluted, DNA/cDNA was then evenly split into two separate plates. One plate finishes the amplification of cDNA by adding a P7 primer reverse complement to the reverse transcriptase primer and a P5 primer reverse complement to the Illumina P5 sequence. The other plate finished the amplification of DNA by adding PCR primers reverse complement to the Illumina P5 and P7 sequences. After amplification is complete, both the DNA and cDNA libraries are separately pooled and bead purified twice with SPRI beads at a 0.8X ratio. PAGE was then performed to confirm successful library generation illustrated by a smear between 200-600 bp. The libraries were sequenced with a MiSeq. The detailed version of this protocol and sequencing scheme is in the supplementary methods.

1.3.7 Accompanying Bioinformatic Methods

Briefly, libraries were first demultiplexed using index 1 used to distinguish cDNA libraries from DNA ones using bcl2fastq. Deindexer was used to demultiplex both DNA and cDNA libraries into individual cell barcode files based on the Tn5 and PCR barcodes. The files were then concatenated while retaining the cell barcode in the read ID of the fastq file. Adapter sequences were then trimmed from both the DNA and cDNA concatenated files using cutadapt. The DNA library was aligned to a concatenated human and mouse genome using bowtie2. Similarly, the RNA library was aligned to a concatenated human and mouse genome using STAR. The dropEst package was then used to collapse the cDNA UMI space and generate a cell barcode x gene counts matrix. We then quantified the amount of human and mouse reads for each cell barcode and then plotted them. The detailed version of this method is in the supplementary methods.

1.3.8 Species Mixing Results

Figure 14 illustrates the workflow described previously with the species mixing plot shown. Here, each point is a recovered cell or nuclei barcode and the coordinates of each point quantify the amount of human and mouse reads for that specific barcode. We observed points that aligned with both the human and mouse axes indicating the presence of single cells for both the DNA and cDNA libraries. However, we identified about 25% of the barcodes were mixed resulting in a high barcode collision rate of about 50%. This means that about half of our datasets were single cells while half of our datasets were doublets. Despite this high collision rate, we demonstrated a promising result that our polyacrylamide gel encapsulation scheme with acrydite modified reverse transcription primers could result in single cell gDNA and RNA libraries co-sequenced from the same cell.

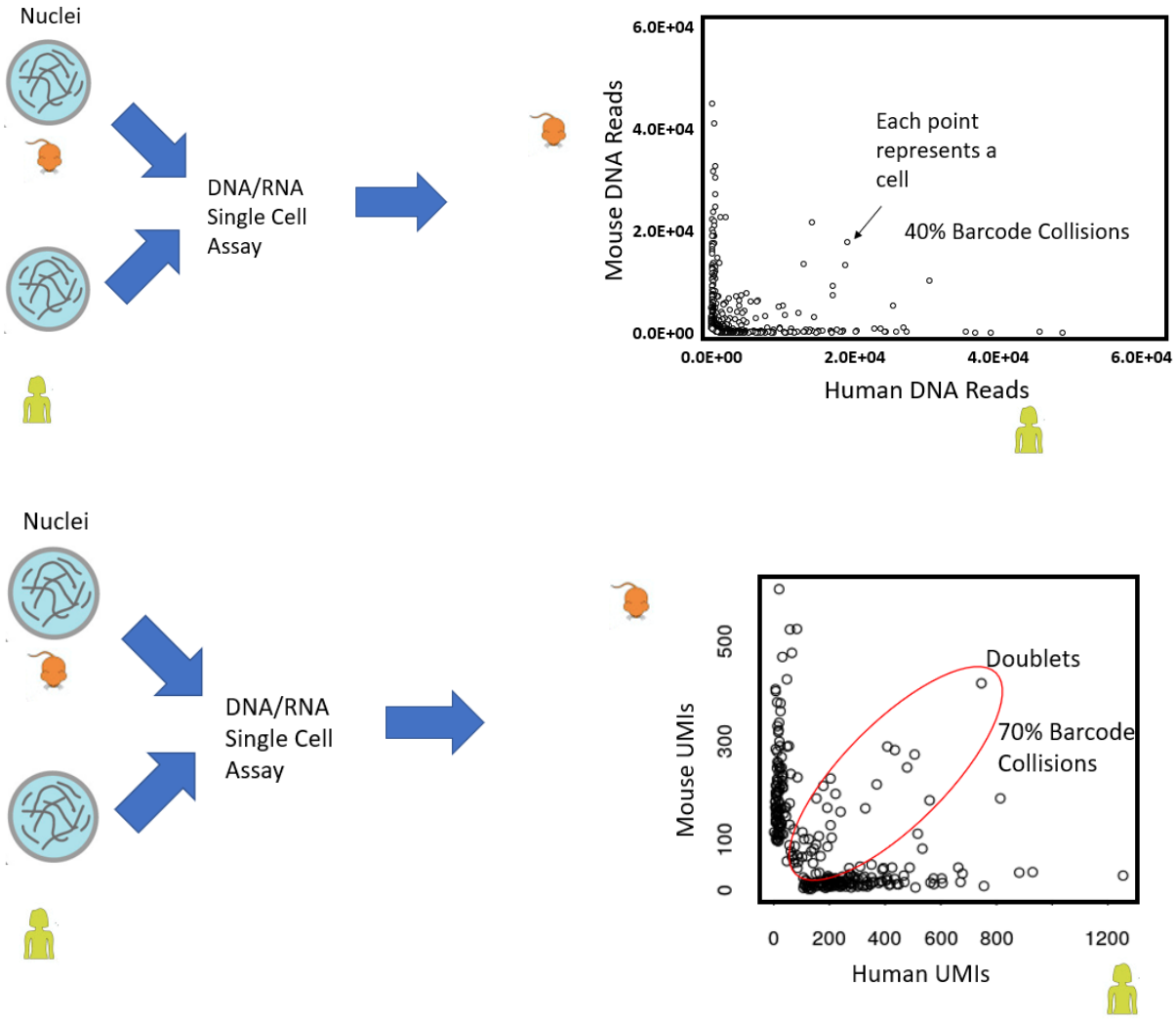


Figure 14: Quantifying the number of human and mouse reads for each barcode for both DNA and cDNA libraries and identifying barcode collision rates.

1.4 Conclusion

Here, we describe the development of an RNA and DNA co-sequencing platform using polyacrylamide gel beads as the combinatorial indexing container. Acrydite modified reverse transcription primers were used as the cDNA immobilizing scheme while DNA was immobilized by the polyacrylamide mesh. We arrived at this final design by screening a variety of nucleic acid containers. The most straightforward approach was to leverage the nucleosome depleted nuclei, but this approach was unreliable due to the low structural integrity of these nuclei. To increase the structural integrity of the nucleic acid container, we tried a hydrogel encapsulation approach. We first used agarose but discovered that cDNA easily diffused out of the gel bead. We then lowered the pore size using a PEG acrylate gel, but still discovered that a noticeable amount of nucleic acid products in the size range of cDNA was lost. This led to the insight of using covalent anchoring of the cDNA using acrydite modified reverse transcription primers and a polyacrylamide gel bead vessel. We then demonstrated the potential of this platform by designing a combinatorial indexing scheme adapted from previous work to co-sequencing DNA and RNA libraries. Unfortunately, the barcode collision rates from this experiment were high likely due to the inherent inaccuracies of estimating the input of 10-20 nuclei into the second 96 well plate during combinatorial indexing. Unlike previously combinatorial indexing methods, the gel beads are too large to be sorted. Thus, some wells in the second plate may contain multitudes higher or lower numbers of nuclei causing higher than expected barcode collisions. Future work could include using a FACS machine with custom settings to account for the additional size of the gel beads or the innovation of a third level of combinatorial indexing.

This powerful platform has the potential to assess copy number variations and RNA from the same cell or nuclei. This may be particularly relevant in the study of high-risk

neuroblastomas where copy number increase of the MYCN oncogene on chromosome 2p occurs in 20% of them.(Dzieran et al. 2018) This MYCN copy number variation typically results in poor prognosis(Dzieran et al. 2018). The single cell gDNA sequencing of neuroblastoma tumors could bioinformatically isolate MYCN copy number amplified tumor cells and profile. The whole transcriptomes of these MYCN amplified tumor cells could then be profiled to potentially identify therapeutic pathways to specifically target MYCN amplified tumor cells.

Portions of Chapters 1 are in part are a reprint of material in submission as it appears in "Ultra-High Throughput Single Cell Co-Sequencing of DNA Methylation and RNA using 3-Level Combinatorial Indexing" The dissertation author was the primary author of this paper along with Andrew Richards and Kun Zhang.

CHAPTER 2: SINGLE CELL METHYLATION SEQUENCING AND RNA INTEGRATION

2.1 Abstract

DNA methylation is sequenced using bisulfite conversion chemistries. Bisulfite conversion presents a few classic challenges that we adapt to our gel bead platform. Firstly, unmethylated cytosines convert into uracil which is sequenced as thymine. Because our combinatorial indexing barcodes are unmethylated, we designed a scheme to methylate them during the gap-filling process of the Tn5 reaction. This methylation of the barcodes is required to prevent barcode collisions that are the result of bisulfite conversion. Secondly, the bisulfite conversion fragments the majority of DNA. In addition, the DNA needs to be extracted post bisulfite conversion with PCR. We designed a bisulfite conversion technique that efficiently carries the beads through each of the steps, and we perform a linear amplification PCR to extract the DNA from the gel beads post bisulfite conversion using an uracil tolerant polymerase. Finally, we explored methods to add an adapter to the 3' end of the DNA fragments to exponentially amplify the DNA and complete library construction. The best method was to use the commercial adaptase protocol. We then validate the integrity of our library by demonstrating high alignment rates 60-70% and recovering expected methylation dynamics over genomic features.

2.2 Introduction

Here, we describe the development of a novel single cell methylation library construction protocol that was designed specifically to tackle the challenges of performing bisulfite conversion in polyacrylamide gel beads. The gold standard method to perform single cell

methylation sequencing employs harsh bisulfite conversion chemistries. There are a few main challenges in developing my protocol around these chemistries. Firstly, bisulfite conversion converts unmethylated cytosine to thymine which results in the cytosines in the unique molecule identifier (UMI) incorporated in the reverse transcription capture primer, required for single cell RNA sequencing, to also be converted to thymine. The cytosines in the Tn5 adapter sequences are also converted resulting in a lowering of the PCR primer annealing temperatures which causes extensive off-target PCR products. Secondly, bisulfite conversion produces extensive DNA fragmentation.(Ahn et al. 2021) For the cDNA library, fragmentations result in the complete loss of the molecule because one end contains the cell barcode while the other end contains the UMI. Because Tn5 inserts in two ends of the DNA library, fragmentations result in the loss of one of the adapters which prevents the addition of Illumina sequencing adapters during PCR. Thirdly, most of the DNA is still contained inside the polyacrylamide beads during the bisulfite conversion process. Typically, DNA is eluted from either a silica column or magnetic bead once bisulfite conversion is completed. Because the DNA hasn't been extracted yet, I needed to design a method that ensure that the gel beads are also moved to the steps beyond the bisulfite reaction.

I first address the methods to protect the Tn5 PCR adapter sequence and the RNA UMI. To complete the Tn5 reaction, the Tn5 must be denatured with 0.1% SDS.(Picelli, Björklund, et al. 2014) After denaturation, the DNA is fragmented into double stranded products with a 5' overhang. This complementary sequence of the overhang can be synthesized with a high fidelity polymerase that is resistant to SDS by extending the recessed 3' end using the 5' overhang as the template strand. This process is called gap filling. To protect the Tn5 adapter sequence from the cytosine to thymine conversion, I created a custom dNTP mixture where the cytosine is replaced

with methylated cytosine. Thus, the newly synthesized DNA from the recessed 3' end through the Tn5 adapter contains methylated cytosine. These methylated cytosines are not converted during bisulfite conversion, retaining the original Tn5 adapter sequence for PCR. To protect the cDNA UMI, I linearly amplify the cDNA using a single PCR primer that hybridizes to the reverse transcription capture primer using the same PCR reaction mix to perform gap filling. This process incorporates methylated cytosine to the newly synthesized cDNA products which protects the whole cDNA strand including the UMI from the cytosine to thymine conversion.

To address the DNA fragmentation issue, I try to optimize the cDNA linear amplification prior to bisulfite conversion so that the subset of cDNA that remains could potentially still reflect the original cDNA library complexity. The gDNA library cannot be similarly amplified prior to bisulfite conversion as the original methylated cytosine profile would be altered. Thus, I explored different post bisulfite adapter tagging methods like scnmC-seq to add an adapter sequence to the 3' end of all the DNA sequences post bisulfite conversion to enable the final PCR required to add Illumina sequencing adapters(Callaway et al. 2021). However, I had to make crucial modifications to allow me to extract the DNA form the gel beads prior to the single-end ligation reaction. In the previous chapter, I mentioned that the method of DNA extraction from the polyacrylamide beads requires a combination of PCR and passive diffusion of DNA products from the gel bead. Thus, I designed a linear amplification PCR step after bisulfite conversion using the protected Tn5 adapter sequence as the priming sequence and uracil tolerant polymerase to extract the DNA from the gel bead. This PCR is distinctive to this method because the template is gel beads coated in the magnetic beads used in the Zymo EZ-96 DNA Methylation MagPrep kit. This modification is required for high DNA library complexity as most DNA sequences are still trapped inside the gel bead.

To test the success of this method, we first spike in lambda phage DNA to ensure that the bisulfite conversion efficiency was 99%. The library was then sequenced to shallow depths to assess the mapping rate to in-silico bisulfite converted genomes. After identifying the best mapping software and settings, we binned the methylation data around reference methylation features to validate the methylation dynamics expected around those features.

2.3 Methods and Results

2.3.1 Adapting Post Bisulfite Conversion PCR Adapter Addition Techniques to sci-Gel

The cytosine to thymine conversion and fragmentation of DNA during bisulfite conversion poses significant library design challenges. Figure 15 illustrates the common WGBS library construction methods. To circumvent the cytosine to thymine conversion of library adapter sequences, conventional bisulfite sequencing involves the addition of methylated adapters. Methylated adapters are typically much more expensive than unmethylated ones. In addition, fragmented sequences resulting from the bisulfite conversion are unrecoverable. The highest library complexity bisulfite sequencing methods involve the addition of adapters post bisulfite conversion which typically involves random priming. At the single cell level, the most effective method was demonstrated in scnmC-seq which first involves cell lysis and bisulfite conversion. Then an initial random priming and extension step like the TruSeq method is performed to synthesize a complementary strand of DNA using the uracil resistant and strand siplacing polymerase, klenow exo-. The strand synthesized by the random primer is then tagged on the 3’ end with an adapter using the adaptase protocol. Illumina sequencing primers are then added to this product using PCR primers complementary to the random primer PCR handle and adaptase adapter. (Luo et al. 2018)

sci-MET takes a slightly different approach. After bisulfite conversion, a random priming and extension step like scnmC-seq is also used. However, this random priming is performed three additional times to increase library complexity. The Illumina sequencing adapters PCR uses primers reverse complementary to the Tn5 adapter and the random priming sequence PCR adapter. The Tn5 adapter sequence is designed to be cytosine depleted and is therefore unchanged through the bisulfite conversion.

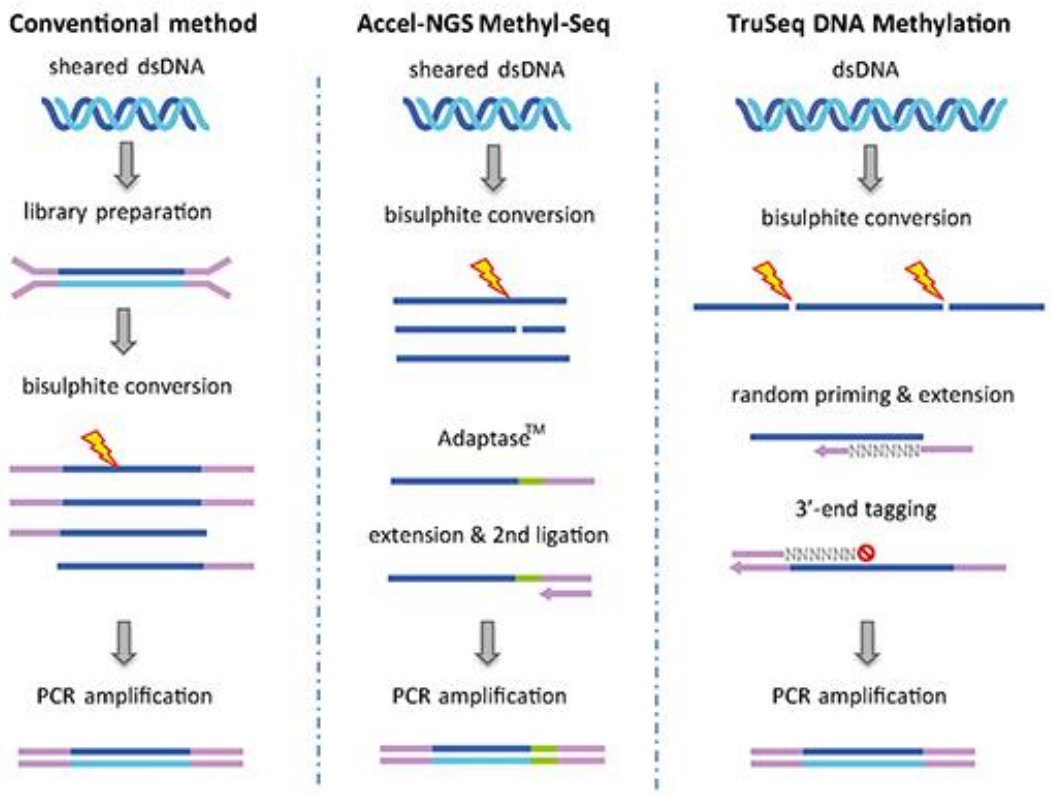


Figure 15: A graphical depiction of common bisulfite conversion techniques

Our technology requires a different approach. Figure 16 illustrates the cDNA library structure prior to bisulfite conversion. Transcriptome sequencing requires the use of UMIs that can clearly distinguish between PCR duplicates and natural gene expression. The design of the UMI is a random sequence of all bases. However, the bisulfite conversion would mutate the UMI

by converting the unmethylated cytosine to thymine. Therefore, we needed to linearly amplify the cDNA with methylated cytosines prior to bisulfite conversion to protect the UMI sequence using a PCR primer that is reverse complement to the reverse transcription primer with a cytosine depleted handle. Post bisulfite conversion, we also needed to design a non-random priming technique since random priming of the cDNA would likely not contain the UMI sequence.

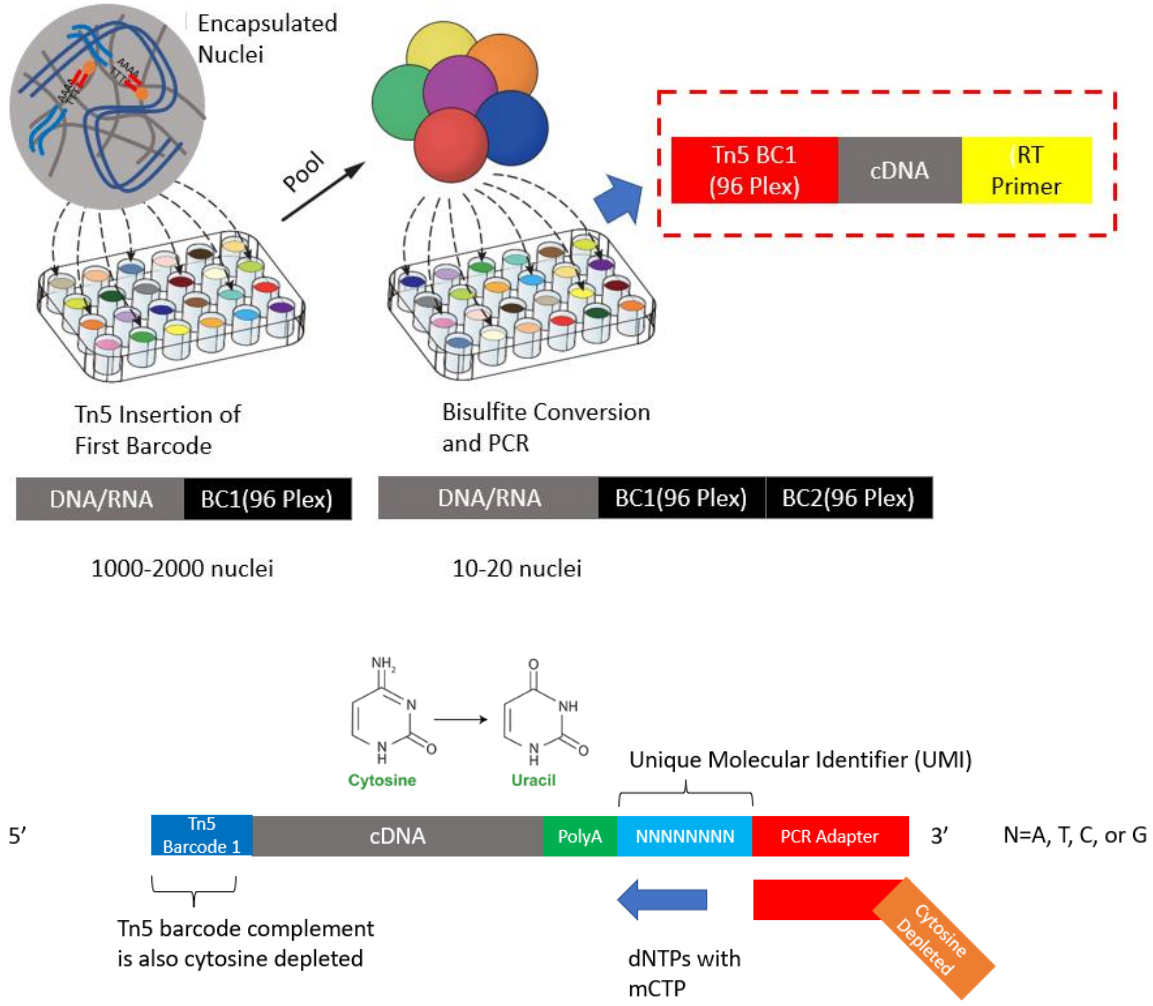


Figure 16: cDNA library post Tn5 insertion and linear amplification with methylated cytosines scheme

The second problem with a random priming protocol is that the gel beads are still intact post bisulfite conversion. Figure 17 shows that the polyacrylamide gel beads coated in the magnetic beads that are used during the bisulfite conversion step. As discussed previously, the DNA needs to be sufficiently amplified to extract the DNA from the gel beads. We designed a post bisulfite linear amplification scheme where the transposon sequence is first gap filled with methylated cytosines instead of unmethylated cytosines. Instead of eluting the DNA from the magnetic beads per the manufacturer's protocol, the magnetic beads containing intact gel beads are transferred to the linear amplification reaction with PCR primers reverse complement to the gap filled transposon sequence that was protected from bisulfite conversion. Figure 18 illustrates this linear amplification process. In the most optimized versions of this protocol, the DNA is linearly amplified for 20 cycles with barcoded primers containing the second cell barcode to complete the combinatorial indexing process and sufficiently extract the DNA from the gel beads. The library is then split where the cDNA is exponentially amplified with PCR primers reverse complement to the cytosine depleted PCR adapter on the reverse transcription primer side of the library and the transposon sequence.

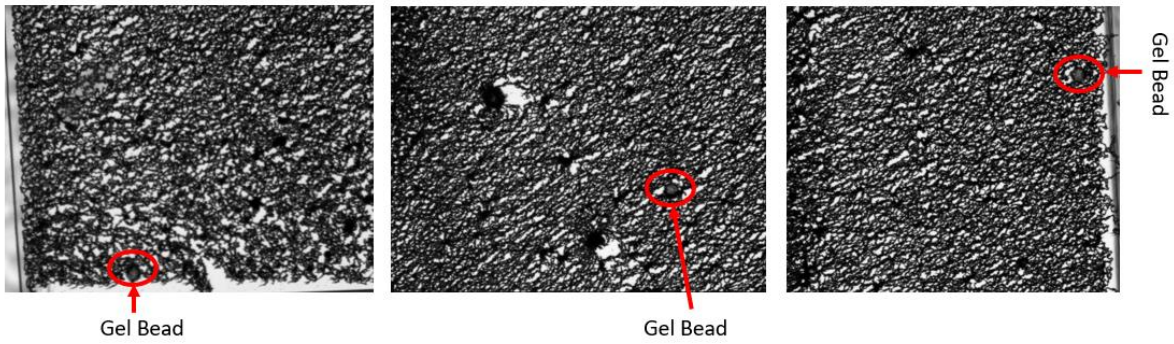


Figure 17: Post bisulfite conversion gel beads coated with magnetic beads

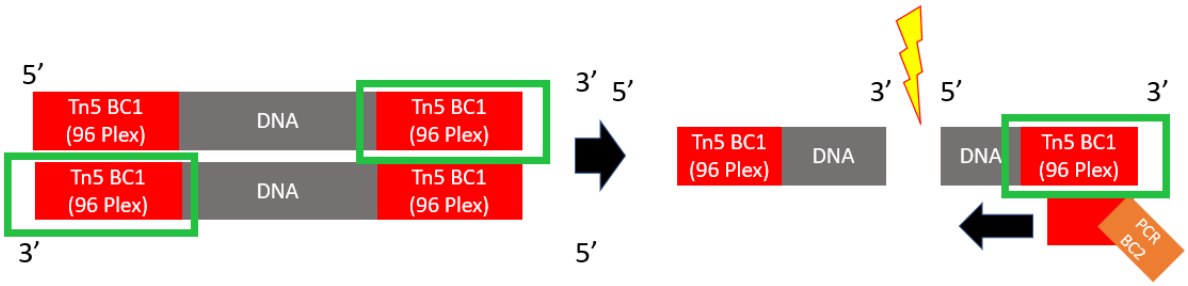


Figure 18: Gap filling and linear amplification scheme post bisulfite conversion

Unfortunately, we lost more cDNA during the bisulfite conversion process than expected. Figure 19 shows how we used qPCR to estimate the loss of cDNA due to fragmentation during bisulfite conversion. We originally hoped that the linear amplification of cDNA prior to bisulfite conversion could compensate for the loss of cDNA due to fragmentation. However, approximately 99% of the cDNA was lost. As a result, we decided to explore exponentially amplifying the cDNA prior to bisulfite conversion or splitting the cDNA and gDNA libraries prior to bisulfite conversion as potential solutions in chapter 3.

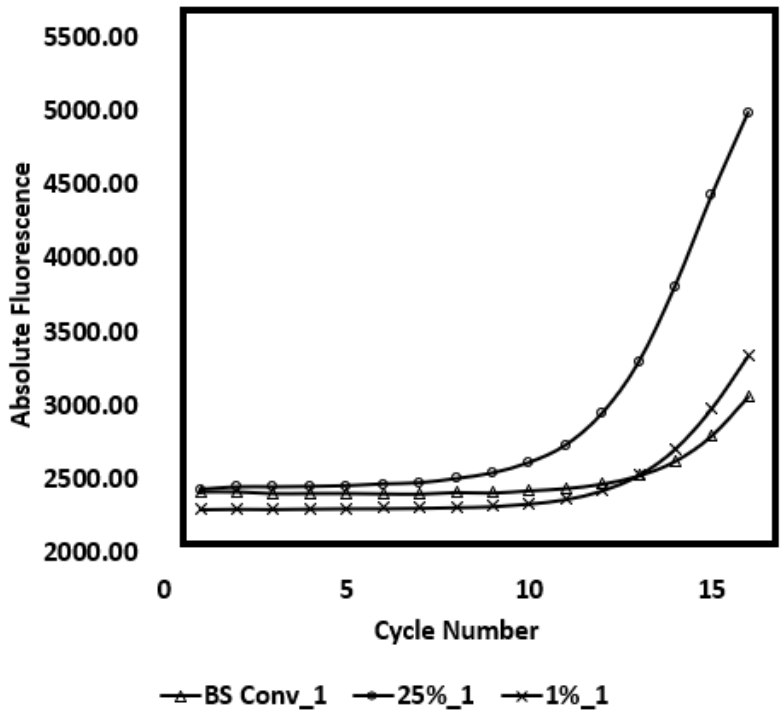


Figure 19: Loss 99% of cDNA during bisulfite conversion

For the DNA library, we tried two different post bisulfite adapter tagging methods to attempt to save costs since the adaptase reaction has an expensive cost of about $20 per reaction. Inspired by a single end ligation design utilizing a modified oligo with 5rapp, we assessed the ligation efficiency between this design and the adaptase reaction.(Wu and Lambowitz 2017) However, the adaptase kit demonstrated substantially higher ligation efficiency as shown in Figure 20. Therefore, we decided that using the adaptase kit was the optimal strategy to generate the methylome libraries as described in Figure 21. After the ligation of our DNA libraries with the adaptase adapter, we performed final PCR to add Illumina sequencing adapters.

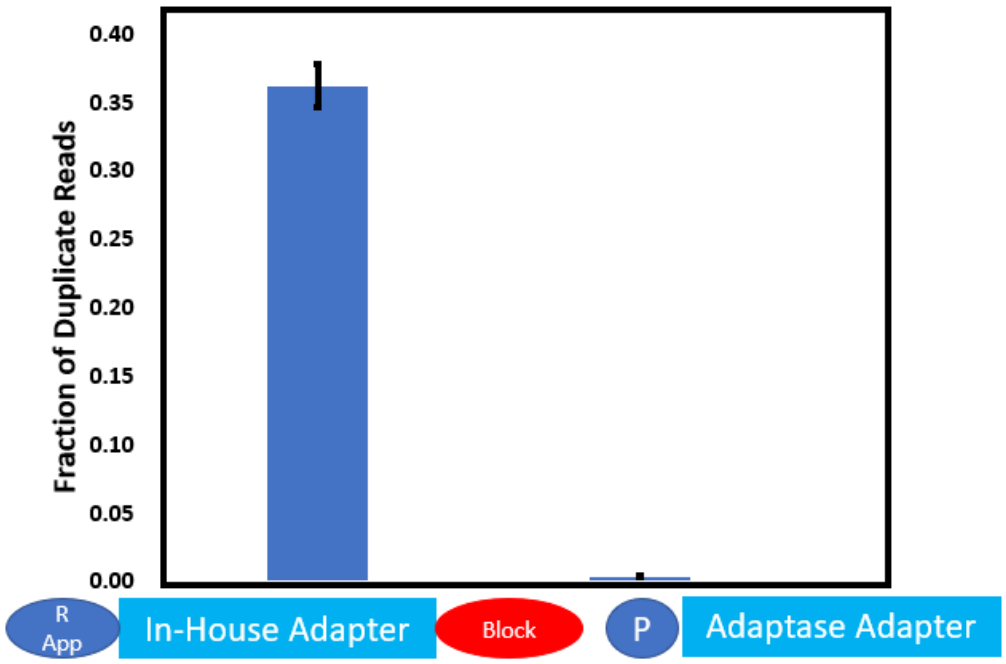


Figure 20: Comparisons between 5rapp ligation and adaptase single end ligation

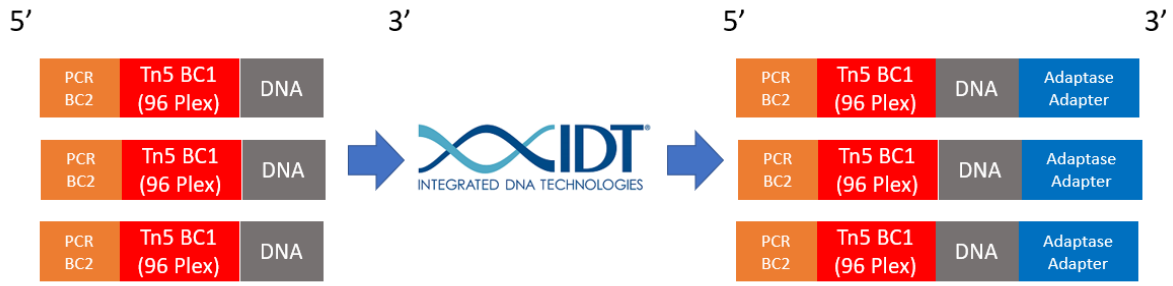


Figure 21: Single end ligation of post bisulfite linearly amplified DNA with adaptase.

2.3.2 sciGel Version 2: Single Cell Methylome Library Formation Protocol

To validate the success of our WGBS method and assess the performance of our assay, we performed single cell WGBS on a colorectal cancer cell line HCT116 and a human kidney tissue. In summary, nuclei undergo reverse transcription, were then encapsulated, 100-200 were split into a 96 well plate and barcoded with Tn5, then 10-20 encapsulated nuclei were split into a second 96 well plate following the same barcoding scheme as described in the whole genome and whole transcriptome co-sequencing assay described in chapter 1. The adaptation for methylome sequencing has a few adaptations. PCR reaction mixture was modified substituting cytosine for methylated cytosine. A cytosine depleted cDNA primer reverse complement to the reverse transcription primer is added. Gap filling takes place as previously followed by 10 cycles of cDNA linear amplification. Bisulfite conversion reagent is then added to each well according to the manufacturer's protocol. The samples are then incubated at 98 C for 8 minutes and 65C for 3.5 hours and then kept at 4C overnight following the standard bisulfite conversion protocol by the manufacturer. Magnetic beads and binding buffer were then added to the bisulfite conversion mixture and transferred to a deep well 96 well plate. The manufacturer's protocol was then followed through the desulphonation step with a modification. After drying the magnetic beads for 25 minutes, 20 μL was added to each well and then incubate at 55C for 5 minutes. Instead of placing the deep well plate back on the magnetic rack, the sample including the magnetic beads were transferred to a 96 well plate. KAPA HiFi Uracil PCR master mix was then added each reaction along with a well specific PCR barcoded primer that is reverse complement to the Tn5 adapter. Linear amplification of the DNA and cDNA then occurs in the presence of the magnetic beads for 20 cycles. Afterwards, rSAP was used to remove the phosphates from residual mosaic end sequences. The samples were then bead purified with SPRI beads at a 1.2X ratio and eluted

into a new 96 well plate. Half of the volume was transferred to a new 96 well plate where KAPA HiFi was used to finish amplifying the cDNA library with PCR primers reverse complement to the cytosine depleted cDNA adapter on the reverse transcription side of the library and Illumina P5 sequences. The DNA half of the library was then incubated at 98C for 3 minutes quickly followed by incubation on ice for 2 minutes to ensure single stranding of the library. The manufacturer's protocol for the adaptase reaction was then performed. After heat inactivation of the adaptase enzymes, KAPA HiFi was used to finish amplifying the DNA library with PCR primers reverse complementary to the adaptase adapter and the Illumina P5 sequences. All the 96 wells for the DNA and RNA plates were pooled and then bead purified with SPRI beads at a 0.8X ratio twice to prepare the library for sequencing. PAGE was performed to check the quality of the library with an expected smear between 200-600 bases. The libraries were sequenced with a MiSeq. The detailed version of this protocol and sequencing scheme is in the supplementary methods. Although the cDNA library was created, we decided not to sequence it due to the low library complexity due to the loss of cDNA from bisulfite conversion as discussed previously.

Table 1 shows that we had strong alignment rates for our bisulfite converted libraries and 99% bisulfite conversion efficiency which is comparable to existing methods. To assess the biological relevance of our technology, we pooled the HCT116 methylome data and binned it across the genomic coordinates of HCT116 H3K4Me3 histone marks based on reference ChIP-seq data. This histone mark is typically hypomethylated and is nearby highly expressed genes. (Sharifi-Zarchi et al. 2017)Figure 22 shows the expected hypomethylation dynamics associated with this feature among others. This validated the integrity of our novel WGBS protocol.

Table 1: Alignment rates assessed by different library preparation conditions and alignment software. The bisulfite conversion efficiency based on a lambda phage DNA spike-in construct.

Tn5 Concentration	Alignment Method	Alignment Rate	Bisulfite Conversion Efficiency
0.1mg/mL	BWA	0.79	99%
0.05mg/mL	BWA	0.68	99%
0.1mg/mL	Bismark PBAT	0.7	99%
0.05mg/mL	Bismark PBAT	0.72	99%

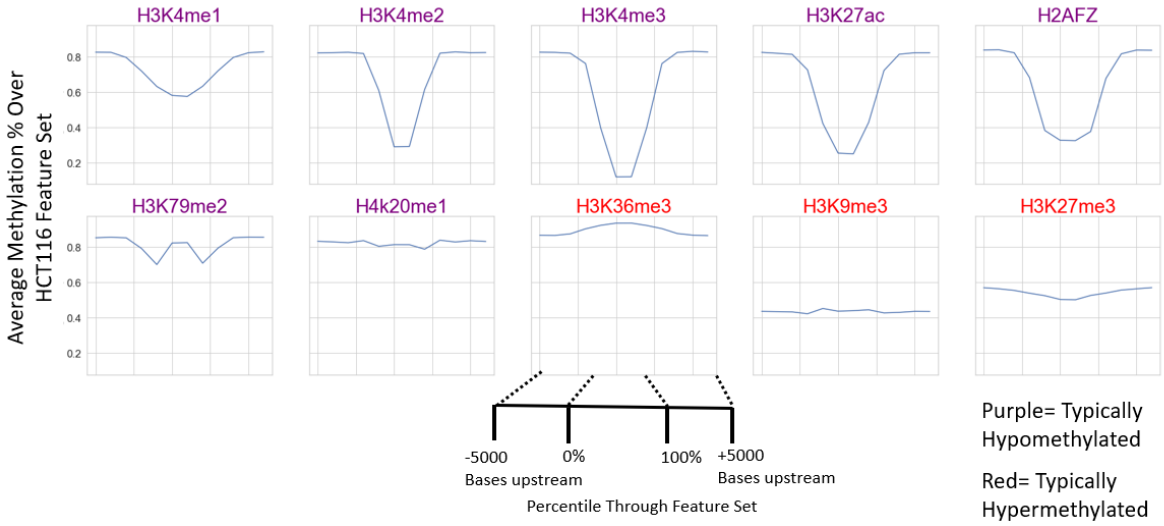


Figure 22: Average methylation over HCT116 feature sets typically hypomethylated (purple) or hypermethylated (red)

We next assessed the library performance at the single cell level for both the HCT116 and kidney tissue libraries. Table 2 summarizes the alignment rates for both libraries showing high alignment rates using the Bismark bowtie2 based method. The CH Methylation level was also low which is expected for terminally differentiated tissue.

Table 2: Sequencing statistics of kidney and cell mixture libraries

Library	Alignment Rate	% CH Methylation
Single Cell Methylome Kidney Library (averaged over 2 libraries)	63%	1.7%
Single Cell Methylome Cell-Line Library (averaged over 2 libraries)	62%	1.5%
Unmethylated DNA Spike-In	-	1%

We plotted the number of unique reads against the fraction of unique reads to identify the cells from empty barcodes shown in Figure 23. Roughly 90% of the barcodes are empty in combinatorial indexing schemes. Thus, barcodes containing single cells can be discriminated by

visually identifying a subset of barcodes with high library complexity. (Mulqueen et al. 2018; 2021)

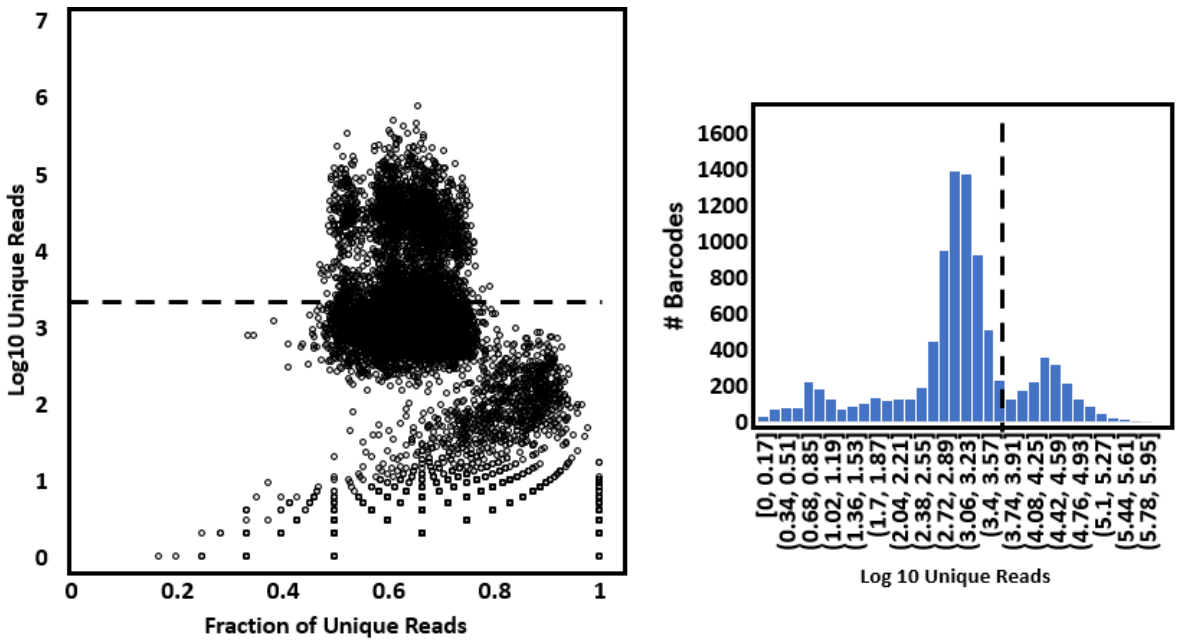


Figure 23: Library complexity analysis of single cell WGBS kidney libraries. Dotted lines indicate read cut-offs separating empty barcodes from occupied ones.

We selected cells with over 100,000 reads for further downstream analysis. Using pre-seq, we estimated the maximal library complexity of our library to 1-2M reads per cell. Table 3 summarizes how our WGBS library compares to existing methods. Currently, our method exhibits similar performance compared to existing single cell high throughput methods. With further optimizations, we believe that we can vastly improve the library complexity to approach the library complexity of single cell per well methods.

Table 3: Comparison of our WGBS gel method (sci-Gel) compared to existing methods. We used pre-seq to estimate the maximal library complexity of our libraries.

Method	Mapping Rate	Duplication Rate	Sequencing Depth/Cell	Number of Unique Reads/Cell
PBAT (Hui et al)	37%	26%	1-2M	-
scM&T (Angermueller et al)	18%	18%	11.5M	3M
SnmC (Luo et al)	50%	48%	5.5M	1.3M
sc-BS (Farlik et al)	11% (Usable Reads)		6M	673K
Sci-MET (Mulqueen et al)	67%	66%	~200k	~1-2M (estimated w/ Pre-seq)
Sci-Gel-M (Single Cell Library)	~60%	30%	~100k	~1-2M (estimated w/ Pre-seq)

Using the cells with over 100,000 reads/cell we binned the reads into 1 megabase windows. This large window size reflects the sparsity of our WGBS datasets. Figure 24 shows that there are roughly 200 CH positions in each bin per cell. Surprisingly, the number of CG positions is roughly 10 fold less. Since CH methylation is very low in kidney tissue, we concluded that this coverage was too sparse for us to confidently perform additional analysis on this dataset. Attempting to perform single cell clustering using kidney tissue is likely impractical without RNA information.

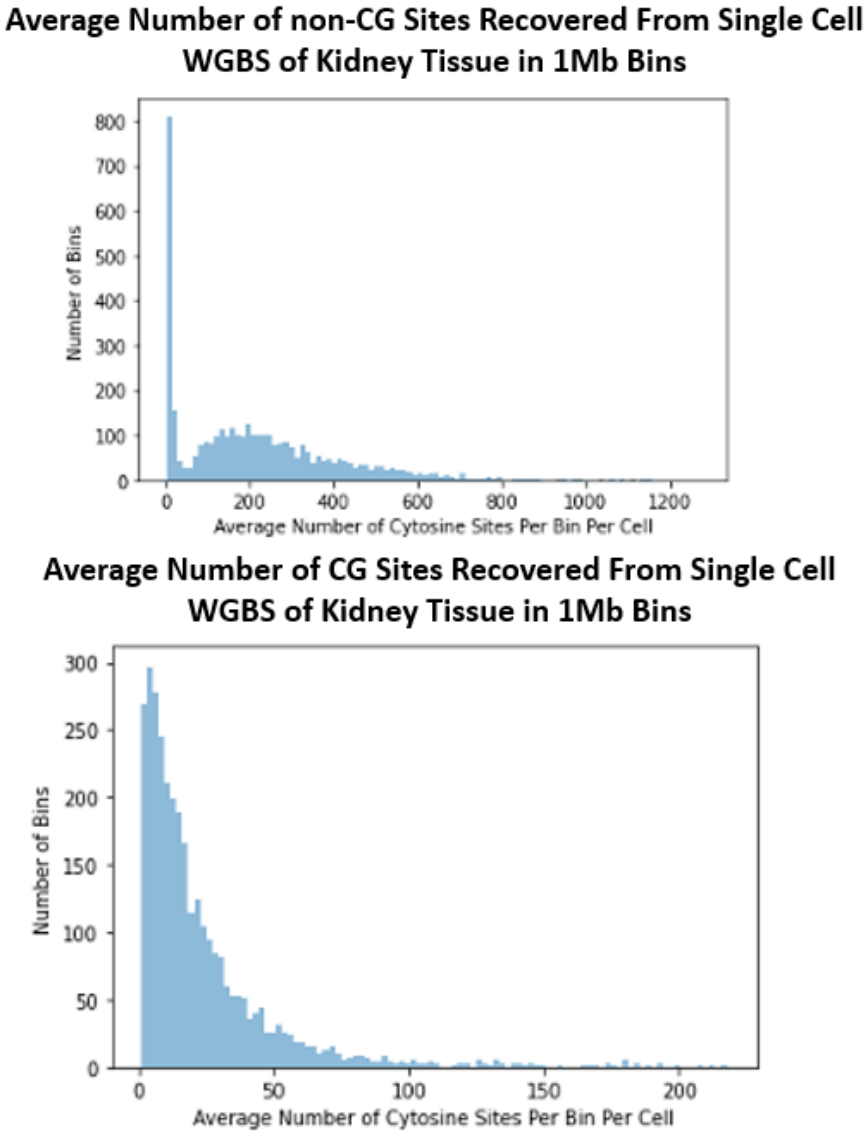


Figure 24: Average CH and CpG positions in each bin per single cell of our kidney WGBS experiment.

2.3.2 Accompanying Bioinformatic Methods

These additional analyses were built on top of the previously described bioinformatic methods in chapter 2. Briefly after demultiplexing into individual barcodes, sequencing reads were aligned with either Bismark (a bowtie2 wrapper for bisulfite sequencing alignment) or BS-Bolt (a bwa-mem wrapper for bisulfite sequencing alignment). A lambda phage DNA construct with cytosine depleted PCR adapters spike in prior to bisulfite conversion was first examined to

ensure high bisulfite conversion efficiency using sanger sequencing. Since this phage DNA is unmethylated, we expected and confirmed that 99% of the cytosines were bisulfite converted. After alignment, the CpG positions were extracted from the aligned reads and the CpG positions were binned based on genomic features such as H3K4Me3 histone marks for methylation dynamics validation. CpG positions can be extracted using either methylpy or the Bsbolt extraction method. The methylation frequency was then calculated as defined as the number of methylated CpG sites divided by the total number of CpG sites recorded in that window. The methylation frequency was then plotted across the features of interest. The detailed version of this protocol can be found in the supplementary methods.

2.4 Conclusion

Here, we developed a new single cell WGBS sequencing method specific for our protocol. We first used methylated dCTPs in the gap filling step to protect the Tn5 adapter and cell barcode sequences from bisulfite conversion. In addition, we included a linear amplification step as an attempt to recover the subset of unfragmented cDNA post bisulfite conversion. However, the yield of cDNA post bisulfite conversion was less than 1%. We concluded that the cDNA library must be split from the DNA library or exponentially amplified prior to bisulfite conversion. We then developed a non-random priming post bisulfite conversion sequencing method to efficiently extract the DNA from the gel beads via PCR and diffusion. We then identified the best post bisulfite adapter tagging method inspired by scnmC-seq which led to the creation of high complexity WGBS libraries. To test the performance of our method, we generated two single cell datasets: an HCT116 cancer cell line dataset and a human kidney dataset. We validated high bisulfite conversion efficiencies using a lambda phage DNA spike in prior to bisulfite conversion and sanger sequencing. Our analysis of the methylation levels over

HCT116 H3K4Me3 histone marks recapitulated hypomethylation dynamics found in other studies.(Sharifi-Zarchi et al. 2017) Furthermore, the kidney pilot study demonstrates the immense difficulty in performing single cell methylation analysis in terminally differentiated tissue. The paucity of CpG sites recovered at the sequencing cost per cell in this study prevents the discrimination of single cells. Since the number of recovered CpG sites is roughly 10 fold less than the number of CH sites. We conclude that this low signal at high sequencing cost necessitates the need for RNA co-sequencing to assist in single cell clustering and cell type calling of terminally differentiated tissues.

Further optimizations at the single cell level of this method will be described in chapter 3. Future work could include separately optimizing this method for bulk WGBS. The higher alignment rates found in Tn5 based combinatorial indexing WGBS methods compared to single cell per well methods as shown in Figure 23 could be because of the efficient Tn5 insertion speculated previously.(Mulqueen et al. 2018) Our method could be a viable alternative to existing bulk WGBS methods.

Portions of Chapters 2 are in part are a reprint of material in submission as it appears in "Ultra-High Throughput Single Cell Co-Sequencing of DNA Methylation and RNA using 3-Level Combinatorial Indexing" The dissertation author was the primary author of this paper along with Andrew Richards and Kun Zhang.

CHAPTER 3: 3-LEVEL INDEXING DEVELOPMENT AND RNA INTEGRATION PART 2

3.1 Abstract

We developed an additional level of combinatorial indexing to lower the barcode collision rates and increase the cell throughput of our assay from 500-1000 cells per experiment to 50,000-100,000 cells per experiment using three 96 well plates. After insertion of the first barcode by Tn5 tagmentation, the transposon overhang presents a target where the ligation of a second barcode can take place. We explored a few ligation approaches and identified the best approach using T7 ligase with a short splint oligo. In parallel, we tackled two major problems with our RNA library construction method: the loss of the library after bisulfite conversion and the high barcode collision rates despite low barcode collision rates in the DNA library. We discovered several pitfalls that are particular to our gel bead system in adapting previous single cell RNA sequencing methods. Thus, we identified an RNA library construction method that is performed completely within the gel bead which resulted in a method with doublet rates less than 10%, like the double rates in the DNA library. We then validate the integrity of our RNA library by recovering the same marker genes for our cell type test samples as is described in previous works. Finally, we describe our on-going efforts to reduce assay variability and discovered key trade-offs. Ultimately, we arrived at the only known strategy to perform DNA methylation and RNA co-sequencing at the scale of 50,000-100,000 cells per experiment.

3.2 Introduction

Despite evidence of single cell resolution through our 2-level combinatorial indexing approach, the number of barcode collisions was inconsistently high. Published combinatorial

indexing protocols typically have barcode collision rates less than 10% while our method had barcode collision rates between 15-40%.(Mulqueen et al. 2021; 2018) The barcode collision rates are typically estimated by performing a human/mouse cell mixture experiment where equal numbers of human and mouse cells are mixed prior to the experiment. After sequencing, a mixed barcodes are identified as any barcode combination that contains at most 80% of reads from one species. The collision rate is then estimated as two times the mixed barcodes rate as doublets from the same species are not observed. We identified two sources of barcode collisions: doublets that arise during the encapsulation process where two are more cells are captured by the same bead and two or more cells that have the same barcoding path. The latter factor is typically controlled by single cell sorting.(Mulqueen et al. 2018; 2021) Because the gel beads are too large to be cell sorted by a typical FACS machine, estimating the concentration of cells prior to plating is required which leads to inherent inaccuracies that could cause barcode collisions. As a result, we developed a third layer of combinatorial indexing to scale the barcode space 100X and increase the tolerance of inaccurate cell number plating using these gel beads at dilute concentration. The increase in barcoding space has additional benefits. It expands the throughput of our technology 100X, vastly decreasing the number of experiments needed to characterize human tissues.

The co-sequencing of the transcriptome and the methylome is complicated by the bisulfite conversion process. Generally, mutli-omic technologies have tackled this problem in two ways: separating the cDNA from the gDNA prior to bisulfite conversion or exponentially amplifying the cDNA with dmCTPs prior to bisulfite conversion. In scNMT-seq, single cells the RNA is separated from the gDNA with reverse transcription primers annealed to magnetic beads.(Clark et al. 2018) In snmCAT-seq, full length cDNA is exponentially amplified with

dmCTPs prior to bisulfite conversion. The cDNA is discriminated form the gDNA library after sequencing as the cDNA library is highly methylated compared to the DNA library. Previously, we attempted to linearly amplify the cDNA with dmCTPs prior to bisulfite conversion. This resulted in an extremely low yield of cDNA libraries post bisulfite conversion. Here, we explore an exponential cDNA amplification method prior to bisulfite conversion like the snmCAT-seq design by designing a combinatorial barcoding approach without Tn5. However, the cDNA was too long to efficiently diffuse out of the gel bead. As a result, we had to split the cDNA prior to bisulfite conversion. Here, we will explore the solutions explored arrive at this conclusion.

By combing two solutions: the splitting of cDNA prior to bisulfite conversion and drastically increasing the combinatorial barcoding space, we arrived at the first workable solution where the transcriptome and methylome are co-sequenced with doublet rate less than 10%. In addition, we optimize each enzymatic reaction in to increase the library complexity of this workable solution over 100X. This solution utilizes an unstable encapsulation and gel formation process. Thus, we will also explore a new solution that increases the consistency of the encapsulation process and subsequent library formation.

3.3 Methods and Results

3.3.1 The Development of 3-Level Combinatorial Indexing

The cutting edge of combinatorial indexing technology development utilizes three or more levels of combinatorial indexing. This development crucially removes the need for cell or nuclei sorting to control barcode collision rates. There are three general methods for three-level combinatorial barcoding that have been demonstrated in single cell DNA accessibility and RNA technologies: 1) The use of Tn5 to insert the first barcode, ligation to the Tn5 overhang to add the second barcode, and PCR to add the third barcode. 2) The use barcoded reverse transcription

primers to add the first barcode, ligation to the reverse transcription primer overhang to add the second barcode, and PCR to add the third barcode. 3) The use of barcoded reverse transcription primers, linear polymerase-based extension to add the second barcode, and PCR to add the third barcode.

3.3.1.1 Tagmentation Based 3-Level Indexing for WGBS

Three-level indexing using Tn5 based DNA accessibility sequencing or ATAC sequencing are at the cutting edge of combinatorial indexing technology. ATAC/RNA co-sequencing methods take advantage of the Tn5 overhanging sequences during Tn5 insertion to allow for a ligation of an additional barcoded adapter, increasing the combinatorial indexing level illustrated in the Figure 25. (C. Zhu et al. 2019; Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnel, et al. 2020; Plongthongkum et al. 2021)

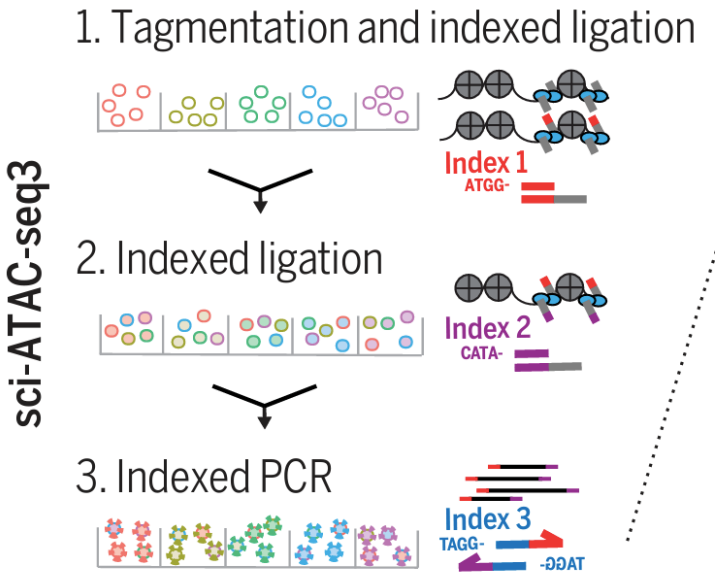


Figure 25: 3-Level sci-ATAC Combinatorial Indexing Scheme

There are two major designs: the 3-level sci-ATAC design or the SPLiT-Seq design which is employed in methods such as SNARE-Seq2 and PAIRED-Seq. In the 3-level sci-ATAC design, T7 ligase and a 15 bp synthetic splint oligo with a 3' blocking modification to prevent extension by polymerases is used to ligate an adapter containing the second cell barcode. In the SPLiT-Seq design, T4 ligase and a 39 bp synthetic splint oligo is used to ligate an adapter containing the second cell specific barcode. Because SNARE-Seq2 was developed by our lab, I decided to first try the SPLiT-Seq design. In summary, Tn5 is first used to insert the first cell barcode in the gel beads. Afterwards, T4 ligase was used to ligate the second cell barcode followed by PCR to add the third barcode using our gel bead platform. Figure 26 showcases the preliminary success of our T4 ligation in-gel design using the SNARE-Seq2 adapters. Our qPCR results show that the ligation was efficient as similar amplification dynamics between ligated and unligated templates were observed. The PAGE also shows the shift in size owing to the ligation of adapters to the transposon overhang.

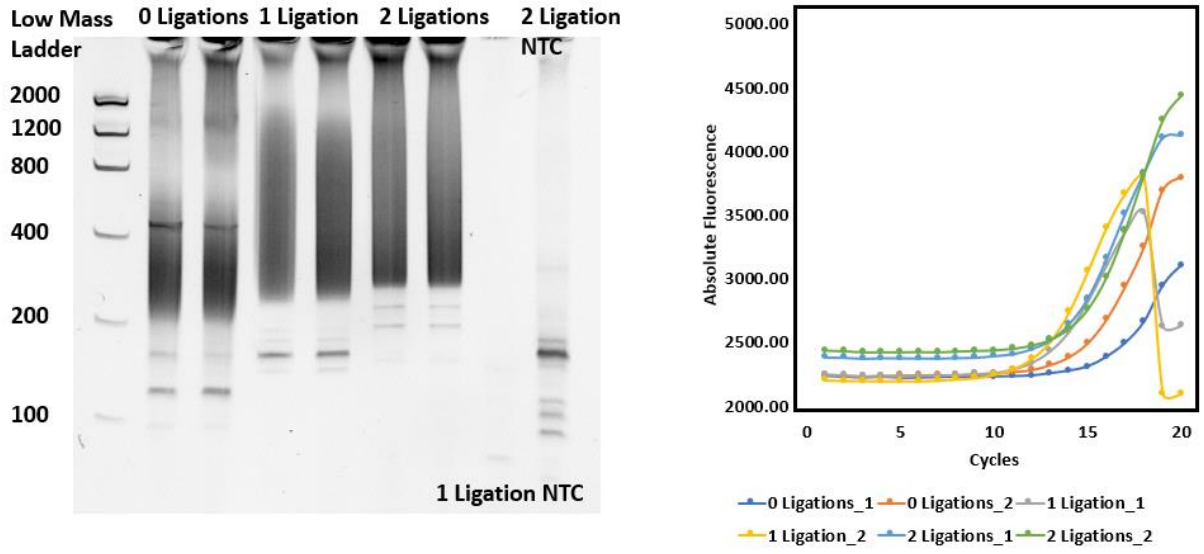


Figure 26: Gel bead T4 ligation with SNARE-Seq2 Adapters

However, the design was not compatible with our WGBS design. Sanger sequencing experiments revealed that one issue was the blunt-end ligation of mosaic end sequences illustrated in the Figure 27. This prompted me to try T7 ligase which has no blunt-end ligation activity. However, I discovered a second problem: the splint oligo was blocking the gap filling step that is required for our WGBS design as discussed in chapter 2. The melting temperature of this splint oligo was too high (calculated to be 80C). In contrast, the mosaic end sequence melting temperature is 54C which allows the mosaic end to unanneal from the transposon sequence during the gap filling step which occurs at 72C. One solution was to switch the polymerase from the high fidelity Q5 NEB polymerase to Taq polymerase as Taq polymerase can displace the splint oligo using a 5' exonuclease capability. In contrast, Q5 polymerase doesn't contain any 5' exonuclease or strand displacing capability. However, Taq polymerase was not compatible with our Tn5 fragmentation protocol. The first step in the gap filling protocol is to denature the Tn5. As previously published, this is typically performed using 0.1% SDS.(Picelli, Björklund, et al. 2014)In my experiments, this SDS needs to be quenched with 2% Triton X prior to gap filling to prevent polymerase inactivation by SDS. Between Taq polymerase and Q5 polymerase, Q5 polymerase displays a much higher resistance to this denaturation and quenching protocol. Taq polymerase is inconsistently active during this protocol. This insight led me to try the 3-level sci-ATAC ligation design.

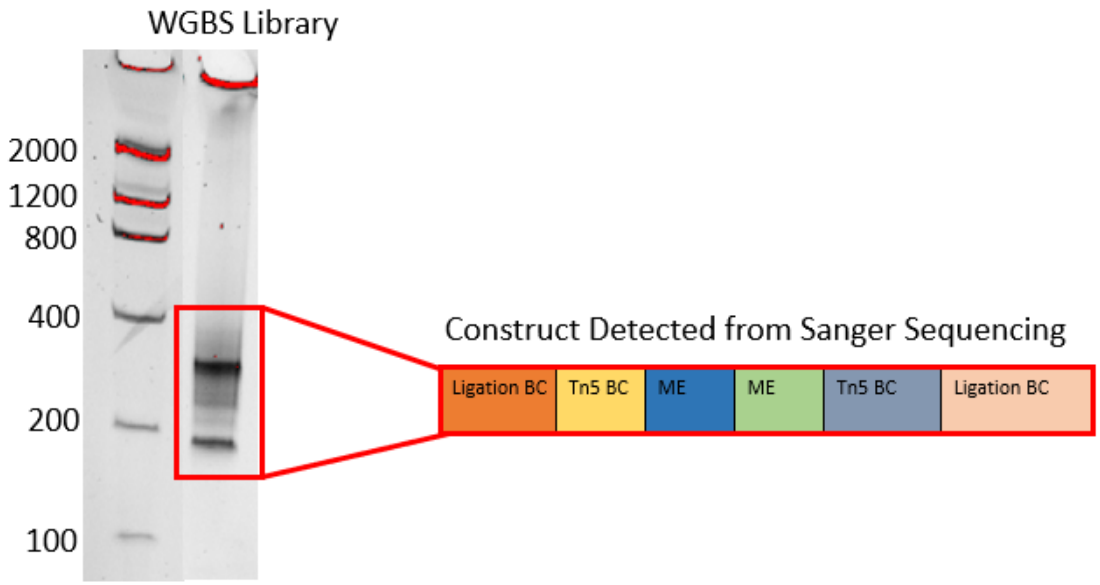


Figure 27: Blunt-end T4 ligation in WGBS experiments using SNARE-Seq2 adapters

The 3-level sci-ATAC design utilizes T7 ligase and, crucially, uses a shorter 15 bp splint oligo with a melting temperature of 58C. This lower melting temperature allows for the splint oligo to easily unanneal from the adapter/transposon junction during gap filling which occurs at 72C. Figure 28 shows the success of our library construction with this method and consistent lower barcode collision rates between the 2-level indexing and 3-level indexing designs. This design shows incredible promise in the development of both a single cell whole genome sequencing and whole genome bisulfite sequencing method at the scale of tens of thousands of cells per experiment with just three 96 well plates. We describe the detailed protocol to generate these libraries in the supplemental methods. Briefly, the encapsulated beads are first split into a 96 well plate containing 100-200 encapsulated beads per well. Following the previous 2-level indexing protocol, the beads are tagmented with Tn5 adding the first cell barcode. The beads are then pooled, washed, and split into a second 96 well plate where the second cell barcode is ligated onto the Tn5 sticky end. The beads are then pooled and then split again to a third 96 well

plate where roughly 40 encapsulated cells or nuclei are input per well. In the case of whole genome sequencing, PCR primers are added after Tn5 fragmentation to amplify the library and add the third cell barcode. In the case of whole methylome sequencing, the same protocol described in chapter 2 is performed but the linear amplification barcoded primer after bisulfite conversion is reverse complement to the ligated adapter.

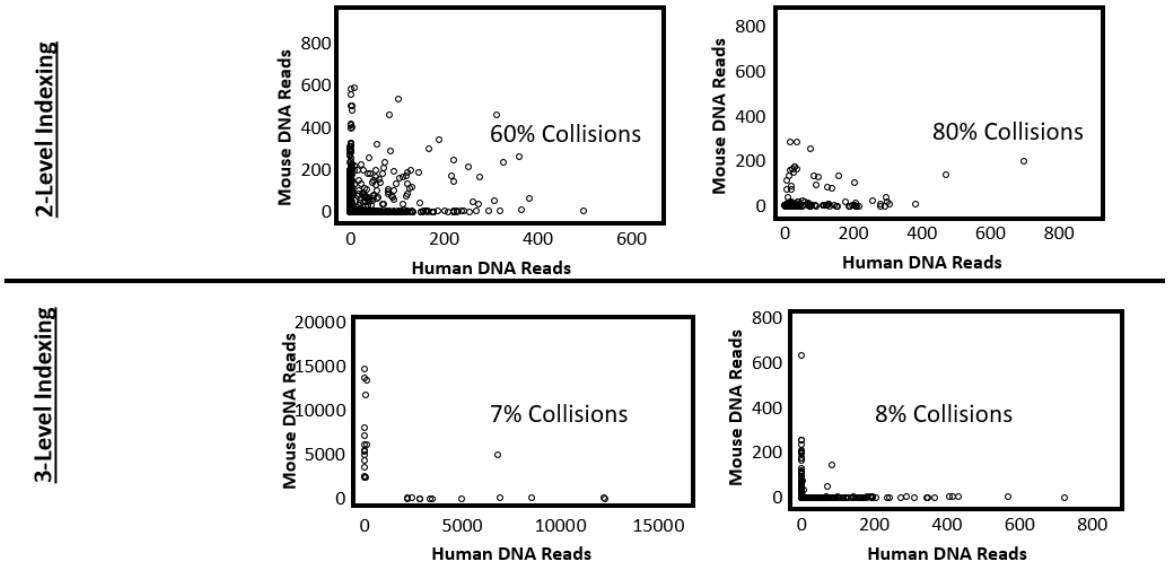


Figure 28: Successful WGBS library construction with 3-level sci-ATAC design adapted to our WGBS protocol.

Figure 29 shows the sequencing statistics at the single cell level using our 3-level combinatorial indexing method. We demonstrate high alignment rates, a mean alignment rate of 62 +/- 8.4%, like the previous 2-level indexing method We achieved Furthermore, we show how the global CG methylation could be used to discriminate single cells in a cell mixture of human cancer HCT116 colorectal cells and mouse fibroblast 3T3 cells. The hypomethylation of HCT116 cancer cells compared to non-cancerous tissue has been described in previous studies.

(Lengauer, Kinzler, and Vogelstein 1997)

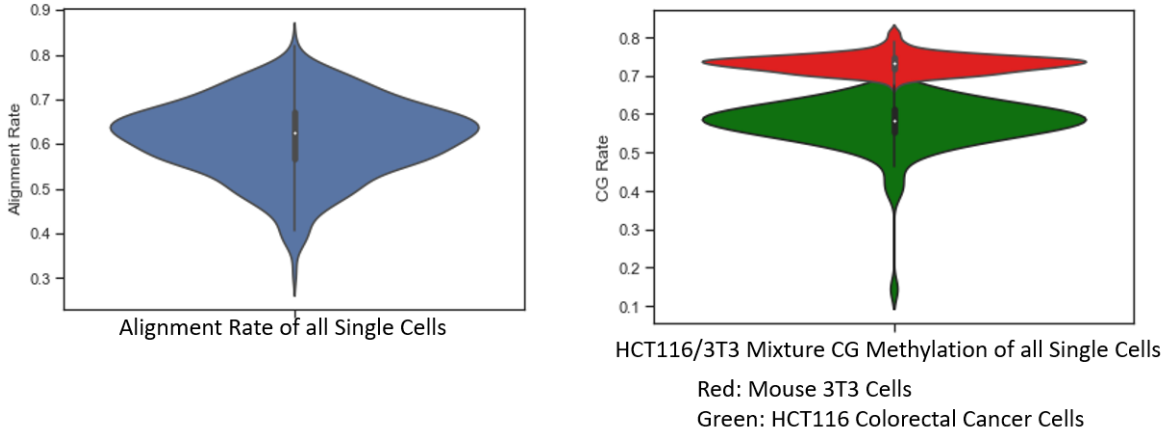


Figure 29: Preliminary sequencing statistics of 3-level WGBS library construction method.

3.3.2 The Development of the cDNA Recovery Method

3.3.2.1 Exponentially Amplifying cDNA Prior to Bisulfite Conversion

From chapter 2, the remaining major challenge of the incorporation of cDNA co-sequencing with WGBS. Building on previous observations in chapter 2, we reasoned that exponentially amplifying cDNA prior to bisulfite conversion could generate enough cDNA product that is recoverable post bisulfite conversion fragmentation. This was inspired by scnmCAT-seq where 10 cycles of full-length cDNA amplification prior to bisulfite conversion resulted in enough intact cDNA fragments to recover the transcriptomes of single cells.(Luo et al. 2022) The challenge was to first generate full length cDNA in the gel bead platform. The exponential amplification of cDNA as demonstrated in SPLiT-Seq, SNARE-Seq2, and PAIRED-Seq relies on the addition of a template switch oligo (TSO) once reverse transcriptase reaches the 5’ end of the RNA. This takes advantage of a feature of reverse transcriptase to often add cytosines on the extended cDNA product. This is beautifully demonstrated in the single cell sequencing technique, SMART-Seq.(Picelli, Faridani, et al. 2014) Typically, reverse transcription with the addition of a TSO requires 90 minutes of incubation at 42C to complete. In

addition, the RNA needs to be free of RNA binding proteins for the reverse transcriptase to reach the 5’ end of the RNA. Together, this requires the nucleus to sufficiently denatured. Thus, combinatorial indexing methods that use the nucleus like SPLiT-Seq require two rounds of reverse transcription. The first round is a partial reverse transcription that ensures that the RNA is binding to the reverse transcription primer is stable. After combinatorial indexing, the nucleus is denatured along with the RNA binding proteins. A second round of reverse transcription is required for the cDNA to be extended to the 5’ end of the RNA and template switching to occur. (Rosenberg et al., n.d.; C. Zhu et al. 2019; Plongthongkum et al. 2021) Secondly, we had to invent a barcoding scheme that protected the cDNA from Tn5 barcoding (necessary for the barcoding of gDNA for WGBS). Tn5 barcoding would fragment the cDNA and prevent the exponential amplification of full-length cDNA using the TSO and capture primer PCR adapter sequences.

In our gel bead platform, we similarly perform partial reverse transcription prior to in-nuclei encapsulation. After nuclear and RNA binding protein denaturation inside of the gel bead, reverse transcription is then completed with a TSO in a similar fashion with a few modifications. TSO based reverse transcription in polyacrylamide gel beads was first documented in a single cell RNA sequencing polyacrylamide gel bead protocol called BAG-Seq.(Li et al. 2020) Instead of the typical 42C for 90 minutes reverse transcription, this protocol utilizes 42C for 60 minutes followed by 50C for 60 minutes to account for reverse transcriptase and TSO diffusion through the gel bead. Utilizing this reverse transcription protocol, we created full length cDNA with the capture primer adapter on one end and TSO adapter on the other.

The next challenge was to invent a barcoding scheme that allows for the full length cDNA amplification prior to bisulfite conversion. Previously published cDNA combinatorial indexing

methods leverage the 5’ overhang on the capture primer to ligate cell barcode adapters. (Rosenberg et al., n.d.; C. Zhu et al. 2019; Plongthongkum et al. 2021) Because the 5’ end of the capture primer in this method is modified with an acrydite, this ligation protocol cannot be used. However, a 3’ cDNA overhang could be created if the RNA and TSO sequence could be removed as shown in Figure 30. The RNA can be digested with RNAseH and the TSO sequence could either also be digested with RNAseH or with brief high temperature heating and blocking with a sequence reverse complement to the TSO to prevent the TSO from reannealing to the single stranded cDNA.

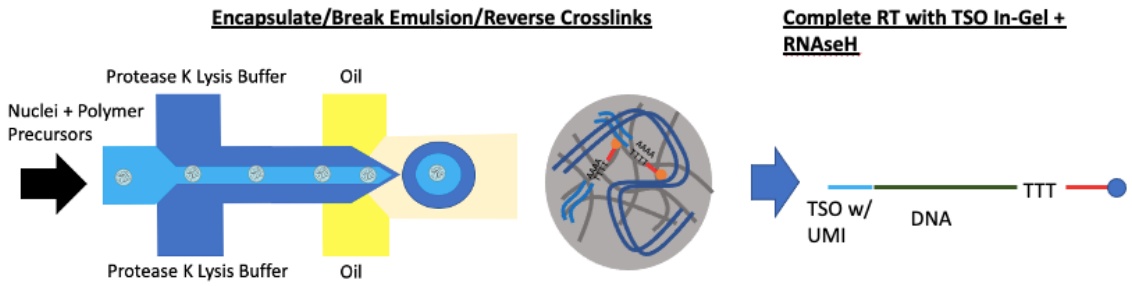


Figure 30: Encapsulation, synthesis of full-length cDNA, and digestion of RNA with RNAseH

With the removal of RNA and the TSO, we first tagmented the DNA to insert the first DNA barcode as described previously. The cDNA is not tagmented because it is single stranded. In the same well, we then ligated an adapter to the TSO end of the cDNA. Although the DNA and cDNA barcode designs are different, the barcode itself is the same. We then pool the beads, split into a second 96 well plate and perform T7 ligation with the same adapter as described previously in the 3-level WGBS method. Figure 31 illustrates this method.

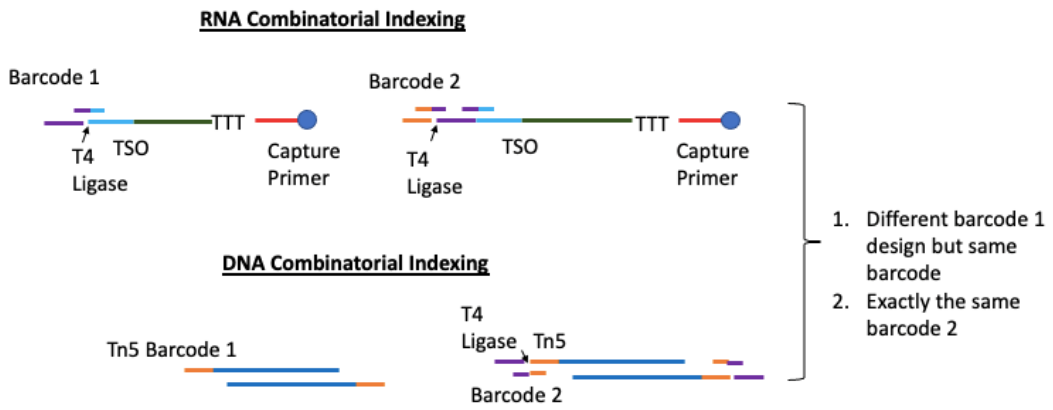


Figure 31: Template switch oligo based combinatorial indexing integrated with the WGBS 3-level indexing protocol.

In our experiments, this approach proved to be a more inefficient way to extract cDNA as the amount of diffusion out of the gel bead was too low to consistently generate high quality libraries. Thus, we had to return to a Tn5 based approach to fragment the cDNA and allow sufficient extraction of these sequencing from the gel bead. Furthermore, the amplification of ligated TSO products produced mostly off-target products. This could be due to the non-specificity of the addition of the TSO sequence during reverse transcription.

3.3.2.2 3-Level Tagmentation-Based cDNA Generation Protocol

With the exponential amplification of cDNA deemed an inviable approach, we decided to split the cDNA library and gDNA prior to bisulfite conversion. In this approach, we first create full length encapsulated cDNA inside of the gel bead after encapsulation like previously described but without the TSO sequence. We then invented a protocol to perform second strand synthesis of the cDNA using a mixture of RNAseH, DNA polymerase I, and DNA ligase slowly degrade the RNA and create a second strand of DNA complement to the one synthesized during reverse transcription. The details of this method are described in the supplementary methods. This double stranded cDNA and DNA are then tagmented with the same barcode followed by ligation with the same barcoded adapters. Prior to bisulfite conversion, the cDNA was then

linearly amplified for 10 cycles as described previously with a few modifications. Firstly, the linear amplification PCR reaction volume was doubled. After linear amplification, each reaction was pelleted at 300g for 2 minutes and vortexed to resuspend the beads twice. This was used to assist in the diffusion of linearly amplified products from the gel beads. Finally, the beads were pelleted, and half of the supernatant was carefully removed without disturbing the bead pellet and transferred into a separate plate. We found that this is crucial as the majority of the gDNA is still inside of the gel bead. Thus, the majority of the gDNA is in the original plate containing the beads while a fraction of linearly amplified cDNA is separated into the separate plate. After splitting the libraries, bisulfite conversion reagent is added to the gDNA plate, and WGBS library construction proceeds as previously described. In the separated cDNA plate, barcoded primers reverse complement to the ligation adapter is added and 7 cycles of exponential amplification are performed. We then performed SPRI bead purification using a 0.8X ratio of each well followed by a second round of exponential amplification with PCR primers containing Illumina sequencing adapters. After this PCR is complete, the libraries were pooled followed by two rounds of SPRI bead purification using a 0.8X ratio to prepare the library for sequencing. The details of this protocol are in the supplementary protocols. To test the success of our combinatorial barcoding scheme, we generated libraries from a cell line mixture of human and mouse cells like described previously. Figure 32 shows that the DNA library had low barcode collision rates. In contrast, the RNA library was completely mixed. We hypothesized that this mixing result was because our method of cDNA synthesis prior to combinatorial indexing generated too much background which would be covalently attached to the gel beads during encapsulation. In contrast, other combinatorial indexing approaches that perform in-nuclei cDNA generation gradually remove the background cDNA as the nuclei are washed between each

combinatorial indexing step. Most of this background occurs during the nuclei isolation as the cytoplasmic RNA can remain after cell lysis. In addition, the standard pelleting and washing of nuclei during these steps typically result in extensive nuclei lysis. Thus, we predict that the success of our method relies on high quality nuclei/cell isolation techniques that minimize background RNA.

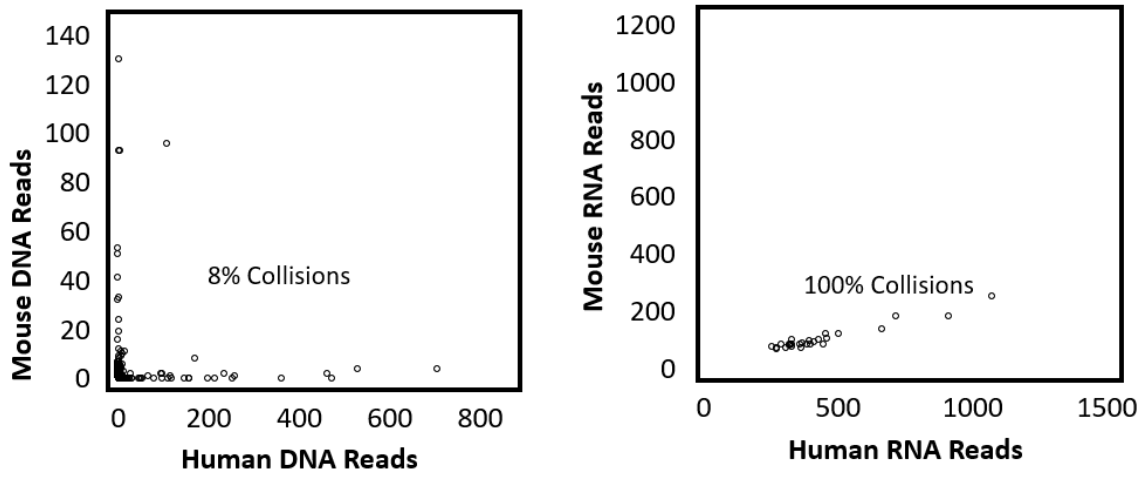


Figure 32: Species mixing results from 3-level Tn5 based library construction.

To minimize background RNA, we decided to first develop our protocol using single cells. As mentioned previously, cell lysis during nuclei isolation generates extensive free RNA which could be covalently attached to our gel beads causing extensive barcode collisions. Instead of performing in-situ reverse transcription, we decided to first encapsulate and lyse the cells with a key modification. The cells would be co-encapsulated with the acrydite modified reverse transcription primers to allow for the capture of RNA polyadenylated bases. The emulsion breaking buffers were modified to include saline-sodium citrate buffer (commonly known as SSC buffer). This high salt buffer enhances the stability of the polyadenylated and reverse transcription primer hybridization to prevent the free diffusion of RNA after encapsulation. Full

length cDNA is then generated as described previously in the gel bead. Figure 33 illustrates this protocol. The details of this protocol are in the supplementary methods.

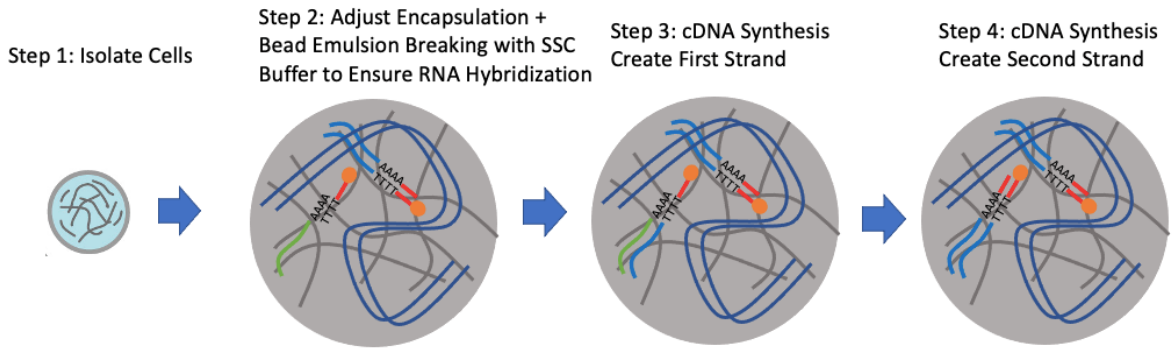


Figure 33: Generation of full-length cDNA within the gel bead

We then generated libraries from a cell line mixture of human and mouse cells using the same protocol described previously to assess the barcode collision rates. Figure 34 shows the success of this method where both DNA and RNA libraries demonstrate low barcode collision rates.

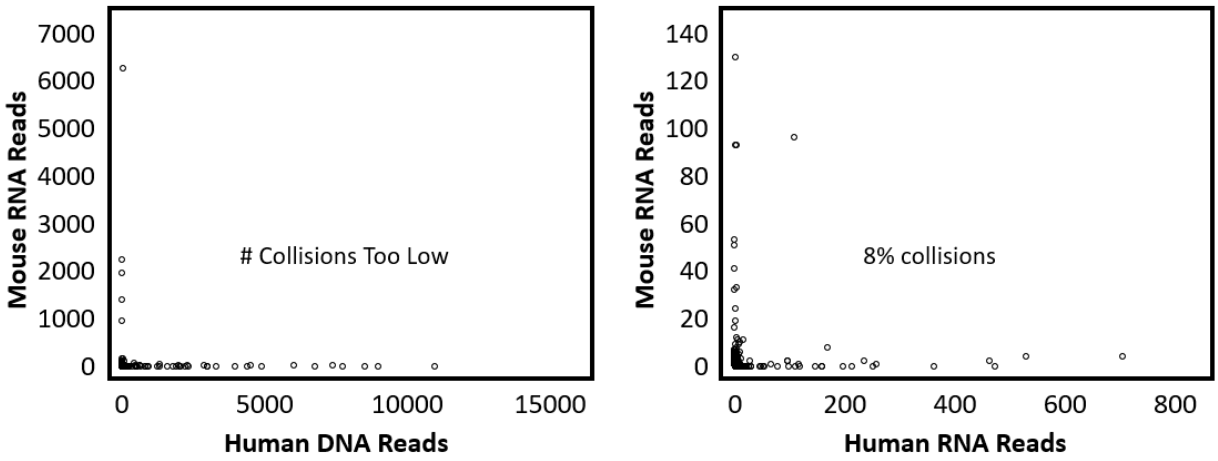


Figure 34: Barcode collision rate assessment of in-gel cDNA synthesis single cell encapsulation approach.

3.3.2.3 Biological validation of RNA Libraries

After identifying the correct RNA sequencing strategy, we assessed the biological relevance of our libraries. We created three RNA libraries using our method: encapsulated HCT116, in-tube HCT116, and in-tube neuroblastoma U87 cells. After sequencing, the gene counts of each library were correlated, and marker genes were identified. Briefly, the single cell resolution encapsulated HCT116 library were first bulked to enhance correlations. The cDNA reads were trimmed, filtered, and then aligned to the human genome using STAR. The htseq package was then used to generate a gene counts matrix. The gene counts matrix was then log normalized using scanpy. Log normalized counts per million of the in-tube HCT116, in-tube U87, and encapsulated HCT116 were then plotted. Marker genes for HCT116 and U87 found in literature were then labeled. The details of how this analysis was performed is documented in the supplementary methods. Figure 35 shows that our gel encapsulation HCT116 RNA sequencing technique recovered the expected marker gene expression. Highly expressed marker genes for the neuroblastoma cells such as Vim are only expressed in brain tissue. The low expression of these gene among other U87 marker genes found in our HCT116 libraries validated the biological relevance of our RNA sequencing method. The high expression of TCFL2 in HCT116 cells also showed hypomethylation in TCFL2 binding sites matching expectations.

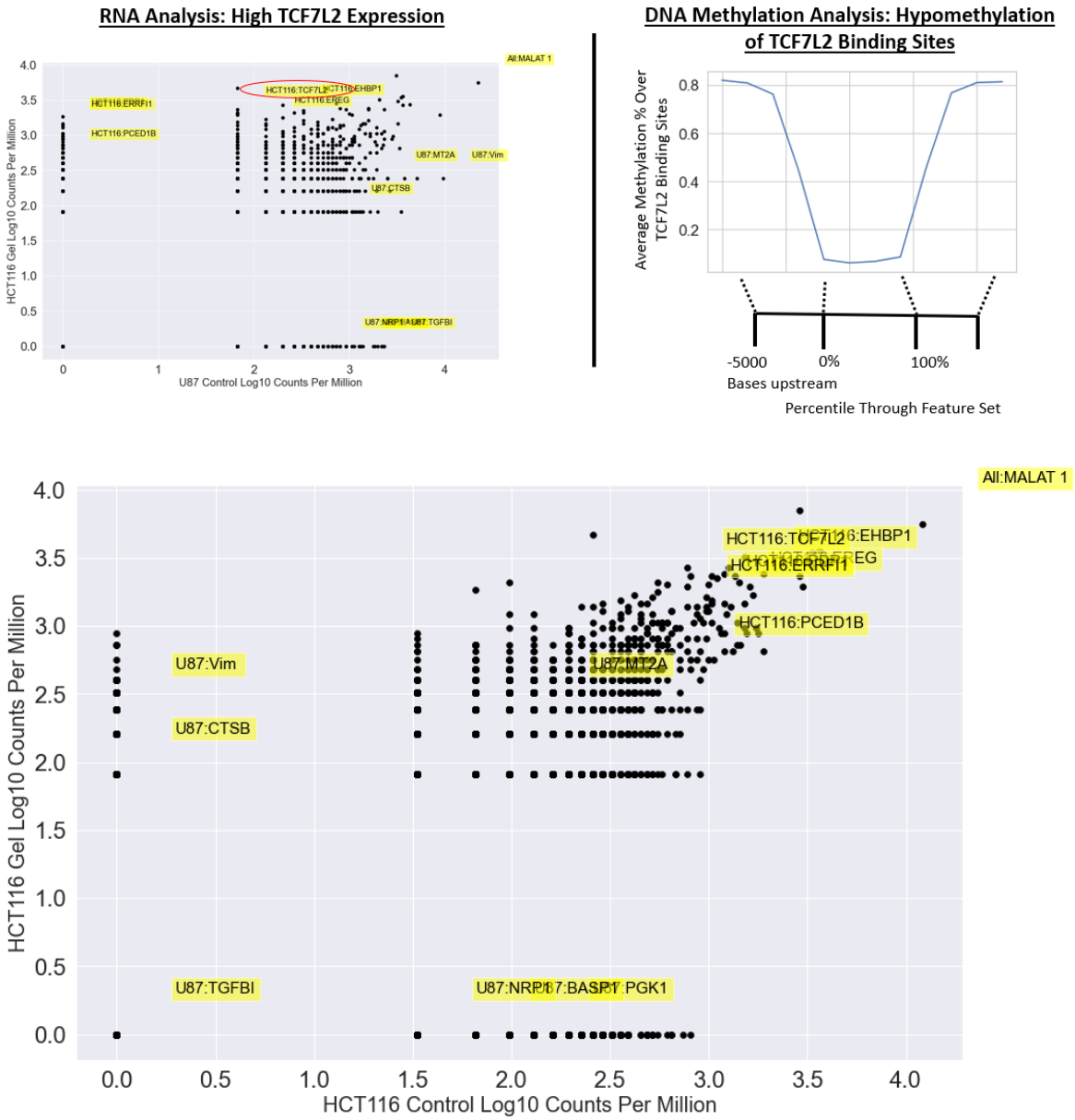


Figure 35: Log normalized counts per million of the U87 in-tube and HCT116 encapsulated sample plotted. DNA methylation at TCF7L2 (HCT116 Marker) binding sites are hypomethylated as expected (top). Log normalized counts per million of the HCT116 in-tube and HCT116 encapsulated sample plotted (bottom). The labeling of genes follows the convention: <Cell type>:<Marker Gene>. MALAT 1 was used as a marker gene and was detected in all libraries at high levels.

3.3.3 Optimizations of Library Formation and Performance

After demonstrating the potential of our 3-level WGBS and RNA co-sequencing method, we next wanted to assess the consistency of the method and library complexity. Table 4 illustrates the variability in barcode collision rates across various experiments. Published combinatorial indexing methods typically result barcode collision rates no more than 10%. (Rosenberg et al., n.d.; Plongthongkum et al. 2021; C. Zhu et al. 2019; Mulqueen et al. 2018)

Table 4: Variability in barcode collision rates with sci-Gel co-sequencing protocol

Experiment	DNA Collision Rate
July 2021 (First L3 Experiment)	6%
November 2021_1	4%
November 2021_2	8%
December 2021	40%
Jan 2022	8%
Mar 2022_1	11%
Mar 2022_2	30%
Mar 2022_3	15%

Thus, we next explored the potential causes of this variability. Across multiple encapsulations, we carefully performed an imaging analysis to correlate potential features with high barcode collision rates. We first noticed that the freeze/thawing process that we typically employed to store the beads to preserve RNA integrity prior to reverse transcription caused extensive aggregation and gel bead destruction as shown in Figure 36.

Figure 36: Gel bead destruction during the freeze/thawing process. Beads were stained with DAPI to identify beads containing single cells/nuclei

We also observed encapsulation quality variability which we identified was caused by two factors: 1) the hydrophobic coating of the microfluidic device and 2) the polymerization of the gel prior to encapsulation. Figure 37 illustrates the inconsistent bead sizes due to the unoptimized hydrophobic coating of the microfluidic device and the non-spherical gel bead products that result from the partial polymerization of polyacrylamide prior to encapsulation.

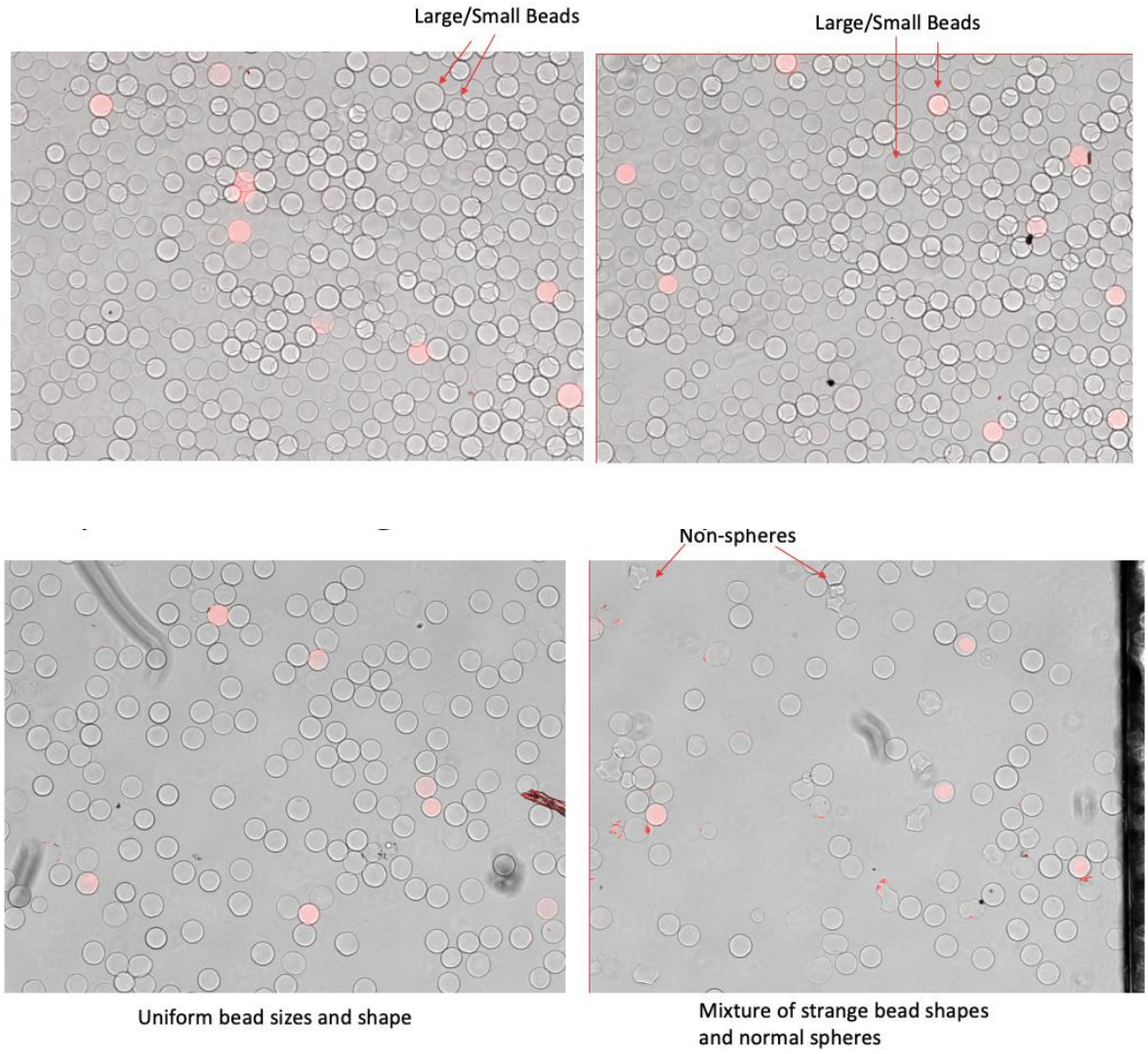


Figure 37: Variable bead sizes (top) and shapes (bottom) due to inconsistent gel bead polymerization and droplet formation.

To address the variability of bead sizes, we noticed that the droplet formation on the microfluidic device was inconsistent. The presence of larger than designed bead sizes leads to the increased probability of multiple cells or nuclei encapsulated in the same bead and heightened barcode collision rates. Originally, we used the microfluidic hydrophobic coating method described in inDrops.(Klein et al. 2015) Briefly, the device is coated with aquapel, air-dried, coated with FC-40, and then air dried. We found that FC-40 doesn't dry easily, and residual FC-40 prevented the proper formation of droplets. We developed an aquapel coating, air dying, coated with isopropyl

alcohol, followed by device drying at 55C for 30 minutes. The details of this method is in the supplementary methods. The isopropyl alcohol maintains the hydrophobicity of the microfluidic device while drying more easily than FC-40.

During the development of the RNA co-sequencing method, we adapted the microfluidic BAG-seq encapsulation scheme to capture the RNA summarized in Figure 38. Interestingly, the polymerization initiator, APS, was mixed the polymer precursor. In our experiments, we found that this encapsulation scheme resulted in gradual polymerization of the acrylamide prior to encapsulation. Thus, the non-spherical beads are the result of non-uniform polymerization of the acrylamide. In addition, this polymerization prior to encapsulation results in cells simply embedded into the gel instead of lysed and uniformly immobilized by the gel bead matrix. These poorly immobilized DNA and cDNA would cause extensive mixing resulting in elevated barcoded collision rates.

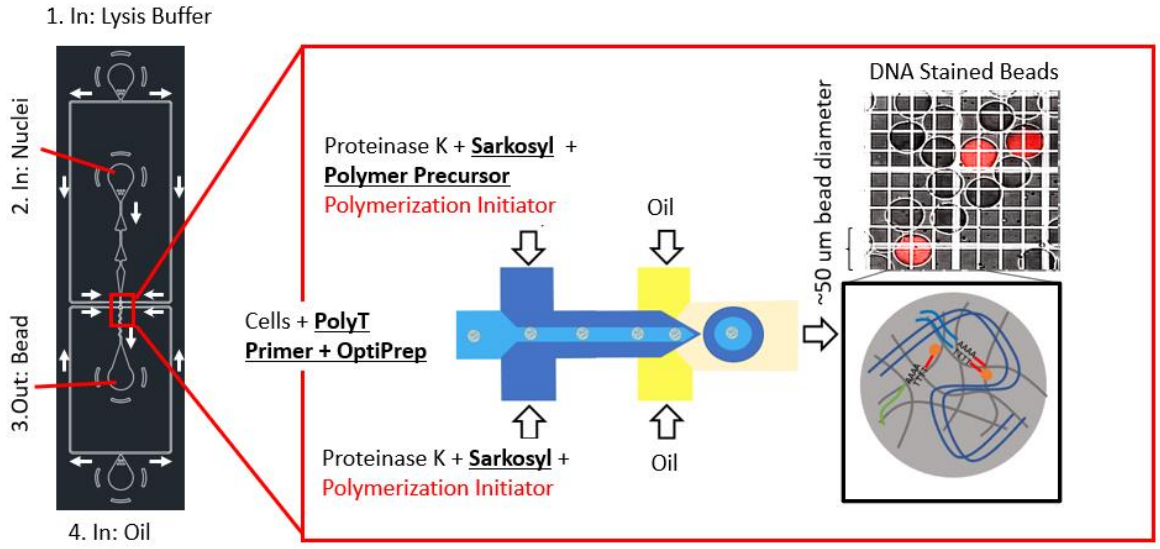


Figure 38: Encapsulation adapted from BAG-seq where the polymerization initiator, APS, is mixed with the polymer precursors

Thus, we separated the polymer precursors and APS in the encapsulation scheme illustrated in Figure 39. Interestingly, this resulted in poor lysis quality. Thus, we had to reoptimize the lysis

detergents to ensure high quality lysis and uniform entanglement of DNA within the polyacrylamide gel matrix. The results of these experiments are illustrated in Figure 40.

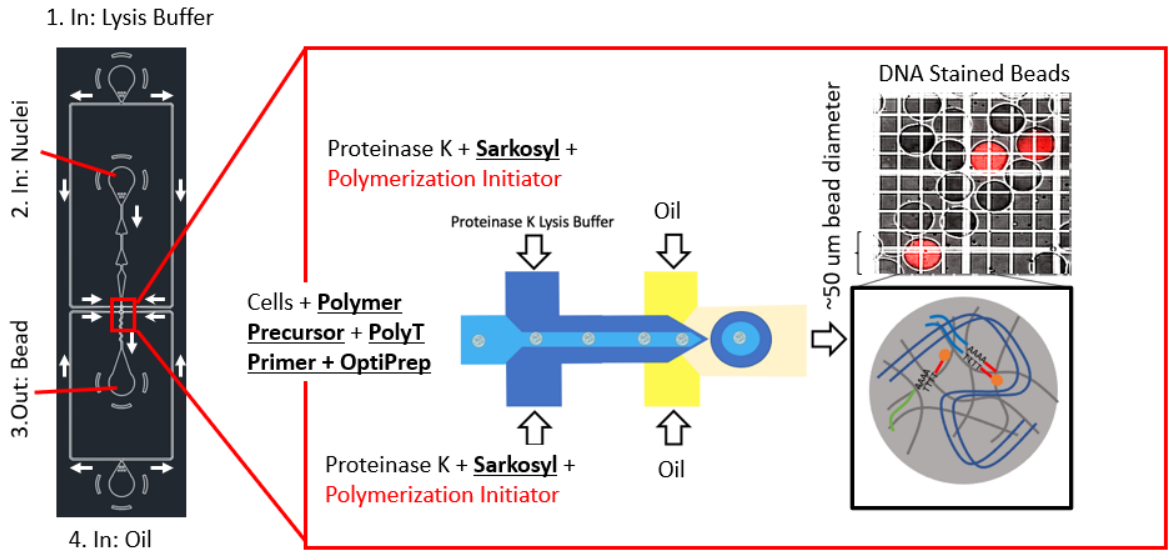


Figure 39: Encapsulation scheme with polymer precursors separated from APS.

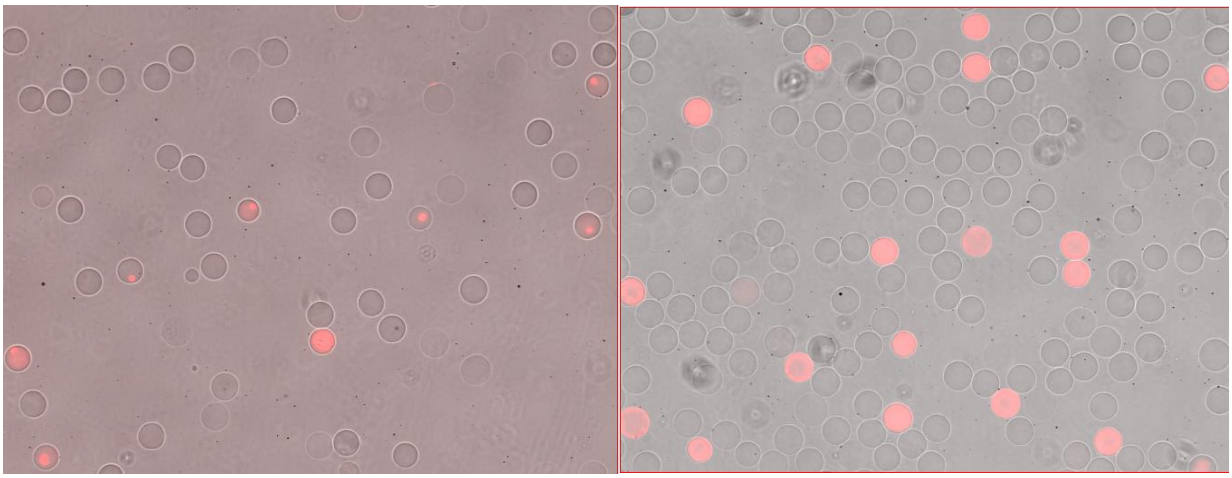


Figure 40: (Left) encapsulated cells using lysis buffers adapted from BAG-Seq. (Right) Encapsulated cells using stronger 0.5% SDS lysis buffer.

These design changes led to the successful lowering of barcode collision rates multiple cell-line mixture encapsulations as shown in Figure 41. However, the RNA library was completely mixed due to the high concentrations of SDS that prevent stable hybridization with the acrydite modified reverse transcription oligos.

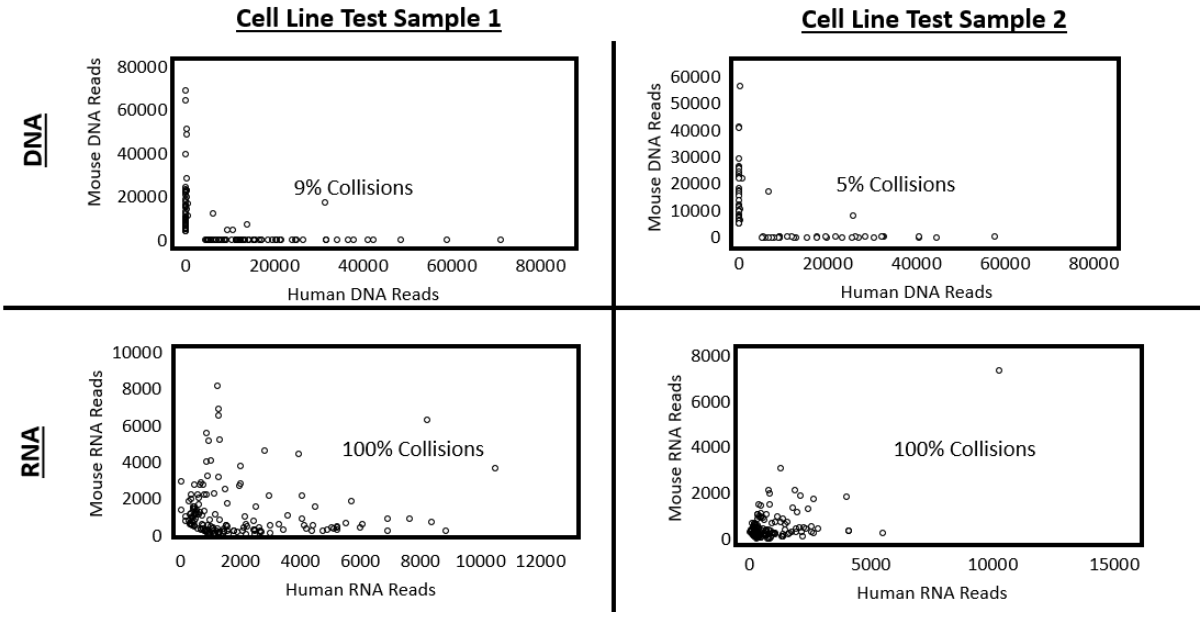


Figure 41: Consistent low barcode collision rates across two cell-line mixture encapsulations for WGBS but not RNA libraries

We next assessed the robustness of our encapsulation method with a human peripheral blood mononuclear cell (PBMC) mixture. To ensure high quality live cells, we first performed a dead cell magnetic separation technique. The specific details of these encapsulation protocols are detailed in the supplementary methods. Figure 42 shows the success of our encapsulation protocol in two PBMC samples.

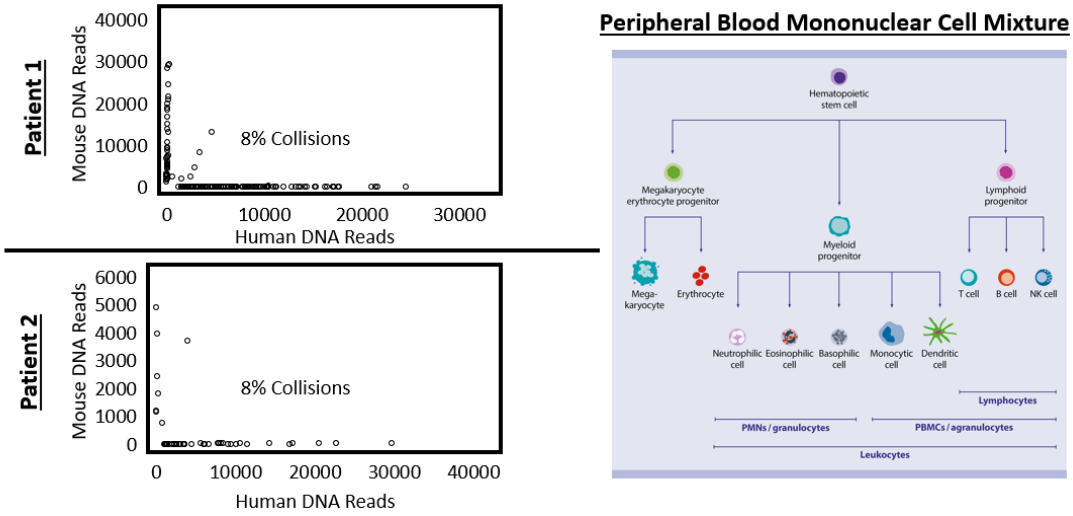


Figure 42: Consistent PBMC encapsulation and low barcode collision rates.

With the improved encapsulation stability and consistent low barcode collision rates, we moved on to improving the library complexity. The most optimized version of this protocol is documented in the supplementary methods. Briefly, we optimized each barcoding reaction: 1) the Tn5 insertion reaction, 2) the ligation reaction, 3) the post bisulfite tagging and PCR reactions. We screened Tn5 reaction concentrations starting at 0.05mg/mL and identified the optimal Tn5 concentration for 100-200 encapsulated cells to be 0.00625mg/mL. The optimal reaction time was found to be 90 minutes. We found that the optimal T7 ligase concentration was 0.75 U/uL or 2.5X higher than standard reaction conditions. Ligation times did not increase library complexity. We found that it was crucial for each well in the final PCR plate to be processed individually even after barcoding was complete. The exponential amplification of each well prior to pooling minimizes stochiometric barcode path biases that are commonly observed given the hundreds of barcoding reactions in the protocol. Figure 43 shows the results of these optimizations at the single cell level. The combination of these optimizations resulted in at least 100X increase in library complexities. For downstream single cell RNA analysis, previous publications used a 200 genes per cell cut-off. For downstream single cell DNA methylation analyses, sciMET uses a 30,000 unique reads per cell cut-off. From these results, we conservatively estimate that our method could detect at least 1000 genes per cell and over 100,000 unique WGBS reads. These library complexity metrics give us the promising preliminary evidence that our method could be used for human tissue profiling.

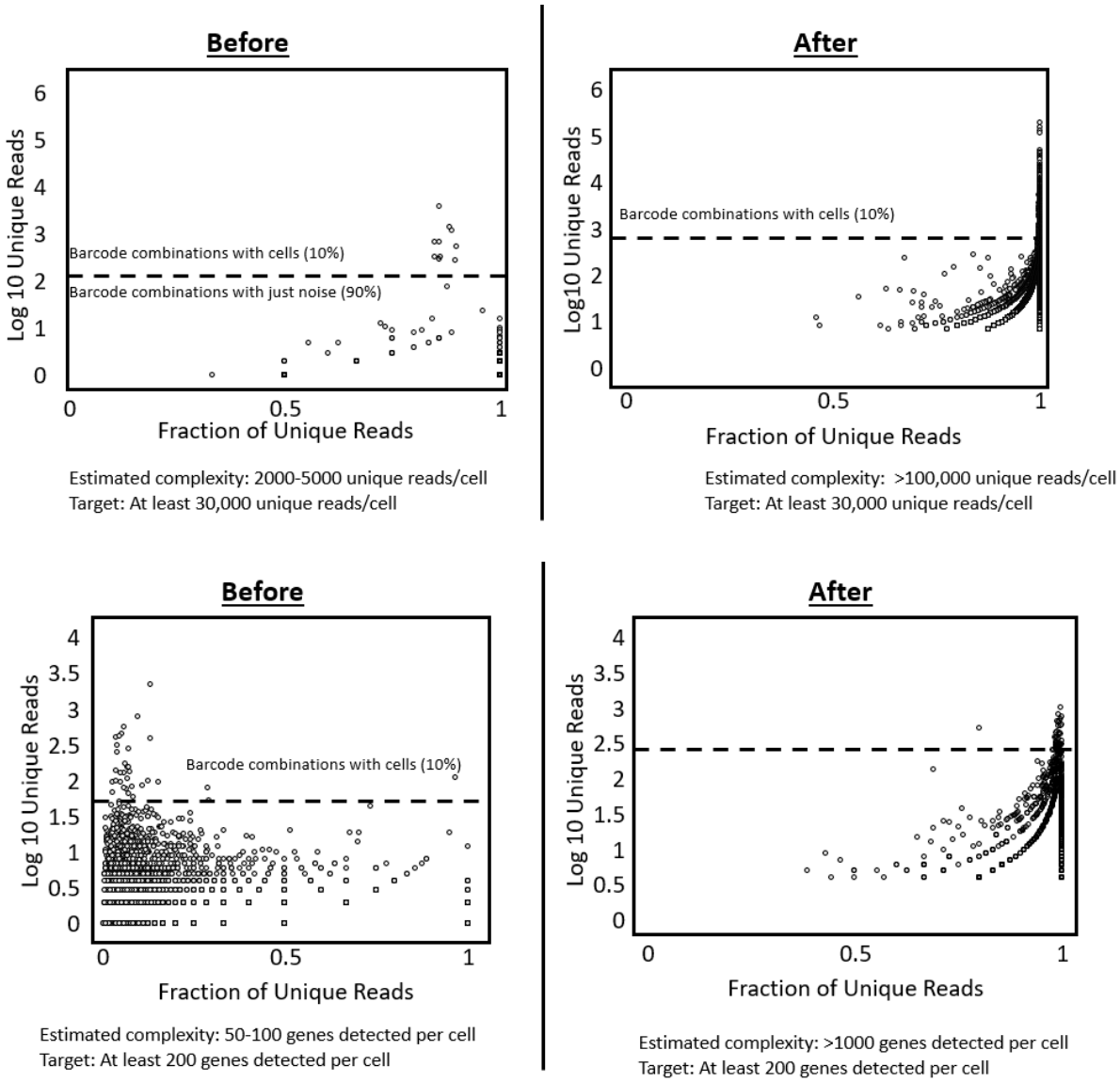


Figure 43: Optimizations of both the DNA and cDNA libraries resulting in 100X increases in library complexity.

3.4 Conclusion

Here, we describe the culmination of foundational works described in chapters 1 and 2 resulting in a successful prototype 3-level WGBS and RNA co-sequencing method. There are currently no methods with the same throughput and co-sequencing capabilities as the one described here. We expanded on our previous 2-level combinatorial indexing protocol to solve

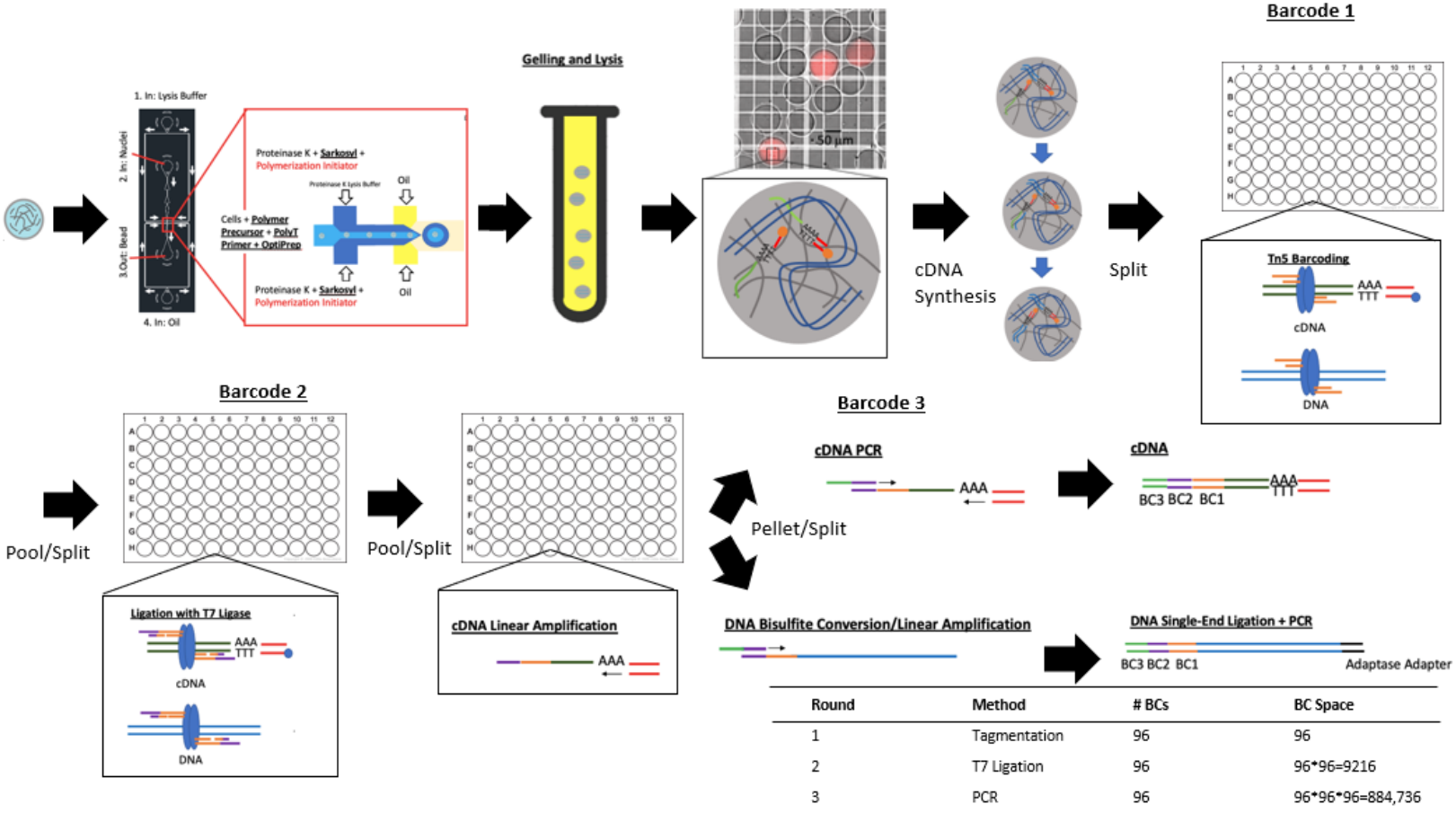
two major problems: inconsistent barcode collision rates and loss of the cDNA library. To make the leap from a 2-levels to 3 levels of combinatorial indexing, we tested two different barcode ligation paradigms: SPLiT-seq T4 ligation and 3-level sci-ATAC T7 ligation.(C. Zhu et al. 2019; Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnel, et al. 2020; Rosenberg et al., n.d.; Plongthongkum et al. 2021) We assess the positive and negative aspects of each approach and explain the reasoning behind choosing the 3-level sci-ATAC T7 ligation approach.

Next, we explored the two potential solutions to the loss of cDNA: splitting the cDNA library prior to bisulfite conversion and exponentially amplifying it. This led to the adaptation of a wide swath of single cell RNA sequencing methodologies: the partial in-nuclei reverse transcription adapted from SPLiT-Seq followed by the SMART-Seq2 full-length cDNA amplification and coupled with the scmCAT-Seq adaptation for bisulfite conversion and the sci-RNA cDNA tagmentation based approach. We found that the full-length cDNA exponential amplification approach couldn't sufficiently generate enough cDNA to diffuse out of our gel beads for sequencing. The most promising approach was to use Tn5 to sufficiently fragment the cDNA to allow for diffusion during the linear amplification step prior to bisulfite conversion. The cDNA and gDNA libraries were then split prior to bisulfite conversion. However, we found that the partial in-nuclei reverse transcription approach created too much free cDNA background that eventually was covalently attached to the polyacrylamide beads after encapsulation. Thus, this background caused extensive barcode collision rates. We then finally arrive at the most promising solution by performing reverse transcription and second strand synthesis in gel. This approach combined with our cDNA splitting approach successfully created both DNA and cDNA libraries with low barcode collision rates. Table 5 summarizes the single cell RNA

sequencing methods adapted and tested in our gel bead platform that guided the development of the correct method. The lessons learned from each preceding chapter in addition to the ones described here culminated in a final working prototype that is graphically summarized in Figure 44.

Table 5: Summary of single cell RNA sequencing methods adapted to our gel bead protocol and subsequent results

Method Molded to Gel Beads	Sequencing	Single Cell?	Outcome/Next Steps
Sci-RNA seq	1% alignment rate	~50% collision rate	Nuclei fixation, no Tn5
SPLiT seq/SNARE2-seq/PAIRED-seq	cDNA too long to diffuse out of gel bead		Re-introduce Tn5, but with full length RNA
SPLiT seq/SNARE2-seq/PAIRED-seq + Sci-ATAC-3-Level	50% alignment rate	100% collision rate	In-nuclei reverse transcription denatures nuclei
In-Gel RT + sci-ATAC 3-Level	48%	8.6% collision rate	Works with whole cell



Round	Method	# BCs	BC Space
1	Tagmentation	96	96
2	T7 Ligation	96	96*96=9216
3	PCR	96	96*96*96=884,736

Figure 44: Graphical summary of our 3-level WGBS and RNA co-sequencing protocol.

We then further optimized our protocol to resolve inconstancies in the polyacrylamide gel bead formation and performed a human tissue a proof of concept with PBMCs. The optimizations of each barcoding reaction that led to over 100X increase in library complexity compared to our initial prototype. Our specific protocol can process 50,000-100,000 cells per experiment with three 96 well plates. With further optimization using 384 well plates could increase the throughput of this platform to 3,000,000-5,000,000 cells per experiment which could be used to profile organ systems. Future work involving the methylome profiling of the PBMCs would showcase the capabilities of this method and be the first multi-omic RNA and DNA methylation study of PBMCs at the single cell level. Furthermore, this work would demonstrate the ultra-high throughput capabilities of our technology. Specifically, the single cell RNA datasets of our PBMC sample could be projected onto the 10X PBMC reference dataset using Seurat. Cell type labels from this reference could be transferred to our single cell RNA datasets to assist in cell type calling and the formation of pseudo bulk methylomes. As mentioned previously, the creation of pseudo bulk methylomes could generate enough methylome coverage for the identification of cell-type specific differentially methylated regions using CG methylation in PBMCs that have never been profiled at the cell-type level. Careful optimization of nuclei isolation methods to minimize cell free RNA could also enable the use of nuclei with this method. The ability to process nuclei would allow this protocol and pseudo bulk methylome analysis framework to add cell type specific methylation annotations to recently published single cell RNA atlases of terminally differentiated tissues such as human kidney and lung. (Travaglini et al. 2020; Lake et al., n.d.)

Portions of Chapters 3 are in part are a reprint of material in submission as it appears in "Ultra-High Throughput Single Cell Co-Sequencing of DNA Methylation and RNA using 3-Level Combinatorial Indexing" The dissertation author was the primary author of this paper along with Andrew Richards and Kun Zhang.

DISCUSSION AND FUTURE DIRECTIONS

The recent development of single cell WGBS tools opens the door to efficiently map the methylome of heterogeneous human tissues.(Luo et al. 2018; 2022; Callaway et al. 2021) However, the innovation of high cell throughput methods has significantly lagged single cell RNA-seq and DNA accessibility sequencing techniques. Currently, snmC-seq is still the only method that has demonstrated single cell methylome libraries of sufficient complexity to map both cell and cell subtypes in the brain. The high throughput of this method relies on liquid handlers which limits further practical scaling beyond thousands of cells. The development of computational methods to interpret the gene activity of methylation features of nearby genes also remains challenging because DNA methylation can be both positively and negatively associated with gene expression depending on the cell type. In contrast, DNA accessibility can be used to predict RNA expression of nearby genes because there are consistently positive correlations between both -omes.(Liu et al. 2020; Pliner et al. 2018) Therefore, single cell WGBS and transcriptome co-sequencing methods are crucial to advancing our understanding of the functional role of DNA methylation and developing analysis frameworks that can predict its effect on gene expression. Multi-omic analysis frameworks recently demonstrated in snmCAT-seq and scNMT-seq leverage the ability to directly correlate variations between the methylome and transcriptome, leading to the ability to cross-validate cell clusters, and observe the coupling of DNA methylation and gene expression at cell-type specificity. (Argelaguet et al. 2018; Luo et al. 2022)

Here, we introduce a method that aims to be the key next step in the development of single cell DNA methylation and transcriptome assays and analysis frameworks. Our technology tackles one of the largest problems with existing DNA methylation and RNA co-sequencing

methods: the throughput. Through the development of this technology, we demonstrate the potential throughput of our method to the scale of between 50,000-100,000 cells per experiment with three 96 well plates. We demonstrate barcode collision rates <10% in line with existing combinatorial indexing split-pool technologies.(Rosenberg et al., n.d.; C. Zhu et al. 2019; Plongthongkum et al. 2021; Mulqueen et al. 2018) Basically, this scales the throughput of existing single cell per well WGBS transcriptome co-sequencing techniques 100X enabling the potential of the method to be performed without capital intensive cell sorting and liquid handling machines.

Most of the work presented here is based on cell-line mixtures. We piloted our method in a couple of human tissues: human kidney and peripheral blood mononuclear cells (PMBCs). However, future work will be needed to assess the consistency of performance metrics of this technology beyond these sample types. In addition, there are a few remaining technical challenges to be solved. Firstly, the method is extremely sensitive to tissue quality and RNA background that is elevated during cell lysis or nuclei extraction. In other combinatorial indexing methods, extra-cellular/nuclear RNA is washed away during each step of the enzymatic reactions. In this method, the RNA background gets covalently attached to the gel beads resulting in high barcode collision rates. For single cells, we demonstrated success in PBMCs by using magnetic dead cell removal techniques. Except for PBMCs which are easily and cleanly isolated, many single cell isolations from solid human tissues are complicated the isolation technique easily ruptures the cytoplasm causing extensive free cytosolic RNA. For most human tissues, nuclei isolation approaches followed by single nuclei RNA sequencing have been shown to be a more promising approach. (Lake et al. 2019)For the use of single nuclei in our protocol, optimized nuclei isolation followed by FACS sorting may be required to reduce free RNA.

Lower amounts of RNA recovered during single nucleus compared to single cell RNA sequencing also need to be considered. In addition, the analysis of nuclear RNA libraries would need to consider the presence of intronic reads as opposed to mature RNA that is sequenced using cellular RNA which is predominantly cytosolic. Secondly, the DNA methylation library has variable quality with clustering issues during sequencing leading to low base calling accuracy and inability to sequence the cell barcode. This may be due to the post bisulfite adapter tagging technique which generates high amounts of off-target concatemer products. This could be potentially solved by optimizing the adaptase reaction conditions or carefully purifying the ligation products while retaining library complexity. Finally, there are no other methods that describe the success of the library chemistries used in polyacrylamide gel beads. This would be the first technique to document DNA Tn5 barcode insertion, ligation, and bisulfite conversion of cells encapsulated in polyacrylamide beads. Additional iteration of this protocol may be needed. We showcase a problem where the gel polymerization was found to be variable causing wide variabilities in cell mixing rates. Our present work aims to make the gel formation process more consistent, but that may affect our downstream optimizations in unintuitive ways.

At these high throughputs, we propose broad avenues of future work applying this method to study developing and terminally differentiated organs. PBMCs may be a good start as there are no single cell methylation profiles of these terminally differentiated cell types. Our pilot studies demonstrate the potential for single cell resolution with our method. Bulk methylome analyses of PBMCs have been associated with knee osteoarthritis and systemic sclerosis, but cell-type specificity was not established.(H. Zhu et al. 2018; M. Dunn et al. 2019) Recent work demonstrated 3-level combinatorial indexing to profile both the transcriptome and DNA accessibility of 15 organs comprising roughly 800,000 sci-ATAC seq and 4,000,000 sci-RNA

seq profiled cells.(Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnel, et al. 2020; Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnell, et al. 2020) Our protocol is currently designed to use three 96 well plates, but the adaptation of this protocol to three 384 well plates can increase the throughput from 50,000-100,000 cells to 3,000,000-5,000,000 per experiment due to the exponential expansion of this combinatorial indexing scheme. Thus, it is plausible to perform coupled single cell RNA and WGBS co-sequencing at organ system scales in the future. This ultra-high throughput technique could also offer the key to practically study DNA methylation in terminally differentiated tissue. Most single cell methylation studies focus on human brain or stem cell tissues that have high levels of CH methylation that assists in single cell clustering at a high sequencing depth of millions of reads per cell.(Luo et al. 2022; Callaway et al. 2021) In the case of terminally differentiated tissue with low CH methylation, the additional sequencing depth required to cluster using CG methylation may be well beyond the technical capabilities of both the existing technologies and funding constraints. At ultra-high throughput, 30X methylome of a cell type could be achieved with 500 cells sequenced at 1 million reads per cell pooled in a pseudo-bulk based on a few hundred RNA reads per cell. Thus, this technology and the accompanying high coverage pseudo-bulk whole methylome analysis strategy could be the most practical way to profile the methylomes of terminally differentiated tissue at the high single cell RNA-seq resolution.

REFERENCES

1. Ahn, Jongseong, Sunghoon Heo, Jihyun Lee, and Duhee Bang. 2021. “Introduction to Single-Cell Dna Methylation Profiling Methods.” *Biomolecules* 11 (7). https://doi.org/10.3390/biom11071013.

2. Angermueller, Christof, Stephen J. Clark, Heather J. Lee, Iain C. Macaulay, Mabel J. Teng, Tim Xiaoming Hu, Felix Krueger 2016. “Parallel Single-Cell Sequencing Links Transcriptional and Epigenetic Heterogeneity.” *Nature Methods* 13 (3): 229–32. https://doi.org/10.1038/nmeth.3728.

3. Argelaguet, Ricard, Stephen J. Clark, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt Andreas Kapourani, Ivan Imaz-Rosshandler 2019. “Multi-Omics Profiling of Mouse Gastrulation at Single-Cell Resolution.” *Nature* 576 (7787): 487–91. https://doi.org/10.1038/s41586-019-1825-8.

4. Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2018. “Multi-Omics Factor Analysis—a Framework for Unsupervised Integration of Multi-omics Data Sets.” *Molecular Systems Biology* 14 (6). https://doi.org/10.15252/msb.20178124.

5. Callaway, Edward M., Hong-Wei Dong, Joseph R. Ecker, Michael J. Hawrylycz, Z. Josh Huang, Ed S. Lein, John Ngai 2021. “A Multimodal Cell Census and Atlas of the Mammalian Primary Motor Cortex.” *Nature* 598 (7879): 86–102. https://doi.org/10.1038/s41586-021-03950-0.

6. Cao, Junyue, Diana R O’day, Hannah A Pliner, Paul D Kingsley, Mei Deng, Riza M Daza, Michael A Zager 2020. “A Human Cell Atlas of Fetal Gene Expression Techniques and Performed Sci-RNA-Seq3 Experiments with Assistance from R HHS Public Access.” *Science* 370 (6518). https://doi.org/10.17504/protocols.io.9yih7ue.

7. Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang 2019. “The Single-Cell Transcriptional Landscape of Mammalian Organogenesis.” *Nature* 566 (7745): 496–502. https://doi.org/10.1038/s41586-019-0969-x.

8. Chen, Song, Blue B. Lake, and Kun Zhang. 2019. “High-Throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell.” *Nature Biotechnology* 37 (12): 1452–57. https://doi.org/10.1038/s41587-019-0290-0.

9. Clark, Stephen J., Ricard Argelaguet, Chantriolnt Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger 2018. “ScNMT-Seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and

Transcription in Single Cells e." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-03149-4.

10. Domcke, Silvia, Andrew J. Hill, Riza M. Daza, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H. Milbank, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnel 2020. "A Human Cell Atlas of Fetal Chromatin Accessibility." *Science* 370 (6518). https://doi.org/10.1126/science.aba7612.

11. Domcke, Silvia, Andrew J. Hill, Riza M. Daza, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H. Milbank, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnell. 2020. "A Human Cell Atlas of Fetal Chromatin Accessibility." *Science* 370 (6518). https://doi.org/10.1126/science.aba7612.

12. Dzieran, Johanna, Aida Rodriguez Garcia, Ulrica Kristina Westermark, Aine Brigette Henley, Elena Eyre Sánchez, Catarina Träger, Henrik Johan Johansson, Janne Lehtiö, and Marie Arsenian-Henriksson. 2018. "MYCN-Amplified Neuroblastoma Maintains an Aggressive and Undifferentiated Phenotype by Deregulation of Estrogen and NGF Signaling." *Proceedings of the National Academy of Sciences of the United States of America* 115 (6): E1229–38. https://doi.org/10.1073/pnas.1710901115.

13. Gu, Hongcang, Ayush T. Raman, Xiaoxue Wang, Federico Gaiti, Ronan Chaligne, Arman W. Mohammad, Aleksandra Arczewska. 2021. "Smart-RRBS for Single-Cell Methylome and Transcriptome Analysis." *Nature Protocols*. Nature Research. https://doi.org/10.1038/s41596-021-00571-9.

14. Heard, Edith, Philippe Clerc, and Philip Avner. 1997. "X-CHROMOSOME INACTIVATION IN MAMMALS." www.annualreviews.org.

15. Hu, Youjin, Kevin Huang, Qin An, Guizhen Du, Ganlu Hu, Jinfeng Xue, Xianmin Zhu, Cun Yu Wang, Zhigang Xue, and Guoping Fan. 2016. "Simultaneous Profiling of Transcriptome and DNA Methylome from a Single Cell." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0950-z.

16. Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 2015. "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells." *Cell* 161 (5): 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

17. Kriaucionis, Skirmantas, and Nathaniel Heintz. n.d. "The Nuclear DNA Base, 5-Hydroxymethylcytosine Is Present in Brain and Enriched in Purkinje Neurons."
18. Lake, Blue B., Song Chen, Masato Hoshi, Nongluk Plongthongkum, Diane Salamon, Amanda Knoten, Anitha Vijayan.2019. "A Single-Nucleus RNA-Sequencing Pipeline to Decipher the Molecular Anatomy and Pathophysiology of

Human Kidneys." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-10861-2.

19. Lake, Blue B., Song Chen, Brandon C. Sos, Jean Fan, Gwendolyn E. Kaeser, Yun C. Yung, Thu E. Duong. 2018. "Integrative Single-Cell Analysis of Transcriptional and Epigenetic States in the Human Adult Brain." *Nature Biotechnology* 36 (1): 70–80. https://doi.org/10.1038/nbt.4038.

20. Lake, Blue B, Rajasree Menon, Seth Winfree, Qiwen Hu, Ricardo Melo Ferreira, Daria Barwinska, Edgar A Otto. n.d. "An Atlas of Healthy and Injured Cell 1 States and Niches in the Human Kidney." https://doi.org/10.1101/2021.07.28.454201.

21. Lan, Freeman, Benjamin Demaree, Noorsher Ahmed, and Adam R. Abate. 2017. "Single-Cell Genome Sequencing at Ultra-High-Throughput with Microfluidic Droplet Barcoding." *Nature Biotechnology* 35 (7): 640–46. https://doi.org/10.1038/nbt.3880.

22. Lengauer, Christoph, Kenneth W Kinzler, and Bert Vogelstein. 1997. "DNA Methylation and Genetic Instability in Colorectal Cancer Cells." *Medical Sciences*. Vol. 94. www.pnas.org.

23. Li, Siran, Jude Kendall, Sarah Park, Zihua Wang, Joan Alexander, Andrea Moffitt, Nissim Ranade. 2020. "Copolymerization of Single-Cell Nucleic Acids into Balls of Acrylamide Gel." *Genome Research* 30 (1): 49–61. https://doi.org/10.1101/gr.253047.119.

24. Liu, Jialin, Chao Gao, Joshua Sodicoff, Velina Kozareva, Evan Z. Macosko, and Joshua D. Welch. 2020. "Jointly Defining Cell Types from Multiple Single-Cell Datasets Using LIGER." *Nature Protocols* 15 (11): 3632–62. https://doi.org/10.1038/s41596-020-0391-8.

25. Luo, Chongyuan, Hanqing Liu, Fangming Xie, Ethan J. Armand, Kimberly Siletti, Trygve E. Bakken, Rongxin Fang. 2022. "Single Nucleus Multi-Omics Identifies Human Cortical Cell Regulatory Genome Diversity." *Cell Genomics* 2 (3): 100107. https://doi.org/10.1016/j.xgen.2022.100107.

26. Luo, Chongyuan, Angeline Rivkin, Jingtian Zhou, Justin P. Sandoval, Laurie Kurihara, Jacinta Lucero, Rosa Castanon. 2018. "Robust Single-Cell DNA Methylome Profiling with SnmC-Seq2." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-06355-2.

27. M. Dunn, Christopher, Michael C. Nevitt, John A. Lynch, and Matlock A. Jeffries. 2019. "A Pilot Study of Peripheral Blood DNA Methylation Models as Predictors of Knee Osteoarthritis Radiographic Progression: Data from the Osteoarthritis Initiative (OAI)." *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-53298-9.

28. Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

29. Mitra, Robi D, and George M Church. 1999. “In Situ Localized Amplification and Contact Replication of Many Individual DNA Molecules.” *Nucleic Acids Research*. Vol. 27.

30. Mulqueen, Ryan M., Dmitry Pokholok, Steven J. Norberg, Kristof A. Torkenczy, Andrew J. Fields, Duanchen Sun, John R. Sinnamon. 2018. “Highly Scalable Generation of DNA Methylation Profiles in Single Cells.” *Nature Biotechnology* 36 (5): 428–31. https://doi.org/10.1038/nbt.4112.

31. Mulqueen, Ryan M., Dmitry Pokholok, Brendan L. O’Connell, Casey A. Thornton, Fan Zhang, Brian J. O’Roak, Jason Link. 2021. “High-Content Single-Cell Combinatorial Indexing.” *Nature Biotechnology* 39 (12): 1574–80. https://doi.org/10.1038/s41587-021-00962-z.

32. Picelli, Simone, Åsa K. Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2014. “Tn5 Transposase and Tagmentation Procedures for Massively Scaled Sequencing Projects.” *Genome Research* 24 (12): 2033–40. https://doi.org/10.1101/gr.177881.114.

33. Picelli, Simone, Omid R. Faridani, Åsa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. “Full-Length RNA-Seq from Single Cells Using Smart-Seq2.” *Nature Protocols* 9 (1): 171–81. https://doi.org/10.1038/nprot.2014.006.

34. Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan. 2018. “Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data.” *Molecular Cell* 71 (5): 858-871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.

35. Plongthongkum, Nongluk, Dinh Diep, Song Chen, Blue B. Lake, and Kun Zhang. 2021. “Scalable Dual-Omics Profiling with Single-Nucleus Chromatin Accessibility and MRNA Expression Sequencing 2 (SNARE-Seq2).” *Nature Protocols* 16 (11): 4992–5029. https://doi.org/10.1038/s41596-021-00507-3.

36. Pluen, Alain, Paolo A Netti, Rakesh K Jain, and David A Berk. 1999. “Diffusion of Macromolecules in Agarose Gels: Comparison of Linear and Globular Configurations.”

37. Quake, Stephen R. 2022. “A Decade of Molecular Cell Atlases.” *Trends in Genetics*. Elsevier Ltd. https://doi.org/10.1016/j.tig.2022.01.004.

38. Rosenberg, Alexander B, † Charles, M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao. n.d. “Single-Cell Profiling of the Developing Mouse Brain and Spinal Cord with Split-Pool Barcoding.” https://www.science.org.

39. Sharifi-Zarchi, Ali, Daniela Gerovska, Kenjiro Adachi, Mehdi Totonchi, Hamid Pezeshk, Ryan J. Taft, Hans R. Schöler. 2017. “DNA Methylation Regulates Discrimination of Enhancers from Promoters through a H3K4me1-H3K4me3 Seesaw Mechanism.” *BMC Genomics* 18 (1). https://doi.org/10.1186/s12864-017-4353-7.

40. Travaglini, Kyle J., Ahmad N. Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene v. Sit, Stephen Chang. 2020. “A Molecular Cell Atlas of the Human Lung from Single-Cell RNA Sequencing.” *Nature* 587 (7835): 619–25. https://doi.org/10.1038/s41586-020-2922-4.

41. Uzun, Yasin, Hao Wu, and Kai Tan. 2021. “Predictive Modeling of Single-Cell DNA Methylome Data Enhances Integration with Transcriptome Data.” *Genome Research* 31 (1): 101–9. https://doi.org/10.1101/gr.267047.120.

42. Wu, Douglas C., and Alan M. Lambowitz. 2017. “Facile Single-Stranded DNA Sequencing of Human Plasma DNA via Thermostable Group II Intron Reverse Transcriptase Template Switching.” *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-09064-w.

43. Xu, Liyi, Ilana L. Brito, Eric J. Alm, and Paul C. Blainey. 2016. “Virtual Microfluidics for Digital Quantification and Single-Cell Sequencing.” *Nature Methods* 13 (9): 759–62. https://doi.org/10.1038/nmeth.3955.

44. Zhu, Chenxu, Miao Yu, Hui Huang, Ivan Juric, Armen Abnousi, Rong Hu, Jacinta Lucero, M. Margarita Behrens, Ming Hu, and Bing Ren. 2019. “An Ultra High-Throughput Method for Single-Cell Joint Analysis of Open Chromatin and Transcriptome.” *Nature Structural and Molecular Biology* 26 (11): 1063–70. https://doi.org/10.1038/s41594-019-0323-x.

45. Zhu, Honglin, Chengsong Zhu, Wentao Mi, Tao Chen, Hongjun Zhao, Xiaoxia Zuo, Hui Luo, and Quan Zhen Li. 2018. “Integration of Genome-Wide DNA Methylation and Transcription Uncovered Aberrant Methylation-Regulated Genes and Pathways in the Peripheral Blood Mononuclear Cells of Systemic Sclerosis.” *International Journal of Rheumatology* 2018. https://doi.org/10.1155/2018/7342472.

SUPPLEMENTAL METHODS

5.1 Summary of the Optimized 3-Level Combinatorial Indexed Co-Sequencing Method

The foundation of our platform is the encapsulation of single cells containing lysis buffer and acrylamide monomer in an oil emulsion using a microfluidic device droplet maker like the one used by 10X Genomics. Reverse transcription primers have 5' acrydite modifications to co-polmyerize with the acrylamide and capture the RNA. After an overnight incubation, each droplet polymerizes into a polyacrylamide bead with the genomic DNA dispersed and intertwined in the polyacrylamide matrix. The acrydite group incorporates the reverse transcription primers to the polyacrylamide back bone. The RNA hybridizes to the reverse transcription primers and are anchored to the gel bead. This polyacrylamide gel bead is accessible to the enzymes critically responsible for cDNA synthesis and combinatorial barcoding. After emulsion breaking, the beads undergo reverse transcription as described in other studies and second strand synthesis overnight (Li et al. 2020).

The DNA and RNA barcoding scheme is like previously published Tn5 based split and pool combinatorial barcoding methods but adapted for polyacrylamide beads as opposed to nuclei. (Domcke, Hill, Daza, Cao, O'Day, Pliner, Aldinger, Pokholok, Zhang, Milbank, Zager, Glass, Steemers, Doherty, Trapnell, et al. 2020; Cao et al. 2019) Briefly, the beads are dispersed into a 96 well plate so that each well contains roughly 200 encapsulated cells or nuclei. Hyperactive Tn5 containing 5' phosphorylated transposons tagment the beads adding the first DNA barcode using optimized reaction conditions found in this work. The beads are then pooled, washed, and split into a second 96 well plate where the second DNA barcode is ligated to the transposon overhang. Finally, the beads are then pooled, washed, and split into a third 96 well

plate. We use linear amplification for 10 cycles to first amplify the cDNA allowing it to diffuse out of the gel bead to split the cDNA libraries from the gDNA using a PCR primer reverse complement to the reverse transcription primer sequence. The beads are then pelleted and 50% of the supernatant containing the cDNA is exponentially amplified for 7 cycles adding the third barcode to the cDNA. The cDNA reaction is then bead purified using SPRI beads at a 0.8X ratio followed by another 10 cycles of PCR using a P5 primer and an i7 primer. Once this reaction is complete, the wells are pooled and 0.8X bead purification was performed twice on the pool.

After linear amplification and extraction of cDNA, the gDNA bisulfite conversion reagent is added to the remaining gDNA for bisulfite conversion. We followed the manufacturers protocol for desulphonation with a key modification. At this point, the magnetic beads coat the gel beads which contain the gDNA. Instead of eluting the DNA from the magnetic beads, we took the magnetic beads along with the gel beads and added them to a PCR reaction where the gDNA is linearly amplified for 20 cycles with primers hybridizing to the ligated adapter. This process allows gDNA to diffuse out of the gel bead. The third barcode is added to the gDNA during this linear amplification process. rSAP is then added to the reaction to remove all 5' phosphates that could potentially interfere with the adaptase protocol. The DNA is then bead purified using SPRI beads at a 1.2X ratio and eluted into the standard adaptase reaction protocol, following the manufacturer's instructions. PCR master mix containing a P5 primer and an i7 primer is then added to the heat inactivated adaptase reaction as described in scnmC-seq.(Luo et al. 2018) We then performed 8 cycles of exponential amplification. The reaction was then bead purified at a 0.8X ratio followed by another 8 cycles of PCR using P5 and P7 primers. Finally, the wells are pooled and 0.8X bead purification was performed twice on the pool.

5.2 Single-Cell 3-Level Detailed mDNA/RNA Gel Bead Sequencing Materials

Equipment

Name	Company	Catalog Number
Inverted microscope	Fisher Scientific Education	29AX
3 mL syringes	Beckton Dickinson	309657
PE-50 tubing	Instech	BTPE-50
Leur stub 22ga	Instech	LS22
Right angle couplers 22ga	Instech	SC22/15RA
PDMS Co-Flow Microfluidic Droplet Device		
30 mm cell strainer	Fischer Scientific	NC9682496
Deep well magnetic plate	Thistle Scientific	VP 771HH-LF

Chemical Reagents/Solutions

Name	**Company**	**Catalog Number**
Nuclei Isolation		
Nuclei Isolation Buffer (10 mM Tris-HCl, 10mM NaCl, 3mM MgCl2, 0.1% Igepal)	-	-
Bovine Albumin Fraction V (7.5% solution)	Fischer Scientific	15260037

SUPERase•In™ RNase Inhibitor (20 U/μL)	Thermo Fisher Scientific	AM2694
Aqueous Phase 2		
Acrylamide	Fisher BioReagents	BP170-500
Bisacrylamide	Sigma-Aldrich	M729-100G
0.1M DTT	Thermo Fisher Scientific	707265ML
Ammonium Persulfate	Sigma-Aldrich	09913-100G
Sodium dodecyl sulfate (SDS)	Thermo Fisher Scientific	28364
Proteinase K	NEB	P8107S
Oil Phase		
HFE-7500 (3M)	Oakwood Chemical	051243
TEMED	Sigma-Aldrich	1.10732
008-FluoroSurfactant	Ran Biotechnologies	008-FluoroSurfactant-5G
Aqueous Phase 1		
Nuclei Buffer (10 mM Tris-HCl, 10mM NaCl, 3mM MgCl2) or PBS for cells	-	-
OptiPrep	Stem Cell Technologies	7820
Nuclei or Cells	-	-
Post-Droplet Generation		
Mineral Oil	Sigma-Aldrich	69794-500ML

1H,1H,2H,2H-Perfluorooctan-1-ol (PFO)	Synquest Laboratories	2101-3-20
Tris-Tween Buffer: 100 mM Tris-HCl pH=8.0, 0.1% v/v Tween 20		
SYTO™ Green Fluorescent Nucleic Acid Stain (300uM)		
Combinatorial Indexing		
Transposon Annealing Buffer: 400 mM Tris-HCl, 500 mM NaCl	-	-
Diagenode Tagmentase	Diagenode	C01070010-20
2X Tagmentation Buffer: 66mM Tris-Acetate, 132mM K-Acetate, 20mM Mg-Acetate, 32% DMF	-	-
2X T7 DNA Ligase Reaction Buffer	NEB	M0318L
T7 Ligase	NEB	M0318L
0.3% SDS	-	-
10% Triton-X	-	-

Q5® High-Fidelity DNA Polymerase	NEB	M0491L
5-methyl-dCTP	NEB	N0356S
Advantage® UltraPure dNTP Combination Kit	ClonTech	639132
EZ-96 DNA Methylation-Direct MagPrep	Zymo	D5044
Post Bisulfite Conversion		
KAPA HiFi HotStart Uracil+ Kit (250 rxn)	KAPA/Roche	KK2802/07959079001
SPRISELECT, 60ML -	Beckman Coulter	B23318
Kapa Hifi Hotstart ReadyMix, KK2602, 6.25ml	Roche	7958935001
xGen™ Adaptase™ Module 96 rxn	IDT	10009826

Shrimp Alkaline Phosphatase (rSAP)	NEB	M0371L

Tables of Required Oligonucleotide Sequences

Sequence Name	**Sequence (5' to 3')**
Reverse Transcription Primer	/5ACryd//iSp18/AAGCAGTGGTATCAACGCAGAGTNNWNNNS TTTTTTTTTTTTTTTTTTTTTTTTTTTTT **(Order with RNAse Free PAGE Purification)**
Mosaic End Sequence	/5phos/CTGTCTCTTATACACATCT **(Order from Eurofins with HPLC Purification)**
Split Oligo For T7 Ligase	CACGAGACGACAAGT/3ddC/ **(Order with HPLC Purification)**
DNA i7 Primer	CAAGCAGAAGACGGCATACGAGAT[Barcode] GTGACTGGAGTTCAGACGTGTGCTCTT
RNA i7 Primer	CAAGCAGAAGACGGCATACGAGATGTTCCTAAGCGTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTAAGCAGTGGTATCAACGCAGAGT
RNA Linear Amplification Primer	AAGCAGTGGTATCAACGCAGAGT
P5 Primer	AATGATACGGCGACCACCG*A
P7 Primer	CAAGCAGAAGACGGCATACG*A

Combinatorial Indexing Sequences

Tn5 Sequences

Name	Splint Oligo Handle	Barcode	ME Sequence
sciGel_Tn5L3_1	/5Phos/GTCTCGTGGGCTCGG	AGAGTCCTGC	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_2	/5Phos/GTCTCGTGGGCTCGG	GGTCGCATTC	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_3	/5Phos/GTCTCGTGGGCTCGG	ACATGACTGA	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_4	/5Phos/GTCTCGTGGGCTCGG	TTCCTGTCAA	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_5	/5Phos/GTCTCGTGGGCTCGG	CTATTGCATG	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_6	/5Phos/GTCTCGTGGGCTCGG	ATAGGTTCAC	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_7	/5Phos/GTCTCGTGGGCTCGG	GTGCATCGGT	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_8	/5Phos/GTCTCGTGGGCTCGG	CAGATGAACT	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_9	/5Phos/GTCTCGTGGGCTCGG	GCTTAGATGA	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_10	/5Phos/GTCTCGTGGGCTCGG	GACGCATGGA	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_11	/5Phos/GTCTCGTGGGCTCGG	ACCTGCTATT	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_12	/5Phos/GTCTCGTGGGCTCGG	TCATGCGCTT	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_13	/5Phos/GTCTCGTGGGCTCGG	GATTGTGCAT	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_14	/5Phos/GTCTCGTGGGCTCGG	TCCATGCCGA	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_15	/5Phos/GTCTCGTGGGCTCGG	AATGTGCAGG	AGATGTGTATAAGAGACAG
sciGel_Tn5L3_16	/5Phos/GTCTCGTGGGCTCGG	CTCTAGGTGA	AGATGTGTATAAGAGACAG

T7 Ligase Sequences & Barcodes

Name	PCR Handle	Barcode Sequence	Splint Oligo Handle
sciGel_Adapter_1	CAGCACGGCGAGACT	GGACGCCTAA	GACTTGTC
sciGel_Adapter_2	CAGCACGGCGAGACT	CAGTTAGACC	GACTTGTC
sciGel_Adapter_3	CAGCACGGCGAGACT	CCGTCTCAAT	GACTTGTC

sciGel_Adapter_4	CAGCACGGCGAGACT	ATGGTACGTT	GACTTGTC
sciGel_Adapter_5	CAGCACGGCGAGACT	GTAACTGAAC	GACTTGTC
sciGel_Adapter_6	CAGCACGGCGAGACT	TAAGGTTAAC	GACTTGTC
sciGel_Adapter_7	CAGCACGGCGAGACT	CTACTACTCC	GACTTGTC
sciGel_Adapter_8	CAGCACGGCGAGACT	TCTCAACCTG	GACTTGTC
sciGel_Adapter_9	CAGCACGGCGAGACT	GTTATTGGTT	GACTTGTC
sciGel_Adapter_10	CAGCACGGCGAGACT	AATAGGTACC	GACTTGTC
sciGel_Adapter_11	CAGCACGGCGAGACT	TGAGGCAGCT	GACTTGTC
sciGel_Adapter_12	CAGCACGGCGAGACT	TACCAACCAA	GACTTGTC
sciGel_Adapter_13	CAGCACGGCGAGACT	CCGATATCAG	GACTTGTC
sciGel_Adapter_14	CAGCACGGCGAGACT	GTTCCATCAA	GACTTGTC
sciGel_Adapter_15	CAGCACGGCGAGACT	CTTCTGGTCC	GACTTGTC
sciGel_Adapter_16	CAGCACGGCGAGACT	GACCTCAGGT	GACTTGTC
sciGel_Adapter_17	CAGCACGGCGAGACT	CTCATTGCAA	GACTTGTC
sciGel_Adapter_18	CAGCACGGCGAGACT	GTCAGTTAGT	GACTTGTC
sciGel_Adapter_19	CAGCACGGCGAGACT	ACCTCTTACC	GACTTGTC
sciGel_Adapter_20	CAGCACGGCGAGACT	TTGCGATTAC	GACTTGTC
sciGel_Adapter_21	CAGCACGGCGAGACT	TTCATCATAT	GACTTGTC
sciGel_Adapter_22	CAGCACGGCGAGACT	CTTCCGTAGG	GACTTGTC
sciGel_Adapter_23	CAGCACGGCGAGACT	TCGGAGAGTC	GACTTGTC
sciGel_Adapter_24	CAGCACGGCGAGACT	ACGTATCTAT	GACTTGTC
sciGel_Adapter_25	CAGCACGGCGAGACT	TTGCTTCATA	GACTTGTC
sciGel_Adapter_26	CAGCACGGCGAGACT	TCGTCTCTAC	GACTTGTC
sciGel_Adapter_27	CAGCACGGCGAGACT	GTTATGCGAA	GACTTGTC
sciGel_Adapter_28	CAGCACGGCGAGACT	GGCGAATCTA	GACTTGTC
sciGel_Adapter_29	CAGCACGGCGAGACT	CCGCGAAGAA	GACTTGTC
sciGel_Adapter_30	CAGCACGGCGAGACT	AGACCAAGAA	GACTTGTC

sciGel_Adapter_31	CAGCACGGCGAGACT	TAATCTATAC	GACTTGTC
sciGel_Adapter_32	CAGCACGGCGAGACT	AGTCATAGTC	GACTTGTC
sciGel_Adapter_33	CAGCACGGCGAGACT	TTCGCGGAGC	GACTTGTC
sciGel_Adapter_34	CAGCACGGCGAGACT	GAATCGTTCC	GACTTGTC
sciGel_Adapter_35	CAGCACGGCGAGACT	ACGAAGGTAC	GACTTGTC
sciGel_Adapter_36	CAGCACGGCGAGACT	AGTCGCATAA	GACTTGTC
sciGel_Adapter_37	CAGCACGGCGAGACT	ACCAACCGTT	GACTTGTC
sciGel_Adapter_38	CAGCACGGCGAGACT	TTCCTTCTAG	GACTTGTC
sciGel_Adapter_39	CAGCACGGCGAGACT	TCCTCCATAC	GACTTGTC
sciGel_Adapter_40	CAGCACGGCGAGACT	GCTACTTACG	GACTTGTC
sciGel_Adapter_41	CAGCACGGCGAGACT	TTGACGACTA	GACTTGTC
sciGel_Adapter_42	CAGCACGGCGAGACT	TCCATACTAC	GACTTGTC
sciGel_Adapter_43	CAGCACGGCGAGACT	TACGTCCATT	GACTTGTC
sciGel_Adapter_44	CAGCACGGCGAGACT	CAGCGAACGG	GACTTGTC
sciGel_Adapter_45	CAGCACGGCGAGACT	ATATTGACTG	GACTTGTC
sciGel_Adapter_46	CAGCACGGCGAGACT	TCAGTCCGAC	GACTTGTC
sciGel_Adapter_47	CAGCACGGCGAGACT	GCGCATGGAA	GACTTGTC
sciGel_Adapter_48	CAGCACGGCGAGACT	CATGCCGTCC	GACTTGTC
sciGel_Adapter_49	CAGCACGGCGAGACT	ACGTTGCTCC	GACTTGTC
sciGel_Adapter_50	CAGCACGGCGAGACT	AGCTAGGACG	GACTTGTC
sciGel_Adapter_51	CAGCACGGCGAGACT	CTACTAATAT	GACTTGTC
sciGel_Adapter_52	CAGCACGGCGAGACT	AGAAGGAACT	GACTTGTC
sciGel_Adapter_53	CAGCACGGCGAGACT	CCTTGAAGGC	GACTTGTC
sciGel_Adapter_54	CAGCACGGCGAGACT	TGAGGCGTTA	GACTTGTC
sciGel_Adapter_55	CAGCACGGCGAGACT	AGACGTATTC	GACTTGTC
sciGel_Adapter_56	CAGCACGGCGAGACT	AAGGCTCATC	GACTTGTC
sciGel_Adapter_57	CAGCACGGCGAGACT	AGTCTCCGTA	GACTTGTC

sciGel_Adapter_58	CAGCACGGCGAGACT	AATGACCTCT	GACTTGTC
sciGel_Adapter_59	CAGCACGGCGAGACT	TAACTGGCCG	GACTTGTC
sciGel_Adapter_60	CAGCACGGCGAGACT	TTAAGCGCTA	GACTTGTC
sciGel_Adapter_61	CAGCACGGCGAGACT	CATAAGGTTG	GACTTGTC
sciGel_Adapter_62	CAGCACGGCGAGACT	TTCGTCGAAG	GACTTGTC
sciGel_Adapter_63	CAGCACGGCGAGACT	CTCGGACTCT	GACTTGTC
sciGel_Adapter_64	CAGCACGGCGAGACT	CGAACCATAG	GACTTGTC
sciGel_Adapter_65	CAGCACGGCGAGACT	GCCAATAGTT	GACTTGTC
sciGel_Adapter_66	CAGCACGGCGAGACT	ACCTCGCAAG	GACTTGTC
sciGel_Adapter_67	CAGCACGGCGAGACT	ACGAGCCGAA	GACTTGTC
sciGel_Adapter_68	CAGCACGGCGAGACT	CAGACTTGAG	GACTTGTC
sciGel_Adapter_69	CAGCACGGCGAGACT	CAGGCCTAAT	GACTTGTC
sciGel_Adapter_70	CAGCACGGCGAGACT	ACGTTAGCCA	GACTTGTC
sciGel_Adapter_71	CAGCACGGCGAGACT	AATATTAGCT	GACTTGTC
sciGel_Adapter_72	CAGCACGGCGAGACT	GAAGATTCCT	GACTTGTC
sciGel_Adapter_73	CAGCACGGCGAGACT	GTCTGGTCTT	GACTTGTC
sciGel_Adapter_74	CAGCACGGCGAGACT	CGCTTATGGT	GACTTGTC
sciGel_Adapter_75	CAGCACGGCGAGACT	GTAATAAGCA	GACTTGTC
sciGel_Adapter_76	CAGCACGGCGAGACT	AATGCTCTAT	GACTTGTC
sciGel_Adapter_77	CAGCACGGCGAGACT	CAAGATAATT	GACTTGTC
sciGel_Adapter_78	CAGCACGGCGAGACT	CGCGCGGAAC	GACTTGTC
sciGel_Adapter_79	CAGCACGGCGAGACT	CGGACTTCGG	GACTTGTC
sciGel_Adapter_80	CAGCACGGCGAGACT	ATCATGGTAA	GACTTGTC
sciGel_Adapter_81	CAGCACGGCGAGACT	GCCATCTATA	GACTTGTC
sciGel_Adapter_82	CAGCACGGCGAGACT	CCAGATATGC	GACTTGTC
sciGel_Adapter_83	CAGCACGGCGAGACT	CTTCTAGAGT	GACTTGTC
sciGel_Adapter_84	CAGCACGGCGAGACT	CATATTCTTG	GACTTGTC

sciGel_Adapter_85	CAGCACGGCGAGACT	CGCGCAGCAG	GACTTGTC
sciGel_Adapter_86	CAGCACGGCGAGACT	CCGTAATATG	GACTTGTC
sciGel_Adapter_87	CAGCACGGCGAGACT	CATTCCGCCG	GACTTGTC
sciGel_Adapter_88	CAGCACGGCGAGACT	TTCAGAAGCA	GACTTGTC
sciGel_Adapter_89	CAGCACGGCGAGACT	AGAAGAATAC	GACTTGTC
sciGel_Adapter_90	CAGCACGGCGAGACT	ATTACTACTG	GACTTGTC
sciGel_Adapter_91	CAGCACGGCGAGACT	GTCTCCGCGG	GACTTGTC
sciGel_Adapter_92	CAGCACGGCGAGACT	GCCGACGAGC	GACTTGTC
sciGel_Adapter_93	CAGCACGGCGAGACT	TGCCTCTAAG	GACTTGTC
sciGel_Adapter_94	CAGCACGGCGAGACT	CGTCTTGGTC	GACTTGTC
sciGel_Adapter_95	CAGCACGGCGAGACT	ACCATCTGCT	GACTTGTC
sciGel_Adapter_96	CAGCACGGCGAGACT	ATGGTTAATT	GACTTGTC

P5 PCR Barcoded Primers

Name	**P5 Adapter**	**Barcode Sequence**	**PCR Handle**
sciGel_P5_PCR_1	AATGATACGGCGACCACCGAGATCTACAC	CAAGGCATTC	CAGCACGGCGAGACT
sciGel_P5_PCR_2	AATGATACGGCGACCACCGAGATCTACAC	GGCCAGTCCG	CAGCACGGCGAGACT
sciGel_P5_PCR_3	AATGATACGGCGACCACCGAGATCTACAC	CCGCTGCCAG	CAGCACGGCGAGACT
sciGel_P5_PCR_4	AATGATACGGCGACCACCGAGATCTACAC	TGGCTGATGA	CAGCACGGCGAGACT
sciGel_P5_PCR_5	AATGATACGGCGACCACCGAGATCTACAC	TAACTGGTTA	CAGCACGGCGAGACT
sciGel_P5_PCR_6	AATGATACGGCGACCACCGAGATCTACAC	CGAATGAGCT	CAGCACGGCGAGACT
sciGel_P5_PCR_7	AATGATACGGCGACCACCGAGATCTACAC	AAGACCGTTG	CAGCACGGCGAGACT
sciGel_P5_PCR_8	AATGATACGGCGACCACCGAGATCTACAC	TCTGATACCA	CAGCACGGCGAGACT
sciGel_P5_PCR_9	AATGATACGGCGACCACCGAGATCTACAC	CGTAGTTACC	CAGCACGGCGAGACT
sciGel_P5_PCR_10	AATGATACGGCGACCACCGAGATCTACAC	TGGCCTGAAG	CAGCACGGCGAGACT
sciGel_P5_PCR_11	AATGATACGGCGACCACCGAGATCTACAC	GTTGAAGGAT	CAGCACGGCGAGACT
sciGel_P5_PCR_12	AATGATACGGCGACCACCGAGATCTACAC	CATTCAATCA	CAGCACGGCGAGACT
sciGel_P5_PCR_13	AATGATACGGCGACCACCGAGATCTACAC	TCGCTAAGCA	CAGCACGGCGAGACT
sciGel_P5_PCR_14	AATGATACGGCGACCACCGAGATCTACAC	GGCGAACTCG	CAGCACGGCGAGACT
sciGel_P5_PCR_15	AATGATACGGCGACCACCGAGATCTACAC	CAAGATGCCG	CAGCACGGCGAGACT
sciGel_P5_PCR_16	AATGATACGGCGACCACCGAGATCTACAC	ACCTCGTTGA	CAGCACGGCGAGACT
sciGel_P5_PCR_17	AATGATACGGCGACCACCGAGATCTACAC	TTCCTAGACC	CAGCACGGCGAGACT

sciGel_P5_PCR_18	AATGATACGGCGACCACCGAGATCTACAC	CCGTTGACTT	CAGCACGGCGAGACT
sciGel_P5_PCR_19	AATGATACGGCGACCACCGAGATCTACAC	CGATTGGTTA	CAGCACGGCGAGACT
sciGel_P5_PCR_20	AATGATACGGCGACCACCGAGATCTACAC	TCAAGCCGAT	CAGCACGGCGAGACT
sciGel_P5_PCR_21	AATGATACGGCGACCACCGAGATCTACAC	ATTGAGATTG	CAGCACGGCGAGACT
sciGel_P5_PCR_22	AATGATACGGCGACCACCGAGATCTACAC	CAAGCAACTG	CAGCACGGCGAGACT
sciGel_P5_PCR_23	AATGATACGGCGACCACCGAGATCTACAC	AGGTTAGCAT	CAGCACGGCGAGACT
sciGel_P5_PCR_24	AATGATACGGCGACCACCGAGATCTACAC	CGGAGATCCG	CAGCACGGCGAGACT
sciGel_P5_PCR_25	AATGATACGGCGACCACCGAGATCTACAC	GGTCGCGTCA	CAGCACGGCGAGACT
sciGel_P5_PCR_26	AATGATACGGCGACCACCGAGATCTACAC	GTTCGTCAGA	CAGCACGGCGAGACT
sciGel_P5_PCR_27	AATGATACGGCGACCACCGAGATCTACAC	TATCATGATC	CAGCACGGCGAGACT
sciGel_P5_PCR_28	AATGATACGGCGACCACCGAGATCTACAC	TCGTAGAGAA	CAGCACGGCGAGACT
sciGel_P5_PCR_29	AATGATACGGCGACCACCGAGATCTACAC	AACCTGCGTA	CAGCACGGCGAGACT
sciGel_P5_PCR_30	AATGATACGGCGACCACCGAGATCTACAC	CCGATTCGCA	CAGCACGGCGAGACT
sciGel_P5_PCR_31	AATGATACGGCGACCACCGAGATCTACAC	TAACTCTTAG	CAGCACGGCGAGACT
sciGel_P5_PCR_32	AATGATACGGCGACCACCGAGATCTACAC	CAGGTATGGA	CAGCACGGCGAGACT
sciGel_P5_PCR_33	AATGATACGGCGACCACCGAGATCTACAC	GCAGACCGGT	CAGCACGGCGAGACT
sciGel_P5_PCR_34	AATGATACGGCGACCACCGAGATCTACAC	GGAGGTTCTA	CAGCACGGCGAGACT
sciGel_P5_PCR_35	AATGATACGGCGACCACCGAGATCTACAC	CTGATGGTCA	CAGCACGGCGAGACT
sciGel_P5_PCR_36	AATGATACGGCGACCACCGAGATCTACAC	TCCTCGAGTC	CAGCACGGCGAGACT
sciGel_P5_PCR_37	AATGATACGGCGACCACCGAGATCTACAC	CGCCTAATGC	CAGCACGGCGAGACT
sciGel_P5_PCR_38	AATGATACGGCGACCACCGAGATCTACAC	CCATAAGTCC	CAGCACGGCGAGACT
sciGel_P5_PCR_39	AATGATACGGCGACCACCGAGATCTACAC	TGAGAACCAA	CAGCACGGCGAGACT
sciGel_P5_PCR_40	AATGATACGGCGACCACCGAGATCTACAC	ACCGGAATTA	CAGCACGGCGAGACT
sciGel_P5_PCR_41	AATGATACGGCGACCACCGAGATCTACAC	CTTGCAGTAG	CAGCACGGCGAGACT
sciGel_P5_PCR_42	AATGATACGGCGACCACCGAGATCTACAC	TACGGCTACG	CAGCACGGCGAGACT
sciGel_P5_PCR_43	AATGATACGGCGACCACCGAGATCTACAC	ACGAAGTCAA	CAGCACGGCGAGACT
sciGel_P5_PCR_44	AATGATACGGCGACCACCGAGATCTACAC	ATTGCGCTGA	CAGCACGGCGAGACT
sciGel_P5_PCR_45	AATGATACGGCGACCACCGAGATCTACAC	GGTACCATAT	CAGCACGGCGAGACT
sciGel_P5_PCR_46	AATGATACGGCGACCACCGAGATCTACAC	GGTACCGGCA	CAGCACGGCGAGACT
sciGel_P5_PCR_47	AATGATACGGCGACCACCGAGATCTACAC	TGGAAGTACC	CAGCACGGCGAGACT
sciGel_P5_PCR_48	AATGATACGGCGACCACCGAGATCTACAC	TAACTCAATT	CAGCACGGCGAGACT
sciGel_P5_PCR_49	AATGATACGGCGACCACCGAGATCTACAC	CTTGCGCCGC	CAGCACGGCGAGACT
sciGel_P5_PCR_50	AATGATACGGCGACCACCGAGATCTACAC	GGCAGGTATT	CAGCACGGCGAGACT
sciGel_P5_PCR_51	AATGATACGGCGACCACCGAGATCTACAC	GCCGTATGCT	CAGCACGGCGAGACT
sciGel_P5_PCR_52	AATGATACGGCGACCACCGAGATCTACAC	TTACCGAGGC	CAGCACGGCGAGACT

sciGel_P5_PCR_53	AATGATACGGCGACCACCGAGATCTACAC	GCAGGTCCGT	CAGCACGGCGAGACT
sciGel_P5_PCR_54	AATGATACGGCGACCACCGAGATCTACAC	CATCAGAATG	CAGCACGGCGAGACT
sciGel_P5_PCR_55	AATGATACGGCGACCACCGAGATCTACAC	TATAGTAAGC	CAGCACGGCGAGACT
sciGel_P5_PCR_56	AATGATACGGCGACCACCGAGATCTACAC	AACCATTGGA	CAGCACGGCGAGACT
sciGel_P5_PCR_57	AATGATACGGCGACCACCGAGATCTACAC	CAATTACCGT	CAGCACGGCGAGACT
sciGel_P5_PCR_58	AATGATACGGCGACCACCGAGATCTACAC	CATACTCCGA	CAGCACGGCGAGACT
sciGel_P5_PCR_59	AATGATACGGCGACCACCGAGATCTACAC	CCAACTAACC	CAGCACGGCGAGACT
sciGel_P5_PCR_60	AATGATACGGCGACCACCGAGATCTACAC	CGTAATGCAG	CAGCACGGCGAGACT
sciGel_P5_PCR_61	AATGATACGGCGACCACCGAGATCTACAC	CCAGGCCGCA	CAGCACGGCGAGACT
sciGel_P5_PCR_62	AATGATACGGCGACCACCGAGATCTACAC	CGAATAGATG	CAGCACGGCGAGACT
sciGel_P5_PCR_63	AATGATACGGCGACCACCGAGATCTACAC	AATCAGCTGC	CAGCACGGCGAGACT
sciGel_P5_PCR_64	AATGATACGGCGACCACCGAGATCTACAC	CGGAAGATAT	CAGCACGGCGAGACT
sciGel_P5_PCR_65	AATGATACGGCGACCACCGAGATCTACAC	CGCGTACGAC	CAGCACGGCGAGACT
sciGel_P5_PCR_66	AATGATACGGCGACCACCGAGATCTACAC	GAGGCATCAA	CAGCACGGCGAGACT
sciGel_P5_PCR_67	AATGATACGGCGACCACCGAGATCTACAC	CCAGTTCCAA	CAGCACGGCGAGACT
sciGel_P5_PCR_68	AATGATACGGCGACCACCGAGATCTACAC	GCCATTCTCC	CAGCACGGCGAGACT
sciGel_P5_PCR_69	AATGATACGGCGACCACCGAGATCTACAC	AAGAATGGAA	CAGCACGGCGAGACT
sciGel_P5_PCR_70	AATGATACGGCGACCACCGAGATCTACAC	TAACCTTCGG	CAGCACGGCGAGACT
sciGel_P5_PCR_71	AATGATACGGCGACCACCGAGATCTACAC	GCTCAGCCGG	CAGCACGGCGAGACT
sciGel_P5_PCR_72	AATGATACGGCGACCACCGAGATCTACAC	GGCTCCTCGT	CAGCACGGCGAGACT
sciGel_P5_PCR_73	AATGATACGGCGACCACCGAGATCTACAC	AACTGATCTT	CAGCACGGCGAGACT
sciGel_P5_PCR_74	AATGATACGGCGACCACCGAGATCTACAC	CCGTTCGGAT	CAGCACGGCGAGACT
sciGel_P5_PCR_75	AATGATACGGCGACCACCGAGATCTACAC	ACCAGCGCAG	CAGCACGGCGAGACT
sciGel_P5_PCR_76	AATGATACGGCGACCACCGAGATCTACAC	TTCCATGGCA	CAGCACGGCGAGACT
sciGel_P5_PCR_77	AATGATACGGCGACCACCGAGATCTACAC	GCGTTCAGCT	CAGCACGGCGAGACT
sciGel_P5_PCR_78	AATGATACGGCGACCACCGAGATCTACAC	AGAACGTCTC	CAGCACGGCGAGACT
sciGel_P5_PCR_79	AATGATACGGCGACCACCGAGATCTACAC	AAGTAGTCAG	CAGCACGGCGAGACT
sciGel_P5_PCR_80	AATGATACGGCGACCACCGAGATCTACAC	GATATCGGCG	CAGCACGGCGAGACT
sciGel_P5_PCR_81	AATGATACGGCGACCACCGAGATCTACAC	TAACGATCCA	CAGCACGGCGAGACT
sciGel_P5_PCR_82	AATGATACGGCGACCACCGAGATCTACAC	ATTCAGGTAC	CAGCACGGCGAGACT
sciGel_P5_PCR_83	AATGATACGGCGACCACCGAGATCTACAC	TGGAGAATTC	CAGCACGGCGAGACT
sciGel_P5_PCR_84	AATGATACGGCGACCACCGAGATCTACAC	AACCTGGTCT	CAGCACGGCGAGACT
sciGel_P5_PCR_85	AATGATACGGCGACCACCGAGATCTACAC	AAGAAGCTAG	CAGCACGGCGAGACT
sciGel_P5_PCR_86	AATGATACGGCGACCACCGAGATCTACAC	GAAGGTTGCC	CAGCACGGCGAGACT
sciGel_P5_PCR_87	AATGATACGGCGACCACCGAGATCTACAC	TTGCTAACGG	CAGCACGGCGAGACT

sciGel_P5_PCR_88	AATGATACGGCGACCACCGAGATCTACAC	GGCAGACGCC	CAGCACGGCGAGACT
sciGel_P5_PCR_89	AATGATACGGCGACCACCGAGATCTACAC	CGGTTGCGCG	CAGCACGGCGAGACT
sciGel_P5_PCR_90	AATGATACGGCGACCACCGAGATCTACAC	AATTAAGACT	CAGCACGGCGAGACT
sciGel_P5_PCR_91	AATGATACGGCGACCACCGAGATCTACAC	CCGTTCCTTA	CAGCACGGCGAGACT
sciGel_P5_PCR_92	AATGATACGGCGACCACCGAGATCTACAC	TAATGAACGA	CAGCACGGCGAGACT
sciGel_P5_PCR_93	AATGATACGGCGACCACCGAGATCTACAC	AATCTGGAGT	CAGCACGGCGAGACT
sciGel_P5_PCR_94	AATGATACGGCGACCACCGAGATCTACAC	AGATATATCG	CAGCACGGCGAGACT
sciGel_P5_PCR_95	AATGATACGGCGACCACCGAGATCTACAC	AGAGCCAGCC	CAGCACGGCGAGACT
sciGel_P5_PCR_96	AATGATACGGCGACCACCGAGATCTACAC	GGTATCCGCC	CAGCACGGCGAGACT

Single-Cell 3-Level Detailed mDNA/RNA Gel Bead Library Preparation

Microfluidic Device Fabrication

Creation of the Microfluidic Device Mold

The creation of the microfluidic device mold follows some standard SU-8 photolithography and microfabrication techniques. I used the same process previously described in Andrew Richard's thesis. This process wholly occurs inside a clean room as ambient dust particles could interfere with the microfluidic device feature formation. Briefly, 4 inch test grad silicon wafers were carefully rinsed with Piranha solution followed by rinsing in acetone, isopropyl alcohol, and finally DI water. This process is required to remove any organic residues off the wafer to ensure stability of the mold. The wafter was then blow-dried with nitrogen. The wafer was then cleaned by oxygen plasma at 5 sccm O2 with 250 W power for 5 minutes using a PE-Etch 100. Su-8 2025 was then spin coated at 500 RPM for 10 seconds accelerated at 100 RPM/second followed by 3000 rpm for 30 seconds accelerated at 300 RPM/second. The wafer was then soft baked at 65C for 2 minutes followed by 95C for 5 minutes. The wafer was then UV-exposed using an EVG 620 mask aligner with a custom photomask. The wafer was exposed in hard contact mode for 12.3 seconds for a total exposure of 160 mJ/cm^2. The custom

photomask was ordered from a commercial vendor (FrontRange PhotoMask) with 10 micron tolerance, dark field background, and right read (chrome) down. The wafer was then carefully post exposure baked at 65C for 1 minute followed by 95C for 5 minutes. Afterwards, the wafers were developed in SU-8 developer by steady agitation until the features appeared. The wafer was periodically rinsed with isopropyl alcohol to check for the presence of unpolymerized SU-8. Undeveloped SU-8 leaves a clearly white residue on the wafer. Continual exposure to the SU-8 developer will eventually remove all the white residue upon exposure to isopropyl alcohol. Typically, this process took 5 minutes. It's important to not over-develop the mold as features will eventually be removed. After the features are clearly seen and no white residue is detected upon rinsing with isopropyl alcohol, the wafer was blow dried with nitrogen. The wafers were then hard baked at 150C for 5 minutes to increase the thermal stability of the features. The wafers were then silanized using fluoro-octyl-trichloro-silane to allow for PDMS stamping using a vacuum chamber for 30 minutes of vapor deposition.

Creation of the Microfluidic Device

The wafers were then transferred to 15 cm petri dishes and ~80g of PDMS mixed with 10% crosslinker was then cast onto the wafer inside the petri-dish, covering the features of the mold. Roughly 10g of PDMS are then added to two 10 cm dishes, covering the bottom surface. The PDMS was then degassed by placing it inside of a vacuum chamber for 5 minutes, relieving the pressure and popping the bubbles with nitrogen gas, and repeating the process twice. The PDMS coated 10cm dishes and mold was then polymerized at 80C for 1 hour. Using an Exacto knife, two devices were cut from a single mold. Subsequent casting requires much less PDMS (roughly 20g of PDMS with 10% crosslinker) just enough to cover the cut-out devices. The inlets/outlets were individually bored out with a 0.75mm biopsy hole punch. 3M tape was then

placed on the devices and then removed twice to remove PDMS debris from the microfluidic features. Next, the PDMS devices and 10cm dishes were then plasma activated with a PE-Etch 100 by placing the devices on the middle rack exposed to 250 W power with 5 sccm O2 for 15 seconds. The bottom of the device was then quickly bonded to the coated 10 cm dishes after plasma activation and lightly pressed to encourage plasma bonding. It's important to not push with too much pressure as the top and bottom of the microfluidic channels may become bonded together preventing fluid flow. The plasma bonded microfluidic devices were then baked at 80C for 20 minutes to finish the bonding process and ensure stability of the bond.

Hydrophobic Coating of the Microfluidic Device

For droplet formation during microfluidic encapsulation to occur, the microfluidic device must be coated with a hydrophobic coating. Aquapel is first filtered through a 30-micron filter to remove dust and precipitates. Using a P20 pipette, carefully pipette aquapel through each of the devices to uniformly coat all the features and incubate for at least 1 minute. Air was then used to push out the aquapel. This was done with a syringe or lab air valve attached to a pipette tip or microfluidic adapter. The device was then washed once with isopropyl alcohol by similarly pipetting it through each of the channels and then pushed out with air similarly as with the aquapel coating. Finally, the microfluidic devices are then dried in a 55C incubator for 30 minutes. It's important to make sure that most of the isopropyl alcohol is dried out.

Microfluidic Device Set Up

1. Figure 50 illustrates the physical set up designed in house to run the microfluidic device. Briefly, we use in-house air to push fluid in the fluid syringes through the

microfluidic device. An air circuit with air pressure regulators allows for the adjustment of fluid flow through each of the fluid syringes individually. The droplets are collected on a heated bed kept at 55C to allow for cell lysis. Figure 51 describes the fluidic circuit and the way cells or nuclei can be encapsulated. Cell lysis and gel bead polymerization can be visualized under a microscope after an overnight incubation. The DNA contents inside of the gel bead can be stained with a DNA stain such as DAPI.

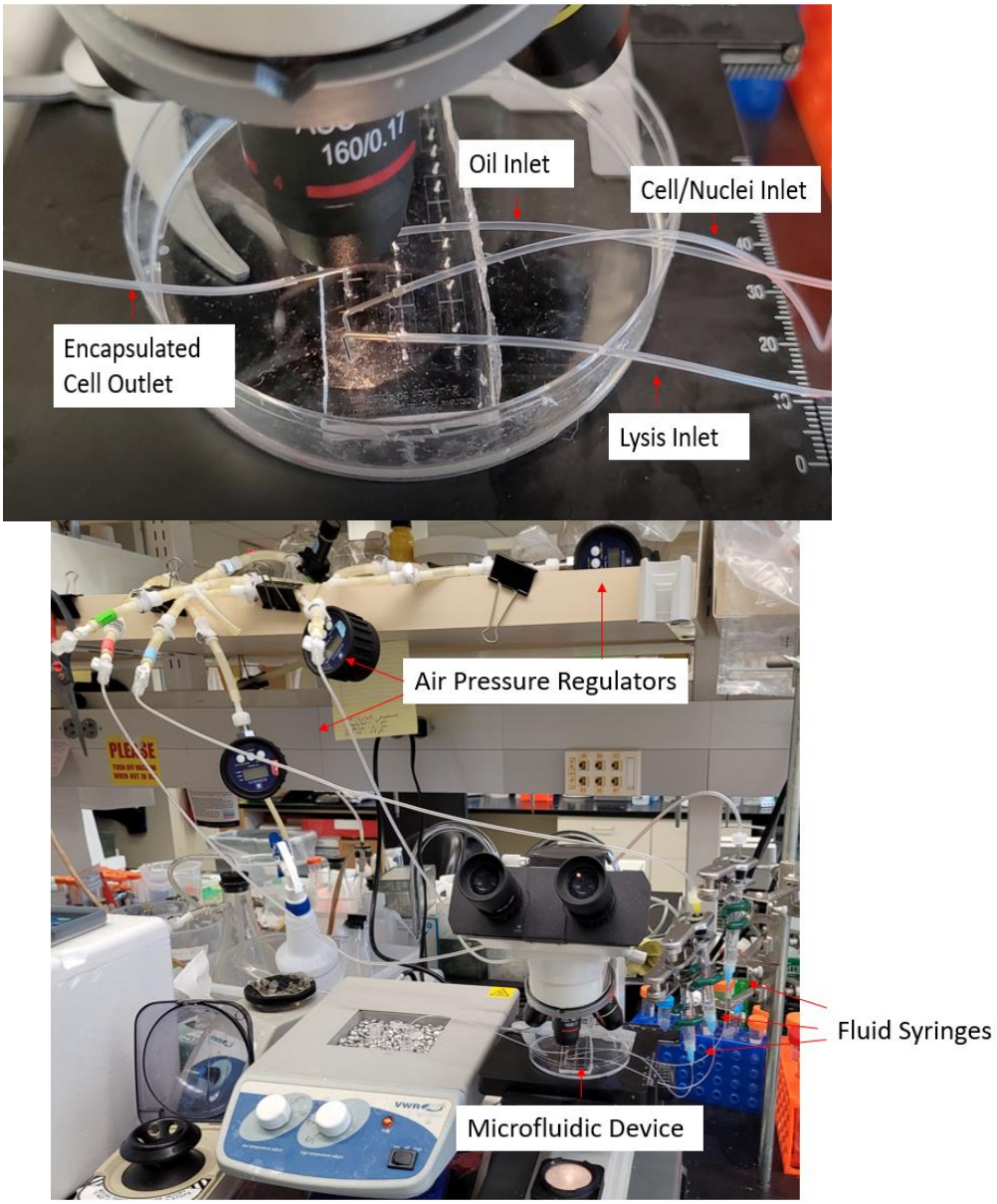


Figure 45: (Top)Cell/nuclei encapsulation bench set up and (Bottom) zoomed in picture of microfluidic device connections.

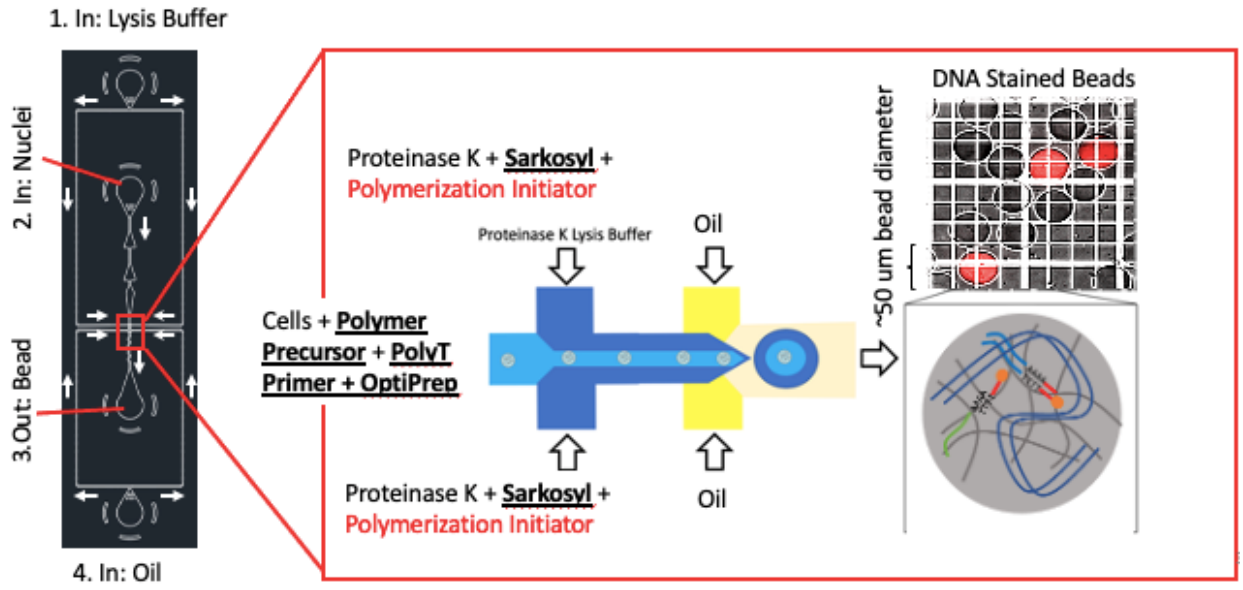


Figure 46: Microfluidic schematic and zoomed in encapsulation scheme of cells or nuclei with lysis buffers and polymer precursors. Polyacrylamide gel beads can then be stained with a DNA stain, DAPI, and imaged.

Cell Encapsulation

1. Trypsonize cells and wash once with 1X PBS by pelleting cells at 300xg for 00:04:00
2. Resuspend cells in 3000 cells/uL in encapsulation buffer: 1X PBS, 40% OptiPrep, 0.75% BSA, 5µM reverse transcription primer, 1% v/v SUPERase RNAseInhibitor
3. Create polyacrylamide buffer:

Polyacrylamide Buffer	**Volume (µL)/Mass**	**Final Concentration**
4M Tris-HCl pH=7.5	175	70mM
Acrylamide	4.5g	45% w/v
Bisacrylamide	44mg	0.9% w/v
H2O	9825	

4. Create lysis buffer

Aqueous Phase 2 Solution	**Volume (uL)**	**Final Concentration**
Polyacrylaide Buffer	447.5	-

20% w/v Sarkosyl	5	0.2%
0.1M DTT	15	3mM
Igepal	5	1%
Ammonium Persulfate (20% w/v)	12.5	0.5%
Proteinase K (0.8U/uL)	10	0.016U/uL
Total	500	-

5. Create oil solution

Oil Solution	**Volume (uL)**	**Final Concentration**
HFE-7500	400	-
008-FluoroSurfactant 20% w/v in HFE-7500	100	2% w/v
TEMED	2	0.4% v/v
Total	502	

6. Turn on the microscope and place the microfluidic device on the microscope stage

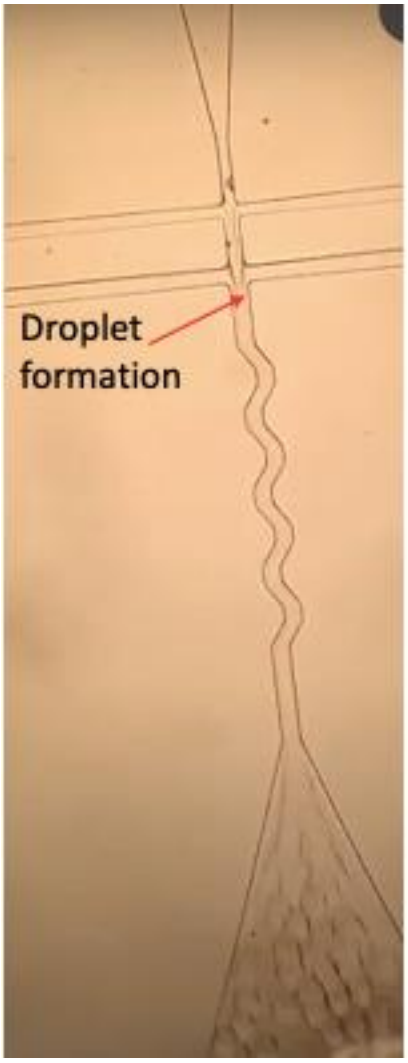


Figure 47: Image of encapsulation taking place in the microfluidic device. Droplet formation occurs right after the oil junction.

7. Assemble the fluidic circuit with 3 syringes connected to the 3 inlets using the tubing and the right-angle couplers. Add a right-angle couple to the outlet and attach tubing to direct the outflow to a 1.5 mL tube containing 150 uL of mineral oil.
8. Set the fluidic pressures to the following settings:
    a. Cell input: 1.2 psi
    b. Lysis input: 1.4 psi

c. Oil input: 1.5 psi

9. Open the pressure valves in the following order waiting one second before opening the next valve: Cell input, lysis input, and oil input
10. Carefully adjust the fluidic pressures to match the fluid profile illustrated in Figure 52.
11. It's crucial that the gel bead size is the same size or larger as the outlet channel otherwise the cells will not be encapsulated
12. After collection is complete, incubate the tube in 55C for 30 minutes and then let the gel polymerize overnight at room temperature

Droplet Breakage

1. Using a pipette, remove the upper mineral oil layer and the lower HFE-7500 layer
2. Add 600 uL of 6X SSC and 150 uL of PFO and vortex the beads briefly to break the gel beads out of the emulsion on ice
3. Centrifuge 300g for 2 minutes at 4C to pellet the beads and remove the top and bottom layers leaving the gel beads in the middle on ice
4. Add another 5 mL of 6X SSC and remove the top and bottom layers leaving the gel beads in the middle and filter through a 100 micron filter
5. Wash once with 5X Reverse Transcription Buffer

cDNA Synthesis

1. Set up reverse transcription reaction:

Reverse Transcription Reaction	**Volume (uL)**	**Final Concentration**

1X Reverse Transcription Buffer	225	1X
Maxima H (200U/uL)	37.5	10U/μL
10mM dNTPs	37.5	0.5μM
Encapsulated Beads	450	-
Total	750	

2. Incubate under rotation for 30 minutes at room temperature and then incubate the reaction at 42C for 60 minutes.
3. Finish the reaction by incubating tubes under rotation at 50C for 60 minutes
4. Add 750 uL of binding buffer and incubate for 5 minutes at room temperature to denature enzymes. Then add 1.5 uL of tween-20 and mix well to prevent beads from sticking to the edge of the tube
5. Then wash twice with Tris-Tween buffer and twice with PBS (for secondary strand synthesis protocol) or tagmentation buffer (no DMF) (for hybrid Tn5 protocol)
6. Set up secondary strand synthesis reaction on ice:

Secondary Strand Synthesis Reaction	**Volume (uL)**	**Final Concentration**
Secondary Strand Synthesis Buffer	80	1X
Secondary Strand Synthesis Enzyme Mix	40	1X
Beads in PBS	680	-
Total	800	

7. Incubate overnight at 16C
8. Add 750 uL of binding buffer and incubate 5 minutes at room temperature to denature enzymes. Then add 1.5 uL of tween-20 and mix well to prevent beads from sticking to the edge of the tube
9. Then wash twice with Tris-tween buffer and wash twice with H2O

Combinatorial Indexing

1. Anneal transposons and mosaic end sequences by setting up the following reaction:

Transposon Annealing Reaction	**Volume (uL)**	**Final Concentration**
Transposon annealing buffer	1	1X
Mosaic End (100μM)	1	10μM
Indexed Transposon (100 μM)	1	10μM
H2O	7	
Total	10	

2. Anneal the transposons with the ramp down protocol:

Transposon Annealing Reaction	**Temperature**	**Time**
1	95C	5:00
2	Cool to 65C	-0.1C/s
3	65C	5:00
4	Cool to 4C	-0.1C/s

3. Set up the transposase reaction:

Transposase Loading	**Volume (uL)**	**Final Concentration**
Annealed Transposons	2	10μM
Unloaded Transposase (0.4mg/mL)	2	0.2mg/mL
Total	4	

4. Load the transposase for 30 minutes at 23C in a thermocycler then add 2 uL of 100% Glycerol
5. Set up multiplexed tagmentation reaction and mix well adding the gel beads last:

Tagmentation Reaction	**Volume (uL)**	**Final Concentration**
Loaded indexed transposons (0.025 mg/mL diluted in 1X tagmentation buffer)	2	0.0025mg/mL
Encapsulated cells (30 cells/uL)	8	-
2X Tagmentation Buffer	10	1X Tagmentation Buffer
Total	20	

6. Incubate samples at 55C shaking at 600 RPM for 90 minutes. Then add 200 uL of Tris-Tween buffer to stop the reaction.
7. Pool all reactions and pellet beads 300g for 2 minutes.
8. Wash with Tris-Tween 20 buffer twice and then wash with H2O twice

9. Set up ligation multiplexed ligation reaction and mix well adding the gel beads last:

Ligation Reaction	**Volume (uL)**	**Final Concentration**
2X T7 Ligation Buffer	10	1X
T7 Ligase (3U/uL)	2.5	0.375U/uL
Splint Oligo (100 μM)	1.8	9μM
Indexed adapter (50 μM)	1.2	3μM
Tagmented Gel Beads	4.5	-

10. Incubate at 25C shaking at 600 RPM for 60 minutes and then heat inactivate the ligase with 65C for 10 minutes
11. Add 150 uL of Tris-Tween buffer and then pool all reactions
12. Wash twice in Tris-Tween buffer and then wash twice in H2O
13. Adjust the bead concentration to 10 cells/uL
14. Split into the final barcoding plate and denature the Tn5:

Tn5 Denaturation	**Volume (uL)**	**Final Concentration**
0.3% SDS	2	0.1%
Gel beads (10 cells/uL)	4	6.6 cells/uL
Total	6	

15. Vortex samples to mix well and incubate 55C for 15 minutes and then add 1.5 uL of 10% Triton-X to quench the SDS.
16. Incubate samples for 55C for 15 minutes and then set up gap filling reaction (change to 20 uL for RNA:

Tn5 Gapfilling Reaction	**Volume (uL)**	**Final Concentration**
5X Q5 Buffer	8	1X

dNTPs replacing the dCTP with dmCTP (7.6 mM)	1.3	0.25mM
Q5 Polymerase (2U/uL)	0.4	0.2U/uL
RNA Linear Amplification Primer (10μM)	2	0.5μM
SDS quenched samples	7.5	
H2O	22.8	
Total	40	

17. Incubate 72C for 10 minutes to run the gap filling reaction

18. Linearly amplify the RNA for 10 cycles to extract out the RNA from the gel beads:

Step	**Temperature**	**Time**
1	72C	10:00
2	98C	3:00
3	98C	0:20
4	59C	1:00
5	72C	2:00
6	GOTO Step 3 x 9	

19. Pellet beads and carefully take 20 uL of the supernatant of the PCR reaction containing the cDNA.

20. Add 1 μL of indexed i5 primer and exponentially amplify the RNA for 7 cycles (fill in later):

Step	**Temperature**	**Time**

1	98C	0:10
2	59C	0:30
3	72C	1:00
4	GOTO Step 1 x 6	

21. Follow the EZ-96 DNA Methylation-Direct MagPrep instructions and add 130 uL of bisuflite conversion reagent to the remaining 20 uL of the gap filled reaction to bisulfite convert the DNA library
22. Perform bisulfite conversion and desulphonation according to the instructions on EZ-96 DNA Methylation-Direct MagPrep. During the elution step, add 20 uL of H2O and mix well
23. Take the whole volume including the magnetic beads and transfer each reaction to a new 96 well plate.

Post Bisulfite Conversion Processing

1. Set up the final barcoding linear amplification for the methylated DNA library:

Linear Barcoding Reaction	**Volume (uL)**	**Final Concentration**
2X KAPA HiFi U+ Master Mix	25	1X
10 μm indexed P5 primer	2.5	0.5μM
Bisulfite Converted Beads	20	-
Total	50	

2. Perform 20 cycles of linear amplification with the following PCR conditions:

Step	**Temperature**	**Time**

1	98C	3:00
2	98C	0:20
3	59C	1:00
4	72C	2:00
5	GOTO Step 2 x 19	

3. Add 2.5 uL of rSAP and incubate all samples at 37C for 30 minutes and heat inactivate 65C for 5 minutes
4. Using a magnetic rack, remove the supernatant from the PCR reactions into a new PCR plate.
5. Perform 1.2X bead purification on all samples and elute 10 uL in a new 96 well plate
6. To set up the adaptase reaction, first incubate the plate at 95C for 3 minutes and then immediately place the plate on ice for 2 minutes.
7. As quickly as possible, set up the adaptase reaction for all samples by adding:

Adaptase Reaction	**Volume (uL)**	**Final Concentration**
Buffer G1	2	-
Reagent G2	2	-
Reagent G3	1.25	-
Enzyme G4	0.5	-
Enzyme G5	0.5	-
Low EDTA TE	4.25	-
Bead purified samples	10	-
Total	20.5	

8. Run the adaptase reaction conditions by first incubating all samples at 37C for 30 minutes followed by heat denaturation of the enzyme by incubating the samples at 95C for 2 minutes.
9. Set up the final qPCR reaction:

Final DNA PCR	**Volume (uL)**	**Final Concentration**
2X KAPA HiFi Master Mix	25	1X
10 μM P5 primer	3	0.6 μM
10 μM DNA i7 primer	5	1 μM
Adaptase reaction	20	-
SYBR Green 100X	0.2	1X
Total	50	

10. Amplify with the following PCR conditions (fill in details later) for 10 cycles:

Step	**Temperature**	**Time**
1	98C	3:00
2	98C	0:10
3	59C	0:30
4	72C	1:00
5	GOTO Step 2 x 9	

11. Bead purify with 0.8X SPRI bead ratio the individually exponentially amplified RNA library and set up the final indexed i7 PCR:

Final RNA PCR	**Volume (uL)**	**Final Concentration**
2X KAPA HiFi Master Mix	25	1X

10 μM P5 primer	2.5	0.6 μM
10 μM RNA i7 primer	2.5	1 μM
RNA exponentially amplified reaction	20	-
SYBR Green 100X	0.5	1X
Total	50	

12. Similarly, bead purify the DNA library with 0.8X SPRI bead ratio individually and set up the final PCR:

Final RNA PCR	**Volume (uL)**	**Final Concentration**
2X KAPA HiFi Master Mix	25	1X
10 μM P5 primer	2.5	0.6 μM
10 μM P7 primer	2.5	1 μM
RNA exponentially amplified reaction	20	-
SYBR Green 100X	0.5	1X
Total	50	

13. Amplify with the following PCR conditions for ~12 cycles right before PCR saturation

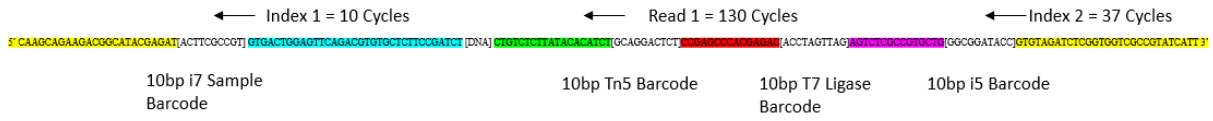
Step	**Temperature**	**Time**
1	98C	3:00
2	98C	0:10
3	59C	0:30

4	72C	1:00
5	GOTO Step 2 x 12	

14. Pool the individual DNA and RNA reactions separately and perform 2 0.8X SPRI bead ratio purifications to clean up the library for sequencing.
15. Methylated DNA libraries were sequenced 130 cycles read 1, 10 cycles index 1, 37 cycles index 2, and 100 cycles read 2. Typically, this library was sequenced to a depth of 500,000 reads/cell
16. RNA libraries were sequenced 130 cycles read 1, 10 cycles index 1, 37 cycles index 2, and 40 cycles read 2. Typically, this library was sequenced to a depth of 10,000 reads/cell

Final Library Structure and Sequencing Scheme

DNA Library Structure and Sequencing Scheme



RNA Library Structure and Sequencing Scheme

Bioinformatic Methods

*A github repository link will be made available here containing the pre-processing and primary analysis modules

Pre-Processing

Libraries were first demultiplexed using index 1 used to distinguish RNA libraries from DNA ones using bcl2fastq. The ligation barcode located in the last 10 bases of the index 2 read was then extracted. Configuration files and barcode lists were assembled according to the formatting required by deindexer. Deindexer was then used to demultiplex the DNA reads and RNA reads by the ligation barcode. In addition, the index 2 read was demultiplexed by deindexer. Both the DNA and RNA reads were then concatenated into a single file but keeping the read ID of each read was edited to the following notation: @xx. Where xx is the ligation barcode number that the read was demultiplexed with. The Tn5 barcode located in the first 10 bases of read 1 were then extracted followed by the PCR barcode located in the last 10 bases of index 2 for both the DNA and RNA libraries. Deindexer was then used to demultiplex the DNA reads and RNA reads by both the Tn5 barcode and PCR barcode. Both the DNA and RNA reads were then concatenated into a single file but keeping the read ID of each read was edited to the following notation: @xx.yy.zz Where xx is the ligation barcode number, yy is the Tn5 barcode, and zz is the PCR barcode. The RNA library was then filtered for the correct construct by looking for a "TTTT" sequence in the 32-36 positions in read 2. In addition, the UMI was extracted from the positions 23-30 in read 2 and the read ID of read 1 was edited to the format: @!xx.yy.zz#UMI. This read ID matches the format required for downstream analyses using the dropEst package. Both the read 1 DNA and RNA libraries were then trimmed for the Tn5 adapter, adaptase adapter, and polyT sequences using cutadapt. An additional 10 bases from the DNA library are trimmed as this is artificially methylated during the gap filling steps. The DNA reads were mapped with the bsbolt package which is a BWA-MEM wrapper for bisulfite converted sequence mapping using the PBAT. In addition, the DNA reads were mapped with

bismark which is a bowtie2 wrapper for bisulfite converted sequence mapping using the PBAT settings. The RNA reads were mapped with STAR. Both DNA and RNA libraries are filtered for high quality reads. The RNA reads were then input into the dropEst package which performs UMI collapse and creates a counts matrix for secondary analysis. The highly methylated reads in the DNA libraries were removed using a G to A conversion cutoff to remove cDNA reads that are artificially methylated prior to bisulfite conversion. The duplicate reads in the DNA library were then removed. Figure 59 illustrates the pre-processing pipeline.

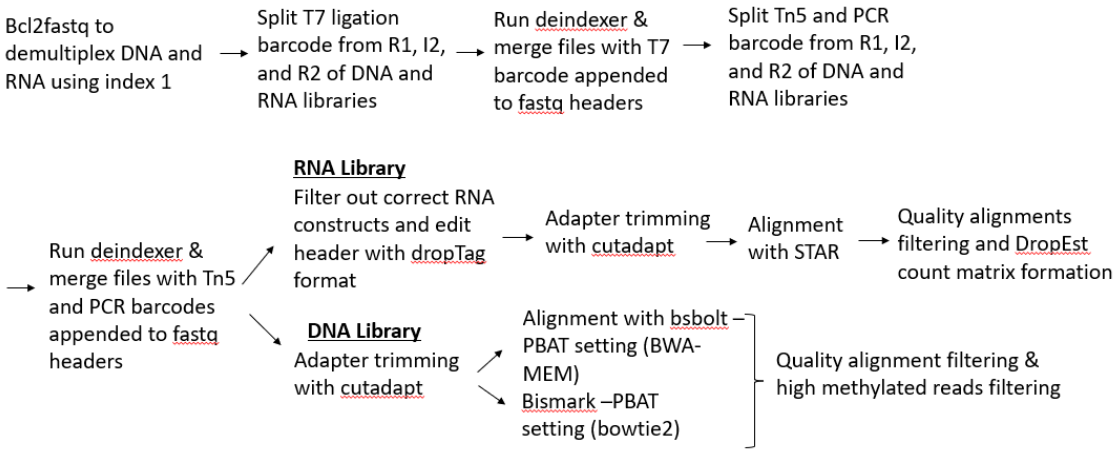


Figure 48: Primary analysis pipeline

Primary Analysis Pipeline

A python dictionary was created to organize all the important sequencing information by cell barcode. The global CG and CH methylation information is encoded in the bismark alignment files under the field "XM". Methylated CH is encoded as H while unmethylated CH is encoded as h. Similarly, methylated CG is encoded as Z while unmethylated CG is encoded as z. These encodings were quantified per cell barcode in the creation of the database. Figure 60 summarizes the database structure

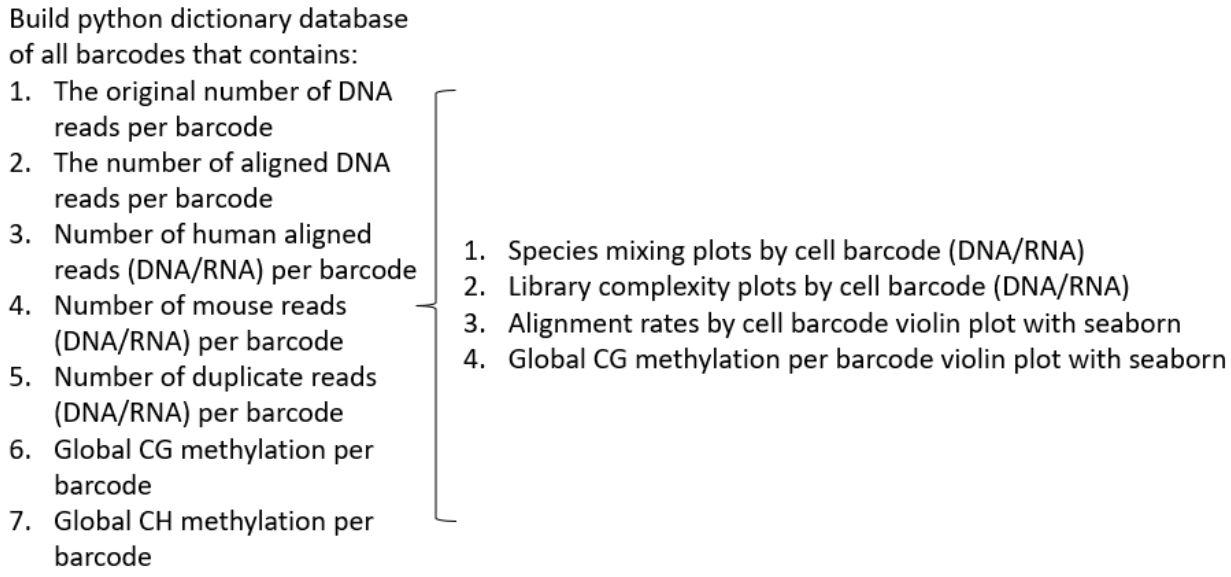


Figure 49: Database structure of libraries used to create sequencing statistic plots

Secondary Analysis Pipeline

Here, we describe the creation of the DNA methylome plots. After running the primary processing pipeline, the alignment files were demultiplexed to single cell barcodes (or kept in bulk). The CG methylation status for each genomic coordinate was extracted with methylpy for each barcode. The resulting tabulated file for each barcode contains the genomic coordinate, methylation status, and read depth at that cytosine position. We then created a 1Mb genomic windows x cell barcode methylation frequency matrix where in each window, we calculated the proportion of methylated CG positions to the total number of CG positions. We created a similar matrix using CH positions. We then calculated the average number of CG and CH positions in each bin per cell and plotted it with the Seaborn python package. To validate the CG methylation across H3K4Me3 features, HCT116 H3K4Me3 ChIP seq coordinates were first downloaded. The coordinates were extended 5000 bases upstream and 5000 bases downstream. Windows were then created in a quartile manner where 4 even windows were created 5000 bases upstream, within the feature coordinates, and 5000 bases downstream. To increase coverage of these

features, the analysis was performed at a bulk level. A similar methylation frequency matrix as previously described was created using these windows and the proportion of methylated CG positions to the total number of CG positions. The average methylation at each quartile for all H3K4Me3 coordinates was then calculated and plotted.

The RNA alignment files were first coordinate sorted and duplicate reads were removed. The htseq software was used to create an RNA gene x sample counts matrix using htseq-count. This counts matrix contained the bulked RNA counts of encapsulated HCT116, RNA counts from an HCT116 in-tube control, and RNA counts from a U87 in-tube control all created by our RNA-seq protocol. The analysis was performed at the bulk level to increase gene coverage. The counts matrices were then input into scanpy where the counts were log normalized and converted to counts per million. We then plotted the log normalized RNA counts of each sample pair-wise and marker genes obtained from literature of each cell type were labeled. At the single cell level, the dropEst counts matrix was input into Seurat. Using Seurat, we filtered barcodes with gene counts < 200 and >1000 (potential doublets). The counts matrix was then similarly log normalized. Further analysis such as clustering and cell type identification follows previously published methods using Seurat.