Microfluidics for Transcriptomics and Epigenomics

Mimosa Sarma

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemical Engineering

Chang Lu, Chair

Abby R. Whittington

Donald G. Baird

Rafael V. Davalos

Richey M. Davis

May 9, 2019

Blacksburg, Virginia

Keywords: microfluidics, single-cell RNA sequencing, immunology, chromatin immunoprecipitation, next-gen sequencing, protein production, protein purification

Microfluidics for Transcriptomics and Epigenomics

Mimosa Sarma

ABSTRACT

A cell, the building block of all life, stores a plethora of information in its genome, epigenome, and transcriptome which needs to be analyzed via various *Omic* studies. The heterogeneity in a seemingly similar group of cells is an important factor to consider and it could lead us to better understand processes such as cancer development and resistance to treatment, fetal development, and immune response. There is an ever growing demand to be able to develop more sensitive, accurate and robust ways to study Omic information and to analyze subtle biological variation between samples even with limited starting material obtained from a single cell. Microfluidics has opened up new and exciting possibilities that have revolutionized how we study and manipulate the contents of the cell like the DNA, RNA, proteins, etc. Microfluidics in conjunction with Next Gen Sequencing has provided ground-breaking capabilities for handling small sample volumes and has also provided scope for automation and multiplexing. In this thesis, we discuss a number of platforms for developing low-input or single cell *Omic* technologies. The first part talks about the development of a novel microfluidic platform to carry out single-cell RNA-sequencing in a one-pot method with a diffusion-based reagent swapping scheme. This platform helps to overcome the limitations of conventional microfluidic RNA seq methods reported in literature that use complicated multiple-chambered devices. It also provides good quality data that is comparable to state-of-the-art scRNA-seq methods while implementing a simpler device design that permits multiplexing. The second part talks about studying the transcriptome of innate leukocytes treated with varying levels of LPS and using RNA-seq to observe how innate immune cells undergo epigenetic reprogramming to develop phenotypes of memory cells. The third part discusses a low-cost alternative to produce tn5 enzyme which low-cost NGS studies. And finally, we discuss a microfluidic approach to carrying out low-input epigenomic studies for studying transcription factors. Today, single-cell or low-input *Omic* studies are rapidly moving into the clinical setting to enable studies of patient samples for personalized medicine. Our approaches and platforms will no doubt be important for transcriptomic and epigenomic studies of scarce cell samples under such settings.

Microfluidics for Transcriptomics and Epigenomics

Mimosa Sarma

GENERAL AUDIENCE ABSTRACT

This is the era of personalized medicine which means that we are no longer looking at one-size-fits-all therapies. We are rather focused on finding therapies that are tailor-made to every individual's personal needs. This has become more and more essential in the context of serious diseases like cancer where therapies have a lot of side-effects. To provide tailor-made therapy to patients, it is important to know how each patient is different from another. This difference can be found from studying how the individual is unique or different at the cellular level i.e. by looking into the contents of the cell like DNA, RNA, and chromatin. In this thesis, we discussed a number of projects which we can contribute to advancement in this field of personalized medicine. Our first project, MID-RNA-seq offers a new platform for studying the information contained in the RNA of a single cell. This platform has enough potential to be scaled up and automated into an excellent platform for studying the RNA of rare or limited patient samples. The second project discussed in this thesis involves studying the RNA of innate immune cells which defend our bodies against pathogens. The RNA data that we have unearthed in this project provides an immense scope for understanding innate immunity. This data provides our biologist collaborators the scope to test various pathways in innate immune cells and their roles in innate immune modulation. Our third project discusses a method to produce an enzyme called 'Tn5' which is necessary for

studying the sequence of DNA. This enzyme which is commercially available has a very high cost associated with it but because we produced it in the lab, we were able to greatly reduce costs. The fourth project discussed involves the study of chromatin structure in cells and enables us to understand how our lifestyle choices change the expression or repression of genes in the cell, a study called epigenetics. The findings of this study would enable us to study epigenomic profiles from limited patient samples. Overall, our projects have enabled us to understand the information from cells especially when we have limited cell numbers. Once we have all this information we can compare how each patient is different from others. The future brings us closer to putting this into clinical practice and assigning different therapies to patients based on such data.

Acknowledgment

I would like to express my heartfelt gratitude to the many individuals who have helped me on this amazing Ph.D. journey right from my childhood to this day. As they say, Rome was not built in a day – so every person I met on the way – contributed a brick to the house. I would like to send a special thanks to my Ph.D. advisor Dr. Chang Lu for always being the encouraging voice in my head and always pushing me to achieve what I thought I couldn't achieve and making me realize that limitations only exist in the head. I would like to say thanks to my entire graduate committee: Dr. Abby Whittington, Dr. Donald Baird, Dr. Rafael Davalos, Dr. Richey Davis for your help, guidance, support and availability throughout this journey. A big thank you to Dr. Song Li and his student Jiyoung Lee for excellent collaboration with providing help and support with the bioinformatics aspect of this work. Thank you, Dr. Liwu Li, for introducing me to the field of innate immunity and for being a collaborator with an amazing disposition. Thank you to my colleagues, Sai Ma (for helping conceive the idea of MID-RNA-seq), Travis Murphy (for setting up a ChIP-seq data analysis program), Yuan-Pang Hsieh & Jerry Liu (for help with TF MOW-ChIP), Dr. Yan Zhu (for ERα data analysis), Chengyu Deng(for help with MOW-ChIP and tn5 production), Bohan Zhu & Qiang Zhang (for general help and support in the lab), Dr. Chen Sun, Dr. Zhenning Cao, Lynette Naler, and Zirui Zhou for providing a positive and affirming environment in the lab so going to work every day was fun and not stressful.

I would like to dedicate this thesis to my parents, Dr. Sailendra K. Sarma and Dr. Nandita Devi who, as doctors, inspired me to work in life sciences research from an early age. I am also immensely grateful that they believed in the ability of a woman in a patriarchal society. Thanks to their inspiration I could be the third ‘Dr.’ in the family. A big thank you to my sister, Sukanya, who has been unconditionally along my side on this journey of life and always has been someone I could talk to. To my teacher from my hometown of Tinsukia, Triveni, who at an early age told me not to be afraid to dream big well beyond the confines of my small town, I am truly grateful.

Thank you to my amazing mentors at Virginia Tech, Dr. Maria del Carmen Caña Jimenez and Dr. Vinodh Venkatesh for playing an important role in my journey at Tech. I would also like send in a thank you to the unending list of friends in Blacksburg and elsewhere whose names will exhaust this acknowledgment, but without whom I truly couldn’t imagine my life as I know it. Muchas Gracias to Roberto Pretelt and Sofia Caceres, my in-laws, for being an absolute blessing in my life and being my second set of parents, for loving me, taking care of me, making me laugh and teaching me how to enjoy and appreciate life. And saving the best for the last, I would like to dump a huge ton of gratitude on my partner, Juan Antonio Pretelt for supporting me all throughout my Ph.D. and pulling me up whenever I wanted to give up. Right from driving me with my samples to different labs across campus and giving me late night rides to the lab, to sitting with me while I write my thesis, you have made an unmeasurable contribution to this thesis and this work.

Table of contents

List of Figures

List of Tables

List of Abbreviations

RNA: Ribonucleic Acid

DNA: Deoxyribonucleic Acid

NGS: Next Generation Sequencing

EST: Expressed Sequence Tags

SAGE: Serial Analysis of Gene Expression

cDNA: Complementary DNA

scRNA-seq: Single Cell RNA Sequencing

dNTP: deoxyribonucleotide triphosphate

UMI: Unique Molecular Identifier

IFC: Integrated Microfluidic Chip

FPKM: Fragments Per Kilobase of transcript per Million mapped reads

LPS: Lipopolysaccharides

TLR4: Toll Like Receptor 4

NET: Neutrophil extracellular traps

IRF5: Interferon Regulatory Factor 5

M-CSF: Macrophage colony-stimulating factor

RA: Rheumatoid Arthritis

PIC: Protease Inhibitor Cocktail

mRNA: messenger RNA

TSS: Transcription Start Site

TF: Transcription Factor

ChIP-seq: Chromatin Immunoprecipitation Sequencing

MOW-ChIP: Microfluidic Oscillatory Washing ChIP

ATAC-seq: Assay for Transposase-Accessible Chromatin

ER α: Estrogen Receptor Alpha

IP: Immunoprecipitation

DSG: Disuccinimidyl Glutarate

FBS: Fetal Bovine Serum

ROS: Reactive Oxygen Species

PS: Penicillin Streptomycin

Chapter 1: Overview of single-cell transcriptomics

1.1 Background importance of transcriptomics

Ribonucleic acid or RNA is a major biological macromolecule (in addition to DNA and protein). The central dogma of molecular biology states that in most living organisms (except some retroviruses) genetic information follows the path of DNA->RNA->Protein. Proteins play important roles in the cell such as catalyzing reactions as enzymes, building tissue blocks, signaling other cells etc. In order to produce a certain protein in a cell, the gene corresponding to that protein is activated and then converted into multiple copies of messenger RNA, or mRNA. This process is called transcription. The mRNA is then used as a template to read groups of the three-letter genetic code to form proteins with the help of ribosomes.

It is evident that the study of the RNA molecules present in a cell (or the transcriptome of a cell) is of utmost importance and is an indication of what genes are active (or are being transcribed) in the cell at that time.[1] This study is called transcriptomics. The main aims of transcriptomics are: to catalog and understand all classes of transcripts, for example by studying their 5′ and 3′ ends, studying alternative splicing events and any modifications to the RNA structure post-transcription; and to measure the change in expression levels of each transcript under different conditions.[1]

Transcriptomics has been around since the early 1990s.[2] An early high throughput method to study the transcriptome was Serial analysis of Gene Expression (or SAGE) which was developed in 1995.[2] Transcript fragments which were generated via Sanger Sequencing were matched to known genes to quantify each transcript[2]. It was an advancement of an earlier method called EST or Expressed Sequence Tags.

The SAGE method produced more information than was possible when sequencing the ESTs, but the preparation of samples for sequencing and analysis of data generated involved a lot of hand on labor and the throughput was still very low.[2] This method was soon phased out and taken over by hybridization microarrays. Microarrays worked on the principle of hybridization between two fluorescently labeled complementary DNA strands by hydrogen bonds formation between complementary nucleotide base pairs. The abundance of a particular transcript was directly proportional to the fluorescence intensity of that particular probe sequence. It was a potent way to study large-scale gene-expression. This technology however came with several limitations like 1) the complementary sequences of the genes being studied had to be known *a priori* 2) there would be chances of cross-hybridization between genes having complementary sequences 3) only highly expressed genes could be studied 4) the exact length and sequence of the RNAs being analyzed was not known 5) it was unable to detect novel transcripts. [3]

Because of these technical limitations, transcriptomics moved on to sequencing-based methods. Briefly, sequencing refers to methods to determine the exact order of nucleotides - Adenine (A), Guanine (G), Thymine (T) or Cytosine(C) in DNA or RNA. The first sequencing technologies used were Sanger sequencing of Expressed Sequence Tag libraries and that has evolved over the years into the current state of the art technology which is called Next-Gen Sequencing of cDNA (this technology is called RNA-Seq).

In conjunction with the development of several high-throughput sequencing technologies, RNA sequencing technology also developed alongside. It had the power to dissect the phenotypic heterogeneity of cells to and can deepen our understanding of the underlying mechanisms governing both health and disease. Over just the last decade, transcriptomics has been profoundly used for the understanding of complex biological systems, giving rise to exceptional understandings in the fields of immunology, neurobiology, and cancer biology.[4]

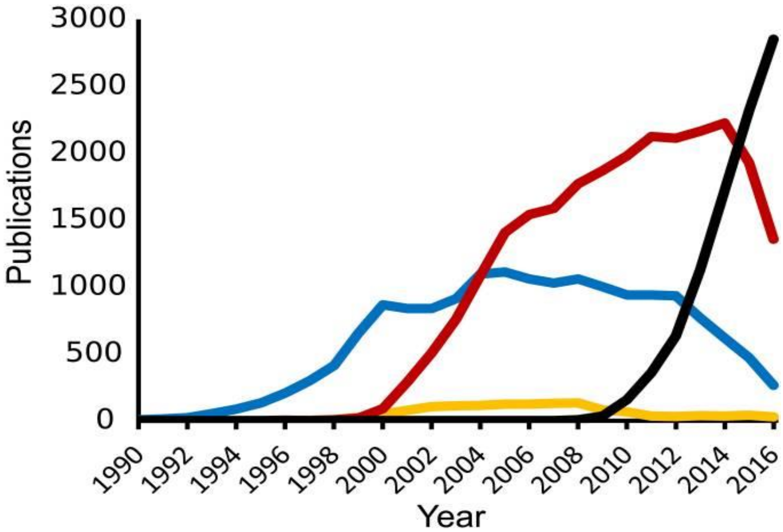

Figure 1: Number of publications involving different transcriptomic methods viz. RNA sequencing (black), RNA microarray (red), expressed sequence tag (blue), and serial/cap analysis of gene expression (yellow). Figure under CC By License used from Lowe et al. (2017)[2]

After the advent of Next-Gen Sequencing(NGS), transcriptomics took a whole new turn. This paper by Mortazavi et al.[5] published in 2008 was one of the first to study RNA sequencing. In his method, Mortazavi used two rounds of poly(A) selection, then fragmented RNA to an average length of 200 nt and then used random priming to convert RNA into cDNA. The resultant cDNA was then prepared into a library for

Illumina/Solexa 1G sequencing. This paper was grounding breaking in its own way however, it talked about bulk RNA sequencing i.e. sequencing RNA from a large number of cells and this method required almost 100ng of purified RNA. Not too long after, around a year later in 2009, Tang et al.[3] came up with a method for RNA-seq from a single cell. This was the beginning of great strides in transcriptomics with many more papers to follow. Though RNA-seq has moved on to much simpler protocols now, the Tang Protocol is still regarded as a classic. In this protocol, single cells were selected using the mouth pipetting method, lysed and then the mRNA was reverse transcribed. Exonuclease I was used to remove any unused primers and a poly (A) tail was attached to the 3' end of the first-strand cDNAs using terminal deoxynucleotidyl transferase. PCR amplified single-cell cDNAs were made into a library and deep sequenced on the SOLiD system. This technique had the ability to capture up to 75% more genes than a microarray. However, this protocol had several disadvantages with a major one being that it was a long and complicated protocol with several steps and the whole process took 6 days to complete. A summary of all transcriptomic techniques can be seen in Figure 1.

1.2 Next Generation Sequencing(NGS) overview

The development of scRNA-seq would not be possible without the development of sequencing. Sequencing in simple words refers to methods to determine the exact order of nucleotides - Adenine (A), Guanine (G), Thymine (T) or Cytosine(C) in DNA or RNA. The specific order of these nitrogenous bases in the DNA chain ultimately give rise to the hereditary and biochemical properties of all species of organisms on earth. Frederick Sanger and his colleagues sequenced the first DNA genome in 1977 using their SANGER sequencing method.[6] This chain-termination technique used some amounts modified dNTPs in solution that lacked the 3′ hydroxyl group that is essential for DNA chains extension. The chance incorporation of such a nucleotide would prevent bond-formation with the 5′ phosphate of the next dNTP thus preventing extension of the DNA chain. This led to the formation of DNA strands of various lengths and by running a gel, the sequence of nucleotides was determined. This method was so accurate, robust and easy that Sanger sequencing is still being used for low throughput sequencing studies in various labs around the world.

Sanger sequencing was a model sequencing method in a group of sequencing technologies called the first-generation sequencing. Sequencing soon moved on to what was called second-generation sequencing techniques. Pål Nyrén and colleagues discovered the second wave of sequencing called Pyrosequencing in 1993. When a pyrophosphate is released in a chain reaction, this technology can detect light.[6] Pyrosequencing technology was purchased by 454 Life Sciences, and it is considered the first successful commercial sequencing platform which contributed to the rapidly decreasing costs of sequencing.

The 454 sequencing platforms (later acquired by Roche) allowed the mass parallelization of sequencing reactions, thus greatly increasing throughput. However, in 2013 however, Roche announced the discontinuation of the 454 sequencing platform because it became non-competitive.

A notable platform that was introduced following the success of 454 platforms was the Solexa platform (later acquired by Illumina). In the Solexa method, adapter-ligated DNA molecules were passed over a flowcell that had a lawn of complementary oligonucleotides. Solid phase PCR produced neighboring clusters from each flow-cell binding DNA strands via ‘bridge amplification’. In this process, DNA strands bend over and attach to primers on the flow cell to amplify forming a bridge. The first Solexa platforms produced very short reads of 35 bp but one of their unique advantages was that the ability to read both ends of DNA to produce paired-end (PE) data. This Solexa platform was what ultimately gave rise to the HiSeq platform by Illumina, which is capable of increased read length, followed by the MiSeq, which provided lower total throughput (with lesser cost) but proved faster sequencing times and longer read lengths up to 2X300 bp. The HiSeq and MiSeq machine dominate the market today.[6]

Another noteworthy second-generation sequencing platform aside from 454 and Solexa/Illumina was the *sequencing by oligonucleotide ligation and detection* (SOLiD)system from Applied Biosystems (later Life Technologies). SOLiD sequenced by ligation using DNA ligase and didn’t involve sequencing by synthesis. The SOLiD platform couldn’t match the read length and throughput depth of Illumina but it wins on a lower cost/base scale.[7]

Another second-generation sequencing platform is the Ion Torrent (Life Technologies product). In this platform, protons (H^+ ions) released during the polymerization reaction cause a change in pH thus helping identify incorporated nucleotides. This technology is very rapid but just like the 454 platform it is not very good at detecting homopolymer sequences when multiple dNTPs of the same kind incorporate into the DNA strand.

The single molecule real-time (SMRT) platform from Pacific Biosciences is what is considered a third-generation sequencing technology. During SMRT runs, DNA polymerization takes place in Nano-Wells called zero-mode waveguides (ZMWs). As DNA chains get extended by DNA polymerase inside the ZMWs, the incorporation of fluorescent nucleotides is monitored in real time. The PacBio systems have some properties that distinguish it from other competitors. Notable ones include its ability to detect modified bases including DNA methylation which is of immense importance in epigenomic studies.[8] These machines are also capable of long reads, much longer than Illumina reads, of up to 10 kb in length with low systematic error rates which is useful for de novo assembly of the genome or for detecting DNA methylation.

One of the most exciting platforms of the third-generation of sequencers is nanopore sequencing. In this process, a single-stranded RNA or DNA strand is driven across a nanopore by electrophoresis. Each time ion flow is blocked, it creates a disruption in the current which helps us identify the particular nucleotide passing through the pore. Oxford Nanopore Technologies (ONT), offers nanopore sequencing platforms called GridION and MinION which are called benchtop sequencers due to their low cost and small size like that of a cellphone. These qualities enable them to be deployed as point-of-care systems, for example Joshua Quick and Nicholas Loman used these machines to sequence Ebola viruses in Guinea within two days after collecting samples.[6]

After looking at all three generations of sequencing technology, we have to delve a little bit more into one of the second generation technologies: Illumina. This by far has the largest market share today of 60% [9] and is the most commonly used technology to sequence a genome. This is the platform that we have used throughout this thesis to sequence our samples and hence warrants a more detailed explanation. Illumina sequencing involves cluster generation and sequencing by synthesis (SBS) technology. Target DNA is usually end-repaired and adaptor-ligated on both ends (this part of the process is called library preparation). After denaturation, single strand DNA fragments containing adaptors are attached to the Illumina flow cell surface which has complimentary adaptors. Each immobilized fragment of DNA forms a bridge structure by binding its free end with another complimentary adaptor thus getting amplified via bridge amplification. Once the primers are attached, fluorescently labeled dATP, dGTP, dCTP and dTTP are incorporated in a complementary fashion to the template strand. These fluorescently tagged nucleotides have the 3′ hydroxyl group modified such that only a single nucleotide is added per cycle. A picture is then taken by a camera and the combined fluorescence from each cluster of the flow cell tells us about the last incorporated nucleotide. Repeating this over many cycles gives us the sequence of interest. The density of the cluster can go up to more than 10 million clusters per square centimeter (Figure 2).

Once sequencing is complete the next step in the NGS pipeline is genome mapping. The obtained sequencing reads need to be aligned to a known reference genome using various software tools. This tells us where our sequencing reads come from. Depending on the quality of the reads, the raw reads might be trimmed and cleaned up of any overrepresented sequences or primers to reduce bias and improve alignment rate. Usually, uniquely mapped reads with a length shorter than 20bp are discarded because they cannot be unambiguously mapped.

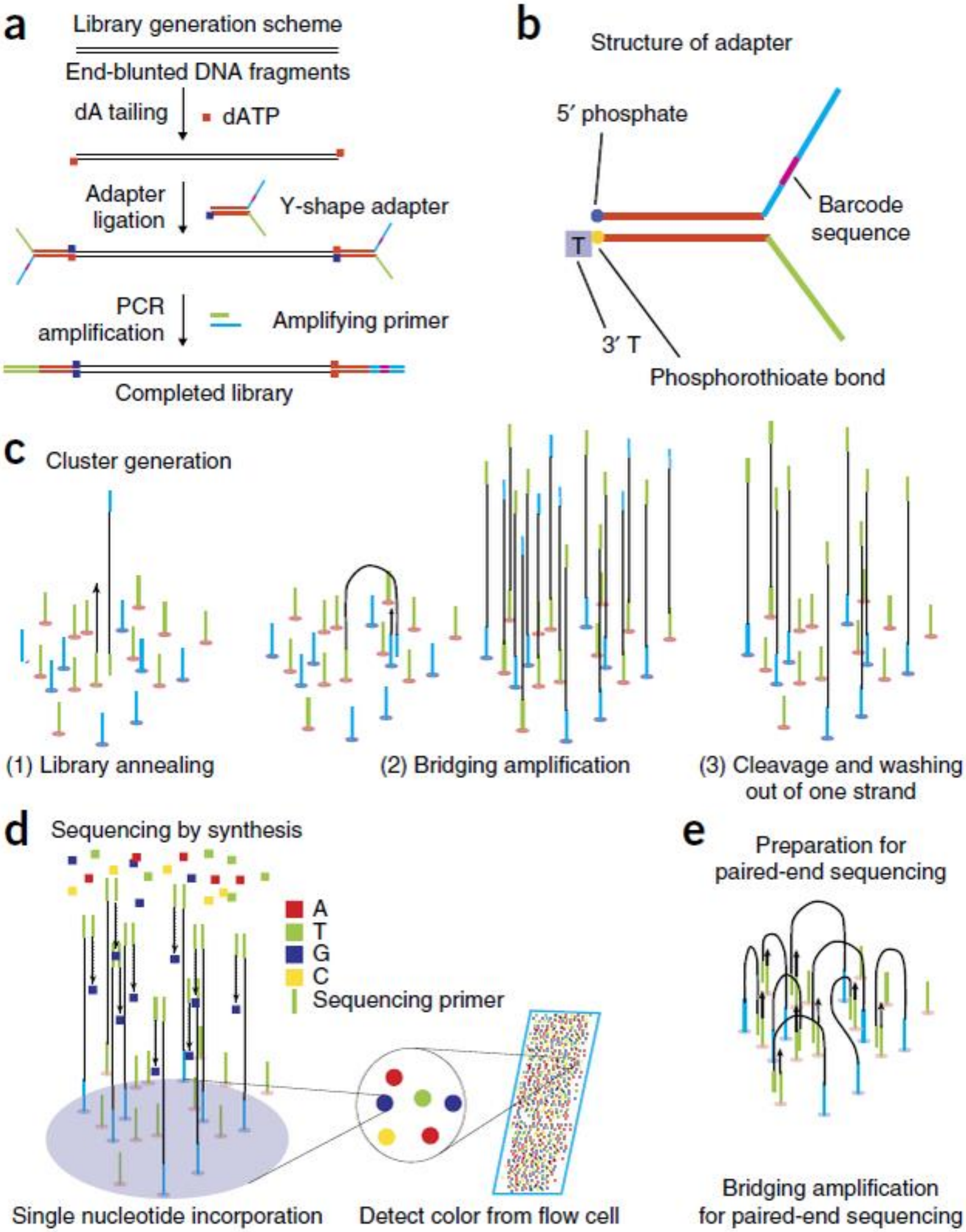
a
Library generation scheme
End-blunted DNA fragments
dA tailing
dATP
Adapter ligation
Y-shape adapter
PCR amplification
Amplifying primer
Completed library
b
Structure of adapter
5′ phosphate
Barcode sequence
T
3′ T
Phosphorothioate bond
c
Cluster generation
(1) Library annealing
(2) Bridging amplification
(3) Cleavage and washing out of one strand
d
Sequencing by synthesis
A
T
G
C
Sequencing primer
Single nucleotide incorporation
Detect color from flow cell
e
Preparation for paired-end sequencing
Bridging amplification for paired-end sequencing

Figure 2: A detailed explanation of the Illumina Sequencing Technology. Figure reprinted with permission from Shin J et al. (2014)[10]

1.3 Single cell RNA-Seq over the years

As referenced in section 1.1, the Tang protocol, despite its various shortcomings, was the first to report the RNA-sequencing of single cell sensitivity. There were two main reasons why it was necessary to increase sensitivity of RNA-Seq to a single cell level instead of bulk RNA seq from millions of cells. The first reason was the need for the analysis of rare cell types (like circulating tumor cells whose concentration is 1-10 cells per 10 ml of blood) or primary cells (cells from patients) for which there may be insufficient starting material. The second reason was the need to study a subpopulation of cells from a larger heterogeneous population (like different subtypes of cells in a tumor) [11]. Studies have shown that transcript levels can vary up to 1000 fold between seemingly similar cells[12] which further necessitates single-cell RNA-seq. Cell-to-cell variation is masked when bulk cells are used and we only get average information of all cells. It is similar to tasting a fruit smoothie and trying to identify all the ingredients that it is made up of. A hint of a mango flavor or the tanginess of yogurt is discernable but overall it tastes like whole blended average of ingredients. This smoothie is equivalent to bulk RNA-seq data from a large number of ground-up cells from a particular organ or tissue say the brain. The brain has many different cell types like neurons, glia, astrocytes, oligodendrocytes etc. The data would ceratinly tell you what transcripts are there in the sample but it wouldn’t tell you which cells those transcripts originated from. It would provide only an average gene expression profile of the mix of cells in the whole brain. Such averages are usually misleading and might lead us to draw false conclusions which is why it is necessary to have technologies to analyze single cells.

Over the past 10 years that have passed since Tang et al.'s paper, several groups have published many new strategies for single cell RNA seq. [13] The most notable ones among these are Smart-seq/Smart-seq2[14, 15], CEL-seq/CEL-seq2[16, 17], STRT-seq[18], Quartz-seq[19], multiple annealing and looping-based amplification cycles (MALBAC)[20], massively parallel single-cell RNA-sequencing (MARS-seq)[21], CytoSeq[22], Drop-seq[23], and inDrop[24]. All of these methods have their own unique advantages and disadvantages. Drop-seq and MARS-seq have the capability of massively parallel high-throughput sequencing of thousands of cells by barcoding each individual cell. CEL-seq2 uses linear in-vitro transcription (IVT) to maintain fidelity during amplification and strand-specificity[25]. The STRT-seq method is another strand-specific protocol and offers the possibility to identify unique transcripts from PCR replicates.

Existing methods for scRNA-seq are not without limitations. For example, although shallow-depth methods are high throughput and cost-effective, these methods detect 50% fewer genes than competing methods[26]. Other methods do not provide full-length transcript information and limit the possibility to detect SNPs or splice variants that are located outside the 5' end [25]. A comparative analysis of some of these methods was published where all the pros and cons were discussed.[26] In this comparison, SMART-seq2 stood out as being the most sensitive process with the least drop-out probability, even coverage of transcript and low variability among replicates. In addition to this SMART-Seq2 was a simpler process that could be completed within a few hours when compared to the 6-day Tang protocol.

1.4 Latest technologies in RNA-Seq

The current methods for scRNA-seq enable us to profile thousands of individual cells in parallel and isolate differences in genotype among single cell populations. One such noteworthy current commercial technology is the Chromium Platform from 10x Genomics.[27] Such high throughput methods have been enabled by advances in droplet microfluidics. Microfluidic devices can generate droplets that can encapsulate a single cell along with a bead coated with barcodes. Once the bead is dissolved each transcript from that cell is specifically labeled with a particular barcode usually called a UMI (Unique Molecular Identifier). Moreover, another barcode exclusive to each bead labels every single cell uniquely. Such labeling helps pool a massive number of cells and combine into one singular amplification and library preparation steps thus greatly reducing costs/cell. Moreover, it helps normalize for PCR amplification bias as each transcript has a unique UMI. Such methods are best for discovering rare cell types in a large population of cells or for understanding the differences in diverse cell collections such as whole tissue, tumor or organ samples.[28] The downside of such ultra-high-throughput technologies includes is that each cell has only a few thousand reads leading to decreased sensitivity or in other words they are unable to identify minor transcriptional differences between cells.[26, 29] However, barcoding cells at early stages followed by low-coverage and low-depth sequencing data are in most cases enough to discriminate different sub-populations of cells. Once such subpopulations are identified, they can be enriched with FACS and a deeper sequencing study, using more sensitive methods like SMART-seq2 can be used to further investigate these subpopulations.

In 2012, the C1 microfluidic platform was introduced by Fluidigm that had a cell capture mechanism that could trap up to 96 cells simultaneously on a single IFC microfluidic chip. All steps of the process were automated and carried out in parallelized nanolitre-sized volumes. [30] The cells captured on the C1 platform could also be evaluated under the microscope for quality control before the reverse transcription and cDNA amplification steps. The minimum cell requirement was 10,000 cells which means that rare cell populations are not well-suited for study on this platform. Another limitation of the Fluidigm C1 is that the cell trapping step is passive and doesn't discard unhealthy or dead cells. There was also an associated size bias on the cells trapped because the C1 Fluidigm only supported cell traps of certain specific diameters [29, 31]. Fluidigm C1 was reported to show higher sensitivity than most competing methods[11] and also the automated microfluidic system saved resources in terms of molecular reagents and labor, however, the large cost of the microfluidic IFC chip itself limits the possibility of carrying out large-scale experiments.

A new direction in which scRNA-seq is evolving is spatial transcriptomics.[32-34] It is a technology to spatially resolve mRNA so in addition to knowing which cell or transcript the mRNA came from; it enables us to know which part of the tissue section it comes from. In this technique sections of tissue from a particular organ, say the brain, are placed on a glass slide that contains barcoded oligo-dT capture probes. After permeabilizing the cells, the PolyA tailed mRNA hybridizes to these barcoded oligonucleotides and is reverse transcribed. After this step, the cDNA is then removed from the slides and all subsequent steps of amplification/library preparation are carried out in a tube. Because the glass slides have positional barcodes, the different tissue sections can be spatially resolved after sequencing. At the moment, spatial features have a diameter of 100 μm that could bind to about 10-100 cells per positional barcode. [35] The because of its ease of use and high throughput, spatial transcriptomics is being adopted widely in the RNA-seq field. It is highly likely that the resolution of this technique will increase very soon and it will lead us to investigate fewer cells per feature barcode. The major challenge for the future would be to integrate this multi-dimensional spatial data with data obtained from other omics technologies.

All the methods for RNA-seq described in this chapter vary greatly in costs, throughput, single-cell isolation methods, sensitivity of gene detection etc. All of these parameters need to be taken into consideration while choosing a method. Depending on the needs of the study, one method may be more suitable than the other. For example, one of the questions that need to be asked is the compromise between cell number and sequencing depth? Studies looking to identify different cell populations with an importance on finding rare cell populations should go for large cell numbers, whereas studies aiming to profile subtle variation in individual genes focusing on differential gene expression should prefer large sequencing depth.

In this thesis work, we aim to develop a novel microfluidic platform for single cell RNA-sequencing which helps to improve on the existing platforms by providing a one-pot system instead of a multi-chambered platform. Our aim is to decrease the volume and size of the scRNA-seq systems and optimize the design such that it permits multiplexing. We hope that this device in addition to being a high throughput device, provides sensitive, accurate and precise scRNA-seq data that compares well to other gold-standard RNA-seq data.

Chapter 2: Diffusion-based microfluidic ‘one-pot’ single cell RNA-seq(MID-RNA-seq)

2.1 Introduction to microfluidics

Towards the end of the last chapter, we talked about microfluidic platforms used for RNA-seq. Most tube-based or plate based platforms for RNA-seq offers a throughput of 50 to 500 single cells per experiment. This throughput can be further increased by use of automation with liquid-handling robotics. Their sensitivity ranges between 5,000–10,000 genes per single cell.[29] However, microfluidic platforms take multiplexing, automation, and sensitivity to a whole new level. Numerous studies [11, 36-38] have shown that doing single cell analysis(or limited starting material analysis) on a microfluidic chip as opposed to conventional in-tube reactions increases the sensitivity and accuracy of these reactions. Such platforms typically perform better than tube-based methods with higher reproducibility due to the reduction of stochastic variation caused by pipetting error and manual handling.[36] In addition, microfluidic isolation leads to less contamination and increases throughput.[36] In addition to this, there is a high scope of multiplexing several reactions and automation.

In the field of microfluidics, we design, manufacture and operate devices and processes that have dimensions of tens or hundreds of micrometers and have volumes of fluid of nanoliters or picoliters. Typically, most microfluidic devices are made out of a master mold which is fabricated via photolithography and the device itself is made via soft lithography. (Figure 3)

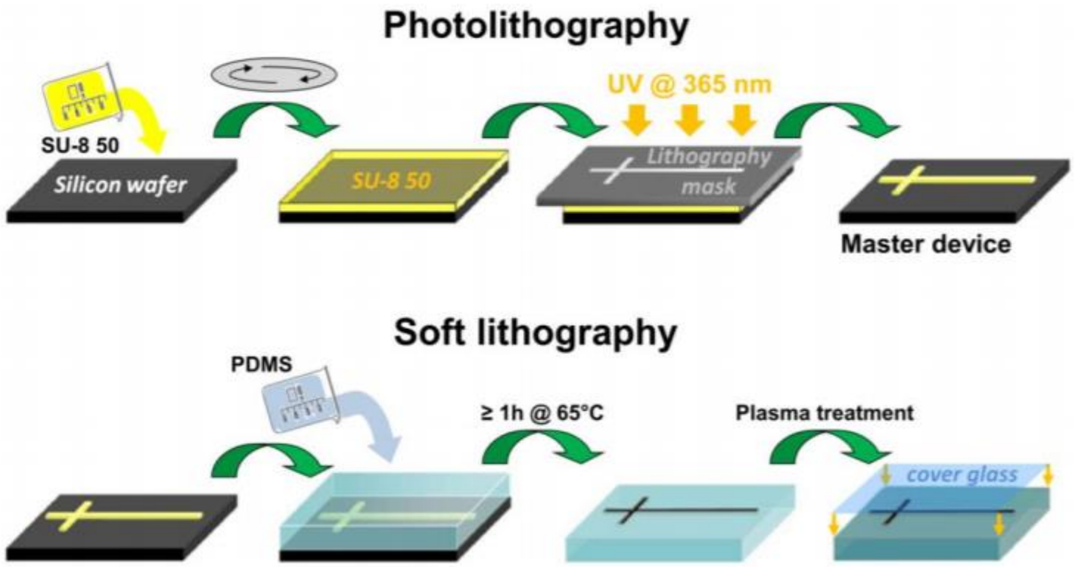

Figure 3: The technique of photolithography and soft lithography to fabricate microfluidic devices. Figure reprinted with permission from Ma et al. (2014)[39]

In this process, an epoxy-based negative photoresist called SU-8 is spin coated onto a wafer made of silicon to the chosen thickness. Next, a transparency mask (which can be pre-designed and printed on any drawing software based on our requirements) is placed on the wafer and the whole arrangement is exposed to UV. The exposed parts of the photo-resist get cross-linked while the unexposed parts can be washed off using SU-8 developer leaving behind a desired pattern mold on the silicon wafer. Next, a polymeric organosilicon compound like Polydimethylsiloxane (PDMS) is poured onto the silicon wafer which can be baked at 80°C and then peeled off the wafer leaving behind a stamp of the original pattern in it. Fluidic channels (i.e. channels connecting chambers in microfluidic devices which need to be closed by valves) can also be manufactured using the same process except in this case a positive photoresist like AZ is used. For AZ-P4620 the opposite takes places where the unexposed parts are insoluble while the parts that are exposed to UV light become soluble to the photoresist developer. All fluidic channels must be made with AZ-P4620 photoresist so they can be reversibly closed and opened by valves. Valves are similar to devices but are made with a thinner layer of SU8 photoresist. The valves can either pass above (push-down) or below (push-up) the fluidic channels and the valves structure themselves are filled with water. Pneumatic pressurization of the valve causes the membrane to deflect up/down into the flow structure thus sealing the channel. The whole device including fluidic and control layer is bound to a glass slide via plasma bonding.

The popularity of microfluidics has surmounted in recent years for its ability to examine minuscule quantities of cell samples (including single cells) and creating highly controlled microenvironments.[40] When it comes to transcriptomics a number of microfluidics platforms have been used over the years. The first protocol used for reverse transcription and amplification via microfluidics used the T7 RNA polymerase, which amplifies mRNA linearly[41]. However, this protocol was demonstrated for microarray analysis and not for NGS. It required between 20 pg to 10 ng purified RNA and thus was close to single cell sensitivity.

Streets et al.[36] was the first group to adopt the single cell Tang protocol[3] to a microfluidic platform. Their microfluidic chip was designed to run 8 parallel reactions and each reaction unit had 6 chambers. Each chamber was dedicated to a particular part of the RNA-seq reaction viz. cell lysis, reverse transcription, poly-A tailing, primer digestion, and second strand synthesis.

Another popular method for RNA-seq, SMART-seq and its successor SMART-seq 2 have been adapted into the C1 platform by Fluidigm by various groups. It has been used to study heterogeneity in brain cells[30, 42] and bone-marrow derived dendritic cells [43]. CEL-seq2 was also adapted to the Fluidigm C1 platform and used to study mouse fibroblast cells and was found to have a significantly higher efficiency than the tube-based CEL-seq2.[17]

The Fluidigm C1 and other microfluidic platforms for RNA-seq often involve a device containing multiple connected chambers [30, 36, 44, 45]. The chambers are kept empty at the beginning of the process before reagents involved in various steps are loaded into the system by opening an increasing number of these connected chambers. Special measures need to be in place to prevent the reagents in earlier steps from inhibiting the reactions in the later steps. For example, chemicals used for cell lysis (such as sodium dodecyl sulfate and Triton X-100) and intracellular molecules such as proteins, polysaccharides, and ions (including Ca^{2+}, Fe^{3+}) in the cell lysate may inhibit PCR by reducing polymerase activity [46-48]. Several strategies such as bead-based methods and non-chemical lysis have been reported to overcome these inhibitory effects [44, 45, 49, 50]. However, these strategies may not produce a complete release of RNA as chemical lysis [51]. Dilution typically has to be used to alleviate interference among reagents [45].

To overcome these limitations of existing microfluidic devices, we demonstrate a one-pot microfluidic device to perform scRNA-seq called MID-RNA-seq. Our approach takes advantage of concentration-gradient-driven diffusion to deliver reagents into the reaction chamber while diffusing out reagents from the previous step, thus eliminating the need for dilution. We show that the results obtained using MID-RNA-seq are comparable to competing scRNA-seq technologies. We demonstrate the utility of this approach with the SMART-seq2 steps and reagents and the device has the capacity for use with other scRNA-seq protocols.

2.2 Results and Discussion

2.2.1 Design and operation of MID-RNA-Seq device

We designed a diffusion-based microfluidic device for scRNA-seq (Figure 4). We applied diffusion-based reagent swapping for reagent loading [52, 53]. The microfluidic device had two layers - a fluidic layer for the chambers and flow lines (indicated in red and pink) and a control layer of pneumatic microvalves (indicated in green). Each of the valves could be independently addressed. The device had two parallel units for processing two single cells simultaneously. Each unit contained three sections: 1) a reaction chamber (80 nl); 2) a loading chamber (200 nl) attached to the reaction chamber; and 3) a cell trapping structure upstream of the reaction chamber. The connection between the reaction and loading chambers could be changed by operating the microvalves that could open or close the diffusion channels in between the chambers.

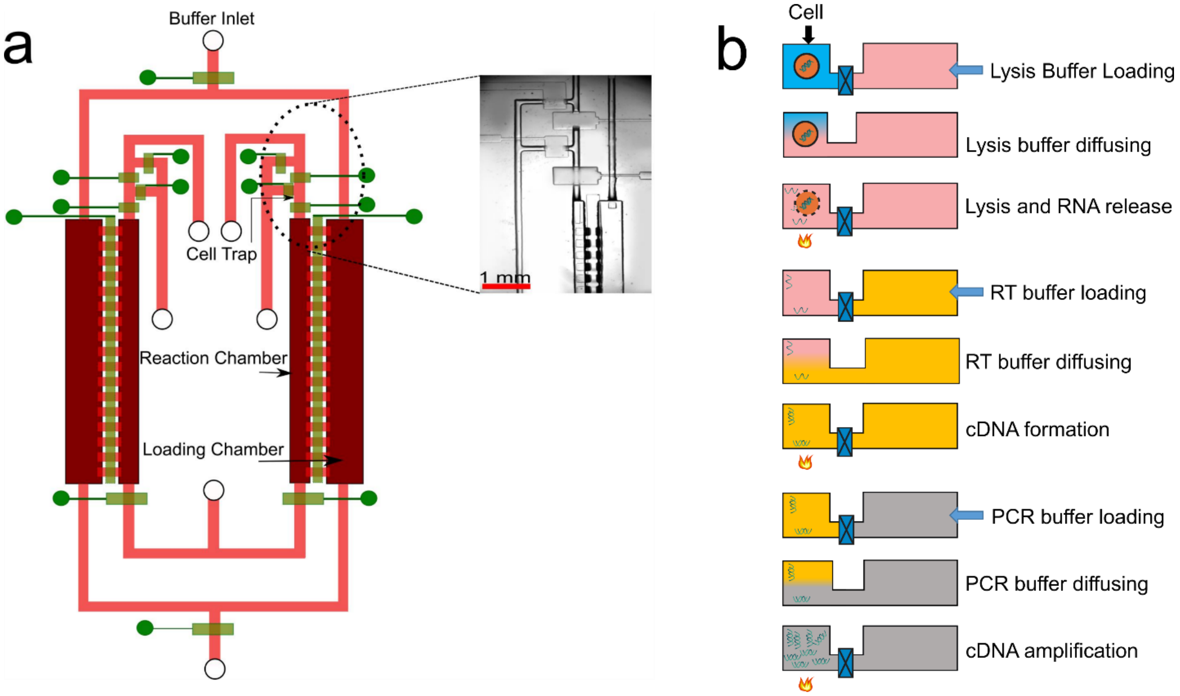

Figure 4: MID-RNA-seq device and operation. a) A schematic illustration of the two-layered microfluidic device in a top-down view (not to scale). Inset: the circled area in Fig. 4a is seen under the microscope. b) Steps involved in MID-RNA-seq in a cross-sectional view.

The single-cell suspension was loaded onto a syringe pump and introduced into the device through the cell inlet. (Figure 4, Figure 5) Single cells were trapped in the cell trapping chambers by operating the surrounding valves(Figure 5). After a single cell was successfully trapped, the upstream channel of the cell-trapping chamber was first rinsed with PBS and then with Lysis Buffer to remove unwanted cells. Next, Lysis Buffer was flowed to push the trapped cell into the reaction chamber by slowly squeezing the air out of the reaction chamber through gas-permeable PDMS while the downstream valve (the valve at the exit of the reaction chamber) and the diffusion valve (i.e. the valve between reaction and loading chambers) were closed. The entire process was monitored under a microscope to ensure the chamber was free of bubbles. The whole microfluidic chip was then mounted on a flat-plate thermocycler for lysis reaction (72℃, 3 min). In the next step, the reverse transcription buffer was filled into the loading chamber via the buffer inlet and the diffusion valve was then opened for 40 min for the RT reagents to diffuse into the reaction chamber. The lysis reagents diffused out during the same period of 40 min to avoid interference with the RT reaction. Because of the relatively large size of mRNA-molecules (average size assumed to be 1.5 Kb [54]), they diffused slowly in the time scale of operation and the loss of RNA by diffusion was small.

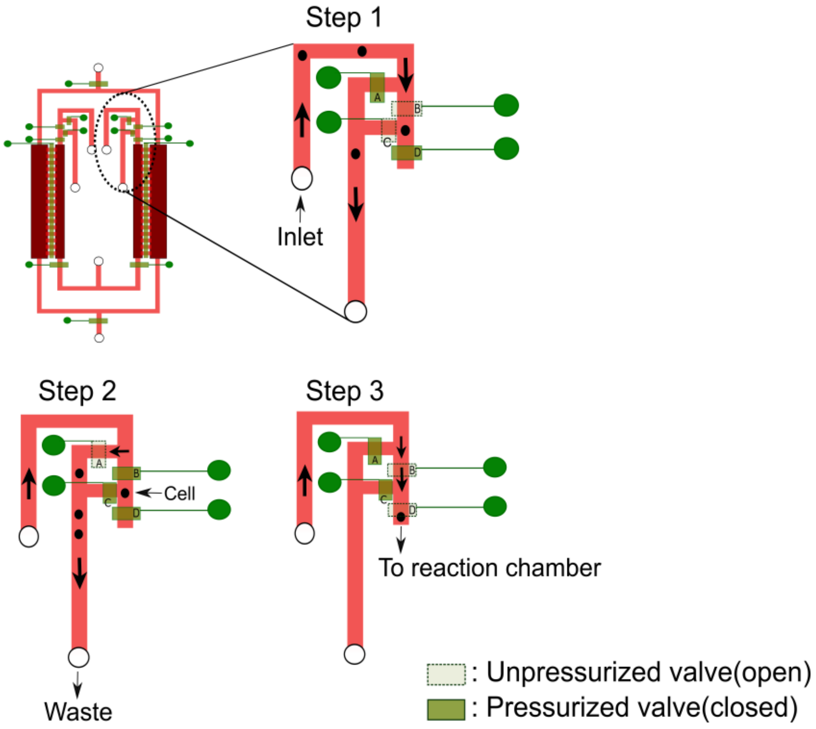

Figure 5: The various steps involved in single cell trapping in MID-RNA-seq. Black arrows show the directions of the flow.

Once the diffusion-based loading was complete, the diffusion valve was closed and the entire chip was placed on the flat-plate thermocycler for the RT reaction. For the PCR amplification step, the reagents in the loading chamber were replaced with fresh PCR buffer and the diffusion valves were again opened for 25 min and the chip was once more placed on the flat-plate thermocycler for PCR reaction to take place. The total reaction volume of the one-pot reaction chamber was 80 nl, which was about ~40% smaller in volume than a Fluidigm-C1-type of device for scRNA-seq [36]. After cDNA synthesis and amplification, the units were independently flushed with elution buffer and the cDNA was collected at the cDNA outlet using a micropipettor and into a tube. The library preparation step was performed in a tube using conventional benchtop techniques. cDNA libraries were sequenced using Illumina HiSeq 4000/Illumina HiSeq X.

The reagent loading/exchange in MID-RNA-seq occurred via concentration-gradient-driven diffusion. The buffer containing reagents for each step was filled into the loading chamber via the buffer inlet (Figure 4) and then the diffusion valve separating the loading and reaction chamber was opened. This allowed for the reagents to diffuse into the reaction chamber and the reagents of the previous step to diffuse out thus preventing the interference of reagents from one step to the other. This process was then repeated for the subsequent steps.

There is a considerable difference in diffusivity (D) between RNA and the other reagents so that the reagents diffuse in and out more quickly than RNA. At 25 °C, mRNA molecule has an average D value of 1 $\mu m^2 s^{-1}$ [55] whereas the RT enzyme (average size 71 KDa) has a D of ~50 μm^2 s^{-1} (estimated from [56]), primers (20-30 bp single-stranded DNA) have D of 70 μm^2 s^{-1}, dNTPs of 370 μm^{2}s^{-1} , Triton X-100(lysis reagent) of 300 μm^{2}s^{-1}, small ions like Mg^{2+}, K^+, Cl^- of 1000 μm^{2}s^{-1} [52]. When delivering RT enzyme which has a small difference in diffusivity compared to RNA, a high concentration of the RT enzyme was applied in the loading chamber to accelerate delivery.

2.2.2 Modelling and optimizing the diffusion-based process

COMSOL Multiphysics was used to model the diffusion process in the MID-RNA-seq device for visualization of the exchange of molecules between the reaction chamber and loading chamber for the reverse transcription step (Figure 6). The 'transport of diluted species' model was used to carry out a time-dependent simulation to analyze the concentration variations of different species in the chambers at 25 °C within the time-frame of 2400 s (40 min). The diffusion process was modelled using Fick's first law $J=-D\ (\partial\varphi/\partial x)$ where J is the diffusion flux (mol m^{-2} s^{-1}), D is the diffusivity or diffusion coefficient (m^2 s^{-1}), φ is the concentration (mol m^{-3}), and $\partial\varphi/\partial x$ is the concentration gradient. No flux at the boundaries was used as a boundary condition. The starting concentrations of Reverse Transcription Enzyme, Primers, dNTP, small ions like Mg^{2+}, K^+, Cl^- in the reaction chamber were set at 0 and their initial concentrations in the loading chamber were set at 100 (arbitrary units). The increase in the species concentration in the reaction chamber over time was modeled. The same approach was used for simulation of species (RNA and Triton-X) diffusing out of the reaction chamber to the loading chamber (by setting their starting concentrations in the reaction chamber as 100 and the ones in the loading chamber as 0). A custom free tetrahedral mesh with maximum element size 200 μm and minimum element size 10 μm with maximum element growth rate of 1.4 was used to model the system. Our results(Figure 6) showed that only 5% of RNA diffuses out whereas 69.53% of Triton-X (lysis reagent) diffused out over 40 min period. Over the same period, the concentrations of reverse-transcription enzyme, primer, dNTP and small ions increased to 18.05%, 24.96%, 61.48% and 65.02% of their starting concentrations in the loading chamber, respectively. This simulation was used to help determine the diffusion time and the loading concentrations of these reagents in order to reach desired concentrations in the reaction chamber.

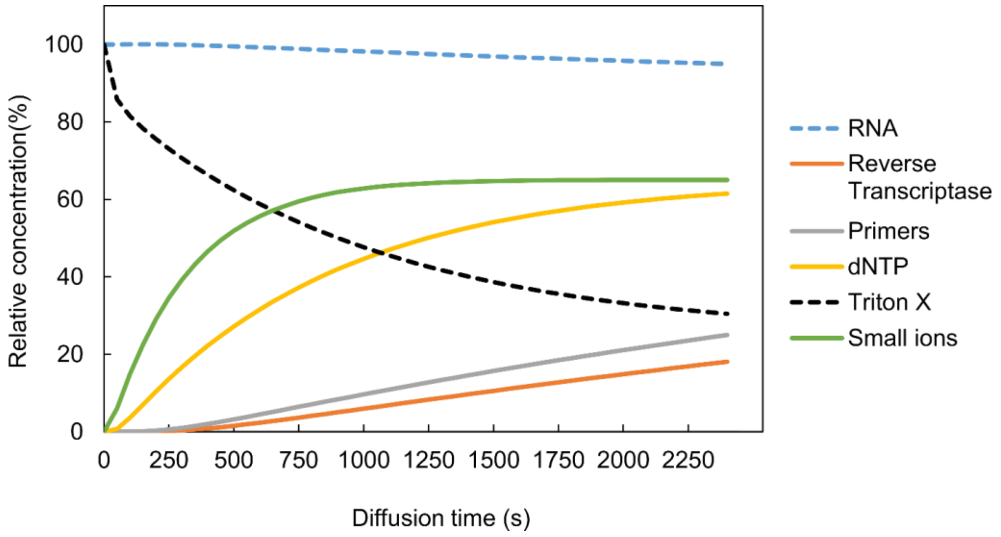

Figure 6: The loading and release of various molecules in and out of the reaction chamber as modeled by COMSOL Multiphysics.

The optimization of the diffusion duration for each step involved a balance between maximizing delivery of reagents and minimizing RNA/cDNA loss. In addition to the COMSOL Multiphysics modeling, we also prepared some samples each containing 10 pg of RNA (equivalent to single-cell amount[15]) for testing various diffusion durations in the 2- unit MID-RNA-seq device. We opened the diffusion valves for various durations to see how the total number of genes with Fragments Per Kilobase of transcript per Million mapped reads (FPKM) > 0 changed with different diffusion durations. There were two steps that involved diffusion – the reagent swapping during the reverse transcription step (denoted by R) and the reagent swapping during the PCR step (denoted by P). Figure 7a shows that the number of genes detected increased as we increased the diffusion durations from 20 to 40 min for R while keeping P constant at 20 min. Figure 7b shows that the number of genes detected peaked at 20 min when the diffusion duration for P increased from 10 to 30 min, while keeping R constant at 40 min. Thus, in conjunction with the results obtained by COMSOL modelling, we chose a diffusion duration of 40 min for R and 25 min for P.

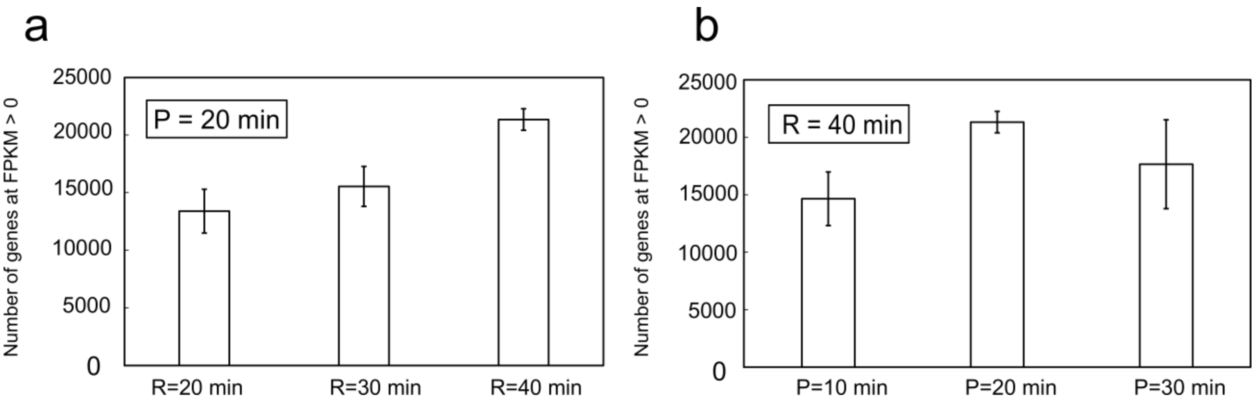

Figure 7: The number of genes detected at FPKM > 0 is plotted against various diffusion durations (time for which the diffusion valve was open) of the reverse transcription (R) and PCR (P) steps in a 2-unit MID-RNA-Seq device. a) R was varied while P was kept constant at 20 min. b) P was varied while R was kept constant at 40 min.

2.2.3 MID-RNA-seq performance benchmarked against competing technologies

In total, we totally produced 41 single-cell data sets using MID-RNA-seq technology on GM12878 human lymphoblastoid and Mouse Embryonic Fibroblast (MEF) cell lines (29 data sets on GM12878 by 5 runs on the 6-unit devices, 6 on GM12878 by 3 runs on the 2-unit devices, and 6 on MEF by 3 runs on the 2-unit device).

Sensitivity in scRNA-seq is usually measured by the total number of genes detected per cell and the number of genes overlapped between the single-cell approach and bulk RNA-seq measurement [11, 36]. We compared our GM12878 data to a number of published datasets by ENCODE (GSM2343071/2, 31 single-cell datasets) and by Marinov et al. which uses SMART-seq (GSE44618, 15 single-cell datasets)[57]. In addition to these, we also compared it to the single-cell data on HCT116 Human Colon Cancer cell line by Fluidigm C1 technique using the SMART-seq2 protocol(GSE51254, 96 single-cell datasets) [11]. The mouse MEF cell line datasets were compared to the ones produced by SMART-seq2[14](GSE49321,7 single-cell datasets), microfluidic scRNA-seq by Streets et al. [36](GSE47835, 8 single-cell datasets) and ones with CEL-seq2 on mouse ear fibroblast cells performed on the Fluidigm C1 platform [17] (GSE78779, 72 single-cell datasets). All of the raw data available were downloaded and processed with the same bioinformatics pipeline with the same number of sub-sampled reads (2 million for each data set).

Figure 8 shows the total number of genes we detect at FPKM > 0 and FPKM > 1 from these datasets. With FPKM >0, our method detected 8908 genes with human datasets and 14726 genes with the mouse datasets. In the case of genes with FPKM>1, our method detected 4762 and 5338 genes respectively.

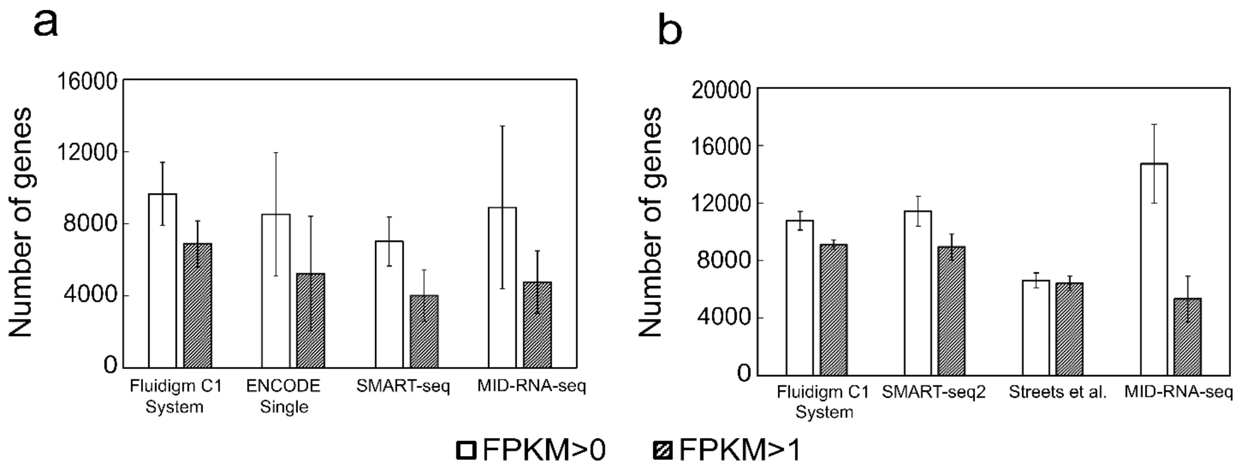

Figure 8: Sensitivity of the scRNA-seq techniques shown as the mean number of genes detected above two thresholds (FPKM of 0 and 1) for data on a) GM12878 cells b) MEF cells.

Figure 9 we further analyzed and compared the quality of the single-cell datasets completed by various methods in terms of their overlap with the bulk RNA-seq data from ENCODE on the same GM12878 cell line (GSE33480, ‘ENCODE bulk’) and produced in our lab using 1000 cells with SMART-seq2 (‘Bulk GM12878’). All datasets were depth matched at 2 million reads and an average profile across all sample replicates (provided by Cuffdiff program which also controls for variability across replicates[58]) was used for comparison. In the first case of comparison with ENCODE bulk data, the percentage overlaps for genes with FPKM >1 were 43.89% for SMART-seq[57], 44.48% for ENCODE-single cell, and 38.18% for MID-RNA-seq. In the case of overlap with Bulk

GM12878 data generated in our lab, the percentage overlaps were found to be 46.58%, 47.08%, 41.69% for SMART-seq, ENCODE-single cell, and MID-RNA-seq respectively. Percentage overlap of single cell data with bulk data is an indicator of the sensitivity of the data.[11] Compared to the literature, 30-40% overlap is typically reported. [11]

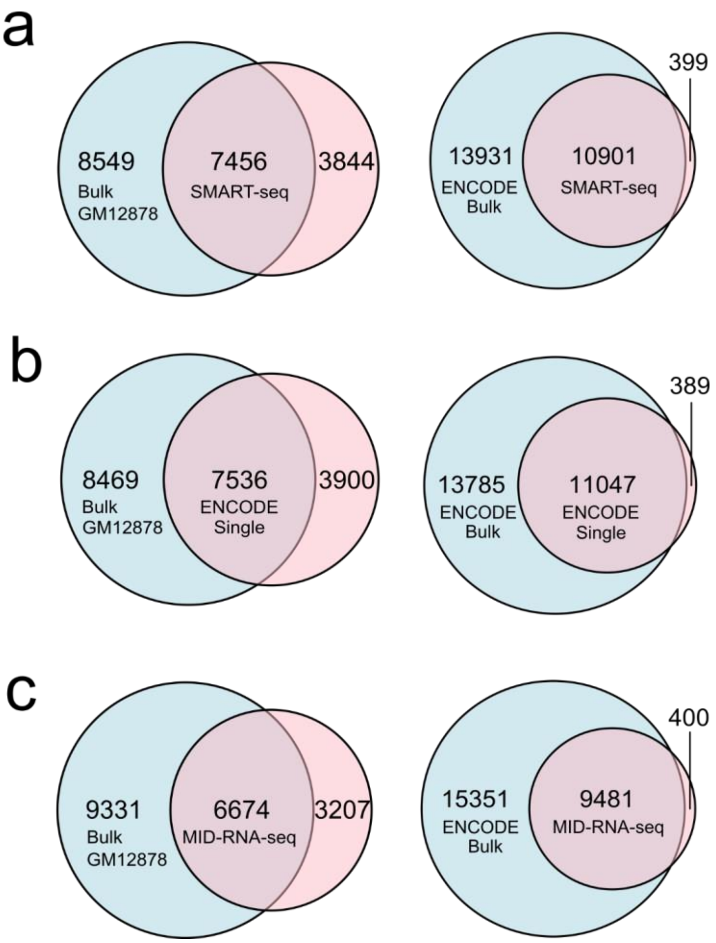

Figure 9: The number of genes that overlap between the scRNA-seq data and the bulk RNA-seq data on GM12878 cell line. The bulk GM12978 data was generated in-house using a commercial RNA-seq kit and compared to different single cell datasets viz. a) SMART-seq single cell data (n=15); b) ENCODE single-cell data (n=31); c) MID-RNA-seq single-cell data (n=35).

In order to compare precision fairly among methods without biases, we calculated dropout probability [26]. We created a set of 18842 human genes that were detected at FPKM>1 in at least one of four single-cell datasets (MID-RNA-seq, ENCODE_single cell, SMART-seq [57], Fluidigm C1 System(human) [11]). We then examined each dataset individually to determine how many genes from these genes were dropped out (FPKM=0). MID-RNA-seq was found to have one of the lowest dropout probability of 0.17 for human cell studies(Figure 10a).

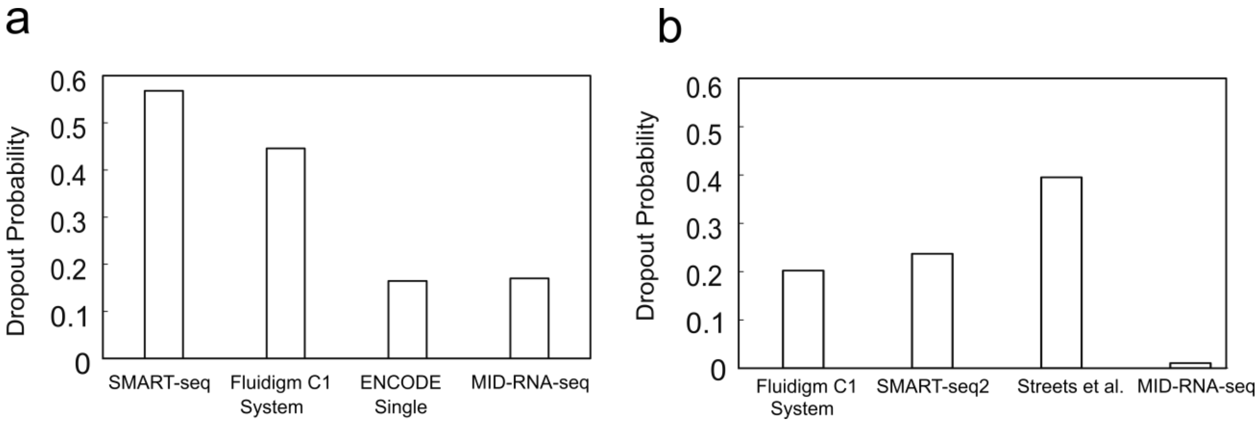

Figure 10: Dropout probability for a) human single-cell datasets b) mouse single-cell datasets taken using MID-RNA-seq.

While comparing the mouse datasets, 13529 genes were included in the common set and MID-RNA-seq was again found to have the lowest dropout probability of 0.01(Figure 10b). The average measurements provided by Cuffdiff across all sample replicates and 2 million randomly-sampled reads were used for examination for fairness. The dropout probability of SMART-seq2 (0.23) and that of Fluidigm C1 for human (0.44) were found to be similar to values reported in previous literature [26].

The correlation between pooled single-cell data (referred to as ‘MID-RNA-seq ensemble’) and bulk RNA-seq data produced in our lab using 1000 GM12878 cells by a commercially available kit (that uses SMART-seq2) is shown in Figure 11 a. Raw reads from 29 single-cell data sets were computationally pooled and then 7 million reads were randomly sampled to form the ensemble. The same number of reads were randomly sampled from the pooled sequencing reads of the bulk RNA-seq data[11]. The Pearson correlation coefficient obtained was 0.72 which showed that the ensemble could partially recapitulate the bulk data. A Loess regression curve was fitted on the data and the curve was found to be almost linear with a $R^2 = 1$. Figure 11 b is another way of visualizing this overlap where 76.36% of the average number of genes detected lies in the common area.

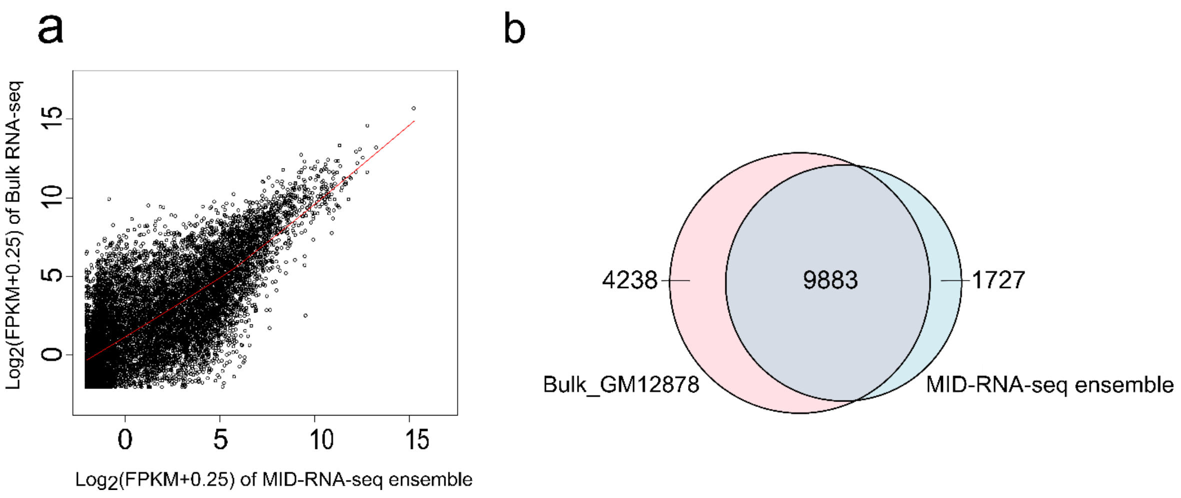

Figure 11: a) Correlation of all genes with FPKM > 0 between pooled single cell data (MID-RNA-seq ensemble) and bulk RNA-seq data both from GM12878 cells. b) The overlap of genes detected between the bulk RNA-seq data and MID-RNA-seq ensemble data.

Figure 12 shows a heat map comparing the expression level of all genes at FPKM >0 across the 29 single-cell data sets (sequenced in the same batch) from 5 rounds of the 6-unit device. The Pearson correlation ranged from 0.34 to 1 among these samples with an average of 0.68. The variation could be attributed to heterogeneity in gene expression among single cells [36].

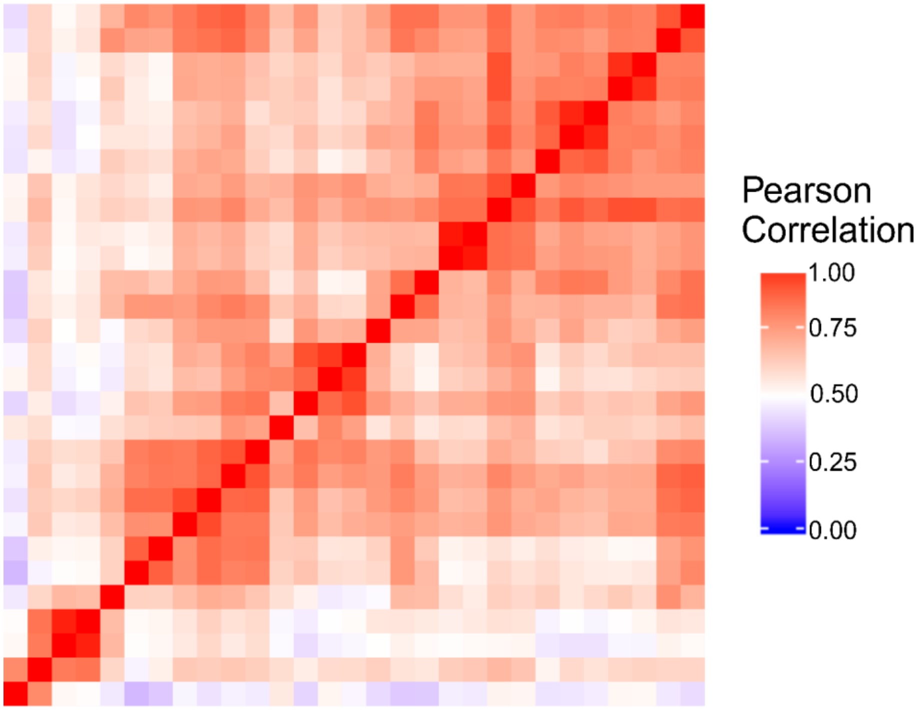

Figure 12: Heat map of Pearson correlation among all genes with FPKM > 0 for the 29 single-cell RNA-seq data sets produced using the 6-unit device.

In order to assess accuracy in the mRNA expression level, exogenous spike-in of 92 polyadenylated synthetic RNA transcripts from the External RNA Controls Consortium (ERCC) was added to the mix. The expression level of each transcript in the spike-in mix was measured in the experiment to determine how well it correlated with its known concentration. The results (Figure 13) showed a strong linear correlation (adjusted R2 ~ 0.9) between measured and original FPKMs of spike-in molecules detected across all 6 single cell datasets performed on a 2-unit device. The adjusted R2 value was used so that the value doesn't depend on the number of data points used. This value of linear correlation corresponded well to values published in literature under similar settings [16, 26, 59] . The number of ERCC molecules detected out of the 92 molecules is N = 84 which also corresponded well to guidelines (general ERCC guidelines indicate that a good quality library usually has an R2 value greater than 0.9 and an N value higher than 60).

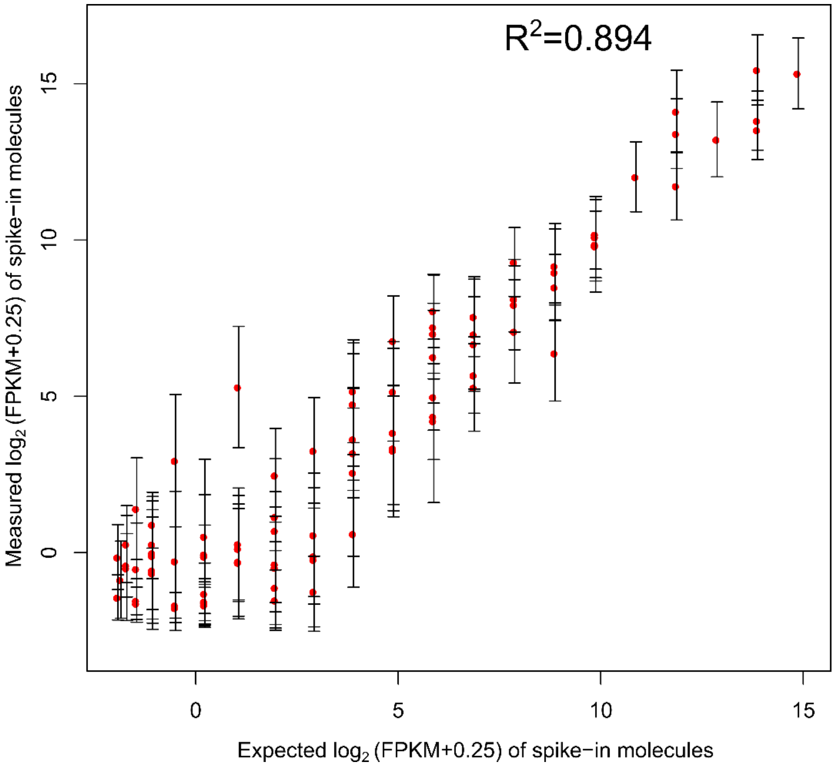

Figure 13: Correlation between measured values (in FPKM) of spike-in molecules detected versus expected values of these molecules

In order to visualize the variability of transcript expression across samples and replicates, the gene expression variance (CV^2, the squared coefficient of variation) was plotted against the mean expression $\log_{10}$FPKM. (Figure 14) The CV^2 value is a measure of variability across replicates and a lower value indicates lesser variation among replicates and is an indicator of data quality. The average profiling across replicates provided by Cuffdiff was used. Six single-cell datasets were randomly picked for each method with each data set containing 2 million randomly sub-sampled reads. In Figure 14a, for human cell samples, the CV^2 for MID-RNA-seq data weaved in between the SMART-seq and ENCODE_single cell data while Fluidigm C1 (human) showed a lower coefficient of variation. In Figure 14b, with the mouse cell line data, MID-RNA-seq shows the lowest CV^2 out of all the technologies compared. Streets et al.[36] (another microfluidic technology) also shows lower CV^2 than other tube-based technologies.

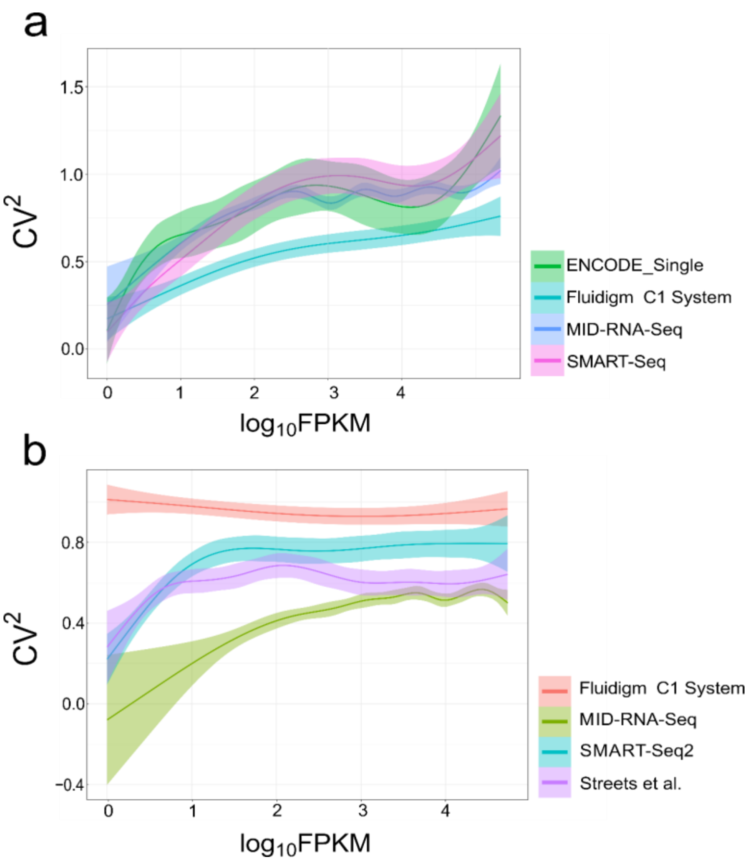

Figure 14: The relationship between variance (CV^2) and FPKM of genes in single cells is shown for a) human and b) mouse cell line samples.

2.2.4 Increasing throughput with 4-unit and 6-unit device

The MID-RNA-seq platforms also offer opportunities for increasing throughput. In addition to the 2-unit device shown in Figure 4, we also tested MID-RNA-seq devices containing 4 units spanning along the horizontal direction (Figure 15a) or 6 units connected in the vertical direction (Figure 15b) to demonstrate the scalability along both directions. As seen in Figure 15a, the diffusion, inlet and outlet valves are combined to ensure simultaneous operation of the four reactions. Four samples of 10 pg RNA each were processed in the device which was the equivalent of one single-cell worth of RNA [15]. The time taken for the operation of the 4-unit device was similar to that of the 2-unit device. Figure 16 shows a compilation of the results obtained from each unit. The results showed that on average 5743 genes (FPKM>1) were detected in each unit and 2503 were reproducibly detected in all units. Between any two units, around 65% of the discovered expressed genes overlapped. In the analysis by Wu et al.[11], a 57-65% reproducibility was typically reported among single cell methods.

In the 6-unit device in Figure 15b, we combined 6-units in the longitudinal direction with all loading chambers connected by common buffer inlet and outlet. Each reaction chamber was isolated to prevent cross-contamination across samples (single cells). There was a common cell trap upstream of the first unit, operating in the same fashion as described in Figure 5. Single cells trapped were individually pushed into each reaction chamber while being observed under the microscope. The amplified cDNA output of each reaction was individually flushed out for subsequent library preparation steps.

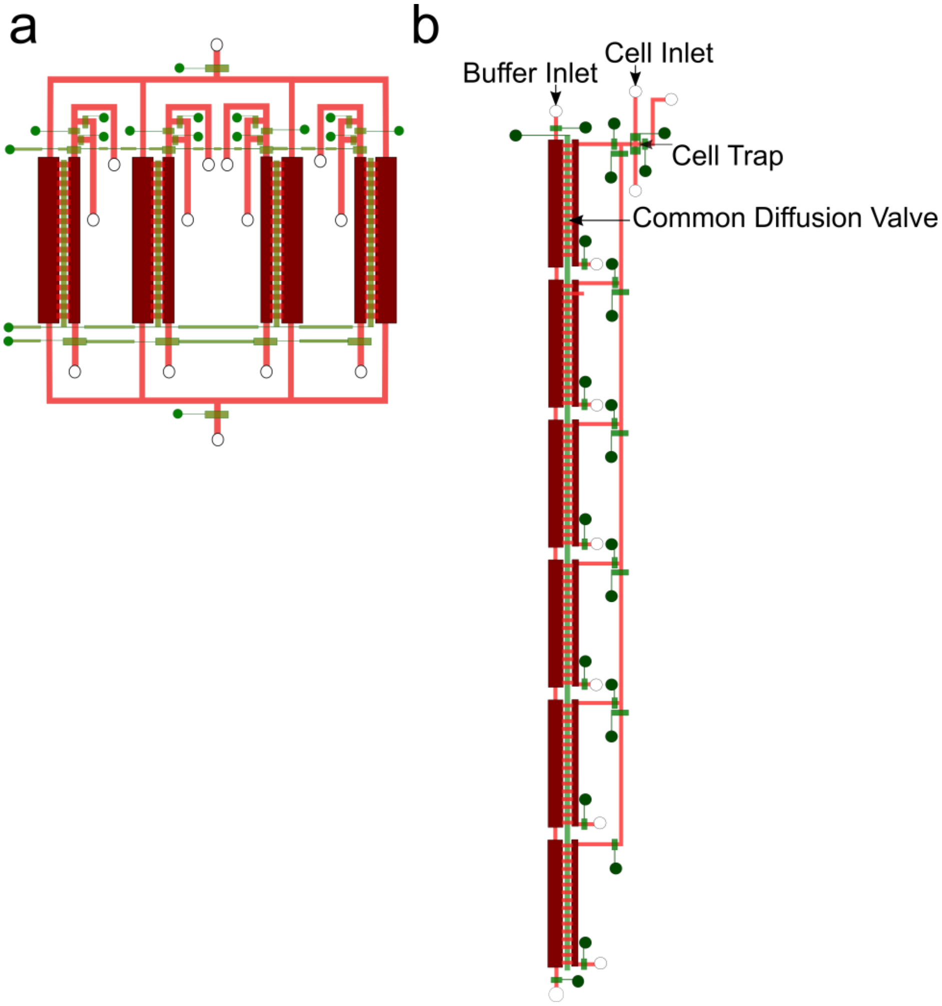

Figure 15: A 4-unit (a) and 6-unit (b) MID-RNA-Seq device with single cell-trapping ability.

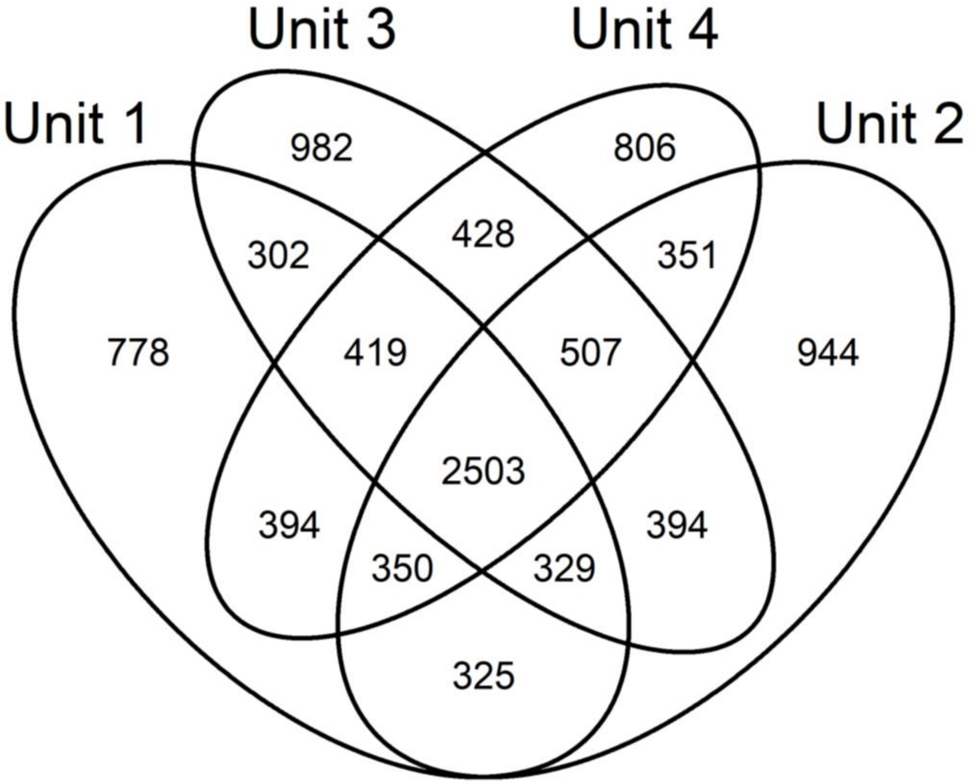

Figure 16: Genes with FPKM>1 detected in the four units of the 4-unit MID-RNA-seq device using RNA from GM12878 cells. The numbers indicate how many genes overlap between each unit.

2.3 Materials and Methods

2.3.1 Microfluidic device fabrication

All microfluidic devices were made using multi-layer soft lithography [60-62]. A photomask containing the desired microscale patterns was designed on Layout Editor and printed on 10000 dpi films (Fineline Imaging, Colorado Springs, CO). The master mask was fabricated by spinning SU-8 2025 (Microchem, Newton, MA) and AZ P9260 (Clariant, Charlotte, NC) on a silicon wafer with the thickness being 50 μm for the reaction/loading chamber (made with SU-8) and 13 μm for fluidic channels (made with AZ P9260). The master was heated to 130℃ for 30 secs to form a rounded cross-sectional profile for the channels made in AZ 9260. The control layer master was fabricated in SU-2025 with 24 μm thickness. The control layer PDMS was made by spinning PDMS (RTV615A: RTV615B=20:1 ratio, R. S. Hughes, Sunnyvale, CA) at 500 rpm for 10s and then at 1500 rpm 30s, which resulted in a thickness of 65 μm. The fluidic layer PDMS was mixed in the ratio of RTV615A: RTV615B =5:1 and had a thickness of ~0.4 cm. Both layers of PDMS were cured at 80 ℃ for 15 min. The two layers were then aligned and thermally bonded for 1 h at 80 ℃. The two-layer PDMS device was then carefully peeled off from the master and access holes were punched to form the inlets and outlets for tubing attachment. Finally, the PDMS structure was bonded to clean thin cover glass (Thickness #1 (0.13 to 0.17mm), Ted Pella Inc.) after plasma oxidation of both surfaces (Harrick Plasma, Ithaca, NY). After bonding to glass, the device was baked at 80℃ overnight to strengthen the plasma bonding.

2.3.2 Cell culture

GM12878 cells were obtained from the Coriell Institute for Medical Research. The cell line was grown in RPMI 1640 media (11875-093, Gibco) supplemented by 15% Fetal Bovine Serum (26140-079, Gibco) and 1% penicillin-streptomycin (15140-122, Gibco) at 37°C and 5% CO_2, passaged every 2-3 days to maintain exponential growth. Before the scRNA-seq experiment, the concentration of the cell suspension was adjusted to 3.2×10^5 /mL in PBS using a hemocytometer to facilitate single cell trapping.

MEF cells were obtained from ATCC (SCRC-1040) and cultured in DMEM (ATCC 30-2002) with 15% FBS and 1%PS at 37°C and 5% CO_2. Cells were harvested at 80% confluence. They were detached by incubating with 0.25% trypsin with 0.1%

EDTA (Thermo Fisher 25200056) for 1 min and then centrifuged at 120 × g for 5 min. Then, the supernatant was discarded, and cells were resuspended in fresh media and then adjusted to a concentration of 3.2×10^5 /ml.

2.3.3 Microfluidic device operation

The control layer (indicated in green in Figure 4, Figure 5) was filled with deionized water before experiments. The pneumatic microvalves were actuated at 30psi by solenoid valves (18801003-12V, ASCO Scientific) that were connected to a compressed air supply. The operation of microvalves was controlled by a LabVIEW program on a computer and a data acquisition card (PCI-6509, National Instruments). We prepared double-stranded cDNA from mRNA of single cells using the protocol as described in Picelli et al.[14] after modification for compatibility with microfluidic platform.

Operation of the two-unit device: The diluted cell suspension was loaded into the microfluidic device by a syringe pump through the inlet (Figure 5). In order to do this, an air plug of about 1 cm long was created in a tubing filled with deionized water by aspiration before a plug of the cell suspension was aspirated into the same tubing. The tubing containing the cell suspension was then plugged into the inlet and the cell suspension was flown into the cell trap at a flow rate of 5 µl/min while valves B and C were open (Figure 5). The cell trapping chamber was observed under the microscope constantly to monitor the arrival of cells. Once a single cell was selected, valve C followed by valve B were immediately closed to trap the cell. Next, the tubing delivering cells was taken out of the inlet and replaced with a fresh tubing delivering Lysis Buffer (0.33 U/µl of RNase inhibitor (Thermo Fisher Scientific, cat. no. N8080119), 0.95% Triton X-100 solution (Sigma-Aldrich, cat. no. T9284), 3.33 µM oligo-dT primer (IDT), 1.66 mM dNTP (Thermo-Fisher cat. no. R0192)). Valve A was opened and any residual cells upstream were flushed out at a flow rate of 30 µl/min. Next, the flow of the syringe pump was halted and valve A was closed. This was followed by the opening of valves B and D. Built-up pressure in the system pushed the trapped cell in lysis buffer into the reaction chamber while pushing out the air in the reaction chamber through gas-permeable PDMS. The outlet valve of the reaction chamber remained closed at all times to ensure no cell escape. Once this was complete, the valves B and D were closed to generate complete isolation of the cell in the reaction chamber. This process was conducted in both units of the device (2-unit device). The device was placed on a flat-plate thermocycler (Techne TC-4000) for lysis to take place at 72 °C for 3 min and then held at 4 °C until the next step. Paraffin oil (18512, Sigma Aldrich) was used to facilitate thermal conduction between the flat plate and the glass substrate of the microfluidic device. After lysis was complete, the reverse transcription (RT) buffer (50U/µl Superscript II (Invitrogen, cat. no. 18064-014), 0.122 %Tween-20 (Thermo-Fisher Scientific cat. no. 85113), 0.5U/µl RNase inhibitor, 1X First Strand Buffer (Invitrogen, cat. no. 18064-014)), 3µM TSO primer(Exiqon), 5mM DTT (Invitrogen, cat. no. 18064-014), 1 M Betaine (Sigma-Aldrich, cat. no. 61962), 10mM $MgCl_2$, 1mM dNTP mix) was then loaded into loading chambers of both units using the syringe pump at a flow rate of 10 µl/min through the buffer inlet. We ensured that no bubbles were trapped anywhere in the device by carefully scanning the device under the microscope. Once the buffer loading was complete, the inlet and the outlet valves seal off the buffer loading chambers. The diffusion valves separating the reaction and loading chambers were then opened for 40 min to allow the diffusion-based exchange to deliver the RT reagents and move out the lysis reagents. After completion of this exchange, the diffusion valves were closed and the flat-plate thermocycler was programmed for reverse transcription to take place at 42°C for 90 min, 10 cycles of (50°C for 2 min, 42°C 2 min), 70°C for 15 min. The device was held at 4 °C until the next step. For the PCR amplification step, the reagent mix in the loading chamber was replaced with fresh PCR buffer (0.3 µM PCR primer, 1X of Fidelity Buffer (Kapa Biosystems cat. No. KK2502), 0.3 mM dNTP, 0.1 U/µl HiFi Hot Start DNA Polymerase (Kapa Biosystems cat. no. KK2502)) using a syringe pump at the flow rate of 10 µl/min. The diffusion valves were then opened for 25 min and closed. The device was placed on the thermocycler and the reaction was carried out at 98 °C for 3 min, 18 cycles of (98 °C for 20 s,67 °C for 15 s, 72 °C for 6 min), 72 °C for 5 min. The device was held at 4 °C until the next step. In order to prevent evaporation of water during long PCR cycles, water droplets were placed on top of the microfluidic device making sure to cover all access holes. Once the PCR reaction was over, for each unit, a tubing filled with Tris-EDTA buffer was attached to the cell inlet and the upstream was flushed out (by opening valve A) to remove any residual lysis buffer/debris at 20 µl/min. Then valves B, D, and outlet were opened while valves A and C were closed and the amplified cDNA was flushed out of the reaction chamber and collected via pipette into an Eppendorf tube. The entire microfluidic process of cDNA preparation from cell trapping to flushing out of the amplified cDNA took about 6-7 h. This was just about 1 h of additional time compared to a standard bench-top SMART-seq2 protocol[63]. A schematic of this process in cross-sectional view can be seen in Figure 4b.

Operation of the four-unit device: The four-unit device (Figure 15a) had no single cell trapping modules attached. The operation was otherwise the same as the two-unit device. The four units were loaded sequentially with RNA solution that was extracted from GM12878 cells that were lysed off-chip. The RNA solution was diluted to 0.125 pg/nl to ensure 10 pg per 80 nl of reaction chamber volume for each unit. The amplified cDNA was extracted sequentially from each unit to prevent any cross-contamination across units.

Operation of the six-unit device: The operation of the 6-unit device(Figure 15b) was similar to those of the 2-unit and 4-unit devices. A common cell trap upstream of the first unit trapped cells in a similar fashion with a side outlet for flushing out excess cells. Once a cell was trapped, a syringe attached to a tubing filled with lysis buffer pushed the trapped cell into each unit individually with the help of a syringe pump. During this process, the other units were sealed off using the microvalves to prevent any interference. The action was repeated 6 times to trap single cells in each unit. Once cells were trapped, buffers were loaded into the outer loading chambers through the buffer inlet. The common diffusion valve was then opened to allow diffusion to take place followed by the closing of the diffusion valve and flat-plate thermocycler based reactions. The time for operation of the 6-unit device was slightly longer than that of the 2-unit or 4-unit devices with additional 10-15 min for trapping 6 cells instead of 2.

Primer sequences used in the process are as given below:

TSO Primer: (5′-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′) from Exiqon

Oligo dT Primer: (5′–AAGCAGTGGTATCAACGCAGAGTACT30VN-3′) from IDT

PCR Primer:(5′-AAGCAGTGGTATCAACGCAGAGT-3′) from IDT.

2.3.4. Library preparation and sequencing

After clean-up of amplified cDNA using Ampure XP Beads (Beckman Coulter A63881), the sample was quantified using Qubit Assay and the appropriate amount of cDNA (100-300 pg) was used for preparing a library using the NEXTERA XT (Illumina, cat. no. FC-131-1096) kit following the manufacturer's protocol. Bulk GM12878 samples were prepared using the SMART-seq v4 Kit from Takara Bio (cat. no. 634894) in the high input mode followed by library preparation using the same NEXTERA XT kit. We checked library fragment size using high sensitivity DNA analysis kit (5067-4626, Agilent) on an Agilent 2200 Tape Station. The libraries were sent out for Illumina HiSeq 4000 sequencing with single-end 50 bp reads with the exception of data produced using the 6-unit device (by Illumina HiSeq X sequencing with 2x150 bp paired-end reads). Averagely, each single-cell data set had a sequencing depth of 11.45 million reads and bulk RNA-seq data sets had a sequencing depth of 3.5 million reads.

2.3.5 Analysis of sequencing data

Quality check of the sequencing data was performed using FASTQC. The sequencing reads were trimmed by Cutadapt 0.4.1 and Trim_galore 1.12 to discard low-quality reads. Adaptors and overrepresented sequences were also trimmed. If trimmed reads were less than 25 bp they were discarded. In order to ensure a fair comparison among methods with differences in sequencing depths, we sub-sampled reads to two million reads each [11, 26]. The cleaned reads were then mapped to the hg19 human genome or the mm9 mouse genome using Tophat v2.1.1. For all libraries, the percentage of mapped reads were between 80-90%. The mapped reads were then converted to fragments per kilobases transcript per million mapped reads(FPKM value) using the well-established *Tuxedo suite* pipeline as described in the paper by Trapnell et al. [64]. R scripts were used to perform further downstream analysis of the Cufflinks and Cuffdiff files to generate other supporting figures. MID-RNA-seq data are deposited under accession number GSE119271.

2.4 Conclusion

In this chapter, we demonstrate a microfluidic platform for scRNA-seq referred to as MID-RNA-seq. The key advantage of MID-RNA-seq is that it eliminates the need of multiple chambers for multi-step treatment by using concentration-gradient-driven diffusion to deliver reagents into the reaction chamber and to remove reagents from previous steps. We also demonstrate that MID-RNA-seq as a scRNA-seq method is sensitive, precise and accurate when benchmarked against the state-of-the-art scRNA-seq techniques. The MID-RNA-seq protocol is complete from start to end: cell trapping, reverse transcription, amplification all take place on the microfluidic chip with very simple structures. Furthermore, the MID-RNA-seq device also offers other advantages associated with microfluidic platforms such as reduced reagent costs, scalability, and multiplexing. Our microfluidic device is also compatible with essentially all scRNA-seq protocols. One drawback of our device is the additional time required for the associated diffusion-based reagent swapping steps. However, this processing time overhead averaged on each assay can be decreased by having a number of units working in parallel.

Chapter 3: Transcriptomic studies of memory of innate leukocytes

3.1 Introduction to innate memory

The immune system is typically divided into two classes – innate immunity and adaptive immunity. Innate immunity is a defense mechanism that is the first round of response to an antigen in the body. The innate immune response is quick and nonspecific. The major players involved in innate immunity include the skin which acts as a physical barrier, chemicals in the blood, and innate immune system cells that destroy foreign cells in the body. In contrast, the adaptive immune response is known to be an antigen-specific response. The adaptive immune system processes and recognizes the antigen first. After antigen recognition, the adaptive immune system secretes an army of immune cells specially designed to target that particular antigen. This adaptive immune response is slow acting and it includes a memory component thus future exposures to the same antigen elicits a stronger response.

For years this has been the classic view of innate vs adaptive immunity – quick vs late, generic vs specific, memory vs. no-memory. However, many studies later proved that the innate immune system also has a 'memory' component. [65-69]. Innate immune memory is different from adaptive memory in a number of ways, for example it lacks gene rearrangements that happens in adaptive immune cells to remember pathogens, it involves epigenomic reprogramming, it involves a different class of cells (leukocytes vs. T and B lymphocytes), and it also involves a different variety of receptors for (selective pattern-recognition receptors (PRR) in innate vs. antigen-specific T cell and B cell receptors in adaptive). Overall, innate memory is still accepted to be a non-specific short-lived response, whereas adaptive memory is accepted to be long-lasting and specific.[69]

A wide body of emerging evidence has shown that following an infection or vaccination, innate immune cells (such as monocytes, neutrophils, or natural killer cells) display long-term changes in function mediated by epigenomic reprogramming.[70] Thus on secondary stimulation by pathogens, there is increased responsiveness, increased production of inflammatory mediators, and improved capacity to eradicate infection in the body. Such epigenomic reprogramming could be mediated by histone modifications[71] , DNA methylation[72] , microRNA[73] and/or long noncoding RNA expression[74]. This leads to changes in the transcriptomic landscape thus changing the innate immune cells' capacity to respond to subsequent stimulation.

One of the standard examples indicative of leukocyte memory is that leukocytes can be 'primed' or in other words, they remember the signal strength and duration of distinct stimulants and modify their phenotype accordingly.[66, 67]Monocytes primed with IFN-γ, β-glucan, or super-low dose endotoxin have been shown to adopt a non-resolving inflammatory state which in turn is favorable to the propagation of chronic inflammatory diseases, notably compromised wound healing[75] and atherosclerosis.[76] On the other hand, 'non-primed' or tolerant monocytes could contribute to homeostasis and inflammation resolution.[68]

Lipopolysaccharide (LPS), a molecule universally found on the surface of Gram-negative bacteria, is a classical pattern recognition molecule of the innate immune system. Innate immune cells react differently to different doses LPS, for example, high doses(1 μg/ml) cause acute but resolving inflammation, whereas low doses(100pg/ml) cause low-grade and chronic non-resolving inflammation.[66] If cells are challenged by high-dose LPS, it changes cells in such a way that upon subsequent challenge, a less robust inflammation response is seen. Priming with a very low dose of LPS however, actually primes the pro-inflammatory response to any successive endotoxin challenge. These two scenarios are summarized in Figure 17. In a practical scenario, slightly elevated levels of LPS(low-dose LPS) are often observed in humans who suffer from chronic diseases and who smoke and drink.[66] Such low-grade inflammation slows and prevents normal wound healing, and also leads to chronic heart disease, diabetes, Parkinson's disease, and rheumatoid arthritis.[77, 78] Thus, a better understanding of the process of how innate leukocytes in such cases get programmed into a chronic, non-resolving state will help us come up with better treatment strategies. In this chapter, we aim to understand the change in the transcriptomic landscape of innate immune cells by varying LPS doses. We do this by carrying out single-cell and bulk RNA-seq studies of neutrophils and monocytes challenged by LPS and interrogate the expression levels of different genes and also interrogate the efficacy of therapeutic molecules like 4-PBA on chronic inflammation.

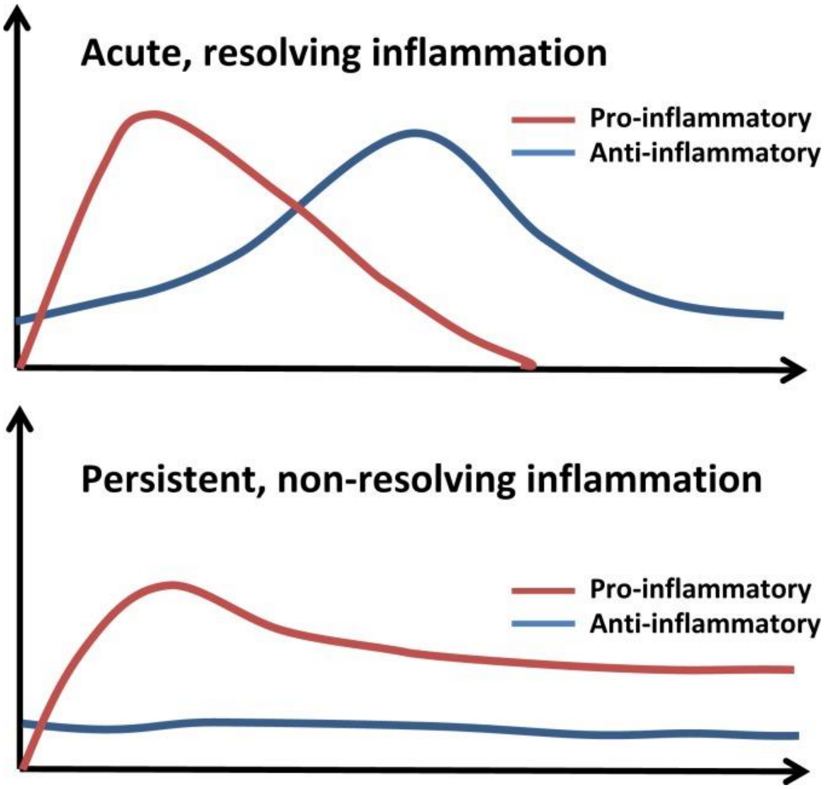

Figure 17: The two types of response of LPS stimulation, tolerant and prime. Figure under CC By License from Morris et al.[66]

3.2 Results and Discussion

3.2.1 Neutrophil adaptation to varying dosages of LPS

Neutrophils are the first responders to any attacking pathogens and are a major player of the innate immune system. Neutrophils perform various functions, which include phagocytosis, chemotaxis, toxin secretion via degranulation, and generation of NET.[79] In this study, we tried to study the transcriptome of neutrophils adapted to high and low doses of LPS over a prolonged period of time and how they may contribute to atherosclerosis. When cells are stimulated by LPS, LPS forms a LPS–LBP protein complex by binding with the LPS-binding protein (LBP). This complex is then presented by CD14 for recognition by the Toll-Like Receptor (TLR4) on the cell surface, which forms a heterodimer with its co-receptor MD-2.[80] Post recognition, the TLR4 signaling cascade is activated which leads to a number of immune responses from leukocytes. We studied the gene expression profiles of murine neutrophils constantly exposed to varying dosages of LPS for a 5-day period.

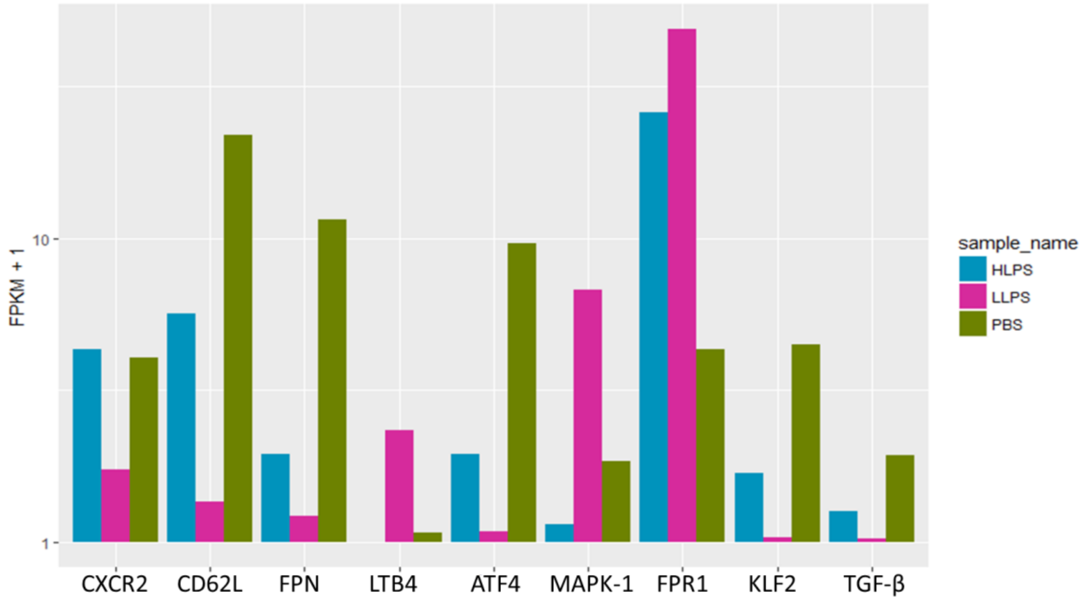

Figure 18: The single cell gene-expression levels (indicated by FPKM) in neutrophils treated with varying LPS doses.

Neutrophils were modified into an anti-inflammatory “tolerant” phenotype after being challenged with high dose LPS (1 µg/ml) and were compared to PBS and low-dose LPS (1 µg/ml) treated neutrophils. This is evident in a significant increase in the FPKM levels of pro-inflammatory genes such as MAPK-1, FPR-1, and LTB4 as seen in Figure 18. The significance testing was performed by the Cuffdiff program and 8 single-cell samples of each type were tested. They also produced lesser levels of CXCR2, FPN, and IRF-5 which aid in abating the inflammatory response. For example, FPN polarizes innate leukocytes into an anti-inflammatory phenotype by varying the iron content of the cells.[81] Leukotriene B4 (LTB4) is an inflammatory lipid mediator and is known to be the reason behind reduced plaque stability in atherosclerosis.[76] Significant elevation of LTB4 gene transcription can be seen levels neutrophils exposed to LPS when compared to the control case with PBS/high LPS. Reduced levels of anti-inflammatory mediator transforming growth factor– β (TGF- β) could be observed in low-dose LPS(Figure 18).

To test whether neutrophils were activated or not, we found significantly reduced levels of CD62L(Sell) a classical indicator of neutrophil activation. Homeostatic transcription factors such as KLF2 and ATF4 transcribe homeostatic molecules like FPN and reduce ROS in cells.[81] The levels of KLF2 and ATF4 were observed to be significantly reduced. Recently, it has been shown that the reduction of homeostatic modulators like KLF2 and ATF4 can polarize neutrophils into a non- resolving inflammatory state favorable to atherosclerosis.[81] The same study has also found evidence of increased atherosclerotic plaque in ApoE −/− mice injected with low-LPS primed neutrophils.[81]

The generation of neutrophil NET plays significant roles in defense against microbes and inflammation.[82] Our transcriptomic study also suggests that neutrophil NET formation, controlled by MAPK-1 gene(Figure 18) may be stimulated by both low and high dose endotoxin. Our transcriptomic study hopefully opens up new avenues to investigate neutrophil function further.

3.2.2 Neutrophils depolarized by 4-PBA

Given that low dose LPS polarizes neutrophils into a primed state via induction of LTB4, CD11b, etc. to promote atherosclerosis, the application of 4-PBA has been shown to attenuate this phenomenon. [81, 83] In our transcriptomics study, this treatment restores the expression of KLF2, LRRC32, TGF- β in neutrophils treated by LPS(Figure 19). It has also been shown that 4-PBA treatment reduced plaque sizes, plaque lipid content and elevated collagen content in *ApoE* −/− mice injected with 4-PBA treatment.[81]

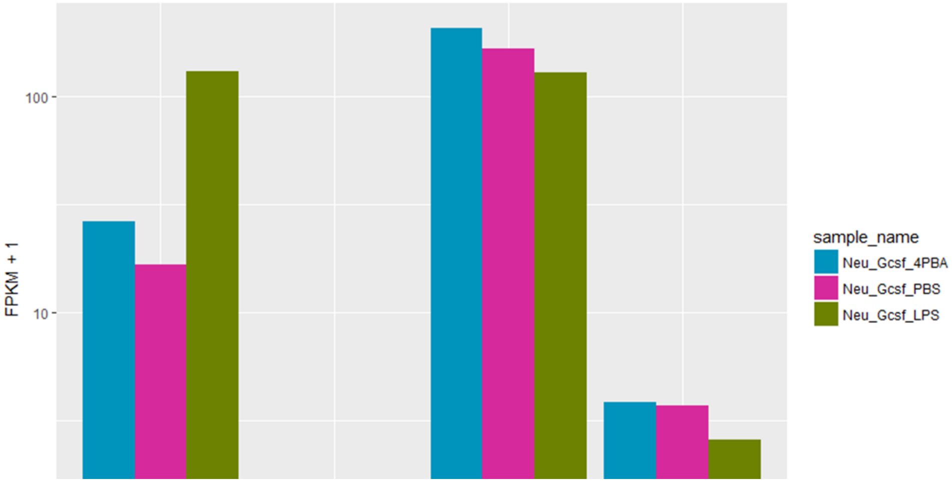

Figure 19: Gene expression in bulk neutrophils treated with 4-PBA vs control

3.2.3Monocyte adaptation to varying dosage of LPS

When monocytes are challenged with IFN-γ, β-glucan, or super-low dose endotoxin can adopt a primed non-resolved inflammatory phenotype.[68, 70] This can occur via epigenomic reprogramming, metabolic reprogramming[84] or via interference with negative signaling regulators, like AKT, ERK, Tollip, and IRAK-M. These negative regulators establish homeostasis and their disruption is theorized to lead to a primed inflammatory state.

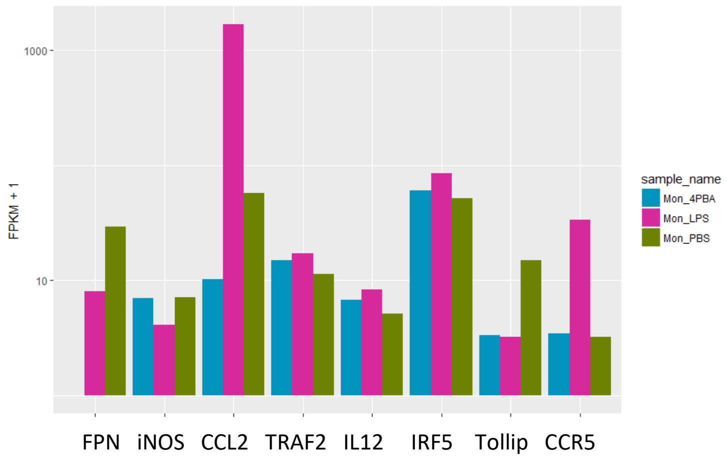

Figure 20:The gene expression levels in bulk murine monocytes treated with varying levels of LPS and 4-PBA

As seen in Figure 20, IRF5, which is a classic indictor of an inflammatory monocyte, is elevated in murine monocytes treated with super-low dose LPS, and restored by the application of 4-PBA. Elevation of IRF5 is often observed in the pathogenesis of chronic inflammatory diseases. The anti-inflammatory phenotype is also reflected in a rise of selected proinflammatory genes such as IL-12 and CCR5. The low-dose LPS treated monocytes also expressed lower levels of homeostatic tissue repair genes such as Nos2(iNOS). These are classic markers of non-resolved inflammation as reported before.[85] TRAF-2 overexpression has also been reported to be associated with increased production of pro-inflammatory genes and can be clearly seen in Figure 20 to be induced by low-level LPS. The levels of negative signaling regulator Tollip and homeostatic regulator FPN have also been decreased by LPS. CCL2 is an important inflammatory cytokine in monocytes which is linked to many inflammatory disorders[86] and its levels are seen to be spiked up when treated by LPS.

In Figure 21a, we can see the results of GO analysis on a number of top genes which are differentially expressed in monocytes treated by LPS when compared to LPS+4-PBA and in Figure 21b, the GO analysis of genes differentially expressed in monocytes treated by LPS when compared to monocytes treated by PBS(control). As we can see the major contenders of genes affected by LPS include membrane receptor binding, G-protein coupled receptor binding and cytokine/chemokine activity. This indicates that important innate immune responses are modulated by the application of LPS and 4-PBA in monocytes. A complete list of up and downregulated genes can be found in Appendix .

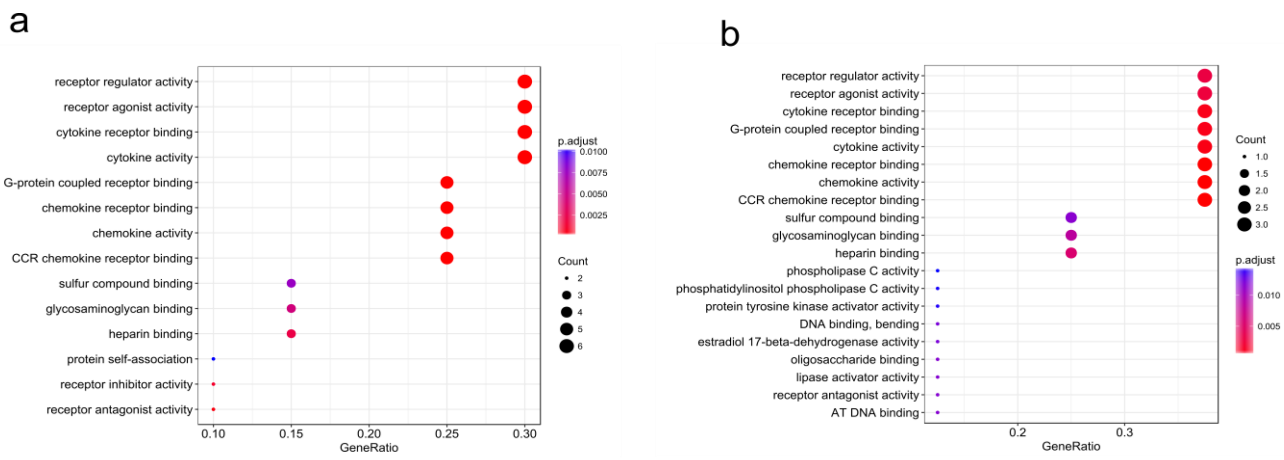

Figure 21: GO analysis of differentially expressed genes in monocytes a) between LPS and LPS+4-PBA treated cells b) between LPS and PBS treated cells

We may just have been skimming the surface of monocyte priming and tolerance while the submerged part may involve complex scenarios of monocyte memory, adaptation, and programming. Many proteomics and genomics studies have revealed multifaceted and varied profiles of differentiated and activated monocytes.[87] This transcriptomic study will likely provide some leads as to where the next efforts to uncover innate memory of monocytes should be focused on.

3.3 Materials and methods

3.3.1 Animals

C57BL/6 were maintained and bred under standard pathogen-free conditions. 6- to 8-week-old male mice were used for the experiments. All animal experiments were approved, prior to the initiation of this study, by the Institutional Animal Care and Use Committee (IACUC) of Virginia Polytechnic Institute and State University.

3.3.2 In vitro neutrophil/monocyte priming

BM neutrophils/monocytes were isolated from C57BL/6 mice using standard protocols as described here[88, 89] and cultured in complete RPMI medium containing 10% fetal bovine serum, 2 mM l-glutamine, and 1% penicillin/streptomycin in the presence of G-CSF (100 ng/ml). PBS, low-dose LPS (100 pg/ml), high-dose LPS (1 μg/ml) or 4-PBA (1 mM) was added to cell cultures. Lipopolysaccharide (*Escherichia coli* 0111: B4) was purchased from Sigma. Murine macrophage colony-stimulating factor (M-CSF) was obtained from PeproTech. Fresh LPS and G-CSF was added to the cell cultures every 2 days. After 5 days, cells were harvested.

3.3.3 RNA-seq studies

All RNA-seq experiments were carried out using the SMART-seq v4 single cell kit from Takara Bio (cat. no. 634894) following the manufacturer’s recommendations. Either a single cell or 1000 cells/reaction each was used. In the case of single cell studies, 8 replicates were used and in the case of bulk (1000 cells) studies, 4 replicates were used. After clean-up of amplified cDNA using Ampure XP Beads (Beckman Coulter A63881), the sample was quantified using Qubit Assay and the appropriate amount of cDNA (100-300 pg) was used for preparing a library using the NEXTERA XT (Illumina, cat. no. FC-131-1096) kit following the manufacturer's protocol. We checked library fragment size using high sensitivity DNA analysis kit (5067-4626, Agilent) on an Agilent 2200 Tape Station. The libraries were sent out for Illumina HiSeq X sequencing with 2x150 bp paired-end reads.

3.3.4 Bioinformatics analysis of RNA-seq data

Quality check of the sequencing data was performed using FASTQC. The sequencing reads were trimmed by Cutadapt 0.4.1 and Trim_galore 1.12 to discard low-quality reads. Adaptors and overrepresented sequences were also trimmed. Trimmed reads were less than 25 bp were discarded. The cleaned reads were then mapped to the mm9 mouse genome using Tophat v2.1.1. For all libraries, the percentage of mapped reads were between 80-90%. The mapped reads were then converted to fragments per kilobases transcript per million mapped reads(FPKM value) using the well-established *Tuxedo suite* pipeline as described in the paper by Trapnell et al.[64] In house R scripts were used to perform further downstream analysis of the Cufflinks and Cuffdiff files to generate other supporting figures.

3.4 Conclusion

Large proportions of the US population are affected by chronic diseases. Here, one in three adults is obese and one in five children (6 and 19 years) fall into the obese category.[66] Slightly elevated levels of LPS is a characteristic of in humans with chronic diseases, obesity, chronic smoking, and drinking. Elevated circulating levels of LPS in the serum could program innate leukocytes into an unresolved low-grade inflammatory state thus leading to a host of other diseases like atherosclerosis, diabetes, reduced wound healing, Parkinson’s disease, and RA. What’s more, even old age could be a factor with elevated levels of circulating endotoxin. Such primed leukocytes lead to low-level pro-inflammatory phase in the body which refuses to resolve. There is a need to understand the immunological response in the body to subclinical doses of LPS. Thus this transcriptomics study of primed neutrophils and monocytes is hoped to shed some light on how low-doses of LPS affect innate immune cells and what mechanisms or pathways are activated when that happens. Moreover, it also is hoped to shed light on the ameliorating effects of the 4-PBA molecule in restoring homeostasis in bodies plagued by low-level persistent inflammation.

Chapter 4: Production and purification of tn5 for low-cost alternatives to commercial tn5

4.1 Introduction to tagmentation and tn5 enzyme

As we discussed in the preceding chapters, DNA sequencing has become a powerful tool to study and interrogate the human genome (along with the transcriptome, epigenome, exome, etc.). As seen in Figure 22, the cost of sequencing has drastically reduced, surpassing Moore's Law prediction. So much has sequencing costs decreased that at one point in late 2018, Veritas Genetics, an at-home sequencing company was offering whole genome sequencing for just $200. (Source: www.veritasgenetics.com). If we take a closer look at the graph in Figure 22, we see a dramatic drop in cost around 2008. This was the time when the sequencing transitioned from first-generation Sanger sequencing to 'second generation' (or 'next-generation') DNA sequencing technologies like Illumina.

With such reasonably low costs, massively parallel DNA sequencing is adapted more and more by the life sciences research community. The holdup, however, is increasing at the front end (i.e. the high cost of preparing sequencing libraries) and at the back end (i.e bioinformatics analysis and interpretation of large volume of sequencing data) rather than in the sequencing itself. Library preparation means making a DNA fragment ready for sequencing and a typical protocol includes fragmentation of DNA (mechanical or enzymatic), end-repair, adaptor-ligation, PCR amplification, and cleanup with size-selection. Most commercially available kits for preparing libraries come with a very heavy price tag attached ranging anywhere between $20-$50/sample. Some bar-coding platforms such as Drop-seq[23] and 10x Genomics[27] reduce this cost by barcoding and processing thousands of cells in parallel. However, such pooling methods only give rise to modest 30,000-50,000 reads/cell and such shallow depths are not effective for functions such as detecting minute changes in gene expression, detecting rare splice variants etc. [90]

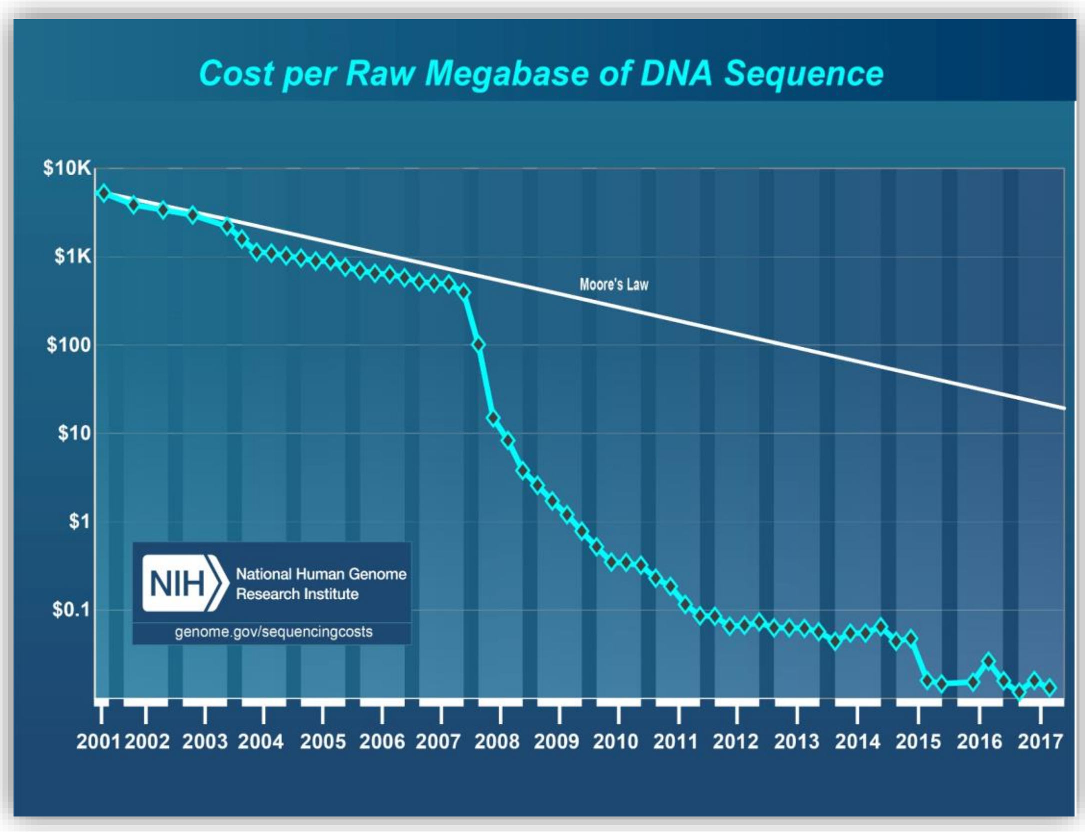

Figure 22: Cost/MB of DNA sequencing from NHGRI Genome Sequencing Program[91]

One of the common library preparation kits is the Nextera library preparation kit by Illumina which costs very close to $50/sample. This kit offers a tagmentation-based library preparation protocol which is pretty robust and yields good quality libraries. It uses the Tn5 transposase enzyme and although efficient, it has a list of undisclosed reagents and high cost which limits the execution of large-scale projects.

The tn5 enzyme is known to be the most expensive component in this kit as most other components of the kit have been deciphered and can be bought individually for a much cheaper library preparation alternative. This enzyme randomly fragments DNA into fragments of smaller size and attaches library adapters in the process. The location where this enzyme cuts DNA has usually a mild preference towards GC rich regions with the target site being AGNTYWRANCT, where N is any nucleotide, R is A or G, W is A or T, and Y is C or T.[92] This preferential bias, however, is low, and almost indistinguishable from mechanical shearing or enzymatic fragmentation of DNA.[93]

In its natural form, a “cut-and-paste” mechanism is used by the Tn5 enzyme to catalyze the translocation of transposable elements in the genome. By natural design, the activity of the Tn5 transposase is low to prevent any accidental damage to DNA. However, over the years, some groups have created a hyperactive version of Tn5 by introducing three point mutations—E54K, M56A, and L372P. Such mutations removes the inhibitory effects of Tn5 activity.[92] The DNA-binding efficiency of Tn5 enzyme is greatly increased by the E54K mutation. This mutated hyperactive form of tn5 is what was used by Illumina to develop their Nextera library preparation kits where the tn5 cuts and ligates the adapter sequences.

Picelli et al.[94] published a protocol outlining the production and purification of hyperactive Tn5 in-house. They used this home-made Tn5 for library preparation. However, Hennig et al.[95] found difficulties in purifying their Tn5 protein construct found low yields and failed to reproducibly obtain libraries for NGS. They published their own modified protocol which was reported to be more reproducible.

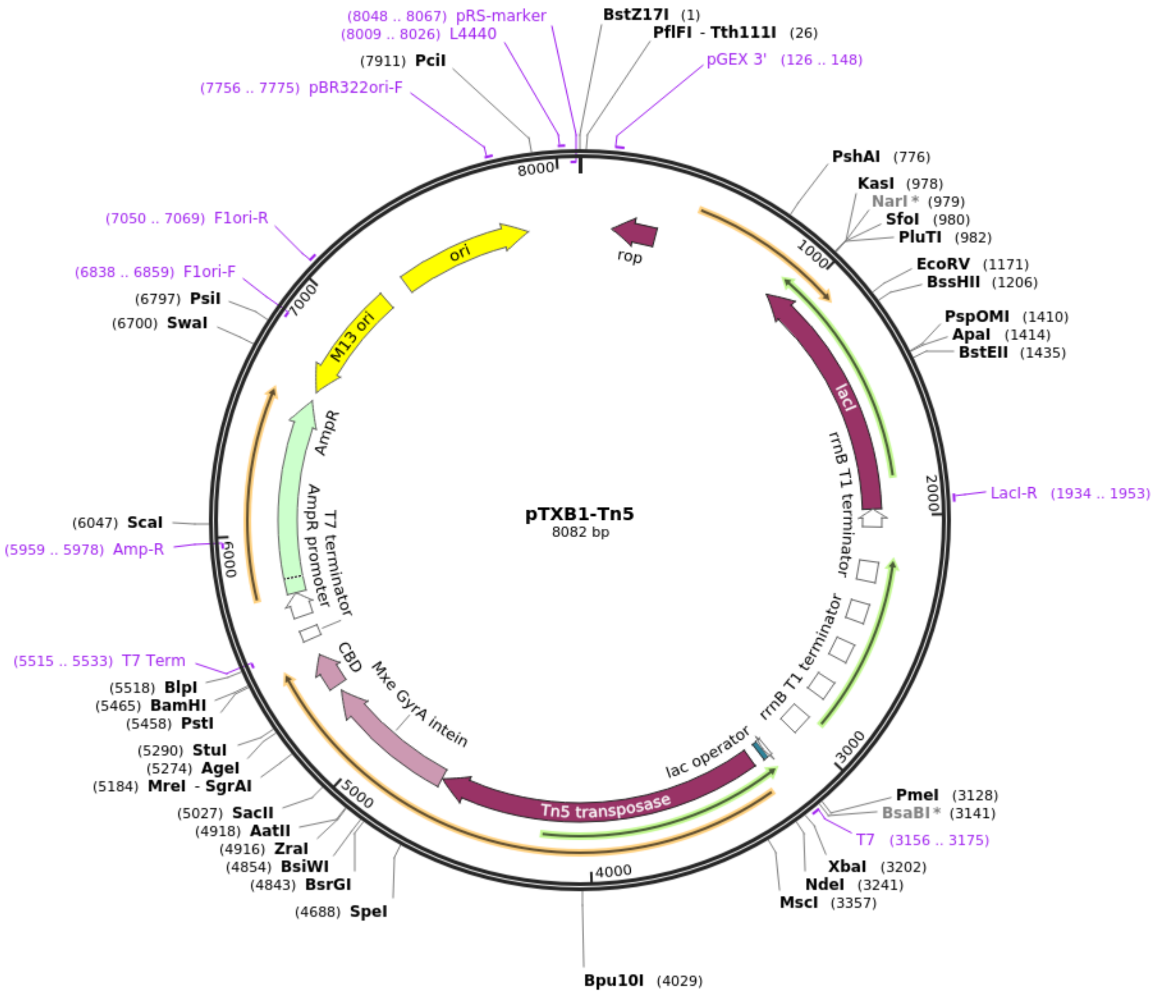

Figure 23: The Addgene full Sequence map for pTXB1-Tn5 plasmid (source:addgene.org)

Here, we tried to recreate and modify the entire library preparation procedure using a homemade tn5 enzyme with an aim to reduce costs for library preparation. We combined the best practices of both Picelli's and Hening's protocol and added on some of our own modifications. A complete and detailed step by step procedure of the process was devised to add in any missing details of the protocol and to document tribal knowledge. We report well-tagmented DNA from a 1ng input which is similar to the input amount suggested by the Nextera XT kit(Illumina) and the reduction of library preparation costs by 50 fold.

4.2 Methodology

4.2.1 Production of tn5 transposase

The pTXB1 plasmid with the Tn5-intein-CBD fusion construct and the hyperactive Tn5 allele with the E54K and L372P mutations was purchased from Addgene (Figure 23, Plasmid # 60240). The stab culture obtained from ADDGENE was streaked onto an LB agar plate, 100 µg/ml Ampicillin plate and grown overnight at 37°C. Colonies were picked and grown in 5 ml of LB medium with 100 µg/ml Ampicillin overnight at 37°C. The vector clones were transformed into T7 Express lysY/Iq Competent E. coli (NEB, # C3013I) following manufacturer's recommendations. The transformed bacteria were streaked on a plate and the excess cells were preserved as glycerol stocks. One colony was picked and grown in 50 ml LB with 100 µg/ml Ampicillin overnight at 37°C,200 rpm. 10ml of the overnight culture was used to inoculate 1 L of LB medium with 100 µg/ml Ampicillin. The culture was grown at 37deg with 200 rpm shaking until the OD@600 reaches 0.5 OR Absorbance@600 reaches 0.9. The culture was then chilled to 10°C for 30min and 250 µl of 1 M IPTG was added to the 1L culture to induce protein expression. The culture was maintained at 23°C for 4 h until the culture reaches A600=3.0. Bacteria were harvested by centrifugation (5000 g, 4°C, 15 min). This is a safe stopping point and pellet can be stored at -80°C.

If frozen, pellets were thawed on ice, resuspended in 80 ml of cold HEGX buffer (20 mM HEPES–KOH, pH 7.2, 0.8 M NaCl, 1 mM EDTA, 0.2% Triton X-100, 10%

Glycerol) with 1% Complete Protease Inhibitor Cocktail (Sigma, # 4693132001). Cells were lysed in 10–12 cycles of 45–50 bursts with 50% duty cycle at output 7 on a Branson sonicator with a 10-mm tip. The temperature during sonication was kept low by cooling in an ice-salt mixture. The lysate was centrifuged in the Beckman JA17 rotor at 15,000 rpm for 30 min at 4°C. The supernatant was collected and 2.1 mL 10% neutralized PEI (must be adjusted beforehand to pH 7.0, from Sigma P3143) was added dropwise on a magnetic stirrer. The precipitate was pelleted by centrifugation at 12,000 rpm for 10 min at 4°C in the Beckman JA17 rotor. Chitin resin was purchased from NEB(S6651S) and packed it into a Pierce Thermo Fischer chromatography column following manufacturer's protocol. The chitin column was washed with 10 column volumes of or HEGX buffer to equilibrate it. The crude cell extract was then loaded onto the column. Typically, 10mL of chitin resin (20mL of chitin slurry) was sufficient to bind 1L of cell culture. The column was washed with 20 to 30 column volumes of HEGX to wash off unbound protein. Next, the chitin resin in the column was submerged in HEGX supplemented with 100 mM DTT. One column volume of solution was let out to make sure the entire resin is surrounded by the DTT buffer. Next, the column outlet was closed and allowed to incubate 36-48 h at 4°C to induce cleavage of the intein tag that binds the protein to the chitin. The column was then washed out with HEGX buffer with 1% PIC in 1 mL aliquots, typically one column volume worth. Each 1mL aliquot was analyzed using a Bradford protein assay but just analyzed visually to determine the intensity of blue color. The samples with the strongest blue color were pooled. The chitin column was cleaned and regenerated following the manufacturer's instructions and can be reused up to 5 times. The protein was dialyzed versus two changes, one after 2 hours and one overnight, using Tn5 dialysis buffer (100 HEPES-KOH at pH 7.2, 0.2 M NaCl, 0.2 mM EDTA, 2 mM DTT, 0.2% Triton X-100, 20% glycerol) at 4°C with constant magnetic stirring. The dialysis was carried out in a Dialysis Cassette (Thermo Fisher, 30K MWCO). After dialysis, the concentration of the protein was measured on Nanodrop and using the dialysis buffer for blanking the device. Usually, further concentration is necessary using a Pierce Protein Concentrator of 30K MWCO at 4°C until the concentration reaches 1.85 mg/mL (A_{280} = 3.0). This protein is very susceptible to degradation so must be maintained at low temperatures continuously. This protein can be stored at -20°C as a glycerol stock which can be prepared by adding 1.1 vol of 100% glycerol (it is recommended to measure glycerol by weight as it is very viscous and volume measurements are inaccurate. The density of glycerol is 1.26 g/cm^3) and 0.33 vol of Tn5 dialysis buffer. For even longer term storage, flash freezing in liquid nitrogen and storage at -80 °C is recommended.

4.2.2 Loading of oligonucleotides and assembly of tn5 transposase

Library adapter oligonucleotides(Tn5ME-A,Tn5ME-B,Tn5MErev) were ordered from Integrated DNA Technologies, with the sequences as mentioned Picelli et al.[94] The oligos must be HPLC purified to ensure good quality libraries and they must be obtained in a lyophilized form. The lyophilized oligos were resuspended in annealing buffer (50 mM NaCl, 40 mM Tris-HCl pH 8.0) to obtain a concentration of 100 µM. Next, one volume of Tn5ME-A oligo or Tn5ME-B oligo was mixed with one volume of Tn5MErev in 1:1 ratio and then the mixture was run on a thermocycler with the following program: 95°C 5 min slowly cool down to 65°C (0.1°C/sec), 65°C 5 min slowly cool down to 4°C

(0.1°C/sec). The annealed linker oligonucleotides could be stored long term at -20°C and are best stored in aliquots.

For assembly of the fresh tn5 enzyme, the recipe used was: 0.125 vol of pre-annealed Tn5MEDS-A and Tn5MEDS-B oligonucleotides mixed in a 1:1 ratio, 0.4 vol of 100% glycerol, 0.12 vol Tn5 dialysis buffer (Section 4.2.1) and 0.36 vol of Tn5 enzyme at the concentration of 1.85 mg/mL. If using glycerol stock enzyme from -20°C (as described in section 4.2.1), the recipe used was changed to 0.143 vol Tn5MEDS-A/B oligonucleotides (mixed in a 1:1 ratio) with 1 volume of the tn5 enzyme. The mixture was incubated for 60 min at room temperature (RT): 23°C under shaking conditions of 350 rpm. tn5 gradually loses activity at 23°C so the assembled protein should be used immediately. If planning to store assembled enzyme at -20°C, the glycerol concentration should be increased from 39.6% (currently in the assembled mix) to 50% final concentration. Note any subsequent library preparation protocol needs to be modified accordingly to account for this change in glycerol percentage. Repeated freeze-thaw should be avoided at all costs. A step by step flowchart of protein production has been shown in Figure 24. A detailed protocol is also included in Appendix .

4.2.3 Library preparation using homemade tn5

In order to prepare a high input reaction for DNA tagmentation, the following recipe was used: 14 µL H_2O, 4 µL TAPS-DMF buffer, 1 µL target DNA at 50 ng/µL, 1 µL of the freshly pre-assembled Tn5(Tn5 volume may change if not freshly pre-assembled). The TAPS-DMF buffer was prepared with the following recipe: 50 mM TAPS-NaOH at pH 8.5, 25 mM $MgCl_2$, 50% DMF. The buffer should be at room temperature and the 50% DMF should be added freshly. This can be accomplished by preparing the buffer at a 2X concentration in advance and diluting 1:1 with DMF before use. Tris-HCl at pH 7.5 can also be used instead of TAPS-NaOH at pH 8.5. The reactions were carried out for 7 min at 55°C. The thermocycler must be pre-heated at 55°C before the reaction and gradual heating should be avoided. 5µL of 0.2% SDS was added to stop the reaction for 7 min at 23°C followed by a 10°C hold. 0.5 µL proteinase K (20 mg/mL) can also be added to stop the reaction for 7 min at 55°C just for testing. Proteinase K is expensive and usually not recommended if trying to reduce costs for library preparation.

In case of low input library preparation (i.e. from 1ng or less DNA), the setup of the reactions was as follows: 4ul of 40% w/v PEG (MW > 4000), 4 µL of TAPS-DMF buffer, 1 µL of DNA, variable amount of Tn5 enzyme (0.01–1 µL, variable amount and depends on the amount of DNA,0.5 µL preferred for 1ng of DNA). Depending on the volume of tn5 added, total volume should be made up to 20 µL with TAPS-DMF buffer. PEG solution was placed at 37°C to make it less viscous. The tagmentation reaction was carried out as before and stopped by SDS or Proteinase K.

Once tagmentation is complete, DNA was analyzed on Agilent Tape Station for correct size distribution. Note if stopping the reaction by SDS, it will interfere with the Tape Station reagents can't be carried out without further cleanup so Proteinase K is recommended. KAPA HiFi HotStart ReadyMix (KK2600) along with the appropriate i7 and i5 indices could be used for further amplification and library preparation flowing manufacturer's recommendations included in the kit manual for PCR amplification.

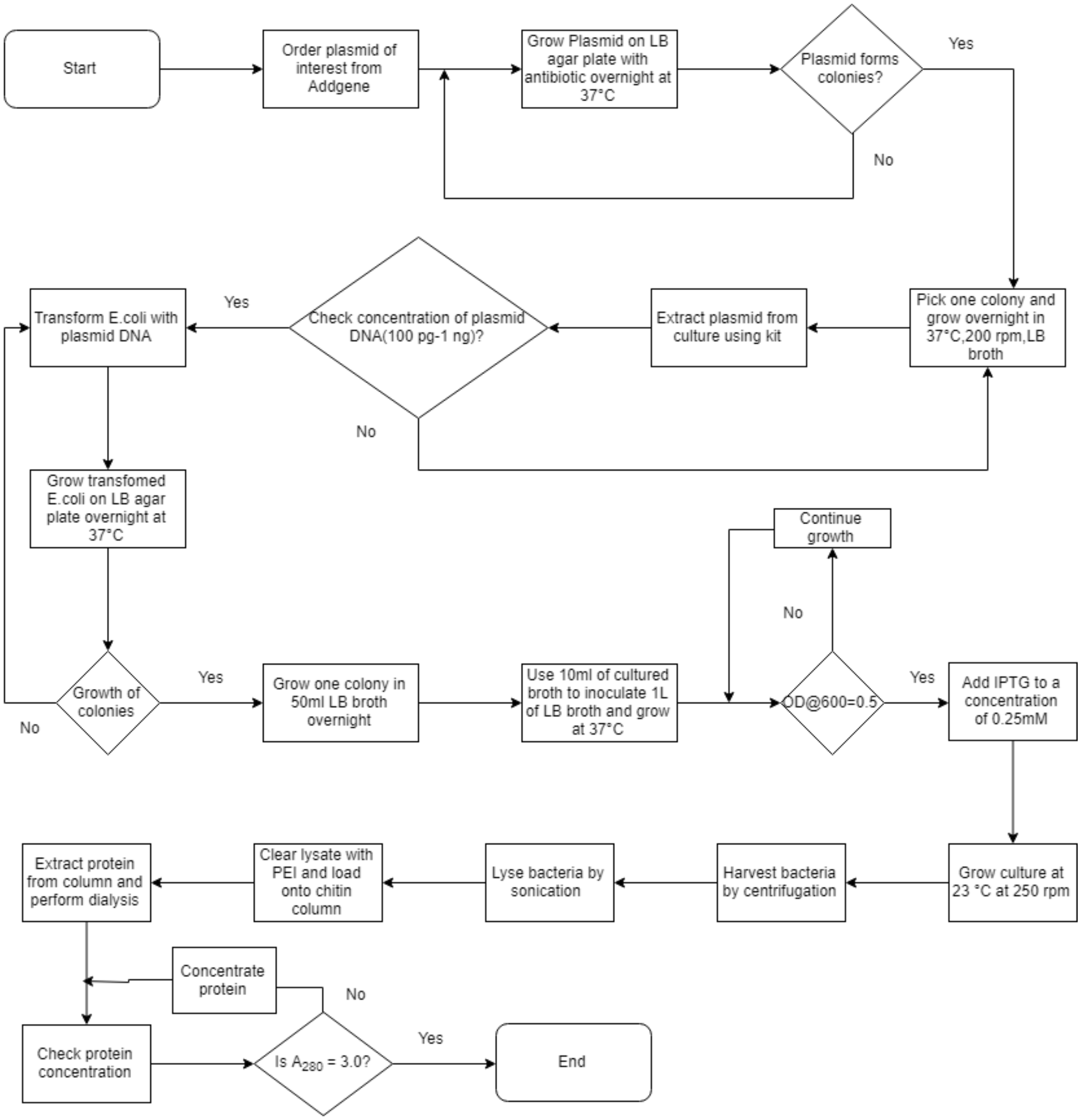
Start
Order plasmid of interest from Addgene
Grow Plasmid on LB agar plate with antibiotic overnight at 37°C
Plasmid forms colonies?
Yes
No
Pick one colony and grow overnight in 37°C,200 rpm,LB broth
Extract plasmid from culture using kit
Check concentration of plasmid DNA(100 pg-1 ng)?
Yes
No
Transform E.coli with plasmid DNA
Grow transfomed E.coli on LB agar plate overnight at 37°C
Growth of colonies
Yes
No
Grow one colony in 50ml LB broth overnight
Use 10ml of cultured broth to inoculate 1L of LB broth and grow at 37°C
OD@600=0.5
No
Continue growth
Yes
Add IPTG to a concentration of 0.25mM
Grow culture at 23 °C at 250 rpm
Harvest bacteria by centrifugation
Lyse bacteria by sonication
Clear lysate with PEI and load onto chitin column
Extract protein from column and perform dialysis
Concentrate protein
No
Check protein concentration
Is A_{280} = 3.0?
Yes
End

Figure 24: A flowchart showing steps involved in producing tn5 protein

4.3 Results and Discussion

To validate tagmentation reactions generated with our Tn5 enzyme and to optimize reaction conditions, we tagmented 50 ng Human DNA (Sigma 11691112001) and observed their tagmentation profile on Tapestation. The tagmentation profiles with different ratios of DNA to the tn5 enzyme are seen in Figure 25. One sample was run with no Tn5 enzyme as control. The tn5 amounts reflect the freshly pre-assembled tn5 at 1.85mg/ml. We also tested tn5 reactions on low-input DNA (1ng) and amplifying it after tagmentation to observe the size distribution(Figure 26). The low input reactions were observed to be more sensitive to the ratio of DNA to tn5. Ultimately, for 50ng reaction, the optimal tn5 amount was found to be 1 µL in a 20 µL reaction (33.3 ng/µL) and for the 1ng reaction, it was found that 0.5 µL of tn5 worked best in a 20 µL reaction (16.65 ng/µL).

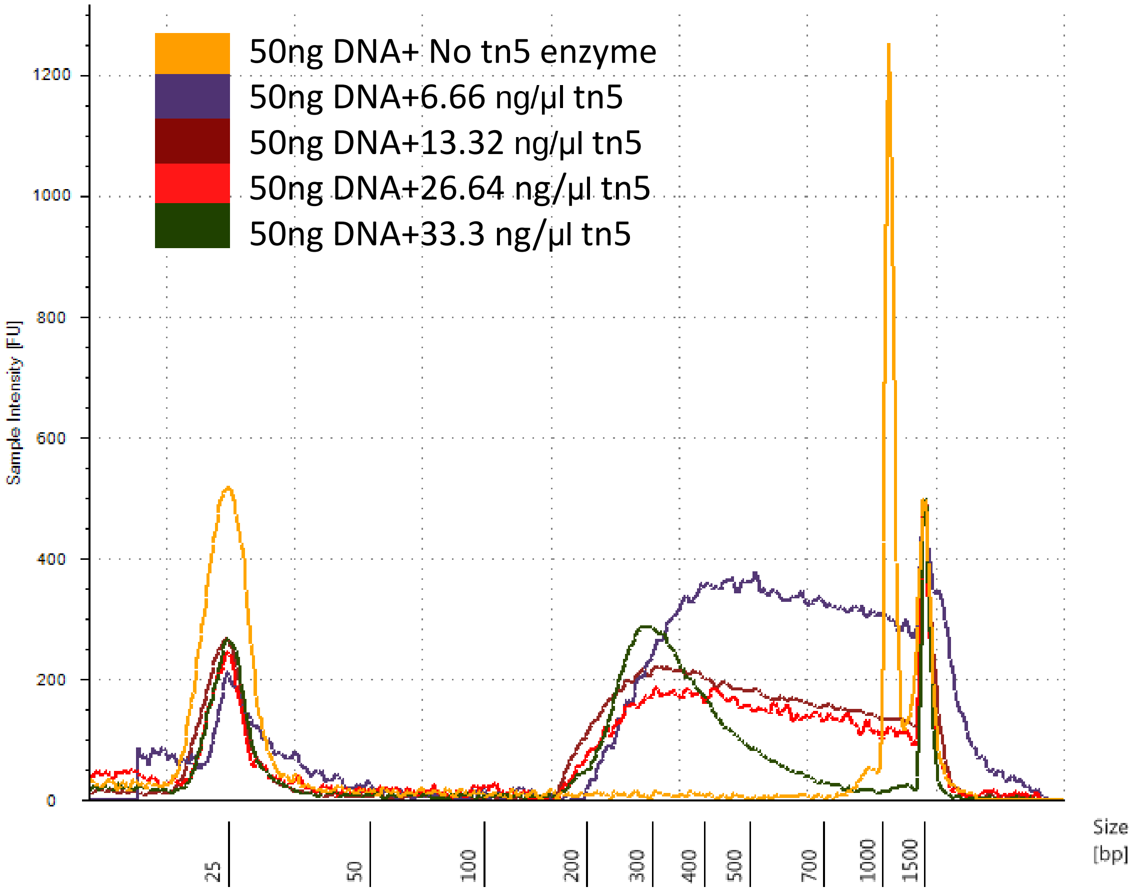

Figure 25: Different tagmentation profiles for 50ng DNA obtained from Tape station

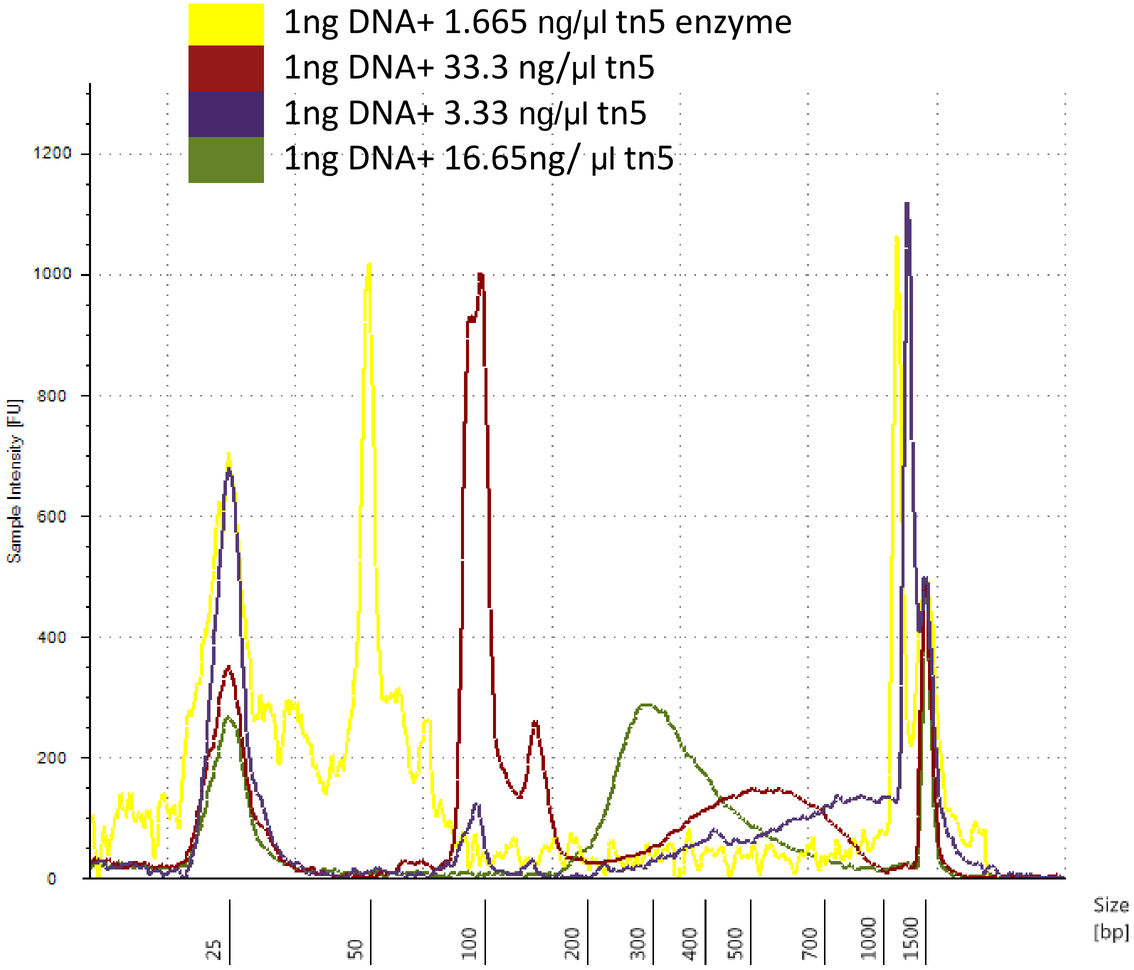

Figure 26: Different tagmentation profiles for 1ng DNA obtained from Tape station

In this chapter, we describe a detailed methodology for the production of in-house Tn5 transposase as a cheap alternative to commercial library kits such as Nextera and Nextera XT by Illumina. The accessibility to home-made Tn5 transposase protein unlocks opportunities for improvement in both library protocol optimization and other applications of Tn5 such as ATAC-seq. We described a successful step-by-step procedure for the production of high-quality Tn5 fragmented DNA which can be used for both low and high input DNA. This will hopefully provide a better idea to researchers trying to recreate this protocol and will remove cost bottlenecks in trying to prepare libraries.

Chapter 5: Low-input epigenomic studies

5.1 Overview of epigenomics

Every cell in an individual more or less has the same genotype but they are quite different in their phenotypes depending on which tissue they come from. This variation rises because each cell has a specific set of instructions on which genes to be turned on and off. This set of instructions is what epigenomics is all about. In Greek, “epi” means above or beyond, thus epigenomics is something beyond the genome. The identity and function of each individual cell depend on this epigenomic influence on its genome. Gene expression controls virtually every process in the body; for example, cell division, development, aging, responses to stimuli, and development of diseases.[96] Epigenomics is what would explain why two identical twins with the exact same genome could have a very different phenotype. Epigenomic processes also have the ability to re-program the genome in response to environmental challenges, such as stress, pollution, exercise, smoking, drinking, diet, etc. Given the power they hold over transcription, the interlacing connections between epigenomic mechanisms and disease are not surprising. Diseases caused by epigenomics could occur due to epigenomic dysfunction caused by mutations or epigenomic mechanisms producing incorrect programming, which are either inherited or programmed from previous life events.

The epigenomic machinery comprises all of the chemical compounds could be added to the DNA as a way to regulate its activity. These chemical compounds are not an active part of the DNA sequence but are usually attached to DNA. Epigenomic modifications or tags are retained as cells divide and can also be inherited through the generations. [97] Although epigenetic tags become more stable during adulthood, they are dynamic and can be modified by changing the environment and lifestyle. There have been abundant examples of how different lifestyle choices can alter epigenetic marks on top of DNA and change health outcomes. The environment strongly controls changes to epigenetic tags and disease susceptibility. For example pollution has been found to be able to alter methyl tags on DNA and give rise to neurological disease.[98] Diet is also an important factor in this regard and the field of nutriepigenomics actually explores how food and epigenetics work in harmony to impact health and disease. Epigenomics controls if the DNA is turned on or off through a variety of mechanisms like histone acetylation, DNA methylation, non-coding RNA, transcription factors and so on.

A striking example of epigenomic marking is that in the study of Weaver et al. (2004) [99] who found that early life experiences can activate the epigenome in ways that can lead to lasting changes in adult health and behavior. In this study, rats that received a loving maternal care in childhood had increased levels of H3 histone acetylation and decreased DNA methylation levels at the promoter of the glucocorticoid receptor which led them to be less stressful adult. The reverse response was seen in rats with low-caring mothers. Thus a number of environmental factors to affect our phenotypes later in life and can even be passed down through generations.

5.2 Different epigenomic mechanisms

A host of non-genetic factors affect cellular phenotypes. Conrad Hal Waddington was the first to coin the term 'epigenetics' way back in 1942 as the mechanism “by which the genes of the genotype bring about phenotypic effects”.[100] Since then, our understanding of how different epigenomic mechanisms work in conjunction to bring about changes has increased manifold. This section is dedicated to understanding them in greater detail.

5.2.1 DNA methylation

The first epigenomic mark to garner attention was DNA methylation. This is still the best studied and researched epigenomic mark. It was discovered in the late 1940s and very early on it was recognized as an important agent controlling gene expression as it was unevenly distributed in the genome and because it could be inherited.[101] Methylation involves the addition of a methyl group to the carbon-5 position of the cytosine pyrimidine ring. Such methylation is usually observed in GC rich regions of the genome called CpG islands. CpG islands are commonly found in the 5′ regulatory regions of genes. Enzymes called DNA methyltransferases (DNMTs) catalyze methylation.[102] The first associations between gene expression and DNA methylation were observed on important model loci (that included chicken and mammalian globin genes, the X-chromosome inactivation centre (XIC), and other regions in the chromosome which were known candidates of genomic imprinting). Experimental associations were confirmed when DNA methyl transferases were knocked out or deleted. Bacteria can cause changes in DNA methylation patterns of infected cells as they are providers of active metabolites such as folate, butyrate, and acetate which catalyze DNA methylation. [102] Many types of cancers have been linked to DNA methylation making this an important indicator of disease.[103] Hypermethylation of CpG islands leads to the downregulation of tumor suppressor genes. On the other hand, global hypomethylation leads to genomic instability and is responsible for the activation of oncogenes.

5.2.2 Chromatin remodeling and histone modification

Histone acetylation is another important epigenetic mechanism which transfers an acetyl group from acetyl coenzyme A (acetyl-CoA) to lysine residues of histone proteins along with the simultaneous production of CoA.[102] Histone acetylation is regulated by the opposing action of two enzymes called histone acetyl transferases (HATs) and histone deacetyl transferases(HDACs). Histone acetylation reduces the electrostatic affinity between histones and DNA, and thus encourages a chromatin structure that is more accessible to transcription. It is thus a transcriptional activator, and the N-terminal tail of histones is an important site for acetylation to occur. Acetylated histone tails are more accessible to transcription factors. Histone deacetylation, however, removes the acetyl (acyl) moiety from lysine residues and thus downregulates transcription. Histone methylation of lysine and arginine residues is also another important epigenetic mechanism which initiates the binding of certain transcription factors thus promoting transcription. Some other histone tail modifications include histone phosphorylation, ubiquitination, sumoylation, histone poly-ADP ribosylation, histone biotinylation, citrullination, and proline isomerization.[102] Each of these tends to affect DNA transcription in distinct ways. On top of that, the cross-talk between histone modification and DNA-methylation has also been observed.[104]

5.2.3 Non-coding RNA

Only a small part, less than 5%, of RNA molecules in a cell account for messenger RNA(mRNA) that codes for proteins. RNA that does not code for a protein falls under the category of non-coding RNA (ncRNA). A major part of the eukaryotic genome is transcribed into ncRNA which is mainly involved in the modulation of messenger RNA (mRNA) translation. ncRNA is classified into small and large ncRNAs.

microRNAs (miRNAs) are about 18–20 nucleotides in length fall under the class of small ncRNAs and are involved in gene regulation.[105] Long ncRNAs, on the other hand, have a length of 200 nucleotides or more. These long ncRNAs have roles in high-order chromosomal dynamics, telomere biology, and structural organization inside cells. [106]

5.3 Transcription Factors and their functions

A fourth epigenomic mechanism is through the involvement of transcription factors(TF). Transcription is a process which takes place in the nucleus in which genetic information contained in a DNA strand is converted into a complementary strand of RNA which is then translated into a protein. Transcription is an interplay of cis- and trans-acting factors within the nucleus that orchestrates gene expression.

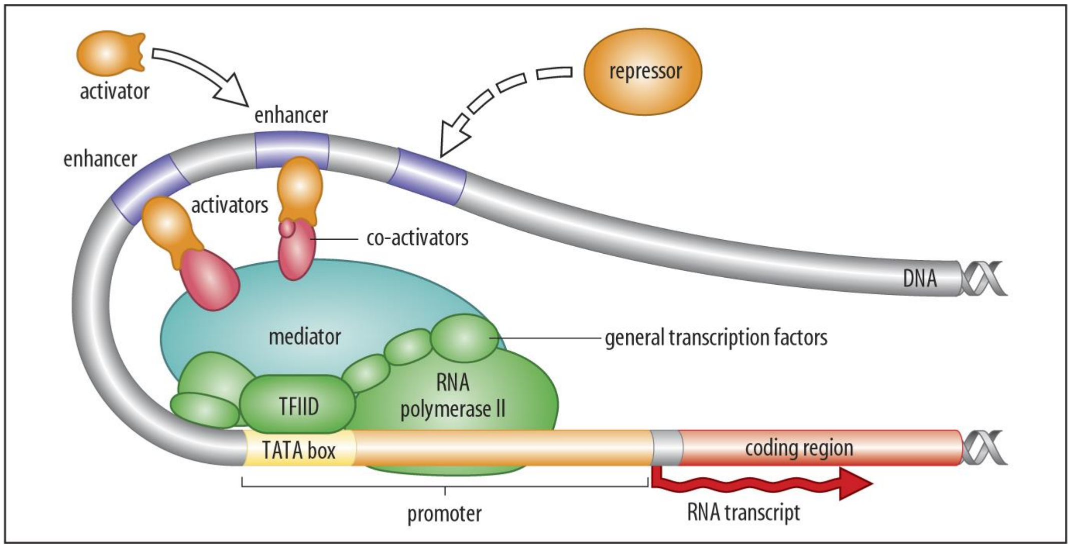

Figure 27: The control of transcription via DNA looping(From www.mun.ca,Dr.Brian E. Staveley)

The cis-acting elements which include promoters and enhancers are non-coding DNA elements located upstream on the same chromosome of the gene. Transcription factors fall under the category of trans-acting factors. The process of transcription begins only when a set of transcription factors and a special protein called RNA Polymerase II binds to the promoter. The promoter is a region upstream of a gene which is further divided into two types - the core promoter and the proximal promoter. The core promoter is within 35bp of the transcription start site (TSS) contains regions where RNA Polymerase II (the enzyme that performs transcription) and the general transcriptional factors bind. The proximal promoter is the site where sequence-specific transcription factors bind and is usually located ~300 bp upstream of the core promoter. A transcription factor that doesn’t bind to the proximal promoter region can bind to a distal region called an enhancer which may be located several Kb away from the gene. Such transcription factors then cause DNA looping to attach to the promoter region thus controlling transcription (*Figure 27*).

About 10% of the human genome codes for transcription factors and there are about 2600 known transcription factors.[107] Mutations in TFs and TF-binding sites are the cause of many human diseases. The same TF can control the function of more than one gene thus hinting towards a dynamic transcriptional regulatory network within the same organism.[108] Understanding the mechanism by which TFs recognize binding sites and modulate transcription is a formidable task. Many TFs bind to DNA by recognizing motifs which are short sequences (6–12 bases) in the genome. There could be multiple motifs for the same TF. About 1,211 of all the human TFs currently have a known binding motif.[108] Most of the TF DNA binding motifs are highly conserved among species.

Transcription factors proteins have a modular structure and contain three main domains: 1) DNA-binding domain (DBD), which attaches to a particular section of DNA either in the enhancer or promoter regions. 2) Trans-activating domain (TAD), which usually binds other proteins such as transcription coregulators. 3) An optional signal-sensing domain (SSD) (also called a ligand binding domain), which binds external signal molecules and activates the transcription factor.

5.4 Low input ChIP-Seq for transcription factors

ChIP-seq[109] has been a revolutionary technique to the study the binding of TFs in vivo. In the ChIP-seq technique, intact cells are 'fixed' with formaldehyde, which creates a reversible protein-DNA cross-link and preserves the TF-DNA binding scenario in vivo. The cells are then lysed and chromatin is subjected to sonication to break them up into smaller fragments. Next, the TF of choice is immunoprecipitated by utilizing specific antibodies conjugated to a magnetic bead. The crosslinks between protein and DNA are then reversed releasing the immunoprecipitated DNA which can then be prepared into a library and sequenced.

Understanding the interaction between TFs and their binding sites in the genome yields important insights into its effectiveness and the transcriptional regulatory mechanisms. There have been a number of molecular techniques for identifying the interaction between regulatory proteins and DNA in general. However, in vitro methods such as electrophoresis mobility shift assays (EMSA) and DNase 1 protection assay cannot properly validate in vivo relevance[110]. In comparison, chromatin immunoprecipitation followed by sequencing(ChIP-seq) has become the preferred method for studying DNA-protein interactions in vivo [111-114]. Common ChIP assays confirm potential interactions between TFs and promoters of interest by conducting locus-specific qPCR analysis. On the other hand, unbiased and genome-wide mapping of TF binding is also possible with techniques such as ChIP-ChIP and ChIP-seq[113]. Such examination provides a snapshot of the dynamic processes of TF recruitment and regulation, without overexpressing any component.

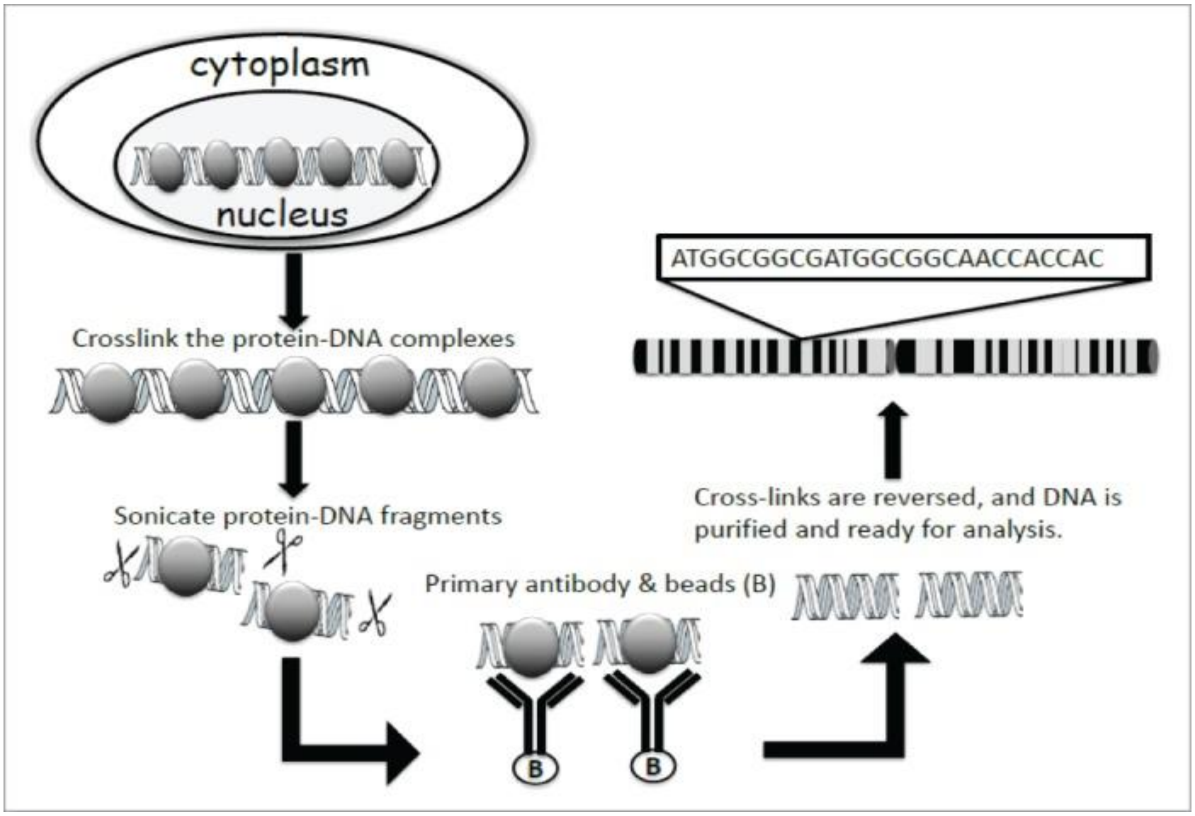
cytoplasm
nucleus
Crosslink the protein-DNA complexes
Sonicate protein-DNA fragments
Primary antibody & beads (B)
B
B
ATGGCGGCGATGGCGGCAACCACCAC
Cross-links are reversed, and DNA is purified and ready for analysis.

Figure 28: Basic steps in ChIP-seq protocol. Figure under CC By License from Mundade et al.(2014)[115]

Although ChIP-qPCR/seq has been used very successfully for epigenomic studies, the technology has certain limitations. First, the main drawback is the requirement of a large number of cells (>10^6 cells per IP for ChIP-qPCR and 10^7-10^8 cells for ChIP-seq)[114, 116, 117]. This is not a big issue when working with cell lines for ChIP-seq assay. However, such a requirement becomes a major hurdle when working with primary cells. For example, there are around 10,000 naturally occurring T regulatory cells per murine spleen, and ~5000 per ml peripheral blood .[118] In metastatic cancer patient samples, only about 1–10 circulating tumor cells can be isolated per ml of blood. Additionally, primary samples are usually a heterogeneous mixture of different cell types. And if one tries to separate a particular type of cell from such a mixture, it leads to a further reduction in the number of available cells. Second, population heterogeneity in a group of cells may lead to significant standard deviation among different trials of a ChIP-seq assay. Lastly, most ChIP-seq assays are very labor-intensive and last about 3–4 days. These unwieldly procedures lead to a loss of precious sample and introduce human errors or technical/systemic errors that lead to low reproducibility among replicates[119, 120]. Thus, there is a need for ChIP-seq assays with high sensitivity, reproducibility and automation.

Some studies notably that by Brind'Amour et al. (2015) [83] have talked about carrying out ultra-low input CHIP-Seq from a 1000 cells using Native ChIP i.e. ChIP without cross-linking with formaldehyde. Here, micrococcal nuclease enzyme is used to cleave the DNA at the specific locations leaving nucleosomes (histones) undamaged and giving rise to DNA fragments spanning the length of one(200bp) to five nucleosomes (1000bp). However, this protocol can only be mainly used for studying histone modifications which are more robust and can work even without cross-linking to the genome. Buenrostro et al. in 2013 came up with a protocol of ATAC-seq[121] where hyperactive tn5 transposase enzyme was used to cut and attach adaptors to accessible regions of chromatin with as low as 500 cells. The use of sample barcoding technology in conjunction with microfluidics has been used by Rotem at al. (2015) to achieve single cell sensitivity on histone modifications.[122] These methods although effective on histone modifications, are not suitable to study TF interactions as they are more transient and less abundant. Because transcription factors bind to a small fraction of the whole genome, most conventional ChIP-Seq techniques for transcription factors require millions of cells below which it is difficult to distinguish signal from noise. The reason behind the requirement of a large number of cells include i) only 1-10% of the cross-linked chromatin is recovered and ii) the wash steps during immunoprecipitation to reduce nonspecific interactions lead to a loss of specific interactions thereby decreasing the signal to noise ratio.[123] The limited availability of patient tissue also doesn't help. There have been several attempts to decrease the amount of starting material required for ChIP-seq. Some groups have reported ChIP-Seq from limited cell numbers like 10,000 cells [124] and 5,000 cells [125]. Zwart et al. (2013) used a carrier molecule (mRNA and Histone H2B mix) which enabled ChIP-Seq from 10,000 cells. How the carrier molecule exactly works to enhance ChIP-Seq signal is not clearly known but it is hypothesized that the carrier molecule being bulky it helps to retain the small amounts of important chromatin throughout the procedure and the carrier also reduces nonspecific binding by acting as a competitor for nonspecific binding sites on the magnetic beads and elsewhere. Shankaranarayanan et al. (2011) developed their own method for studying TF binding for as low as 5000 cells using their own amplification strategy called Single-tube linear DNA amplification (LinDA). In this method, DNA was ligated with poly T and in vitro transcribed to RNA. The RNA was then reverse transcribed and amplified using the T7 promoter-BpmI-oligo (dA) 15 primer. In both of these tube-based methods, numerous manual handling steps lead to significant user error and loss of ChIP DNA. Under such circumstances, microfluidics provides a powerful platform for parallel experiments and minimized manual hands-on interference. It also reduces sample and material consumption and improves the sensitivity of the ChIP-Seq assay. Several groups have tried to utilize these advantages by performing

CHIP-Seq on microfluidic devices from limited input. Several groups have successfully performed low-input ChIP-Seq of histone modifications. Shen et al. (2015)[126] claimed to achieve a sensitivity of 1000 mouse early embryonic cells to study histone marks and Cao et al. (2015) [127] reached sensitivity of a 100 cells to study histone modification in GM12878 cells and fetal liver cells. However, not much progress has been made in the field of low-input ChIP-Seq for TFs. As late as the last year 2018, Dahl et al. have reported that no microfluidic platform has been developed to study TF binding.[128] Thus this field provides us ample opportunities for development and research for limited input studies of TFs.

Chapter 6: Low input ChIP-seq of Estrogen Receptor (ER) α using MOW-ChIP

6.1 Introduction to Estrogen Receptor α ChIP-seq

Breast cancer today has become a worldwide catastrophe and in the US, a woman has a 12.3%, or a 1 in 8, lifetime risk of being detected with breast cancer.[129] More than 80% of breast cancer tumors are categorized as 'ER-positive' tumors. (source: www.nationalbreastcancer.org). Estrogen binding to the ERα receptors stimulates proliferation of the mammary cells which leads to increased DNA replication thus increasing the probability of mutations leading to breast cancer. The treatment strategy for such tumors is usually via endocrine treatment where receptor activation is inhibited in some way or form, each of which results in an inhibition of ERα-driven cell proliferation. One of the most commonly used breast cancer drugs called Tamoxifen is a Selective Estrogen Receptor Modulator (SERM) which induces an alternative conformation of the ligand-binding domain that results in the repression of transcription instead of activation. Some other breast cancer therapies, especially in postmenopausal women, work by inhibiting the secretion of estrogen hormone in the body. However, resistance to endocrine treatment is very commonly observed in ERα positive breast cancers. This resistance can take place through different mechanisms, some of them are the differential expression of kinases, coregulators, and transmembrane receptors.[124] It has also been reported that intrinsic ERα/chromatin interactions can give rise to resistance to treatment. About one-half of the patients with estrogen-receptor-positive breast cancer fail on tamoxifen.[130] It is very important therefore to study the mechanism of ER function especially in the case of patients who have displayed resistance to therapy.

Microfluidic oscillatory washing based ChIP-seq (MOWChIP-seq) with the sensitivity of a 100 cells was developed in our lab to study histone modification in GM12878 cells and fetal liver cells.[127] In this process, a packed bed of beads for ChIP when combined with oscillatory washing for removing nonspecific adsorption/trapping led to an exceptionally high yields of desired ChIP DNA. On the other hand, Zwart et al.[124] used a carrier molecule to carry out low-input TF ChIP-Seq to increase the signal to noise ratio from 10000 cells. A third strategy reported by Singh et al.[131] involves the use of double fixation using DSG and formaldehyde for ERα ChIP-seq from postoperative tumor tissue.

In this chapter, we present a combination of these three strategies to further lower the input cell number required for ER α ChIP-seq. We present some preliminary data with 10k, 7.5k and 5k cells of with a potential to further decrease cell number and data quality. Such sensitivities will be several orders of magnitude higher than the prevailing good quality ER α ChIP-seq assays. By conducting the immunoprecipitation (IP) in a tiny microfluidic chamber and due to the high degree of automation, we drastically shortened the time for IP to less than 1.5 hours (compared to overnight in most ChIP protocols). Our microfluidic technology is superior to conventional ChIP-seq for a couple of reasons. First, high concentrations from trace amounts of molecules could be built up inside the tiny volumes offered by the microfluidic chamber. Adsorption kinetics and completeness was facilitated by such high concentration. Second, amount of IP beads take up a large fraction of the tiny volume so that the surface area/volume ratio (15-40%) is tremendously improved when compared to 5% in the conventional ChIP[132]. The presence of beads in the local vicinity significantly increased the efficiency and rate for attachment of chromatin to the magnetic bead's surface. This was attributed to the presence of short diffusion lengths among beads.[118] The rapid adsorption kinetics of chromatin on beads was because travel time follows the relationship of $\tau_D \sim w^2/D$, where w is diffusion distance between two beads, and D is diffusivity.[127] Third, by using microfluidic technique uniquely suited for bead manipulation at the microscale, we effectively remove nonspecific adsorption after high-efficiency adsorption using microfluidic oscillatory washing. This is critical for producing high-quality ChIP DNA that preserves desired biological information. Finally, the microfluidic device integrates various steps and minimizes material loss among steps. The use of a carrier mix further improves IP efficiency. The carrier can be easily removed after IP. Formaldehyde is typically used in protein-DNA cross-linking. However, TF proteins are characterized by rapid binding dynamics their interactions with chromatin is very transient. Thus TFs are usually not cross-linked efficiently, which negatively affects ChIP efficiency.[133] Also, formaldehyde being is a short-range cross-linker is not very efficient at stabilizing protein-protein interactions. As a large number of proteins constitute the ERα transcription factor complex, using DSG as an additional fixative to stabilize the entire complex alongside standard Formaldehyde fixation would lead to more stable protein-DNA interactions.

Thus the combination of all three strategies is indicative of yielding unprecedented sensitivities in ERα ChIP-seq will permit some transformative applications of ChIP-seq technology to primary samples. Transcriptional regulation and mechanism during disease development from limited cell samples of a patient that used to be not accessible to the researchers and clinicians due to the technological limitation could then become attainable. Armed with such information, one can better understand the role of epigenetics in the diagnosis and prognosis of various diseases which will in turn help devise strategies for personalized medicine.

6.2 Materials and methods

6.2.1 Fabrication of the microfluidic ChIP device

MOW-ChIP device[127] is composed of a microfluidic chamber (~800 nl), connecting channels, and a micromechanical valve that can be partially closed to stop magnetic beads while allowing liquid to pass. The main chamber is an elliptical shaped chamber with a major axis of 6 mm, a minor axis of 3 mm, and a depth of 40 µm. 27 micro-pillars are spotted inside the main chamber to prevent the upper layer from collapsing.[118]

Multilayer soft lithography was used to fabricate the microfluidic ChIP device. Briefly, two photomasks (one for fluidic layer, and one for control layer) were designed with computer-aided design software Layout Editor and printed on high-resolution (5,080 d.p.i.) transparencies. To make fluidic layer master (~40 µm thick), photoresist (SU-8 2025, Microchem) was spun on a 3-inch silicon wafer (978, University Wafer) at 500 rpm for 10s and 2500 rpm for 30 s followed by soft baking at 65°C for 1 min and 95°C for 7 min. To make control layer master (~50 µm thick), SU-8 was spun at 500 rpm for 10s and 1500 rpm for 30s and followed by same soft bake condition. Each master covered with its photomask was UV exposed for 17s at 580 mW exposure intensity and followed by a post-exposure bake at 65°C for 1 min and 95°C for 7 min. Masters were then developed in SU-8 developer for 2-3 min, rinsed with IPA and air blown to dry. To make fluidic layer stamp, PDMS (General Electric silicone RTV 615, MG chemicals) with a mass ratio of A: B = 5:1 was thoroughly mixed and vacuumed for 1 hr. it is then poured onto the fluidic layer master in a Petri dish to a height ~5 mm thick. To make control layer stamp, PDMS with a mass ratio of A: B = 20:1 was mixed, vacuumed for 60 min, and spun onto the control layer master at 1100 rpm for 35s to a height of 108 µm thick. Both stamps were partially cured at 80 °C for 30 min. The fluidic layer was then peeled off from the fluidic layer master and aligned to the control layer. Two-layer PDMS were thermally bonded by baking at 80 °C for 60 min and then peeled off from control layer master. Inlets and outlets of the device were punched by a 2 mm hole puncher. The cleaning solution used for glass slides was H_2O: 27% NH_4OH: 30% H_2O_2 = 5:1:1, volumetric ratio, 75 °C for 2 h. Slides were then rinsed with ultrapure water and carefully air dried. Finally, the two-layer PDMS device and a pre-cleaned glass slide were bonded together via plasma bonding using an oxygen plasma cleaner (PDC-32G, Harrick Plasma). Device was then baked at 80 °C for 60 min to strengthen the bonding between PDMS and glass.

6.2.2 Setup of the microfluidic device

The microfluidic experiment was observed by a charge-coupled device (CCD) camera (ORCA-285, Hamamatsu) attached to the port of an inverted microscope (IX 71, Olympus). The experiment started with prefilling the control channel with water to prevent air bubble defusing into the fluidic channel. The reagents were flown into the inlet via perfluoroalkoxy alkane (PFA) high-purity tubing (1622L, ID: 0.02 in. and OD: 0.0625 in., IDEX Health & Science) driven by a syringe pump (Fusion 400, Chemyx). The micromechanical valve was activated by a solenoid valve (18801003-12V, ASCO Scientific), which was connected to a pressure source (a gas cylinder or a compressed air outlet) and controlled by a data acquisition card (NI SCB-68, National Instruments) and a LabVIEW (LabVIEW 2012, National Instruments) program for its switching function. The pressure (30 – 35 psi) that was applied to control channel deformed PDMS membrane between the fluidic channel and control channel and formed a partially closed valve to stop beads while allowing fluids to pass. The oscillatory washing was conducted by connecting the inlet and outlet of the microfluidic device to solenoid valves via PFA tubing. A digital pulse signal was first created in LabVIEW program, and it got converted to an electric signal by the data acquisition card and then it was sent out to solenoid valves to achieve automation.

6.2.3 Cell Culture

The cell line used for this ChIP-Seq study was MCF-7. MCF-7 is a breast cancer cell line isolated in 1970 from a 69-year-old Caucasian woman. MCF-7 cells are useful for *in vitro* breast cancer studies because the cell line has retained several ideal characteristics of the mammary epithelium which include the ability to process estrogen in the form of estradiol via estrogen receptors in the cell cytoplasm. Thus MCF-7 cell line is an estrogen receptor (ER) positive cell line. It is noteworthy that MCF-7 is also progesterone receptor positive and HER2 (Human Epidermal Growth Factor) negative. This cell line has been studied for over 40 years and is a well-established cell line for studying breast cancer. In order to test our microfluidic low input ChIP-Seq, this was the cell-line of choice. MCF-7, an adherent cell line, was cultured in DMEM (ATCC) with 10% FBS and 1%PS at 37°C and 5% CO_2. Cells were harvested at 90% confluence. In case of treatment with Estrogen, Phenol Red Free DMEM with 5% charcoal-stripped FBS and 1% PS was prepared. At 80% confluence, the original media was discarded and the cells were rinsed with PBS twice. The hormone starvation media was added, and cells were starved for at least 72 hours. 20ug/mL of 17 beta-estradiol(Sigma) stock solution was prepared by adding 1 ml absolute ethanol to 1 mg β-estradiol; gently swirled to dissolve, and then 49 ml of sterile phenol-red free media was added while mixing. The MCF-7 cells were treated with 10 nM of E2 solution (by appropriate dilutions of stock solution with phenol-red free media) for 1 hr. prior to the ChIP-seq procedures.

6.2.4 Preparation of ERα chromatin

Stock chromatin was prepared from 1 million cells. Media was aspirated and the cells were washed twice with PBS at room temperature by centrifugation and resuspension. Formaldehyde was freshly diluted to 1% using PBS and the cells were incubated at room temperature for 10min on a rocking platform. For cells fixed with DSG + formaldehyde, the medium was aspirated and the cells were covered with PBS containing 2 mM DSG for 35 min at room temperature, after which formaldehyde was added to 1% final concentration, and the adherent cells were further incubated for another 10 min at room temperature. In both cases, formaldehyde was quenched with by adding glycine (the glycine shouldn't be prepared more than 1 month in advance) to a final concentration of 0.125M and shaken for 5 min-10 min at room temperature on the rocker. Cells were washed twice with PBS to remove any residual formaldehyde and scraped into PBS with (+1X PIC +5mM sodium butyrate) and collected by centrifugation (1600g for 5 min at 4℃). The cell pellet was then centrifuged and resuspended in 130μL Sarkosyl Buffer (0.1% SDS, 1% Triton X-100,10 mM Tris HCl--pH7.4, 1mM EDTA-pH 8.0, 0.1% Na Deoxycholate, 0.25% Sarkosyl, 0.3M NaCl + 1X PIC+ 5 mM sodium butyrate) and were then put on ice for 10-15 min. Post the cross-linking process, chromatin was sonicated with a Covaris M220 with peak incident power 50, duty factor

20%, cycles per burst 200, time, temp 20°C for 15 minutes. After sonication, the solution was centrifuged for 10 min at 4°C, 14,000 x g. The supernatant was carefully transferred to a new tube. This lysate contained 1 million cells. An appropriate fraction was used for each reaction. For example, for 10,000 cells reaction, 1/100 of this solution was used in a pool-split manner.

6.2.5 Preparation of immunoprecipitation (IP) beads

On ice, 2.5 μL of Protein A (Thermo Fisher Cat. No. 10001D) and 2.5 μL of Protein G (Thermo Fisher Cat. No.100003D) were mixed together and placed on a magnetic rack. The supernatant was removed and the beads were washed 3 times with 125 μL of ice-cold PBS+ 0.5% BSA. The beads were then resuspended in 125 μL of PBS-BSA. ERα Antibody (Santa Cruz Biotechnology Cat. No. sc-8002) were added to the bead solution and the whole mixture was gently incubated with the antibody at 4°C on a rotator mixer at 24 rpm overnight. The antibody coated beads are then rinsed twice with 125 μL of PBS plus BSA (ice-cold). They are then resuspended in 5μL of Sarkosyl Buffer.

6.2.6 MOW-ChIP procedure

The various steps involved in the process are described in Figure 29. The circular structure denoted by 1 is the inlet and the structure denoted by 2 is the outlet and the valve 3 is a sieve valve. The microfluidic device was first rinsed with the Sarkosyl Buffer. The antibody-coated magnetic beads were then flowed into the microfluidic chamber through inlet 1 using the pressure-driven flow of a syringe pump. A magnet was also used to assist the flow of beads into the device. After the magnetic beads were loaded and packed into a bed, 50μl of sonicated chromatin fragments mixed with 1.05 µl of carrier [20:1 ratio of recombinant histone 2B (M2505S; New England Biolabs) and human mRNA (Invitrogen)] suspended in Sarkosyl Buffer are flowed in via syringe pump. This chromatin+carrier mixture flowed through the packed bed of beads at a flow rate of 1.5 µl/min. After ChIP, the beads are washed using an oscillatory washing mechanism sequentially with three ice-cold buffers: low salt RIPA 0 buffer (0.1%SDS, 1% Triton--X100, 10 mM Tris HCl pH7.6, 1 mM EDTA, 0.1% NaDOC), high salt RIPA 0.3M NaCl buffer (0.1%SDS, 1% Triton--X100, 10 mM Tris HCl pH 7.6, 1mM EDTA, 0.1% NaDOC, 0.3M NaCl) and a LiCl Buffer (250mM LiCl, 0.5% NP40, 0.5% Na DOC, 1mM EDTA, 10mM Tris HCl pH 8.1). After each oscillatory washing, the beads were held in place by the magnet while the washed off debris was rinsed out of the microfluidic chamber by a flow of fresh corresponding washing buffer of the subsequent step flow at a flow rate of 2 µl/min. In the end, the beads were flowed out of the microfluidic chamber with TE Buffer (pH 8) at a flow rate of 50 µl/min and collected into an Eppendorf tube. A step by step protocol of the entire procedure has been included in Appendix .

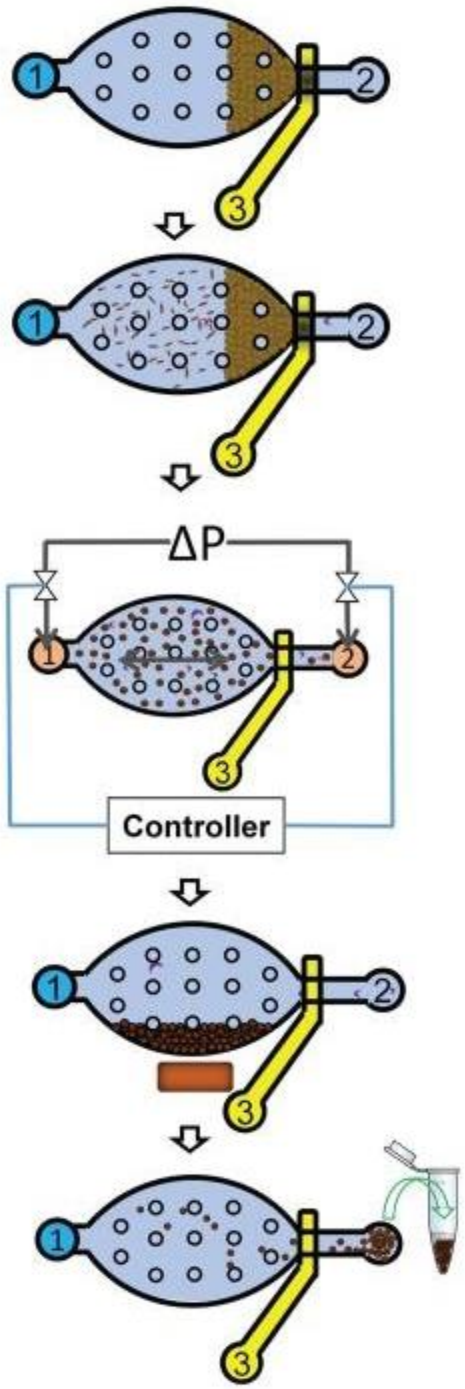

Figure 29: A step by step process outlining the functioning of Carrier MOW-ChIP.

Adapted with permission from Cao et al.(2015)[127]

6.2.7 Extraction of ChIP DNA and input DNA

The beads were placed on a magnetic rack and the buffer was replaced by 100 µl of fresh TE Buffer. RNase (10 mg/mL, Roche) was added at a volume of 1 µl and the tube was incubated at 37 °C for 30 min to remove any residual RNA from the carrier mix. In the case of the input samples, because there were no beads involved, the total volume of chromatin was made up to 100 µl with TE buffer and 1 µl of RNase was added. After the RNase digestion of immunoprecipitated samples and input control samples was over, 5 µl of Proteinase K (Thermo Fisher Cat. no. EO0492) was added to each sample and the tube was placed at 65 °C for 8 hrs. for reverse crosslinking. The reverse cross-linked DNA was then purified using standard phenol-chloroform extraction.

6.2.8 Construction of libraries and sequencing

Sequencing libraries were prepared by Accel-NGS 2S Plus DNA Library Kit (21096, Swift Biosciences). This kit provides high complexity next-generation sequencing libraries and compatibility with ultra-low inputs (~10 pg). The library preparation process involved two steps of repairs and two steps of ligations to repair both 5' and 3' termini and sequentially attach Illumina adapter sequences to the ends of fragmented double-stranded DNA. Bead-based SPRI clean-ups were used to remove oligonucleotides and small fragments, and to change enzymatic buffer composition between steps. PCR amplification (98 °C for 30 s, followed by 98 °C for 10 s, 60 °C for 30 s, 68 °C for 60 s for each cycle) was conducted to increase the yield of indexed libraries. Agilent 2200 TapeStation was used to check library size using a high-sensitivity DNA analysis kit (5067-4626, Agilent). The Kapa library quantification kit

(KK4809, Kapa Biosystems) gave the correct library concentrations required for pooling. The libraries were sequenced on an Illumina HiSeq 2500 with single-end 50-nt reads. On an average, each library was sequenced to a depth of 15–20 million reads.

6.2.9 Sequencing Data Analysis

ChIP sequencing reads were mapped to the human genome (hg19) using BWA (v0.7.17) with default parameter settings. Peaks of each ChIP sample were called against input by SPP (v1.14) with -npeak=300,000. 300,000 peaks were further filtered by IDR (v2.0.4) with -idr_thresh set at different thresholds.

ChIP-seq signals were normalized to input based on signal per million reads. Normalized ChIP signals in all promoter regions in the genome were mined. Promoter regions were defined as upstream 2000 bp and downstream 2000 bp around the TSSs. Regions with no signals the data sets were excluded for calculating Pearson correlation coefficient (remove zeroes).

Samtools (-F 1804) was used to remove reads unmapped, not primary alignment, reads failing platform quality checks, and duplicates. (-q 30) was used to remove multi-mapped reads that have low mapping quality score. PCR duplicates were marked by Picard’s MarkDuplicates.

6.3 Results and Discussion

For our preliminary data, we have MCF-7 data untreated with estrogen and with only formaldehyde cross-linking. Since ERα binds 200000 bp around transcription starting site[134], we compared microfluidic ERα untreated ChIP-seq data to ENCODE untreated data at TSS +/- 200000 bp regions (Figure 30). ENCODE (Encyclopedia of

DNA Elements) was started in 2003 and was the first worldwide project that compiled data from large-scale epigenomic studies. The ENCODE project was a pioneer in establishing many of the essential technologies required to study epigenomic marks and mainly dealt with cell lines instead of tissues or primary cells for cell lines provide more consistent data. The ENCODE is thus a complete repository of cell-line epigenomic data and we chose ChIP-seq data from the MCF-7 cell line. We compared two replicates of 10 K ERα untreated ChIP-seq have averaged ~0.86 Pearson correlation to the ENCODE data and ~0.98 self-correlation. Two replicates of 7.5 K ERα untreated ChIP-seq have averaged ~0.86 Pearson correlation to the ENCODE data and ~0.97 self-correlation. Two replicates of 5 K ERα untreated ChIP-seq have averaged ~0.78 Pearson correlation to the ENCODE data and ~0.88 self-correlation. Two replicates of 2.5 K ERα untreated ChIP-seq have averaged ~0.72 Pearson correlation to the ENCODE data and ~0.87 self-correlation.

When we compared the data to the existing low-input datasets from Zwart et al. (2013) who performed ERα CHIP-Seq for breast cancer biopsy for 10,000 cells and Shankaranayan et al. (2011) who published data on 5000 cells on ERα albeit on a different cell line (H3396 cell line). *Table 1* shows some comparison between these datasets. We were able to obtain a higher total number of peaks from our datasets with 10k, 7.5k and 5k cells when compared to other low-input datasets. Fraction of reads in peaks (or FRiP) is the fraction of all mapped reads that lie in the called peak regions and is used as an indicator of ChIP-seq quality. Our FRiP score (Fraction of reads in peaks) scores were also considerably better than competing technologies.

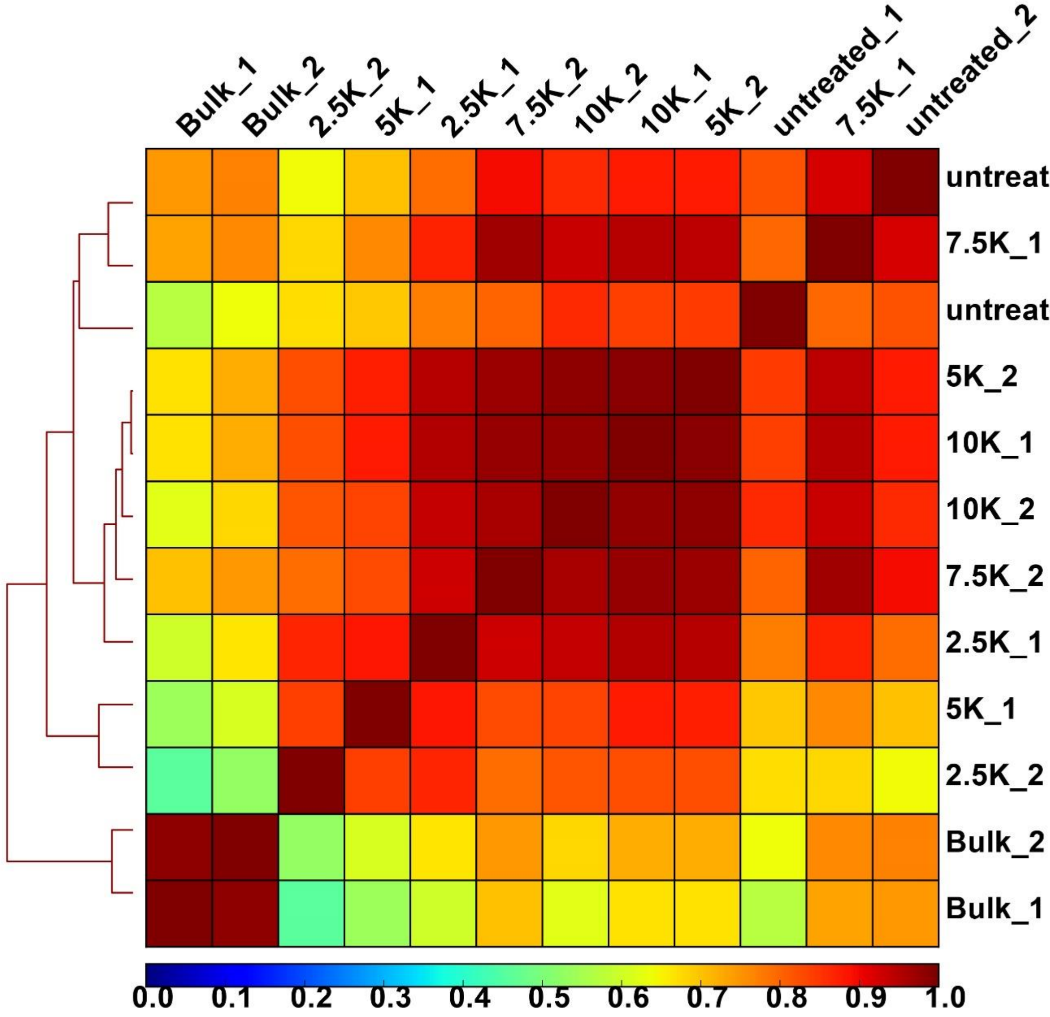

Figure 30: Microfluidic ERα Carrier MOW ChIP-seq data on MCF-7 cell line. Genome-wide Pearson correlations among ChIP-seq data sets of various input cell numbers. ChIP-seq genomic coverage profiles at TSS +/- 200000 bp regions were used for computing correlations.

Name	Description	Average no. of peaks	FRiP score
Zwart et al. (2013)[124]	10000 cells Breast cancer Biopsy sample	3366	1.03
Shankaranarayan et al.(2011)[125]	5000 cells H3396 cell line	3583	0.37
Carrier MOW-ChIP	10,000 cells MCF7 cell line	7030	3.12
Carrier MOW-ChIP	7,500 cells MCF7 cell line	5647	2.74
Carrier MOW-ChIP	5,000 cells MCF7 cell line	4822	1.69

Table 1: Comparison of MOW-ChIP with ERα to other competing technologies. Each dataset is 2 replicates.

Based on these preliminary results, the further work planned in this project revolves around gradually decreasing the cell number to 5000 cells, 2500 cells and then 1000 cells while still maintaining data quality. The DSG cross-linking along with formaldehyde will also be used. Once we get good quality data from a 1000 cells, the next challenge is to repeat the experiment using cross-linked chromatin extracted from just

1000 cells instead of cross-linking chromatin from a million cells and using 1000 cell equivalents of chromatin in a pool-split manner.

Using ChIP-seq it is possible to study how transcription factors interact with DNA at a genome-wide level. As sequencing technology improves, the resolution of ChIP-seq data has reached such a level that the difference between actual and predicted binding location of the transcription factor is usually within ~50 bp.[135] An ERα ChIP-seq experiment generally yields between hundreds to thousands of predicted binding locations of ERα (called 'ChIP-seq peaks') which can then be used for motif discovery. After finding statistically significant motifs, a Gene Ontology analysis can be performed to understand what physiological roles are being controlled by the ERα transcription factor. If a particular TF binding motif occurs upstream of genes related to a particular function (for example, say mammary cell proliferation), it usually implies that the transcription factor that binds the motif may regulate mammary cell proliferation. This is a valuable treasure trove of information for understanding processes such as gene regulation, drug resistance or metastasis in breast cancer tumors. Most of these tests have to be carried out on clinical patient samples and thus the need arises that this technology is made possible from a limited input. Our research shows a promising packed-bed microfluidic platform which could be combined with efficient carrier chromatin and a two-step cross-linking to develop sensitive ERα CHIP-Seq assays for as low as 1000 cells of a cell line MCF-7. With slight tweaks, this technology can easily be applied to other transcription factors and to limited clinical patient samples which is what makes it powerful. This technology can be combined with automated platforms and be used for large-scale patient epigenetic screens in clinical settings. In the context of personalized medicine, such screens will help doctors and clinicians understand disease prognosis and help in better treatment decisions.

Chapter 8: Summary and Outlook

Omic technologies have become an indispensable part of our healthcare. The speed, accuracy, and decreasing costs of next-generation sequencing (NGS) have helped accelerate the research on precision medicine. Patients can now be stratified according to their *omic* profile and a custom-made treatment plan can be devised for every individual. The presence of NGS and personalized medicine has mostly been felt in oncology where physicians are trying to understand the omic landscape of their patients' tumors to assign them therapies particularly suited for the factors driving the tumor's growth.

The governments of many countries around the globe such as Australia, the UK, and Finland are sponsoring projects to get genomic data from its population. In the US, the 'All of Us' program has been funded by the NIH to gather genomic data from 1 million individuals from diverse lifestyles, environments and biology, and this data would no doubt be a very useful resource for developments in precision medicine.

In this thesis, we discussed a number of projects which we can contribute to advancement in this field. Our first project, MID-RNA-seq offers a novel platform for one-pot single cell RNA-seq with great multiplexing ability. This platform provides good quality scRNA-seq data and can be used for transcriptomic studies of patient samples. This platform has enough potential to be scaled up and automated into an excellent platform for deep sequencing of rare or limited cell samples. The second project involves transcriptomic studies of innate leukocytes. The transcriptomic data that we have unearthed in this project provides an immense scope for understanding innate immunity and also provides hope in using therapeutic molecules like 4-PBA to reverse the damage caused by chronic inflammation. This data also provides our biologist collaborators the scope to test various pathways in innate immune cells and their roles in innate immune modulation. This data in conjunction with epigenomic data has immense potential into providing useful insights into the function of neutrophils and monocytes. Our third project discusses a method to produce tn5 enzyme for tagmentation of DNA. This has enabled us to greatly reduce costs of library preparation and is currently being used by some members in the lab for large scale library preparation and also for other processes that use tagmentation such as Chipmentation.[136] The fourth project discussed involves low-input transcription factor ChIP-seq where we have done some preliminary work to optimize the MOW-ChIP platform for transcription factors. With the initial data we have, there is immense scope for further optimization by future students in the lab. This would enable us to study epigenomic profiles from limited patient samples.

The next big challenge for the community would be to develop integrated multi-*omic* technologies with the genome, transcriptome, epigenome, metabolome, and proteome analyzed from the same cell with the analytical powers to process all this data simultaneously. This would open doors for unprecedented omic profiling of patients and individuals and would truly herald in the multiomic revolution.

Appendix A

Here is a step-by-step procedure to prepare tn5 tagmentation enzyme in-house.

1. Streak stab culture obtained from ADDGENE onto an LB agar plate with 100 µg/ml Ampicillin plate.
2. Grow overnight at 37°C.
3. Pick one colony were picked and grow in 5 ml of LB medium with 100 µg/ml Ampicillin overnight at 37°C with 250 rpm shaker.
4. Extract plasmid DNA from culture using the QIAPrep Spin Miniprep kit (Cat No. 27104).
5. Measure concentration of plasmid DNA using Nanodrop. The amount obtained should be between 100pg-1ng for successful transformation.
6. Transform the T7 Express lysY/Iq Competent E. coli (NEB, # C3013I) using the plasmid DNA following manufacturer's recommendations.
7. Streak transformed bacteria on a plate and preserve excess cells as glycerol stocks. Grow overnight at 37°C.
8. Pick a colony and grow in 50 ml LB broth supplemented with 100 µg/ml Ampicillin overnight at 37°C,200 rpm.
9. Use 10ml of this overnight culture to inoculate 1 L of LB medium containing 100 µg/ml Ampicillin.
10. Grow culture grown at 37°C,200 rpm until OD@600=0.5 OR A@600=0.9.
11. Chill culture to 10°C for 30min.
12. Add 250 µl of 1 M IPTG to the 1L culture to induce protein expression.
13. Grow culture at 23°C for 4 h at 250 rpm until the culture reaches A600=3.0.
14. Harvest bacteria by centrifugation (5000 g, 4°C, 15 min).

SAFE STOPPING POINT: Pellet can be stored at -80°C.

15. If frozen, thaw pellets on ice and resuspended in 80 ml of cold HEGX buffer freshly supplemented with 1% Complete Protease Inhibitor Cocktail (Sigma, # 4693132001).
16. Lyse cells with 10–12 cycles of 45–50 bursts with 50% duty cycle at output 7 on a Branson sonicator with a 10-mm tip. Maintain cells in an ice-salt mixture using to prevent overheating during sonication.
17. Pellet lysate in the Beckman JA17 rotor at 15,000 rpm for 30 min at 4°C.
18. Add 2.1 ml of 10% neutralized PEI (must be adjusted beforehand to pH 7.0, from Sigma P3143) dropwise on a magnetic stirrer.
19. Remove the white precipitate containing the bacterial nucleic acids by centrifugation at 12,000 rpm for 10 min at 4°C in the Beckman JA17 rotor.
20. Pack chitin resin (NEB, S6651S) was packed it into a Pierce Thermo Fischer chromatography column following manufacturer’s protocol.
21. Wash chitin column with 10 column volumes of or HEGX buffer to equilibrate it.
22. Load the crude cell extract onto the column. 10mL of chitin resin (20mL of chitin slurry) is sufficient to bind 1L of cell culture.
23. Was column with 20 to 30 column volumes of HEGX(with 1% PIC) to wash off unbound protein.
24. Submerge chitin resin in the column in HEGX(1% PIC) supplemented with 100 mM DTT.
25. Release 1 column volume of solution to make sure the entire resin is saturated by the DTT buffer.

26. Close column outlet and incubate 36-48 h at 4°C to induce cleavage of the intein tag that binds the protein to the chitin.
27. Was column with HEGX buffer (1% PIC) and collect protein in 1 mL aliquots, up to one column volume worth of liquid.
28. Visually analyze each 1mL aliquot using a Bradford protein assay to determine the intensity of blue color. Pool the samples with the strongest blue color.
29. Clean and regenerate chitin column following the manufacturer's instructions. Column can be reused up to 5 times.
30. Dialyze protein versus two changes, one after 2 hours and one overnight, using Tn5 dialysis buffer at 4°C with constant magnetic stirring.
31. Measure protein concentration using a Nanodrop spectrophotometer and use the dialysis buffer for blanking the device. Do not use water for blanking.
32. Concentrate protein further using a Pierce Protein Concentrator of 30K MWCO at 4°C until the concentration reaches 1.85 mg/mL (A_{280} = 3.0). Protein must be maintained at low temperatures continuously.
33. Store protein at -20°C as a glycerol stock which can be prepared by adding 1.1 vol of 100% glycerol (it is recommended to measure glycerol by weight as it is very viscous and volume measurements are inaccurate. The density of glycerol is 1.26 g/cm^3) and 0.33 vol of Tn5 dialysis buffer.
34. For even longer term storage, flash freezing in liquid nitrogen and storage at -80 °C is recommended. While using flash frozen protein, thaw rapidly in 37°C.

<u>Tn5 protein assembly</u>

35. Resuspended lyophilized oligos were resuspended in annealing buffer (50 mM NaCl, 40 mM Tris-HCl pH 8.0) to obtain a concentration of 100 µM.
36. Add one volume of Tn5ME-A oligo with one volume of Tn5MErev in 1:1 ratio and then run the mixture on a thermocycler with the following program: 95°C 5 min slowly cool down to 65°C (0.1°C/sec), 65°C 5 min slowly cool down to 4°C (0.1°C/sec). Repeat process with Tn5ME-B oligo.
37. Make aliquots of annealed linker oligonucleotides and store long term at -20°C.
38. If using fresh enzyme, add 0.125 vol of pre-annealed Tn5MEDS-A and Tn5MEDS-B oligonucleotides mixed in a 1:1 ratio to 0.4 vol of 100% glycerol, 0.12 vol Tn5 dialysis buffer and 0.36 vol of Tn5 enzyme at the concentration of 1.85 mg/mL.
39. If using glycerol stock enzyme from -20°C, add 0.143 vol Tn5MEDS-A/B oligonucleotides (mixed in a 1:1 ratio) with 1 volume of the tn5 enzyme.
40. Incubate mixture for 60 min at room temperature (RT): 23°C under shaking conditions of 350 rpm. tn5 gradually loses activity at 23°C so the assembled protein should be used immediately.
41. If planning to store assembled enzyme at -20°C, the glycerol concentration should be increased from 39.6% (currently in the assembled mix) to 50% final concentration. Repeated freeze-thaw should be avoided at all costs.

High-input tagmentation using homemade tn5

42. In order to prepare a high input reaction for DNA tagmentation, the following recipe was used: 14 µL H_2O, 4 µL TAPS-DMF buffer, 1 µL target DNA at 50 ng/µL, 1 µL of the freshly pre-assembled Tn5 from step 40. The TAPS-DMF

buffer should be at room temperature and the 50% DMF should be added freshly.

43. Incubate mixture for 7 min at 55°C. The thermocycler must be pre-heated at 55°C before the reaction.
44. Add 5µL of 0.2% SDS to stop the reaction by incubating at 7 min at 23°C followed by a 10°C hold.

Low-input tagmentation using homemade tn5

45. In case of low input library preparation (i.e. from 1ng or less DNA), the setup of the reactions was as follows: 4ul of 40% w/v PEG (MW > 4000), 4 µL of TAPS-DMF buffer, 1 µL of DNA, variable amount of Tn5 enzyme (0.01–1 µL, variable amount and depends on the amount of DNA,0.5 µL preferred for 1ng of DNA).
46. Make up the total volume to 20 µL with TAPS-DMF buffer.
47. Carry out the tagmentation reaction as mentioned in step 43 and step 44.

Appendix B

Here is a step-by-step protocol for the carrier MOW-ChIP process. All measurements are for a 75 cm^2 flask and can be scaled accordingly.

1. Grow MCF-7 cells in 15ml media to 80% confluence in D-MEM/F-12 supplemented with 10% FBS and 1%PS.
2. Cells can be sub-cultured in a 1:3 ratio by trypsinization with 0.25% Trypsin-EDTA, 37°C for 1min. During sub culture, if cells clump up they can be passed through a 40 µm FLOWMI cell strainer (Cat. No. H13680-0040).
3. Wash confluent cells with PBS (37°C, sterile) and change media to phenol red-free D-MEM/F-12 medium supplemented with 5% charcoal-dextran stripped FBS and 1% PS for 48 h for hormone starvation.
4. Prepare E2 stock solution by dissolving it in 100% ethanol and then making up the volume with the hormone-starvation media. Add E2 to the hormone-starved culture flask at a final concentration of 10 nM for 1 hr.
5. Aspirate media, rinse the flask with PBS twice. PBS should be at 37°C/or room temperature.
6. Cover the cells with 15ml of the following solution: (50 mM Hepes-KOH, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA,2 mM DSG) for 35 min at room temperature on a rocking platform.
7. Add formaldehyde to 1% final volume.
8. Incubate for 10 min further at room temperature on the rocking platform.
9. Quench formaldehyde by adding Glycine to a final concentration of 0.125M and incubate for 10 min under rocking at room temperature.
10. Rinse flask twice with ice-cold PBS.

11. Add 1ml of (ice-cold PBS+1x PIC+ 5mM Sodium Butyrate) and scrape cells using a cell scraper into this solution. Add additional volume if necessary.
12. Spin cells at 2000 rpm, 2min at 4°C.

SAFE STOPPING POINT: Cell pellet can be stored at -80°C if required, but not recommended.

13. Resuspend cell pellet in appropriate volume of Sarkosyl Buffer freshly supplemented with 1x PIC.
14. Let it sit on ice for 10-15 min.
15. The volume of Sarkosyl buffer used will depend on the number of cells and the sonication volume of Covaris (50 µl/130 µl) and it is recommended not to overcrowd the sonication volume. The volume, cell number and sonication duration needs to be optimized.
16. Typically, 1 million cells can be sonicated in 130 µl volume in Covaris M220 with a Peak Incident Power of 50, Cycles per burst 200, time 900 secs, temperature 4°C.
17. Clear lysate by spinning down at full speed at 4°C for 15 min and transfer supernatant to a new tube. Aliquot and snap-freeze chromatin at -80 °C.
18. On ice, take 2.5 µl each of Protein A and G beads mixed in 1:1 ratio and remove supernatant.
19. Wash 3 times with 125 µl of ice-cold PBS + 0.5% BSA.
20. Resuspend in 125 µl of PBS-BSA.
21. Add 3 µl of Antibody ER (Santa Cruz Biotechnology sc-8002) to the solution.
22. Rotate at 4°C for overnight (or 2 hrs. if in a hurry).

23. Put beads on a magnetic stand and remove supernatant. Rinse beads twice with 125 µl of ice-cold PBS-BSA.
24. Resuspend beads in 5 µl of Sarkosyl Buffer with 1x PIC freshly added.
25. Take 50 µl of chromatin from desired amount of cells. Volume can be adjusted with Sarkosyl Buffer-1x PIC
26. Add 1.05 µl of carrier mix (Histone 2B and RNA).
27. Rinse MOW-ChIP device with Sarkosyl Buffer and make sure there are no bubbles.
28. Load the 5 µl beads solution at 20 µl /min with a magnet underneath. Make sure no beads are stuck in the inlet. A packed bed of beads form near the sieve valve.
29. Load the chromatin at 1.5 µl /min with a magnet underneath.
30. Attach tubing filled with wash buffers at both ends of the device. Use very low pressure of <0.5 psi for oscillatory washing.
31. Wash 1: Ice-cold RIPA 0 buffer (5 min oscillatory washing) - flush out debris @ 2 µl /min while holding beads in place with a magnet.
32. Wash 2: Ice-cold RIPA 0.3 buffer (5 min oscillatory washing) - flush out debris @ 2 µl /min.
33. Wash 3: Ice-cold LiCl Buffer (5 min oscillatory washing) - flush out debris @ 2 µl /min.
34. Flush out beads from the chip with TE buffer.
35. Discard supernatant on a magnetic rack
36. Resuspend the beads with 100 µl TE buffer.
37. Treat with 1 µl of 10mg/ml of RNAase 30 min,37°C.

38. Add 5 µl of 20mg/ml Proteinase K, incubate @ 65°C for 16 hrs.

39. Extract ChIP-DNA using standard ethanol precipitation.

Appendix C

Here is a list of genes in monocytes that are upregulated/downregulated in cells treated with LPS alone vs cells treated by LPS and 4-PBA:

Upregulated by LPS	Downregulated by LPS
Psap	Rps2
Ccl2	Rps14
Ccl12	Rpl7a
Ccl4	Tubb4b
Wfdc17	H2afz
Ccl5	Plac8
Ccl9	Rpl32
Rsad2	
Ctsl	
Ctsb	
Irg1	
Ifit1	
AW112010	
Il1rn	
Sirpa	
Fabp5	
Fcgr1	
Prdx1	
C3ar1	

Saa3 Ctsd	

Here is a list of genes in monocytes that are upregulated/downregulated in cells treated by LPS vs control cells treated by PBS:

Upregulated by LPS	Downregulated by LPS
Lilrb4a Ccl2 Ccl4 Ccl9 Rsad2 Pnp Irg1 Mx1 H2-D1,H2-L,LOC547349 Ms4a4c Ifit3 AW112010 Ms4a6d Il1rn Fcgr1	Crip1 Lgals1

Saa3 Ifitm3	

References

1. Z. Wang, M. Gerstein and M. Snyder, *Nat Rev Genet*, 2009, **10**, 57-63.
2. R. Lowe, N. Shirley, M. Bleackley, S. Dolan and T. Shafee, *PLoS Comput Biol*, 2017, **13**, e1005457.
3. F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao and M. A. Surani, *Nat. Methods*, 2009, **6**, 377-382.
4. F. Valdes-Mora, K. Handler, A. M. K. Law, R. Salomon, S. R. Oakes, C. J. Ormandy and D. Gallego-Ortega, *Front Immunol*, 2018, **9**, 2582.
5. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer and B. Wold, *Nat Methods*, 2008, **5**, 621-628.
6. J. M. Heather and B. Chain, *Genomics*, 2016, **107**, 1-8.
7. M. L. Gonzalez-Garay, *Per Med*, 2014, **11**, 523-544.
8. B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach and S. W. Turner, *Nat Methods*, 2010, **7**, 461-465.
9. M. Eisenstein, *Nature Biotechnology*, 2012, **30**, 295-296.
10. J. Shin, G. L. Ming and H. J. Song, *Nat Neurosci*, 2014, **17**, 1463-1475.
11. A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke and S. R. Quake, *Nat. Methods*, 2013, **11**, 41-46.
12. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas and S. Tyagi, *PLoS Biol*, 2006, **4**, e309.
13. L. Wen and F. Tang, *Genome Biol*, 2016, **17**, 71.
14. S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg and R. Sandberg, *Nat. Methods*, 2013, **10**, 1096-1098.
15. D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth and R. Sandberg, *Nat. Biotechnol.*, 2012, **30**, 777-782.
16. T. Hashimshony, F. Wagner, N. Sher and I. Yanai, *Cell Rep.*, 2012, **2**, 666-673.
17. T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev and I. Yanai, *Genome Biol.*, 2016, **17**, 77.
18. H. Hochgerner, P. Lönnerberg, R. Hodge, J. Mikes, A. Heskol, H. Hubschle, P. Lin, S. Picelli, G. La Manno, M. Ratz, J. Dunne, S. Husain, E. Lein, M. Srinivasan, A. Zeisel and S. Linnarsson, *Sci. Rep.*, 2017, **7**.
19. Y. Sasagawa, I. Nikaido, T. Hayashi, H. Danno, K. D. Uno, T. Imai and H. R. Ueda, *Genome Biol*, 2013, **14**, R31.
20. A. R. Chapman, Z. He, S. Lu, J. Yong, L. Tan, F. Tang and X. S. Xie, *PLoS One*, 2015, **10**, e0120889.
21. D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay and I. Amit, *Science*, 2014, **343**, 776-779.
22. H. C. Fan, G. K. Fu and S. P. A. Fodor, *Science*, 2015, **347**.
23. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev and S. A. McCarroll, *Cell*, 2015, **161**, 1202-1214.
24. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz and M. W. Kirschner, *Cell*, 2015, **161**, 1187-1201.
25. S. Picelli, *RNA Biol.*, 2017, **14**, 637-650.
26. C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann and W. Enard, *Molecular cell*, 2017, **65**, 631-643.e634.
27. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson and J. H. Bielas, *Nat Commun*, 2017, **8**.

28. X. Zhang, T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang and J. Wang, *Molecular cell*, 2019, **73**, 130-142 e135.

29. E. Papalexi and R. Satija, *Nat. Rev. Immunol.*, 2018, **18**, 35-45.

30. A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp Ii, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein and J. A. A. West, *Nature Biotechnology*, 2014, **32**, 1053.

31. R. Zilionis, J. Nainys, A. Veres, V. Savova, D. Zemmour, A. M. Klein and L. Mazutis, *Nat. Protoc.*, 2017, **12**, 44-73.

32. P. L. Stahl, F. Salmen, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, A. Borg, F. Ponten, P. I. Costea, P. Sahlen, J. Mulder, O. Bergmann, J. Lundeberg and J. Frisen, *Science*, 2016, **353**, 78-82.

33. T. Nawy, *Nature Methods*, 2018, **15**, 30-30.

34. K. Thrane, H. Eriksson, J. Maaskola, J. Hansson and J. Lundeberg, *Cancer Res*, 2018, **78**, 5970-5979.

35. S. Giacomello and J. Lundeberg, *Nature Protocols*, 2018, **13**, 2425-2446.

36. A. M. Streets, X. Zhang, C. Cao, Y. Pang, X. Wu, L. Xiong, L. Yang, Y. Fu, L. Zhao, F. Tang and Y. Huang, *Proceedings of the National Academy of Sciences*, 2014, **111**, 7048-7053.

37. R. Kurita and O. Niwa, *Lab Chip*, 2016, **16**, 3631-3644.

38. A. R. Wu and S. R. Quake, *Cold Spring Harb Protoc*, 2016, **2016**, pdb prot084996.

39. Y. Ma, J. Thiele, L. Abdelmohsen, J. Xu and W. T. Huck, *Chem Commun (Camb)*, 2014, **50**, 112-114.

40. S. Ma, T. W. Murphy and C. Lu, *Biomicrofluidics*, 2017, **11**, 021501.

41. J. G. Kralj, A. Player, H. Sedrick, M. S. Munson, D. Petersen, S. P. Forry, P. Meltzer, E. Kawasaki and L. E. Locascio, *Lab on a Chip*, 2009, **9**, 917-924.

42. S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. H. Gephart, B. A. Barres and S. R. Quake, *P Natl Acad Sci USA*, 2015, **112**, 7285-7290.

43. A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. N. Lu, P. L. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. H. Wang, R. H. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May and A. Regev, *Nature*, 2014, **510**, 363-+.

44. N. M. Toriello, E. S. Douglas, N. Thaitrong, S. C. Hsiao, M. B. Francis, C. R. Bertozzi and R. A. Mathies, *Proceedings of the National Academy of Sciences*, 2008, **105**, 20173-20178.

45. A. K. White, M. VanInsberghe, O. I. Petriv, M. Hamidi, D. Sikorski, M. A. Marra, J. Piret, S. Aparicio and C. L. Hansen, *Proceedings of the National Academy of Sciences*, 2011, **108**, 13999-14004.

46. I. G. Wilson, *Appl. Environ. Microbiol.*, 1997, **63**, 3741-3751.

47. A. Trombley Hall, A. McKay Zovanyi, D. R. Christensen, J. W. Koehler and T. Devins Minogue, *PLoS One*, 2013, **8**, e73845.

48. W. A. Al-Soud and P. Rådström, *J. Clin. Microbiol.*, 2001, **39**, 485-493.

49. J. S. Marcus, W. French Anderson and S. R. Quake, *Anal. Chem.*, 2006, **78**, 3084-3089.

50. L. A. Legendre, J. M. Bienvenue, M. G. Roper, J. P. Ferrance and J. P. Landers, *Anal. Chem.*, 2006, **78**, 1444-1451.

51. T. Gilbert, T. Sellaro and S. Badylak, *Biomaterials*, 2006, DOI: 10.1016/j.biomaterials.2006.02.014.

52. S. Ma, D. N. Loufakis, Z. Cao, Y. Chang, L. E. Achenie and C. Lu, *Lab Chip*, 2014, **14**, 2905-2909.

53. S. Ma, M. de la Fuente Revenga, Z. Sun, C. Sun, T. W. Murphy, H. Xie, J. González-Maeso and C. Lu, *Nature Biomedical Engineering*, 2018, **2**, 183-194.
54. F. Miura, N. Kawaguchi, M. Yoshida, C. Uematsu, K. Kito, Y. Sakaki and T. Ito, *BMC Genomics*, 2008, **9**, 574-574.
55. R. Milo and R. Phillips, *Cell Biology by the Numbers*, Garland Science, 2015.
56. M. E. Young, P. A. Carroad and R. L. Bell, *Biotechnol. Bioeng.*, 1980, **22**, 947-955.
57. G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers and B. J. Wold, *Genome Res.*, 2014, **24**, 496-510.
58. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter, *Nature Biotechnology*, 2012, **31**, 46.
59. S. R. Krishnaswami, R. V. Grindberg, M. Novotny, P. Venepally, B. Lacar, K. Bhutani, S. B. Linker, S. Pham, J. A. Erwin, J. A. Miller, R. Hodge, J. K. McCarthy, M. Kelder, J. McCorrison, B. D. Aevermann, F. D. Fuertes, R. H. Scheuermann, J. Lee, E. S. Lein, N. Schork, M. J. McConnell, F. H. Gage and R. S. Lasken, *Nat. Protoc.*, 2016, **11**, 499-524.
60. T. Geng, Zhan, Y., Wang, J., Lu, C., *Nature Protocols*, 2011, **6**, 1192-1208.
61. S. Ma, Hsieh, Y.-P., Ma, J., Lu, C., *Science advances*, 2018, **4**, eaar8187.
62. T. W. Murphy, Y. P. Hsieh, S. Ma, Y. Zhu and C. Lu, *Anal Chem*, 2018, **90**, 7666-7674.
63. S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser and R. Sandberg, *Nature Protocols*, 2014, **9**, 171.
64. C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter, *Nat. Protoc.*, 2012, **7**, 562-578.
65. M. Morris and L. Li, *Arch Immunol Ther Exp (Warsz)*, 2012, **60**, 13-18.
66. M. C. Morris, E. A. Gilliam and L. Li, *Front Immunol*, 2014, **5**, 680.
67. S. K. Biswas and E. Lopez-Collazo, *Trends Immunol*, 2009, **30**, 475-487.
68. C. Lee, S. Geng, Y. Zhang, A. Rahtes and L. Li, *J Leukoc Biol*, 2017, **102**, 719-726.
69. D. Boraschi and P. Italiani, *Front Immunol*, 2018, **9**, 799.
70. M. G. Netea, L. A. Joosten, E. Latz, K. H. Mills, G. Natoli, H. G. Stunnenberg, L. A. O'Neill and R. J. Xavier, *Science*, 2016, **352**, aaf1098.
71. M. A. Hamon and P. Cossart, *Cell Host Microbe*, 2008, **4**, 100-109.
72. J. Lee, T. Zhang, I. Hwang, A. Kim, L. Nitschke, M. Kim, J. M. Scott, Y. Kamimura, L. L. Lanier and S. Kim, *Immunity*, 2015, **42**, 431-442.
73. S. Monticelli and G. Natoli, *Nat Immunol*, 2013, **14**, 777-784.
74. W. Ahmed and Z. F. Liu, *Frontiers in Immunology*, 2018, **9**.
75. R. X. Yuan, S. Geng, K. Q. Chen, N. Diao, H. W. Chu and L. W. Li, *J Pathol*, 2016, **238**, 571-583.
76. S. Geng, K. Q. Chen, R. X. Yuan, L. Peng, U. Maitra, N. Diao, C. Chen, Y. Zhang, Y. Hu, C. F. Qi, S. Pierce, W. H. Ling, H. B. Xiong and L. W. Li, *Nat Commun*, 2016, **7**.
77. R. Sturm, *Health Aff (Millwood)*, 2002, **21**, 245-253.
78. L. Qin, X. Wu, M. L. Block, Y. Liu, G. R. Breese, J. S. Hong, D. J. Knapp and F. T. Crews, *Glia*, 2007, **55**, 453-462.
79. T. N. Mayadas, X. Cullere and C. A. Lowell, *Annu Rev Pathol-Mech*, 2014, **9**, 181-218.
80. B. S. Park and J. O. Lee, *Exp Mol Med*, 2013, **45**, e66.
81. S. Geng, Y. Zhang, C. Lee and L. Li, *Sci Adv*, 2019, **5**, eaav2309.
82. H. Yang, M. H. Biermann, J. M. Brauner, Y. Liu, Y. Zhao and M. Herrmann, *Frontiers in Immunology*, 2016, **7**.
83. E. Lynn, S. Lhotak and R. Austin, *Atherosclerosis*, 2014, **235**, E126-E126.
84. E. Lachmandas, L. Boutens, J. M. Ratter, A. Hijmans, G. J. Hooiveld, L. A. B. Joosten, R. J. Rodenburg, J. A. M. Fransen, R. H. Houtkooper, R. van Crevel, M. G. Netea and R. Stienstra, *Nat Microbiol*, 2017, **2**.

85. R. Yuan, S. Geng and L. Li, *Front Immunol*, 2016, **7**, 497.
86. K. R. Higgins, W. Kovacevic and L. Stokes, *Mediat Inflamm*, 2014, DOI: Artn 293925

10.1155/2014/293925.
87. T. Satoh, K. Nakagawa, F. Sugihara, R. Kuwahara, M. Ashihara, F. Yamane, Y. Minowa, K. Fukushima, I. Ebina, Y. Yoshioka, A. Kumanogoh and S. Akira, *Nature*, 2017, **541**, 96-+.
88. M. Swamydas and M. S. Lionakis, *Jove-J Vis Exp*, 2013, DOI: ARTN e50586

10.3791/50586.
89. M. Wagner, H. Koester, C. Deffge, S. Weinert, J. Lauf, A. Francke, J. Lee, R. C. Braun-Dullaeus and J. Herold, *J Vis Exp*, 2014, DOI: 10.3791/52347.
90. G. Heimberg, R. Bhatnagar, H. El-Samad and M. Thomson, *Cell Syst*, 2016, **2**, 239-250.
91. W. KA, *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, 2018, **Available at: www.genome.gov/sequencingcostsdata, Accessed 3-25-2019**

92. I. Y. Goryshin, J. A. Miller, Y. V. Kil, V. A. Lanzov and W. S. Reznikoff, *Proc Natl Acad Sci U S A*, 1998, **95**, 10716-10721.

93. A. Adey, H. G. Morrison, Asan, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C. Caruccio, X. Zhang and J. Shendure, *Genome Biol*, 2010, **11**, R119.

94. S. Picelli, A. K. Bjorklund, B. Reinius, S. Sagasser, G. Winberg and R. Sandberg, *Genome Res*, 2014, **24**, 2033-2040.
95. B. P. Hennig, L. Velten, I. Racke, C. S. Tu, M. Thoms, V. Rybin, H. Besir, K. Remans and L. M. Steinmetz, *G3 (Bethesda)*, 2018, **8**, 79-89.
96. L. J. Hawkins, R. Al-Attar and K. B. Storey, *PeerJ*, 2018, **6**, e5062.
97. W. Burggren, *Biology (Basel)*, 2016, **5**.
98. N. Grova, H. Schroeder, J. L. Olivier and J. D. Turner, *Int J Genomics*, 2019, **2019**, 2085496.
99. I. C. G. Weaver, N. Cervoni, F. A. Champagne, A. C. D'Alessio, S. Sharma, J. R. Seckl, S. Dymov, M. Szyf and M. J. Meaney, *Nat Neurosci*, 2004, **7**, 847-854.
100. C. H. Waddington, *Int J Epidemiol*, 2012, **41**, 10-13.
101. S. H. Stricker, A. Koferle and S. Beck, *Nature Reviews Genetics*, 2017, **18**, 51-66.
102. B. Paul, S. Barnes, W. Demark-Wahnefried, C. Morrow, C. Salvador, C. Skibola and T. O. Tollefsbol, *Clin Epigenetics*, 2015, **7**, 112.
103. M. Kulis and M. Esteller, *Adv Genet*, 2010, **70**, 27-56.
104. Y. Kondo, *Yonsei Med J*, 2009, **50**, 455-463.
105. R. Kala, G. W. Peek, T. M. Hardy and T. O. Tollefsbol, *J Clin Bioinforma*, 2013, **3**, 6.
106. T. R. Mercer, M. E. Dinger and J. S. Mattick, *Nature Reviews Genetics*, 2009, **10**, 155.
107. A. H. Brivanlou and J. E. Darnell, Jr., *Science*, 2002, **295**, 813-818.
108. S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes and M. T. Weirauch, *Cell*, 2018, **175**, 598-599.
109. D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, *Science*, 2007, **316**, 1497-1502.
110. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular biology of the cell*, Garland Science, New York, 4th edn., 2002.
111. C. E. Massie and I. G. Mills, *Embo Rep*, 2008, **9**, 337-343.
112. P. J. Park, *Nat Rev Genet*, 2009, **10**, 669-680.
113. P. J. Farnham, *Nat Rev Genet*, 2009, **10**, 605-616.
114. P. Collas, *Mol Biotechnol*, 2010, **45**, 87-100.
115. R. Mundade, H. G. Ozer, H. Wei, L. Prabhu and T. Lu, *Cell Cycle*, 2014, **13**, 2847-2852.

116. L. P. O'Neill, M. D. VerMilyea and B. M. Turner, *Nat Genet*, 2006, **38**, 835-841.
117. J. D. Nelson, O. Denisenko and K. Bomsztyk, *Nat Protoc*, 2006, **1**, 179-185.
118. Z. Cao, Doctor of Philosophy, Virginia Polytechnic Institute and State University, 2015.
119. L. P. O'Neill and B. M. Turner, *Methods Enzymol*, 1996, **274**, 189-197.
120. V. A. Spencer, J. M. Sun, L. Li and J. R. Davie, *Methods*, 2003, **31**, 67-75.
121. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang and W. J. Greenleaf, *Nat Methods*, 2013, **10**, 1213-1218.
122. A. Rotem, O. Ram, N. Shoresh, R. A. Sperling, A. Goren, D. A. Weitz and B. E. Bernstein, *Nat Biotechnol*, 2015, **33**, 1165-1172.
123. H. Hao, H. Liu, G. Gonye and J. S. Schwaber, *J Neurosci Methods*, 2008, **172**, 38-42.
124. W. Zwart, R. Koornstra, J. Wesseling, E. Rutgers, S. Linn and J. S. Carroll, *BMC Genomics*, 2013, **14**, 232.
125. P. Shankaranarayanan, M. A. Mendoza-Parra, M. Walia, L. Wang, N. Li, L. M. Trindade and H. Gronemeyer, *Nat Methods*, 2011, **8**, 565-567.
126. J. Shen, D. Jiang, Y. Fu, X. Wu, H. Guo, B. Feng, Y. Pang, A. M. Streets, F. Tang and Y. Huang, *Cell Res*, 2015, **25**, 143-147.
127. Z. Cao, C. Chen, B. He, K. Tan and C. Lu, *Nat Methods*, 2015, **12**, 959-962.
128. J. A. Dahl and G. D. Gilfillan, *Brief Funct Genomics*, 2018, **17**, 89-95.
129. C. E. DeSantis, C. C. Lin, A. B. Mariotto, R. L. Siegel, K. D. Stein, J. L. Kramer, R. Alteri, A. S. Robbins and A. Jemal, *CA Cancer J Clin*, 2014, **64**, 252-271.
130. Y. Pawitan, J. Bjohle, L. Amler, A. L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. M. Shaw, J. Smeds, L. Skoog, S. Wedren and J. Bergh, *Breast Cancer Res*, 2005, **7**, R953-964.
131. A. A. Singh, K. Schuurman, E. Nevedomskaya, S. Stelloo, S. Linder, M. Droog, Y. Kim, J. Sanders, H. van der Poel, A. M. Bergman, L. F. Wessels and W. Zwart, *Life Sci Alliance*, 2019, **2**, e201800115.
132. S. P. Chellappan, *Methods Mol Biol*, 2015, **1288**, V-Vi.
133. L. Schmiedeberg, P. Skene, A. Deaton and A. Bird, *PLoS One*, 2009, **4**, e4636.
134. J. S. Carroll, C. A. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoute, A. S. Brodsky, E. K. Keeton, K. C. Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. A. Fox, P. A. Silver, T. R. Gingeras, X. S. Liu and M. Brown, *Nat Genet*, 2006, **38**, 1289-1297.
135. T. L. Bailey, *Bioinformatics*, 2011, **27**, 1653-1659.
136. C. Schmidl, A. F. Rendeiro, N. C. Sheffield and C. Bock, *Nat Methods*, 2015, **12**, 963-965.