Preface

When I graduated high school with a degree in Latin and English, I could not have imagined that six years later, I would be working in one of today's most rapidly advancing technological fields. And yet, this has become a reality. During my master thesis, I worked together with some of the brightest engineers, biologists and bioinformaticians in the field, on the development of technology which I believe will truly revolutionise biology and improve the lives of many. I am indebted to many people for where I am today, but I am especially grateful to my parents and my brother, who have continuously supported me throughout my career. Moeke, Vake, Freek - I may not always say it out loud, but you all share a special place in my heart. Without you, I would have been nowhere.

Several people were involved in bringing this work to a successful ending. First, I want to thank Valerie and Samira. The two of you helped me in some of the most important experiments performed in this project, even when your time was limited. Second, Gert, Kris and Katina. You provided excellent support in processing our sequencing data, and reassured me that I had not (completely) screwed up the experiments. Stein, Suresh, in working together with you closely for the past year, I have started to look up to you not just as great scientists, but also as great people. Thank you for sharing your vast knowledge and experience with me, and for devoting a large part of your time and patience to this work. You inspire me.

Finally, Jasper. When you were proofreading my thesis right before submission, I showed you a mock version of this preface where you had not received nearly as much praise as you deserve. Joke's on you, I dedicated a whole paragraph to you in the final version. You supported me from day one until the end - when things were going well, and when they were not. We spent many long hours in the lab and in the office, and you answered my panic-driven questions at 2 am in the morning, the weekend before the deadline. You are one of the most extraordinarily kind, intelligent and patient people I have ever met, and I consider you a mentor and a dear friend. Thank you for everything.

This thesis is dedicated to my biological parents. I do not know where you are, or what you look like, but I love you. Your loss has not been in vain.

Summary

What drives cancer? How do our brains work? How is an embryo formed, and where can it go wrong? These questions are at the heart of biology, but due to the limits of the previous decade's technology, have not yet been answered. Researchers were constrained to taking a tissue sample, processing it, and analysing the resulting dataset as originating from a single entity. Such an approach leads to an enormous loss of information, as tissue consist of millions of individual cells, each with their own role. Today, several technological leaps have made it possible to profile the gene expression of thousands of single cells at a reasonable cost within a realistic time frame, an unthinkable feat a decade ago.

However, much remains to be improved. Today's single-cell assays are prone to noise, demand high amounts of labour and are still too expensive for routine use. This project focuses on improvement in the technological aspect of the single-cell landscape. The first part of this thesis describes how we combined elements of inDrop and Drop-seq, two droplet-microfluidic single-cell RNA-sequencing platforms, into a new protocol. Several new additions and features, both on the physical and the chemical/molecular scale, were tested and refined, producing a final sequencing library which we then analysed (and troubleshooted) computationally. In the second part, we implemented the assay for transposase-accessible chromatin (ATAC-seq) on our custom microfluidic platform to produce a droplet-based single-cell ATAC-seq protocol, a protocol which does not exist in open-source form yet.

While a number of technical limitations prohibited us from efficiently separating our data to the single-cell level, a problem which is examined in great detail in the third and final part of this work, the positive bulk characteristics of the data encouraged us to continue optimisation and refinement of our method. We have since diagnosed part of the problem, and are ready to move forward to an improved prototype. More importantly, due to the experience and knowledge we gained here, we are now well-equipped to be on the front line of single-cell technology. The future still holds several exciting challenges - from the development of new single-cell assays, to the integration of spatial information and finally, the combination of the two. By directly contributing to the solution of these challenges, we hope to push single-cell technology beyond what is possible now, and advance (systems) biology as a whole. Data generated with this technology will finally help us understand those processes where cellular heterogeneity is important, and will bring us one step closer to answering biology's great unsolved mysteries.

Samenvatting

Wat veroorzaakt kanker? Hoe werken onze hersenen? Hoe wordt een embryo gevormd, en waar kan het fout lopen? Deze vraagstukken staan in het centrum van de biologie, maar bleven onbeantwoord door de limieten van technologie uit de afgelopen decennia. Onderzoekers waren beperkt tot het homogeniseren van een weefselstaal, het "in bulk" te behandelen, en de resulterende data te verwerken als komende van een enkele entiteit. Deze aanpak leidt tot een enorm verlies aan informatie, want weefsels bestaan in werkelijkheid uit miljoenen individuele cellen, elk met hun eigen rol. Vandaag hebben een aantal technologische sprongen het mogelijk gemaakt om de genexpressie van duizenden individuele cellen te profileren, en dit binnen een redelijk tijdsbestek en aan een betaalbare prijs - een onmogelijke taak meer dan tien jaar geleden.

Er blijft echter veel plaats voor verbetering. De single-cell technieken van vandaag genereren veel ruis, vereisen grote hoeveelheden manueel werk en zijn nog steeds te duur om op routinebasis uitgevoerd te worden. Het eerste deel van deze thesis beschrijft hoe we elementen van inDrop en Drop-seq, twee druppel-gebaseerd microfluïdische RNA-sequencing platforms, combineerden in een nieuw, geïntegreerd protocol. Een aantal nieuwe addities, zowel op het fysische en chemische vlak, werden getest en verfijnd, leidend tot een sequencing dataset die dan computationeel geanalyseerd werd. In het tweede deel implementeerden we het "assay for transposase-accessible chromatin" (ATAC-seq) op het zelfde microfluidische platform om een druppel-gebaseerde single-cell ATAC-seq protocol te ontwikkelen, een techniek die vandaag nog niet in open-source vorm bestaat.

Hoewel een aantal technische limitaties ons verhinderden onze datasets efficient te splitsen tot het single-cell niveau, een probleem dat uitvoerig onderzocht wordt in het derde en laatste deel van de thesis, moedigden de positieve bulk eigenschappen van onze data ons aan om het optimisatie- en verfijningsproces van onze methode verder te zetten. Inmiddels hebben we een deel van de problemen kunnen identificeren, en zijn we klaar om verder te gaan naar een verbeterd prototype. Bovendien zijn we nu, door de expertise en kennis opgedaan in dit project, goed uitgerust om aan de frontlijn van het "single-cell" onderzoek te staan. De toekomst bevat een aantal spannende uitdagingen - van de ontwikkeling van nieuwe single-cell assays, tot de incorporatie van spatiale informatie, en uiteindelijk de combinatie van de twee. Door direct mee te werken aan het oplossen van deze uitdagingen hopen we single-cell technologie verder te brengen dan wat vandaag mogelijk is, en om de grenzen van de (systeem-) biologie te verleggen. Data gegenereerd via single-cell technologie zal ons helpen om processen te begrijpen waar cellulaire heterogeniteit belangrijk is, en zal ons een stap dichter brengen bij de antwoorden op de grote onopgeloste vragen in de biologie.

Contents

List of Figures

*

Context

As a technology, microfluidics seems almost too good to be true: it offers so many advantages and so few disadvantages. But it has not yet become widely used. Why not? Why is every biochemistry laboratory not littered with 'labs on chips'? Why does every patient not monitor his or her condition using microfluidic home-test systems? The answers are not yet clear.

— George M. Whitesides, Nature 2006

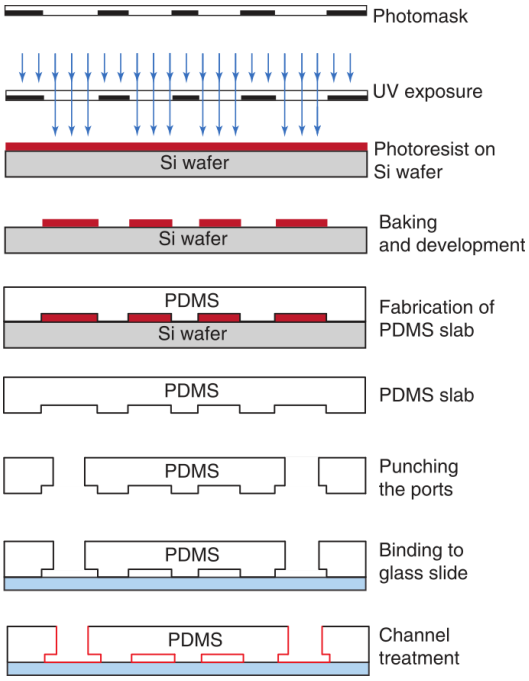


Figure C.1: PDMS soft lithography. Taken from Mazutis et al. (2013)

In 2006, when PDMS was still a rare sight in most labs, George M. Whitesides pondered why the seemingly inconspicuous polymer had not yet revolutionised molecular research. It had been eight years since the many benefits of PDMS soft lithography and its possible applications in microfluidics were originally proposed (Xia and Whitesides, 1998). Today, more than a decade after Whitesides' 2006 Nature Insight article, soft elastomer lithography has become the most widespread technique for microfluidic device manufacturing and the 1998 paper has been cited more than 11 000 times. The technique has found many applications in various branches of science. In biotechnology, where its low toxicity, inherent hydrophobicity, transparency and inert nature pose many advantages, PDMS soft lithography has become a staple technique. Droplet microfluidics, a subset of microfluidics where pico- to nanolitre droplets are produced by focusing two immiscible phases, has been used extensively in bioscience in the past years. The ultra-low volumes associated with droplet microfluidics reduce reagent usage, leading to a reduction in overall cost, and can in some cases increase reaction efficiency (Wu et al., 2014). The technology has been used in a wide array of biological applications: as a platform for high-throughput drug screening assays (Brouzes et al., 2009), in the analysis of protein expression of single cells (Huebner et al., 2007) and even in the engineering of stem cell growth niches (Allazetta and Lutolf, 2015).

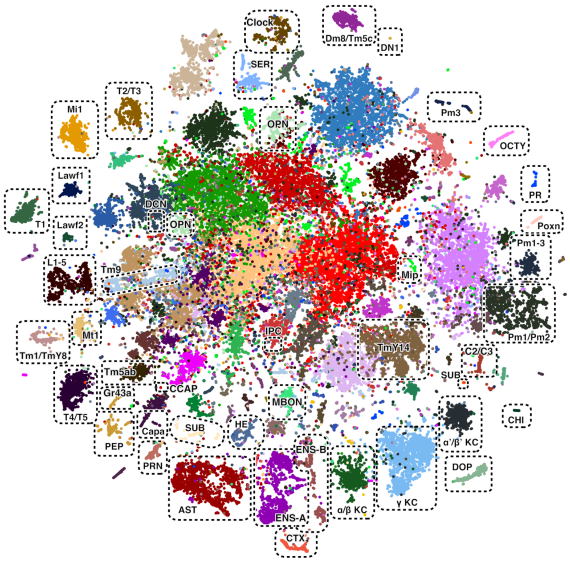


Figure C.2: Single-cell atlas of Drosophila brain. Transcriptomes of 57k single cells are sequenced and computationally separated to discern different cell types. (Davie et al., 2018)

In the past few years, droplet microfluidics have also been applied in single-cell technology. Single cells are co-encapsulated into nanolitre droplets together with various reagents and barcoded hydrogel beads in order to capture their mRNA content, which is then molecularly barcoded. Due to the barcoding process, mRNA from single cells can be sequenced together and computationally separated. An example of such an application is given in figure C.2. Here, 57 000 fly brain cells were processed using droplet-microfluidic single-cell technology, and computationally analysed to form a comprehensive atlas of the different cell types present in the Drosophila brain.

This thesis revolves around optimisation and innovation in the technical/molecular side of the single-cell analysis field. The work can be split in two main lines: the first line (chapter 2) focuses on the implementation and optimisation of inDrop, a pre-existing droplet-microfluidic single-cell RNA sequencing platform (Klein et al., 2015; Žilionis et al., 2017). In the second line (chapter 3), the droplet microfluidics and molecular engineering expertise acquired in the first part is applied to develop an entirely new protocol. Here, we attempt to prototype a microfluidic approach for ATAC-seq ("Drop-ATAC"), which as of today only exists as a commercial package. As both the optimised inDrop and Drop-ATAC are situated on the far end of the throughput scale, they pose high potential in screening assays where large numbers of cells need to be analysed rapidly.

The end-goal is to explore the new possibilities posed by combining droplet microfluidics and modern molecular cell-analysis. This combination has only recently emerged as a powerful tool in (systems) biology, but is already revolutionising the field. As the droplet-microfluidic single-cell techniques are only in their infancy, there is still a lot of room for improvement and innovation, an opportunity we address in this work.

1 | Single-Cell Omics: State of the Art

All life originates from and operates as a massive, interconnected network of molecules. The term 'omics' refers to the systematic characterisation of these elements in order to elucidate their relation to development, function and disease. Since the advent of genomics in the mid 90s shortly after the first genomes of small organisms were sequenced, several new 'omics' fields have emerged. Transcriptomics focuses on the quantification and characterisation of all RNA transcripts. Epigenomics revolves around chromatin changes and their relation to gene expression. Since the focus of omics is a complete overview of an entire class of molecules, the availability of high-throughput analysis methods is critical in these fields (Hood et al., 2004; Patti et al., 2012; Patterson and Aebersold, 2003)

Since the mid 2000s, several bulk omics analysis techniques have allowed researchers to explore the complete transcriptomes and epigenomes of pre-defined cell populations. These techniques often require explicit cell pre-selection, for example using fluorescence-activated cell sorting (FACS) and established marker genes. In the bulk approach, the sampled population is isolated as a group and further treated as a single, homogeneous entity. Any information on heterogeneity within that population is disregarded, masking the presence of previously undiscovered and rare cell types (Bengtsson et al., 2005; Wang and Bodovitz, 2010; Kolodziejczyk et al., 2015; Grün and Van Oudenaarden, 2015).

For decades, however, researchers have acknowledged that morphologically homogeneous cells may exhibit vastly different transcriptional profiles. One of the earliest mentions of this idea occurred in 1992, when Eberwine et al. discovered that in a sample of 15 rat hippocampus pyramidal cells, 2 cells exhibited significantly different messenger RNA (mRNA) profiles than the rest of the population (Eberwine et al., 1992). Realising the power of their discovery, the authors suggested that analysing single cells could help identify new cell types, reveal gene transcript subtypes and even directly determine the influence of external stimuli on a cell's gene expression. However, the single-cell analysis methods employed by Eberwine et al. were expensive and laborious, prohibiting large-scale application of the technique. Indeed, the challenges associated with analysing single cells are numerous and complex:

1. Single cells need to be isolated from complex tissue matrices while preserving their natural state as closely as possible.
2. The minute nucleic acid content of single cells (pg) calls for sensitive library generation protocols before sequencing can occur. This problem is exacerbated by the relative mRNA transcript levels of different genes, which can span orders of

magnitudes between and within single cells from the same tissue (Bengtsson et al., 2005).

3. Due to the large number of cells, the aforementioned challenges need to be tackled in ways that permit high throughput and low reaction volumes to reduce time, cost and labour while maintaining sufficient efficiency and sensitivity.

In the past few years, technological progress in microfluidics and next-generation sequencing have enabled researchers to overcome many of these challenges. Using advanced separation techniques, single cells can now be isolated and analysed individually with increasing accuracy. Ultra-sensitive amplification techniques and transcript barcoding allow sequencing of a single cell's DNA or mRNA content. Since these single-cell sequencing techniques are usually dependent only on basic reagents and machinery, many research institutes have started to adopt them, leading to a rapid increase in the number of publications using them (figure 1.1).

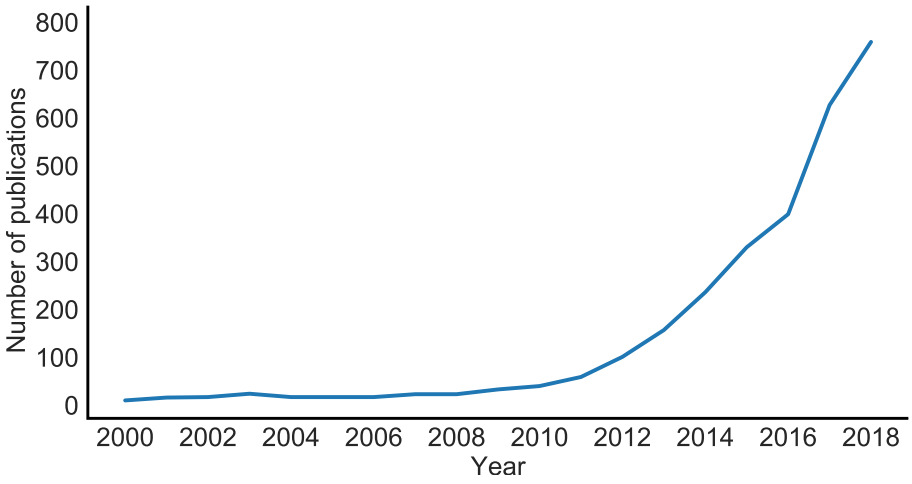


Figure 1.1: Number of abstracts published in scientific journals containing the words "single-cell" and "sequencing". Data gathered from Scopus.

Single-cell analysis techniques have been used to uncover transcriptional heterogeneity in tissues previously deemed homogeneous, to identify new transcripts, map cell state trajectories in (pseudo)time, to study the effect of gene knockdowns and to unravel gene regulatory networks (Tang et al., 2011).

In this literature study, the current state of the art in single-cell research is examined. First, a selection of popular single-cell analysis techniques is explained using their cell separation strategy as a basic classification criterion. Figure 1.2 shows the general subdivision of the reviewed methods. We start with the most straightforward approach: applying established bulk analysis methods on single-cells sorted in microwells. Then, two techniques that automatically load cells into microfluidic arrays in order to decrease labour are

discussed. Third, we look at how droplet microfluidics lead to a 100-fold increase in cell throughput. Fourth, a number of ultra-high throughput well-based protocols is discussed briefly. After the state of the art is established, we go over some of the most important breakthroughs in biology fueled by single-cell research. Finally, a short conclusion is drawn on single-cell research and possible future perspectives of single-cell omics are posed.

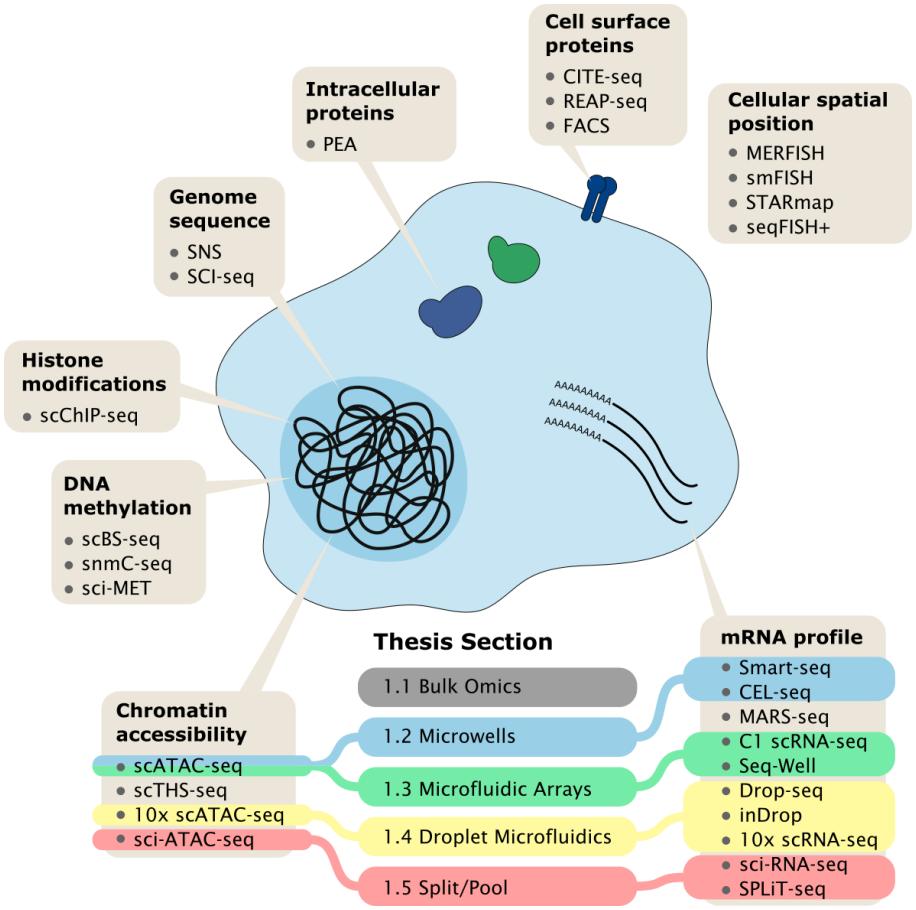


Figure 1.2: Current single-cell technology landscape. An array of analysis techniques can be used to study genomics, transcriptomics, proteomics and epigenomics at single-cell resolution. In this literature study, a number of important single-cell techniques are explained and compared. Modelled after Stuart and Satija (2019).

1.1 Bulk Omics Techniques

Before single-cell analysis techniques can be discussed, the bulk techniques which they are founded on need to be thoroughly understood. The following section will briefly explain RNA-seq and ATAC-seq, which have both become a major point of focus in the single-cell omics landscape and will play a key role in the experimental part of this work.

RNA-seq

The transcriptome comprises all mRNA transcripts, the functional elements of the genome, in a cell population. Transcriptomics focuses on the quantification of these mRNA transcripts and how their levels change during disease and development. Previous generations of transcriptomic techniques, such as expression microarrays, suffer from hybridisation artefacts, cannot detect splice variants or new genes, have a low dynamic range and yield semi-quantitative data due to the limitations of fluorescence (Wang et al., 2009; Tang et al., 2011). RNA-seq, first described in 2008, was the first comprehensive and simple protocol to offer accurate quantification of a cell population's gene expression without requiring cloning of the sample RNA (Mortazavi et al., 2008).

RNA-seq evades the pitfalls of hybridisation and fluorescence-dependent methods by sequencing the cDNA of captured mRNA (fig. 1.3). By using oligo(dT) magnetic beads to capture poly-A tails (repetitions of adenosine nucleotides that is incorporated in all mature mRNA transcripts), transcripts can be detected without explicit previous knowledge of their sequence. In the next step, fragmentation of the captured mRNA destroys secondary structures and mitigates transcript length variation, improving upon random hexamer reverse transcription priming. Optionally, Arabidopsis and phage lambda RNA standards can be co-processed with the sample in order to allow absolute transcript quantification. These known RNA sequences are added at set concentrations, yielding a linear standard curve which can be used to relate read count and transcript concentration. In the 140 million reads generated by RNA-seq, Mortazavi et al. detected alternative splice variants for 3 462 genes and identified 596 new candidate transcripts in mouse brain, liver and muscle.

RNA-seq offers uniform transcript coverage, accurate quantification of transcripts up to 1 transcript/cell, a dynamic quantification range of five orders of magnitude, and is able to detect transcripts outside of the prior reference transcriptome. Generated data is highly replicable and correlates well to RNA-microarray data. RNA-seq allows researchers to detect differences in transcriptional profile in response to certain conditions or along development, to catalogue different species of transcripts and to determine the transcriptional structure of genes. Due to its simplicity and low cost, RNA-seq remains one of the main techniques used in transcriptomics to date. However, the method is still sensitive to strongly related sequences in the exome such as gene duplications and paralogs (Wang et al., 2009).

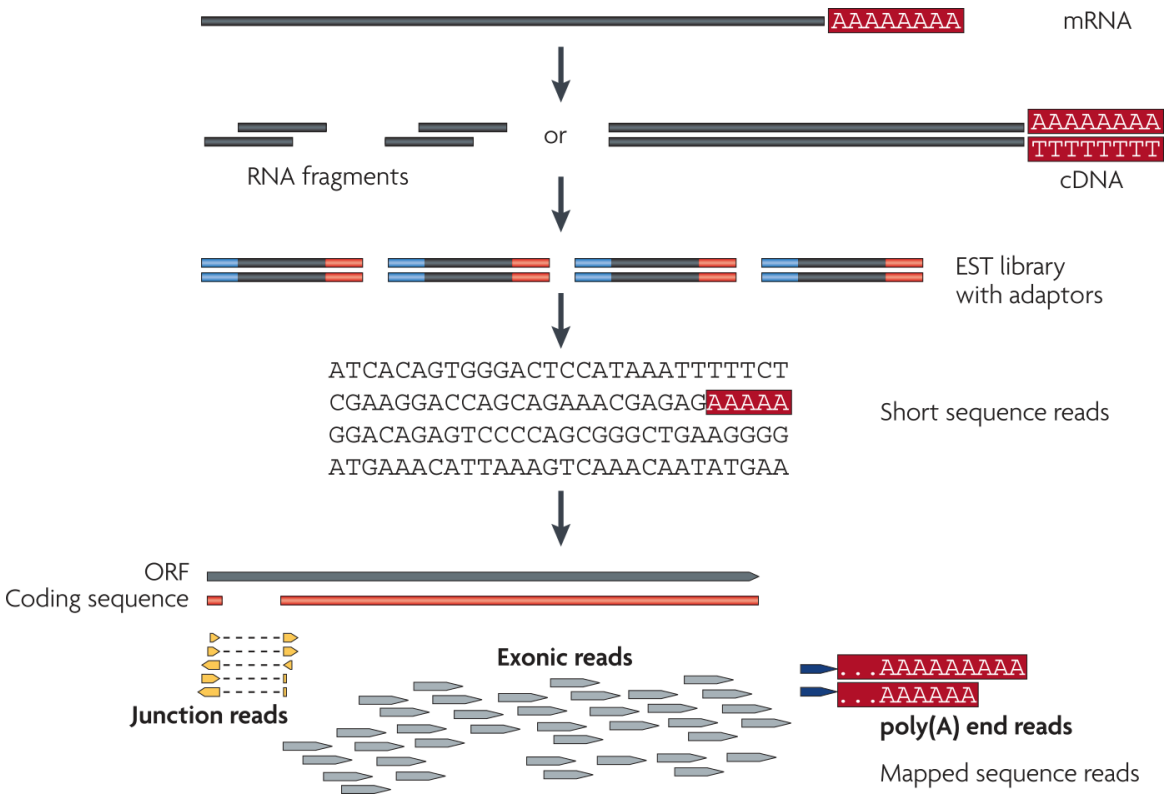


Figure 1.3: RNA-seq workflow. Processed mRNA transcripts are captured using poly-T magnetic beads and hydrolised to 200-300 bp fragments to remove secondary structures. After reverse transcription using random hexamer primers and second strand synthesis, the fragments are processed into a sequencing library. Sequencing and alignment to a reference transcriptome or exome yields an accurate overview of the sample tissue's expression profile. Additionally, clusters of reads that do not map to previously known exons can be organised into candidate exons in order to identify newly transcribed regions. Adapted from Wang et al. 2009

In 2012, a method for absolute transcript quantification was developed (Kivioja et al., 2012). Here, each reverse transcription primer carries a sequence of 10 additional random nucleotides, called the unique molecular identifier (UMI). This UMI is incorporated into each cDNA transcript. After PCR, each amplification product can now be traced back to its transcript of origin by its UMI, allowing absolute quantification of the original number of transcripts present in the sample.

ATAC-seq

The DNA of eukaryotic cells is organised into chromatin, an organised complex of nucleic acid and histone nucleoproteins. This form of compaction allows the cell's entire genome to fit into the nuclear subspace (Kornberg, 1974). When DNA is packed tightly around these nucleoproteins, the underlying genes sequences are inaccessible to transcription factors, inhibiting transcription of the restricted DNA. However, using an array of mechanisms such as DNA methylation, nucleosome positioning and histone modification, the cell can locally remodel the steric accessibility of chromatin to allow gene transcription. Chromatin accessibility state is therefore a leading indicator of which genes are actively expressed in a cell and, together with DNA methylation and histone modification, forms the physical basis of epigenomics (Jaenisch and Bird, 2003; Kouzarides, 2007; Schones and Zhao, 2008; Bannister and Kouzarides, 2011). Epigenomic variation has been shown to be highly variable in time, between cell populations and across cell generations and thus effectively provides a layer of information "on top of" the cell's genome. Studying epigenomic phenomena can thus help construct models of gene regulatory pathways. A thorough understanding of these pathways may ultimately help answer the question of how different cell phenotypes can arise from genetically identical precursor cells (Johannes et al., 2008).

The chromatin state of cell populations has previously been studied using techniques such as FAIRE-seq and DNase-seq (Giresi et al., 2007; Song and Crawford, 2010; Gaulton et al., 2010; Song et al., 2011). However, these methods involve steps such as chloroform extraction and gel purification which may lead to loss of sensitivity. The assay for transposase-accessible chromatin using sequencing (ATAC-seq), a technique published by Buenrostro et al. in 2013, allows researchers to investigate the chromatin condensation state of a cell population without such potentially loss-prone steps. ATAC-seq relies on the activity of a hyperactive, mutated Tn5 transposase which (near)-randomly fragments only sterically accessible chromatin regions. In contrast to DNase, another enzyme that randomly fragments accessible chromatin, Tn5 ligates oligonucleotide adapters where it cleaves DNA. Strands fragmented by Tn5 are thus flanked by two adapter sequences. These adapters can then be used in PCR as hybridisation sites for specialised sequencing primers, simplifying post-fragmentation processing (fig. 1.4).

Random fragmentation of accessible DNA results in many short fragments of the accessible chromatin and fewer but longer fragments of the inaccessible regions. The fragment length and count can therefore be used to determine the "accessibility" of a given chromatin region. Areas of significant enrichment in the read distribution are detected computationally in a process called peak calling, yielding information on which regions of the genome are accessible in the cell population.

ATAC-seq requires a starting number of cells 2 - 3 orders of magnitude lower than DNase-seq and FAIRE-seq (~50 000 versus ~50 000 000). The protocol takes only ~5 hours from sample collection to sequencing compared to 3.5 days for DNase-seq, but the resulting data is of similar quality (Buenrostro et al., 2013). In 2017, a more widely-

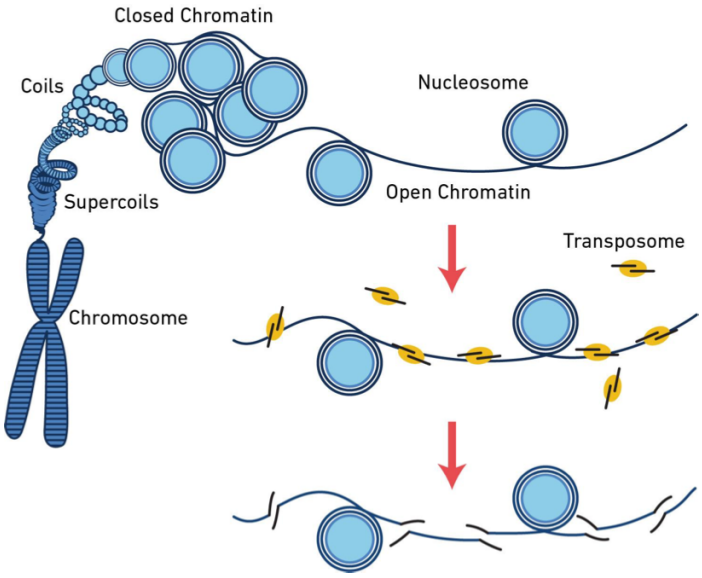


Figure 1.4: ATAC-seq concept. Tn5 transposase fragments DNA only in sterically accessible chromatin and ligates adapters. A subsequent PCR uses these adapters to incorporate sequencing adapters and sample indexes. After sequencing, reads are mapped to a reference. Source: 10x Cell Ranger ATAC Algorithm brochure

applicable bulk ATAC-seq protocol, dubbed "Omni-ATAC-seq", was released. This updated version is applicable to a broader range of cell types and yields a higher fraction of reads in peaks than the standard ATAC-seq. Importantly, the improved protocol also efficiently removes mitochondrial DNA from the transposition reaction, hereby reducing sequencing costs (Corces et al., 2017). As will be shown later, recent efforts have scaled ATAC-seq up to single-cell resolution (Buenrostro et al., 2015; Chen et al., 2018).

1.2 Microwell-based Single-Cell Omics Techniques

Some of the earliest efforts to analyse the nucleic acid content of single cells simply applied bulk analysis techniques to single cell lysate suspended in small compartments. For example, Eberwine et al. carried out reverse transcription of mRNA by injecting a live cell with viral reverse transcriptase and oligo-dT primers and aspirating the cell contents into the glass electrode. Due to the high fixed labour cost and limits of technology at the time, their approach led to extremely low throughput compared to today's standard: 5 days of work for 1 cell then, compared to 5 days of work for 10 000 cells today (Hashimshony et al., 2012). The following section shows recent continuations on the microcompartment approach, which have become widespread due to their ease of use.

Smart-seq

Switching mechanism at the 5' end of RNA template sequencing (Smart-seq), originally formulated by Ramsköld et al. in 2012, was one of the first single-cell mRNA-seq protocols and the first to provide full transcript coverage. The technique is modelled after an earlier mRNA-seq protocol by Tang et al. (2009), who sequenced the transcriptome of a single mouse blastomere.

Ramsköld et al. first applied Smart-seq to 42 human cells manually picked from dissociated tissues using microscope-assisted micromanipulation. Single cells were further treated separately in microwells. Figure 1.5 shows the steps involved in Smart-seq library generation. An important addition of Smart-seq compared to the Tang et al. protocol is the use of template switching. During reverse transcription, the Moloney murine leukemia virus (MMLV) reverse transcriptase adds extra cytosines to the 3' end of the cDNA strand, allowing a so-called template-switching oligo (TSO) to hybridise. MMLV reverse transcriptase then continues cDNA synthesis using the TSO as a template in a process called template switching. The sequence complementary to the TSO is thus incorporated in the cDNA library and is used as a PCR priming site in further amplification steps. Importantly, the SMART primer can hybridise to both ends of the cDNA transcript. The PCR product of short templates can thus form loops which impede further amplification. This mechanism corrects the natural short-fragment bias associated with PCR amplification. The template-switching reverse transcription approach is more user-friendly and time-efficient than linear amplification by in-vitro transcription. Compared to a regular reverse transcription, template-switching also produces cDNA transcripts of the complete RNA template. This means that any splicing information carried by the very distal ends of the transcript is retained after sequencing.

An improvement on Smart-seq, Smart-seq2, was published by Picelli et al. in 2013. Here, single cell isolation was performed on 262 human and mouse cells. In the improved Smart-seq2 protocol, cDNA-yield was increased twofold by incorporating a locked nucleic

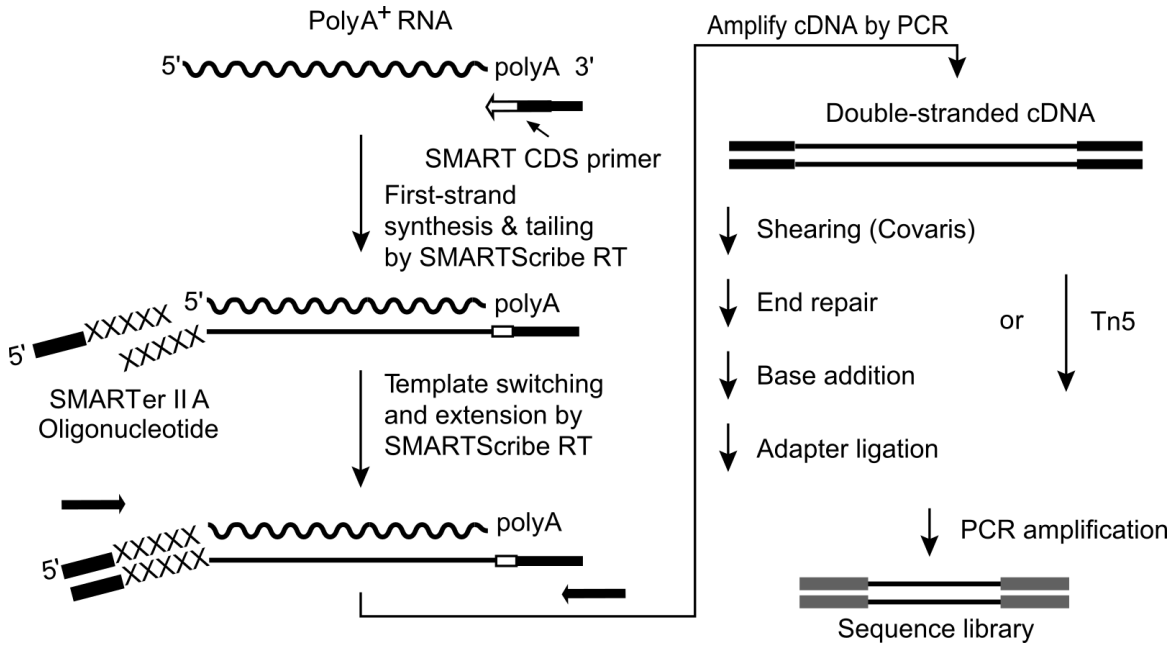


Figure 1.5: Smart-seq library generation. Single cells are manually isolated from dissociated sample tissue and placed in separate microwells. The cells are then lysed and their polyadenylated mRNA content is captured by oligo(dT) reverse transcription primers. Full transcript cDNA is synthesised using MMLV reverse transcriptase and pre-amplified by ISPCR. Double-stranded cDNA is then fragmented and tagged with sequencing primers and adapters by Tn5. Alternatively, shearing and adapter ligation can be used. After further PCR amplification, the library is ready for next-generation sequencing. Adapted from Ramsköld et al., 2012.

acid (LNA) in the TSO. This small change led to better thermal stability of the TSO-cDNA duplex. Further reagent concentration optimisations resulted in an overall increase in sensitivity and accuracy relative to Smart-seq. Picelli et al. report that Smart-seq2 detects ~12k genes from HEK cells compared to ~10k genes detected by first generation Smart-seq. Additionally, cells were isolated by distributing µl volumes of strongly diluted cell suspensions into microwells instead of relying on manual cell picking. This allows for the use of automated liquid handling, but incurs an additional increase in reagent cost due to empty wells. An overview of the Smart-seq2 library generation workflow is given in figure 1.6.

Smart-seq and Smart-seq2 yield full transcript information and thus allow researchers to study both distal and proximal splicing events, novel exon detection, single-cell SNP detection, and allele-specific gene expression (Kolodziejczyk et al., 2015). However, using microwells reduces throughput and increases reaction volumes, leading to an overall increased cost per cell. Cell selection based on dilution also inhibits the detection of rare

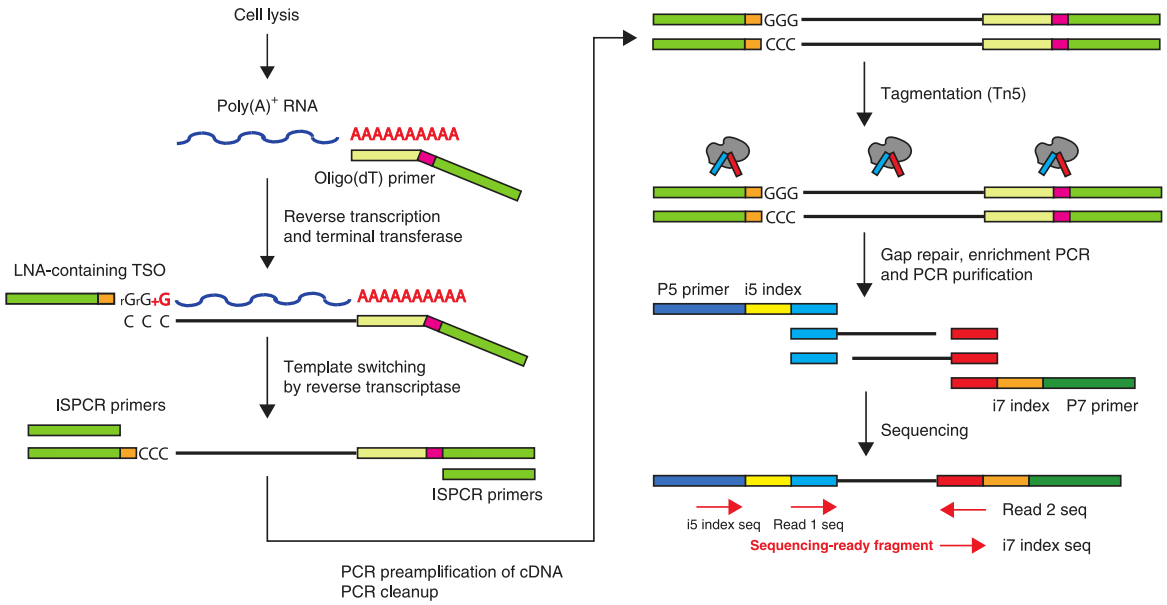


Figure 1.6: Smart-seq2 library generation. Single cells from a dilute cell suspension are distributed into separate microwells. The cells are then lysed and their polyadenylated mRNA content is captured by oligo(dT) reverse transcription primers. After full-transcript reverse transcription by template switching using an LNA-TSO, the resulting cDNA is pre-amplified and subsequently tagmented with sequencing adapters using Tn5 transposase. The library is then further amplified and sequenced. Adapted from Picelli et al., 2014b.

cell types in large samples. The Smart-seq and Smart-seq2 protocols do not incorporate UMIs into the cDNA, making absolute quantification of transcripts impossible. The use of double-stranded cDNA also discards information about transcript strand specificity. Moreover, the random fragmentation nature of the Tn5 tagmentation reaction leads to a reduced sequencing coverage of the very 5' ends of transcripts, which carry the transcription start site (TSS) and 5' untranslated region (5'UTR). Studying these important regions is therefore difficult using Smart-seq. In conclusion, Smart-seq2 is a highly sensitive, but low-throughput method best suited for small cell populations where the complete mRNA transcript is needed to investigate distal splicing or presence of SNPs. Since Smart-seq2 requires only off-the shelf reagents and equipment, it has become a widespread single-cell RNA-seq protocol (Picelli, 2017).

CEL-seq

Cell expression by linear amplification and sequencing (CEL-seq), published in 2012 by Hashimshony et al., is a single-cell transcriptomics technique revolving around mRNA barcoding followed by in-vitro transcription (IVT) for linear amplification. IVT leads to more reproducible and sensitive results compared to exponential PCR amplification, but requires ~400 pg of cDNA compared to the average eukaryotic cell's mRNA content of ~1 pg (Tang et al., 2011; Hashimshony et al., 2012). CEL-seq satisfies this requirement by pooling barcoded cDNA from different cells of origin before applying IVT. Due to the in vitro transcription step which generates only sense RNA from antisense cDNA, CEL-seq generates strand-specific sequencing libraries. For a detailed overview of the CEL-seq protocol, see (Hashimshony et al., 2012). A second generation protocol, CEL-seq2, was published in 2016. Figure 1.7 shows the generalised CEL-seq2 workflow.

CEL-seq2 increases reverse transcription efficiency by shortening the reverse transcription primer. This leads to a higher fraction of mRNA transcripts being detected. Incorporating Illumina sequencing adapters during reverse transcription eliminates a PCR-ligation step, increasing read mappability from 60.9% to 93.8%. CEL-seq2 also incorporates a UMI into every cDNA strand, enabling absolute transcript quantification. These incremental improvements together lead to a 30% increase in number of genes detected compared to the first generation CEL-seq. It is also shown that CEL-seq2 can be readily performed on the Fluidigm C1 microfluidic platform for increased gene detection and reduced labour cost (Hashimshony et al., 2016).

Compared to Smart-seq, CEL-seq2 yields a near 100% increase in genes detected on the same cell sample (Hashimshony et al., 2016). CEL-seq2, however, is strongly 3'-biased and can therefore not provide information on distal splicing as opposed to Smart-seq. Importantly, CEL-seq's in-vitro transcription amplifies at a lower rate and is more contrived than Smart-seq's PCR, taking 13 hours per sample in the original CEL-seq protocol. Regardless, CEL-seq's sensitivity and absolute quantification possibilities often outweigh its disadvantages and make it suitable technique for many transcriptomics applications.

Microwell scATAC-seq

A scATAC-seq approach which does not explicitly require the use of microfluidics was published in 2018 by the Teichmann Group (Chen et al., 2018). Their protocol relies on an interesting peculiarity about Tn5: when Tn5 binds to accessible chromatin, it cleaves chromosomal DNA and inserts its adapters at the 5' end of each opposite strand. However, during this process, Tn5 remains clamped on the DNA strand until it is denatured by either SDS or a heat shock (Picelli et al., 2014a). During this process, the cell is lysed, but its DNA content stays confined to the nucleus. Only upon denaturation of Tn5, the fragmented DNA is released from the enzyme. In the Teichmann protocol, bulk tagmented

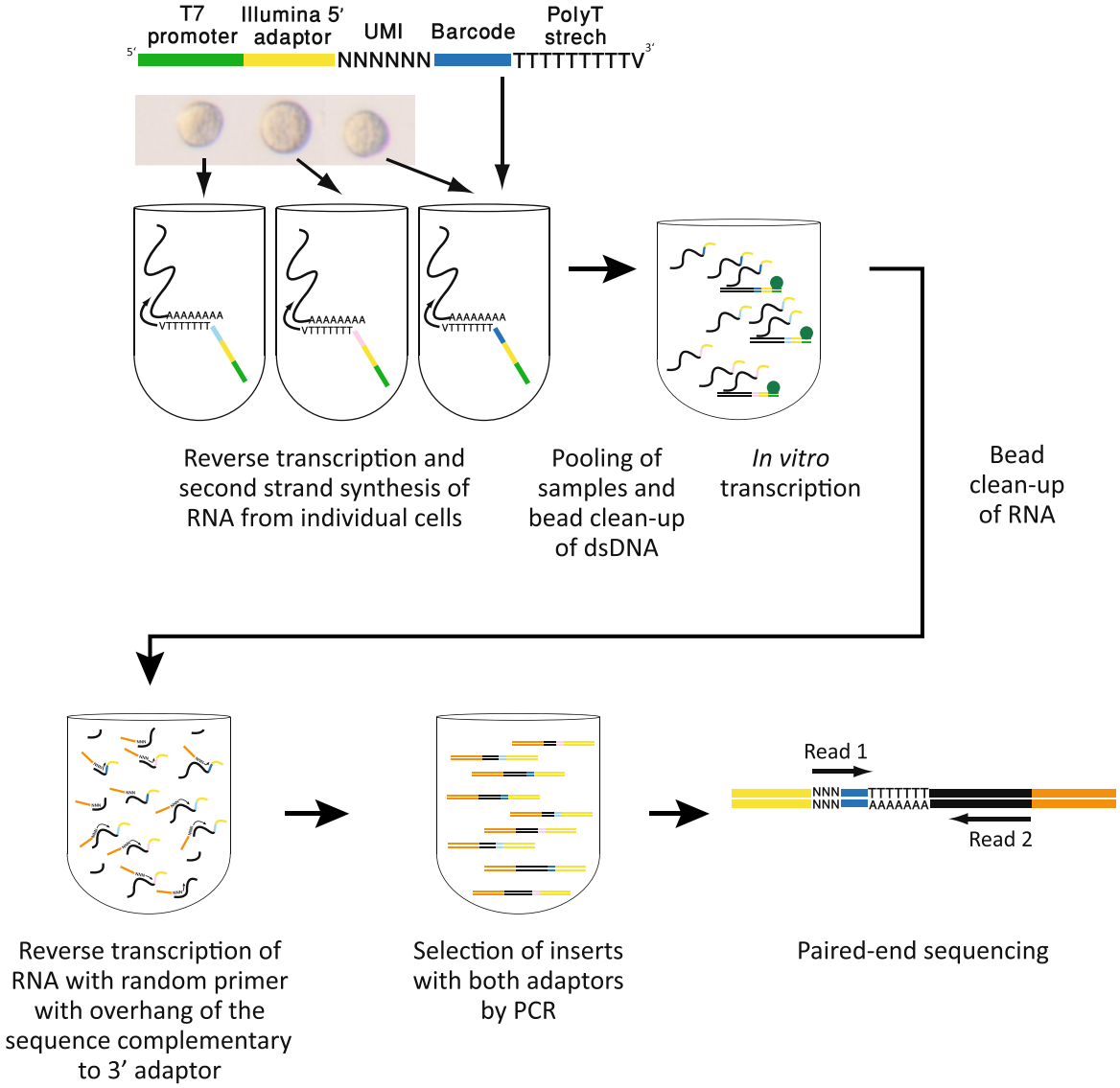


Figure 1.7: CEL-seq2 library generation. Single cells are lysed and poly-A$^+$ mRNA is reverse transcribed using a poly-T primer that includes a cell-specific barcode, a transcript-specific UMI, an Illumina 5' adapter and a T7 promoter. Pooled cDNA is transcribed in vitro and the amplified RNA is purified. Then, using a random primer with overhang complementary to the 3'-adapter, inserts flanked with a 3' Illumina sequencing adapter are generated which are then purified and processed for paired-end sequencing. Adapted from Hashimshony et al. (2016)

nuclei are sorted into a microtiter plate containing lysis buffer. The tagmentation reaction is then stopped using SDS and further reactions are performed separately in each microwell. An overview of the technique is shown in figure 1.8.

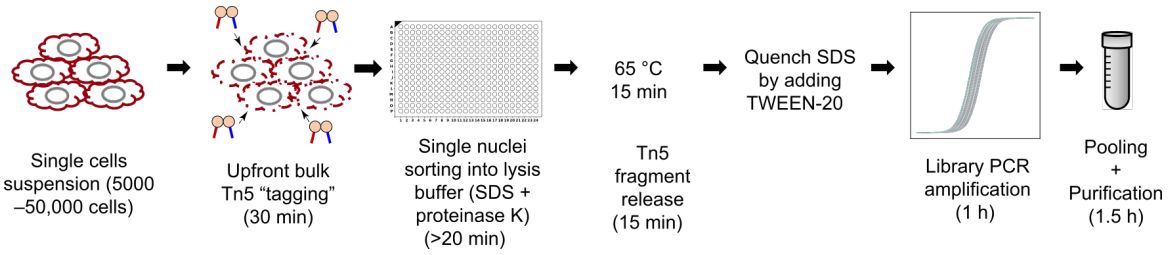


Figure 1.8: Teichmann scATAC-seq workflow. Tagmentation is performed in bulk by incubating 50 000 cells with Tn5 transposase. "Tagmented" nuclei are sorted into 384-well plates containing lysis buffer with SDS and proteinase K, which inactivate Tn5. After Tn5 is inactivated and DNA fragments are released, SDS is quenched by addition of Tween-20, preventing non-specific inactivation of downstream enzymes. Fragments generated by Tn5 are then indexed using indexing PCR, amplified, pooled and sequenced. Taken from Chen et al. (2018).

In contrast to earlier scATAC-seq protocols, the Teichmann protocol does not require intermediate purification of the DNA, simplifying the tagmentation workflow. Chen et al. benchmarked their scATAC-seq method against the Buenrostro et al. C1 scATAC-seq protocol (which will be shown in section 1.3) and found that their own protocol generated higher quality data in a shorter time frame. The whole procedure takes place in the same plate and does not require intermediate purification steps, as opposed to previous scATAC-seq protocols which employ bead purification of DNA after tagmentation. Chen et al. also show that immunostained cells retain their staining after bulk tagmentation, meaning FACS can be used to filter rare cell types from the sample post-tagmentation. Additionally, the bulk tagmentation approach cuts cost of Tn5 per cell considerably but the subsequent microcompartment-based approach still incurs a substantial cost due to the 20 µl PCR reaction volume.

Microwell Approaches: Key Takeaway

Due to their user-friendliness and modest requirements in terms of equipment, the use of plate-based techniques has become widespread. However, microcompartment-based single-cell analysis techniques suffer from a number of inherent disadvantages:

1. Isolating and/or sorting single cells into microwells and performing reactions on them is tedious, especially when performed manually. FACS and automated liquid handling may be used to partially solve these problems.
2. Performing single-cell reactions in a microwell plate leads to large reagent volumes per cell, resulting in a high cost and, more importantly, decreased reaction efficiency due to dilution effects (Wu et al., 2014).
3. Handling reaction plates can form a throughput bottle neck in parallel processing of a large number of cells, for example during reverse transcription, where one heat block can only process 96 cells.

Despite these drawbacks, microcompartment-based techniques have proven to be viable tools to extract information on single-cell resolution DNA methylation, chromatin condensation, gene expression and genomic profile. Generally speaking, microcompartment techniques such as Smart-seq and CEL-seq yield data of higher quality than the high-throughput methods shown later (Ziegenhain et al., 2017). Together, the protocols explained in this sections have led to a number of groundbreaking results of which a limited selection will be shown in 1.6.

The greatest advantage that microwells offer can be found not in accuracy, cost or throughput, but in flexibility. It is straightforward to manipulate and sample the contents of a microwell plate to optimise the reactions carried out within. Moreover, multi-step methods involving sequential addition and purification can easily be carried out on an accessible well platform. Many high-throughput techniques shown in the following sections were born in a microwell, and most likely some of the more complicated techniques we will see in the future will start in a well as well.

1.3 Microfluidic Arrays

The previous section has shown how single-cell data can be obtained by simply applying bulk analysis methods to a single cell in a microwell. A major bottleneck in the application of these assays is the distribution of single cells into individual microcompartments. Common methods include manual picking, FACS or Poisson-distributed dilution, each with their own set of disadvantages. In the following section, a number of techniques which separate or trap single cells into (sub)nanolitre volume compartments are introduced.

Fluidigm C1

The C1 comprises a benchtop microfluidic controller that contains the pneumatic hardware, and a disposable integrated fluidic circuit (IFC) (figure 1.9), which hosts the microfluidic channels through which the sample is routed. Together, they form a platform that can isolate single cells from dilute cell suspensions into nanolitre capture sites. In section 1.2, we briefly mentioned that the Smart-seq2 protocol can be implemented on the C1. Preceding Smart-seq2, Buenrostro et al. published a C1-based scATAC-seq protocol (Buenrostro et al., 2015). Here, tagmentation and PCR occur on the C1 IFC, after which single-cell libraries are collected and further amplified with cell-identifying barcoded primers. For an overview of the Buenrostro et al. scATAC-seq method, see figure 1.10.

The greatest advantage of the C1 compared to conventional microwells are ease of use and semi-automation at a minimal loss of data quality. During cell trapping, tagmentation and pre-amplification, minimal human interference is necessary. However, barcoding and subsequent amplification of the single-cell libraries still need to happen in a microwell plate. The C1 approach is therefore practically an automated cell-capture and nucleic acid capture method followed by microcompartment-style barcoding. As the IFC can only process 96 cells in parallel, high-throughput application of the C1 is difficult. Moreover, the C1 IFC only accommodates three cell size classes per chip design (5-10, 10-17 and 17-25 µm in diameter), leading to a size-biased selection. The very nature of the C1's cell trapping mechanism also complicates the capture of non-spherical cells. Macosko et al., authors of a competing microfluidic scRNA-seq method covered in section 1.4, report that 30% of C1-generated libraries contain mixed-species contamination, which is unusually high for single-cell methods (Macosko et al., 2015). Shalek et al. report a more forgiving multiplet rate of 11%. Additionally, the 96-cell IFC takes an input sample of at least 1000 cells, imposing an effective capture efficiency of 1-10% (Žilionis et al., 2017). This last drawback prohibits the application of C1-based cell capture on protocols where information on rare cell types is key, such as de novo cell typing. It is due to these limitations and strong competition from droplet microfluidics approaches that the C1 has not been able hold a large market share, appearing in 228 publications over a span of 8 years (Fluidigm Website, 2019), compared to 391 publications for the 10x Chromium, which was released

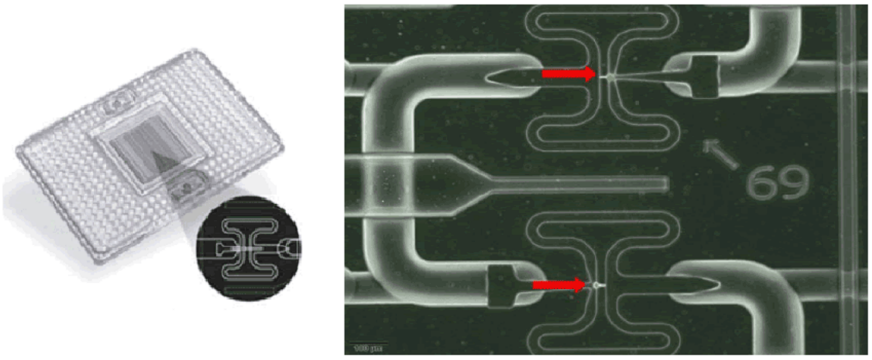


Figure 1.9: Fluidigm C1 IFC. A suspension of single cells is added to the IFC's sample inlet and inserted into the C1 control unit. The control unit pneumatically pumps the suspension through a serpentine channel containing a series of capture sites. Once a cell occupies a capture site, access to the site's collection chamber is blocked, routing subsequent cells to the next capture site. This results in a series of capture sites along the path each containing a single cell. After a run time of 1 hour, cell occupancy is checked using (fluorescence) microscopy. Then, lysis reagents are routed through the serpentine channels, lysing cells and clearing the obstructed path to the collection chambers, resulting in a flow of lysate from each capture site to its respective collection chamber. Next, reverse transcription or amplification reagents can be routed to the chamber. The reaction products from individual cells are then collected into a separate microwell. During each microfluidic unit operation, a set of peristaltic valves seal off every capture site in order to minimise contamination. Adapted from Azizi et al. (2014).

in 2016. Still, the C1 has found niche applications in the automisation of specialised and complicated protocols such as sci-CAR (see section 1.5).

Seq-Well

A rather straightforward microfluidic cell loading strategy dubbed Seq-Well was published in 2017 (Gierahn et al., 2017). In this method, cells and beads carrying barcoded poly-dT primers are loaded into an array of picolitre wells. Here, the barcoded poly-dT primers capture cellular mRNA. The beads are then pooled for bulk reverse transcription. An overview of the Seq-Well method is given in figure 1.11.

Seq-Well improves on earlier picowell methods by Fan et al. (2015) and Yuan and Sims (2016) by chemically sealing the array of picolitre wells using semi-permeable membrane which allows for buffer exchange, but prevents cross-contamination between wells. The inner surface of the picowells is also treated to prevent non-specific mRNA binding. Similarly

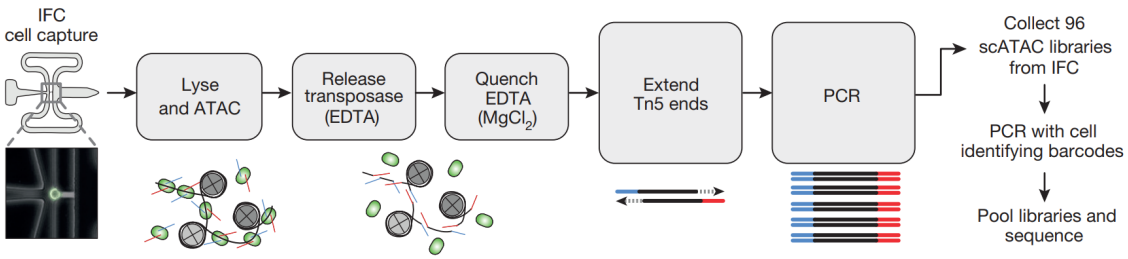


Figure 1.10: Buenrostro scATAC-seq. 5 µl of a 300 cells/µl suspension is placed into the IFC sample inlet. Single cells are trapped into capture sites and lysed. In an individual reaction chamber, each cell is tagmented in 13.5 nl of tagmentation mix. After EDTA-mediated release of Tn5, the tagged fragments are pre-amplified by PCR on-chip. After amplification, individual cell samples are collected and individually barcoded using PCR. Only then are the single-cell libraries collected and pooled for sequencing. Taken from Buenrostro et al. (2015).

to Smart-seq, Seq-Well uses a template-switching library generation strategy to facilitate amplification and to produce full-length transcripts. In short, Seq-Well is a simple and portable protocol for loading a large number of cells into microwells rapidly. However, Žilionis et al. remark that not all mRNA is capture by the barcoded beads, possibly leading to contamination at the pooling step after hybridisation (Žilionis et al., 2017). Gierahn et al. also observed a multiplet rate of 11.4% when 20 000 of the 83 200 available picowells were loaded with cells. 77.5% of the Seq-Well reads could be mapped to a reference exome and ~6 000 human genes were detected. This is higher than droplet-microfluidics techniques such as Drop-seq (~5 000 human genes) and 10x Chromium (~4 600 human genes) but lower than a true microwell protocol such as Smart-seq2 (~12 000 human genes).

Microfluidic Arrays: Key Takeaway

Microfluidic array methods have succeeded in reducing the significant labour costs associated with single-cell experiments. As always, a trade-off is made between data complexity and throughput. Single-cell experiments on the Fluidigm C1 platform offer high quality data consistent with explicit well-based techniques, but at a high cost and low throughput while Seq-Well offers high-throughput, portability and low cost at the expense of sensitivity and specificity. Importantly, both methods still allow customisation and adaptation of the reactions performed on the isolated cells. Microfluidic array techniques therefore form an intermediate between well-based protocols and the droplet-microfluidic techniques shown in section 1.4.

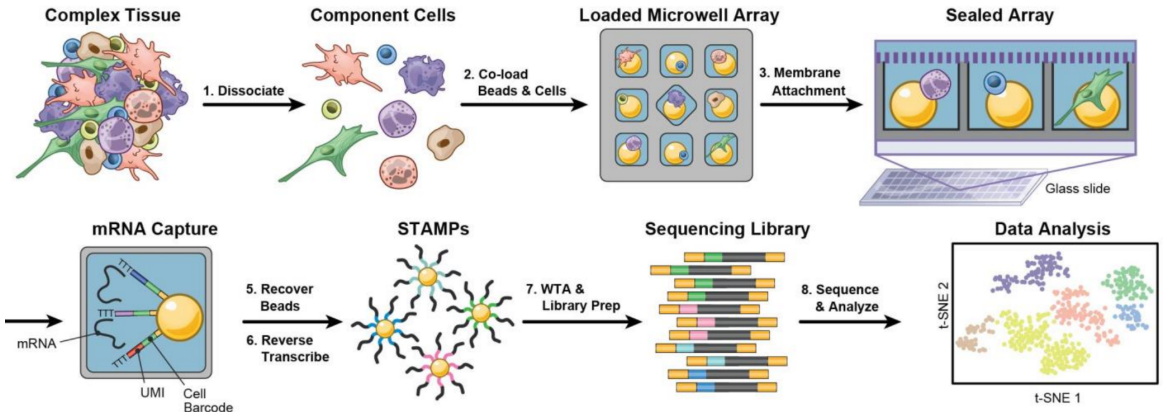


Figure 1.11: Seq-Well cell loading and library generation. Cells are co-loaded together with barcoded microbeads in a PDMS array containing 83 200 picowells. Each bead oligo holds a cellular barcode, a UMI and a poly-dT tail to capture mRNA poly-A tails. The wells are then sealed using a semi-permeable membrane and submerged in lysis buffer. In each well, released cellular mRNA is captured by the poly-dT primers on the barcoded beads. The plates are unsealed, and beads are pooled. A bulk template-switching reverse transcription reaction then generates the single-cell transcriptomes attached to microparticles (STAMPs), which are then amplified and sequenced. Adapted from Gierahn et al. (2017).

1.4 Droplet Microfluidic Single Cell Omics Techniques

In 2015, two remarkable single-cell RNA sequencing methods debuted in the same issue of Cell (Klein et al., 2015; Macosko et al., 2015). Two research groups both affiliated with Harvard university had collaboratively come up with the idea of barcoding cells by encapsulating them together with solid primer carriers in microscopic droplets. The methods, aptly named inDrop and Drop-seq, both relied on encapsulating single cells in tiny water-in-oil droplets together with a microbead carrying the barcoded primers used to index every cell's mRNA content. The following section will briefly cover inDrop and Drop-seq and, as the experimental section of this thesis revolves around a hybrid version between the two, their differences will be examined closely.

Drop-seq

Drop-seq, Macosko et al.'s approach to droplet-based single-cell RNA sequencing, is based around co-encapsulating single cells with a barcoded resin bead on a microfluidic chip. Inside every droplet, the cell's mRNA content is captured by the bead's barcoded transcription primers. The ensuing reverse transcription reaction incorporates the primer's barcode and unique molecular identifier (UMI) in the cDNA, which allows researchers to demultiplex the cDNA by cell and mRNA transcript of origin. Drop-seq's bead production process, microfluidic chip and workflow are given in figures 1.12-1.14.

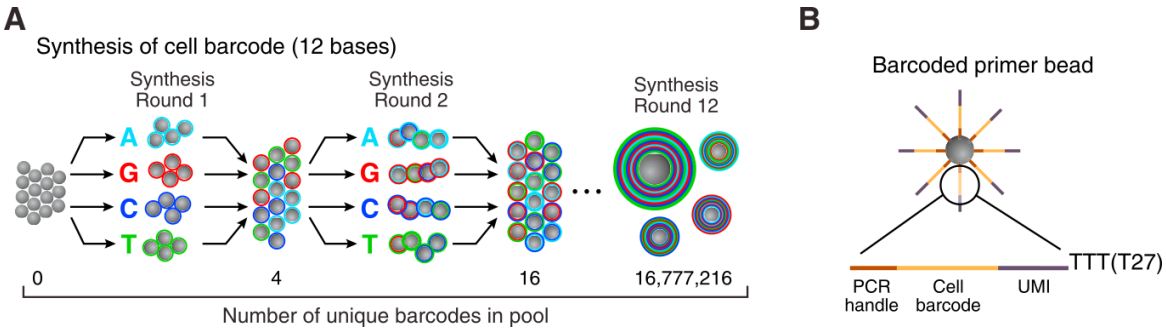


Figure 1.12: Drop-seq barcoded bead generation. (A) Resin beads undergo successive rounds of splitting, oligo synthesis and pooling to generate $2^{12} = 16\ 777\ 216$ possible uniquely barcoded beads. Then, eight steps of degenerate synthesis append an eight nucleotide UMI to the cell barcodes, followed by addition of a poly-T tail which can capture cellular mRNA. (B) Overview of the final bead barcode structure. Taken from Macosko et al. (2015).

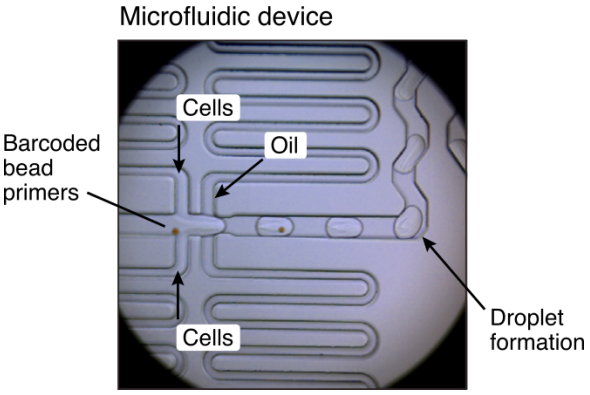


Figure 1.13: Drop-seq microfluidic chip. A flow of barcoded beads suspended in lysis buffer is joined by a single-cell suspension and emulsified into nanolitre droplets by an oil flow. Taken from Macosko et al. (2015).

Using Drop-seq, a single researcher can prepare thousands of single cell transcriptomes for sequencing in a single day, orders of magnitude faster than the 96-well plate assays discussed in section 1.2. In the original Drop-seq paper, Macosko et al. profiled 44 808 single-cell mouse retina transcriptomes in just 4 days, at a fraction of the cost of

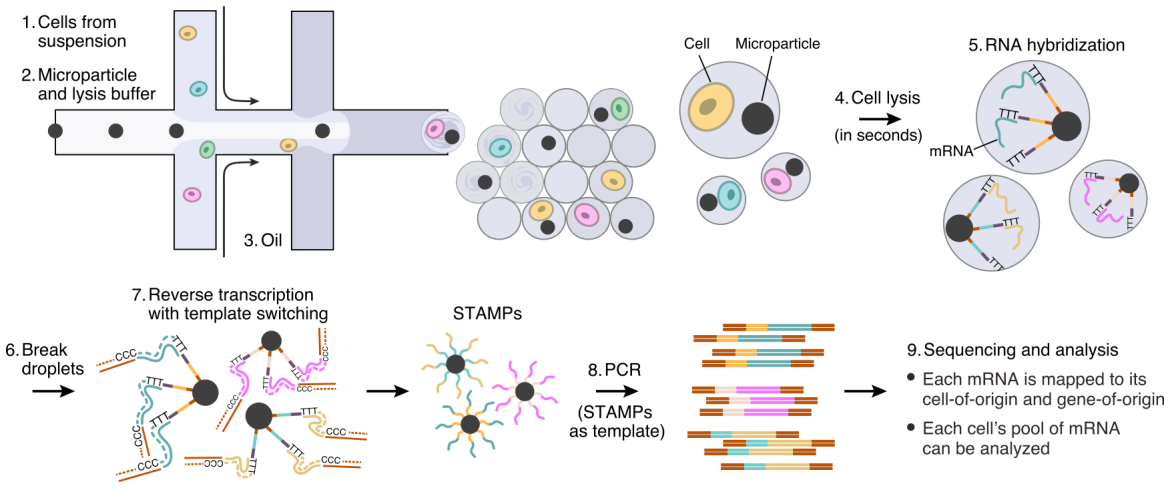


Figure 1.14: Drop-seq library generation. Single cells are encapsulated in a nanolitre volume droplet together with a barcoded resin bead and lysis buffer. The surface of each bead is covered in millions of copies of a bead-specific barcoded poly-dT primer which captures mRNA released by the cells in the droplets. The beads are then pooled in a large volume of water, which prevents further hybridisation outside of the droplet. Immediately after, the hybridised mRNA libraries are copied to the bead primers by template-switching reverse transcription, generating a library of single-cell transcriptomes attached to microparticles (STAMPs). The STAMP libraries are ISPCR-amplified, fragmented, PCR-amplified again and sequenced. The last PCR preferentially amplifies 3' fragments, leading to a 3'-enriched sequencing library. After sequencing, the cellular barcode is used to identify each transcript's cell of origin, and the UMI is used to count individual transcripts. Taken from Macosko et al. (2015).

conventional microwell-based due to the ultra low reaction volumes (~1 nl). Whereas the price of Smart-seq is estimated at $3 - $30 per cell before sequencing, a Drop-seq run indexes single cells at $0.1 - $0.65 per cell (Macosko et al., 2015; Ziegenhain et al., 2017). Drop-seq's incredible throughput potential is paired with a number of important drawbacks, a major one being reduced mRNA capture efficiency and sensitivity compared to well-based protocols. Using spiked-in RNA standards, Macosko et al. estimate that Drop-seq captures ~12% of the cellular mRNA and can detect an average of 44 295 mRNA transcripts from 6 722 genes in HEK cells. This is significantly lower than Smart-seq2, which captures ~20% of cellular mRNA and detects 12k genes in the same cell line (Picelli et al., 2013). It is important to note that both Drop-seq and Smart-seq use the same path of template-switching RT followed by ISPCR and subsequent Illumina Nextera XT sequencing library preparation. However, Drop-seq selectively PCR-amplifies 3' fragments since this is where the cellular barcode is located. Other fragments can simply not be assigned to a cell of origin and are therefore suppressed in the final sequencing library.

This means that Drop-seq can absolutely quantify transcripts by mapping the 3' ends to a reference exome, but cannot give information on distal transcript regions.

Another important limitation of Drop-seq is low cell capture efficiency. Since co-encapsulation of two independent particles is a double-Poissonian process, both cell and bead concentrations need to be kept low in order to minimise multiplet encapsulation. At the concentrations used by Macosko et al., the chance for a droplet to receive either a bead or a droplet is ~1-5%. This leads to a dual occupancy rate of only ~0.1% and an effective cell capture efficiency of just ~5% of the input sample. Macosko et al. suggest that Drop-seq's low capture efficiency can be compensated by simply brute-force processing a large number of cells. This is shown in the original publication, where rare cells constituting 0.1%-0.9% of the population are successfully characterised when thousands of cells are analysed.

inDrop

Published concurrently with Drop-seq, inDrop ("*ind*exing *drop*lets") is a different approach to the same concept. Again, single cells are encapsulated in nanolitre droplets together with reverse transcription reagents and beads carrying barcoded poly-dT primers (Klein et al., 2015; Žilionis et al., 2017). The inDrop bead production process, microfluidic chip and workflow are shown in figures 1.15-1.17.

Though both methods share a central concept, several important differences distinguish inDrop from Drop-seq. One of the most important differences can be found in the nature of the primer-loaded beads - whereas Drop-seq uses hard, 30 µm plastic beads, inDrop uses 70 µm soft, deformable hydrogels. These hydrogels are loaded into the microfluidic chip at concentrations ~100 times higher than Drop-seq's beads, allowing them to stack inside the chip's microfluidic channels. Near the end of the funnel, a single, lined-up file of beads is formed and pushed towards the cell flow. As the release of beads can be controlled directly, flows can now be tuned so that nearly 100% of droplets contain at least one bead (Klein et al., 2015; Abate et al., 2009). Due to this "super-Poissonian" stochastic bead loading, inDrop attains a cell-capture rate of near 95%, compared to Drop-seq's 5%. This allows inDrop to be applied to scenarios where the sample does *not* permit brute-force analysis of thousands of cells, for example in a clinical context.

Another important difference between Drop-seq and inDrop is the post-encapsulation library preparation. Whereas Drop-seq is strongly Smart-seq inspired, inDrop essentially follows the CEL-seq/MARS-seq (a high-throughput implementation of CEL-seq) library preparation process (Hashimshony et al., 2012; Jaitin et al., 2014). Like CEL-seq, the inDrop cDNA libraries are pooled, in-vitro transcribed for linear amplification and fragmented/PCR enriched for sequencing. Due to this overnight in-vitro transcription step, inDrop has a higher fixed time cost than Drop-seq. Using ERCC spike-ins, Klein et al. estimate inDrop's mRNA capture efficiency at 7.1% - higher than CEL-seq's 3.4%, but

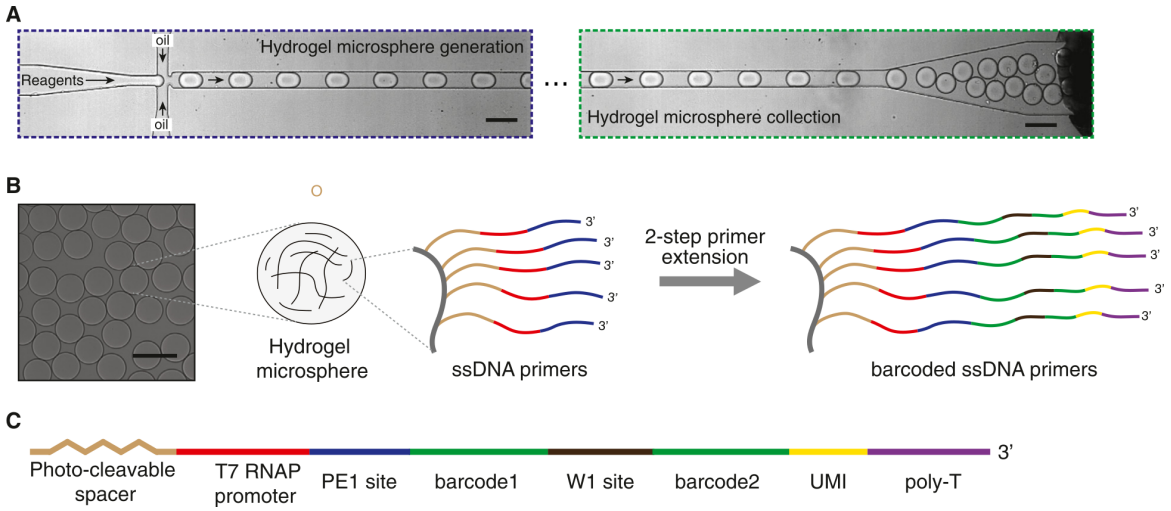


Figure 1.15: inDrop barcoded bead generation. (A) Millions of polyacrylamide microspheres are produced by droplet microfluidics and polymerised to hydrogel beads. (B) Short DNA primers in the acrylamide matrix serve as a PCR priming site and allow a photocleavable spacer, T7 RNA promoter to be appended. After two successive steps of splitting and pooling PCR in 384 wells, each hydrogel carries one of 147 456 (= 384 x 384) possible cellular barcodes. (C) The hydrogel primers also carry a UMI and a poly-T tail. Taken from Klein et al. (2015).

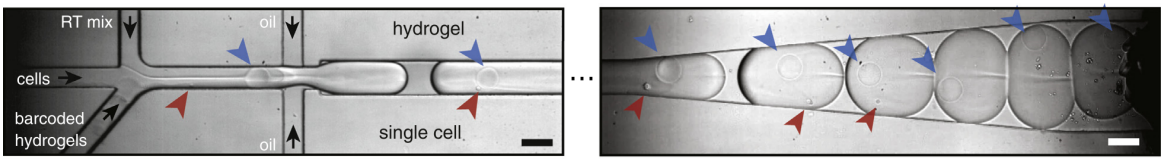


Figure 1.16: inDrop microfluidic chip. A dilute cell suspension is joined by a flow of reverse transcription reagents and stacked hydrogels, and further emulsified by flow-focusing. Taken from Klein et al. (2015).

lower than Smart-seq2 and CEL-seq2's ~20% (Klein et al., 2015; Grün et al., 2014; Picelli et al., 2013; Hashimshony et al., 2016). inDrop is thus plagued by the same low sensitivity issues as Drop-seq.

Droplet Microfluidic Techniques: Key Takeaway

As shown above, droplet-microfluidic single-cell techniques offer several advantages, most notably throughput. A single researcher can now process 10 000 single cells for sequencing in 12 hours, at a fraction of the cost of microwell-based assays. As the actual cell encapsulation step takes only minutes, throughput can be increased further by simply

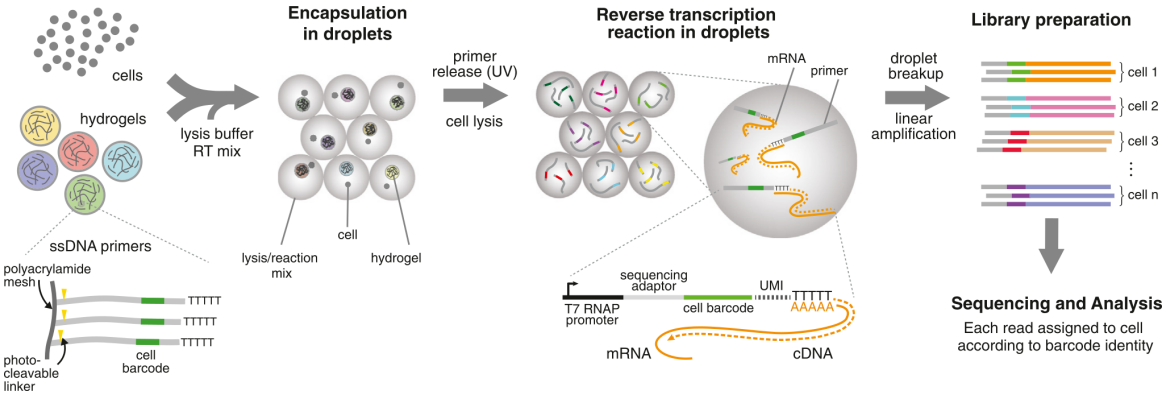


Figure 1.17: inDrop library generation. Similarly to Drop-seq, single cells are encapsulated in a nanolitre volume droplet together with a barcoded hydrogel bead. Each hydrogel bead carries a payload of bead-specific barcoded poly-dT primers covalently incorporated in the gel matrix. These primers are released from the hydrogel using a UV-cleavable chemical linker. The cell is lysed and its mRNA content is reverse transcribed with the released hydrogel primers in the droplet. The resulting barcoded cDNA fragments are pooled and linearly amplified using IVT, similar to CEL-seq. After sequencing, each transcript's cellular barcode and UMI are used to identify cell and transcript of origin. Taken from Klein et al. (2015).

encapsulating more cells and processing their emulsions together. The ultra-low reaction volumes (nl) associated with microfluidics strongly reduce cost, which can be further minimised when cell/bead dual occupancy are optimised. Moreover, these low reactions volumes have been shown to reduce technical noise and increase product yield (Streets et al., 2014). However, despite this supposed increase in reaction efficiency, all droplet microfluidic single-cell techniques exhibit significantly lower sensitivity when compared to their microwell counterparts. Two considerations need to be made regarding this observation. First, sequencing depth, a deciding factor in the data quality of single-cell experiments, is usually much lower in high-throughput methods due to the high number of cells processed per sample. It is often difficult or prohibitively costly to sequence thousands of single-cell transcriptomes from the same tissue sample to saturation. Hopefully, the ongoing decrease of NGS price will mitigate this limitation in the future. Second, Drop-seq and inDrop are often compared with newer generation of microwell methods such as Smart-seq2 and CEL-seq2. Over the years, these methods have been heavily optimised in terms of reagent concentrations, enzymes, incubation times and so on. Several efforts have already been made to improve the sensitivity and capture efficiency of droplet microfluidic methods. We expect that in the coming years, droplet microfluidic single-cell techniques will undergo a similar process of optimisation.

So far, the microfluidics involved in droplet-based single-cell technologies have been very simple, usually consisting of just one round of encapsulation. Žilionis et al. remark

that microfluidic modalities such as droplet splitting, merging and sorting or even multiple encapsulation could lead to droplet versions of more complicated protocols such as ATAC-seq, ChIP-seq, or a combination of pre-existing modalities (Žilionis et al., 2017; Ahn et al., 2006a,b).

1.5 In-Situ Cellular Indexing by Splitting & Pooling

So far, every technique shown has depended on molecular reactions performed on physically separated individual cells. Split-pool methods rely on an entirely different concept - here, groups of cells are in-situ barcoded together in successive rounds of redistributing/splitting and pooling. In these approaches, the cells are never truly individually separated from each other - the nucleic acid content is contained within in the cell itself. Split/pool approaches pose specific advantages and risks, both which will be explained below.

Combinatorial Indexing

In 2015, Shendure Lab produced an interesting cell barcoding strategy which they dubbed single-cell combinatorial indexing (sci) (Cusanovich et al., 2015). In sci-based protocols, a large number of cells is split into microwells where their nucleic acid content is barcoded using in-situ PCR. Then, the cells are pooled and redistributed into new wells where a second barcode is appended to the first. If a sufficiently high number of wells is used in each step, the majority of cells pass through a unique combination of wells, resulting in a unique set of two barcodes nucleic acid content. A visual overview is shown in figure 1.18.

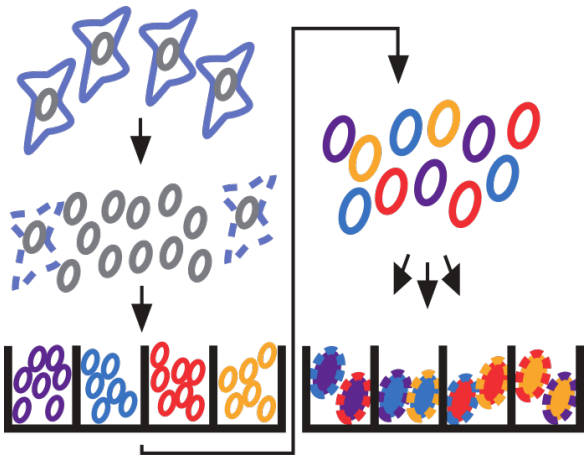


Figure 1.18: Single-cell combinatorial indexing. A suspension of nuclei is equally distributed into wells where they undergo a well-specific barcoding reaction. The nuclei are then pooled and redistributed into a second array of wells, where a second barcode is appended to the first. The majority of nuclei undergo a unique barcoding trajectory, each receiving a unique cellular index that is later used to demultiplex the resulting libraries after sequencing. Adapted from Cusanovich et al. (2015)

The sci approach was first used in Cusanovich et al.'s single-cell combinatorial indexing ATAC-seq (sci-ATAC-seq) method (Cusanovich et al., 2015). In sci-ATAC-seq, a population of ~2000 nuclei is distributed equally into 96 microwells. Here, the nuclei undergo tagmentation with a well-specific custom Tn5 transposase which carries a well-specific barcode. This step results in a chromatin library of which the fragments are tagged with

a first barcode. The tagmented nuclei are then pooled and redistributed into 96 new wells where they undergo PCR amplification with a barcoded primer, resulting in single-cell ATAC-seq libraries. The protocol was later adapted for to sci-RNA-seq (Cao et al., 2017) by performing reverse transcription with well-specific barcodes instead of well-specific Tn5 tagmentation. sci-RNA-seq starts from either single nuclei or fixed and permeabilised cells. The sci approach can thus be used for both scATAC-seq an scRNA-seq.

As with the high-throughput droplet-microfluidic methods described in section 1.4, the sensitivity of sci-based protocols is sub-par. Fiers et al. remark that sci-ATAC-seq retrieves ~2500 chromatin fragment reads per cell compared to ~73 000 reads per cell from the Buenrostro et al. scATAC-seq protocol (see section 1.3) (Fiers et al., 2018). This difference can in part be explained by the difference in per-cell sequencing coverage between the two methods.

A major single-cell multiomics breakthrough was made in 2018 when Cao et al. extracted both RNA-seq and ATAC-seq data from the same nucleus using a sci-based approach (Cao et al., 2018). The method, called single-cell combinatorial indexing for chromatin accesibility and RNA (sci-CAR), simultaneously barcodes the RNA and chromatin content of single nuclei before splitting the sample for both sci-ATAC-seq and sci-RNA-seq processing. Since the cDNA and chromatin reads of the same cell share the same barcode, both libraries can be assigned to their cell of origin. An overview is given in figure 1.19.

Sci-CAR was applied to extract joint transcriptomic and chromatin accessibility data from 4 825 and 11 296 nuclei, a major increase from other multiomics protocols by e.g. Hou et al. (2016) and Clark et al. (2018) respectively, which could only be applied to fewer than 100 cells. To date, sci-CAR remains the only protocol able to simultaneously profile the transcriptome and chromatin accessibility of high numbers of single cells. Despite the intrinsic value of such datasets in the study of for example gene regulatory networks, sci-CAR has not been applied yet by other labs. A hindrance to widespread adoption is the requirement for a custom Tn5 transposase, which is not available commercially. Importantly, sci-CAR also discards half of the available cellular material. Sci-approaches generally require a large number of cells to initiate (10^5 to 10^6 cells), and also retrieve about 7-50% of the sample input, with a high incidence of doublets. A final drawback is that cells/nuclei need to be fixed for sci, prohibiting the use of fresh samples. Ideally, a multiomic method would barcode and sequence the whole transcriptome and chromatin accessibility profile of cells completely separately. Such a protocol has not been published yet.

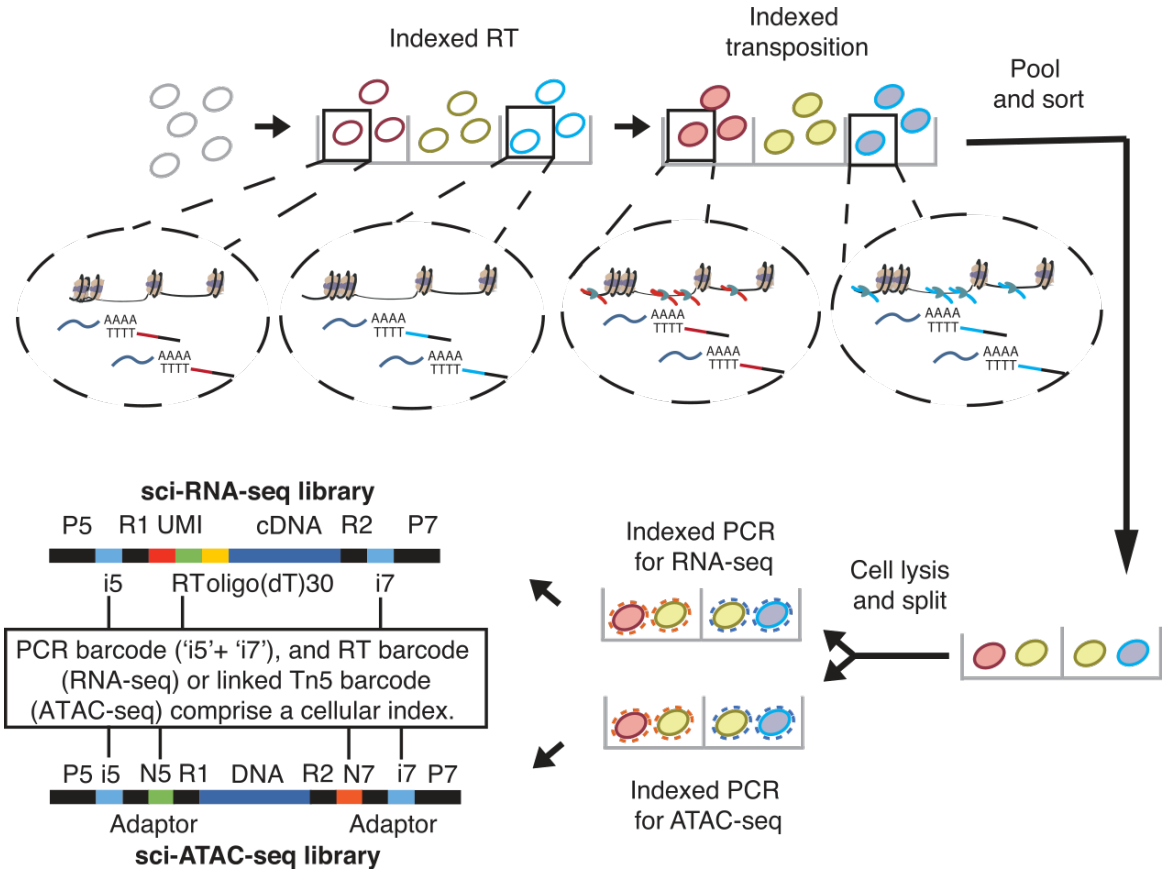


Figure 1.19: Sci-CAR concept. Single nuclei are distributed into a 96 well plate at 5 000 nuclei per well, where their RNA content is barcoded by a reverse transcription reaction with a well-specific primer. Next, the nuclei's chromatin content is fragmented and barcoded in the same well by a custom Tn5 carrying a well-specific adapter. In a crucial step, nuclei are first pooled and then FACS-sorted into 576 new wells where the second cDNA strand synthesis occurs. The nuclei are then lysed and the sample is split equally (without pooling) into dedicated portions for subsequent sci-RNA-seq or sci-ATAC-seq library generation. Both the RNA-seq and ATAC-seq reads can be assigned to the cell of origin based on their barcode combination. Taken from Cao et al. (2018)

SPLiT-seq

In 2018, Rosenberg et al. published split-pool ligation-based transcriptome sequencing (SPLiT-seq), a scRNA-seq technique conceptually identical to the Shendure Lab's single-cell combinatorial indexing scRNA-seq (Cusanovich et al., 2015; Rosenberg et al., 2018). Like sci-RNA-seq, SPLiT-seq indexes fixed cells using successive split/pool in-situ reverse transcription and PCR. However, the addition of two barcode ligation steps leads to four barcoding opportunities per cell, as opposed to sci-RNA-seq's two. This increases overall hands-on time, but allows SPLiT-seq to scale to large numbers of cells more rapidly than sci-RNA-seq. An overview of the SPLiT-seq barcoding process is given in figure 1.20.

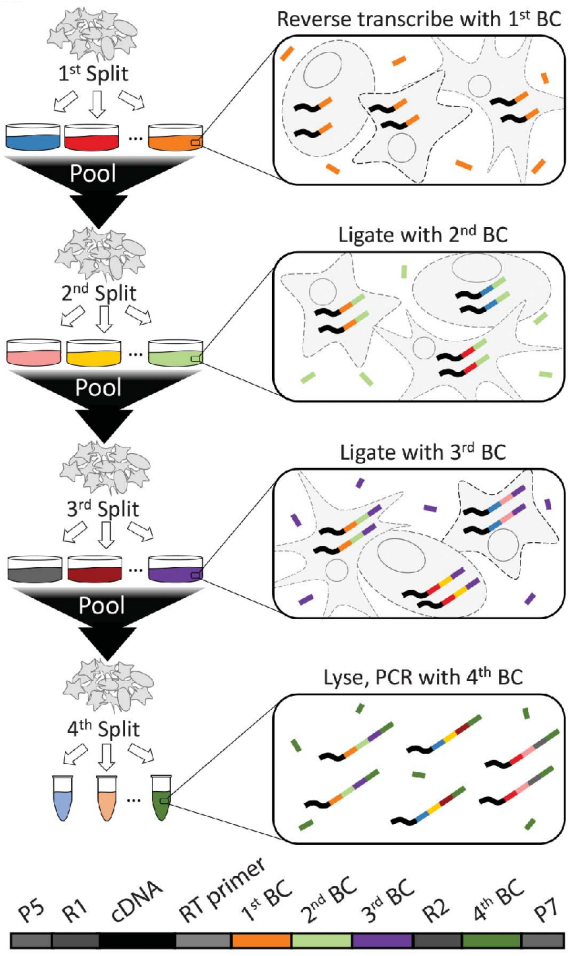


Figure 1.20: Split-pool ligation-based transcriptome sequencing (SPLiT-seq). Methanol-fixed and permeabilised cells are split into 48 wells where they undergo in-situ reverse transcription with a well-specific barcode. Two additional barcodes are in-situ ligated to the cellular cDNA libraries in subsequent splitting and pooling steps. The cells are then lysed and a final barcode is introduced in the sequencing library indexing PCR. Taken from Rosenberg et al. (2018).

SPLiT-seq's main advantage is unprecedented throughput. In theory, SPLiT-seq can barcode more than 6 million cells using just four different microwell plates. It is not unthinkable that future analysis of large cell samples will necessitate experiments of this magnitude. Today, however, the current cost of sequencing and modern computational possibilities prohibit the execution of such ultra-high throughput experiments.

In-Situ Cellular Indexing: Key Takeaway

The ultra-high throughput capabilities of combinatorial indexing approaches hold promises for future experiments where 10^4-10^5 cells are analysed. Due to the limits on computation and sequencing technology, such dazzling numbers are unthinkable today. In a sense, the main limiting factor to contemporary single-cell research is by no means the number of cells analysed in a single run. Rather, high-throughput (10^3-10^4 cells), high-sensitivity assays seem to be a more important goal. It is for this reason, combined with the low sensitivity associated with combinatorial indexing and the use of proprietary reagents, that sci and SPLiT-seq have not been extensively applied outside of their respective labs yet.

1.6 Applications of Single-Cell Omics

Cell Typing

One of the major applications of single-cell RNA-seq is de-novo cell typing. The *Tabula Muris* is an example of such a large scale effort. The Tabula Muris Consortium sequenced the individual transcriptomes of 55k mouse cells spread over all major organs using 10x Chromium scRNA-seq, and another 45k cells using FACS + Smart-seq2. The resulting cells were computationally separated (figure 1.21) and analysed, identifying several distinct new cell types and transcription factor networks, and revealing previously undiscovered roles of known genes.

Analysing such a large amount of cells spanning over multiple organs allows for direct comparison of the resulting data, but is a major undertaking using the current state of the art technology. The tiered approach employed here, where low sequencing coverage droplet microfluidic methods are used to rapidly form a global picture, and more plate-based methods are used to analyse carefully filtered pre-defined populations at higher sequencing depth, will most likely be the key to undertaking large scale sequencing operations such as these in the future.

Development

During embryogenesis, a single totipotent cell will divide, and its descendants will gradually lose potency and differentiate into the cell types that make up the organism. The acquisition of cell identity, function and morphology is for a large part controlled through differential gene expression (Farrell et al., 2018), making high-throughput scRNA-seq a valuable tool in our effort to understand this complex process.

Farrell et al. analysed the individual transcriptomes of 39k zebrafish cells from 12 different embryonic stages using Drop-seq and developed novel computational methods

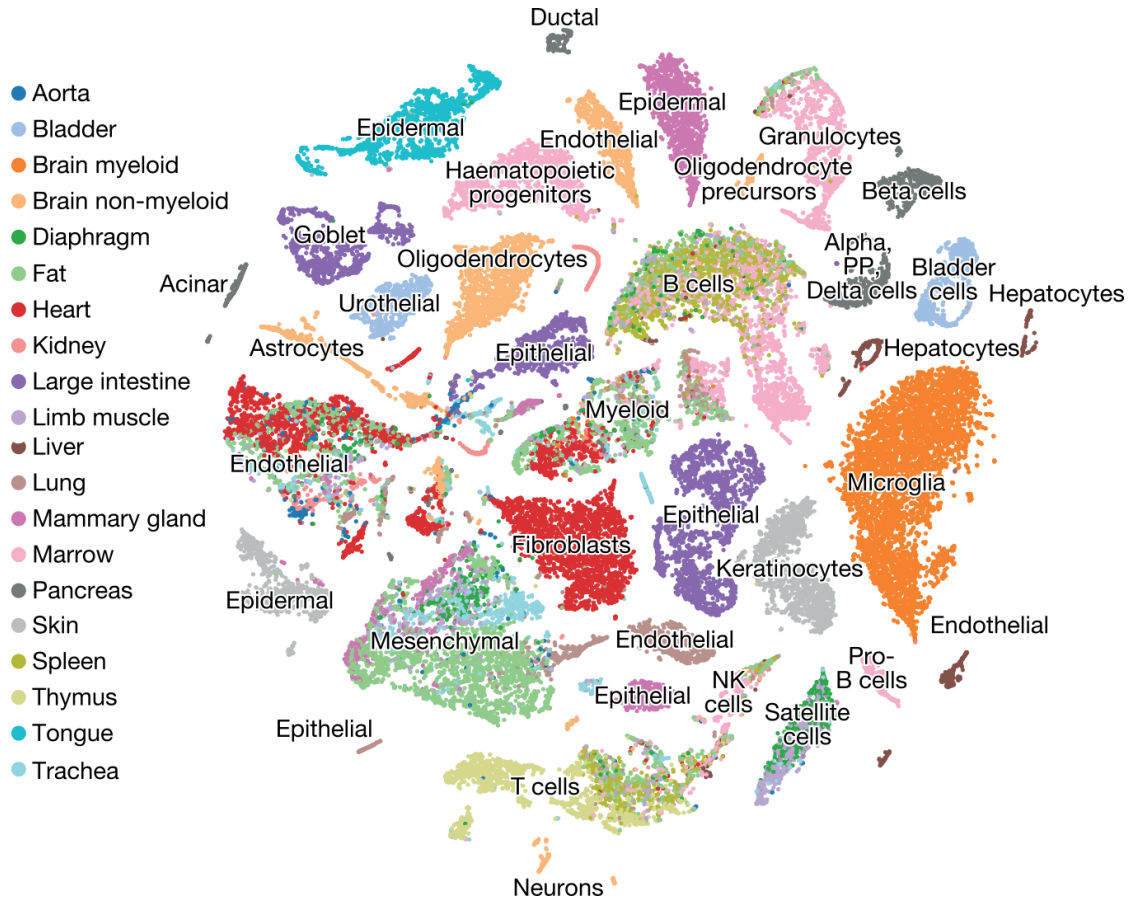


Figure 1.21: Murine cell atlas. tSNE visualisation of 45k single cell transcriptomes from 20 mouse organs. Taken from Tabula Muris Consortium (2018)

(a combination of URD and Seurat) to map the differentiation process of 25 cell lineages during embryonic development (figure 1.22). The tips of the resulting tree corresponded to previously known cell types in terms of marker gene expression, and much of what was already known about embryonic development was reflected in the tree's branching structure.

In addition to agreement with the canonical embryological knowledge, the systems biology approach revealed new candidate regulators of the differentiation process and how the spatial organisation of the developing embryo may be decided earlier in development than previously thought. Importantly, the computational framework developed by Farrell et al. can be used for the reconstruction of the developmental trajectories of other biological model organisms.

Another new computational tool was pioneered by La Manno et al. in 2018, who realised that the time derivative of gene expression profiles across samples taken on multiple time points could be inferred from the identity of unspliced and mature mRNA transcripts.

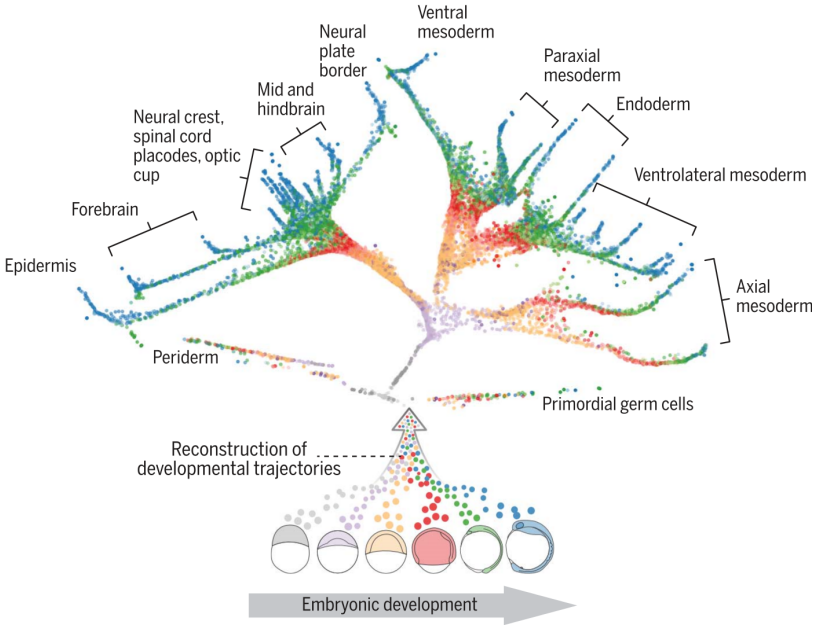


Figure 1.22: Developmental trajectories in Zebrafish. Taken from Farrell et al. (2018)

They found that in several open-access single-cell RNA-seq datasets from samples taken at different time points, the population of immature mRNA transcripts was present as mature transcripts at the next time point. La Manno et al. then used this information to project the developmental flow on regular tSNE plots (figure 1.23).

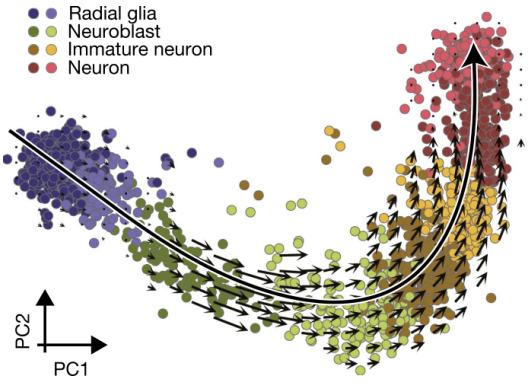


Figure 1.23: RNA velocity in human neurogenesis. Taken from La Manno et al. (2018)

As this method was first performed on previously published data, this publication proves that new information can often be extracted from the datasets generated by single-cell sequencing experiments when approaching them from a different angle. When a sequencing dataset is generated, it exists permanently as a vector space in which future in-silico "experiments" can be performed, as opposed to classical experiments where samples may have a limited lifetime and interrogating the same sample at a later time point may be difficult. In the future, we may thus see more examples where discoveries are made in older datasets thought to be exhausted of information.

Disease

The final application of single cell technology, and its main accelerator today, is human medicine. Several areas of medicine will directly benefit from single-cell resolution data, notably those where cell diversity is most impactful - such as cancer and brain disease. Today's leading effort in pushing single-cell technology to applications in medicine is the human cell atlas (HCA) (Regev et al., 2017). This project, which aims to approach the Human Genome Project in magnitude and scope, aims to comprehensively map all cells present in the human body. Such an atlas could be used as a reference for patient sample comparisons and help understand the cellular mechanisms behind disease. Figure 1.24 shows the key impact aims of the HCA.

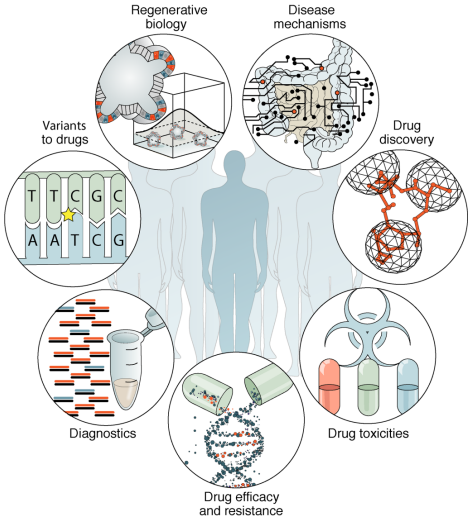


Figure 1.24: Impact areas of the human cell atlas. Taken from The Human Cell Atlas Consortium (2017)

A major application of the HCA will be targeted drug discovery. Comparing the genetic profile of healthy cell samples to the reference cell atlas would provide leads for possible new drugs. Single-cell experiments could also be used to compare in-vitro generated cell cultures with the reference profile of the human tissue they attempt to mimic, accelerating the development of engineered tissues for regenerative medicine. The project's first draft, published in 2017, profiles a subset of cells from the tissues and organs that hold the most promise for immediate application. The first draft includes 'only' 30 to 100 million cells, a fraction of the projected 10 billion for the final atlas (The Human Cell Atlas Consortium, 2017).

1.7 Single-cell Omics: Current Progress and Future Perspectives

As shown in the previous sections, the past few years have seen an explosion in single-cell research and technology. Single-cell transcriptomics, genomics and epigenomics are already a fact, and proteomics and spatial omics are looming over the horizon. Single cell technology has allowed us to capture snapshots of complex cellular processes such as gene regulation, gene expression, tissue development, and the origins of disease. Even in its infancy, single-cell technology has proven its value in almost every area of biology involving cells. We have uncovered regulatory pathways and mapped the (partial) transcriptome atlases of several model organisms, and are rapidly moving forward to achieving the same in humans. However, much remains to be done. It is painstakingly clear that the vast majority of today's single-cell technology can be further improved upon. Our most cutting-edge high-throughput techniques are able to capture only a small fraction of the information embedded in a single cell, with low reproducibility and high noise levels. Indeed, we are asking much of our bioinformatician colleagues. Equally concerning is how virtually every single-cell technique to date can only interrogate a single omics modality. It is therefore difficult to, for example, systematically relate a cell's transcriptome to its epigenome. The clear-cut next step is therefore to fine-tune existing methods and to move forward to single-cell multi-omics. Extracting information on several omics modalities from the same single cells, at high fidelity *and* close spatial resolution will help us understand the absolute fundamentals of life - and how we may improve it.

2 | Optimising the inDrop Single-Cell RNA-seq Platform

2.1 Redesigning inDrop

As shown in chapter 1, droplet microfluidics have been applied extensively in high-throughput single-cell technology. Single cells are co-encapsulated with various reagents and barcoded hydrogel beads in order to capture their nucleic acid content (figure 2.1). The first part of this thesis revolves around the optimisation of inDrop, a leading droplet microfluidic single-cell RNA-seq technique (Klein et al., 2015). We implemented a number of changes to the original inDrop protocol, most of which were inspired by other droplet microfluidic protocols such as Drop-seq and 10x Genomics' commercial scRNA-seq solution. First, we physically altered the microfluidic protocol to improve quality of life and ease of operation. Next, we changed the reverse transcription reaction and library preparation steps to a Smart-seq approach. An overview of the major changes is given in figure 2.2.

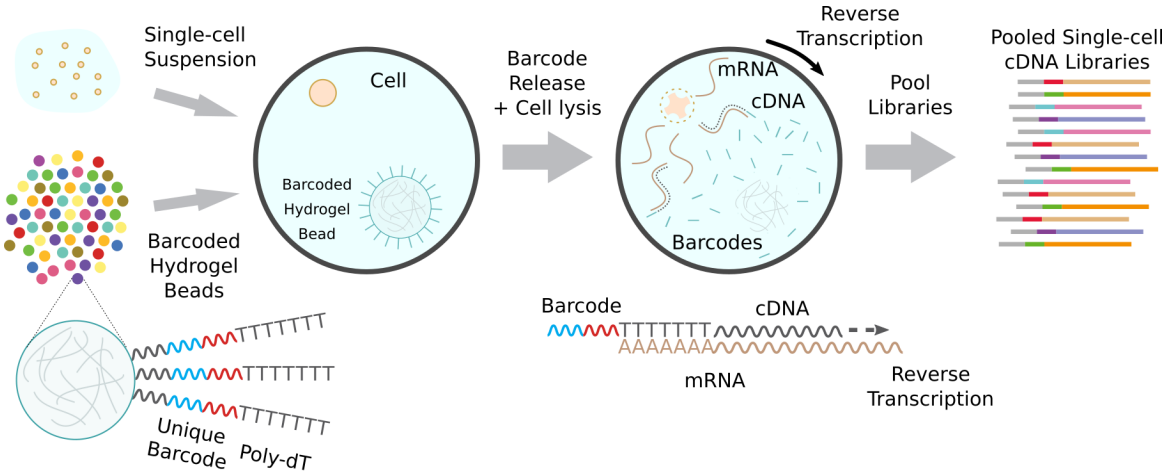


Figure 2.1: General overview of an inDrop experiment. Single cells are encapsulated into a nanolitre volume droplet together with a barcoded hydrogel bead. The cell is lysed, and its mRNA content is tagged in a reverse transcription reaction using the hydrogel's barcoded primers. The cDNA libraries of thousands of cells are then pooled and processed for sequencing.

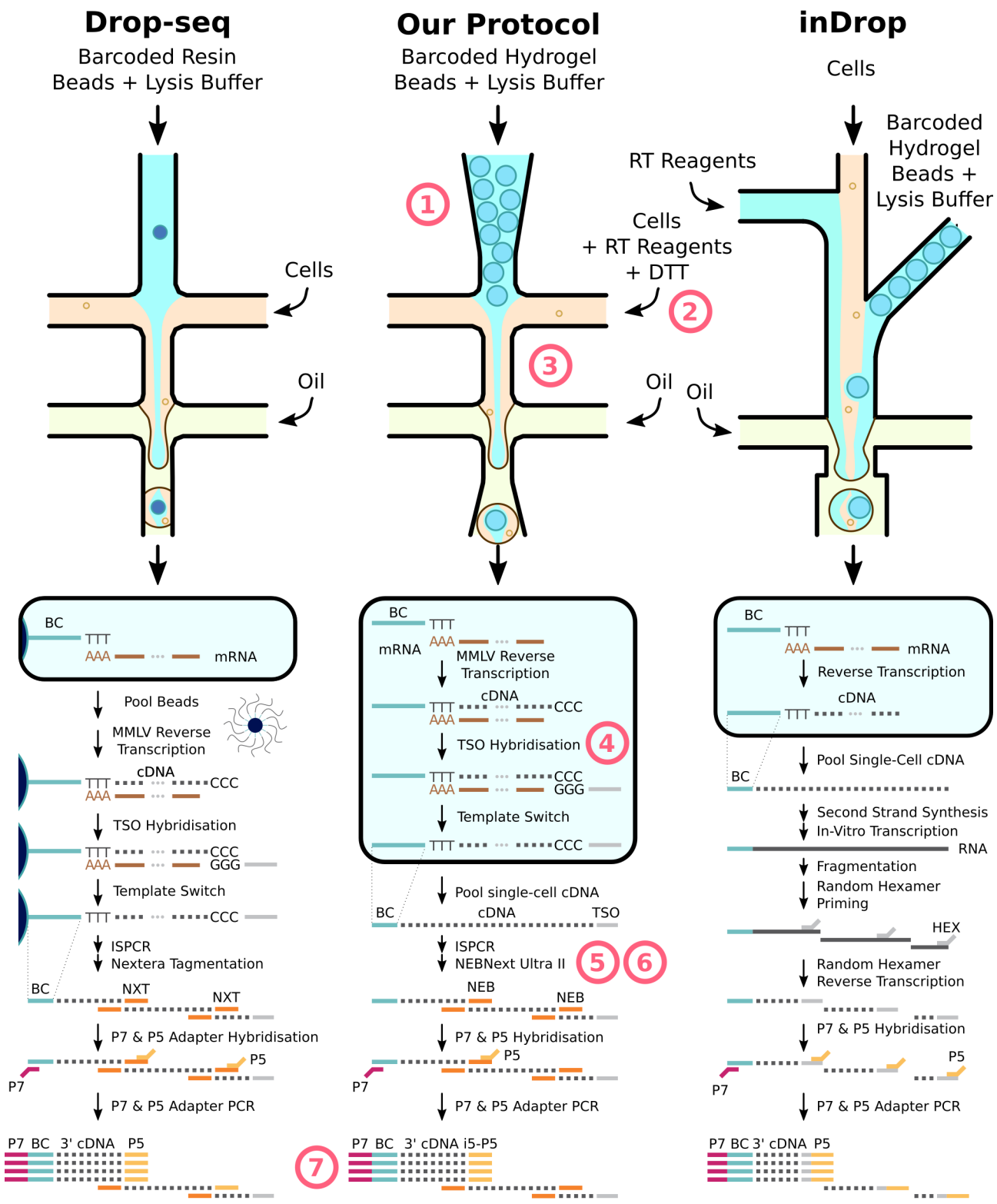


Figure 2.2: Molecular changes to the inDrop protocol. Red markers indicate major differences or improvements.

1. We retained inDrop's soft, deformable hydrogel beads in order to keep the super-Poissonian bead loading, but scaled them down to a smaller size (50 µm instead of 70 µm). With this change in size, we aimed to produce beads compatible with the 10x Chromium microfluidic chips.

2. inDrop beads release their primers into the cell droplet by cleavage of a UV-sensitive linker. We replaced this linker with a disulfide bond which can be cleaved by DTT-mediated reduction in the droplet. This change greatly improves handling of beads during all steps of the process as they do not need to be shielded from ambient light any more.

3. inDrop employs a system of 3 input flows - one each for cells, hydrogel beads and RT/lysis reagents. While such a setup allows for fine control over all flows, it complicates microfluidic operation. We adopted Drop-seq's microfluidic design by merging lysis buffer flow with bead flow and RT reagent flow with cell flow.

4. Like Drop-seq, we opt for a Smart-seq-like template-switching reverse transcription which incorporates a template-switching oligo (TSO) at the 3' end of the cDNA transcripts. In our case, the TSO bears a sequence complementary to the 5' end of the barcode. Template switching thus induces end-complementarity into the transcript which will play an important role in step 5. The template-switching reverse-transcription reaction takes place *in* the droplet, as opposed to Drop-seq's pooled single-cell transcriptome attached to microparticle (STAMP) approach.

5. We bypass inDrop's time-consuming in-vitro transcription based amplification protocol by using inverse suppressive PCR (ISPCR). ISPCR employs a single primer which hybridises to both ends of the cDNA transcript. Short fragments form hairpins, obstructing primer hybridisation and thus amplification, reducing classical PCR bias towards shorter fragments.

6. We replace inDrop's random hexamer reverse transcription library prep with the off-the-shelf NEBNext ligation protocol. This approach is very similar to Drop-seq's classic Nextera library preparation, which uses Tn5 tagmentation and adapter PCR. NEB and NXT indicate the NEB and Nextera adapters.

7. Whereas inDrop sequences the cell barcode in one dedicated sequencing read on the Illumina NGS platforms, our cell barcode is read in two separate reads. We thereby avoid sequencing the fixed sequence in-between both barcode halves and dedicate these sequencing cycles to the cDNA instead. This will be detailed in chapter 4.

The result of these changes is a more streamlined protocol for high-throughput droplet-microfluidic single-cell RNA-seq.

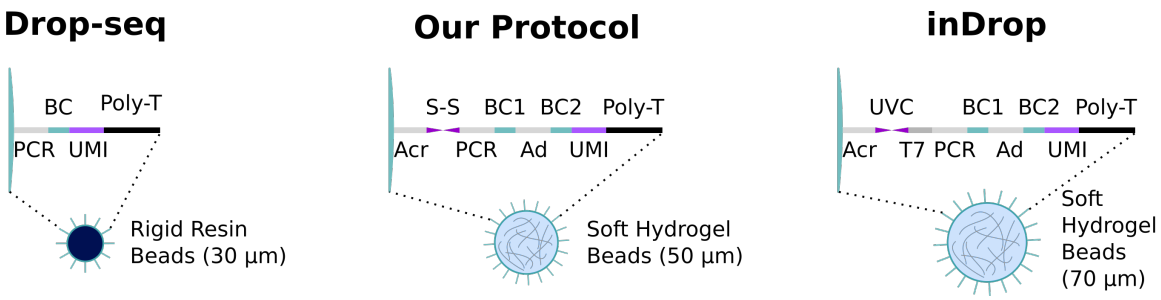


Figure 2.3: Hydrogel bead barcode structure. For Drop-seq - PCR: PCR primer site, BC: barcode, UMI: unique molecular identifier. For our protocol and inDrop - Acr: Acrydite linker, S-S: disulfide bond PCR: priming site for isothermal amplification during barcoding, BC1 and BC2: both halves of the barcode used in split-pool barcode generation, Ad: PCR adapter used in split-pool barcode generation, T7: T7 promoter used in in-vitro transcription. UVC: inDrop's UV-Cleavable spacer.

Methodology and Work Plan

We systematically set up and streamlined our inDrop workflow according to a set plan illustrated in figure 2.4. First, we produced a standardised stock of barcoded hydrogel beads to use during all following experiments. We then performed a number of dry inDrop runs to find the right flow rates for optimal bead and cell loading. Next, we progressed onto "true" inDrop runs with high concentrations of cells. These runs were progressed up to the library preparation stage. We performed several intermediate quality control steps on these sequencing libraries, such as DNA quantification and fragment length distribution analysis, and made adequate changes to the library preparation protocols. Finally, we performed a complete inDrop run that was sequenced on the Illumina NextSeq 500 platform. The resulting data is shown and analysed in chapter 4.

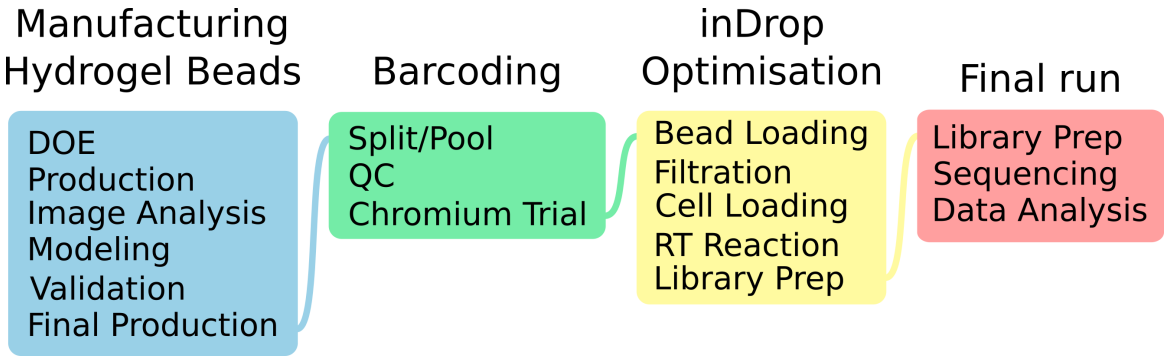


Figure 2.4: inDrop workplan.

2.2 Manufacturing Barcoded Hydrogel Beads

Both the scRNA-seq method described this chapter and the scATAC-seq method of chapter 3 will make extensive use of hydrogel beads. These hydrogel spheres serve as carriers for the barcoded oligonucleotide payload used to index a cell's nucleic acid material. Because inDrop hydrogels are soft and deformable, they can be loaded into microfluidic channel at high concentrations, leading to a stacking effect that allows for near-deterministic control of hydrogel flow. Such a highly-controllable flow essentially removes a large amount of stochasticity from the cell encapsulation process, leading to more efficient sample loading and lower cost (Klein et al., 2015; Abate et al., 2009). This section describes the general workflow for the production of these hydrogel beads, as well as a deterministic model for estimating the microfluidic parameters required for a given hydrogel diameter.

Hydrogel Bead Production

First, hydrogel beads carrying short DNA stubs are produced using standard polyacrylamide chemistry adapted for microdroplet generation on a microfluidic chip (Žilionis et al., 2017). These chips are produced according to standard protocols described in M.1.2. A flow of mixed monomer/crosslinker/acrydite oligo is emulsified into monodisperse droplets in fluorinated oil. The fluorinated oil phase contains TEMED, which catalyses the polymerisation reaction by accelerating free radical formation from APS. The resulting emulsion is incubated overnight at 65 °C during which the individual monomer droplets polymerise into spherical hydrogels. After extensively washing to remove all traces of apolar solvents and unused reagents, the hydrogel beads are ready for barcoding. The detailed protocol is described in M.1.3. Figure 2.5 shows the dimensions of the microfluidic chip and the polymerisation reaction involved in the process.

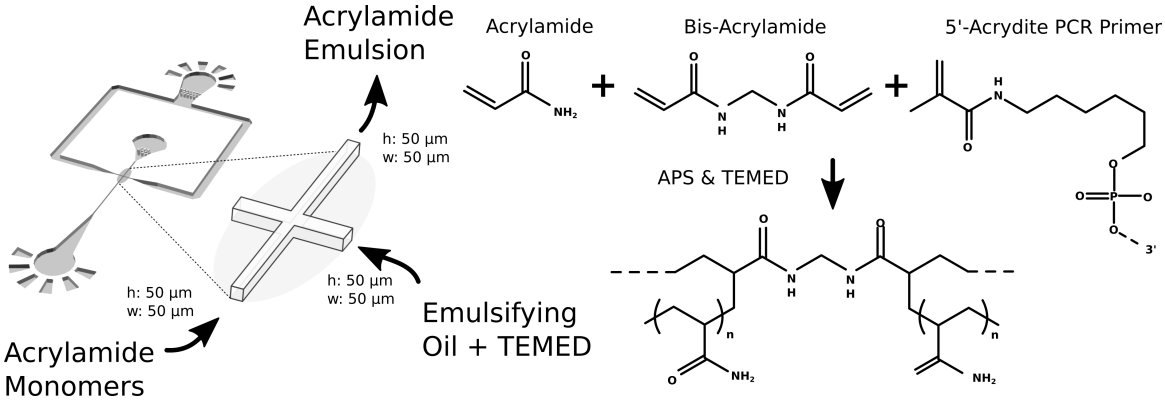


Figure 2.5: Hydrogel bead generation chip and chemistry.

Modelling Hydrogel Bead Diameter

The diameter of the hydrogel beads can be tuned within a range of 30-80 µm by adjusting the flow rates of oil (Q_{oil}) and monomer/crosslinker (Q_{oil}) during droplet generation. The size of the hydrogel beads directly determines the amount of barcoded primer they can carry, which impacts the efficiency of reverse transcription (for inDrop) or PCR (for drop-ATAC). Moreover, loading only hydrogels of controlled size and shape into the microfluidic device will facilitate microfluidic operation, preventing non-linear behaviour such as shockwaves and oscillations. Having a monodisperse distribution of diameter size is thus crucial for our application. We chose a set bead diameter of 50 µm, which is similar in size to 10x's hydrogel beads. Having a similar bead diameter would allow us to benchmark the beads without accounting for variability in microfluidic operation. Additionally, such beads could be used to develop new protocols for the 10x Chromium, which is more user-friendly than in-house microfluidic setups.

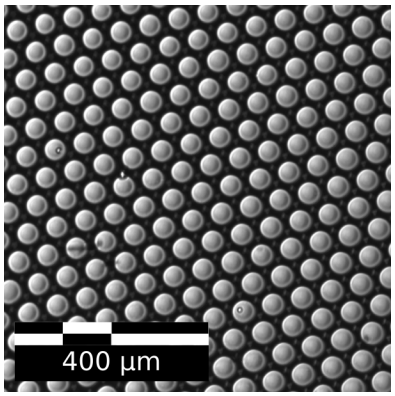


Figure 2.6: Hydrogel emulsion before polymerisation.

As the diameter of the beads increases by up to 25% during polymerisation and subsequent washing steps, predicting bead diameter based on emulsion droplet diameter is difficult. In order to find the flow parameter settings that would produce monodisperse beads of 50 µm, we formulated a model that can predict the diameter of the final beads based on the flow rates used during bead generation. We used an experimental design approach to maximise the amount of information we could extract from the labour and costs associated with producing the hydrogel beads. We designed a blocked response surface experiment with 2 blocks of 4 combinations according to the methodology described in Goos and Jones (2011). We proposed a quadratic model as they are a relatively simple class of models and often sufficiently explain variation in a process. Observations were blocked by the microfluidic chip used to account for variation between individual chips. During the experiments, it became apparent that 2 of the 8 combinations did not produce any droplets, but instead resulted in laminar oil and monomer co-flows. This occurred when the ratio of monomer to oil flow was greater than ~1.5. We therefore augmented the initial design with two additional blocks of each 3 runs constrained to oil/monomer $>$ 1.5, for a total of 12 samples (M.1.3).

In order rapidly and efficiently measure bead diameter, a Python image analysis script was written which could batch analyse a number of microscopy images from the beads. The script uses the `opencv` library to fit circular objects within set distances from each other, and automatically returns their diameters and locations within the frame. Figure 2.7 shows an example of the script's accuracy. The script itself can be found in S.1.1.

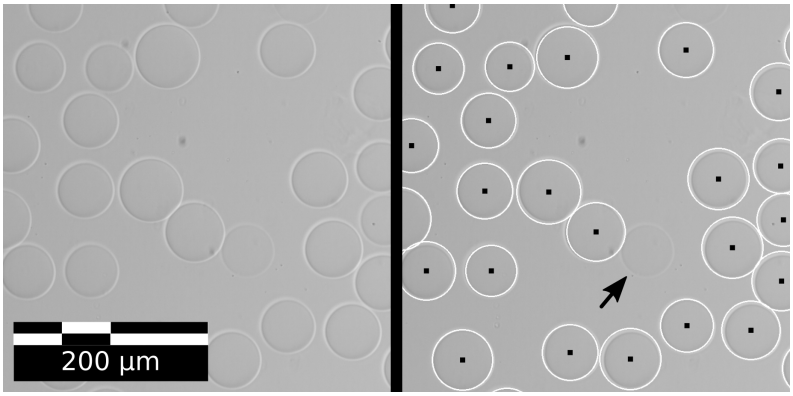


Figure 2.7: Bead diameter script example. Arrow indicates a bead that went undetected by the image analysis script.

We then produced the 12 batches of beads according to the parameters dictated by our experimental design and measured them after all washing steps were completed. The resulting dataset comprised ~8000 beads unequally divided over the 12 parameter combinations, shown in supplementary figure S.1 (sample 16 was omitted, as it was a statistical outlier and located at the edge of what physically resulted in beads). Based on this dataset, two separate models were estimated - one relating flow rates and bead diameter (eq. 2.1), and the second relating flow rates and the *variation on* bead diameter (eq. 2.2). As bead diameter was (by design) heteroscedastic over the parameter space and our dataset was heavily imbalanced, we used restricted maximum likelihood (REML) to estimate the diameter model parameters (M.1.4). We then used mixed stepwise regression control to iteratively select significant effects ($p < 0.05$). Heredity was invoked for Q_{oil} as it occurs in significant interaction effects. Units are µm and µl/h for diameter and flow respectively:

$$\begin{aligned} d = 53.6\,\mu m &- 12.6 \times 10^{-3}\frac{\mu m\,h}{\mu l} \times Q_{mono} - 569 \times 10^{-3}\frac{\mu m\,h}{\mu l} \times Q_{oil} \\ &- 3.43 \times 10^{-6}\frac{\mu m\,h^2}{\mu l^2} \times Q_{mono} \times Q_{oil} \\ &+ 11.1 \times 10^{-6}\frac{\mu m\,h^2}{\mu l^2} \times Q_{mono}^2 - 612 \times 10^{-9}\frac{\mu m\,h^2}{\mu l^2} \times Q_{oil}^2 \end{aligned} \quad [\mu m] \quad (2.1)$$

$$\begin{aligned} \sigma_d = 3.84\,\mu m &+ 996 \times 10^{-9}\frac{\mu m\,h}{\mu l} \times Q_{mono} - 1.62 \times 10^{-3}\frac{\mu m\,h}{\mu l} \times Q_{oil} \\ &- 256 \times 10^{-9}\frac{\mu m\,h^2}{\mu l^2} \times Q_{mono} \times Q_{oil} \end{aligned} \quad [\mu m] \quad (2.2)$$

Equation 2.1 was then set to 50 µm and used as a constraint on flow rates in a non-linear minimisation of equation 2.2. This resulted in a set of flow rates that will yield a predicted

diameter of 50 µm at the lowest possible diameter variation. The entire modelling process is visualised in fig. 2.8 and MATLAB code can be found in S.1.2.

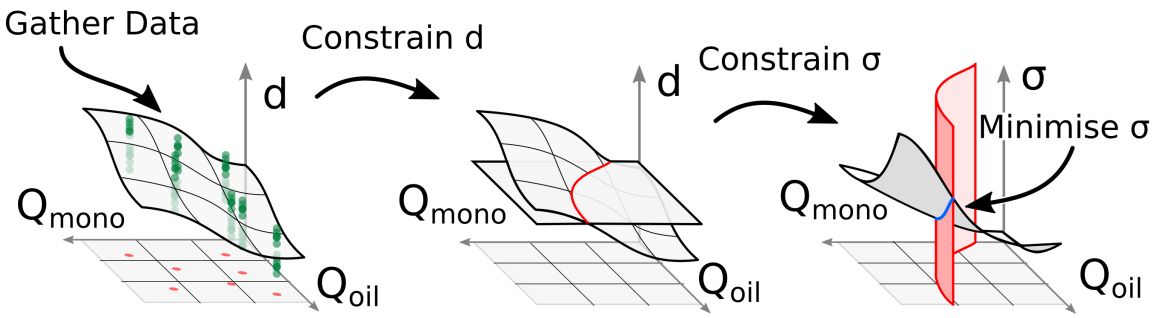


Figure 2.8: Bead parameter modelling. A surface response is fitted through diameter data gathered on the experimental design flow parameters. Diameter is then fixed to the desired amount, and the intersection is used as a constraint space in a minimisation problem on the second model, which relates standard deviation and flow.

For our desired diameter of 50 µm, the models suggested flows of 1461 µl/h and 1500 µl/h for Q_{oil} and Q_{mono} respectively. Using these parameter combinations, we produced a batch of hydrogel beads which will be used throughout this thesis. These hydrogels averaged a diameter of 53 µm with a standard deviation of 5 µm, which was sufficiently close to the target diameter for us to continue to the barcoding step. As producing hydrogels takes 2-3 work days, we did not further validate the model for other diameters. The model will be reused (and validated) on new microfluidic chips to produce beads of different diameters in the near future, when we may require beads of a different size for different protocols.

Barcoding the Beads

In order to equip every single hydrogel bead with clonal copies of a pseudo-unique barcode, we performed two sequential split/pool reactions according to Žilionis et al. (2017), giving each bead in the stock one out of 147 456 possible barcodes. In a stock of millions, the beads are thus not unique ("pseudo-unique"), but in a single experiment using ~100 000 beads, the chance of collisions is low. The barcoding is shown graphically in figure 2.9 and the detailed protocol is given in M.1.5.

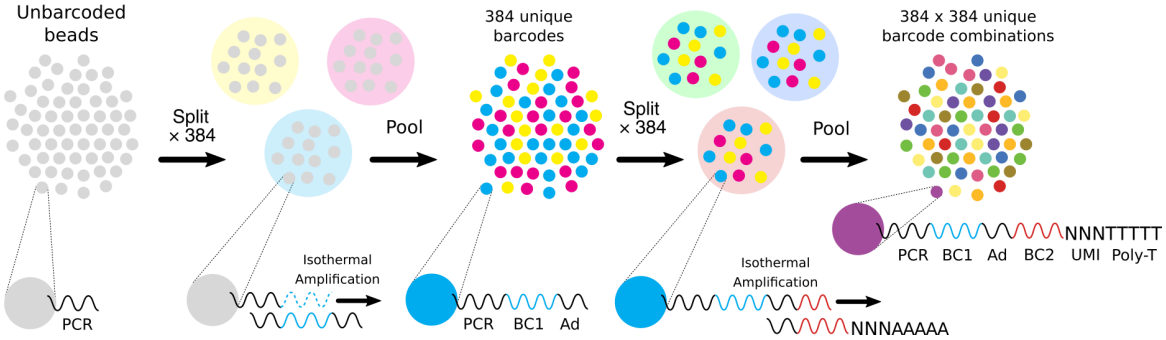


Figure 2.9: Hydrogel bead barcoding process. Unbarcoded hydrogel beads are split into 384 wells where a first well-specific, barcoded primer hybridises to the acrydite stub via a PCR handle (PCR). Barcode 1 (BC1) is then appended to the acrydite stub by isothermal amplification. The beads are re-pooled and re-split in 384 new wells where a second well-specific barcode (BC2) is appended to the first via an adapter sequence (Ad). This adapter sequence will also serve as a sequencing primer site later on. The split/pool process leads to 384 x 384 = 147 456 possible paths that a bead may have travelled, and thus an equal number of barcode combinations. Adapted from Žilionis et al. (2017)

We then performed two simple quality control experiments on the barcoded hydrogel beads (BHBs). First, we performed a series of FISH-like experiments during different stages of the barcoding process to asses if the previous step had succeeded (figure 2.10). After every step in the barcoding process, beads that had just undergone isothermal amplification were incubated with a fluorescent FAM probe complementary to the newly appended barcode part. As a negative control, beads that did not undergo the last step were also incubated with the fluorescent probe. The results of these tests were as expected - only those beads that had undergone barcoding showed a fluorescent signal. A background signal is detected in 2.10-b, which could be explained by the change in camera setup or non-specific binding. Theoretically, it is possible to quantify the light intensity signal from the beads to estimate the amount of primer on the oligo. We did not apply this method as our camera setup was not fixed, but plan to do this once a fixed camera epifluorescence microscope is available.

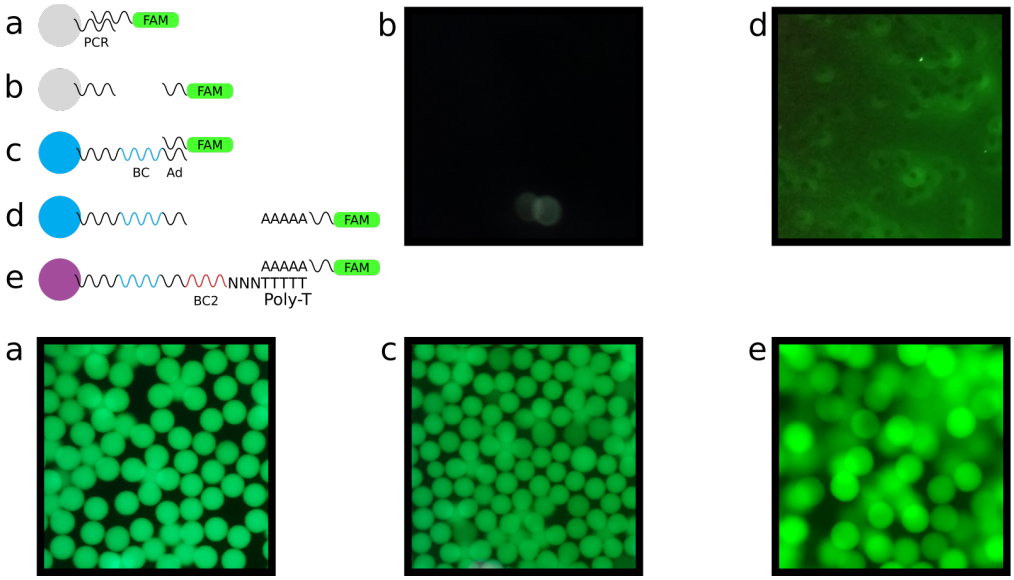


Figure 2.10: BHB FISH quality control. (Un)barcoded hydrogel beads were incubated with a fluorescent FAM probe, washed, and examined under a microscope. The following combinations were used: a) unbarcoded beads with an acrydite stub PCR primer probe, b & c) unbarcoded and partially barcoded beads with an adapter probe, d & e) partially and fully barcoded beads with a poly-A probe.

Second, we tested whether the disulfide-cleavage mediated release of primers from the beads worked as planned. Here, beads were incubated in various concentrations of DTT, after which the presence of released oligo in the supernatant was detected using qPCR (M.1.6). The concentration of beads was equal to the ratio of bead to water inside the droplet during an inDrop run. The resulting data showed that there is no significant difference between the amount of primers released by 10 mM or 125 mM of DTT (figure 2.11).

We therefore continued our experiments with the lowest concentration of DTT in order to not disturb the reverse transcription reaction. It can be seen that even at 0 mM DTT, there is a significant concentration of primer in the supernatant. This can be explained by either contamination of the sample - that is, supernatant still containing beads after centrifugation - or contamination of the bead stock with free-floating primers. As these free-floating primers would barcode the cell's nucleic acid without sharing a barcode with that cell's bead oligos, they would negatively impact the specificity of the inDrop scRNA-seq results. As of now, the true impact of possible stock contamination and how to reduce it is unclear.

After sieving them through a 70 µm filter, we attempted to use the freshly barcoded beads on the 10x Chromium microfluidic chip as originally intended. We loaded the beads onto the chip as per chip manufacturer's instructions, and started the run. Sadly, the

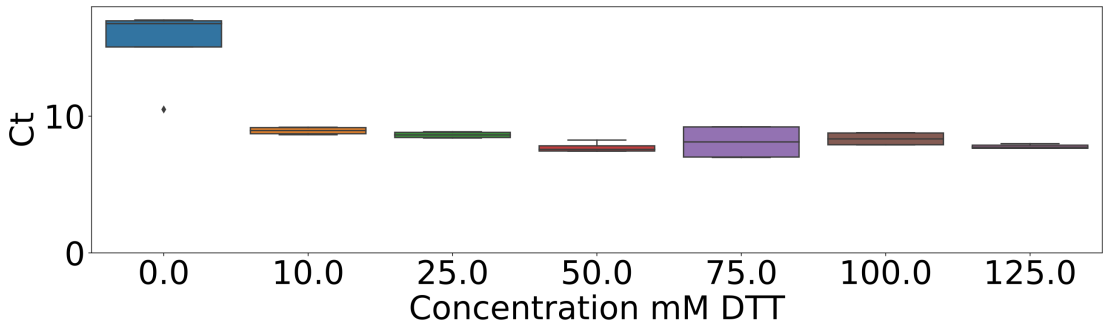


Figure 2.11: Bead oligo release qPCR results.

Chromium did not generate droplets with beads in them - a very crude emulsion with large droplets was formed in most output wells instead. The single output well which had formed a monodisperse emulsion, did not contain any beads. This suggests that dust in the bead stock had blocked the chip's fine microchannels before our beads could reach the flow focusing point, resulting in a beadless emulsion (figure 2.12). This dust would go on to become a major hurdle we had to overcome when producing a working microfluidic assay.

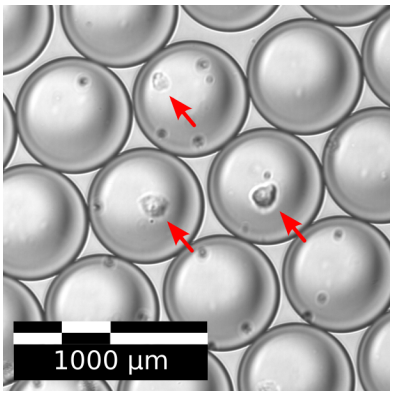


Figure 2.12: Failed 10x run with custom beads. Arrows indicate cells.

2.3 Execution of the Improved inDrop Protocol

This section will detail all of the (partial) inDrop runs we performed and the rationale between the specific parameter settings employed.

Preliminary Bead and Cell Loading Tests

First, we wanted to get acquainted with the microfluidic chip. We performed a number of dry trial runs with only beads to explore the flow rate settings that would lead to sufficient bead loading. The cell flow was replaced by mock flow of PBS at $777\,\mu l\,h^{-1}$. During the first run, dust from the bead stock immediately began blocking the fine microfluidic channels (figure 2.13), halting the entire process and in some cases leading to delamination of the PDMS from the glass slide due to overpressure.

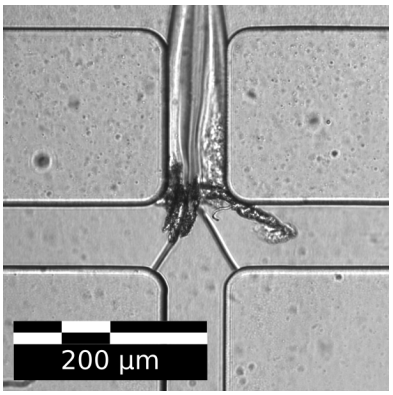


Figure 2.13: Dust blocking microfluidic channel.

We suspect this dust was introduced during the barcoding process, as the 4 96-well reaction plates are exposed to the open air for 10-30 minutes during manual addition of the barcoded primers. After barcoding, the BHBs are washed several times and filtered through a 70 µm filter. This filtering process was repeated a number of times after the dust was observed, but did not completely eliminate all dust from the stock. The filtering step also leads to small losses, so it was not repeated after two more times.

To solve or alleviate the dust issue, we came up with three possible solutions:

1. Eliminate dust from the barcoding process by using an automated liquid handler in an enclosed space. We started working on automated barcoding protocols for the Hamilton Microlab STAR early on, before the first round of barcoding took place. During initial tests, we found that the viscous properties of the bead solution and the low volumes used during barcoding would require careful fine-tuning of the STAR protocol. We therefore barcoded the first batch of beads manually. Automation of the barcoding is an ongoing project.

2. Filtering the beads by simply running them through an inDrop chip, and switching to a fresh chip when dust blocks the channels. This strategy produced a very pure bead solution, but took an exhausting amount of time at $300\,\mu l\,h^{-1}$. Due to high viscosity of the bead stock, higher flow rates would usually lead to delamination problems or produce pressures high enough to push dust through the channels.

3. Produce a filtering inDrop chip design. In the first filter design, a number of tightly-packed columns blocked both large and small dust filaments from entering the microfluidic channels while allowing beads and cells to pass. In practice, this design often suffered from delamination as the very fine filter acted as a point of high resistance for the bead stock to pass. We therefore adjusted the design with a set of more streamlined filters, which provided a functional balance between filtering capacity and fluid resistance. Figure 2.14 shows the different filter designs.

Out of all options considered, the filtering inDrop chip design yielded the best practical result, so we continued down that path.

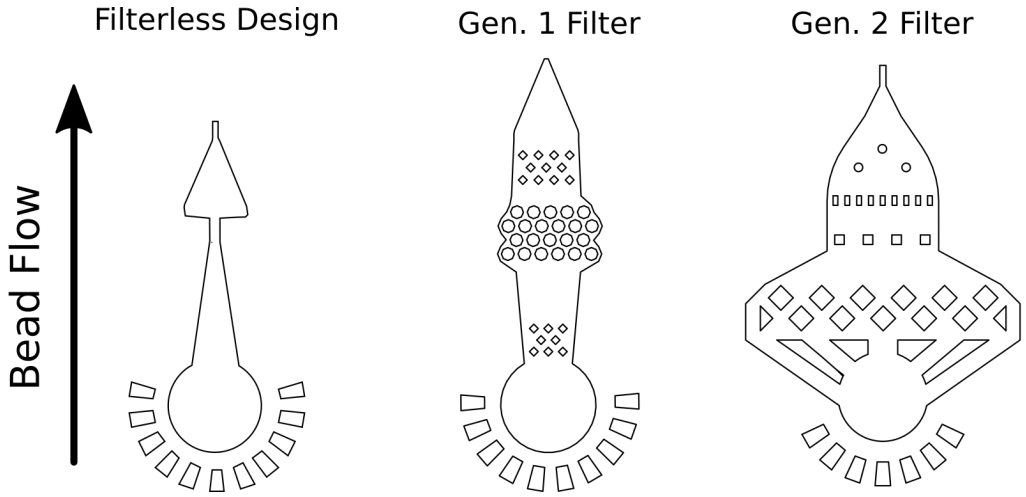


Figure 2.14: inDrop filter design evolution.

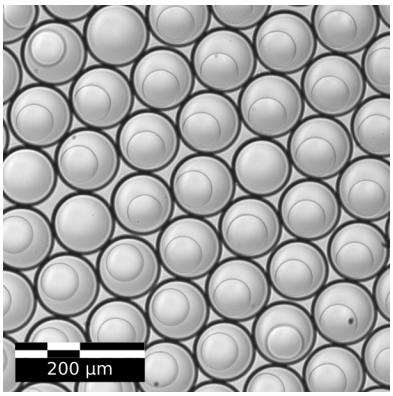


Figure 2.15: inDrop dry run bead coverage.

Once the dust issue was under control, we started fine-tuning the flow rates used for bead/cell encapsulation, starting with beads only (M.2). After a few rounds of trial and error, we found that flow rates of 300 µl h^{-1} for mock cell flow (PBS), 200 µl h^{-1} for beads and 450 µl h^{-1} for oil achieved a ~90% coverage for beads in droplets. In later runs with cells, we were not able to achieve such high coverages due to low sample volumes not permitting adjustment before the run ended. In the final run that was eventually sequenced, we also used higher average flow rates to increase speed and combat cell sedimentation/sticking to the chip.

We then started performing inDrop runs with beads and cells to try to achieve good droplet coverage of both cells and beads. During these trials, and the next, we used either melanoma (MM087 or MM074) or breast cancer

(MCF7) cells, as they were readily available to us in the lab and we were interested in the physical properties of the system rather than the biological properties of the cells. The RT enzyme was replaced by an equal volume of 50% glycerol to reduce cost and retain the flow properties of the RT mix. Cells are rapidly lysed after they come into contact with the bead/lysis mix, making it difficult to asses how well the cells have loaded. As we were still in the process of setting up the protocol in pilot runs, we decided to keep cell concentrations high during all experiments described in this thesis (initially up to 800 000/ml, gradually scaled down to 300 000/ml) to ensure a signal and to combat cells sticking to the microfluidic chip and tubing. A side-effect of this approach is low per-cell signal after sequencing, as the number of reads is limited per sequencing run. In a true single-cell experiment, we would scale the cell concentration down by a factor 10-100 to increase per-cell sequencing depth.

During these trials, we also found that the cells formed an insoluble precipitate together with the master mix, which again led to complete obstruction of the microfluidic channel. After a number of replacements and additions to the RT master mix, we succeeded in reducing - but not completely eliminating - cell precipitation. The main change was the removal of BSA from the cell suspension. BSA was originally incorporated here to prevent cells from sticking to each other and to the microfluidic tubing. We also tried to pre-incubate the tubes instead of coating the cells with BSA. This reduced cell-sticking to a minimum, but the film of BSA delaminated from the tubing when brought into contact with the RT mix. We therefore removed all BSA from the protocol, which led to higher cell sticking, but strongly reduced precipitation.

Generating cDNA Libraries

After the cell suspension precipitation was reduced, we started performing inDrop runs with beads, cells and functional RT mix. The emulsions generated here were incubated on a heat block for reverse transcription and, if they passed cDNA quantification on the Qubit, were further processed for ISPCR. Figure 2.16 shows a still of a single cell about to be encapsulated in the microfluidic chip.

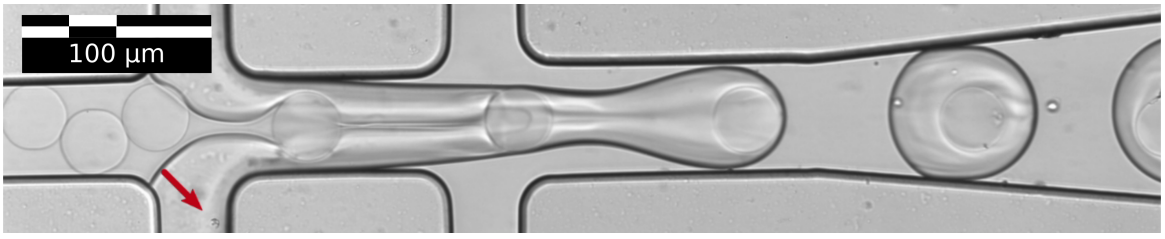


Figure 2.16: inDrop snapshot. Arrow indicates a cell about to be encapsulated.

In the first trial (M.2.2), we used an RT master mix containing 11.25% PEG, which is present for molecular crowding of the RT reaction, and no BSA. After reverse transcription, we pooled the libraries by breaking the droplets. Here, we measured 259 ng of DNA. Half of the library was treated with Exo I to remove residual primers left in the sample. Both samples were ISPCR-amplified using Terra polymerase. After a second round of clean-up, including Ampure bead filtering of small fragments, the Exo I treated sample yielded only 12 ng of DNA, and the untreated sample 57 ng. This was deemed insufficient, and we did not process the libraries further.

In the second trial (M.2.3), we wondered if the Terra enzyme may have been at fault, producing such low amount of DNA in the first trial. We therefore decided to exactly repeat the first trial, but process half of the sample with Terra polymerase, and half of the sample using KAPA HiFi polymerase. We did not perform any Exo I treatment, as it seemed to had negatively affected the yield in the previous run. Now, the Terra PCR run produced 104 ng of DNA, and the Kapa PCR run produced only 22 ng of DNA after Ampure bead cleanup. However, this library failed the library preparation stage as described in the next section.

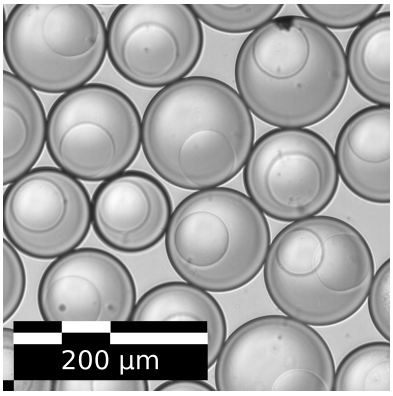


Figure 2.17: Final inDrop run droplets.

The final run that was eventually processed for sequencing was performed by a senior lab scientist, and did not incorporate PEG in the RT suspension. This run, which was performed on MM087 melanoma cells, passed all quality control stages, and was also successfully processed for sequencing. Figure 2.17 shows the droplets generated in this final run. These droplets were generated at higher flow rates than the previous runs (800, 1100 and 1200 µl h^{-1} for cells, beads and oil), as this reduced runtime, meaning less time for the cells to sediment and the enzymes to denature.

Library Preparation Troubles

After we had achieved a sufficient yield after the ISPCR stage, moved on to sequencing library preparation. As we wanted to sequence the libraries using Illumina's sequencing by synthesis technology, we needed to append P5 and P7 adapters to all barcoded fragments. These adapters are used on the Illumina flow cell to generate clusters using bridge amplification. In our case, the P5 adapter comes with a sample index which can be used to demultiplex pooled samples after sequencing.

Initially, we planned to perform adapter tagging using a classical Illumina Nextera XT kit, which first tagments the library and appends P5 and P7 adapters to the Tn5 adapters using PCR. Due to unknown reasons, this library prep approach failed multiple times on our library, producing no or low yields of amplified DNA, and an unfavourable fragment

length distribution (figure S.2). We therefore tried the NEBNext Ultra II Library Prep Kit, which did produce a sequencing library. This kit works by fragmenting the cDNA and ligating a special hairpin loop adapter to both ends of the fragments. This hairpin is then enzymatically cleaved open to produce non-complementary strand ends which can be used to selectively amplify one strand. The NEBNext kit thus utilises a hairpin loop to produce strand-specific sequencing libraries, which can be used to identify antisense transcripts. These transcripts may play an important role in regulating gene expression (Mills et al., 2013). The entire final library preparation process is outlined in figure 2.18 and, together with the inDrop run protocol, in M.3.

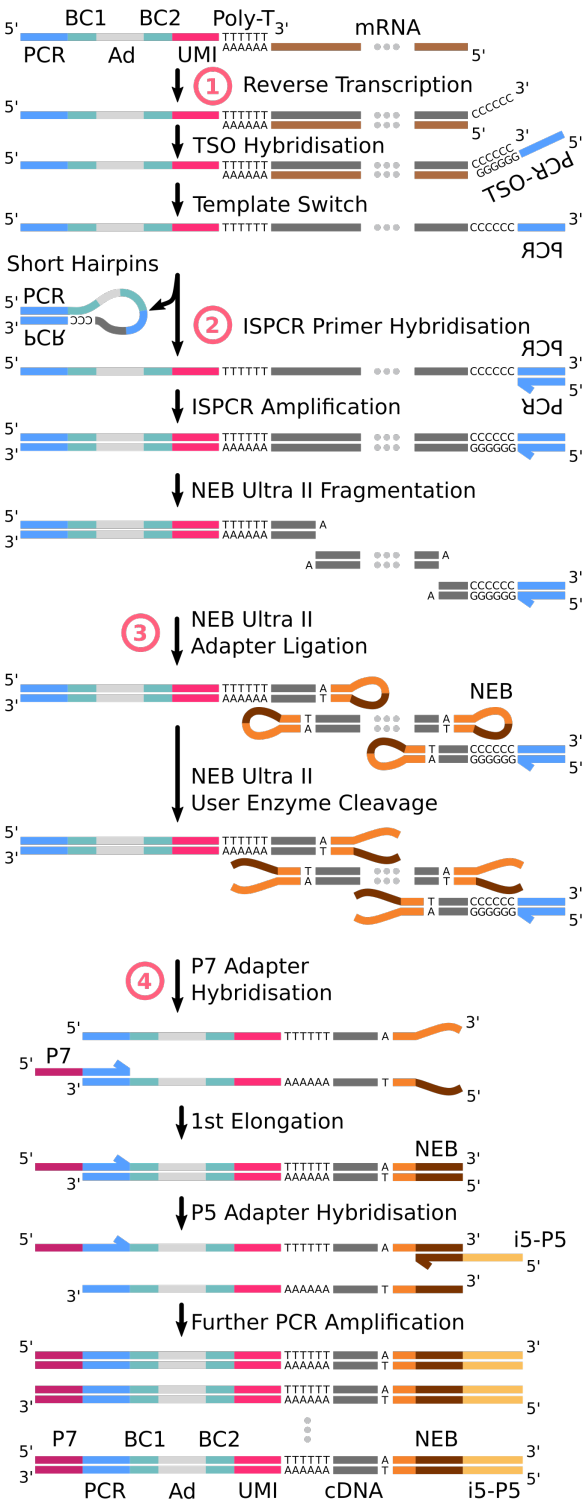


Figure 2.18: inDrop sequencing library preparation.

1. Barcoded poly-T primers are released into the droplet by the hydrogel bead and hybridise to Poly-A$^+$ mRNA from the lysed cell. An MMLV reverse transcriptase catalyses the reverse transcription and adds a string of cytosine nucleotides to the 3' end of the cDNA transcript which serves as a hybridisation site for the template-switching oligo (TSO). The TSO is used as a template for further polymerisation by the MMLV reverse transcriptase and contains the reverse sequence of a PCR primer that is also present at the 5' end of the barcode, leading to complementarity of both ends.

2. The single-cell cDNA libraries are then pooled and undergo ISPCR, where we add a single PCR primer which can hybridise on either of the transcript. Fragments that contain a short cDNA strand can form hairpin loops, preventing hybridisation and thus reducing the size-selective bias associated with classical PCR.

3. The cDNA is then fragmented, end-repaired and dA-tailed using the NEB Ultra II Illumina Library prep protocol. The NEB hairpin adapter is then ligated and cleaved, resulting in two non-complementary primers on the sense and antisense strands.

4. Lastly, the adapter-ligated library is PCR-amplified with Illumina P7 and P5 adapters. First, a P7-PCR-primer hybridises to the 3' end of the antisense strand, and elongation occurs. Only now, the i5-P5-PCR-primer hybridisation site is present, and exponential amplification can occur. This PCR is thus selective for the antisense cDNA strand of the barcoded fragment.

Sequencing Results

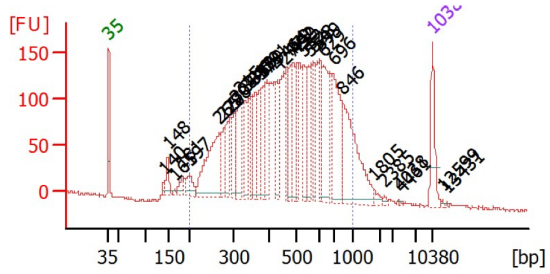


Figure 2.19: inDrop electropherogram.

The final library preparation led us to a library that had an average fragment size of 400 bp as indicated by the Bioanalyzer electropherogram (figure 2.19). This library was sequenced on the Illumina NextSeq 500. The resulting sequencing data was mapped using STAR, leading to a 56% mapping percentage, which is relatively low. We were pleased to see that the library showed a classical RNA-seq profile characterised by exon/transcript agreement and enrichment for transcripts mapping to housekeeping genes. The library was also strongly 3' enriched, showing that our library preparation succeeded in selectively amplifying the barcoded transcript ends only (figure 2.20).

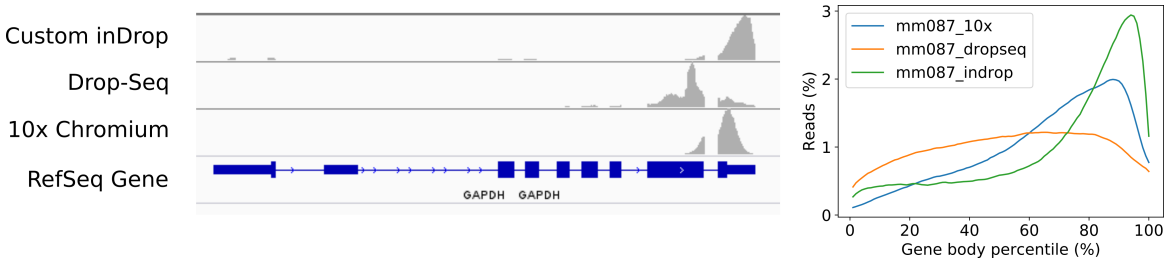


Figure 2.20: Gene coverage comparison. Gene coverage of our custom inDrop library is compared to a Drop-seq and 10x Chromium scRNA-seq dataset on the same MM087 cell line. GAPDH is a housekeeping gene and used as a reference point. Data are scaled to account for sequencing depth.

However, we quickly discovered that the cell barcode could not be retrieved for the vast majority of transcripts. It appeared that only the cDNA, but not the barcodes, had been sequenced well on the Illumina platform. The lack of cell barcodes had effectively turned our single-cell dataset into a (very expensive) bulk RNA-seq. The ramifications of this discovery and how we attempted to explain it will be further detailed in chapter 4.

3 Development and Execution of a Single-Cell ATAC-seq Protocol

3.1 Designing Drop-ATAC

This part of the thesis documents how we designed and performed a droplet microfluidics-based single-cell ATAC-seq protocol, which we named Drop-ATAC. By making a number of changes to our custom inDrop protocol, we effectively extended the functionality of our microfluidic framework to the assay for transposase-accessible chromatin. Conceptually, Drop-ATAC and inDrop are similar, but molecularly they are very different. Drop-ATAC isolates single nuclei in nanolitre droplets together with a barcoded hydrogel bead. These barcoded hydrogel beads are produced in a similar process as described in section 2.2. However, the barcode sequence now captures cellular DNA instead of mRNA which is then barcoded and amplified using droplet PCR. Figure 3.2 shows the rough workflow of Drop-ATAC, while figure 3.3 shows the manipulations a DNA strand undergoes in the protocol.

Similarly to the process described in chapter 2, we iteratively performed Drop-ATAC runs until we were confident in the quality metrics of the final ATAC library, which was then processed for sequencing. The roadmap that we travelled during this process is given in figure 3.1.

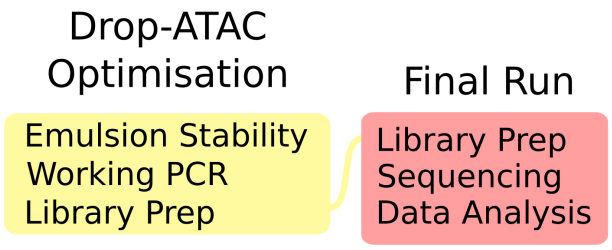


Figure 3.1: Drop-ATAC workplan.

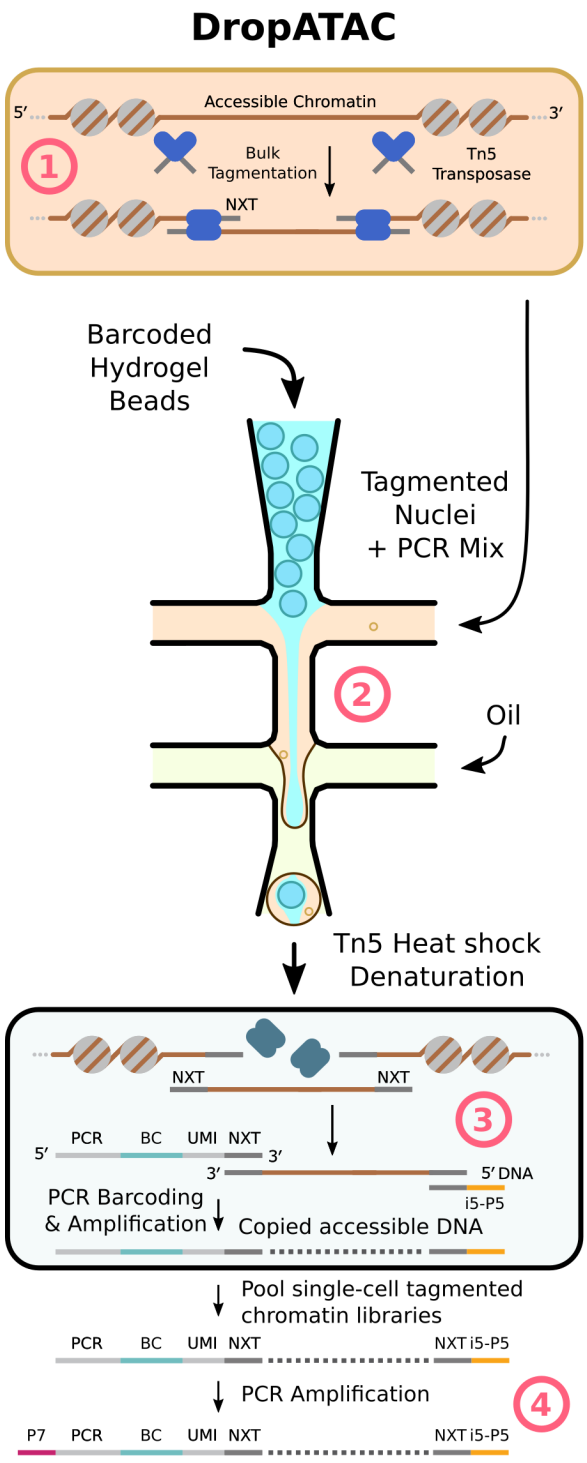


Figure 3.2: Drop-ATAC protocol overview

1. Isolated nuclei are tagmented in bulk. Tn5 simultaneously cleaves accessible chromatin regions and inserts its adapters at the cleavage site. The Tn5-DNA complex remains intact until denatured.

2. The suspension of "tagmented" nuclei is then run through the same microfluidic device used in our custom inDrop. Here, the single nuclei are now encapsulated together with a barcoded hydrogel bead. The nuclei suspension contains PCR reagents and DTT, which mediates release of the barcodes from the hydrogel bead. In the Drop-ATAC protocol, however, the barcode does not contain a poly-T tail to capture poly-A^+ mRNA, but the sequence complementary to the Nextera Tn5 adapter.

3. The resulting bead-nuclei emulsion then undergoes a heat shock which denatures the Tn5-DNA complex followed by PCR cycling. During this step, the Drop-ATAC barcodes capture the newly released tagmented DNA fragments and, together with indexed Illumina P5 adapters, act as PCR primers. In this emulsion PCR, the tagmented DNA is thus barcoded and i5 indexed.

4. The single-nucleus ATAC libraries are now pooled and tagged with Illumina P7 adapters for sequencing in a final adapter PCR.

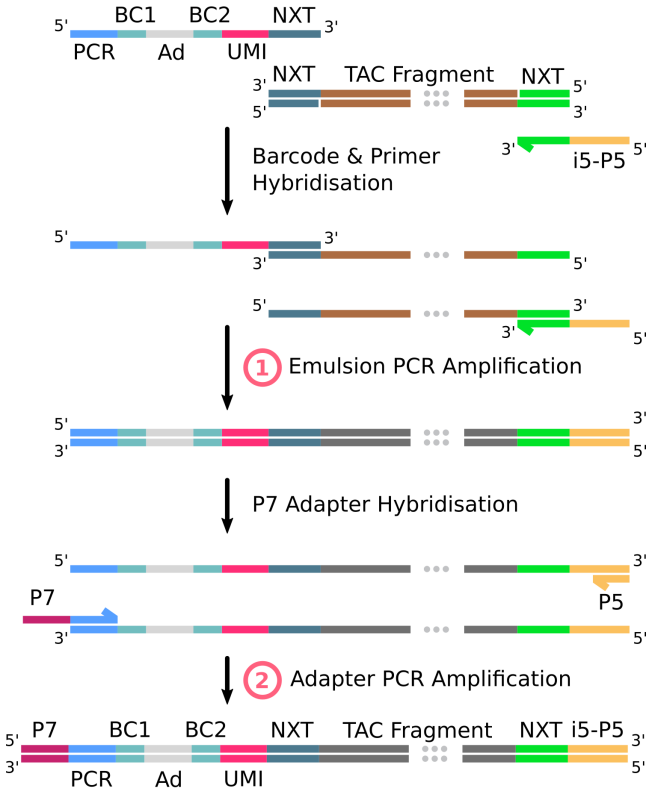


Figure 3.3: Drop-ATAC sequencing library preparation.

What happens during the multiple PCR steps can be summarised as follows:

1. In the emulsion PCR step, transposase accessible chromatin (TAC) fragments are captured by barcoded primers released by the hydrogel bead and tagged with an Illumina indexed P5 primer.

2. The barcoded single-cell ATAC libraries are pooled and undergo a second PCR reaction. Here, the barcoded TAC fragments receive a P7 primer, resulting in a sequencing-ready library after subsequent bead purification.

3.2 Execution of Drop-ATAC

Before we arrived at the protocol described in M.5, we went through a number of trials in order to find those parameter combinations that would lead to an ATAC library ready for sequencing. The reagents used in the preliminary/optimisation stage are described in M.4.

PCR Emulsion Stability Trials

Previous experiments carried out by a senior lab scientist had already shown that the many successive temperature cycles involved in PCR destabilise the bead-nucleus emulsion. Since the polymerase mix constitutes a significant part of the droplet volume, we tested whether different polymerase mixes would yield different results in droplet stability. Two PCR mixes were tested: Takara Bio e2TAK and NEB Q5®. Both PCR mixtures were equal in terms of added electrolytes, dNTPs, and PEG - except for the manufacturer's 5x PCR buffer supplied with the two polymerases. Droplets were generated according to the general protocol outlined in M.5, but without nuclei (dry run). When comparing the droplets pre- and post-PCR, it became apparent that many of the droplets had merged during thermocycling (figure 3.4), but the emulsion was more stable than previous experiments where all droplets had merged.

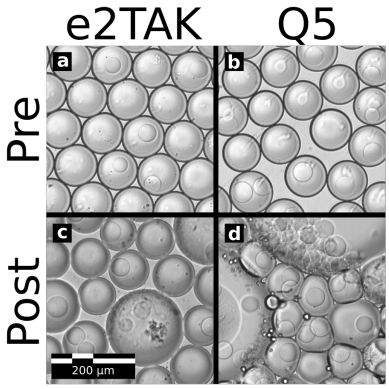


Figure 3.4: Emulsion PCR stability with PEG.

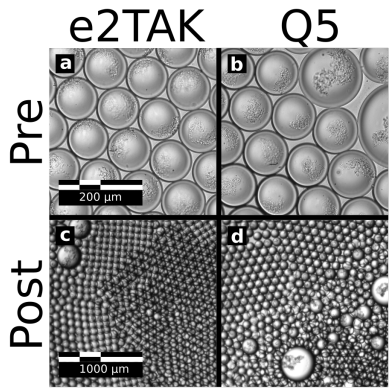


Figure 3.5: Emulsion PCR stability with OptiPrep.

There was no immediate difference between both PCR mixes, aside from a marginally better droplet profile in the e2TAK mixture (figure 3.5). However, shortly after preparing the polymerase mixtures, we noticed an occlusion in the e2TAK mix. The precipitate was clearly visible under a microscope and posed danger to the microfluidic operation by clumping nuclei together. We hypothesized that this precipitation could be caused by a reaction between DTT and PEG, which are both present in the nuclei/PCR mixture.

Based on these observations, we decided to repeat the trial, but moved PEG from the nuclei-PCR mix to the BHB solution. We also added 15% OptiPrep to the nuclei-PCR mix (final droplet concentration v/v) in order to reduce nuclei sedimentation and aggregation in the absence of PEG, and 0.1% BSA to the lysis mix (final droplet concentration w/v) to further improve droplet stability. The

resulting emulsion was very stable, but we still observed precipitation in the nuclei/PCR mix (figure 3.5). In a final droplet stability and precipitation trial, we omitted all BSA and PEG, but retained the 15% OptiPrep and noticed that the precipitation had been heavily reduced. We decided to move forward with the protocol and test the PCR reaction efficiency.

A Number of Bulk Runs

After the droplet stability and nuclei aggregation issues were reduced (but not completely mitigated), we tested the functionality of the in-droplet PCR reaction by performing several "bulk" Drop-ATAC runs. In essence, we did not use BHBs here, but rather encapsulated the nuclei with the PCR reagents while the PCR primers were already dissolved in the nuclei suspension. Such a trial performed by a senior scientist had provided an encouraging Bioanalyzer plot. This run was bare-boned: it omitted most non-necessary additives from the nuclei mix entirely to prevent precipitation or side-effects, using a pre-made NEBNext mix instead. Due to an operation error, the repeated "bulk" run was continued for a moment after the bead and cell suspensions had run out. As the syringes are primed with PBS, this resulted in the generation of PBS-filled droplets in the sample emulsion. Despite this error, we decided to process this sample for PCR, and were pleased to find that the resulting emulsion was very stable post-PCR, even moreso than the original bare-bones run. A possible explanation for this effect is that the PBS-filled droplets stabilise the rest of the emulsion. The resulting "bulk" library had a very strong ATAC-like electropherogram profile (figure 3.6). The different peaks and dales show a pitch of around 200 bases, which corresponds to the discrete number of nucleosomes associated with inaccessible DNA. We therefore kept the generation of the PBS droplets in the following Drop-ATAC runs.

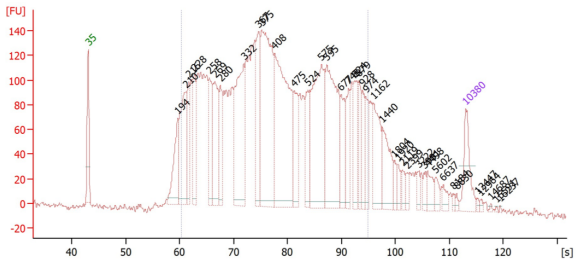


Figure 3.6: Bulk Drop-ATAC Bioanalyzer electropherogram.

The first "bulk" Drop-ATAC run used NEBNext primers in the emulsion PCR in order to guarantee that the primer aspect of the reaction was not at fault if the PCR did not generate DNA. We then performed a second bulk Drop-ATAC run, but replaced the standard NEBNext PCR primer with our custom barcoded primer. All primers had the same BC1 and BC2 sequence, so there was no variation between the primers used. This run produced DNA, but only about 30% the amount of the first run (with the NEBNext primer), indicating that the custom PCR primers do introduce some inefficiency in the reaction. So far, the nuclei remained intact throughout the whole ATAC-seq procedure - being present even after PCR cycling. Before proceeding, we therefore performed another

run where the we added Triton X-100 to the bead mix (leading to a final concentration of 1.8% w/w in the droplets). Such high concentrations of Triton X-100 will lyse the nuclei, and we hoped that this change would lead to an increased yield of DNA post-PCR. While the nuclei did lyse due to the addition of Triton X-100, there was no DNA post-PCR. We therefore re-ran with a only 0.36% Triton X-100 in the droplets, which did produce 80 ng of DNA. This was too low, and combined with a viscous precipitation that appeared only after we started adding Triton X-100 to the mix, we decided to omit Triton X-100 from the final protocol.

Final Sequencing Library Preparation

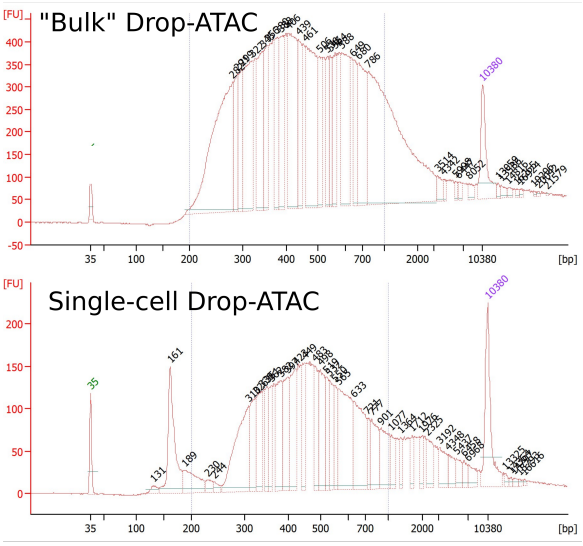


Figure 3.7: Electropherogram of final libraries.

The final two runs, of which the libraries were both sequenced, had the following properties:

1. In the final "bulk" Drop-ATAC run, a suspension of tagmented MCF7 nuclei was emulsified with a liquid primer mix consisting of the non-uniquely barcoded bead oligo and the indexed P5 primer, as described in figure 3.3. We used MCF7 cells as they were readily available to us in the lab. There was no addition of BSA, PEG or Triton X-100, and we observed no strong precipitation. As described earlier, half the volume of the resulting suspension consisted of PBS droplets, and the resulting emulsion was very stable post-PCR.

2. In the true single-cell Drop-ATAC run (M.5), tagmented MCF7 nuclei were co-encapsulated with barcoded hydrogel beads. The beads were washed using the same buffers as used in our custom inDrop protocol. The nuclei suspension now contained 20 mM of DTT in order to release the barcoded primers from the bead.

After final P7 adapter PCR and Ampure purification, the PCR yielded 1075 ng of DNA, but the single-cell Drop-ATAC produced only 193 ng of DNA. This difference in yield can be explained by the reduced PCR efficiency we observed in previous bulk runs with the custom barcoded primer. The washing buffers used to treat the beads before a high concentration of sodium chloride, which may have affect the PCR reaction. This washing step, originally tailored for the inDrop protocol, needs to be reviewed and adapted to the Drop-ATAC protocol.

The electropherograms of the two sequencing libraries did not exhibit a strong ATAC profile as observed with the earlier "bulk" Drop-ATAC run (figure 3.7). The single-cell library also showed a strong small fragment peak. This peak was removed in an additional Ampure bead cleaning step.

3.3 Sequencing Results

76% of the single cell Drop-ATAC and 78% of the "bulk" Drop-ATAC reads were mapped to the reference genome using STAR. Figure 3.8 shows the coverage of GAPDH. Both datasets showed enriched coverage of the gene's 5' end, indicating an accessible TSS, as expected for a housekeeping genes such as GAPDH. When aggregated, the single cell Drop-ATAC data thus showed good bulk properties.

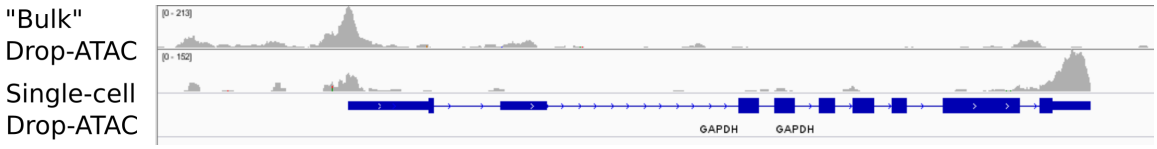


Figure 3.8: Drop-ATAC gene coverage.

However, the single-cell Drop-ATAC dataset also had a strong enrichment of 3' fragments which is *the* defining characteristic of our inDrop library. There are a number of ways this can be explained, the most likely one being contamination during library preparation. Since we performed many of the inDrop and Drop-ATAC experiments in parallel, it is possible that a contamination of indexing primers happened during the inDrop library prep. Such a mistake could have led to index collisions, as the inDrop and Drop-ATAC libraries were sequenced on the same flow-cell. We do not believe that the Drop-ATAC library truly contains a large number of 3' fragments, as our protocol cannot capture or amplify them in such large quantities.

Similarly to the inDrop library, we were unable to demultiplex the single-cell Drop-ATAC library due to lack of barcode reads. This problem will further be explored in the next chapter.

4 The quest for cell barcodes

In the previous chapters, I showed how we modified the inDrop microfluidic framework to a) accomodate a Smart-seq/Drop-seq-like scRNA-seq and b) implement a novel Drop-ATAC protocol. The normal course of action would now be to separate the fragments from the pooled libraries based on their cell barcode, and account for PCR induced bias by counting UMIs (figure 4.1). However, while the bulk characteristics of the resulting datasets met our expectations, we were unable to retrieve cellulcar barcodes, making it impossible to demultiplex our single cell libraries. In this chapter, I will explain how we attempted to diagnose the problem in our libraries, its possible origins and how to solve it in the future.

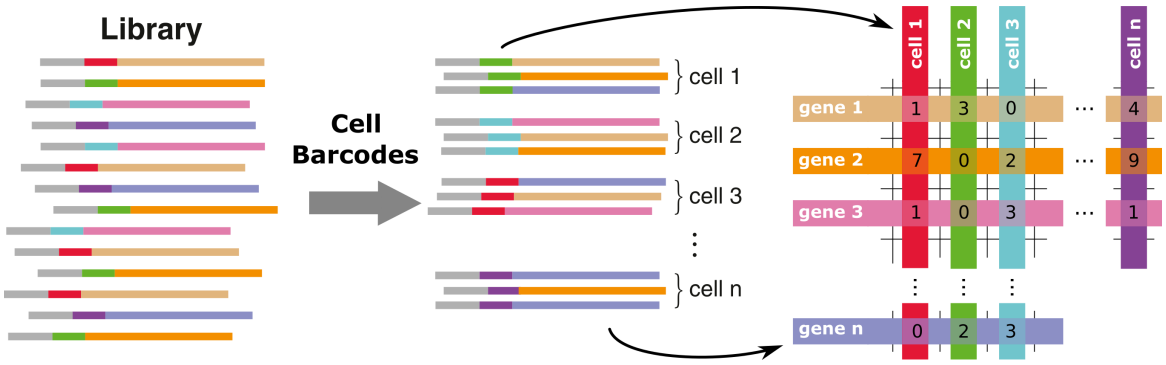


Figure 4.1: Library demultiplexing. Transcripts in a pooled single cell library can be demultiplexed by assigning them to a common cell of origin based on their cell barcode. Transcripts can be quantified by counting UMIs.

4.1 Illumina Next-Generation Sequencing

In recent years, the San Diego-based Illumina has dominated the sequencing market, operating at a de facto monopoly (Greenleaf and Sidow, 2014). The company's sequencing adapters have been used extensively in single cell omics, often being directly incorporated into published methods and protocols. Two advantages of Illumina sequencing are throughput and accuracy. A single fragment is clonally amplified to produce a fluorescent signal that can be used to call bases with high accuracy. However, quality drops progressively as the sequencing goes on due to cumulative errors. The read length of sequencing by synthesis is thus limited. A schematic overview of a DNA strand undergoing the different

steps of Illumina sequencing is given in figure 4.2. For a more detailed overview of bridge amplification on a glass substrate and sequencing by synthesis chemistry, see Fedurco et al. (2006) and Harris et al. (2008).

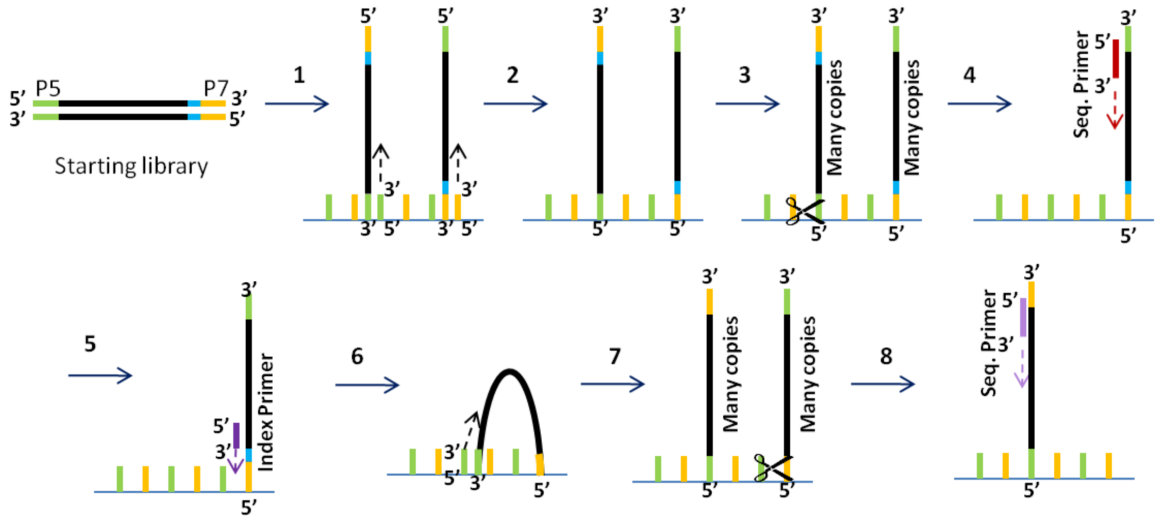


Figure 4.2: Strand progression during Illumina sequencing. **(1)** The fragment's P5 and P7 regions hybridise to the flow cell surface. **(2)** Complementary strands are synthesized and the original library is washed away, leaving only the freshly synthesised complement. **(3, 4)** Bridge amplification occurs, amplifying the single DNA strand to a densely packed clonal cluster. The P5 region is then cleave, leaving only P7-bound strands to be sequenced. **(5)** After the sequencing primer attaches to the P5 end, sequencing by synthesis (SBS) occurs. The combined fluorescent signal of all strands in a clonal cluster is optically detected as a bright spot on the flow cell surface and computationally translated to a base call. **(6)** Each fragment's sample index is sequenced using an index primer. **(7, 8)** In paired-end sequencing, a second round of bridge amplification occurs. The P7 end is now cleaved, leaving P5-bound strands to be sequenced. Adapted from Harvard MGH Sequencing Core Resources.

We sequence our scRNA-seq and scATAC-seq libraries differently from classical inDrop and Drop-seq libraries, which read their whole barcode sequence in a single go. Due to the split-pool PCR reaction we use to barcode the hydrogels, both halves of the barcode are separated by a 32 bp PCR primer site. This primer site is used as a read primer site for both halves of the barcodes (figure 4.3). By not reading the PCR adapter inbetween both barcode halves, and using it as a read primer site instead, we increase the cycle budget for the cDNA by 32 cycles. The process illustrated in figure 4.3a.

The Drop-ATAC library has a similar structure to the inDrop library, but also has a staggered barcode sequence due to mistake in the barcode design (figure 4.3b). This

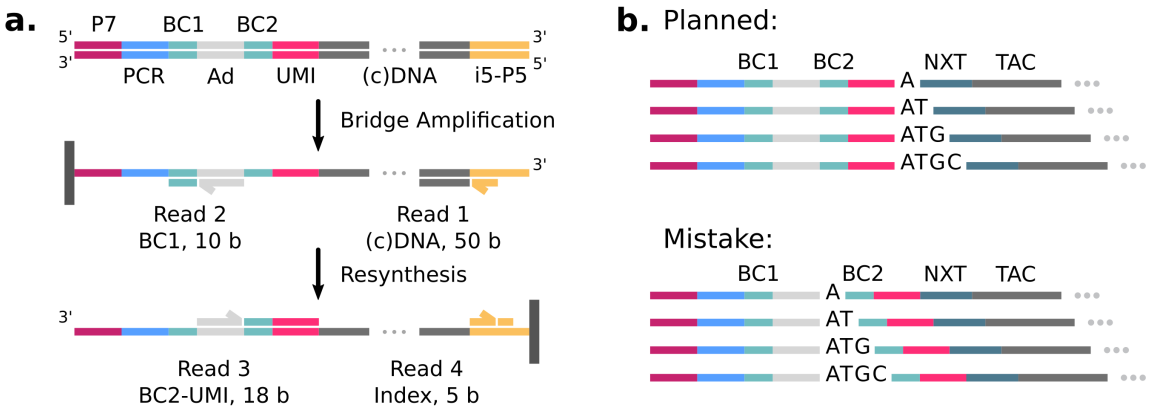


Figure 4.3: Custom sequencing process and stagger sequence. a. After the library is bridge-amplified, the first two reads will read the 3' cDNA (RNA-seq libraries) or DNA (ATAC-seq libraries) fragment and the first half of the barcode. The reverse complement is then synthesized, and the second half of the barcode and index are read by read 3 and 4. **b.** Incorrect placement of stagger sequence in Drop-ATAC library.

stagger sequence was originally intended to be located after the UMI, right before the Nextera Tn5 adapter sequence. Such a stagger would shift the Tn5 adapter by one nucleotide and, enabling us to extend the read cycles beyond BC2 without risking low base-call accuracy associated with highly similar sequences. Paired-end reads would allow us to map the exact length of the fragments. By mistake however, the stagger sequence was placed before BC2 instead, complicating the barcode identification process by requiring a slight change to the whitelist for barcode 2.

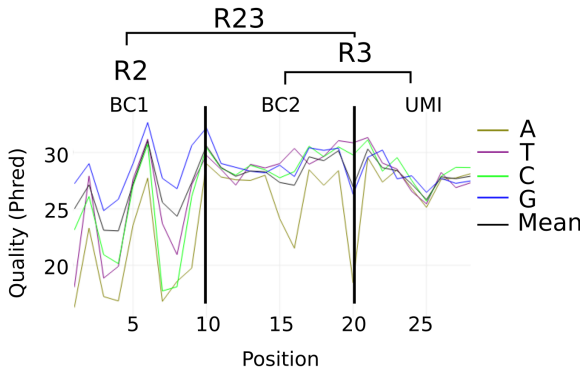


Figure 4.4: inDrop barcode read quality.

The sequencing data was mapped to the reference genome and examined. As mentioned in chapters 2 and 3, the bulk properties of the library were satisfactory, but we could not cell-demultiplex the transcripts. The quality of the barcode reads (Read 2 and Read 3, which are in-silico concatenated to a single virtual read, referred to as R23) was low and inconsistent compared to the cDNA reads (figure 4.4). Only 723 combined R23 reads had a mean Phred quality of > 34 (corresponding to a 99.96% average base call certainty), and out of these 723 reads, only 2 matched our barcode whitelist. Unfiltered, less than 0.01% of the R23 reads matched the 384 x 384 barcode whitelist completely. When looking at R2 and R3 individually, 3% and 10.9% of reads matched the whitelist respectively. The vast majority of R23 contained poly-G noise, illustrated in figure 4.5.

Similarly to the inDrop library, our single cell Drop-ATAC library could not be cell-demultiplexed. The problem is even worse - less than 1% of all R23 carry even one whitelisted half-barcode. We did not observe poly-G, but the single cell library's BC2 showed a high percentage of A bases. Even in the "bulk" Drop-ATAC library, which is produced using a single dissolved barcoded primer in all droplets, no significant amount of barcodes was found. Here, we would expect the single barcode used in the experiment to dominate the distribution - but it did not.

We briefly attempted to apply the BLAST-like alignment tool (blat) algorithm to R23 to see if we could retrieve any reads that resembled a whitelisted barcode (within 2 Hamming distance). However, due to the large database of 147k whitelisted barcodes, the algorithm was inefficient. We therefore tried to use `blat` on R2 and R3 separately - where we had to match only 384 whitelisted barcodes per read - but found no significant number of hits. We later on discovered that the blat algorithm is ill-suited for matching sequences of 25 bp or fewer, which at least partially explains the output we had achieved. In section 4.5, we use the `cutadapt` algorithm to look at a MinION sequencing dataset, a different algorithm which could be used to extract more information out of the "failed" inDrop dataset in the future.

We highly suspect that the Illumina sequencing process itself may have failed, and not our inDrop/Drop-ATAC protocols. The split-pool barcoding process, should not generate, under any circumstances, the poly-G or random sequence barcodes that are read. During the second isothermal amplification step, the second barcode half carrying the poly-T tail can only be appended to the bead primer if a BC1 was was appended there prior. If the primer released by the bead has no poly-T, it cannot capture a poly-A^+ mRNA fragment. It

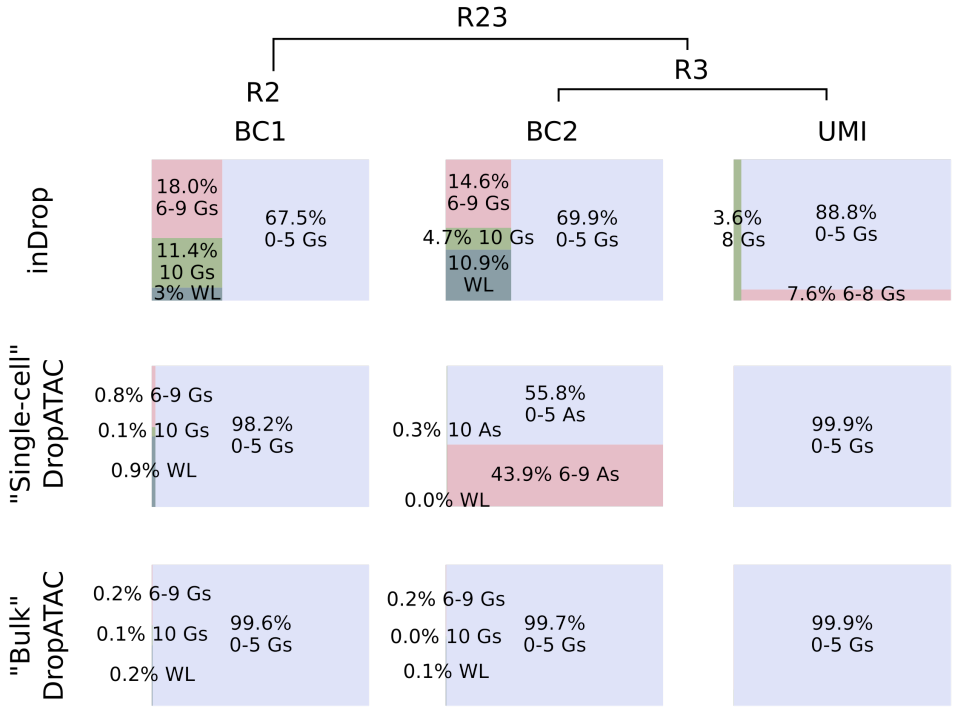


Figure 4.5: Barcode whitelist percentages. R2 and R3 refer to Read 2 and Read 3, R23 to the in-silico concatenation of the two. WL refers to the percentage of reads that could be retrieved from the whitelist, and 10Gs, 6-9Gs and 0-5Gs denote the number of Gs in the non-whitelisted read.

is therefore unlikely, especially at the scale we observe, that partial or incorrectly synthesised barcoded primers on the BHBs led to the effects we observe here.

Since the cDNA read (R1) and the sample index read (R4), which both use standard Illumina sequencing primers, were sequenced fine, our prime suspect became the custom sequencing primers used to read both barcode halves (R2 and R3). In our experiments, we used the Illumina NextSeq 500 platform, which uses a 2-dye colour system to identify the four different bases (figure 4.6). Poly-G reads, which are indistinguishable from *no signal* on the NextSeq 500 platform (figure 4.6), are most likely caused by the custom sequencing primer not hybridising. When this happens, sequencing by synthesis cannot occur, and the base calling software will interpret the lack of signal as a G nucleotide. Inversely, the non-whitelisted, non-poly-G reads could be explained by the custom sequencing primer hybridising where we do not expect it. However, we extensively searched through the theoretical fragment produced by our protocol and did not find any (close) match, and at a length of 32 base, it is unlikely to hybridise to the cDNA.

Finally, it is also possible that this NextSeq 500 run in particular yielded bad results due

to technical failure. The low-quality of some reads can be attributed to overclustering. Our sequencing run had a cluster density of 312 000 mm^{-2} and a filter pass rate of 61.5%, far beyond the flow cell's recommended 200 000 mm^{-2}. When overclustering occurs, base calling becomes less precise due to overlap in signals from different clusters.

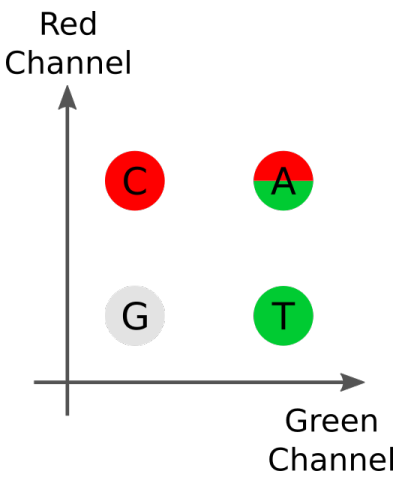


Figure 4.6: Illumina 2 channel technology.

The indications above prompted us to take the following course of action:

1. Sequence a number of inDrop and Drop-ATAC fragments by classical Sanger sequencing. Sanger sequencing can process the whole library fragment in one read, and does not rely on our custom sequencing primers. It therefore provides the true sequence of a low number of purified fragments, but cannot give us insights in macro-scale qualities of the sequencing libraries.

2. Re-sequence the libraries on the NextSeq 500 to eliminate technical failure, this time using a reduced concentration of input DNA and an increased percentage of spiked-in PhiX. PhiX is an Illumina spike-in standard that can be used as a control for sequencing accuracy and clustering efficiency.

3. Produce a new batch of BHBs that do not require custom sequencing primers, perform a new custom inDrop run and sequence on the NextSeq 500.

4. Sequence the library on the Oxford Nanopore Minion. This platform is also capable of sequencing the library's complete fragment length, but at a much larger scale than Sanger sequencing. Since it is a single-molecule sequencing technology, it does suffer from low single-base accuracy.

4.2 Sanger Sequencing

In order to sequence a small number of fragments with high fidelity, we reached back for the gold standard approach: Sanger sequencing. To amplify the fragments, which have a mean length of 400 bp, we used TOPO cloning followed by DNA purification. We sequenced 9 fragments from the inDrop library, 5 from the single cell Drop-ATAC library and 10 from the two "bulk" Drop-ATAC libraries.

The results are illustrated in figure 4.7. The Sanger sequencing showed that 4 out of the 9 sampled inDrop fragments perfectly matched the theoretical fragment we envisioned. The other fragments had a number of minor defects, but in total 15 of the 18 half-barcodes sequenced matched the whitelist. The Drop-ATAC fragments showed similar characteristics. Out of the 10 "bulk" Drop-ATAC fragments, 9 carried the barcode combination used in the experiment. The single cell Drop-ATAC library also had 5/5 fully whitelisted barcodes. One of the fragments did not contain a captured DNA fragment. It is unclear how this happened during the library preparation, and how it escaped the Ampure short-fragment filtering.

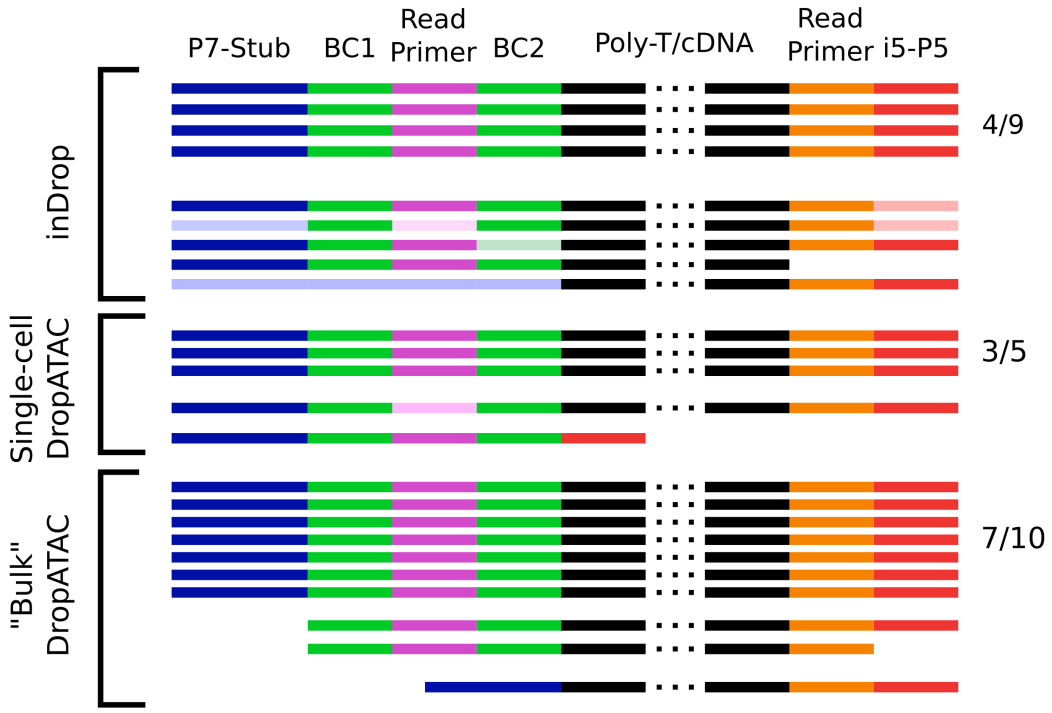


Figure 4.7: Sanger sequencing results. Lighter colour denotes slight deviation from the expected sequence. The bottom inDrop fragment was inverted in the TOPO cloning, leading to a barcode at the utmost end of the Sanger sequencing run, leading to low base-calling accuracy.

The Sanger sequencing confirmed our suspicions that the Illumina sequencing run was not representative of the true library. While the library had a number of true defects, the majority of sequenced fragments had a whitelisted barcode and an intact read primer site. Theoretically, it is possible that these well-behaved fragments were picked by chance, but given that TOPO cloning can be assumed to be unbiased, it's unlikely that our library is

not enriched with the good fragments detected here. We could thus narrow the problem down to flawed sequencing, most likely associated with our custom read primers.

4.3 NextSeq 500 Re-sequencing

We re-sequenced the inDrop and single cell Drop-ATAC libraries on the NextSeq 500 to eliminate the possibility of technical failure during the first run. This time, the spiked-in PhiX fraction was strongly increased from 1% to 20% and the amount of input DNA was reduced. These factors led to a lower clustering density of 205 000 mm^{-2} with 83.1% of clusters passing filtration - marginally overclustered, but not alarming. The overall sequencing quality of the re-sequencing run was better than the first run. Compared to the first sequencing run, the mapping percentages remained the same for both inDrop (56%) and Drop-ATAC (77%). Figure S.4 shows the gene coverage of the re-sequencing run, which was also near-identical to the first sequencing run.

In the inDrop library, we still found poly-G reads, but not to the extent of the first sequencing run. The quality of the inDrop barcode reads also improved to 66% of the BC1 reads and 28.3% of the BC2 reads matching to entries in their respective 384 barcode whitelist (figure 4.8). Since both halves are needed for a full barcode, the number of full barcodes is lower - 23.6% of the BC1 + BC2 combinations matched the master whitelist of 384 x 384 combinations.

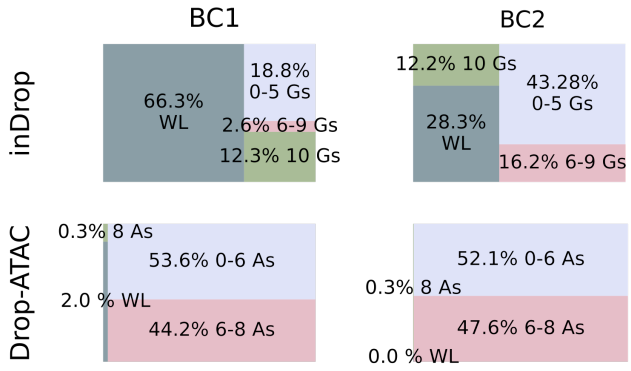


Figure 4.8: Re-sequencing run inDrop and Drop-ATAC barcode distribution. The original inDrop and single cell Drop-ATAC libraries were re-sequenced on the NextSeq 500, but at reduced clustering density. The inDrop library showed a strongly improved fraction of reads mapping to the barcode whitelist, but the Drop-ATAC library did not.

As for the Drop-ATAC library, the new sequencing data did not provide any more information. In fact, whereas only the second barcode half had a poly-A problem in the

first sequencing run, both halves showed poly-A reads in the second sequencing run. In a sense, the poly-A problem can be seen as the inverse of the poly-G problem, as A is encoded by the combination of red and green signal on the NextSeq platform (figure 4.6). As of now, we have no explanation for the poly-A reads, given that they are not "true" reads as indicated by the Sanger sequencing, and that they were not present in the first sequencing run of the same library. At any rate, so far, we have been unsuccessful in retrieving barcodes from the Drop-ATAC system due to the sequencing errors we experienced. The logical next step is to re-synthesize the BHB barcodes using Illumina's TruSeq adapters instead of our custom read primers. Such a preliminary trial was performed for the inDrop library, and will be discussed in the following section.

As mentioned in chapter 2, we systematically overloaded the inDrop run with cells to ensure an RNA-seq signal was present during the library preparation stages. This approach leads to severe undersequencing of the single cell transcriptomes. The number of cells with more than 1000 UMI counts (corresponding to 1000 unique transcripts captured) is only 699, and only 283 cells detected more than 750 genes. The mean number of UMIs was 222, and the mean number of genes detected was 158. These metrics are low compared to what can be expected from a 10x Chromium scRNA-seq run (figure 4.9), but should improve when the sequencing depth per cell is increased. The general approach for selecting cells for further analysis is to find a steep drop in UMI count in the barcode rank plot (figure 4.10). This sharp drop indicates the transition from true cells to noise associated with barcodes released into droplets without cells. In our dataset, no such steep drop was observed until the very end, which indicates that there is a gradual decline in UMI count. We can again explain this phenomenon by the cell overloading. Clearly, cell loading needs to go down in the future to increase per-cell sequencing depth.

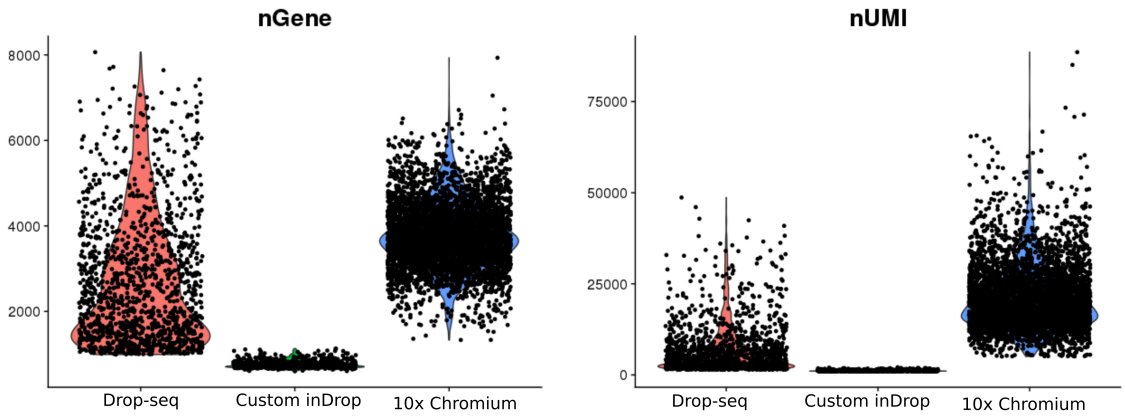


Figure 4.9: Gene and UMI count comparison. Our method underperformed on the single cell level compared to Drop-seq and 10x Chromium scRNA-seq.

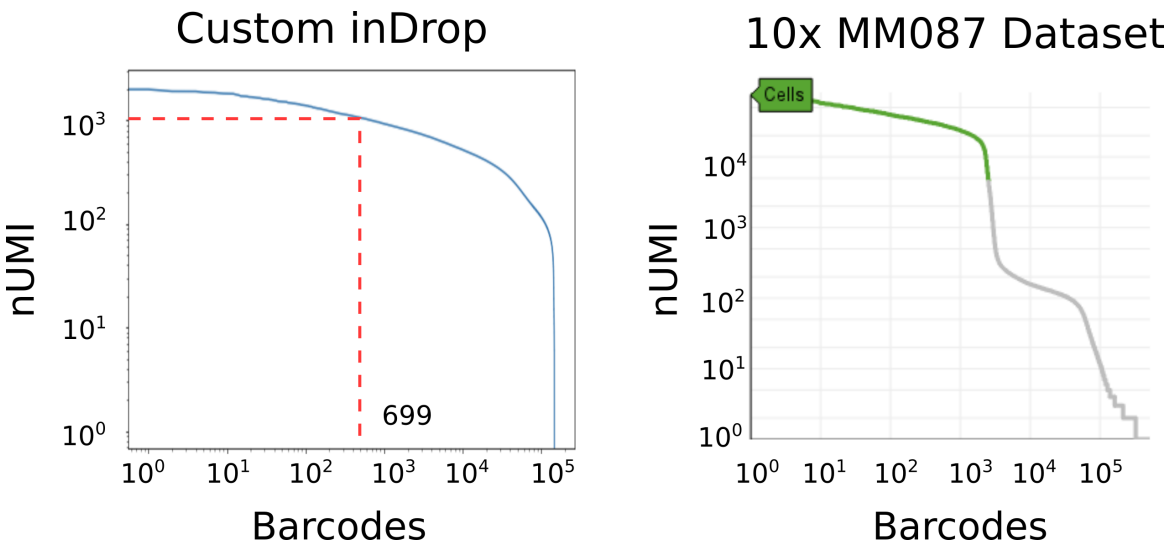


Figure 4.10: Barcode rank plots. Barcodes are ranked by the number of UMIs they were associated with in the dataset. The steep decline in the 10x Chromium MM087 dataset indicates the transition from "true" cells to droplets containing only a barcoded hydrogel bead, and no cell.

We were interested to see if the new inDrop library showed any resemblance to previous scRNA-seq datasets from the same MM087 cell line. We therefore performed canonical correlation analysis (CCA) on the 699 inDrop cells that had a UMI count of $>$ 1k combined with MM087 Drop-seq data and 10x Chromium scRNA-seq data from several MM cell lines (figure 4.11). The Seurat CCA algorithm attempts to find shared sources of variation in gene expression (canonical correlation) between the different samples. The canonical components that explain most variation in the data are selected and further "aligned" (normalised) so they can be directly compared. The dimension-reduced correlation is then visualised using t-distributed stochastic neighbour embedding (tSNE), which further reduces the aligned canonical correlation components until two dimensions remain, which can be plotted. The resulting plot shows a systematic clustering of cells that share common sources of variation. We hypothesized that our MM087 scRNA-seq dataset would cluster together with the 10x and Drop-seq datasets of the same cell line, but not with the other cell lines. In essence, this would mean that the datasets from the three techniques yield the same information. As is evident in the graph, this was not the case. Our cells mixed together with the Drop-seq dataset in the middle of the plot. This indicates that the inDrop and Drop-seq data did not contain enough information to be co-clustered with the 10x MM087 dataset. The lack of information is also evident in figure 4.9.

We then briefly assessed the bulk properties of the re-sequenced library. We took all the datasets used in the previous analysis and removed the cell identifiers, transforming them from single cell datasets to bulk datasets. We then performed a regular correlation analysis

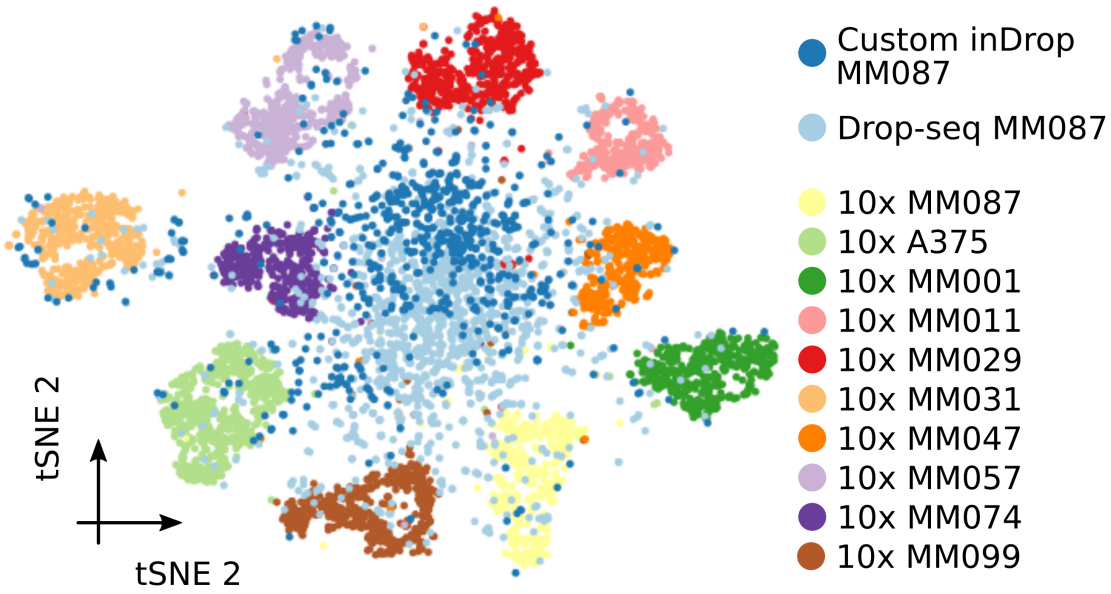


Figure 4.11: MM087 canonical correlation analysis. Our custom inDrop dataset does not co-cluster with the 10x Chromium MM087 dataset.

between all the datasets and found that the inDrop dataset was in in fact significantly more correlated with the 10x MM087 dataset ($p < 0.01$, figure S.3).

4.4 Reduced Complexity inDrop Repetition

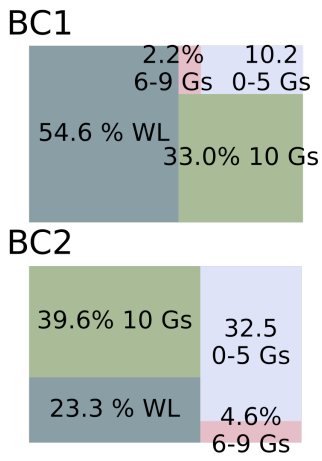


Figure 4.12: 4 x 4 inDrop library barcode distribution.

The entire process of producing and barcoding hydrogel, and using them in the custom inDrop protocol was repeated on MM074 cells. This time, we replaced the adapter sequence between both barcode halves with Illumina's standard TruSeq read primer sequence, allowing us to sequence the barcodes using well-defined and tested read primers. In order to save time and cost, only 4 x 4 different barcode combinations were used. The new 4 x 4 library was sequenced together with the previous 384 x 384 library on the same NextSeq 500 run and shared many of its improved sequencing properties. We were able to match a higher number of barcode reads to their respective whitelist, but there was again a high number of pure G-reads. We did not explore the single cell properties of this library since it would only be able to separate 16 virtual "cells".

The reduced complexity inDrop run, together with the Sanger sequencing, confirmed our suspicions that many of the problems we encountered could be mitigated by using standard sequencing primers. Whereas the dual barcode read system was originally put in place to assign more sequencing cycles to the cDNA, we now realise that this approach causes more problems than it solves. In the current system, the fragment is re-sequenced after the first barcode half is read, and the second half of the barcode is read on the newly synthesized complement. Physically, the two strands that are being read are not the same, which manifests itself as a steep drop in quality between both reads. We are therefore thinking about a redesigned system with a shorter adapter sequence in-between both barcodes, allowing the barcode to be sequenced in a single read. The problem of the poly-G reads remains, and can most likely be reduced further, but never completely. Other single cell sequencing techniques also filter out a large portion of the reads. A last resort solution could be to use a 4-channel sequencing system such as the Illumina MiSeq. In this system, every base is encoded by its own fluorescence signal, meaning we could discern true G reads from an absence of signal.

4.5 MinION Sequencing

The final measure we used to extract the last drop of information from the original 384 x 384 inDrop library was Oxford Nanopore MinION sequencing. The MinION sequencer works by pulling single DNA molecules through a nanopore and inferring base calling information from the change in electrical resistance over the pore. It can sequence long reads (up to 200 kbp), but with low single-base accuracy (Bowden et al., 2019). We ran the MinION sequencer for 12 hours, during which the 1200 functional pores generated 276k reads which passed MinKNOW base-calling software's internal filter. The individual base quality is low - the mean read quality was 8.6, and from the 126 megabases read, only 17% had a Phred quality score of > 10 (corresponding to a base call confidence of 90%). A 1 in 10 chance for incorrect base calls is too low to attempt to match barcodes to the whitelist. Figure 4.13 shows the average average quality and length of each read. The mean read length of 457 bp corresponds to what we expected from our library's electropherogram. The longer length band that separates from the rest of reads is most likely a remnant of a CHEQ-seq library which was sequenced on the MinION device prior to our run. The MinION flow-cell can be reused multiple times, but will suffer from sample carry-over in doing so.

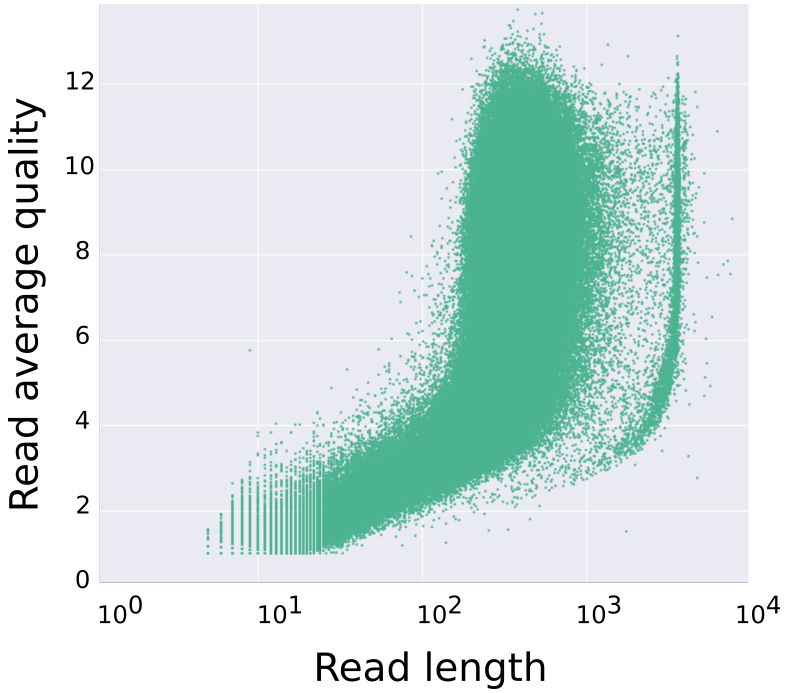


Figure 4.13: MinION read length and quality. "Read average quality" denotes the average base calling quality of each single read, expressed as a Phred quality score.

Since Nanopore sequencing yields full length fragments, we had to extract the barcode locations ourselves. For this task, we used the `cutadapt` adapter trimming tool (Martin, 2011), which can (partially) match and find adapter sequences and manipulate the sequences attached to them. First, we removed all the reads that did not contain the bead primer sequence, which already eliminated 63% of the reads. We then looked at only those reads that also contained the custom read site and the poly-T tail. All reads now contain the sequences that flank both barcode halves, meaning we could now look at the length of the barcodes. We found that there was a general adherence to the expected size, with 28% of the BC1 sequences being exactly 10 bp, and 50% of BC2 + UMI being exactly 18 bp in length, corresponding to the 10 bp barcode half and 8 bp UMI (figure 4.14). The MinION sequencer is prone to both insertions and deletions, so some deviation is to be expected. We did not observe any such insertions or deletions in the Sanger sequencing data. We also observed that the minION sequencing did not yield any significant poly-G stretches, once again confirming that the poly-G reads produced by the NextSeq 500 were not true reads.

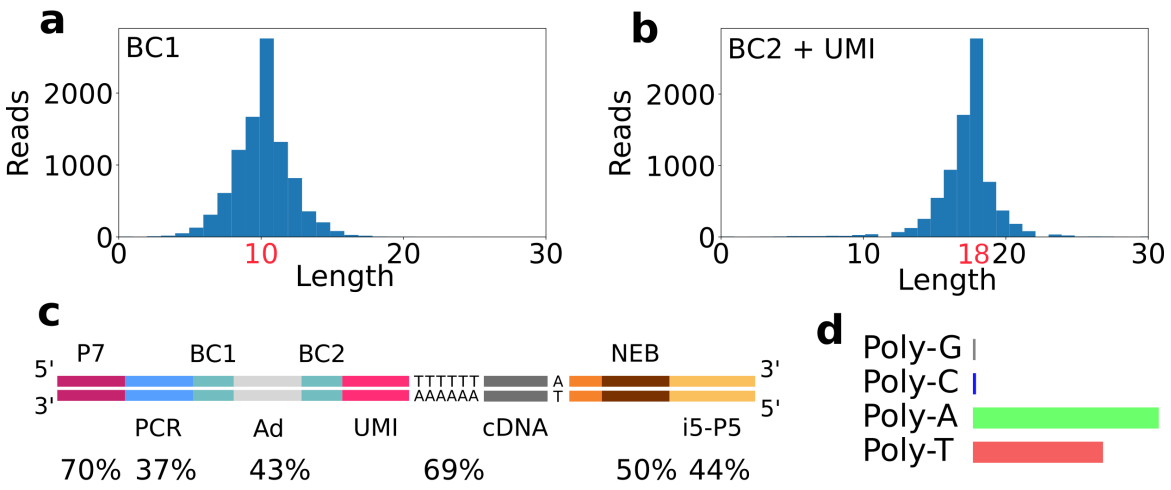


Figure 4.14: MinION sequencing barcode feature lengths distributions. **(a, b)** Distribution of stretch between PCR site and Adapter (BC1) and between Adapter and Poly-T/A (BC2+UMI). **(c)** Idealised fragment, percentage indicates number of MinION reads that have that particular feature. **(d)** Fraction of fragments that have at least a single Poly-N sequence, defined as an uninterrupted stretch of 10 identical nucleotides.

5 | Conclusion and Future

While our custom inDrop and Drop-ATAC framework did not immediately produce the single-cell data we envisioned, our preliminary results form a strong foundation for future experiments. We were successful in producing functional barcoded hydrogel beads, a droplet microfluidic cell encapsulation system and both a working RNA-seq and ATAC-seq reaction in droplets. In terms of bulk assay properties, both our Drop-seq/inDrop hybrid and the Drop-ATAC prototype showed desirable qualities. Furthermore, we have a number of strong indications that additional fine-tuning of the cell barcodes and sequencing process will bring us to a completely functional scRNA-seq and/or scATAC-seq platform.

The next step is to redesign the barcoding process, most likely rolling back the dual-read system to a single-read process or opting for the Illumina TruSeq approach. On the physical level, we want to introduce disulfide linkers in the hydrogel matrix itself in order to dissolve the bead and thereby increase diffusion within the droplet. Furthermore, there is a clear need for optimisation of the single-cell suspension to prevent cell aggregation, which impedes true single-cell measurements. Lastly, we aim to automate several of the steps in the protocol using automated liquid handlers in the future.

In this work, we started on the physical level by producing well-defined barcoded hydrogel beads and setting up a microfluidic system, transitioned into the molecular-biological sphere when optimising our RNA-seq and ATAC-seq assays, and finally ended in the computational domain when analysing the resulting sequencing data. In a span of just 8 months, we were able to set up a complete microfluidic/molecular workflow, showing that Whitesides' original projection has aged well. We have now arrived at a point where graduate students can, under close and careful supervision, independently produce and operate microfluidic systems and put them to use in a rapidly advancing field such as single-cell technology. The field advances rapidly indeed - shortly before this work was submitted, two new droplet-based ATAC-seq techniques were published by the Chang and Buenrostro labs (Satpathy et al., 2019; Lareau et al., 2019). Both techniques are being pushed to the market as ready to use packages, showing the translation potential of single cell technology.

The demand for and application potential of single-cell technology is high, but it will take more innovation and large-scale co-operation to get there. With the experience and knowledge gained in this project, our lab is now well-equipped to be part of this effort and help advance single-cell technology to its full potential. Once the barcoding issue is solved, we will look forward to combine scRNA-seq and scATAC-seq into a single droplet-based assay. This hypothetical multi-omic technique could find direct application in large-scale CRISPR perturbation screening assays, where the RNA-seq modality can retrieve the exact nature of the CRISPR gene edit, and the ATAC modality can provide valuable

information on that edit's impact on gene regulation. Developing such a combined, multi-modal technique poses several technical challenges, notably concerning the preservation of cellular identity throughout the multi-step protocol that will most likely be required.

Bibliography

Abate, A. R., C. . Chen, J. J. Agresti, and D. A. Weitz
2009. Beating poisson encapsulation statistics using close-packed ordering. *Lab on a Chip*, 9(18):2628–2631.

Ahn, K., J. Agresti, H. Chong, M. Marquez, and D. A. Weitz
2006a. Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. *Applied Physics Letters*, 88(26).

Ahn, K., C. Kerbage, T. P. Hunt, R. M. Westervelt, D. R. Link, and D. A. Weitz
2006b. Dielectrophoretic manipulation of drops for high-speed microfluidic sorting devices. *Applied Physics Letters*, 88(2):1–3.

Allazetta, S. and M. P. Lutolf
2015. Stem cell niche engineering through droplet microfluidics. *Current opinion in biotechnology*, 35:86–93.

Azizi, F., S. Clouthier, and M. S. Wicha
2014. The promise of single cell omics for onco-therapy. *Journal of Molecular and Genetic Medicine*, 8(3):121–122.

Bannister, A. J. and T. Kouzarides
2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395.

Bengtsson, M., A. Ståhlberg, P. Rorsman, and M. Kubista
2005. Gene expression profiling in single cells from the pancreatic islets of langerhans reveals lognormal distribution of mrna levels. *Genome research*, 15(10):1388–1392.

Bowden, R., R. W. Davies, A. Heger, A. T. Pagnamenta, M. D. Cesare, L. E. Oikkonen, D. Parkes, C. Freeman, S. Y. Patel, N. Popitsch, C. L. C. Ip, H. E. Roberts, G. Lunter, J. C. Taylor, D. Buck, and M. A. Simpson
2019. Sequencing of human genomes with nanopore technology. *Nature Communications*, Pp. 1–9.

Brouzes, E., M. Medkova, N. Savenelli, D. Marran, M. Twardowski, J. B. Hutchison, J. M. Rothberg, D. R. Link, N. Perrimon, and M. L. Samuels
2009. Droplet microfluidic technology for single-cell high-throughput screening. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14195–14200.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf
2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218.

Buenrostro, J. D., B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf
2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

Cao, J., D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell, and J. Shendure
2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385.

Cao, J., J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, and J. Shendure
2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667.

Chen, X., R. J. Miragaia, K. N. Natarajan, and S. A. Teichmann
2018. A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*, 9(1).

Clark, S. J., R. Argelaguet, C. . Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik
2018. Scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells e. *Nature Communications*, 9(1).

Corces, M. R., A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf, and H. Y. Chang
2017. An improved atac-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, 14(10):959–962. Cited By :43.

Cusanovich, D. A., R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure
2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914.

Davie, K., J. Janssens, D. Koldere, M. De Waegeneer, U. Pech, Å. Kreft, S. Aibar, S. Makhzami, V. Christiaens, C. Bravo GonzÃąlez-Blas, S. Poovathingal, G. Hulselmans,

K. I. Spanier, T. Moerman, B. Vanspauwen, S. Geurs, T. Voet, J. Lammertyn, B. Thienpont, S. Liu, N. Konstantinides, M. Fiers, P. Verstreken, and S. Aerts
2018. A single-cell transcriptome atlas of the aging drosophila brain. *Cell*, 174(4):982–998.e20.

Eberwine, J., H. Yeh, K. Miyashiro, Y. Cao, S. Nair, R. Finnell, M. Zettel, and P. Coleman
1992. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 89(7):3010–3014.

Fan, H. C., G. K. Fu, and S. P. A. Fodor
2015. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222).

Farrell, J. A., Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier
2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392).

Fedurco, M., A. Romieu, S. Williams, I. Lawrence, and G. Turcatti
2006. Bta, a novel reagent for dna attachment on glass and efficient generation of solid-phase amplified dna colonies. *Nucleic acids research*, 34(3).

Fiers, M. W. E. J., L. Minnoye, S. Aibar, C. B. González-Blas, Z. K. Atak, and S. Aerts
2018. Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, 17(4):246–254.

Fluidigm Website
2019. Fluidigm Website - https://www.fluidigm.com/publications/c1.

Gaulton, K. J., T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K. L. Mohlke, J. D. Lieb, and J. Ferrer
2010. A map of open chromatin in human pancreatic islets. *Nature genetics*, 42(3):255–259.

Gierahn, T. M., M. H. Wadsworth, T. K. Hughes, B. D. Bryson, A. Butler, R. Satija, S. Fortune, J. Christopher Love, and A. K. Shalek
2017. Seq-well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods*, 14(4):395–398.

Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb
2007. Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6):877–885.

Goos, P. and B. Jones
2011. *Optimal design of experiments a case study approach*. Wiley.

Greenleaf, W. J. and A. Sidow
2014. The future of sequencing: Convergence of intelligent design and market darwinism. *Genome biology*, 15(3).

Grün, D., L. Kester, and A. Van Oudenaarden
2014. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640.

Grün, D. and A. Van Oudenaarden
2015. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810.

Harris, T. D., P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. DiMeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss, and Z. Xie
2008. Single-molecule dna sequencing of a viral genome. *Science*, 320(5872):106–109.

Hashimshony, T., 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
2016. Cel-seq2: Sensitive highly-multiplexed single-cell rna-seq. *Genome biology*, 17(1).

Hashimshony, T., F. Wagner, N. Sher, and I. Yanai
2012. Cel-seq: Single-cell rna-seq by multiplexed linear amplification. *Cell Reports*, 2(3):666–673.

Hood, L., J. R. Heath, M. E. Phelps, and B. Lin
2004. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696):640–643.

Hou, Y., H. Guo, C. Cao, X. Li, B. Hu, P. Zhu, X. Wu, L. Wen, F. Tang, Y. Huang, and J. Peng
2016. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26(3):304–319.

Huebner, A., M. Srisa-Art, D. Holt, C. Abell, F. Hollfelder, A. J. DeMello, and J. B. Edel
2007. Quantitative detection of protein expression in single cells using droplet microfluidics. *Chemical Communications*, 12:1218–1220.

Jaenisch, R. and A. Bird
2003. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature genetics*, 33(3S):245–254.

Jaitin, D. A., E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit
2014. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

Johannes, F., V. Colot, and R. C. Jansen
2008. Epigenome dynamics: A quantitative genetics perspective. *Nature Reviews Genetics*, 9(11):883–890.

Kivioja, T., A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale
2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74.

Klein, A. M., L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner
2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Kolodziejczyk, A. A., J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann
2015. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620.

Kornberg, R. D.
1974. Chromatin structure: A repeating unit of histones and dna. *Science*, 184.

Kouzarides, T.
2007. Chromatin modifications and their function. *Cell*, 128(4):693–705.

La Manno, G., R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko
2018. Rna velocity of single cells. *Nature*, 560(7719):494–498.

Lareau, C. A., F. M. Duarte, J. G. Chew, V. K. Kartha, Z. D. Burkett, A. S. Kohlway, D. Pokholok, M. J. Aryee, F. J. Steemers, R. Lebofsky, and J. D. Buenrostro
2019. Droplet-based combinatorial indexing for massive scale single-cell epigenomics. *bioRxiv*.

Macosko, E. Z., 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
2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Martin, M.
2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12.

Mazutis, L., J. Gilbert, W. L. Ung, D. A. Weitz, A. D. Griffiths, and J. A. Heyman
2013. Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, 8(5):870–891.

Mills, J. D., Y. Kawahara, and M. Janitz
2013. Strand-specific rna-seq provides greater resolution of transcriptome profiling. *Current Genomics*, 14(3):173–181.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold
2008. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7):621–628.

Patterson, S. D. and R. H. Aebersold
2003. Proteomics: The first decade and beyond. *Nature genetics*, 33(3S):311–323.

Patti, G. J., O. Yanes, and G. Siuzdak
2012. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269.

Picelli, S.
2017. Single-cell rna-sequencing: The future of genome biology is now. *RNA Biology*, 14(5):637–650.

Picelli, S., Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg
2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1100.

Picelli, S., Å. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg
2014a. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research*, 24(12):2033–2040.

Picelli, S., O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg
2014b. Full-length rna-seq from single cells using smart-seq2. *Nature Protocols*, 9(1):171–181.

Ramsköld, D., S. Luo, Y. . 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
2012. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777–782.

Regev, A., S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea,

G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. Van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, and H. Participants
2017. The human cell atlas. *eLife*, 6.

Rosenberg, A. B., C. M. Roco, R. A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. T. Graybuck, D. J. Peeler, S. Mukherjee, W. Chen, S. H. Pun, D. L. Sellers, B. Tasic, and G. Seelig
2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182.

Satpathy, A. T., J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, P. Shah, J. C. Bell, D. Jhutty, C. M. Nemec, J. Wang, L. Wang, Y. Yin, P. G. Giresi, A. L. S. Chang, G. X. Zheng, W. J. Greenleaf, and H. Y. Chang
2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. *bioRxiv*.

Schones, D. E. and K. Zhao
2008. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9(3):179–191.

Shalek, A. K., R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May, and A. Regev
2014. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369.

Song, L. and G. E. Crawford
2010. Dnase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 5(2).

Song, L., Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B. . Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey
2011. Open chromatin defined by dnasei and faire identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10):1757–1767.

Streets, A. M., X. Zhang, C. Cao, Y. Pang, X. Wu, L. Xiong, L. Yang, Y. Fu, L. Zhao, F. Tang, and Y. Huang
2014. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19):7048–7053.

Stuart, T. and R. Satija
2019. Integrative single-cell analysis. *Nature Reviews Genetics*.

Tabula Muris Consortium
2018. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372.

Tang, F., 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
2009. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Tang, F., K. Lao, and M. A. Surani
2011. Development and applications of single-cell transcriptome analysis. *Nature Methods*, 8(4 SUPPL.):S6–S11.

The Human Cell Atlas Consortium
2017. The human cell atlas whitepaper.

Wang, D. and S. Bodovitz
2010. Single cell analysis: The new frontier in 'omics'. *Trends in biotechnology*, 28(6):281–290.

Wang, Z., M. Gerstein, and M. Snyder
2009. Rna-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Whitesides, G. M.
2006. The origins and the future of microfluidics. *Nature*, 442(7101):368–373.

Wu, A. R., 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
2014. Quantitative assessment of single-cell rna-sequencing methods. *Nature Methods*, 11(1):41–46.

Xia, Y. and G. M. Whitesides
1998. Soft lithography. *Angewandte Chemie - International Edition*, 37(5):550–575.

Yuan, J. and P. A. Sims
2016. An automated microwell platform for large-scale single cell rna-seq. *Scientific Reports*, 6.

Ziegenhain, C., B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard
2017. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643.e4.

Žilionis, R., J. Nainys, A. Veres, V. Savova, D. Zemmour, A. M. Klein, and L. Mazutis
2017. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1):44–73.

Materials and Methods

M.1 Manufacturing Barcoded Hydrogel Beads

M.1.1 Required Buffers

Several buffers and solutions are needed for completion of the following protocols, all of which are filtered through a 0.2 µm membrane after mixing (Žilionis et al., 2017).

Tris-buffered EDTA Triton (TBSET) buffer	
Nuclease-free H_2O	822 ml
1 M NaCl	137 ml
0.5 M EDTA	20 ml
1 M Tris-HCl (pH 8.0)	10 ml
10% (v/v) Triton X-100	2.58 ml
2 M KCl	1.35 ml

Tris-EDTA-Tween (TET) buffer	
Nuclease-free H_2O	480 ml
0.5 M EDTA	10 ml
1 M Tris-HCl (pH 7.0)	5 ml
10% Tween-20	5 ml

STOP-25	
Nuclease-free H_2O	880 ml
0.5 M EDTA	50 ml
2 M KCl	50 ml
1 M Tris-HCl (pH 8.0)	10 ml
10% (v/v) Tween-20	10 ml

4x Acrylamide/bis-acrylamide	
Nuclease-free H_2O	3.82 ml
40% (w/w) Acrylamide/bis-acrylamide (AA/BIS)	3.6 ml
40% (w/w) Acrylamide (AA)	2.58 ml

STOP-10	
Nuclease-free H_2O	1820 ml
2 M KCl	100 ml
0.5 M EDTA	40 ml
1 M Tris-HCl (pH 8.0)	20 ml
10% (v/v) Tween-20	20 ml

Denaturation solution	
Nuclease-free H_2O	242 ml
10 M NaOH	3.75 ml
30% (w/w) Brij-35	4.2 ml

Neutralisation buffer	
Nuclease-free H_2O	385 ml
1 M Tris-HCl (pH 8.0)	350 ml
1 M NaCl	50 ml
0.5 M EDTA	10 ml
10% (v/v) Tween-20	5 ml

Hybridisation buffer	
Nuclease-free H_2O	407 ml
2 M KCl	82.5 ml
1 M Tris-HCl (pH 8.0)	5 ml
0.5 M EDTA	100 µl

QC Buffer	
Nuclease-free H_2O	24 ml
2 M KCl	25 ml
0.5 M EDTA	500 µl
1 M Tris-HCl (pH 8.0)	250 µl
10% (v/v) Tween-20	250 µl

M.1.2 Producing Microfluidic Chips

Silicon wafers were imprinted with an SU-8 mask by a senior lab scientist according to standard protocols (Xia and Whitesides, 1998). Then, a mixture of 10 parts polydimethylsiloxane (PDMS) and 1 part of curing agent (SYLGARD^{TM}) was prepared, mixed well and degassed under vacuum for 45 min to remove air bubbles. The bubble-free PDMS was poured into removable PMMA dam surrounding every single mask on the chip at a rate of ~12 g per chip, leading to an average height of ~1 cm. The PDMS was cured on-chip by heating at 80 °C for 2 hours. The PDMS slabs were then cooled and removed from the Si wafer. Inlets were punctured using 1 mm diameter biopsy needles. A glass slide was cleaned with acetone and IPA and plasmatreated together with the chip for 30 s. After plasma treatment, the chip was pressed onto the glass slide. Immediately after, the channels on the chip were flushed with 2% trichloro(1H, 1H, H, 2H-perfluorooctyl)silane in HFE-7500 (3M^{TM}) to make the channel surface hydrophobic. The chips were then baked for an additional 4 hours at 80 °C and further stored with the channels taped shut.

M.1.3 Generating Hydrogel Microspheres

Hydrogel microspheres were produced according to Žilionis et al. (2017). Emulsion oil was prepared by spiking EvaGreen QX200^{TM} fluorinated oil with 2% TEMED (v/v). The emulsion oil was fixed into a CETONI 290N neMESYS syringe pump and connected to the hydrogel bead generation chip using 1 mm diameter tubing. The following monomer mix prepared was loaded into a syringe, with a thin layer of HFE to nullify dead volume, and the syringe was similarly fixed into the syringe pump and connected to the chip.

Monomer Mix	
100 µM Acrydite DNA primer	500 µl
4x AB solution	250 µl
Nuclease-free H_2O	120 µl
TBSET	100 µl
APS	30 µl

Both the TEMED-spiked oil and monomer mix were then pump through the microfluidic chip at the rates calculated in section 2.2. The resulting emulsion was collected under a layer of protective mineral oil and incubated overnight at 65 °C.

Flow Rates	
Monomer Mix	1500 µl h^{-1}
Oil	1461 µl h^{-1}

The hydrogel beads must now be washed thoroughly to remove all traces of the oil phase. After removing the mineral and carrier oil phases, an equal volume of TBSET was added, and the emulsion was broken with 20% (vol/vol) PFO. The PFO-phase was removed and this step was repeated until the beads appeared to be semi-transparent. After adding an equal volume of 1% SPAN-80 in hexane, the beads were vortexed and then centrifuged at 5 000 g for 30 s. The top hexane layer was then removed and this step was repeated. The beads were then transferred to 15 ml tubes and topped off with TBSET. After centrifugation, the hexane traces were removed using a syringe. This step was repeated an additional 3 times, after which the beads were measured using the Python script (S.1.1) and stored for barcoding.

M.1.4 Bead Diameter Model

A blocked, I-optimal response surface design was generated for 8 samples and 2 blocks using SAS JMP. 2 out of the 8 flow combinations did not physically generate droplets, so they were omitted from the dataset, and the design was augmented with an additional 2 blocks of 3 runs each was added with the added parameter constrain that $Q_{oil} > 1.5 \times Q_{mono}$. Beyond this constraint, the monomer flow becomes too strong for the oil flow to generate a proper emulsion. Each block of runs was performed on a set of different microfluidic channels to account for variation in the chip. The 12 runs resulted in 12 batches of emulsions with various droplet sizes, which were individually washed and photographed under a microscope. A Python image analysis script (S.1.1) was used to calculate the bead sizes of the beads in the images. Finally, a number of different modelling approaches were tested and in-silico validated on previous data. The best model resulted from REML regression estimation of independent quadratic models for diameter and bead size standard deviation on the non-averaged dataset. Sample 16 was omitted as it was an outlier. Significant model factors were selected using mixed stepwise selection. The diameter was then fixed and used as a non-linear constraint in a minimisation problem for bead standard deviation. The solution of this minimisation problem is the set of parameter combinations (flows) that will produce beads of a given diameter, and with the least possible deviation from the mean. The resulting data is shown in table S.1 and visualised in figure S.1.

M.1.5 Barcoding Hydrogel Beads

Beads were washed three times in HBW. Washing comprises filling the 15 ml tubes with HBW, vortexing, centrifuging at 15 g for 3 min and removing the supernatant. Per ml of beads, 0.241 ml of Isothermal Amplification Buffer was added. The beads were vortexed, centrifuged at 1000 g, and supernatant was taken off. Due to increased ionic strenght, the HBs have now shrunk to ~80% of the original volume.

HBs were then mixed into a hydrogel bead mix as follows:

Hydrogel Bead Mix	
H_2O	2.3 ml
HB suspension	9.4 ml
10 mM dNTPs	0.812 ml

The hydrogel bead mix was then aliquoted into a 96-well plate at 125 µl per well. The first set of barcoded primers (4 96-well plates) was then thawed, centrifuged for 1 min at 1000 g and placed in a PCR machine for denaturation (Lid at 105 °C, 70 °C for 40 s followed by 4 °C for 20 °C). Extra care must be taken when removing the seal to avoid cross-contamination.

4 reaction plates were prepared by mixing 30 µl of hydrogel bead mix with 13.5 µl of one of the 384 primers. These plates were then incubated on heat blocks at 85 °C for 2 min, 60 °C for 20 min with a lid temperature of 100 °C. During this step, the primers hybridise to the acrydite stubs in the hydrogel matrix.

After the hybridisation has ended, 15 µl of isothermal reaction mix (prepared and stored on ice) was added to every well on the reaction plates.

Isothermal Reaction Mix	
nuclease-free H_2O	6.85 ml
10x isothermal amp. buffer	0.8 ml
Bst 2.0 DNA 8000 U ml^{-1}	0.35 ml

Plates were then sealed and incubated on heat blocks at 60 °C for 60 min and 4 °C indefinitely with a lid temperature of 100 °C. During this step, the acrydite primer is extended with the barcoded primer by Bst 2.0 isothermal amplification.

The plates were then cooled on ice, and 40 µl of STOP-25 buffer was added to each reaction well to stop the isothermal amplification reaction. The contents of all wells were pooled and wells were rinsed with an additional 100 µl of STOP-25. The pooled product was incubated in a rotator at 20 rpm for 30 min.

The half-barcoded beads now need to be washed thoroughly to remove all traces of unused primers, salts, dNTPs and enzymes before the second round of barcoding can take place. The half-barcoded bead stock was centrifuged at 300 g for 15 min, and the supernatant was collected and centrifuged again to collect all beads. The beads were then split equally into four separate 15 ml tubes and washed three times with STOP-10 buffer. Washing encompasses filling the tubes with STOP-10 buffer, rotating for 10 min at room temperature, centrifuging at 300 g for 3 min and removing the supernatant. The half-barcoded HBs were then washed 4 times in denaturation solution, 2 times in neutralisation buffer and finally 2 times in TET buffer, after which the beads

were stored in 4 °C overnight for the next barcoding step.

The half-barcoded beads were then washed 3 times in HBW, and processed for a second round of barcoding as described above - but with the Barcode 2 primer plates instead of barcode 1 - and washed using the same washing steps. The resulting stock of barcoded beads is stored at 4 °C.

M.1.6 qPCR

7 samples were prepared in triplicate with 1.25 µl of packed BHBs and incubated in the appropriate concentration of DTT. At a total volume of 4 µl per sample, the concentration of beads in the DTT mix was equal to the "concentration" of beads in a true droplet (taken as the inverse of the droplet volume per bead).

Sample	Bead	H_2O	0.25 M DTT
0 mM	2.75 µl	1.25 µl	0 µl
10 mM	2.59 µl	1.25 µl	0.16 µl
25 mM	2.35 µl	1.25 µl	0.4 µl
50 mM	1.95 µl	1.25 µl	0.8 µl
75 mM	1.55 µl	1.25 µl	1.2 µl
100 mM	1.15 µl	1.25 µl	1.6 µl
125 mM	0.75 µl	1.25 µl	2 µl

The beads were incubated for 30 minutes, after which they were centrifuged (5000 rcf, 2 min), and the supernatant was processed for qPCR using ready-to-use hot-start PCR mix containing FastStart Taq DNA polymerase, reaction buffer, dNTP mix with dUTP instead of dTTP, Sybr Green I dye and $MgCl_2$.

qPCR mix	
FastStart Taq mix	5 µl
Bead supernatant	2.5 µl
Ultra-pure H_2O	1.5 µl
10 µM PCR primer	1 µl

The following qPCR program was used on the LightCycler 480:

ISPCR Program		
	95 °C	5 min
45 cycles of:	95 °C	10 s
	72 °C	35 s
Melting curve:	95 °C	5 s
	65 °C	1 min.
	97 °C	cont.
	40 °C	10 s

M.2 inDrop Optimisation

The optimisation trials followed the same steps as the third and final protocol, which is described in greater detail M.3, but stopped where failure occurred either physically on the chip or molecularly by not yielding cDNA.

M.2.1 Bead/Cell Loading Experiments

180 µl of bead stock was washed twice using HBW and dissolved into 300 µl of 2X Lysis buffer. Beads were then loaded onto the chip, with cell flow replaced with PBS. $Q_{bead}, Q_{cell}, Q\_oil$, were each varied around the limits of 150 µl/h and 1000 µl/h. Bead coverage was observed by taking pictures of the suspension in the outlet funnel, as well as of the resulting suspension on a glass slide. Cell loading trials were performed by similarly testing different concentrations of input cell suspension (300/µl, 600/µl and 1200/µl. To achieve an ultra-pure bead solution, beads were run through an inDrop chip, replacing both cell and oil flow with TET buffer. When dust stopped the flow, the process was stopped and continued on a fresh chip.

M.2.2 First Custom inDrop Trial

MM074 cells were resuspended in the following master mix and co-encapsulated with lysis buffer and BHBs on the inDrop system. Final cell concentration of the master mix was 600 µl, and flow rates of 300, 200 and 450 µl/h were used for cells, beads and oil respectively.

RT Mix	
5x RT Buffer	36 µl
40% PEG-8000 in H_2O	33.75 µl
H_2O	13 µl
10 µM dNTPs	12 µl
200 U µl^{-1} Maxima hRT	9 µl
4.29 M DTT	6.3 µl
40 U µl^{-1} RNAse Inhibitor	4.5 µl
100 µM TSO	4.5 µl

The sample emulsions were collected, aliquoted into a 96-well PCR plate and heat-cycled for reverse transcription. Half of the sample was then Exo I treated according to standard Exonuclease clean-up instructions. After a subsequent ISPCR using Terra polymerase, the libraries were purified using 0.6 x volume Ampure beads and Qubit-quantified.

Terra ISPCR Master Mix	
2x Terra Direct PCR buffer	25 µl
Purified cDNA Library	20 µl
100 µM SMART PCR Primer	2.5 µl
Terra DNA polymerase	2 µl
Ultrapure H_2O	0.5 µl

ISPCR Program	
98 °C	3 min
14 cycles of: 98 °C	20 s
65 °C	45 s
72 °C	4 min
72 °C	10 min
4 °C	inf.

ISPCR Master Mix	
2x KAPA HiFi PCR Mix	50 µl
Purified cDNA Library	30 µl
100 µM SMART PCR Primer	2.5 µl
Ultrapure H_2O	17.5 µl

M.2.3 Second Custom inDrop Trial

MM087 cells were processed using the identical master mix (cell density 600 μl^{-1}) and inDrop parameters as the first trial, but split between two different polymerases for ISPCR. Terra polymerase was carried out as previously described, and the KAPA HiFi PCR was performed as follows:

ISPCR Program	
98 °C	3 min
4 cycles of: 98 °C	20 s
65 °C	45 s
72 °C	3 min
10 cycles of: 98 °C	20 s
67 °C	20 s
72 °C	3 min.
72 °C	10 min
4 °C	inf.

M.3 Final Custom inDrop Protocol

This final trial was first performed by a senior lab scientist and taken as standard protocol as it produced the most favourable library fragment size distribution. This protocol was repeated using the reduced complexity beads described in chapter 4.

M.3.1 Preparations

All the microfluidic peripherals were set up before preparing the cells, as the time between resuspending the cells into the RT master mix and starting cell encapsulation needs to be minimised. The PDMS microfluidic chip was fixed under a microscope, and a syringe was filled with oil and connected to the correct chip inlet via 1 mm tubing.

The BHB stock was centrifuged (1000 rcf, 3 min) and the supernatant was removed. 60 µl of condensed bead stock was added to 940 µl of HBW, vortexed well and centrifuged (2000 rcf, 1.5 min). The supernatant was removed, and replaced with an equal volume of HBW and the wash was repeated. Supernatant was removed, leaving 200 µl of condensed bead stock. The beads were incubated in an equal volume of 2x lysis buffer for 10 min on ice.

2x Lysis Buffer	
1 M Tris HCl	9 µl
5 M NaCl	1.8 µl
10% IGEPAL	20 µl
Nuclease-free H_2O	469.2 µl

The beads were then centrifuged (5000 rcf, 1 min) and supernatant was removed until 75 µl was left. The beads were then carefully centrifuged with a mini benchtop centrifuge to pull down beads sticking to the walls.

A single cell suspension was prepared by collecting sample tissue or culture and dissociating according to appropriate protocols. 70 000 cells were distributed into a 1.5 ml eppendorf, centrifuged (300 rcf, 5 min) and the supernatant was carefully removed. Cells were washed with ice-cold PBS and centrifuged (300 rcf, 5 min). Supernatant was taken, and the PBS wash was repeated. The resulting cell pellet was resuspended in 120 µl of reverse-transcription master mix:

RT Mix	
5x RT Buffer	57.6 µl
10 µM dNTPs	21.48 µl
200 U μl^{-1} Maxima hRT	14.28 µl
40 U μl^{-1} RNAse Inhibitor	7.2 µl
100 µM TSO	7.13 µl
4.29 µM DTT	6.72 µl
H_2O	5.59 µl

Cells are now ready to proceed to the inDrop run.

M.3.2 inDrop Run

Cell and bead mix were pre-loaded into 1 mm tubing by filling a syringe with PBS, connecting it to the tube, and carefully aspirating the solutions into the tube. The syringes were then fixed into a CETONI neMESYS

290N syringe pump, and tubing was connected to the correct inlet on the microfluidic chip. Similarly, the chip outlet was connected to a 1.5 ml eppendorf tube in a cold block.

All flows were simultaneously primed at 500 µl h^{-1} until they had all stabilised. Then, final flow rates were set as follows:

Flow rates	
Cell	800 µl h^{-1}
Beads	1100 µl h^{-1}
Oil	1200 µl h^{-1}

Bead and cell loading must be monitored continuously under the microscope and can be tuned by adjusting flow rates. Care must be taken not to inflate the droplets with either component during flow adjustment. Cell encapsulation ended after cells ran out, and the resulting emulsion was redistributed into a 96-well plate at 70 µl of emulsion with 30 µl of EvaGreen QX200™. Reverse transcription was then carried out according to the following program:

RT Program	
42 °C	90 min
85 °C	5 min
4 °C	inf.

M.3.3 Post-RT Clean-up and ISPCR

150 µl of droplet breaking solution was added to every well and mixed by pipetting.

Droplet Breaking Solution	
HFE 7500	700 µl
PFO	300 µl

The broken emulsions were then transferred into Eppendorf DNA lo-bind PCR tubes, and the bottom oil layer was removed and discarded. 125 µl of QIAGEN PB buffer was added to each PCR tube, and the solution was mixed well by pipetting 15 times. The contents were transferred into a 1.5 ml Eppendorf tube and centrifuged (5000 rcf, 1 min) and the supernatant was removed. The cDNA was then purified using QIAGEN MinElute columns according to manufacturer's instructions. Final elution was performed in 31 µl, and the concentration was measured using an Invitrogen Qubit fluorometer.

The purified cDNA library was then subjected to ISPCR using the following mix and PCR-program.

ISPCR Master Mix	
2x KAPA HiFi PCR Mix	50 µl
Purified cDNA Library	30 µl
100 µM SMART PCR Primer	6 µl
Ultrapure H_2O	14 µl

ISPCR Program	
98 °C	3 min
4 cycles of: 98 °C	20 s
65 °C	45 s
72 °C	3 min
10 cycles of: 98 °C	20 s
67 °C	20 s
72 °C	3 min.
73 °C	5 min
4 °C	inf.

The ISPCR-amplified cDNA library was then purified with 0.6 volumes of SPRI beads according to manufacturer's instructions. Final elution was performed in 17.5 µl of ultrapure H_2O. The library was then quantified using an Invitrogen Qubit fluorometer and quality-controlled using Agilent Bioanalyzer capillary electrophoresis.

The amplified cDNA library was then end-tagged and strand-purified using the NEBNext Ultra ^{TM} II DNA Library Prep Kit for Illumina.

M.4 Drop-ATAC Preliminary Trials

The preliminary Drop-ATAC trials followed the same steps as the final protocol, which is described in greater detail M.5, but stopped where failure occurred either physically due to emulsion destabilisation or molecularly by not yielding a DNA library.

Q5 PCR mix	
Q5 Buffer	44 µl
Ultra-pure H_2O	25.4 µl
40% PEG	16.5 µl
100 mM $MgCl_2$	11 µl
10 mM dNTPs	5.5 µl
Q5 Polymerase	1.1 µl

M.4.1 Droplet Stability Trials

Two different PCR master mixes were tested by emulsifying them together with 150 µl of beads pre-soaked in 150 µl of 2x lysis buffer. No nuclei were used in these trials.

Both emulsions were processed on the same PCR program.

2x Lysis Buffer	
Nuclease-free H_2O	1015.2 µl
5 M NaCl	72 µl
1 M Tris HCl	60 µl
10% IGEPAL	48 µl
10% SDS	4.8 µl

PCR Program		
	72 °C	5 min
	98 °C	5 min
14 cycles of:	98 °C	10 s
	60 °C	30 s
	72 °C	20 s
	4 °C	inf.

e2TAK PCR mix	
5x e2TAK Buffer	44 µl
Ultra-pure H_2O	25.4 µl
40% PEG	16.5 µl
100 mM $MgCl_2$	11 µl
10 mM dNTPs	5.5 µl
e2TAK Polymerase	1.1 µl

Emulsion stability was then assessed under a microscope.

The previous trial was repeated but with the PEG shifted from the PCR mix to the lysis mix, 0.2% BSA added to the lysis mix (resulting in 0.1% final droplet concentration), and with 15 % OptiPrep added to the PCR mix. In the repetition, no nuclei and no beads were used (2x lysis buffer was diluted with an equal volume of H_2O).

1x Lysis Buffer	
Nuclease-free H_2O	2107.2 µl
5 M NaCl	72 µl
1 M Tris HCl	60 µl
40% PEG	60 µl
10% IGEPAL	48 µl
10% BSA	48 µl
10% SDS	4.8 µl

e2TAK PCR mix	
5x e2TAK Buffer	44 µl
Ultra-pure H_2O	26.4 µl
Optiprep	16.5 µl
100 mM $MgCl_2$	11 µl
10 mM dNTPs	5.5 µl

Q5 PCR mix	
Q5 Buffer	44 µl
Ultra-pure H_2O	26.5 µl
Optiprep	16.5 µl
100 mM $MgCl_2$	11 µl
10 mM dNTPs	5.5 µl

Both emulsions were processed on the same PCR program as the first run.

PCR Program		
	72 °C	5 min
	98 °C	5 min
14 cycles of:	98 °C	10 s
	60 °C	30 s
	72 °C	20 s
	4 °C	inf.

Due to precipitation, this run was replicated completely, but without PEG and BSA (volumes replaced with water).

M.4.2 Bulk Runs

Here, the bead/lysis mix is replaced by an "empty" bead side mixture that does not contain BHBs, but readily dissolved primers instead. This mixture was co-encapsulated with 100 000 bulk tagmented nuclei resuspended in 120 µl of pre-made NEBNext PCR master mix. The first run was a repetition of a successful run by senior scientist which used standard NEBNext forward and reverse PCR. The nuclei suspension originated from an MM087 single-cell suspension bulk-tagmented by a lab scientist according to standard OmniATAc protocol (Corces et al., 2017).

"Bead" side mix	
Nuclease-free H_2O	96 µl
10 µM NEBNext fw primer	12 µl
10 µM NEBNext rv primer	12 µl

The resulting emulsion was PCR processed according to the following program:

PCR Program		
	72 °C	5 min
	95 °C	7 min
	98 °C	3 min
14 cycles of:	98 °C	10 s
	59 °C	10 s
	72 °C	1 min
	10 °C	inf.

Droplets were broken using droplet breaking solution, and the DNA contents were purified using Minelute columns, and processed on the Bioanalyzer.

This exact same protocol was repeated using our custom Drop-ATAC reverse primer instead of the NEBNext reverse primer:

"Bead" side mix	
Nuclease-free H_2O	96 µl
10 µM i5-NEBNext fw primer	12 µl
10 µM Drop-ATAC primer	12 µl

The third "bulk" Drop-ATAC run repeated all steps in the previous run, but now also had 1.8% Triton X-100 added to the "bead" side mix to completely dissolve the nuclei.

"Bead" side mix	
Nuclease-free H_2O	71.4 µl
10 µM i5-NEBNext fw primer	12 µl
10 µM Drop-ATAC primer	12 µl
10% Triton X-100	21.6 µl

The final bulk run was identical to the single-cell protocol described in M.5, but used the dissolved custom primer instead of beads. DTT is added to mimic the composition of the droplets in the single-cell experiment with beads.

"Bead" side mix	
Nuclease-free H_2O	96 µl
10 µM i5-NEBNext fw primer	12 µl
10 µM Drop-ATAC primer	12 µl

PCR Master Mix	
NEBNext 2x PCR Mix	120 µl
1 M DTT	2.4 µl

All other steps and ingredients are identical to M.5.

M.5 Final Drop-ATAC Protocol

M.5.1 Preparations

All the microfluidic peripherals were set up before preparing the cells, as the time between resuspending the cells into the RT master mix and starting cell encapsulation needs to be minimised. The PDMS microfluidic chip was fixed under a microscope, and a syringe was filled with oil and connected to the correct chip inlet via 1 mm tubing.

The BHB stock was centrifuged (3000 rcf, 3 min) and the supernatant was removed. 100 µl of condensed bead stock was added to 900 µl of HBW, vortexed well and centrifuged (5000 rcf, 3 min). 800 µl supernatant was removed, and replaced with an equal volume of HBW and the wash was repeated. Supernatant was removed, leaving 200 µl of condensed bead stock. The beads were incubated in an equal volume of 2x lysis buffer for 10 min on ice.

2x Lysis Buffer	
1 M Tris HCl	9 µl
5 M NaCl	1.8 µl
Nuclease-free H_2O	489.2 µl

The beads were then centrifuged (5000 rcf, 1 min) and supernatant was removed until 100 µl was left. The beads were then carefully centrifuged with a mini benchtop centrifuge to pull down beads sticking to the walls. Beads were then mixed with an appropriately i5-indexed NEBNext P5 PCR primer to prevent index collision on the sequencing run.

Bead Mix	
Washed BHBs	108 µl
10 µM i5-NEBNext Primer	12 µl

Two batches of each 50 000 cells were bulk tagmented according to the Omni-ATAC protocol by Corces et al. (2017). Tagmentation was stopped by adding 50 µl of suspension buffer to each 50 µl sample of nuclei and incubating on ice for 10 min..

Tagmentation Stop Buffer	
Nuclease-free H_2O	197 ml
1 M Tris-Hcl	2 µl
0.5 M EDTA	1 µl

Then, 100 µl of PBS was added to the nuclei suspension and the mixture was centrifuged (500 rcf, 5 min, 4 °C) and the supernatant was carefully removed. The nuclei pellet was then resuspended in 120 µl of PCR master mix with 20 mM DTT.

PCR Master Mix	
NEBNext 2x PCR Mix	120 µl
1 M DTT	2.4 µl

The nuclei are now ready to proceed to the Drop-ATAC run.

M.5.2 Drop-ATAC Run

Nuclei and bead mix were pre-loaded into 1 mm tubing by filling a syringe with PBS, connecting it to the tube, and carefully aspirating the solutions into the tube. The syringes were then fixed into a CETONI neMESYS 290N syringe pump, and tubing was connected to the correct inlet on the microfluidic chip. Similarly, the chip outlet was connected to a 1.5 ml eppendorf tube in a cold block.

All flows were simultaneously primed at 500 µl h^{-1} until they had all stabilised. Then, final flow rates were set as follows:

Flow rates	
Cell	900 µl h^{-1}
Beads	650 µl h^{-1}
Oil	1500 µl h^{-1}

When the nuclei suspension ran out, the microfluidic setup was kept running until an equal number of nuclei-free "dud" droplets was formed. These droplets contain only PBS and beads (no nuclei) and serve to stabilise the emulsion during PCR cycling. The resulting emulsion is redistributed into a 96-well plate at 70 µl of emulsion with 30 µl of EvaGreen QX200™ using a blunt pipette tip in order to not damage the droplets. A number of pre-PCR pictures were taken by pipetting a small amount of emulsion onto a glass slide to compare to droplet shape and stability post-PCR. The plate was then sealed, and PCR was carried out according to the following program:

PCR Program	
72 °C	5 min
95 °C	7 min
98 °C	3 min
18 cycles of: 98 °C	10 s
63 °C	30 s
72 °C	1 min
10 °C	inf.

Post-PCR pictures were taken of the emulsion and the library was cleaned up if droplets were still intact.

M.5.3 Post-PCR Clean-up

150 µl of droplet breaking solution was added to every well, and mixed well by pipetting.

Droplet Breaking Solution	
HFE 7500	700 µl
PFO	300 µl

The broken emulsions were then pooled into Eppendorf DNA lo-bind PCR tubes, and the bottom oil layer was removed and discarded. 175 µl of QIAGEN PB buffer was added to each PCR tube, and the solution was mixed well by pipetting 15 times. The contents were pooled into a 1.5 ml Eppendorf lo-bind tube, and sodium citrate was added in increments of 10 µl to adjust pH until the solution was vibrant orange in color. The mixture was then centrifuged (5000 rcf, 1 min) and the supernatant was removed. The DNA was then purified using QIAGEN MinElute columns according to manufacturer's instructions. Final elution

was performed in 21 µl, and the concentration was measured using an Invitrogen Qubit fluorometer.

If the DNA amount was satisfactory, the library was cleaned using 1.4 x volume Ampure beads according to manufacturer's protocol, and eluted in 12.5 µl of nuclease-free H_2O. Another Qubit quanitification was performed, and the library was further PCR amplified and tagged with Illumina P7 adapters according to NEBNext Ultra ™ II DNA Library Prep Kit specifications.

M.6 Sequencing and Data Analysis

M.6.1 Illumina NextSeq 500 Sequencing

All libraries were sequenced on the Illumina 500 platform. We used a 75 + 16 cycle kit which is originally designed for 75 sample DNA cycles and 2 x 8 index read cycles. We used a custom program and custom read primers to redirect these cycles as follows: 55 cycles are devoted to the cDNA read, 10 cycles for the first barcode read, 18 for the second barcode + UMI read, and 8 to the index read. Libraries were and custom read primers for the barcode were loaded onto the machine according to manufacturer's instructions by the lab manager, together with 1% PhiX in the first run and 20% PhiX in the second run. The NextSeq 500 libraries were then mapped to Hg38 and Hg19 by a lab bioinformatician using STAR. Barcode distribution plots were made using `bashscript` and `awk` commands, and visualised using `matplotlib`.

M.6.2 Sanger Sequencing

TOPO cloning and transformation of ThermoFisher OneShot chemically competent E. coli cells was performed according to manufacturer's protocol, but with all reaction volumes halved. 9 inDrop and 15 ATAC-seq fragment colonies were picked, and the amplified vector was purified using Qiagen Miniprep columns. Final elution occurred in 50 µl of H_2O and was sent for Sanger sequencing. Bases were called in ApE.

M.6.3 Seurat CCA

CCA was performed using the `R` package Seurat. 699 MM087 cells from the inDrop re-sequencing run with 1000 UMIs or more were analysed together with a previously generated Drop-seq data from MM087 cells, and 10 10x datasets from various MM cell lines. The first 70 CCs were calculated and CC1 - CC15 were selected after heat map visualisation. CC1 - C15 were then aligned using time-warping, and ultimately dimension-reduced using tSNE and UMAP.

M.6.4 MinION Sequencing

MinION sequencing was performed using the SQK-LSK-109 sequencing kit according to manufacturer's protocol on a flow-cell previously used for CHEQ-seq sequencing. The MinION was ran for 12 hours overnight and washed according to manufacturer's instructions after use. Barcode length distributions were obtained by first finding reference PCR sequence, Adapter and Poly-A elements in all reads using the `cutadapt` tool, and calculating the distributions of the distance between them in Python.

M.6.5 Typography and Figures

This document was written in LaTeXusing the `neovim` text editor. All figures were made in Inkscape.

Supplementary scripts, data and figures

S.1 Manufacturing Barcoded Hydrogel Beads

S.1.1 Python Script for Automatic Diameter Measurement

```
import numpy as np, pandas as pd, cv2, os, re

path_in = "../ims"
path_out = "./results.csv"
sigma = 0.33 # percentage upper/lowerbound for canny
conversion_factor = 0.474687961 # um/px
dmin = 20 # minimum beadd iameter in um
rmin = int(np.round(dmin/conversion_factor/2, 0))
rdmin = 2*rmin#minimum distance between centerpoints
dmax = 60 # maximal bead diameter
rmax = int(np.round(dmax/conversion_factor/2, 0))

def bhb_diam(im, rmin, rmax, rdmin):
    rerun = 1
    while rerun =  = 1:
        eq = cv2.equalizeHist(im) # equalize histogram to
           improve contrast
        blurred = cv2.GaussianBlur(eq, (3, 3), 0) #
           gaussian blur to smoothe
        v = np.median(blurred) # take median intensity of
           blurred image
        lower = int(max(0, (1.0-sigma)*v)) # lower bound
           for Canny filter
        upper = int(min(255, (1.0+sigma)*v)) # upper bound
            for Canny filter
        edged = cv2.Canny(blurred, lower, upper) # apply
           Canny filter to detect edges
```

```
        circles = cv2.HoughCircles(edged, cv2.
            HOUGH_GRADIENT, 1, rdmin, param1 = 30, param2 =
            30, minRadius = rmin, maxRadius = rmax)
        nbeads = int(np.round(np.divide(circles.size, 3),
            2)) # calculate total number of beads
        dmean = np.round((np.mean(circles, axis = 1)[0,
            2]*conversion_factor*2), 2)
        dstd = np.round((np.std(circles, axis = 1)[0, 2]*
            conversion_factor*2), 2)
        reldstd = np.round(np.multiply(np.divide(dstd,
            dmean), 100), 2)
        im = cv2.fastNlMeansDenoising(im, 10, 10, 9, 25)

# apply the function to all.tiff files in path_in
total_results = np.empty((0, 11))
for file in os.listdir(path_in):
    if file.endswith(".tif"):
        path_im = path_in+"/"+file
        im = cv2.imread(path_im, -1)
        output, nbeads, dmean, dstd, reldstd = bhb_diam(im
            , rmin, rmax, rdmin)

        settings = re.split('_', file)
        settings2 = re.split('-', settings[3])
        r = int(settings[2])/int(settings[1])
        Q = (int(settings[2])+int(settings[1]))/1000]
        im_results = np.array([settings[0], settings[1],
            settings[2], r, Q, settings2[0], re.sub(".tif",
            "", settings2[1]), nbeads, dmean, dstd, reldstd
            ])
        total_results = np.append(total_results, [
            im_results], axis = 0)
        cv2.imwrite(str(re.sub(".tif", "", file)+"
            _processed.tif"), output)

cols = np.array(['chem', 'ap', 'oil', 'r', 'Q', 's', 'pic'
    , 'n', 'dmean[um]', 'dstd', 'reldstd'])
df = pd.DataFrame(data = total_results, columns = cols)
df_sorted = df.sort_values(by = ['s', 'pic'])
print(df_sorted)
```

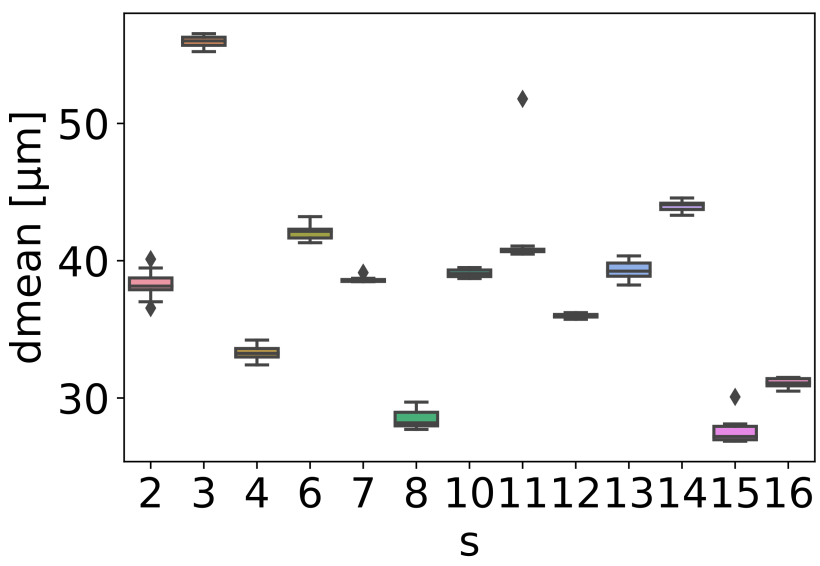
S.1.2 MATLAB Script for Bead Model Optimisation

```
syms x diam std

% define standard deviation function taken from JMP
std = @(x)2.4936441 + 0.4981291 * x(1) - 0.810586 * x(2)...
    +  0.7852652 * x(2) * x(2);

nonlcon = @constr % call the non-linear constraint defined
    in the function below

% minimise std based on the non-linear constraint
x = fmincon(std, [0.982 -0.65085714285], [], [], [], [],...
    [-1 -1],...
    [1 1],...
    nonlcon)

% convert x from coded parameter space to physical space
ap = x(1)*500 + 1000
oil = x(2)*1750 + 2250

% define constraint function
function [c,ceq] = constr(x)
    wanteddiam = 50
    c = []
    ceq = 39.963418 + 0.95277 * x(1) - 11.82268 * x(2)...
    - 3.002797 * x(1) * x(2)...
    + 2.7799737 * x(1) * x(1) - 1.875135 * x(2) * x(2)...
    - wanteddiam
end
```

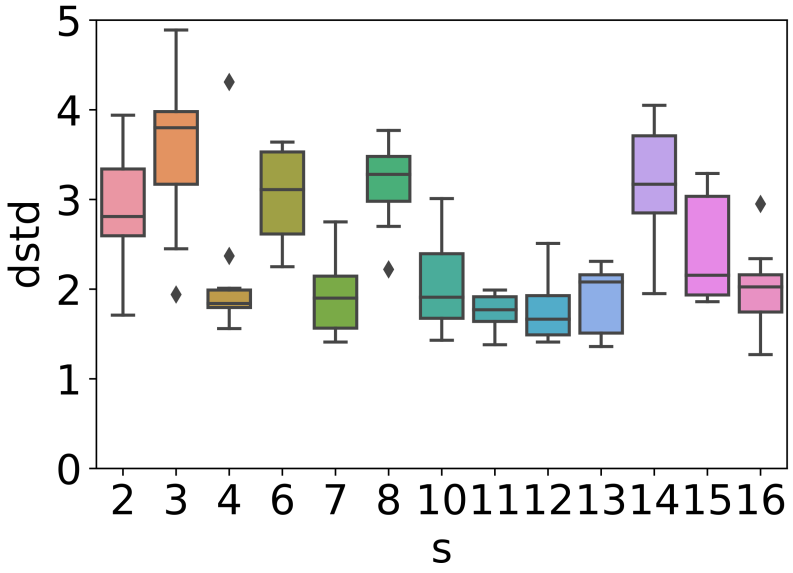
S.1.3 Model Data

block	run	AP	Oil	s	d
1	2	1500	3405	3.018628	38.169132
1	3	500	500	3.637247	55.816757
1	4	500	4000	1.978469	33.250258
2	8	1000	4000	2.953969	42.05399
2	7	500	2250	2.050949	38.767815
2	6	1000	1603	3.363416	28.716258
3	10	1150	1725	2.214742	39.066216
3	11	1500	2250	2.531675	40.80637
3	12	732	2600	1.851823	35.963218
4	13	500	2250	2.153793	39.167843
4	14	1150	1725	3.252607	44.006613
4	15	1200	3708	2.277241	27.377586
5	16	500	4000	1.986091	31.077456

Table S.1: Bead model data. Data generated by producing batches of hydrogel beads, taking images, analysing the bead diameters using the custom Python script and averaging over all batches. True modelling was performed on non-averaged dataset where each bead formed a single data point.



Sample	dmean (µm)
2	38.246311
3	56.004108
4	33.279067
6	42.060990
7	38.646020
8	28.562103
10	39.068467
11	42.031105
12	36.003394
13	39.185321
14	44.021055
15	28.373916
16	31.092614



Sample	stdev (µm)
2	2.790667
3	3.586338
4	2.106177
6	3.242827
7	1.887602
8	3.275265
10	2.046121
11	3.898428
12	1.880386
13	1.943205
14	3.188352
15	5.335753
16	2.089830

Figure S.1: Hydrogel bead model dataset.

S.2 inDrop

S.2.1 Failed Nextera Library Preparation Electropherogram

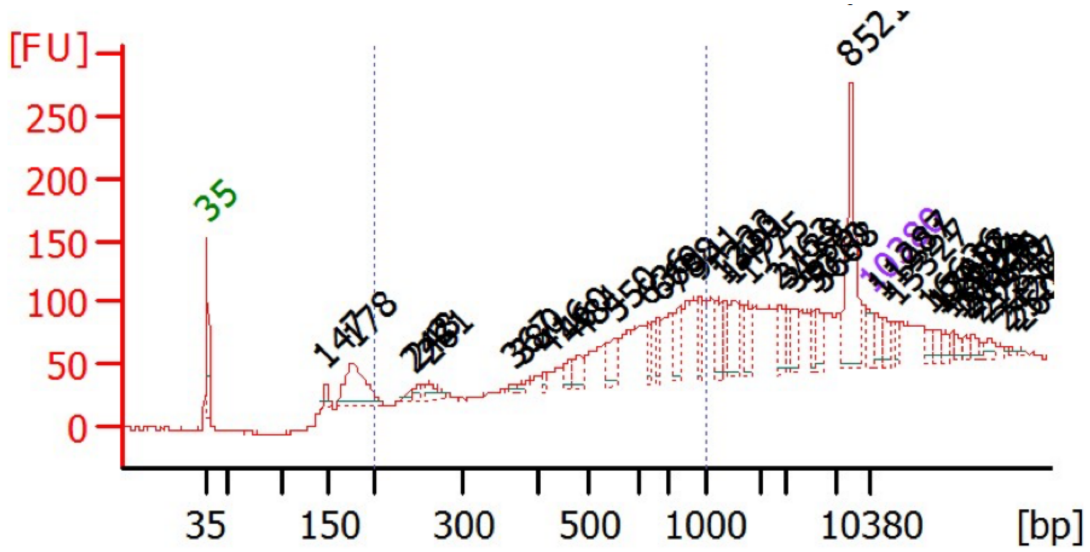


Figure S.2: Failed Nextera library preparation electropherogram. The fragment distribution is too broad, too long and has an enrichment of small fragments, indicating residual primers. Generated by senior lab scientist.

S.3 Sequencing

S.3.1 RNA-seq Datasets Correlations

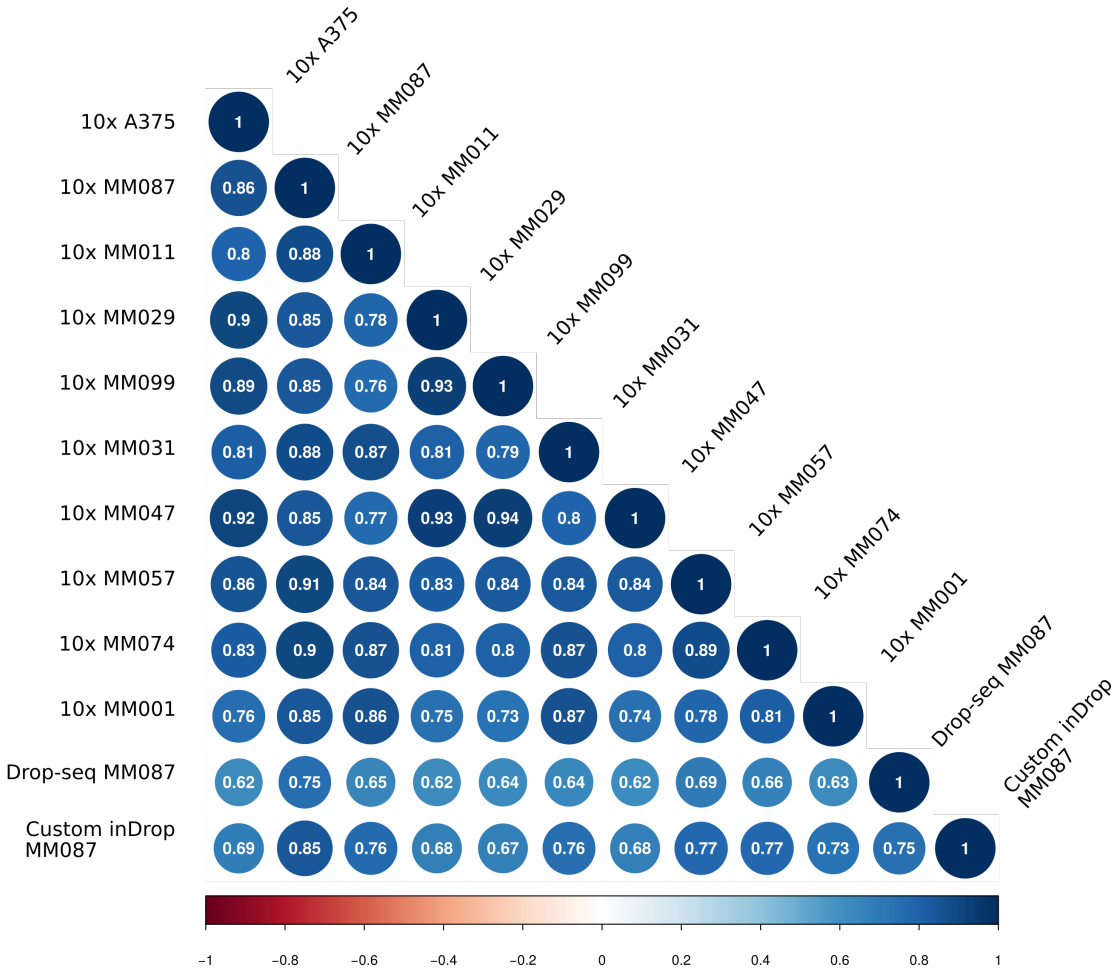


Figure S.3: Correlation between different single-cell RNA-seq datasets. Our inDrop MM087 dataset is slightly more correlated with the 10x MM087 dataset ($p < 0.01$).

S.3.2 Re-sequencing Run Gene Coverage

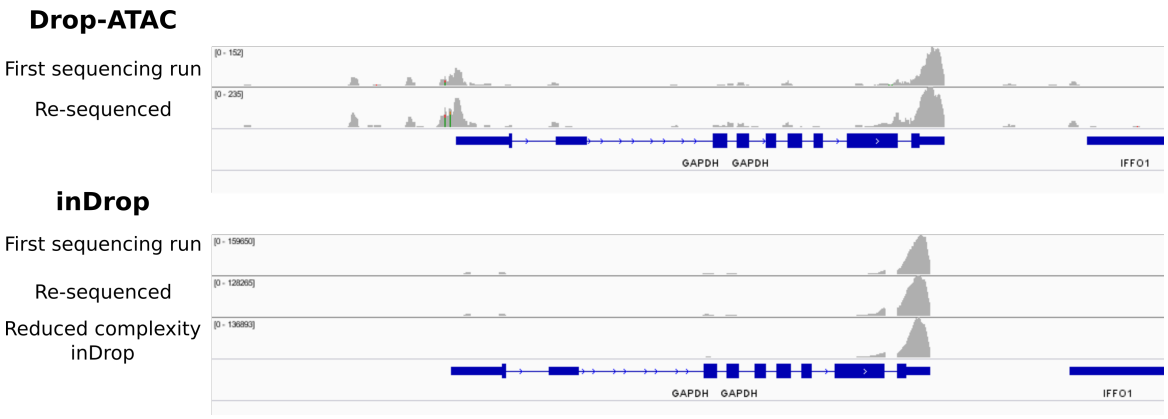


Figure S.4: Re-sequencing run gene coverage. Gene coverage of reads, visualised in IGV.

Vulgariserende Samenvatting

Wat veroorzaakt kanker? Hoe werken onze hersenen? Hoe wordt een embryo gevormd, en waar kan het fout lopen? Deze centrale vraagstukken in de biologie kunnen niet beantwoord worden via de verouderde analysetechnieken van de vorige decennia. Onderzoekers namen toen een weefselstaal, behandelden het met complexe technieken en analyseerden alle resultaten tezamen. Deze aanpak leidt tot een enorm verlies aan informatie, want weefsels bestaan in werkelijkheid uit miljoenen individuele cellen, elk met hun eigen rol. Met onze oude technologie kan geen onderscheid gemaakt worden tussen de verschillende cellen. Vandaag hebben een aantal technologische sprongen het echter mogelijk gemaakt om duizenden cellen individueel te behandelen, en dit binnen een redelijk tijdsbestek en aan een betaalbare prijs - een onmogelijke taak meer dan tien jaar geleden.

Er blijft echter veel plaats voor vordering. De "single-cell" technieken van vandaag staan nog niet op punt. Ze vereisen grote hoeveelheden manueel werk en zijn nog steeds te duur om op routinebasis uitgevoerd te worden. In een nieuwe klasse van single-cell technieken worden individuele cellen opgesloten in een druppel 1 miljoenste van een milliliter groot. De cellen worden elk in hun eigen druppeltje opgelost, en hun inhoud wordt individueel verwerkt. Doordat deze druppels zo klein zijn, kunnen we er vele duizenden tegelijk maken, en hoeven we niet te veel geld te spenderen aan dure reagentia. In dit project combineerden we twee reeds bestaande druppelgebaseerde technieken, wat leidde tot een verbeterde methode. Ook herschreven we een andere methode reeds bestaande methode zodat deze ook in druppels uitgevoerd kan worden.

Data gegenereerd met deze technieken kan ons helpen om processen te begrijpen waar de verschillende rollen van verschillende celtypes belangrijk is, en zal ons een stap dichter brengen bij he antwoord op de grote onopgeloste vragen in de biologie. Bovendien zijn we nu, Door de expertise en kennis opgedaan in dit project, goed uitgerust om aan de frontlijn van het "single-cell" onderziek te staan. De toekomst bevat een aantal spannende uitdagingen, van de ontwikkeling van volledig nieuwe technieken tot het combineren van reeds bestaande methodes. Door direct mee te werken aan het oplossen van deze uitdagingen hopen we single-cell technologie verder te brengen dan wat vandaag mogelijk is, en om zo mee de grenzen van de biologie te verleggen.