2. Single-cell RNA sequencing

This chapter provides a short introduction to the most widely used single-cell ribonucleic acid (RNA) sequencing assays and associated basic molecular biology concepts. Multimodal or spatial assays are not covered here, but are introduced in the respective advanced chapters. All sequencing assays have individual strengths and limitations which must be known by data analysis to be aware of possible biases in the data.

2.1. The building block of life

Life, as we know it, is the characteristic that distinguishes living from dead or inanimate entities. Most definitions of the term life share a common entity - cells. Cells form open systems which maintain homeostasis, have a metabolism, grow, adapt to their environment, reproduce, respond to stimuli, and organize themselves. Therefore, cells are the fundamental building block of life which were first discovered in 1665 by the British scientist Robert Hooke. Hooke investigated a thin slice of cork with a very rudimentary microscope, and to his surprise noticed that the slice appeared to resemble a honeycomb. He named these tiny units ‘cells’.

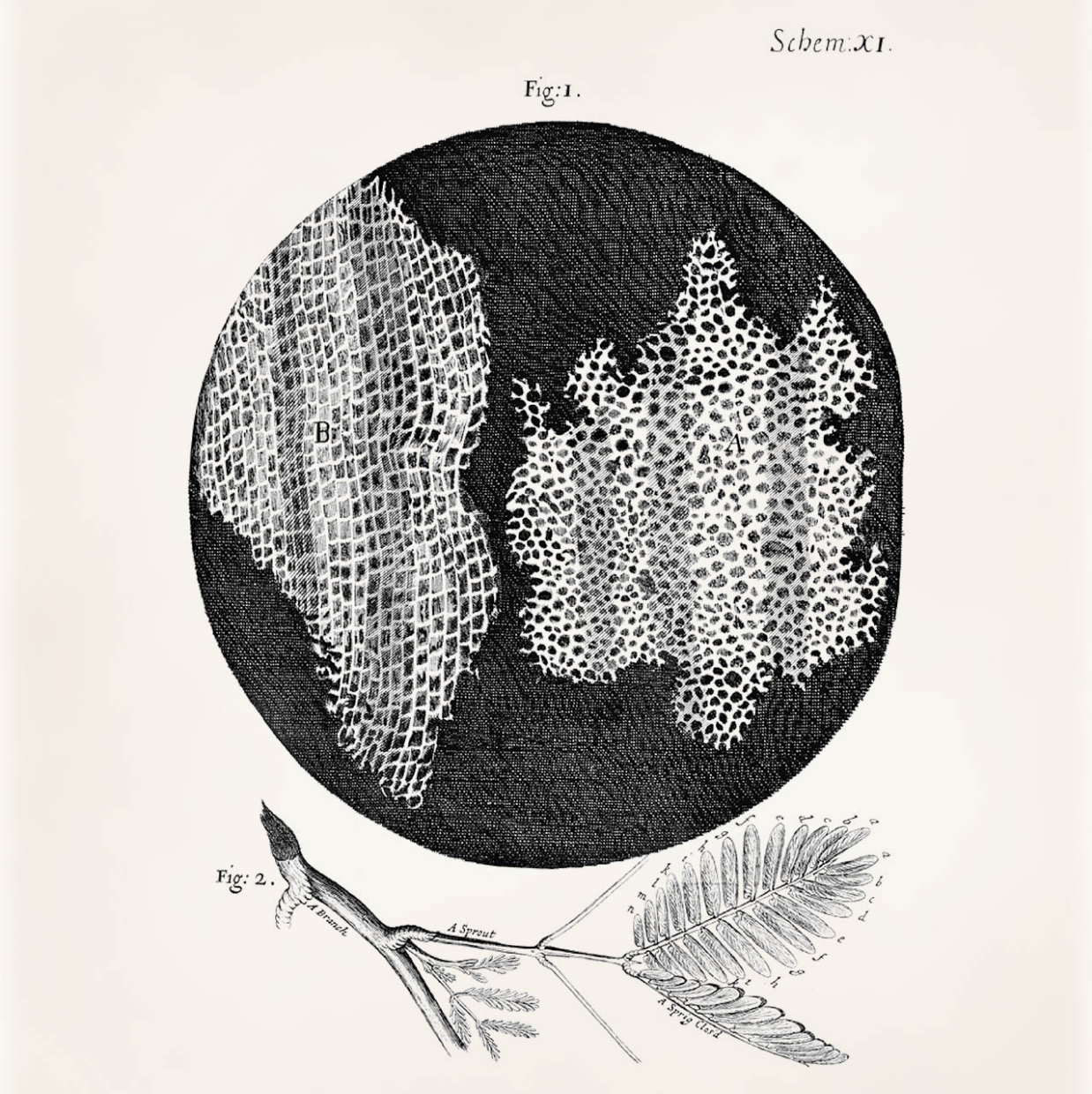

Fig. 2.1 Robert Hooke's drawing of cork cells. Image obtained from Micrographia.

In 1839, Matthias Jakob Schleiden and Theodor Schwann first described Cell Theory. It describes that all living organisms are made up of cells. Cells act as functional units that by themselves originate from other cells, making them the basic units of reproduction.

Since the early definition of cell theory, researchers discovered that there exists an energy flow within cells, that heredity information is passed from one cell to another in the form of DNA and that all cells have almost the same chemical composition. Two general types of cells exist, eukaryotes and prokaryotes. Eukaryotic cells contain a nucleus, where the nuclear membrane encapsulates the chromosomes; while prokaryotic cells only have a nucleoid region, but no nucleus. The nucleus hosts the cells' genomic deoxyribonucleic acid DNA and is the reason for the eukaryotes' name: *Nucleus* is Latin for kernel or seed. Eukaryotes are organisms composed of a single cell (unicellular) or multiple cells (multicellular), whereas prokaryotes are single-celled

organisms. Eukaryotic cells are further distinguished from prokaryotic cells by their high degree of compartmentalization, i.e. membrane-bound organelles are carrying out highly specialized functions and providing crucial support for cells.

Compared to prokaryotic cells, eukaryotic cells have on average about 10,000 the volume with a rich mix of organelles and a cytoskeleton constituted of microtubules, microfilaments, and intermediate filaments. The DNA replication machinery reads the hereditary information that is stored in the DNA in the nucleus to replicate themselves and keep the life cycle going. The eukaryotic DNA is divided into several linear bundles called chromosomes, which are separated by the microtubular spindle during nuclear division. Understanding the hereditary information hidden in DNA is key to understanding many evolutionary and disease-related processes. *sequencing* is the process of deciphering the order of DNA nucleotides and is primarily used to unveil the genetic information that is carried by a specific DNA segment, a complete genome, or even a complex microbiome. DNA *sequencing* allows researchers to identify the location and function of genes and regulatory elements in the DNA molecule and the genome, and uncovers genetic features such as open reading frames (ORFs) or CpG islands, which indicate promotor regions. Another very common application area is evolutionary analysis, where homologous DNA sequences from different organisms are compared. DNA sequencing can additionally be applied for the associations between mutations and diseases or sometimes even disease resistance, deeming one of the most useful applications.

A very popular example is sickle cell disease, a group of blood disorders, which results from an abnormality in the oxygen-carrying protein hemoglobin in red blood cells. This leads to serious health issues including pain, anemia, swelling in the hands and feet, bacterial infections and strokes. The cause of sickle cell disease is the inheritance of two abnormal copies of the β-globin gene (HBB) that makes hemoglobin, one from each parent. The gene defect is caused by a single nucleotide mutation where a GAG codon changes to a GTG codon of the β-globin gene. This results in the amino acid glutamate being substituted by valine at position 6 (E6V substitution) and henceforth the above-mentioned disease. It is unfortunately not always possible to find such “simple” associations between single nucleotide mutations and diseases, due to most diseases being caused by, for example, complex regulatory processes.

2.2. A brief history of sequencing

2.2.1. First generation sequencing

Although DNA was already first isolated in 1869 by Friedrich Mietscher, it took the scientific community more than 100 years to develop high throughput sequencing technologies. In 1953, Watson, Crick and Franklin discovered the structure of DNA; and in 1965 Robert Holley sequenced the first tRNA. Seven years later, in 1972, Walter Fiers was the first to sequence a complete gene (the coat protein of bacteriophage MS2) using RNAses to digest the virus RNA, isolate oligonucleotides and finally

separate them with electrophoresis and chromatography[JOU *et al.*, 1972]. In parallel, Friedrich Sanger developed a DNA sequencing method using radiolabeled, partially digested fragments termed “chain termination method”, which is more commonly known as “Sanger Sequencing”. Although Sanger Sequencing is still used even today, it suffered from several shortcomings, including lack of automation and time-consuming. In 1987, Leroy Hood and Michael Hunkapiller developed the ABI 370, an instrument that automates the Sanger Sequencing process. Its most important innovative accomplishment was the automatic labeling of DNA fragments with fluorescent dyes instead of radioactive molecules. This change not only made the method safer to perform, but also allowed for computers to analyze the acquired data[Hood *et al.*, 1987].

Strengths:

- Sanger sequencing is simple and affordable.
- If done correctly, the error rate is very low (<0.001%).

Limitations:

- Sanger methods can only sequence short pieces of DNA of about 300 to 1000 base pairs (bp).
- The quality of a Sanger sequence is often not very good in the first 15 to 40 bases, because this is where the primers bind.
- Sequencing degrades after 700 to 900 bases.
- If the sequenced DNA fragment has been cloned, some of the cloning vector sequence may find its way into the final sequence.
- Sanger sequencing is more expensive than second or third generation sequencing per sequenced base.

2.2.2. Second generation sequencing

Nine years later, in 1996, Mostafa Ronaghi, Mathias Uhlen, and Pål Nyŕen introduced a new DNA sequencing technique called pyrosequencing, introducing the age of second generation sequencing. Second generation sequencing, also known as next-generation sequencing (NGS), was primarily made possible by further automation in the lab, the usage of computers, and the miniaturization of reactions. Pyrosequencing measures luminescence that is generated by pyrophosphate synthesis during sequencing. This process is also commonly known as “sequencing-by-synthesis”. Two years later, Shankar Balasubramanian and David Klenerman, developed and adapted the sequencing-by-synthesis process for a new method which utilizes fluorescent dyes at the company Solexa. Solexa’s technology also forms the basis of Illumina’s sequencers, which dominate the market today. The Roche 454 sequencer developed in 2005, was the first sequencer to fully automate the pyrosequencing process in a single,

automated machine. Many other platforms were introduced such as SOLiD systems' "sequencing-by-ligation" (2007) and Life Technologies' Ion Torrent (2011) that uses "sequencing-by-synthesis" to detect hydrogen ions when new DNA is synthesized.

Strengths:

- Second generation generation sequencing is often the cheapest option with respect to required chemicals.
- Sparse material can still be used as input.
- High sensitivity to detect low-frequency variants and comprehensive genome coverage.
- High capacity with sample multiplexing.
- Ability to sequence thousands of genes simultaneously,

Limitations:

- The sequencing machines are expensive and often need to be shared with colleagues.
- Second generation sequencers are big, stationary machines and not designed for field work.
- Generally, second generation sequencing results in many short sequencing fragments (reads) which are hard to use for novel genomes.
- The quality of sequencing result is dependent on the reference genome

2.2.3. Third generation sequencing

The third generation of sequencing, nowadays also known as next-generation sequencing, brought two innovations to the market. First, long-read sequencing, which describes the ability to obtain nucleotide fragments of longer lengths than the usual Illumina short-read sequencers generate (order of 75 to 300 base pairs depending on the sequencer). This is especially important for the assembly of novel genomes without an available reference genome. Second, the ability to sequence in real time is another major advancement in third generation sequencing. Combined with portable sequencers, which are small in size and do not require further complex machines for the chemistry, sequencing is now "field-ready" and can be used even far away from laboratory facilities to collect samples.

Pacific Biosciences' (PacBio) introduced zero-mode waveguide (ZMW) sequencing in 2010, which uses so-called nanoholes containing a single DNA polymerase. This allows incorporation of any single nucleotide to be directly observed by detectors attached below the nanoholes. Each type of nucleotides is labeled with a specific

fluorescent dye that emits fluorescent signals during the incorporation process, which are subsequently measured as sequence readout. Reads obtained from PacBio sequencers are usually of 8 to 15 kilobases (kb), with possibilities for up to 70kb.

Oxford Nanopore Technologies' introduced the GridION in 2012. The GridION and its successors MinIO and Flongle are portable sequencers for DNA and RNA sequencing which produce reads of more than 2 Mb. Notably, such a sequencing device even fits into a single human hand. Oxford Nanopore sequencers observe changes in the electrical current that occur when nucleic acids pass through protein nanopores, to identify the nucleotide sequence[Jain *et al.*, 2016].

Strengths:

- Long reads will allow for the assembly of large novel genomes.
- Sequencers are portable, allowing for field work.
- Possibility to directly detect epigenetic modifications of DNA and RNA sequences.
- Speed. Third generation sequencers are fast.

Limitations:

- Some third generation sequencers exhibit higher error rates than second generation sequencers.
- The reagents are generally more expensive than second generation sequencing.

2.3. Overview of the NGS process

Even though a variety of NGS technologies exist, the general steps to sequence DNA (and therefore reverse transcribed RNA) are largely the same. The differences lie primarily in the chemistry of the respective sequencing technologies.

1. **Sample and library preparation**: As a first step, a so-called library is prepared by fragmenting the DNA samples and ligating them with adapter molecules. They act in the hybridisation of the library fragments to the matrix and provide a priming site.
2. **Amplification and sequencing**: In the second step, the library gets converted into single strand molecules. During an amplification step (such as a polymerase chain reaction), clusters of DNA molecules are being created. All of the clusters perform individual reactions during a single sequencing run.

3. **Data output and analysis**: The output of a sequencing experiment depends on the sequencing technology and chemistry. Some sequencers generate fluorescence signals which are stored in specific output files, and others may generate electric signals which are stored in corresponding file formats. Generally, the amount of generated data, the raw data, is very large. Such data requires complex and computationally heavy processing. This is further discussed in the raw data processing chapter.

2.4. RNA sequencing

So far, we have only introduced sequencing with the unmentioned assumption that the DNA is being sequenced. However, knowing the DNA sequence of an organism and the positions of its regulatory elements tells us very little about the dynamic and real-time operations of a cell. For example, by combining different mRNA splicing sites and exons from the same mRNA precursor, one gene can code for multiple proteins. This alternative splicing event is naturally occurring and commonly seen in eukaryotes; however, a variant could potentially result in a non-functional enzyme and an induced disease state. This is where RNA sequencing (RNA-Seq) comes into play. RNA-Seq largely follows the DNA sequencing protocols, but includes a reverse transcription step where complementary DNA (cDNA) is synthesized from the RNA template. Sequencing RNA allows scientists to obtain snapshots of cells, tissues or organisms at the time of sequencing in the form of expression profiles of genes. This information can be used to detect changes in disease states in response to therapeutics, under different environmental conditions, when comparing genotypes and other experimental designs. Modern RNA sequencing allows for an unbiased sampling of transcripts in contrast to for example microarray based assays or RT-qPCR, which require probe design to specifically target the regions of interest. The obtained gene expression profiles further enable the detection of gene isoforms, gene fusions, single nucleotide variants, and many other interesting properties. Modern RNA sequencing is not limited by prior knowledge and allows for the capture of both known and novel features, resulting in rich data sets that can be used for exploratory data analysis.

2.5. Single-cell RNA sequencing

2.5.1. Overview

Sequencing of RNA can be mainly conducted in two ways: Either by sequencing the mixed RNA from the source of interest across cells (bulk sequencing) or by sequencing the transcriptomes of the cells individually (single-cell sequencing). Mixing the RNA of all cells is in most cases cheaper and easier than experimentally complex single-cell sequencing. Bulk RNA-Seq results in cell-averaged expression profiles, which are generally easier to analyze, but also hide some of the complexity such as cell expression profile heterogeneity, which may help answer the question of interest. Some drugs or perturbations may affect only specific cell types or interactions between cell

types. For example, in oncology, it is possible to have rare drug resistant tumor cells causing relapse, which is difficult to identify by simple bulk RNA-seq even on cultured cells.

To uncover such relationships, it is vital to examine gene expression on a single-cell level. Single-cell RNA-Seq (scRNA-Seq) does, however, come with several caveats. First, single-cell experiments are generally more expensive and more difficult to properly conduct. Second, the downstream analysis becomes more complex due to the increased resolution, and it is easier to draw false conclusions.

A single-cell experiment generally speaking, follows the same steps as a bulk RNA-Seq experiment (see above), but requires several adaptations. Just like bulk sequencing, single-cell sequencing requires lysis, reverse transcription, amplification, and the eventual sequencing. In addition, single-cell sequencing requires cell isolation and a physical separation into smaller reaction chambers or another form of cell labeling to be able to map the obtained transcriptomes back to the cells of origin later on. Hence, these are also the steps where most single-cell assays differ: single-cell isolation, transcript amplification, and, depending on the sequencing machine, sequencing. Before explaining how the different approaches to sequencing work, we will now discuss transcript quantification more closely.

2.5.2. Transcript quantification

Transcript quantification is the process of counting the hits of the sequenced transcripts against the gene sequences. These counted hits eventually make it into the count table. More details on this computational process will be described in the next chapter. There are two major approaches to transcript quantification: full-length and tag-based. Full-length protocols try to cover the whole transcript uniformly with sequencing reads, whereas tag-based protocols only capture the 5’ or 3’ ends. The transcript quantification method has strong implications on the captured genes, and analysts must therefore be aware of the used quantification process. Full-length sequencing is restricted to plate-based protocols (see below) and the library preparation is comparable to bulk RNA-seq sequencing approaches. An even coverage of transcripts is not always achieved with full-length protocols and therefore specific regions across the gene body may still be biased. A major advantage of full-length protocols is that they allow for the detection of splice variants. Tag-based protocols only sequence the 3’ or 5’ ends of the transcripts. This comes at the cost of not (necessarily) covering the full gene length, making it difficult to unambiguously align reads to a transcript and distinguishing between different isoforms[Archer *et al.*, 2016]. However, it allows for the usage of unique molecular identifiers (UMIs), which are useful to resolve biases in the transcript amplification process. The transcript amplification process is a critical step in any RNA-seq sequencing run, to ensure that the transcripts are abundant enough for quality control and sequencing. During this process, which is typically conducted with polymerase chain reaction (PCR), copies are made from identical fragments of the original molecule. Since the copies and the original molecules are indistinguishable, determining the original number of molecules in samples becomes challenging. The

usage of UMIs is a common solution to quantify the original, non-duplicated molecules. The UMIs serve as molecular barcodes and are also sometimes referred to as random barcodes. These 'barcodes' consist of short random nucleotide sequences that are added to every molecule in the sample as a unique tag. UMIs must be added during library generation before the amplification step. The ability to accurately identify PCR duplicates is important for downstream analysis to rule out - or be aware of amplification biases[Aird *et al.*, 2011]. Amplification bias is a term for the RNA/cDNA sequences which are preferentially amplified and will therefore be sequenced more often, resulting in higher counts. It can have a detrimental effect on any gene expression analysis, because the not-very-active genes may suddenly appear to be highly expressed. This is especially true for sequences which are amplified at a later stage of the PCR step, where the error rate may already be comparably higher than earlier PCR stages. Although it is computationally possible to detect and remove such sequences by removing reads with identical alignment coordinates, it is generally advised to always design the experiment with UMIs, if possible. The usage of UMIs further allows for normalization of gene counts to be performed without a loss of accuracy[Kivioja *et al.*, 2012].

ADD A FIGURE HERE.

2.5.3. Single-cell sequencing protocols

Currently, three types of single-cell sequencing protocols exist, which are grouped primarily by their cell isolation protocols: Microfluidic device-based strategies where cells are encapsulated into hydrogel droplets; well plate based protocols where cells are physically separated into wells and finally, the commercial Fluidigm C1 microfluidic chip based solution which loads and separates cells into small reaction chambers. These three approaches differ in their ability to recover transcripts, the number of sequenced cells, and many other aspects. In the following subsections, we will briefly discuss how they work, their strengths and weaknesses, and possible biases that data analysts should be aware of regarding the respective protocols.

2.5.3.1. Microfluidic device based protocols

Microfluidic device based single-cell strategies trap cells inside hydrogel droplets allowing for compartmentalisation into single-cell reaction chambers. The most widely used protocols inDrop[Klein *et al.*, 2015], Drop-seq[Macosko *et al.*, 2015] and the commercially available 10x Genomics Chromium[Zheng *et al.*, 2017] are able to generate such droplets several thousand times per second. This massively parallel process generates very high numbers of droplets for a relatively low cost. Although all three protocols differ in details, nanoliter-sized droplets containing encapsulated cells are always designed to capture beads and cells simultaneously. The encapsulation process is conducted with specialized microbeads with on-bead primers containing a PCR handle, a cell barcode and a 4-8b bp-long unique molecular identifier (UMI - see below) and a poly-T tail. Upon lysis the cell's mRNA is instantaneously released and captured by the barcoded oligonucleotides that are attached on the beads. Next, the

droplets are collected and broken to release single-cell transcriptomes attached to microparticles (STAMPs). This is followed by PCR and reverse transcription to capture and amplify the transcripts. Finally, tagmentation takes place where the transcripts are randomly cut and sequencing adaptors get attached. This process results in sequencing libraries that are ready for sequencing as described above. In microfluidic based protocols only about 10% of the transcripts of the cell are recovered[Islam *et al.*, 2014]. Notably, this low sequencing is sufficient for robust identification of cell types.

All three microfluidic device-based methods result in characteristic biases. The material of the used beads differs between the protocols. Drop-seq uses brittle resin for the beads and therefore the beads are encapsulated with a Poisson distribution, whereas the InDrop and 10X Genomics beads are deformable resulting in bead occupancies of over 80%[Zhang *et al.*, 2019]. Moreover, capture efficiency is likely influenced by the use of surface-tethered primers in Drop-Seq. InDrop uses primers which are released with photocleavage and 10X genomics dissolves the beads. This disparity also affects the location of the reverse transcription process. In Drop-seq, reverse transcription occurs after the beads are released from the droplets, while reverse transcription takes place inside the droplets for the InDrop and 10X genomics protocols[Zhang *et al.*, 2019].

A comparison from Zhang et al. in 2019 uncovered that inDrop and Drop-seq are outperformed by 10X Genomics with respect to bead quality, as the cell barcodes in the former two systems contained obvious mismatches. Moreover, the proportion of reads originating from valid barcodes was 75% for 10X Genomics, compared to only 25% for InDrop and 30% for Drop-seq.

Similar advantages were demonstrated for 10X Genomics regarding sensitivity. During their comparison, 10X Genomics captured about 17000 transcripts from 3000 genes on average, compared to 8000 transcripts from 2500 genes for Drop-seq and 2700 transcripts from 1250 genes for InDrop. Technical noise was the lowest for 10X Genomics, followed by Drop-seq and InDrop[Zhang *et al.*, 2019].

The actual generated data demonstrated large protocol biases. 10X Genomics favored the capture and amplification of shorter genes and genes with higher GC content, while Drop-seq in comparison preferred genes with lower GC content. Although 10X Genomics was shown to outperform the other protocols in various aspects, it is also about twice as expensive per cell. Moreover, except the beads, Drop-seq is open-source and the protocol can more easily be adapted if required. InDrop is completely open-sourced, where even the beads can be manufactured and modified in labs. Hence, InDrop is the most flexible of the three protocols.

Strengths:

- Allows for the cost-efficient sequencing of cells in large quantities, to identify the overall composition of a tissue and characterize rare cell types.
- UMIs can be incorporated.

Limitations:

- Low detection rates of transcripts compared to other methods.
- Captures 3’ only and not full transcripts, because the cell barcodes and PCR handles are only added to the end of the transcript.

2.5.3.2. Plate based

Plate based protocols typically separate the cells physically into microwell plates. The first step entails cell sorting by, for example, fluorescent-activated cell sorting (FACS), where cells are sorted according to specific cell surface markers; or by micro pipetting. The selected cells are then placed into individual wells containing cell lysis buffers, where subsequently reverse transcription is carried out. This allows for several hundreds of cells to be analyzed in a single experiment with 5000 to 10000 captured genes each. Plate based sequencing protocols include, but are not limited to, SMART-seq2, MARS-seq, QUARTZ-seq and SRCB-seq. Generally speaking, the protocols differ in their multiplexing ability. For example, MARS-seq allows for three barcode levels, namely molecular, cellular and plate-level tags, for robust multiplexing capabilities. SMART-seq2 on the contrary, does not allow for early multiplexing limiting cell numbers. A systematic comparison of protocols by Mereu et al in 2020 revealed that QUARTZ-seq2 is able to capture more genes than SMART-seq2, MARS-seq or SRCB-seq per cell[Mereu *et al.*, 2020], which means QUARTZ-seq2 is able to capture cell-type specific marker genes well, allowing for confident cell type annotation.

Strengths:

- Recovers many genes per cell, allowing for a deep characterization.
- Possible to gather information before the library preparation e.g. through FACS sorting to associate information such as cell size and the intensity of any used labels with well coordinates.
- Allows for full-length transcript recovery.

Limitations:

- The scale of plate-based experiments is limited by the lower throughput of their individual processing units.
- Fragmentation step eliminates strand-specific information [Hrdlickova *et al.*, 2017].
- Depending on the protocol, plate based protocols might be labor-intensive with many required pipetting steps, leading to potential technical noise and batch effects.

2.5.3.3. Fluidigm C1

The commercial Fluidigm C1 system is a microfluidic chip, which loads and separates cells into small reaction chambers in an automated manner. The CEL-seq2 and SMART-seq (version 1) protocols are using the Fluidigm C1 chips in their workflow, allowing the RNA extraction and library preparation steps to be conducted together, thereby decreasing the required manual labor. However, the Fluidigm C1 requires rather homogeneous cell mixtures, since the cells will reach different locations on the microfluidic chip based on their size, which could introduce potential location bias. Since the amplification step is carried out in individual wells, full-length sequencing is possible, effectively reducing the 3' bias of many other single-cell RNA-seq sequencing protocols. The protocol is generally also more expensive and is therefore primarily useful for an extensive examination of a specific cell population.

Strengths:

- Allows for full-length transcript coverage.
- Splicing variants and T/B cell receptor repertoire diversity can be recovered.

Limitations:

- Only allows for the sequencing of up to 800 cells[Fluidigm, 2022].
- More expensive per cell than other protocols.
- Only about 10% of the extracted cells are captured, which makes this protocol unsuitable for rare cell types or low input.
- The used arrays only capture specific cell sizes, which may bias the captured transcripts.

2.5.3.4. Nanopore single-cell transcriptome sequencing

Long-read single-cell sequencing approaches rarely used UMI [Singh *et al.*, 2019] or did not perform UMI correction [Gupta *et al.*, 2018] and therefore assigned novel UMI reads to novel UMIs. Due to the higher sequencing error rate of long-read sequencers this causes serious issues [Lebrigand *et al.*, 2020]. Lebigrand et al. introduced ScNaUmi-seq (Single-cell Nanopore sequencing with UMIs) which combines Nanopore sequencing with cell barcode and UMI assignment. The barcode assignment is guided with Illumina data by comparing the cell bar code sequences found in the Nanopore reads with those recovered from the Illumina reads for the same region or gene [Lebrigand *et al.*, 2020]. However, this effectively requires two single-cell libraries. scCOLOR-seq computationally identifies barcodes without errors using nucleotide pair complementary across the full length of the barcode. These barcodes are then used as guides to correct the remaining erroneous barcodes [Philpott *et al.*, 2021]. A modified UMI-tools directional network based method corrects for UMI sequence duplication.

Strengths:

Recovers splicing and sequence heterogeneity information

Weaknesses:

- Nanopore reagents are expensive.
- High cell barcode recovery error rates.
- Depending on the protocol, barcode assignment is guided with Illumina data requiring two sequencing assays.

2.5.3.5. Summary

In summary, we strongly recommend that wet lab and dry lab scientists select the sequencing protocol based on the aim of the study. Is a deep characterization of a specific cell type population desired? In this case one of the plate-based methods may be more suitable. On the contrary, droplet based assays will capture heterogeneous mixtures better, allowing for a more broad characterization of the sequenced cells. Moreover, if the budget is a limiting factor, the protocol of choice should be more cost-effective and robust. When analyzing the data, be aware of the sequencing assay specific biases. For an extensive comparison of all single-cell sequencing protocols, we recommend the “Benchmarking single-cell RNA-sequencing protocols for cell atlas projects” paper by Mereu et al[Mereu *et al.*, 2020].

2.5.4. single-cell vs single-nuclei

So far we have only been discussing single-cell assays, but it is also possible to only sequence the nuclei of the cells. Single-cell profiling does not always provide an unbiased view on cell types for specific tissues or organs, such as, for example, the brain. During the tissue dissociation process, some cell types are more vulnerable and therefore difficult to capture. For example, fast-spiking parvalbumin-positive interneurons and subcortically projecting glutamatergic neurons were observed in lower proportions than expected in mouse neocortex[Tasic *et al.*, 2018]. On the contrary, non-neuronal cells survive dissociation better than neurons and are overrepresented in single-cell suspensions in the adult human neocortex[Darmanis *et al.*, 2015]. Moreover, single-cell sequencing highly relies on fresh tissue, making it difficult to make use of tissue biobanks. On the other hand, the nuclei are more resistant to mechanical force, and can be safely isolated from frozen tissue without the use of tissue dissociation enzymes[Krishnaswami *et al.*, 2016]. Both options have varying applicability across tissues and sample types, and the resulting biases and uncertainties are still not fully uncovered. It has been shown already that nuclei accurately reflect all transcriptional patterns of cells[Ding *et al.*, 2020]. The choice of single-cell versus single-nuclei in the experimental design is mostly driven by the type of tissue sample. Data analysis however should be aware of the fact that dissociation ability will have a strong effect on the potentially observable cell types. Therefore, we strongly encourage discussions between wet lab and dry lab scientists concerning the experimental design.