University of Catania

Dipartimento di matematica e informatica

MASTER'S DEGREE IN COMPUTER SCIENCE (LM-18)

Single-cell sequencing: a new frontier for personalized medicine

THESIS

Supervisor:
Alfredo Ferro

Co-supervisor:
Salvatore Alaimo

Co-supervisor:
Giovanni Micale

Candidate:
Locicero Giorgio

Serial number:
1000024196

ACADEMIC YEAR 2021-2022

Abstract

Single-cell sequencing is one of the most significant new cutting-edge research of biological analysis and bioinformatics since it results in high resolution and accurate data associated with cells in a tissue sample or a group of cells. This means that the results obtained will be very precise, detailed and personalized. Along with the details of single cells, we can examine and infer cell-to-cell interaction, spatial transcriptomics, processes and behavior or how the cell(or a group of cells) influences the environment where the cells are. We can also work in conjunction with multi-omics and combine the data from different sources to create models that are very accurate. In recent years, there has been a rapid increase in the use of single-cell sequencing along with most of the important and innovative methods of AI applied to this field. Data obtained from single-cell sequencing is used for personalized medicine and to build innovative solutions to problems related to the field of biology, medicine and computer science as well. This research will focus on how to use the material obtained from multi-omics and especially from single-cell sequencing to infer information from the expression of single cells in comparison and combination with classic bulk RNA-seq data. This research will aim at applying single-cell sequence data to biological pathways, see if the final results obtained from pathway activation and activity analysis will differ between two groups with distinct features and how the results will be different from those obtained from bulk RNA-seq. The research will not stop at the biological side of the various systems (through data analysis and the use of already established tools), but the project will aim at addressing the use and theory behind it with a lot of attention to the definition of the problems and the formalization from a mathematical and CS point of view. A methodology will be presented to treat single-cell data in combination with bulk data and topological information obtained from biological networks, and this methodology will be tested on real data.

Contents

Chapter 1

Introduction

Personalized medicine is the forefront new branch of medicine that uses an individual's genetic profile to guide decisions made in regard to the prevention, diagnosis, and treatment of disease. In personalized medicine, the objective is to be more precise and have predictability and powerful health care that is customized for individual patients. With the growing understanding of genetics and genomics and how they influence health, disease and drug responses in each person, the individual study of patients or groups of patients with genetic profiles is enabling doctors to provide better disease prevention, more accurate diagnoses, safer drug prescriptions and more effective treatments for the many diseases and conditions that affect health. Also known as individualized medicine or genomic medicine, personalized medicine aims at tailoring health care to each person's unique genetic makeup.

To reach the objectives that are at the core of personalized medicine, a lot of research for techniques that use sequence and expression databases is being done, some of the research done even at this moment is the following:

- Systems genetics is a comprehensive approach to studying and understanding this biological complexity [[91],[48]].
- Mining vast databases (big data) with powerful tools from statistics, genetics, mathematics, physics and computer science [8].
- Statistical methods that focus on smaller, specific and well-defined subgroups, sought to provide guidance in clinical decision making based on individual differences, and have attempted to achieve better risk minimization and benefit maximization for single-stage and multi-stage clinical data. For the latter, it involves **dynamic treatment regimes** to control the treatment over all the stages and make decisions based upon statistical analysis and a decision support system [171].
- With help of CT scan (DICOM images), computer vision is used for a preliminary diagnosis and taking into consideration the condition of the patients when deciding for a treatment [29].
- The use of deep learning techniques for drug development, disease characteristic identification, and therapeutic effect prediction, while considering applied methods in detail and offering insights into their pros and cons while also taking into attention challenges and the future of deep learning in this field [169].

The objectives of this thesis are related to statistical analysis and inference along with the creation of AI models and Graph models to model the response to treatment, disease and prevention by simulation, subgrouping and creating links between different

-omics and forefront technologies (single-cell sequencing in particular). All source code, data, workflows, methodologies and theories fall under the name of the project that is **SCAPE** (Single Cell Analysis and Pathway Embedding) and the source code is open source and available in my repository [73].

Across human tissues, there is an incredible diversity of cell types, states, and interactions. To better understand these tissues and the cell types present, single-cell RNA-seq (scRNA-seq) offers a glimpse into what genes are being expressed at the level of individual cells. Often, cells in the same tissue are not independent of each other in their functions since they communicate and cooperate to carry out the tissue function as a whole. This consideration is really important when studying the parts of a complex system since the function of a single tissue (and of the single part isolated) is not enough to describe the system itself and needs to take into account how the single parts communicate and change their functions based on the needs of the whole tissue itself. The techniques of scRNA-seq are not enough to describe a system though, since they have a lot of technical limitations and most of them are only at the start of their life. Information acquired through single-cell techniques should be integrated with classical approaches like RNA-seq (for the whole tissue, otherwise called **bulk**) and other sources of data and information like pathways, spatial genomics, etc.

Today it is possible to obtain genome-wide transcriptome data from single cells using high-throughput scRNA-seq. The main advantage of scRNA-seq is that the cellular resolution and the genome-wide scope make it possible to address issues that are intractable using other methods, such as bulk RNA-seq or combined methods that use data from multi-omics to get biomarkers (after the construction of the models and analysis of the big data resulting from these sources) and useful insight. However, to analyze scRNA-seq data, novel methods are required and some of the underlying assumptions for the methods developed for bulk RNA-seq experiments are no longer valid.

The resolution of scRNA-seq has some problems due to technical noise, imprecision of devices used, and the methods used both for the scRNA-seq preparation to get the sequences of reads and the computation of values done with computers since these kinds of procedures are derived from common mechanisms used for bulk sequencing.

The objectives of this thesis are to obtain information about biological networks for single cell analysis and use this information to:

- estimate traits about patients
- compare groups and carry out an enriched differential analysis, this could also be useful for treatments and responses to drugs or to inflammation.
- locate or estimate latent features that could be used to understand properties of biological systems
- create unique information for patients taken one by one that can be used for treatment analysis and diagnostic
- generate embeddings that could be used for additional analysis and inference. These features need to encode the information about the complex system that they address (for biological pathways, these features need to consider the gene pathways' structure and signaling information while also considering the singular behaviour of patients and single cells).

Strategies that can effectively and quickly predict the disease progression and strat-

ify patients for appropriate health care arrangements are urgently needed, as it has been possible to see in the past years with the covid-19 epidemic.

The research about single-cell sequencing is not confined to RNA-seq but it is the most prolific and useful way of using single-cell since technologies and platforms that perform single-cell sequencing are focused and precisely built to find accurate estimates of amounts of mRNA in particular. Ribonucleic acid (RNA) has multiple forms and plays a critical role in cell growth and differentiation and it is at the center of research since transcriptomic is at the core of biological properties and the functions of a tissue. RNA transcription and stability are tightly regulated in response to physiological and pathological stimuli. Abnormal expression of RNA is frequently associated with human cancer initiation, development, progression and metastasis. In addition to the mutation of tumor suppressor genes and oncogenes, gene expression could be overactivated or epigenetically silenced which could lead to uncontrolled tumor cell growth and proliferation. Aberrant activation of cell growth signaling pathways and/or transcription factors could lead to high-level expression of genes associated with tumor development and progression so knowing the properties and the activation of signaling pathways and expression is fundamental to treating and understanding disease and treatments. Different gene expression profiles may reflect different cancer subtypes, the stage of cancer development or tumor microenvironment. Therefore, RNAseq is a powerful tool for understanding the molecular mechanisms of cancer development and developing novel strategies for cancer prevention and treatment. Aside from coding RNA(mRNA that goes through translated in proteins), non-coding RNA (ncRNA) are thought to be key parts of gene regulatory(regulation of cell activity and proteins production) processes and their single-cell expression patterns may help dissect the biological function of single-cell variability. Technology for measuring ncRNA in single cells is still in development and most of the current single cell datasets have reliable measurements for only lncRNA. Although protein-coding sequences have thus far been the most highly studied sequences in the genome, non-coding sequences play crucial roles in a wide variety of cellular functions. Non-coding RNA is implicated in many different processes and pathways, including imprinting, differentiation, and cell cycle regulation. Since ncRNA are so important for pathways, regulation, differentiation and also communication among cells, there will be a further discussion about them in the context of the objectives previously defined in chapter 9.

This research will start by explaining both sequencing technologies (single-cell and bulk) by primarily introducing bulk with an **overview of bulk sequencing** that will give the first look at the bulk sequencing theory, methodology, and workflow in chapter 2. During the overview of bulk sequencing, there will be a focus on the differences with single-cell sequencing by highlighting the limitations and strengths of bulk sequencing in modern medicine and computational biology. To be specific, there will be an analysis of the applications of bulk sequencing and the impact of the new technologies, with a lot of attention to details and the latest problems and considerations related to the topic. Since most of the technologies and protocols used in bulk sequencing are reused in single-cell sequencing (with some proper changes made to address some properties of single-cell protocols), the research will delve into the details that are important for the single-cell methodology, in particular Next Generation Sequencing (NGS), sequencing

library preparation and a comparison of the latest technologies used in bulk and single-cell sequencing.

On the same note, single-cell sequencing will be introduced but, in any case, also deepened since it is the main topic of the thesis and also the less stable one (since methodologies and techniques are coming out even at this moment and, at the same time, standardization is not really popular for new technologies and leaves the place to the hegemony of the market and the procedures). This project will focus on modern scRNA-seq protocols to capture the expression of mRNA since the main application of single-cell sequencing is in the estimation of mRNA (and most of the platforms and databases available to do single-cell analysis use data collected from mRNA sequencing). With the wide range of technologies available, it is becoming harder for users to select the best scRNA-seq protocol/platform to address their biological questions of interest. In this research, there will be a discussion about the advantages and limitations of commonly used scRNA-seq platforms and methodologies in order to clarify their suitability for different experimental applications. The research will also address how the datasets generated by different scRNA-seq platforms and methodologies can be integrated, and how to identify unknown populations of single cells using unbiased bioinformatics methods. Single-cell methodologies and considerations will be seen in chapter 3. Along the lines of bulk sequencing, the research will highlight both the limitations and strengths of single-cell sequencing by also introducing some of the most advanced ideas and methodologies (along with some personal methodologies for differential expression analysis, pre-processing, and treatment of the data and other workflows that will be presented in chapters 3.5, 7 and 8 where the previously mentioned methods will be used on real data).

The rapid development of scRNA-seq technologies has led to the emergence of many methods for removing systematic technical noises, including imputation methods that will be introduced in section 3.3.2, which aim to address the increased sparsity observed in single-cell data. Although many imputation methods have been developed, there is no consensus on how methods compare to each other and on what criterion the method should be chosen.

Other technologies and protocols, such as bulk RNA-seq, could be used in combination with scRNA-seq. In this way, the information obtained from established, documented, and trustworthy methods like bulk sequencing can be combined with high accuracy and details obtained from scRNA-seq, flow cytometry, and other cutting-edge sources that have arisen in recent years. For example, bulk sequencing can be attributed to scRNA-seq via the use of **deconvolution** and other techniques that infer the hidden structural, quantitative, and qualitative information. These techniques will be seen in chapter 4.

As an additional overview of the theory for another important subject of interest that will be used for the objectives of this thesis, **biological networks** will be introduced in chapter 5, where the characteristics and a general presentation of diverse biological networks will be presented and linked to their use in personalized and translational medicine and computational biology. Algorithms related to the use of biological networks for embedding and extrapolation will be seen in chapter 6.

At the end of this research, the focus will shift to new techniques to gain knowledge

on scRNA-seq, bulk sequencing and expression analysis for genes and pathways (where the expression is substituted by the activity and the dysregulation of the pathway). This knowledge will be obtained with workflows and pipelines that will be built to find latent features that encode additional information on cell-to-cell interaction, topological information on signaling pathways, and how they differ from cell to cell (in reality from a group of cells to another group of cells) and spatial information as well. All of these discussions will be seen more in-depth in chapter 7.

In addition to delving into the methodology and theory of embedding and getting information in pathways from data coming from scRNA-seq and other sources, there will be an experimental analysis where real data will be analyzed and compared with traditional techniques that use bulk sequencing datasets. The experimentation will be seen in chapter 8, where a variety of sequencing and expression databases will be used to test the techniques and methodologies discussed during this research. The research will not end here with this project since new data need to be used for new field of application and the data that was used for this project was not consistent for the objectives of this project, more about the new data that will be used for future projects can be seen in the final chapter 9.

Chapter 2

Bulk sequencing overview

2.1 Bulk RNA sequencing

Since the start of the human genome project, bulk sequencing has become the most valuable and extensively used tool in understanding biology and biological processes. In particular, bulk RNA-seq is one, if not the most important instrument that can be used in translational and personalized analyses for patients and disease studies(cancer biology [[58],[158],[59]], inflammation[[24],[164]], autoimmunity[[101],[112]], etc).

Its diverse translational research and potential clinical applications have been well reviewed in the past and are used nowadays in every laboratory as a standard procedure to obtain useful data about patients and groups of patients. This section does not intend to discuss technologies, methodologies or their applications in detail. Instead, this research will highlight the bottlenecks of its clinical translation and the recent progress toward their solutions.

By RNA species, bulk RNA-seq involves sequencing two types of libraries: mRNA-only library and whole transcriptome library that includes all RNA species except for rRNA. By sequencing type, the most frequently used bulk RNAseq is a single-end short sequencing focused on differentially expressed genes to understand molecular mechanisms implicated in various stages of tumor genesis. This type of sequencing is simple and cost-effective, largely focused on mRNA only. The less routinely used type is paired-end longer sequencing aimed at additional knowledge on alternative splicing, point mutations, novel transcripts, long non-coding RNAs and gene fusions. This type of bulk RNAseq normally sequences rRNA-depleted libraries for more comprehensive information (Figure 2.1)

2.2 Objectives and techniques

Next-generation sequencing capabilities are improving at a shocking rate, especially after the pandemic of Covid-19 has started. RNA-seq studies continue to provide knowledge about the quantitative and qualitative aspects of transcriptomes in both prokaryotes and eukaryotes. Advances have been made in areas including the cataloging of sense and antisense transcripts, alternative splicing events, fused transcripts and transcription initiation sites in physiologically normal and anomalous settings.

Although several technological barriers still remain even after so many years of the standardization of RNA-seq protocols[114](especially for old protocols and platforms that have remained the same since the start of the millennium), major advances towards

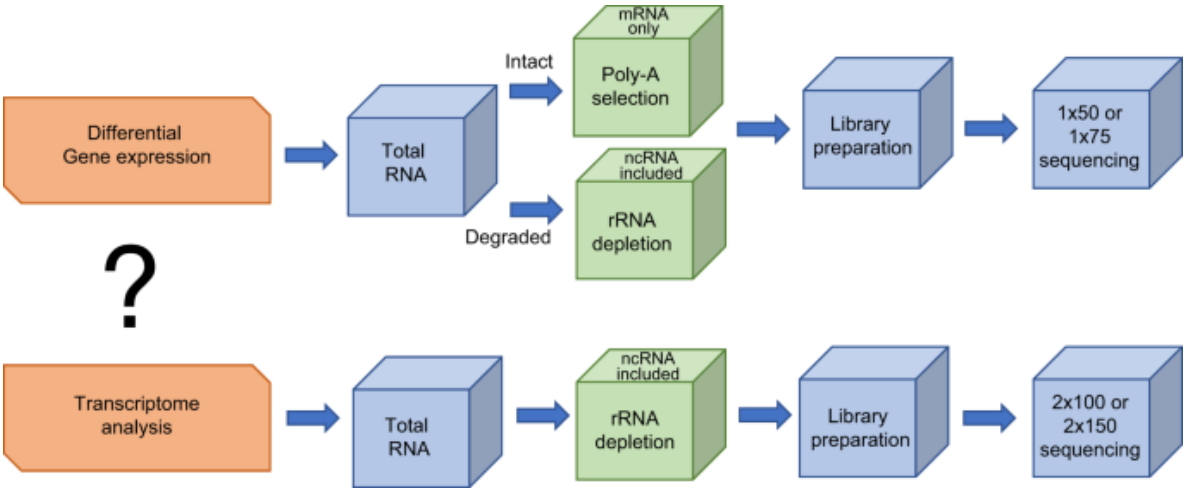


Figure 2.1: Two types of bulk RNA-seq libraries preparation

reliable analyses of RNAs from limited cell quantities have been achieved not too long ago[145], paving the way towards transcriptome profiling at the single-cell level. The standardization of modern protocols will be directly used in single-cell protocols since the true sequencing part of scRNA-seq is directly taken from bulk sequencing protocols and platforms, while the pipelines to isolate cells (pre-sequencing) and to generate expression levels for transcriptome (post-sequencing) are directly dependent on the platforms used and will be highly variable and not standardized in a general way. More about scRNA-seq protocols and pipelines will be seen in chapter 3, for now, the research will focus on explaining the most common protocols and platforms used for bulk RNA-seq.

The bulk RNA-sequencing techniques usually start by taking a sample tissue by biopsy that will be used as the starting point to the whole RNA-seq protocol (and also for single-cell protocols, even though the process will be different for the next steps). A biopsy is the removal of tissue in order to examine it. The tissue samples can be taken from any part of the body. Biopsies are performed in several different ways. Some biopsies involve removing a small amount of tissue with a needle while others involve surgically removing an entire lump, or nodule, that is the object of study (a part of the body, a tumor, etc.).

Often, the tissue is removed by placing a needle through the skin (percutaneously) to the area, but the methods are diverse and depend on the tissue that is sampled and the treatment/analysis that will be done (if the experiment needs a large amount of tissue, a larger section will be performed).

2.2.1 Sequencing library preparation

After the biopsy of the tissue, the treatment of the sample starts with the preparation of the sequencing library which involves taking the sample as input and results in a set of sequencing libraries composed of fragments of cDNA complete of **adapters** that have different roles and functions (PCR adapters, TSO, primers that will be used

for the sequencing step, UMI that are a type of molecular barcoding that provides error correction and increased accuracy during sequencing, **barcodes** related to the sample/cell used for multiplexing libraries, etc.).

The library preparation is composed by the following steps to obtain the sequencing library:

1. **RNA isolation**: during this step, the cells that are in the tissue are lysed and the RNA in the cells (and **outside the cells** as well, this is an important remark to do since single cells protocols isolate only RNA inside the cells) is isolated. There are some methods to isolate RNA [132], and nowadays there is a need to limit and avoid **confounders**(batch effects, different techniques used for experiments that should be the same for differential analysis, a change in the environment where the RNA-sequencing is performed), especially for single-cells protocols, but the common factor among the methods is the need to use an isolation kit (Qiagen [122], ThermoFisher [151], OmegaBiotek [113]) that also involve RNA purification (free of contaminants and inhibitors) and other pre-treatment of RNA to get better quality out of the experiment.
2. **RNA selection/depletion**: The isolated RNA can either be kept as is, filtered for RNA with 3' polyadenylated (poly(A)) tails to include only mRNA, depleted of ribosomal RNA (rRNA), and/or filtered for RNA that binds specific sequences to analyze signals of interest. The RNA with 3' poly(A) tails are mainly composed of mature, processed, coding sequences. Poly(A) selection is performed by mixing RNA with poly(T) oligomers covalently attached to a substrate, typically magnetic beads (Qiagen [121], Lexogen [86]). Poly(A) selection has important limitations in RNA detection since many **RNA biotypes are not polyadenylated**, including many non-coding RNA (ncRNA) and histone-core protein transcripts, or are regulated via their poly(A) tail length (e.g., cytokines) and thus might not be detected after poly(A) selection. This constraint about poly(A) detection is often one of the most important to consider, even more important for single-cell sequencing since the majority of techniques and platforms for single-cell sequencing use the detection of poli(A) as a mandatory step used as the base for building the primer sequence. More about this process will be seen in 3.1.
3. **cDNA synthesis**: This step is RNA is reverse transcribed to cDNA because DNA is more stable and to allow for amplification (which uses DNA polymerases) and leverage more mature DNA sequencing technology. This is also a mandatory step since the following steps require the sequences of cDNA. While direct sequencing of RNA molecules is possible, most RNA-Seq experiments are carried out on instruments that sequence DNA molecules due to the technical maturity of commercial instruments designed for DNA-based sequencing. Therefore, cDNA library preparation from RNA is a required step for RNA-Seq. The cDNA library preparation method varies depending on the RNA species under investigation, which can differ in size, sequence, structural features and abundance.
4. **Fragmentation and size selection** to fragment the RNA or cDNA in smaller sequences and filter out the sequences that are too long or too small, this is performed to purify sequences that are the appropriate length for the sequencing machine. The RNA, cDNA, or both are fragmented with enzymes, sonication, or

nebulizers. Fragmentation of the RNA reduces 5' bias of randomly primed-reverse transcription and the influence of primer binding sites, with the downside that the 5' and 3' ends are converted to DNA less efficiently. Fragmentation is followed by size selection, where either small sequences are removed or a tight range of sequence lengths are selected.

5. **Amplification**: is the process of amplification of the nucleotide sequence. This process is usually done via PCR or Linear Amplification. This step subsequent to reverse transcription results in various problems that should be addressed properly:
   - **Loss of strandedness**, which can be avoided with chemical labeling or single molecule sequencing.
   - **Amplification bias**, that is the amplification of the cDNA sequences will be exponential to the number of fragments that are already present, resulting in not detecting cDNA fragments that are not in high numbers in the sample (that means that the mRNA that is expressed in low quantities will not be detected, resulting in **dropouts**, a serious problem, especially in single-cell sequencing), which can be controlled with the identification of the single molecules pre-amplification (via UMI [[80], [140]] or other techniques that will be at the core of single-cell sequencing [67]) or by using different methods than PCR (e.g. Linear Amplification [57]).
   - **Contamination and contribution to confounding factors**, since amplification is usually done in a very specific environment while the other steps are done in different settings, the final results will be impacted by the contamination caused and confounding factors will be introduced. To account for contamination, there are some established and tested techniques pre-amplification and post-amplification, to get more in-depth see [[14],[165]].
6. **Indexing (OPTIONAL)** The cDNA for each experiment can be indexed with a hexamer or octamer barcode, so that these experiments can be pooled into a single lane for multiplexed sequencing.
7. **Normalization (OPTIONAL)** Normalization in next-generation sequencing (NGS) is the process of equalizing the concentration of DNA libraries for multiplexing. Multiplexing helps maximize the use of the capacity of NGS technology, enabling to run multiple (thousands) of libraries on a single **flow cell**, and driving down costs [28]. During this step, the possibility of amplification bias is also lowered. This step and most of the previous steps can be done with a single kit that does what is called tagmentation [63] (which uses bead-linked transposomes that bind to DNA fragments, more about tagmentation will be seen later in this chapter and also in the next chapter 3, where the method is an integral part of the methods of scRNA-seq that will be presented).

It is very important to understand that small RNA molecules (miRNA, siRNA, snRNA, etc.) are not in the final sequencing library with the method described above, there are a lot of methods to also take into consideration these molecules [[16],[17],[103]] but none of them will be seen here since the scope of the project is to build the foundations that will be used for single-cell sequencing and analysis, and single-cells protocols use specifically the poly(A) selection to also introduce other fundamental steps to the

protocol in the pipeline.

In the next chapter 3 about **Single cell sequencing**, the steps of preparation of the sequencing library will change and will be adapted for the **isolation** of single cells and the **identification** of the cells.

An important NGS library preparation protocol that needs to be introduced here and will also be used extensively in single-cell sequencing is **tagmentation on microbeads**(TOM, also referred to simply as tagmentation). This protocol consists of **Bead-linked transposome** (BLT , also called Tn5 transposome) chemistry that integrates DNA extraction, fragmentation, library preparation, and library normalization steps. This method is capable of recovering long-range information through tagmentation mediated by microbead-immobilized transposomes. Using transposomes with **DNA barcodes** to identically label adjacent sequences during tagmentation, the method can restore the inter-read connection of each fragment from the original DNA molecule by fragment-barcode linkage after sequencing. Transposases exist in both prokaryotes and eukaryotes and catalyze the movement of defined DNA elements (transposons) to another part of the genome in a "cut and paste" mechanism. Taking advantage of this catalytic activity, transposases are widely used in many biomedical applications like the one previously seen used for tagmentation. In this setting, an engineered and hyperactive Tn5 transposase from E. coli can bind to synthetic 19 bp mosaic end-recognition sequences appended to Illumina sequencing adapters (the "Tn5 transposome" as stated before) and will be utilized in an in vitro double-stranded DNA (dsDNA) tagmentation reaction (namely simultaneously fragment and tag a target sequence with sequencing adaptors) to achieve rapid and low-input library construction for next-generation sequencing.

The available product to conduct the tagmentation is only from Illumina [63] (as the main supplier and distributor of the tagmentation kit and protocol documentation under the identifier of **Nextera XT** or **Nextera Flex**).

At the end of the protocols to obtain the fragment of cDNA (from mRNA or RNA in general even though this research has not visited deeply the possibility of preparing the sequencing library with ncRNA and small RNA), the result is a set of sequencing libraries ready to be sequenced or that could pass through additional steps before the sequencing (they could be pooled/multiplexed if the molecules are identified by a barcode for the single sample, or before doing the sequencing analysis, a preliminary analysis could be done on the sample to measure an estimate of the quality and accuracy of the sequencing library).

The types of sequencing libraries are dictated by the goal of the experiment:

- the single short-read libraries are generally for differential gene expression and expression analysis in general since it is not important if the whole transcriptome is covered or not if the final objective is to obtain the number of reads for a gene
- the paired long-read libraries are for whole transcriptome analysis, including splicing variant and point mutation analysis in addition to analysis of differentially expressed genes.

2.2.2 Sequencing

After the library preparation, the fundamental step to do is the sequencing itself. DNA/RNA sequencing is the process of determining the nucleotide sequence (nucleic acid sequence) that is the order in which they appear in DNA or RNA. It includes any method or technology that is used to determine the order of the four bases: adenine, guanine, cytosine, and thymine (uracil for direct RNA sequencing). The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery. Direct RNA sequencing is less common than usual but is getting more recognition in recent years because converting RNA into cDNA, ligation, amplification, and other sample manipulations have been shown to introduce biases and artifacts that may interfere with both the proper characterization and quantification of transcripts. Since the methods used nowadays are standardized (documented and focused on maintaining the integrity of the sample) and proven to be pretty accurate [[110],[40]] (for next-generation bulk DNA-sequencing at least, especially when using some of the steps introduced in the previous section like indexing, purification and normalization along respecting the workflow defined by the supplier of RNA-seq kits, more research should be done on the accuracy of modern technologies as well though).

This research will not delve too much into the details of the protocols for DNA-sequencing, especially for old protocols like **Sanger sequencing** or such that are good for understanding the history and the base procedure but are incomparable with the latest methods in NGS that provide with high-throughput and accuracy with lower costs, even though Sanger sequencing and similar protocols still have the advantage over short-read sequencing technologies (like Illumina) in that they can produce DNA sequence reads of > 500 nucleotides and maintain a very low error rate with accuracies around 99.99

Sanger's method involved four extensions of a labeled primer by DNA polymerase, each with trace amounts of one chain-terminating nucleotide, to produce fragments of different lengths. The sizes of fragments present in each base-specific reaction were measured by electrophoresis on polyacrylamide slab gels, which enabled the separation of the DNA fragments by size with single-base resolution. The gels, with one lane per base, were put onto X-ray film, producing a ladder image from which the sequence could be read off immediately, going up the four lanes by size to infer the order of bases. The original method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotide triphosphates (dNTPs), and modified dideoxynucleotide triphosphates (ddNTPs), the latter of which terminate DNA strand elongation. These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. The ddNTPs may be radioactively or fluorescently labelled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP), while the other added nucleotides are ordinary ones. The concentration of reagents should be up to the sample that is sequenced

(both in quantity and in quality taken into consideration). The single reaction for the bases are done on different strips of gel. The DNA bands may then be visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation.

Right after Sanger sequencing came the shotgun sequencing protocol, that is sequencing of random clones followed by sequence assembly based on the overlaps. With shotgun sequencing and sequencing of fragments of DNA, there was a need for **alignment** tools that could take into consideration large number of data in a multi-alignment setting to recreate the original sequence. More about alignment will be seen in section 2.2.3.

Even though Sanger sequencing is one of the most precise methods for sequencing, the need to sequence large amount of DNA in a short amount of time and with a low cost is really important, especially for personalized medicine where the sequencing is done per patient and **scaling longitudinally** (by tracking the same sample at different points in time).

Since the need for speed of sequencing and low cost, the NGS technologies were being developed to address these issues. Also called "massively parallel", next-generation DNA sequencing (NGS) sharply depart from electrophoretic sequencing in several ways, but the key change is **multiplexing**. Instead of one tube per reaction, a complex library of DNA templates is densely immobilized onto a two-dimensional surface, with all templates accessible to a single reagent volume. Rather than bacterial cloning, in vitro amplification generates copies of each template to be sequenced. Finally, instead of measuring fragment lengths, sequencing comprises cycles of biochemistry (for example, polymerase-mediated incorporation of fluorescently labelled nucleotides) and imaging, also known as "sequencing-by-synthesis" (SBS). Although amplification is not strictly necessary the dense multiplexing of NGS, with millions to billions of immobilized templates, was largely enabled by clonal in vitro amplification.

Some techniques perform the amplification step on beads (emulsion PCR) while other techniques use a flow cell composed of surfaces covered with primers that bind to the single-stranded fragments and form bridges between the adapters of the strand (bridge PCR). The amplification step is very important to the whole sequencing since the steps of sequencing cycles are dependent on the medium used to trap and amplify the fragments.

For amplification on beads, single fragments are trapped with a single bead and the fragment is copied around the whole surface of the bead covered with primers, and after the beads are covered with the same fragments, they are trapped in a micro-well to undergo sequencing (details about sequencing during this step vary with the platform used, for more information use the original documentation of the platform).

For bridge PCR amplification, the strands form clusters of the same strands and the process of sequencing is paired (the final reads will be paired). The process of reading the nucleotide sequences is done automatically with computer vision.

A note about emulsion PCR is that it is a method that is very similar to what

will be seen in chapter 3.1 since one of the methods most used nowadays for single-cell sequencing is **microfluidic-droplet-based** that uses a similar concept to emulsion PCR, that is isolating cells (and not single DNA fragments like emulsion PCR) with beads in an emulsion of water-in-oil droplets.

Emulsion PCR is used in platforms like **IonTorrent** [152] and only supports single-strand sequencing for longer reads than Illumina platforms. Bridge PCR is used in platforms like **Illumina MiSeq** [64] and supports high throughput and large numbers of reads with shorter reads.

A visualization of the two methods of sequencing can be seen in figure 2.2 for the IonTorrent platform (even though it is taken from research about Roche/454 sequencer the principle are the same at least until the detection of the nucleotide during sequencing) and in figures 2.3 for the procedure of the Illumina platforms. For more information about sequencing see [[102],[79],[11],[136]] and the documentation provided by the sequencing platform providers.

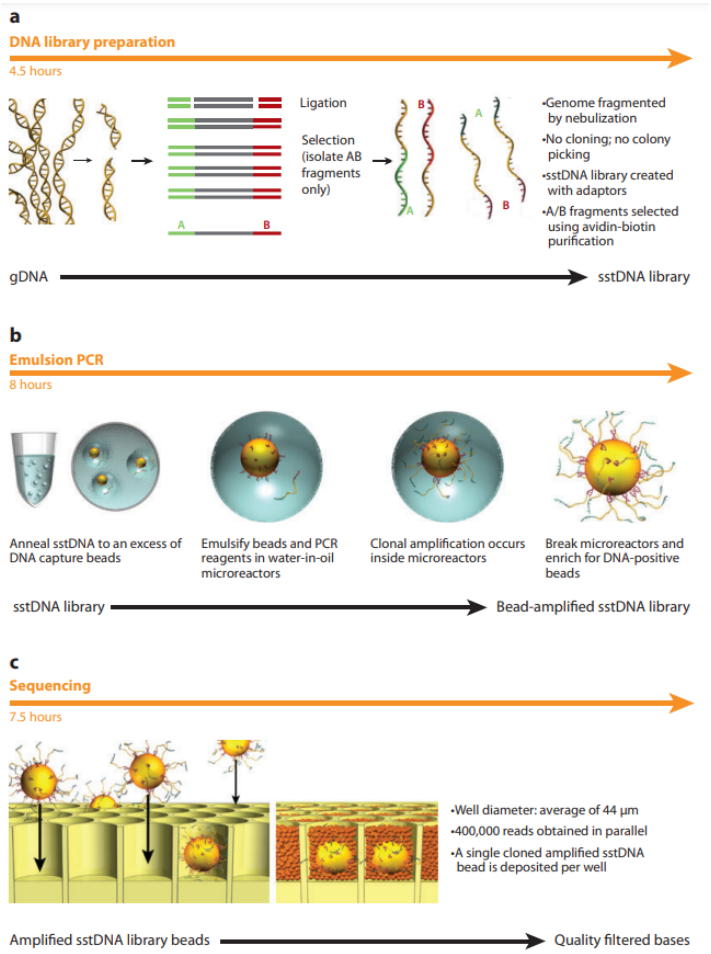
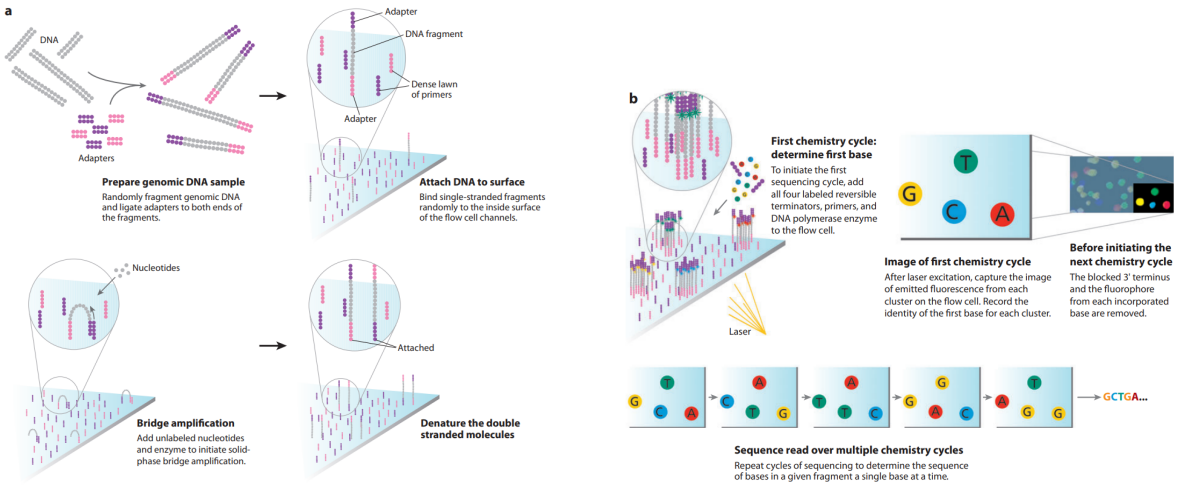


Figure 2.2: The method used by the Roche/454 sequencer to amplify single-stranded DNA copies from a fragment library on agarose beads, this method is really similar to the one used in IonTorrent platforms



(a) First step of Illumina sequencing

(b) Second step of Illumina sequencing

Figure 2.3: The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3 -OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that affect the cleavage of the fluorescent labels and the 3 -OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation

Other methods of sequencing exist (since the technology has advanced to address the need for higher accuracy and coverage, especially in single-cell sequencing), but they will not be seen here. Some technologies for third-generation sequencing are **PacBio** which is the sequencing platform of single molecule real-time sequencing (SMRT), based on the properties of zero-mode waveguides, and **Oxford Nanopore** which involves passing a DNA molecule through a nanoscale pore structure and then measuring changes in an electrical field surrounding the pore. These new technologies are aimed toward real-time sequencing and long reads. Long reads and real-time sequencing will be really important for single-cell sequencing and they will be addressed directly when talking about the challenges of single-cell sequencing in today's methodologies.

A figure to summarize NGS technologies and generations is 2.4, where the third generation of NGS can be seen.

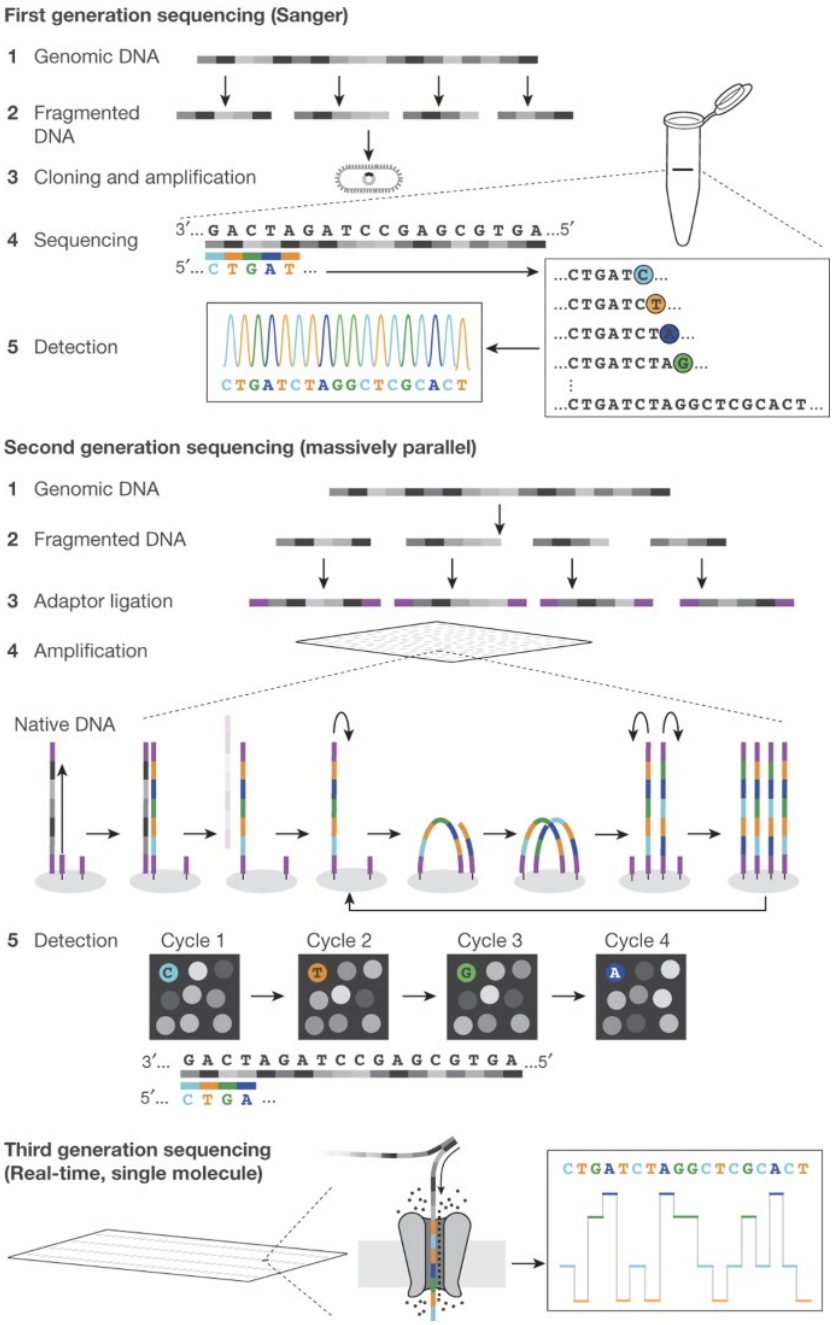


Figure 2.4: NGS technologies from first generation (Sanger sequencing), second generation (massively parallel with bridge PCR seen on the image, but also emulsion PCR) and third generation (also called long-read sequencing, under development right now)

After the DNA sequencing, the platform used will output a file in **FASTQ** that will contain the reads of the fragments along some additional information, the structure of a FASTQ file can be seen in figure 2.5. Also the composition of the header of FASTQ can be seen in figure 2.6.

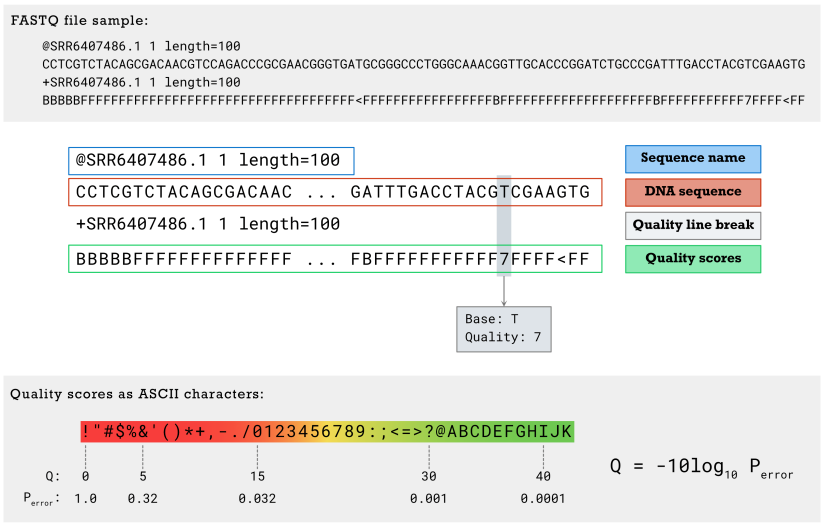


Figure 2.5

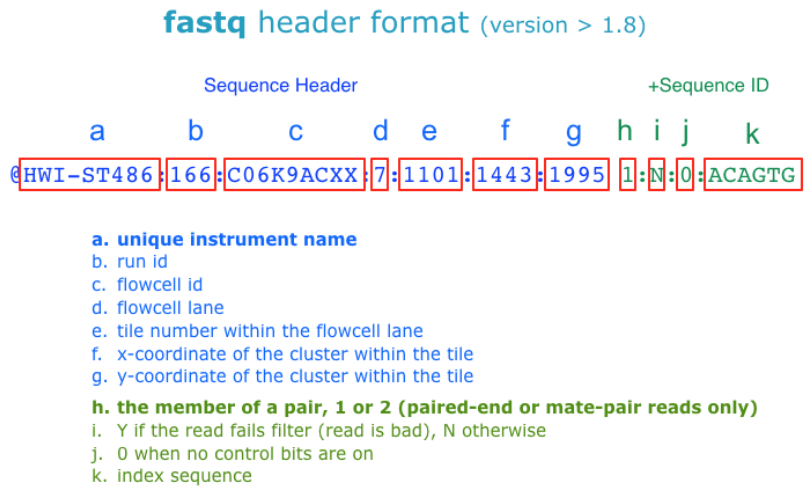


Figure 2.6

The next step after obtaining the reads of the fragments as a FASTQ file is alignment(without considering some additional steps like using other sources to integrate the data and filter/integrate the reads obtained with additional information, more about some conjunction methods to obtain better and accurate results will be discussed in section 2.5).

2.2.3 Alignment

To understand the process of RNA-seq and to summarize what was introduced here in this research as a first look into sequencing, there is a need to understand how the whole pipeline works and how the genes in DNA are expressed, this is summarized in figure 2.7.

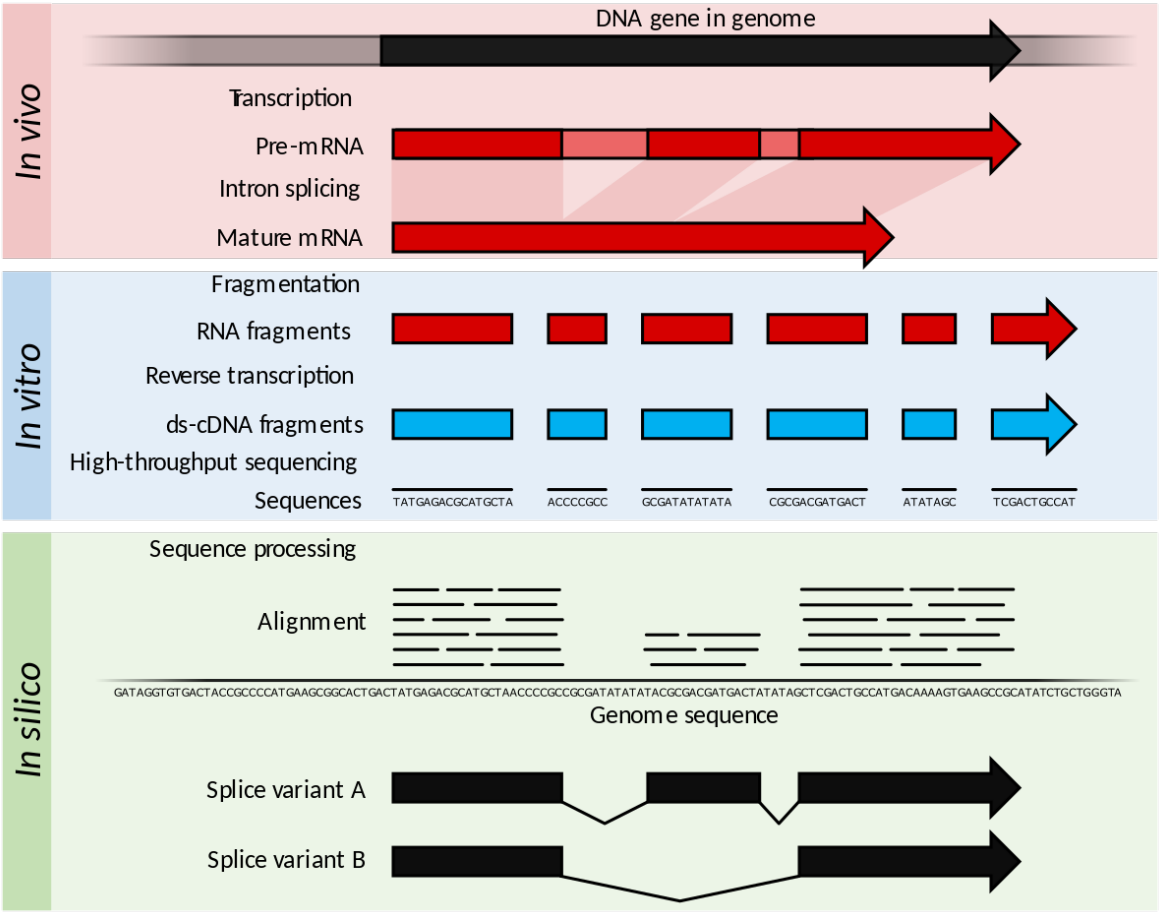


Figure 2.7: Summary of RNA-seq: Within the organism, genes are transcribed and spliced (Eukaryotic) to produce mature mRNA transcripts. The mRNA is extracted from the organism, fragmented and copied into stable ds-cDNA (via reverse transcriptase). The ds-cDNA is sequenced using high-throughput, short-read sequencing methods (to compute expression levels, otherwise long-read protocols are used). These sequences can then be aligned to a reference genome sequence to reconstruct which genome regions were being transcribed. This data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants. Source [163]

Rapidly evolving sequencing technologies produce data on an unparalleled scale. A central challenge to the analysis of this data is **sequence alignment**, whereby **multiple sequence reads must be compared to a reference and aligned**, this is done by aligning the unknown sequence with one or more known database sequences to predict the common portions. A wide variety of alignment algorithms and software have been subsequently developed over the past years since the start of the genome mapping project. During this section, the focus of the project will be to systematically review the current development of these algorithms and introduce their practical applications with particular attention to new types of data such as single-cell data. The research will also consider the future development of alignment algorithms with respect to emerging long sequence reads.

The objectives of the alignment are not only to align nucleotide sequences to a reference but involve the steps after the alignment that are:

- **Molecular Phylogeny**: that involves the comparative analysis of the nucleotide sequences of genes and the amino acid sequences and structural features of proteins from which evolutionary histories and relationships, and in some cases also functions, can be inferred.
- **Genome evolution and mutations**: that is also related to molecular phylogeny but is aimed at finding the specific differences and their consequences
- **Protein characterization**: to characterize the functions of proteins (both already known proteins and unknown proteins), formulating hypothesis on the functions and behaviour of proteins based upon known functions of other proteins or some other models.

An important point to make is that alignment uses a concept of **Similarity** between sequences that is also related to the concept of **Biological Omology** (Evolutionary distance known as similarity due to shared ancestry between a pair of structures or genes). These concepts will be used directly during alignment and depending on the final objective, one of the two will be better to use (even though they are quite similar in the definition, Homology is more of a **distance in graph theory** while Similarity is related to the concept of **edit distance**).

The optimum alignment arranges two or more sequences in such a way that a maximum number of identical or similar residues are matched. The sequences may be nucleotide sequences (DNAs or RNAs) or amino acid sequences (Proteins), for this research only nucleotides will be considered, but the algorithms, tools and methods are almost the same. The rearrangement process may introduce one or more spaces or gaps in the alignment. A gap indicates a possible loss or gain of a residue; thus, evolutionary insertion or deletion (indel that is usually marked by the - sign), translocations and inversion events can be observed in Sequence alignment.

The method of Sequence Alignment can be of two types: **global alignment** and **local alignment**. Global alignment is done when the similarity is counted over the entire length of the sequences (the sequences are aligned over their entire length). Several MSA techniques accomplish global alignment, but difficulties arise when sequences are only homologous over local regions where a clear block of ungapped alignment is common to all of the sequences or if there is the presence of shuffled domains among the related sequences. In such cases, local alignment is performed to know the local similar regions among the sequences. When there is a large difference in the lengths of the sequences to be compared, local alignment is generally performed.

An important part of alignment methods and algorithms is the scoring method used to assign scores to the compared sequence and alignment. Different scoring methods are used in the sequence alignment to know the level of identity or similarity. Nucleotide scoring is a simple identification scheme where identical bases in both sequences are assigned positive scores. In contrast, for protein, a similarity score is also counted (along with an identity score) denoting the amino acids having similar physicochemical properties. The substitution matrices mostly consulted for protein sequence alignment are Point Accepted Mutation (PAM) [43], and BLOcked SUbstitution Matrix (BLOSUM) [161].

Alignment uses a scoring function to find the best alignment. One of the most used scoring functions is the sum of the similarities between the aligned sequences, that is

formally:

Definition 2.2.1. (Pairwise alignment scoring function). Given an alphabet Ω, two sequences $S_1 \in \Omega^n$ and $S_2 \in \Omega^n$ (the two sequences must be of the same length), a similarity function $\sigma : \Omega x \Omega \longrightarrow \mathbb{R}$, and given that $|S_1| = |S_2| = n$, the scoring function is the following:

$$\Theta : \Omega^n x \Omega^n \longrightarrow \mathbb{R}$$

$$\Theta(S_1, S_2) = \sum_{i=0}^{n} \sigma(S_1[i], S_2[i])$$

Some important things to take into consideration when creating the scoring function (a scoring function that takes into consideration the neighbors' sequences and not the punctual comparison of sequences, so not like the function defined in 2.2.1) are the gap penalties. Creating a gap should have a lesser value than expanding a gap since the final alignment will be more compact and the gaps can be thought of as introns or parts of the DNA that are not transcribed or are spliced. Also, since the DNA is fragmented in continuous sections of nucleotides from the original strand, the final reads should have continuous gaps before or after the read since the strand was fragmented in different **continuous** sections.

The alignment of two sequences (not globally or locally optimal by design) is defined formally as follows:

Definition 2.2.2. (Pairwise alignment). Given an alphabet Ω that contains the INDEL symbol **-** and two sequences $S_1 \in \Omega^n$ and $S_2 \in \Omega^m$, an alignment of the two sequences is a function $Pair : \Omega^n x \Omega^m \longrightarrow \Omega^l x \Omega^l$ that maps $S_1 \longrightarrow S_1' \in \Omega^l$ and $S_2 \longrightarrow S_2' \in \Omega^l$ such that

- $|S_1'| = |S_2'| = l$
- Removing the INDEL symbols in S_1' and S_2' results in S_1 and S_2

The set of pairwise alignment over the sequences of length n and m is defined as $setPair_{n,m} = \{f : \Omega^n x \Omega^m \longrightarrow \Omega^n x \Omega^n | f \text{ is a pairwise alignment}\}$

A globally optimal alignment is defined as follows

Definition 2.2.3. (Globally Optimal Pairwise alignment). Given an alphabet Ω that contains the INDEL symbol **-**, two sequences $S_1 \in \Omega^n$ and $S_2 \in \Omega^m$ and a scoring function, a globally optimal alignment of the two sequences is a pairwise alignment function $OptPair : \Omega^n x \Omega^m \longrightarrow \Omega^n x \Omega^n$ such that

- $OptPair \in setPair_{n,m}$
- $OptPair(S_1, S_2) = (S_1', S_2')$
- $\underset{(S_1*, S_2*) = f(S_1, S_2) | f \in setPair_{n,m}}{max} \Theta(S_1*, S_2*) = \Theta(S_1', S_2')$

Locally optimal alignment is done on subsequences of the two sequences. A tool that is very important for local alignment is **BLAST** (Basic Local Alignment Search Tool) [[7],[157]]. BLAST performs comparisons between pairs of sequences looking for regions of similarity, rather than global alignment between entire sequences, so it

performs local alignment in different subsequences of the two sequences. BLAST can perform thousands of comparisons between sequences in a few minutes and in a short time it is possible to compare a query sequence with the entire database to search for all sequences similar to it.

Alignment is also required for Gene annotation done by analyzing short RNA-seq reads derived from mRNA and mapping them to the reference genome. The sequence reads must evenly cover each transcript along both ends. But the most challenging part is the Sequence reads are much shorter than the biological transcripts. Therefore, short reads are needed to align themselves to find the common end sequences among the reads and thus assemble them to construct the entire transcript and its coverage in the reference genome along with the splice sites.

Sequence alignment at amino acid level is more relevant than nucleotide level as protein is the key functional biological molecule and hence carries structural and/or functional information, but the results that these research needs are transcript level expression that will be used to compare and generate some differentially expressed genes that will be used in the creation of signatures from meta-pathways for single cells and bulk samples. Transcript level expression could be obtained as well from protein alignment but it is not the scope of the project for now (for more information and future outlooks see 9)

For PSA, Dynamic Programming (DP) always provides the optimum alignment for a given objective function by finding the optimal alignment path using the trace-back process (Needleman-Wunsch), an example of pairwise alignment can be seen in figure 2.8. The objective function is used for assessing the quality of alignment of a set of input sequences.

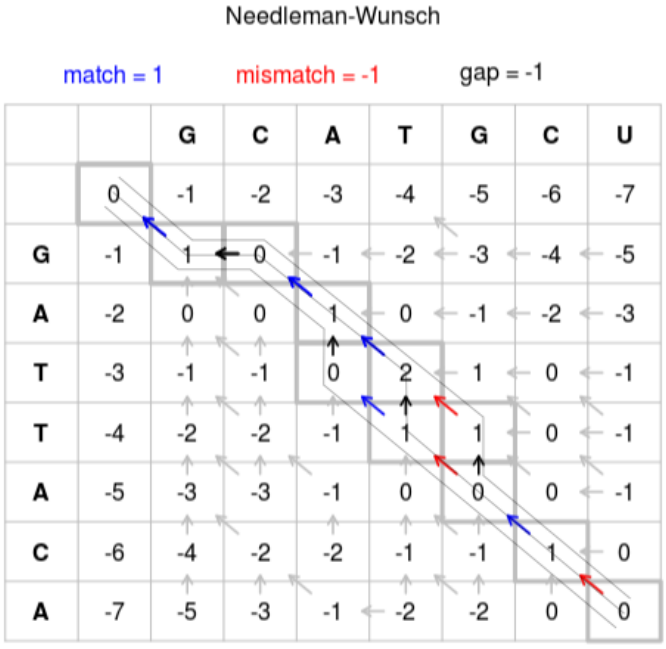


Figure 2.8: Example of Needleman-Wunsch pairwise sequence alignment

The algorithm for Needleman Wunsch uses Dynamic programming and considers subsequences alignment with the following recursive rule to compute the F matrix:
Basis:

$$F_{0,j} = \sigma(" - ", S_2[j]) * j$$
$$F_{i,0} = \sigma(S_1[i], " - ") * i$$

Recursive, principle of optimality

$$F_{i,j} = max(F_{i-1,j-1} + \sigma(S_1[i], S_2[j]), F_{i-1,j} + \sigma(" - ", S_2[j]), F_{i,j-1} + \sigma(S_1[i], " - "))$$

The algorithm for Needleman-wunsch can be seen in 1

Algorithm 1: Needleman-Wunsch pairwise alignment algorithm

Input: Sequence S_1, Sequence S_2, Similarity function σ
Output: F matrix
$F_{0,j} \longleftarrow \sigma(" - ", S_2[j]) * j, \forall j \in [0, |S_2|]$;
$F_{i,0} \longleftarrow \sigma(S_1[i], " - ") * i, \forall i \in [0, |S_1|]$;
for *i=1 to* $|S_1|$ **do**
  for *j=1 to* $|S_2|$ **do**
    Match $\longleftarrow F_{i-1,j-1} + \sigma(S_1[i], S_2[j])$;
    Mismatch1 $\longleftarrow Fi - 1, j + \sigma(" - ", S_2[j])$;
    Mismatch2 $\longleftarrow Fi, j - 1 + \sigma(S_1[i], " - ")$;
    $F_{i,j} \longleftarrow$ max(Match, Mismatch1, Mismatch2)
  end
end

The variable for *Mismatch1* and *Mismatch2* are also called **deletion** and **insertion** since they could be seen as a deletion of a character(nucleotide) in the first sequence or an insertion of another character always in the first sequence.

The algorithm to get the alignment for the two sequences can be seen in 2

Algorithm 2: Get the two aligned sequences

Input: Sequence S_1, Sequence S_2, matrix F from Needleman-Wunsch, similarity function σ

Output: S'_1 , S'_2

$S'_1 \longleftarrow$ "";
$S'_2 \longleftarrow$ "";
$i \longleftarrow |S_1|$;
$j \longleftarrow |S_2|$;
while $i > 0 \vee j > 0$ **do**
  if $i > 0 \wedge j > 0 \wedge Fi, j == F_{i-1,j-1} + \sigma(A_i, B_j)$ **then**
    $S'_1 \longleftarrow S_1[i] + S'_1$;
    $S'_2 \longleftarrow S_2[j] + S'_2$;
    $i \longleftarrow i - 1$;
    $j \longleftarrow j - 1$;
  end
  else if $i > 0 \wedge F_{i,j} == F_{i-1,j} + \sigma(" - ", S_2[j])$ **then**
    $S'_1 \longleftarrow S_1[i] + S'_1$;
    $S'_2 \longleftarrow " - " + S'_2$;
    $i \longleftarrow i - 1$;
  end
  else if $j > 0 \wedge F_{i,j} == F_{i,j-1} + \sigma(S_1[i], " - ")$ **then**
    $S'_1 \longleftarrow " - " + S'_1$;
    $S'_2 \longleftarrow S_2[j] + S'_2$;
    $j \longleftarrow j - 1$
  end
end

The optimized alignment function is rarely biologically optimum when more than three sequences are considered as in **MSA** since globally optimal alignment for multiple sequences is NP-complete [159] and therefore the exact method is not applied in MSA for all but unrealistically small datasets. The computation is also a complex task and demands high computer resources. Moreover, DP suffers from high-dimensional problems in MSA, as the number of sequences is equal to the number of dimensions. If two or more optimal paths are available and need to trace backward, the complexity of the back tracing grows exponentially. Most MSA procedures are therefore heuristics or approximate in nature, which provides feasible alignment solutions within a short and limited timeframe.

As available heuristics do not provide the best solution and because of rapidly growing database sizes, the development of new high-performing MSA methods to find good sequence alignment is still under research. MSAs are often hard to achieve because of the complex relationship that often exists among related sequences, and sometimes because of a lack of evolutionary history.

Three categories of approaches are frequently used in MSA:

- **exact alignment** that usually delivers high-quality alignment that is very close to optimal. It tries to **simultaneously align multiple sequences** and thus

needs to depend on DP. Due to the drawback of the exact method in alignment as stated above, most MSA follows the other two categories.

- **progressive alignment** is a heuristics approach where complex MSA problem is separated into subproblems. This solves the direct MSA problem indirectly with PSA. This approach assembles all sequences progressively where the best pairwise alignment is first taken into account. Progressive alignment uses **guide tree** to solve MSA problem where each leaf represents a sequence to be aligned [38]. To build the tree, UPGMA or neighbor-joining methods are applied. Each visited internal node is associated with an MSA of the sequences in its corresponding subtree. Finally, MSA of all considered sequences is associated with the root node (an example of a guide tree can be seen in figure 2.9, where the first alignment chosen is the one on s4 and s5, then pairwise alignment is computed on the s1 and s3, and after that MSA is done on the alignments of s3 and s1, and on the sequence of s2, and so on). Progressive alignment technique is used in several alignment programs such as MULTAL, MAP, CLUSTAL [56], and others. Among them, the most widely used method is ClustalW [[153],[62]]. It first performs the global pairwise alignment of the sequences and develops a distance matrix. It then builds a guide tree based on the matrix values. Finally, it generates a consensus alignment by gradually adding sequences following the guide tree where the closest sequence pairs (smallest branch length in the guide tree) are aligned first and thus, it gradually adds the next sequences. However, the greedy nature of these approaches cannot allow for modifying the gaps and hence, the alignment cannot be altered in the later stage. Some drawbacks of progressive alignment are that local minima are traps to the algorithms since the procedures use a greedy algorithm (it is a heuristic that finds a good solution but not the best solution to the problem of MSA), another drawback is that any progressive MSA is influenced by the initial alignment. As a result, any error made at that stage is propagated to the final MSA results. On the other hand, the iterative alignment methods iteratively modify the alignment by realigning the sequences or sequence groups and thus, overcome the drawback of the progressive method.
- **iterative alignment** generally performs post-processing by making changes in the alignment made by progressive methods. The methods of iterative alignment usually modify the construction of the guide tree. They compute new distance matrices using an MSA obtained by progressive alignment. This, in turn, constructs a new guide tree that leads to a second round of progressive alignment. Some iterative methods perform by repeatedly dividing the aligned sequences into two groups and then realigning those groups until the alignment process is converged

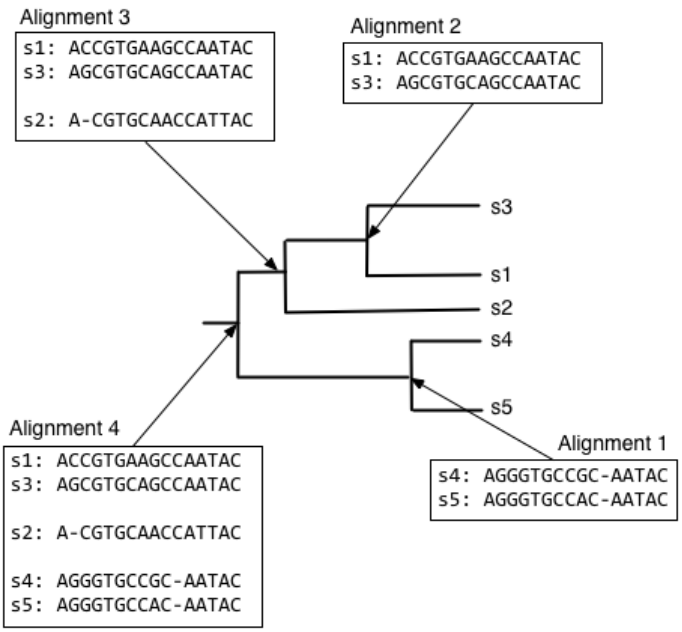


Figure 2.9: Example of a guide tree used for progressive alignment

Other models of alignment can be found and are also under development since the interest for the field of searching for better algorithms and methodologies in MSA is always really in high demand. Some of these additional alignment models are **Hidden Markov model-based alignments**, **Genetic Algorithms** and **Stochastic approaches**, for more information about these methodologies see [[88],[23],[26]].

One thing to take into consideration during alignment is the final results of the sequencing process, that is paired-end reads or single-end reads. If the final reads are paired the alignment process should consider one of the two paired ends as the inverse and maintain this behavior through the alignment of the reads.

It is also possible to align the RNA reads directly on the reference transcripts. One of the most modern and powerful algorithms that allow for the alignment of the RNA reads directly on the reference transcripts is Salmon[115]. Salmon allows both alignments on transcripts and counting of reads. It is used for quantifying transcript abundance from RNA–seq reads. Given a FASTQ file in input, a matrix is obtained directly with the counts of the reads for each gene transcribed. This procedure is faster than the alignment on the genome and it is recommended to use Salmon for mRNA analysis and this tool will be used directly during the experimentation in chapter 8. More about the computation of gene expression directly from the unaligned reads will be seen in the following chapter 2.3.

After aligning the sequences to a reference (genome or transcriptome) the resulting file will be a **SAM** (sequence alignment map) or **BAM** file that will contain the reads aligned to the reference. After obtaining the aligned reads, the next step should depend on the objectives (mutation identification, phylogeny, etc.) but for this project, the objective is to obtain the expression of genes (that will be seen in the next section) and use them to compare cells (and groups of patients) to obtain the differentially expressed

genes.

2.3 Gene expression

Estimating transcript abundance is a fundamental task in genomics. These estimates are used for the classification of diseases and their subtypes, understanding expression changes during development, and tracking the progression of cancer. Accurate and efficient quantification of transcript abundance from RNA–seq data is an especially pressing problem owing to the wide range of technical biases that affect the RNA–seq fragmentation, amplification, sequencing process, the exponentially increasing number of experiments, and the adoption of expression data for medical diagnosis as was already introduced during the previous chapters when talking about technical noise/bias and confounding factors.

Gene expression analysis provides quantitative information about the population of RNA species in cells and tissues. It is an exceptionally powerful tool of molecular biology that is used to explore the basic biology of specific tissues and understand the differentiation among different samples(tissues or samples of the same tissue but different patients with different conditions). Expression analysis is among the most commonly used methods in modern biology. There are over 750,000 expression data sets in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) public database.

In the past, gene expression was measured by the **DNA Microarray** technologies which have been most frequently used for expression analysis, consisting of millions of individual oligonucleotide probes fixed to a solid surface. The oligonucleotide probes typically have sequences representative of known RNA species and are generally used to compute the quantity of the relative levels of RNA species that hybridize with the probes. NGS technologies seen in the previous sections, on the other hand, provide a measure of the frequency of RNA species through sequencing of RNA-derived cDNA populations. Other approaches, such as digital molecular barcoding, represent a fusion of the hybridization and counting approaches. For instance, the nCounter digital quantification platform relies on the hybridization of labeled probes to RNA molecules and single-molecule imaging to provide a measurement of the frequency of particular RNA species.

Using NGS technologies output as the base (the BAM files), the gene expression of a single sample is computed by counting the genes. The counting starts from the aligned reads and a reference annotation genome(or transcriptome) that is contained in a **GTF/GFF** file that contains the position of the genes and other features(name, variants, versions, etc.). When dealing with different types of sequencing technologies, one needs to take into account how the reads are composed and pre-process the data in accord (if the sample was barcoded and pooled, the reads should be demultiplexed, and if the sample had UMI sequences appended as well, these are needed for aligning and especially counting, since they can be used to normalize the expression directly and limit the effects of amplification bias). Tolls such as the ones that will be presented now have some options to take into account the presence of adapters, barcodes and UMIs

(almost all of them have the possibility to see if the reads have barcodes), but most of the time some pre-processing with other tools and hand-written code is necessary.

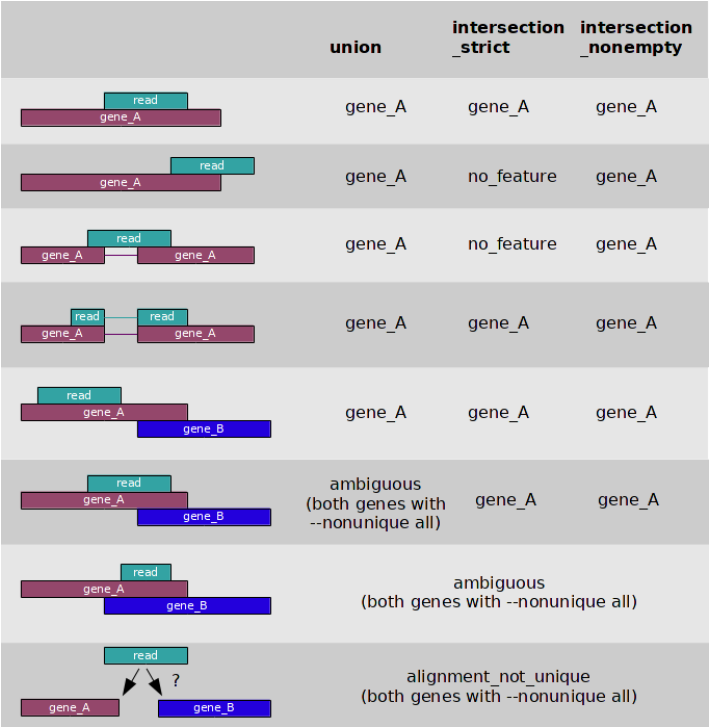


Figure 2.10: Htseq-count modes. Source [61]

Another thing to take into account is the quality of the reads and of a set of reads. Under this problem fall the practices of **Quality control** that will be very important in case of contamination of some sort or when the final reads are not representative of the sample due to technical limitations or constraints depending on the types of experiment done. Quality control will be very important for single-cell sequencing since the methods introduced in this chapter and used for bulk RNA-seq are not very effective for scRNA-seq. A complete discussion about Quality control for scRNA-seq will be seen in 3.3.

Read counting algorithms allow the count of the number of reads that are aligned in a specific genomic region and take as an input the two files described previously (along with some other files like indexes or some additional options). A variety of algorithms to count the reads exists, the three more popular algorithms are the following

- **Htseq-count** [9]: a tool from a Python library (HTseq) that preprocesses RNA-Seq data for differential expression analysis by counting the overlap of reads with genes. A representation of the modes in which the genes expression is assigned from aligned reads can be seen in figure 2.10. For more information about the modes see [61].
- **FeatureCounts** [92]: a read summarization program suitable for counting reads generated from either RNA or genomic DNA sequencing experiments. featureCounts implement highly efficient chromosome hashing and feature blocking tech-

niques. It works with either single or paired-end reads and provides a wide range of options appropriate for different sequencing applications.

- **STAR** [32]: A tool for alignment and also for counting the number of reads per gene. With *–quantMode GeneCounts* option STAR will count the number of reads per gene while mapping. A read is counted if it overlaps (1nt or more) one and only one gene. Both ends of the paired-end read are checked for overlaps. The counts coincide with those produced by htseq-count with default parameters.

Another tool that is used for quantification (and not directly counting) is **Salmon**[115]. As already introduced before at the end of the alignment section, salmon is a tool that aligns the reads to the transcript and quantifies the gene expression based on a complex model. Salmon combines a dual-phase parallel inference algorithm and feature-rich bias models with an optimized("ultra-fast" as the authors say) read mapping procedure. It is the first transcriptome-wide quantifier to correct for fragment GC-content(Guanine Cytosine content in the transcriptome) bias. If GC-content bias is not taken into account, these biases can lead to undesired effects, for example, a loss of false discovery rate (FDR) control in differential expression.

Salmon consists of three components that carry out different functions of the algorithm to reach good results for the alignment and counting/quantification process: a lightweight mapping model, an online phase that estimates initial expression levels and model parameters, and an offline phase that refines expression estimates. This two-phase inference procedure allows Salmon to build a probabilistic model of the sequencing experiment that incorporates useful information, like terms contributing to the conditional probability of drawing a fragment of a given transcript. Salmon is capable of either mapping sequencing reads itself by using a fast and lightweight procedure called quasi-mapping(with pseudo-alignment and a reference transcriptome) or accepting precomputed read alignments in the form of a SAM or BAM file, even though it is important that the precomputed read alignments are aligned to the same **transcriptome** as the one used for counting/quantification.

Salmon will be used in the experimental analysis during chapter 8 to align and count both bulk data(from samples) and reads from isolated cells.

2.3.1 Normalization

The quantification of the reads results in a digital measure of the abundance of transcripts. The normalization of these units is necessary to remove technical errors in the sequencing data such as the different depths of the sequencing (more depth produces more read counts for a gene expressed at the same level as another) and the different lengths of the different transcripts analyzed (different lengths in genes generate an unequal number of reads counts for genes expressed at the same level, longer genes will have more read counts). Some methods for normalization are the following:

- **RPKM or FPKM** The measure RPKM (reads per kilobase of exon per million reads mapped) was devised as a within-sample normalization method; as such, it is suitable to compare gene expression levels within a single sample, rescaled to correct for both library size and gene length while FPKM stands for fragments per kilobase of exon per million mapped fragments. It is analogous to RPKM

and is used specifically in paired-end RNA-seq experiments. The formula is the following:

$$RPKM_i \text{ or } FPKM_i = \frac{q_i}{\frac{l_i}{10^3} * \frac{\sum\limits_j q_j}{10^6}} = \frac{q_i}{l_i * \sum\limits_j q_j} * 10^9$$

where q_i are raw read or fragment counts, l_i is feature (i.e., gene or transcript) length, and $\sum\limits_j q_j$ corresponds to the total number of mapped reads or fragments

- **TPM** stands for transcript per million, and the sum of all TPM values is the same in all samples, such that a TPM value represents a relative expression level that, in principle, should be comparable between samples

$$TPM_i = \frac{\frac{q_i}{l_i}}{\sum\limits_j \left(\frac{q_j}{l_j}\right)} * 10^6$$

where q_i denotes reads mapped to transcript, l_i is the transcript length, and $\sum\limits_i (\frac{q_j}{l_j})$ corresponds to the sum of mapped reads to transcript normalized by transcript length.

The only difference between RPKM and FPKM is that FPKM considers the read count in one of the aligned mates if paired-end sequencing is performed. TPM is a modification of RPKM in which the sum of all TPMs in each sample is consistent across samples (exonic read×mean read length×106/exon length×total transcript). This approach makes comparisons of mapped reads for each gene easier than PKM/FPKM-based estimates because the sum of normalized reads in each sample is the same in TPM. These library-size-based normalization methods may be insufficient, however, when detecting differentially expressed genes. Consider the case when two genes are being expressed in two conditions (A and B). In condition A, the two genes are equally expressed, whereas in condition B, gene B has two-fold higher expression than gene A. If this absolute expression is converted into relative expression, one might conclude that gene A is differentially expressed, although this effect is only a consequence of its comparison with gene B. As observed previously, if a particular set of mRNAs is highly expressed in one condition and not in the other, non-differentially genes may be falsely identified as consistently down-regulated.

For more information about other normalization methods see [170].

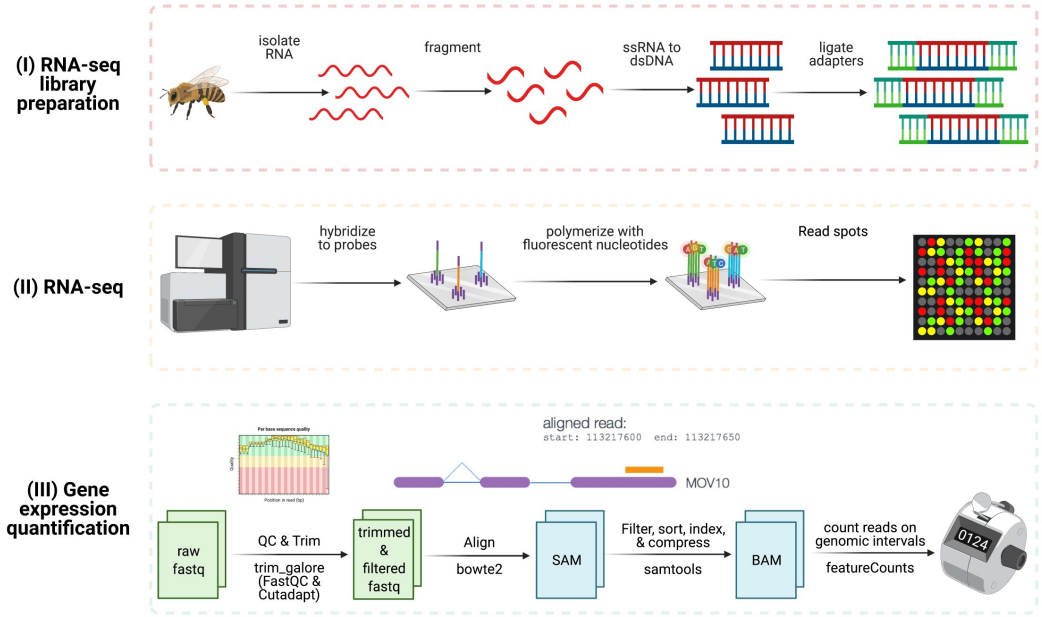


Figure 2.11: Example of the full workflow of RNA-seq, a note to keep in mind is that bowtie[85] is not ideal for RNA-seq so it is advised to use instruments like salmon or STAR [32]

The methods used for bulk sequencing are quite diverse and also take into account the characteristics of the sample. In the next chapter about single-cell, the methods introduced in this section will not be enough to satisfy some conditions of single-cell but, at the same time, will be enough for most applications since the perfect normalization for single-cell need a lot of different data to be able to work properly and generate consistent normalized expression values across different cells and samples. More about scRNA-seq normalization will be seen in 3.3.

A full workflow that starts from a sample and ends up in the gene or transcript expression can be seen in figure 2.11

In the next section, the focus will change from how to obtain transcript expression levels to how to compare different samples from the levels obtained in this section.

2.4 Differential expression

The comparison between groups of samples allows the identification of differentially expressed genes and we, therefore, speak of differential expression analysis. Differential expression analysis is a field that aims at comparing two or more groups of people to find how a single group of people differ from the other groups in terms of gene expression within and outside of the group. The identified genes (or transcripts) that are identified as differentially expressed are characterized by a **fold change (FC)** which indicates the intensity of the alteration of the expression compared to another sample (the sample denomination is usually **control** for the normal sample that serves as the main group to confront other samples that are usually called **altered**, but these definitions are only when the differential expression analysis is done on these kinds of

samples, and that is not always the case as it will be seen in 3.4, to keep the difference between the two kinds of sample they will be called **target** sample since it is used as a reference for the "controls" and **source** sample for the "altered"). Fold change is computed as a proportion of the increment from source to target, that is, given two values $source, target \in \mathbb{R}$, $\frac{source}{target}$, for groups of values in two different categories, the fold-change is usually computed as the ratio of the means(geometric or other central tendencies metrics). This fold-change measures the dysregulation of the gene from the compared sample and its sign identifies **down-regulated (- sign)** or **up-regulated (+ sign)** genes. Log2-fold-change is actually used more since it translates the ratio of change in a more controllable metric. One thing to take into consideration is that the dysregulation is anti-commutative, which means that if a source sample is dysregulated for some genes, these genes will also be dysregulated for the target if the point of reference is the source, the dysregulation of these genes will be opposed though, that means that up-regulated genes in the source sample will be down-regolated in the target sample if the point of reference is the source sample. This consideration seems obvious but, at the same time, it is really important to understand that when the samples are not "control" and "altered", the differential expression analysis results will depend on the point of reference.

The FC alone does not give a measure of significance, which is given by a p-value, which is usually corrected to limit the number of false positives (type I errors) in multiple tests, with the Benjamini-Hochberg correction method for the False Discovery Rate (FDR).

FC and corrected p-value are calculated from a statistical analysis that takes into consideration the sample (or samples) of interest with respect to a target(control) and varies according to the number of **technical**(repeated measurements of the same sample) or **biological**(Measurements of biologically distinct samples) replicas available and the experiment was done and what are the objectives. It is essential to start from correct expression values (usually normalized counts on the total or TPM) to allow a correct comparison between the samples.

Some of the most used algorithms to perform differential expression are the following:

- **LIMMA** [[125],[142]]: An R/Bioconductor software package that provides an integrated solution for analysing data from gene expression experiments. It contains rich features for handling complex experimental designs and for information borrowing to overcome the problem of small sample sizes. It is a package for differential expression analysis of data arising from microarray experiments. The package is designed to analyze complex experiments involving comparisons between many RNA targets simultaneously while remaining reasonably easy to use for simple experiments. It contains particularly strong facilities for reading, normalizing, differential splicing analyses, and exploring such data. The central idea is to fit a linear model to the expression data for each gene. The expression data can be log-ratios, or sometimes log-intensities, from two-color microarrays or log-intensity values from one-channel technologies such as Affymetrix [65]. More information about it in the references previously cited.
- **DESeq2** [97]: An R package that integrates methodological advances with several additional features to facilitate more quantitative analysis of comparative RNA-

seq data using shrinkage estimators for dispersion and fold change. This package is really important since it is the one used for the experiments in 8, both for the bulk data and the single-cell data.

- **edgeR** [126]: It is a Bioconductor software package for examining differential expression of replicated count data. An overdispersed Poisson model is used to account for both biological and technical variability. Empirical Bayes methods are used to moderate the degree of overdispersion across transcripts, improving the reliability of inference. The methodology can be used even with the most minimal levels of replication, provided at least one phenotype or experimental condition is replicated.

The three protocols of limma, DESeq2 and EdgeR are similar but have different steps among the processes of the analysis. For example, a linear model is used for statistics in limma, while the negative binomial distribution is used in edgeR and DESeq2. Additionally, the normalized RNA-seq count data is necessary for EdgeR and limma but is not necessary for DESeq2.

A very useful resource that can be seen is [93] where the three methods are compared and their field of usage are stated. In that research, the results of the three methods showed that DESeq2 and EdgeR obtain more DEGs than limma. The reason for this difference is that edgeR and DESeq2 are based on the negative binomial model, which contributes to large numbers of false positives. On the contrary, limma only uses the variance function and does not show excessive false positives, as is the case with a variance stabilizing transformation followed by linear model analysis with limma.

All three methods have their own advantages, and the choice is just dependent on the type of data. For example, if there is microarray data, limma should be given with priority, but when it is the next-generation sequencing data, DESeq2 and EdgeR are preferred.

There are also other packages and methodologies for differential expression analysis (for bulk) but they will not be seen here (in contrast with the methods that will be seen in single-cell DE, where there is no state of the art and it depends on the origin and characteristics of the data).

2.5 Additional methods

Knowledge of cell type composition in disease-relevant tissues is an important step towards the identification of cellular targets of disease. Cell types and population is straightforward with scRNA-seq, but these features could also be estimated with bulk RNA-seq with the use of **deconvolutions** methods. Deconvolution methods use bulk RNA-seq data to estimate the cell population in a tissue. The methods of deconvolution are, most of the time, paired with scRNA-seq to validate and integrate the results obtained with both of the two type of sequencing data. A method that use bulk RNA-seq data and scRNA-seq data will be seen in 4.

A methodology that was only stated in this chapter was **Polymorphism/mutation** detection. The workflow is almost the same as the one to get the quantification of cDNA, but after aligning the reads to the genome(especially to the genome since polymorphism

detection is done in reference to a genome and not to a transcriptome), there are some additional step of filtering for known sites and mutations and **variant calling** to see if the mutations were present in the sequenced tissue. There are some additional steps and methodologies used to variant calling and polymorphism detection but they will not be seen here. For more information see [166].

Another field that involves sequencing and differential analysis is **biomarker creation and identification**. The tools that do this kind of research will take an expression matrix where samples(patients in most cases) are associated with a vector (gene expression vector for every patient), a class (control and altered), some characteristics that can allow the grouping of the samples(like pathology or sample tissue type) and other features that can be used to compute/enrich the biomarkers. The output of these tools will be a series of features or a model that will identify the markers for every group. One of these tools is MIDCLass [46].

Nowadays (and in the future as well), other -omics technologies, such as proteomics and metabolomics, are now often incorporated into the everyday methodology of biological research. The "omics" at the end of a molecular term implies a comprehensive, or global, assessment of a set of molecules. The omics field has been driven largely by technological advances that have made possible cost-efficient, high-throughput analysis of biologic molecules. A list of some -omics is the following:

- **Genomics** is the most mature of the omics fields and is the focus of this research along with transcriptomics. In the realm of medical research, genomics focuses on identifying genetic variants associated with disease, response to treatment, or future patient prognosis.
- **Epigenomics** focuses on genome-wide characterization of reversible modifications of DNA or DNA-associated proteins, such as DNA methylation or histone acetylation. Covalent modifications of DNA and histones are major regulators of gene transcription and subsequently of cellular fate. Those modifications can be influenced both by genetic and environmental factors, can be long lasting, and are sometimes heritable. The importance of epigenomics in biological processes and disease development is evident from many epigenome-wide association studies.
- **Transcriptomics** examines RNA levels genome-wide, both qualitatively (which transcripts are present, identification of novel splice sites, RNA editing sites) and quantitatively (how much of each transcript is expressed as already seen before in this whole chapter). The central dogma of biology viewed RNA as a molecular intermediate between DNA and proteins, which are considered the primary functional read-out of DNA. Other examples of RNA function, such as structural (e.g., ribosomal complexes), or regulatory (e.g., Xist in ChrX inactivation) have often been regarded as odd exceptions to the general rule.
- **Proteomics** is used to quantify peptide abundance, modification, and interaction. The analysis and quantification of proteins has been revolutionized by MS-based methods and, recently, these have been adapted for high-throughput analyses of thousands of proteins in cells or body fluids.
- **Metabolomics** simultaneously quantifies multiple small molecule types, such as amino acids, fatty acids, carbohydrates, or other products of cellular metabolic functions. Metabolite levels and relative ratios reflect metabolic function, and

out of normal range perturbations are often indicative of disease. Quantitative measures of metabolite levels have made possible the discovery of novel genetic loci regulating small molecules, or their relative ratios, in plasma and other tissues.

- Microbiomics is a fast-growing field in which all the microorganisms of a given community are investigated together. Human skin, mucosal surfaces, and the gut are colonized by microorganisms, including bacteria, viruses, and fungi, collectively known as the microbiota (and their genes constituting the microbiome). The human microbiome is enormously complex; for example, the gut contains roughly 100 trillion bacteria from 1000 different species. There are substantial variations in microbiota composition between individuals resulting from seed during birth and development, diet and other environmental factors, drugs, and age.

Each type of omics data, on its own, typically provides a list of differences associated with the disease. These data can be useful both as markers of the disease process and to give insight as to which biological pathways or processes are different between the disease and control groups. However, analysis of only one data type is limited to correlations, mostly reflecting reactive processes rather than causative ones. Integration of different omics data types is often used to elucidate potential causative changes that lead to disease, or the treatment targets, that can be then tested in further molecular studies. More about the different types of -omics field, their characteristics and how to integrate them together can be found in [53].

An additional technique for multiple sequence alignment to take into consideration since it is one of my expertise is **motif finding** or **profile analysis**. Motif finding constructs global multiple sequence alignments that attempt to align short conserved sequence motifs among the sequences in the query set. This is usually done by first constructing a general global multiple sequence alignment, after which the highly conserved regions are isolated and used to construct a set of profile matrices. The profile matrix for each conserved region is arranged like a scoring matrix but its frequency counts for each amino acid or nucleotide at each position are derived from the conserved region's character distribution rather than from a more general empirical distribution. The profile matrices are then used to search other sequences for occurrences of the motif they characterize. In cases where the original data set contained a small number of sequences or only highly related sequences, pseudo-counts are added to normalize the character distributions represented in the motif. This type of sequence alignment is very similar to local sequence alignment but is also one of the most interesting methodologies since it implements concept from dynamic programming and statistical inference, aside from the curiosity and novelty of motif finding in MSA, these methods will not be used.

When doing differential analysis, the final results of the differentially expressed genes should also be validated by using new technologies and additional methods that use independent biological replicates and high-throughput quantitative reverse-transcription PCR (qPCR), also called real-time PCR [123]. Real-time PCR is a method that simultaneously amplifies and quantifies DNA. Also, NGS of the third generation could be used in conjunction to get real-time RNA quantities. Other methodologies that also take into account validation of DE analysis can be seen in [120].

2.6 Applications and impact

Genomics has revolutionized cancer research. Conventional classifications of disease, in terms of which organs and tissues it affects, are being divided into subtypes defined by the specific mutations that drive the disease. Developing a genetically targeted therapy is no easy task. It can be tricky to identify which genetic mutations are driving cancer and which are passengers(genes that are statistically linked, but that do not cause cancer and are probably only one of the consequences of cancer and other gene mutations). And although developers of targeted therapies focus mainly on mutations to a subset of genes called oncogenes, there is more to malignancy. Most genetic alterations in cancer are not oncogenes, they're tumour-suppressor gene alterations and these mutations inactivate genes that usually help to guard against cancer, such as those responsible for repairing DNA damage or controlling programmed cell death. And because the proteins encoded by these genes are often not produced in the cancer cells, they are difficult to target. For this reason, understanding transcriptomics for tissue is fundamental and always had an impact in medicine.

Bulk RNA-seq had already many consequences in medicine and biology and especially had an impact for cancer research (like breast tumours with mutations in genes HER2 or other genes related to regulation and cell cycle and function).

One thing to consider is how bulk RNA-seq can be used for **personalized** medicine, since the majority of the work is a long-term and experimentally heavy protocol that tries to understand the correlations of information of a group of patients(e.g. breast cancer patients), how these patients differ from some controls reference organisms that do not have the disease and see how some treatments will interact with the transcriptome(with the tumour cells and non-tumour cells) of the patients. The process is not quick and is personalized in reference of large enough groups of patients, so it is a consensus treatment more than a personalized methodology for specialized groups of patients.

This research is focused on the identification and acquisition of differential data from the sequencing database that will be used to generate some embeddings using graphs and the features inferred from the differential expression. Aside from this research, the applications and impact of bulk sequencing are very diverse and are at the center of transcriptome analysis. Applications such as isoform and gene fusion detection, digital gene expression profiling, targeted sequencing and many others are currently done where needed for research and direct or simulated treatment.

Pharmacogenomics has revealed compelling genetic signals associated with variability in drug response. Gene expression studies represent an additional approach to identifying candidate genes accounting for drug response variability. RNA-seq represents a useful tool to use in drug response studies and treatment in general. Personalized medicine, as the tailoring of clinical interventions, is mostly pharmacological, based on a person's ability to respond favorably; for pharmacological agents, this entails the metabolic capability to process them. For drugs with a narrow therapeutic range, such as blood-thinning agents, a small functional activity change can result in either a too low or a too high physiological effect that can lead to health complications. While pharmacogenetics for common drugs detects germline variants, cancer pharmacogenetics is

for selecting small molecule inhibitors and analyzing somatic variants from tumor cells. As cancer is predominantly a genetic disease, tumor DNA analysis is routinely deployed for molecular characterization of the cancer cells, as well as treatment prognostics and monitoring. In recent years, liquid biopsy has been successfully applied to obtain tumor circulating free DNA. It is now possible to use liquid biopsy for early cancer detection, prognostics, and treatment selection and monitoring. Unfortunately, the cost of cancer genetic tests and targeted treatments are still very high, making them inaccessible in less developed countries.

Transcriptomics provides researchers with a more dynamic view of a cell, but RNA is mainly an intermediary for biology's most fundamental players, the proteins. Proteomics was measured and it is measured today with indirect means of counting the RNA-seq expression profiles for genes and translating the genes into protein expression.

High costs and limitations in terms of technologies have remained the main barriers for the greater omics-based implementation of personalized medicine. AI-driven machines, are being deployed to cut costs, especially in overcoming the enormous volume of collected patient data. AI is most commonly deployed at two levels within clinical bioinformatics: in silico gene damage scoring and prioritization and phenotype scoring, where various text mining algorithms are adopted. Other than that, AI is still a research tool until large longitudinal data, and more robust informatics frameworks are available. It is worth mentioning that one of the main strengths of AI in clinical practice is the area of image recognition(segmentation and identification of regions), even though it is another source of data, it can be used in combination with transcriptomics and other sources that use the bulk RNA-seq workflow.

2.7 Considerations

The methods presented in this chapter for bulk sequencing (and bulk RNA-seq especially) are tested in the majority of the work that is done today on personalized medicine. These methods are proven in all kinds of environments and have been honed for a variety of applications in medicine and biology. The impact of bulk RNA-seq for these applications has been incredible and the practices for bulk sequencing and what to do with the data that is obtained were standardized and proven with concrete examples during all these years. The methods and platform used to sequence a sample are available to the public and **highly standardized** and provide accurate results (for large tissue samples and high amount of genetic material in the sample) while maintaining a low cost for the protocol and the experiment itself. Deep sequencing has been revolutionizing biology and medicine in recent years, providing single base-level precision for the understanding of nucleic acid sequences in high throughput fashion. Sequencing of RNA is now a common method to analyze gene expression and uncover novel RNA species. Aspects of RNA biogenesis and metabolism can be interrogated with specialized methods for cDNA library preparation, so the field is always thriving and growing even after single-cell analysis was born.

The problem with bulk sequencing lies in the granularity of the protocols and methods that were seen here. Bulk sequencing requires a tissue sample and it will find the

population of fragments of DNA in that sample (RNA via cDNA for RNA-seq). This populations of DNA/RNA will come from the **whole population of cells and the environment** and the complex system of cells and other products will be treated as a **single unit** that will be sequenced and analyzed, while the reality of the interactions in the tissue is really diverse and complex (patient-related, environment-related, treatment-related and so on).

The gene expression resulting from the process of RNA-seq will be approximately the **average amount of gene expression for the population of cells** so it will have low details and it will not make any differentiation with tissue composition (for the standard methods used for everyday research at least, so not methods that use a different kind of data and features, also called "multi-omics data", with a holistic approach, like the ones introduced in 2.5). Methods that use multi-omics data will be seen in 4 and also in 3.5, where bulk and single-cell workflows and protocols become one in ensemble models that take different data and combines them to reach a better and more accurate outcome.

As already said before, one of the limitations of bulk RNA-seq is that when the quantities of RNA are very low (like in single cells), there are a lot of problems with the whole workflow since amplification could lead to bias in the fragments and the final results will very much likely have a lot of **dropouts genes**. A review of applications and options to take when having low quantities of RNA can be seen in [111]. This problem with bulk sequencing will be directly inherited by single-cell since most of the technologies used today to do single-cell sequencing (the true sequencing part of the protocols) is the same as the ones used for bulk and have the same problems as well.

Bulk deconvolution should be seen another instruments that could be used in collaboration with single-cell true populations to see how the populations are different from each other and understand if some population in single-cell where mislabelled(in the cluster) or where mistyped. While bulk deconvolution is a multimodal spectrum decomposition (seeing which genes are majorly expressed in the data and deconvolute the sequencing data itself), single-cell isolate the cells and sequences them as it will be seen in the next chapter.

Everything presented here was in sight for introducing the subject and topics treated in the next chapter about **single-cell sequencing**.

Chapter 3

Single cell sequencing

Single-cell sequencing is the cutting edge of high-resolution cell phenotyping and characterization for complex and heterogeneous samples. Despite the fact that this methodology requires a lot of expertise and a considerable amount of physical tools (many reagents, machines for the initial stages, a cost not to be underestimated ranging from $ 0.30 to $ 20 per sequenced cell for scRNA-seq), it is practically applied extensively in research in recent years and is becoming a necessary tool to have higher accuracy for sequencing experiments. Furthermore, there is still no standardization of some methods and the instrumentation used defines the quality and quantity of information that are obtained at the end of the process. For this reason, single-cell methodologies are mainly guidelines to be followed during the study and analysis of the data obtained from them, referring to the specific documentation of platforms and tools used as the main source of procedures and workflows.

3.1 Single cell RNA-seq

As already seen and discussed during the bulk sequencing analysis, one of the most important applications of single-cell sequencing is scRNA-seq and in particular mRNA sequencing. It is really important to characterize the cell from the mRNA that it produces since the activity of the cell itself is predominantly composed of the production of transcript mRNA and its consequent translation in proteins.

The single-cell RNA sequencing protocols characterize the single cell according to the mRNA sequencing data of the isolated and lysed cell. With the methods that will be presented in this presentation it is not possible to sequence other types of non-coding RNA (although there are methods to do so) The sequencing process starts from an input tissue sample and results in an expression matrix where each row represents the gene that has been sequenced and each column a cell that has been isolated.

3.2 Objectives and techniques

Unique phenotypic alterations in specific cell types, visualized as varying RNA expression levels (both coding and non-coding), have been identified as crucial factors in the pathology underlying conditions. Recent advances in single-cell RNA sequencing (scRNA-seq) have elucidated a new realm of cell sub-populations and transcriptional variations that are associated with normal and pathological physiology in a wide variety

of diseases along widening the field of genes expression analysis with specific results for the components of a sample. This breakthrough in the phenotypical understanding of our cells has brought novel insight into many different fields of medicine and biology.

scRNA-seq allows for the separation of widely distinct cell subpopulations which were simply averaged together(along with other genetic and transcriptional material) with bulk-tissue RNA-seq until the advent of this technology and methodology. scRNA-seq has been used to identify novel cell types in a variety of tissues and environments that could be implicated in a variety of disease pathologies and other biological activities. Furthermore, scRNA-seq has been able to identify significant heterogeneity of phenotypes within individual cell subtype populations. The ability to characterize single cells based on transcriptional phenotypes allows researchers the ability to map the development of cells and identify changes in specific sub-populations due to diseases at very high throughput.

scRNA-seq has given the field of biology and personalized medicine the ability to profile the phenotypes of single cells leading to discoveries of new cellular sub-populations that could contribute to many different pathologies in different tissues. However, because of the complexity of the data, scRNA-seq findings may be difficult to interpret by clinicians or fit into current knowledge bases.

While there are numerous methods for performing scRNA-seq, all of them follow a general workflow composed of common and optional steps needed that will be seen in details during this chapter:

1. **Sample of the tissue**: the targeted tissue or micro-environment is sampled/dissected via the same procedures introduced in 2.2.
2. **Dissociating cells in the tissue**: this is done via enzymes or physically induced (mechanically induced). Widely used dissociating enzymes in scRNA-seq(especially in Drosophila but the reagents should be similar for other research) include trypsin, collagenase, papain, liberase, and elastase, and in most cases, these enzymes are used in combination to improve the dissociation efficiency. When choosing a dissociation method, it is best to test multiple methods as their efficiency can vary significantly, depending on the cell type, tissue type, and developmental stage. Another important factor that needs to be considered is cell viability. Since tissue dissociation is a harsh process, it can cause cellular stress and transcriptional changes. Thus, if two methods can both adequately dissociate the desired tissue, the less damaging one with higher cell viability should be used. This step is necessary to obtain a suspension of dissociated cells which is mandatory for the next steps. Suspensions should be filtered with appropriately sized cell strainers (pore size larger than cell diameter) to remove clumps and debris. Treatment of cell dissociation varies on the type of cells treated, the objective of the project and the fragility of the cells that needs to be isolated and studied. The use of some methods can induce stress-related variation in gene expression. If cell dissociation is impossible or very hard to do (e.g. neural tissue that is extremely interconnected), there are some methods that do not isolate the cells but the nuclei, these methods fall under the name of snRNA-seq. However, nuclei have lower amounts of mRNA compared to cells and are more challenging to enrich or deplete for specific cell types of interest. The mRNA of nuclei has

also higher number of pre-mRNA that was not spliced (introns are present), and have no altered nucleotide in 5' end (capping) or the polyadenylated tail in the 3' end.

3. **Selection of the cells (OPTIONAl)**: especially removing dead cells via different technologies like flow cytometry, FACS sorting and filtering and the use of membrane markers or phosphors(colorants) like fluorescent transgenes or staining dyes.
4. **isolating single cells**: this is done in micro-wells or droplets nowadays, while at the start of the scRNA-seq research it was done via pipetting the single cells one by one(as can be seen in figure 3.2, where the advancement in scRNA-seq technology over the years is presented). Micro-wells methods are characterized by the use of sorting techniques that can isolate the cells and (if possible) associate them with some meaningful information like size, count, functions, biomarkers, etc.
5. **extraction and capture of RNA**: uses the same techniques that were seen previously in 2.2, although it is more focused on the capturing of mRNA that have the poli(A) nucleotide sequence, even though other methods exists to isolate other types of RNA (like lncRNA) [[135],[139]]. In principle, scRNA-seq applications are not restricted to specific species as long as poly(A)-tailed RNA is present. However, some organisms might require additional processing steps to efficiently release molecules into the reactions (e.g., cell wall removal for plant material).
6. **reverse transcribing RNA to cDNA**: this step is the same as the one described in 2
7. **amplifying the cDNA**: this step uses also the same techniques seen in 2 for cDNA amplification, but it has a lot more importance since **amplification bias** is greater and the quantity of RNA is very low for single cells generally. To resolve some of the problems with amplification bias **UMI** will be used especially for the **10x genomics** platform and protocol [4].
8. **library preparation**: as the whole process to obtain the sequencing library, for single-cells techniques such as pooling, multiplexing and associating unique molecular identifiers to identify the fragments become really important for the whole protocols used. The result of the library preparation will be the fragments elongated with adapters, fragmented (with loss of the fragmented region or the full region), and identified via molecular barcodes(Illumina indexes, UMI, etc.).
9. **DNA sequencing**: This step is the same as the one for bulk sequencing seen in 2. Additional attention will be given to demultiplexing and UMI detection since the scRNA-seq protocols use them very frequently. Usually, paired-end is used
10. **Data analysis**: Is done on the sequenced reads and is partially really similar to the analysis done in bulk sequencing(at least for the first and the last parts of the workflow).

It is within each of these steps that the variety of scRNA-seq protocols make use of different available technologies and methodologies, which are specifically advantageous to particular experimental designs and goals and are at the core of the whole final results and how to handle the data to account for variations, noise, accuracy and integration with different data. For instance, at the single cell isolation step, a manual isolation

technique, like micromanipulation(FACS or related technologies to order/select the cells and isolate them one by one in an orderly fashion), has advantages for samples of few precious cells, while the more high throughput and cost-effective microfluidics(bead-based) isolation techniques are advantageous for experiments with a large number of cells and where the resulting number of cells sequenced need to be really high as well since micromanipulation is done on a very low number of cells compared to microfluidics methods and platforms. Another experimental detail to consider when selecting a scRNA-seq protocol is the desired transcript coverage. Protocols that sequence full-length transcripts, like Smart-seq2[119], are more advantageous if the goal of the experiment is to analyze isoform usage or allelic expression. Whereas, tag-based sequencing, as used in protocols like the 10x Genomics Chromium droplet-based method, allows for higher throughput and cost-effectiveness, which makes it more advantageous if the experimental goal is cell subpopulation detection. This is because a a plethora of cells that could be sequenced with this method increases resolution for improved cell subpopulation detection thereby making detection of rare cell populations more likely. Additional factors to consider when selecting a scRNA-seq protocol include the type of RNA of interest in the experiment (polyA + or -), whether spike-ins or Unique molecular identifier (UMI) can be used, and cost. New methodologies that try to integrate both full-length transcripts and tags(UMI specifically for their properties and the problem addressed) exists and the field is thriving and expanding even at this moment, for example Smart-seq3 [50].

It is very important to understand that the precautions to take for bulk RNA-seq (about contamination, maintaining a good environment for the workplace and keeping the same protocols over all the experiments) are exponentially more important in scRNA-seq since variation in the resulting data will be immensely more impactful compared to bulk RNA-seq.

The **preparation of high-quality single-cell suspensions** is key to successful single-cell studies. Depending of the starting material, the condition of the cells is critical for efficient cell capture and optimal performance of the scRNA-seq protocols. Although most methods use fresh viable single cells, alternatives include preserved samples and nuclear RNA from frozen tissue.

Despite the varying advantages of different scRNA-seq protocols, there are shared challenges. For instance, accounting for library preparation costs and confounding batch effects remain significant barriers for all scRNA-seq protocols. One current improvement to scRNA-seq protocols which addresses these limitations is multiplexing (methods that label cells at the sample level, allowing for pooling and processing of all the cells in one run). Recently developed techniques include:

- **lipid oligonucleotides** which contain sample-level barcodes that incorporate into the plasma membrane of live cells,
- **combinatorial barcoding**, where cells are identified through the unique combination of barcodes acquired from multiple rounds of random barcoding
- using **naturally variant single nucleotide polymorphism genotypes** to distinguish cells' sample of origin.

Most centrally, these many scRNA-seq methods all function in accomplishing the same aim: to profile which genes are expressed and their level of expression for individ-

ual cells, which allows for novel data analyses important for investigating fundamental biological questions.

Regardless of the exact protocol followed, all scRNA-seq data present unique interpretational challenges because of the high levels of technical and biological noise. RNA capture efficiency, batch effects, transcriptional kinetics, cell cycle stage, and, most significantly, a large amount of amplification owing to the small amount of starting material are a few such sources of noise. Therefore, to ensure that the signal of interest in the data is not masked by unwanted variation, experimental and computational normalization methods need to be performed to adjust for these sources of noise in the sequencing data. Typically, multiple methods for the removal of noise are used in conjunction because each method accounts for a specific source of bias/noise. For instance, the experimental integration of UMI sequences allows for the **detection** and **removal of polymorphism/mutations introduced during the PCR** step and **amplification duplicates**, thereby adjusting read counts across samples and therefore avoiding the effects of amplification bias and uneven counts of the true transcriptome of the cell.

Common steps required for the generation of scRNA-seq libraries include cell lysis, reverse transcription into first-strand cDNA, second-strand synthesis, and cDNA amplification. In general, cells are lysed in a hypotonic buffer, and poly(A)+ selection is performed using poly(dT) primers to capture messenger RNAs (mRNAs) (Fig. 1g). It has been well established that due to Poisson sampling, only 10–20% of transcripts will be reverse transcribed at this stage. This low mRNA capture efficiency is an important challenge that remains in existing scRNA-seq protocols and necessitates a highly efficient cell lysing strategy.

Another important source of noise is variation in sequencing depth between cells. Initially, sequencing depth normalization was accomplished through bulk-RNA-seq established methods, like Reads Per Kilobase Million (RPKM), Trimmed Mean of M-values (TMM), and quantile normalization as seen in 2.3.1, but these techniques were not that efficient for single-cell since they were primarily created for bulk RNA-seq and a large amount of transcript. Recently, scRNA-seq data-specific methods have been developed, like Single-Cell Differential Expression (SCDE)[76] or Model-based Analysis of Single-cell Transcriptomics (MAST) [104], which take into account attributes unique to single-cell expression data, like the high rate of dropout events (genes erroneously reported to have zero expression because of missed RNA capture). These normalization methods are an important preprocessing step to improve the quality of downstream analyses. Both topics of barcoding/UMI association and normalization can be visualized in figure 3.1.

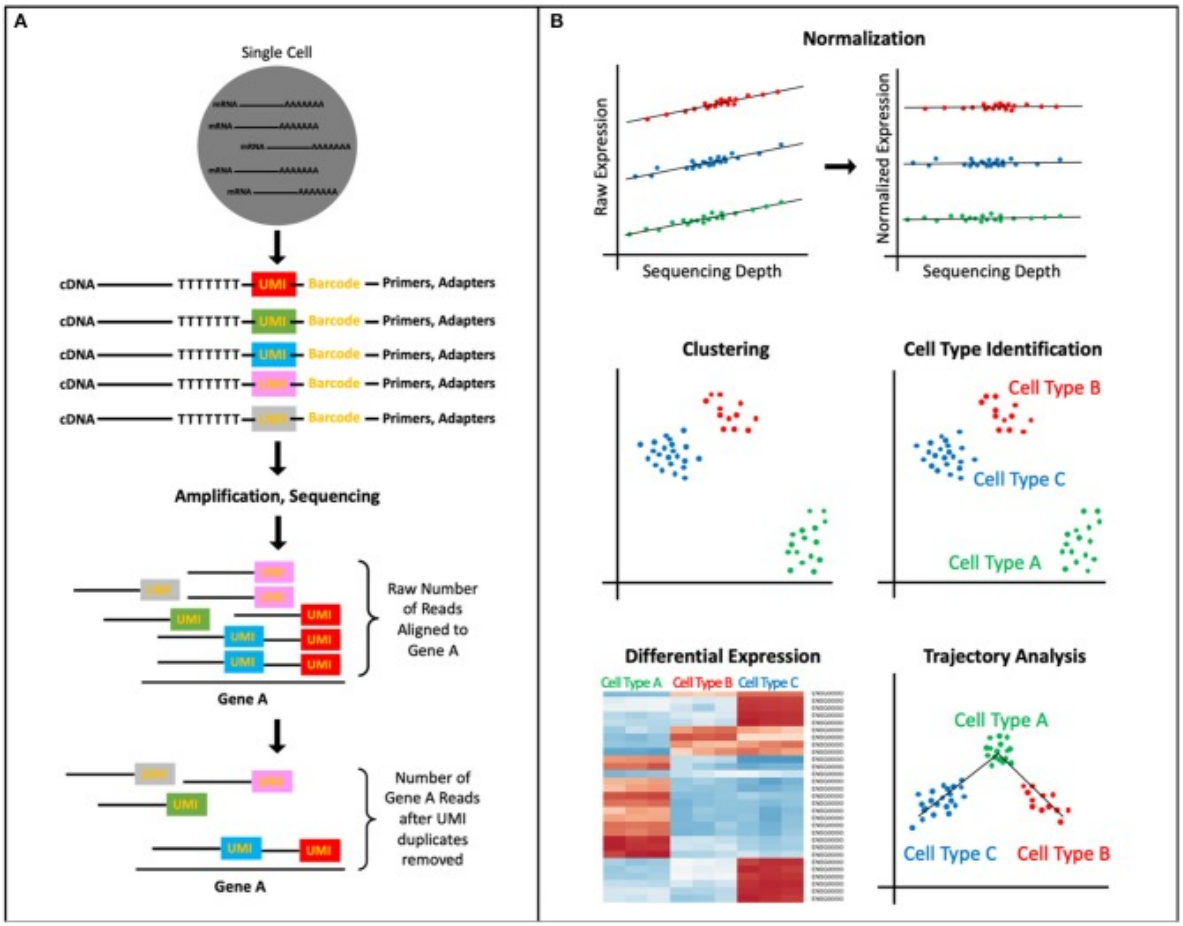


Figure 3.1: **scRNA-seq Data Processing and Analysis**. (A) UMIs, short DNA sequences tagged to cDNA fragments before amplification, identify unique reads that will avoid PCR duplicates and exponential bias, thereby normalizing the transcript counts. (B) A common analysis pipeline for scRNA-seq data includes: normalizing data to account for sources of technical and biological noise (sequencing depth), clustering cells to identify novel and known cell types as well as sub-populations, ordering cell types and states into trajectories, and performing differential expression analysis, which can allow for identification of biomarkers that can bea used to infer or generate additional information about the cells groups and the system as a whole, assigning function to cell cluster. More about the next steps of the pipeline will be seen in 6

For cDNA preparation, an engineered version of the **Moloney murine leukemia virus reverse transcriptase** with low RNase H activity and increased thermostability is typically used in first-strand synthesis. Second strands can be generated using either poly(A) tailing or by a template-switching mechanism (TSO or another mechanism of template-switching). This latter approach ensures uniform coverage without loss of strand-specificity compared to the former. The small number of synthesized cDNAs are then further amplified using conventional PCR or in vitro transcription. The in vitro transcription method can amplify templates linearly but is time-consuming, as it requires an additional reverse transcription, which may lead to 3' coverage biases. **Smart-seq2** (improved version of Smart-seq)[119] generates full-length transcripts and is thus suitable for the discovery of alternative-splicing events and allele-specific expression using single-nucleotide polymorphisms. Currently, the Illumina platforms are

widely used for the sequencing step.

The data analysis workflow for scRNA-seq data, as implemented by software packages like Cell Ranger[1], STARsolo[74], Seurat[131], SingleR[137], and Monocle3[27], includes:

- Demultiplexing of barcodes, resolution of UMI and alignment to reference, along the creation of expression profiles
- Quality control and filtering for cells and genes
- Elucidating cells' heterogeneity via clustering cells based on gene expression profiles
- Characterizing cell clusters by assigning cell type or functionality via biomarkers and/or differential expression analysis
- Organizing defined cell types/states into a trajectory and create a pseudo-time for the cells sequenced

The first step to identifying underlying patterns among the transcriptomes of the single cells is to perform dimensionality reduction. Dimensionality reduction tools, like PCA, tSNE, and more recently (and importantly), UMAP, project the high dimensional scRNA-seq data (the expression levels of thousands of genes per cell, in thousands of cells) into lower dimensional space, thereby collapsing the data and effectively identifying and preserving only the features that contributed to the structure of the original high-dimensional data. The cells can then be separated into populations based on the similarity of their gene expression profiles through clustering algorithms that employ one of four main methods: k means, hierarchical clustering, density-based clustering, or graph-based clustering. The identified clusters can then be visualized via a scatterplot that translates individual cells into data points, where cluster membership is indicated by the physical proximity of the points on the plot. The dimensionality reduction is done only for the clustering part since the genes need to be expressed to get the differential expression. Also, UMAP is used also for visualization in 2D-3D, but the clustering and annotation of cell groups should be done in higher dimensions.

The processed scRNA-seq data is then suitable for use in analysis applications, such as characterizing the cell clusters, and trajectory inference. The most straightforward method for characterizing cell clusters is to identify cluster-specific expression of cell type biomarker genes within and between the cell groups by differential expression coupled with additional techniques of statistical analysis of the genes that are significantly down-regulated or up-regulated between groups and have low variance within the group of cells itself. Other methods to get the biomarkers use **MAST**, as previously stated, or other statistical analysis techniques that use the variance of the groups or other measures and hypothesis testing. In the case where cells cannot be identified via biomarkers, differential expression analysis, of which the different methods are highly abundant and well-established, presents an alternative method for cell cluster characterization. Differential expression analysis identifies sets of genes significantly more highly expressed within clusters compared to all other cells, which can provide clues as to the identity of the cell type/state or the functionality of the cells within the cluster. Another analysis of scRNA-seq data is trajectory inference "**pseudotime**" analysis. This analysis works by ordering cells along a trajectory based on the similarity of their gene expression profiles and establishes a differentiation timeline of the cells in the sample(could also

be seen as an evolution such as in tumors where cells that are precursors of new tumor cells have fewer mutations overall). Numerous methods have been developed to perform pseudotime/trajectory analysis, but the most important factor when selecting a method is the type of biological trajectory that is expected (for instance, a linear, bifurcated, or multifurcated cell type differentiation trajectory). One of the most important tool to do trajectory analysis is **monocle** [27]. Despite the trajectory topology of the method, pseudotime analyses face some limitations that are aimed to be addressed by future method developments, these include:

- accounting for other processes (like cell cycle stage) that may mask the gene expression patterns of the process of interest
- incorporating other types of information (such as location, chromatin state, and post-translational modifications) that contribute to a cell's state in addition to transcriptome

While scRNA-seq data require careful processing to achieve interpretability, the end result of elucidating the cell types/states present in a sample and their relationships to each other (i.e., which genes are differentially expressed between them, and in what order they occur in a dynamic process) has critical implications for deepening the understanding of disease pathologies.

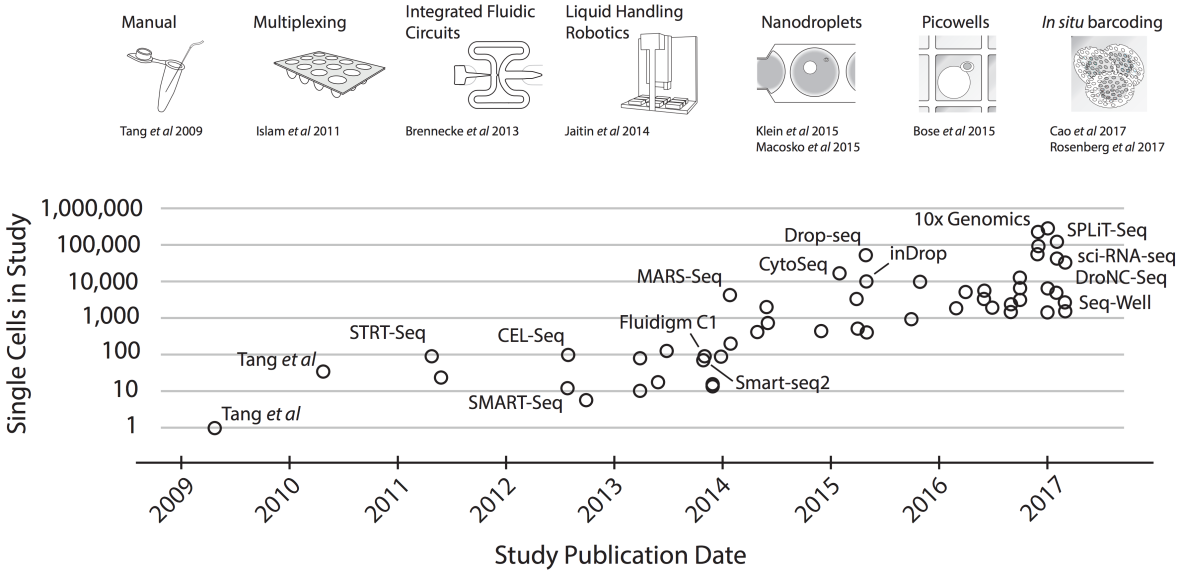


Figure 3.2: Evolution of scRNA-seq protocols. Source [148]

There are currently a wide diversity of protocols for preparing scRNA-seq data, each with its own strengths and weaknesses, the most known and documented protocols are seen in figure 3.3.

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay	10^5 10^4 10^3 10^2										


Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

Figure 3.3: scRNA-seq protocols. Source [25]

These methods can be categorized in different ways, but the two most important aspects are **cell capture or isolation** and **transcript quantification**. These two steps are very important throughout the process as they classify the sequencing method and how the data is handled.

There are other phases that make up the process but the two listed above are the most important since they establish the characteristics and final precautions to be taken into consideration at the end of the sequencing process when you have the data obtained from the process (transcriptome for a single cell, cell-related characteristics such as cell size and phase, etc.).

3.2.1 Cell isolation

Single-cell isolation is the first step in obtaining transcriptome information from an individual cell. Cell isolation/capture methods are really important for the process of scRNA-seq since they will establish the quality of the experiment and results overall. Cell isolation methods contribute to:

- throughput of the experiment as the number of isolated cells;
- the quality of the cells isolated (no dead cells, status of the cells retained from the sample when they are isolated, little to no alteration in gene expression);
- how cells are selected before being sequenced;
- what other information can be obtained in addition to the sequencing information

The field of cell isolation for single-cell sequencing is a vast and always growing field with a lot of methods applied to reach some compromise on the quality of the results and the additional information obtained, as stated previously. Some of the most important isolation methods can be seen in figure 3.4. Limiting dilution (Fig. 3.4a) is a commonly used technique in which pipettes are used to isolate individual cells by dilution. Typically, one can achieve only about one-third of the prepared wells in a well plate when diluting to a concentration of 0.5 cells per aliquot. Due to this statistical distribution of cells, this method is not very efficient. Micromanipulation (Fig. 3.4b) is the classical method used to retrieve cells from early embryos or uncultivated microorganisms, and microscope-guided capillary pipettes have been utilized to extract single cells from a suspension. However, these methods are time-consuming and have low throughput. More recently, flow-activated cell sorting (FACS, Fig. 3.4c) has become the most commonly used strategy for isolating highly purified single cells. FACS is also the preferred method when the target cell expresses a very low level of the marker. In this method, cells are first tagged with a fluorescent monoclonal antibody, which recognizes specific surface markers and enables the sorting of distinct populations. Alternatively, negative selection is possible for unstained populations. In this case, based on predetermined fluorescent parameters, a charge is applied to a cell of interest using an electrostatic deflection system, and cells are isolated magnetically. The potential limitations of these techniques include the requirement for large starting volumes (difficulty in isolating cells from low-input numbers ¡10,000) and the need for monoclonal antibodies to target proteins of interest. Laser capture microdissection (Fig. 3.4d) utilizes a laser system aided by a computer system to isolate cells from solid samples. Microfluidic technology (Fig. 3.4e) for single-cell isolation has gained popularity due to its low sample consumption and low analysis cost together with the fact that it enables precise fluid control. Importantly, the nanoliter-sized volumes required for this technique substantially reduce the risk of external contamination. Microfluidics was initially utilized in a small number of biochemical assays for the analysis of DNA and proteins. However, complex arrays have now been developed that permit individual control of valves and switches, thus increasing their scalability. Microfluidics allows higher-throughput scRNA-seq workflows, thus eliminating the technical constraints on scalability associated with microtiter plates. Moreover, reducing reaction volumes from microliters to nanoliters reduces costs and technical variability while improving cDNA yield. There are three strategies for capturing cells: **IFCs, droplets** and **nanowells**, all of which increase the number of capture sites relative to that achieved with microtiter plates. The first microfluidics system used for scRNA-seq was designed as an automated array solution (Fluidigm C1) in which single cells enter a fluidics circuit and then are immobilized in hydro-dynamic traps, lysed, and processed in consecutive nanoliter reaction chambers via a modified Smart-seq2 protocol (that will be seen in the next section).

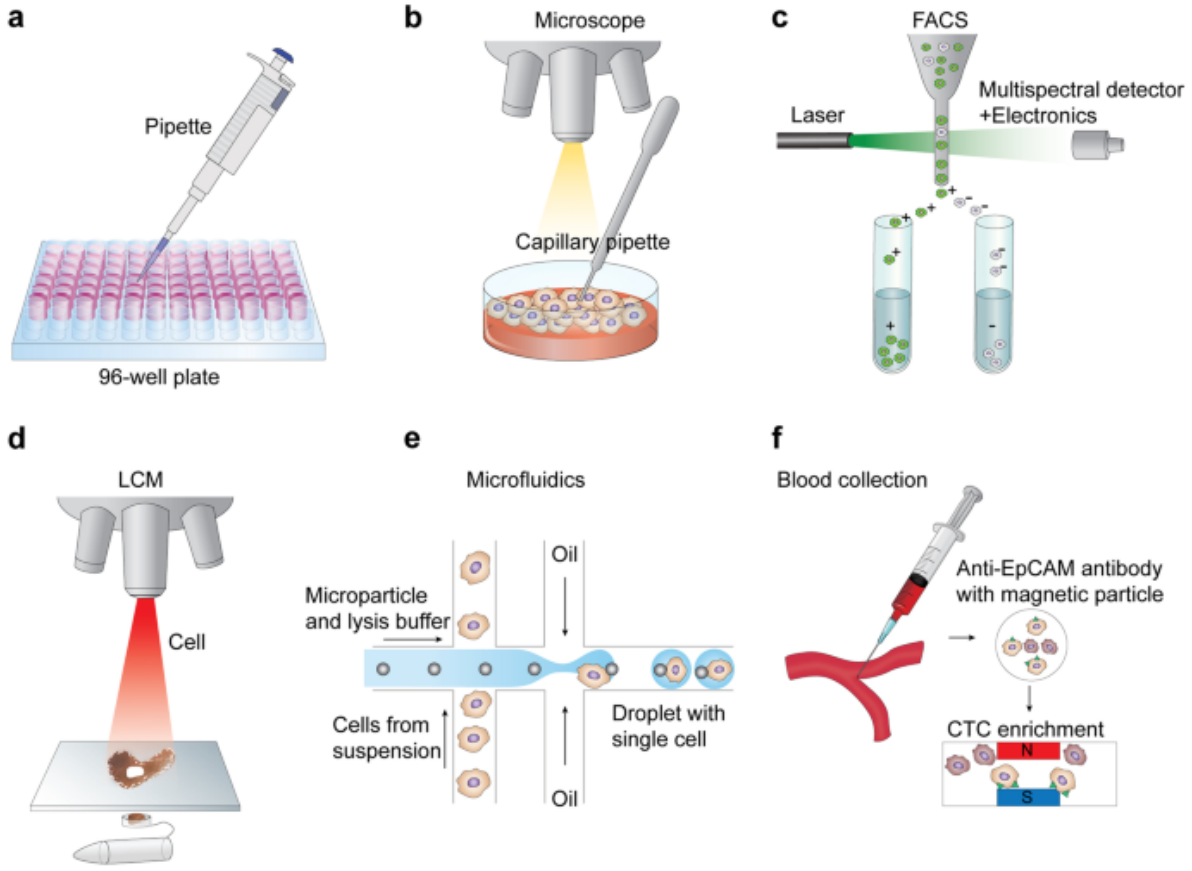


Figure 3.4: **Different techniques of cell isolation**: **a)** The limiting dilution method isolates individual cells. **b)** Micro-manipulation involves collecting single cells using microscope-guided capillary pipettes. **c)** FACS isolates highly purified single cells by tagging cells with fluorescent marker proteins. **d)** Laser capture microdissection (LCM) utilizes a laser system aided by a computer system to isolate cells from solid samples. **e)** Microfluidic technology for single-cell isolation requires nanoliter-sized volumes. An example of in-house microdroplet-based microfluidics (e.g., Drop-Seq). **f)** The CellSearch system enumerates CTCs from patient blood samples by using a magnet conjugated with CTC binding antibodies.

Another isolation technique not seen in figure 3.4 is **MACS**, a technique similar to FACS that uses magnetic beads attached to the cells for sorting.

This research will focus on methodologies that use FACS or microfluidics (10x genomics chromium platform [3] in particular and Fluidigm C1 platform[39]). The two (three since the two microfluidics platform use two different technologies, that is **Microfluidic-array** for Fluidigm C1 and **Microfluidic-droplets** for 10x chromium platform)isolation technologies will be seen in detail since they are really important for most research that is done today on scRNA-seq.

The strategy used for capturing cells determines the throughput of the experiment (i.e. how many cells are isolated), how the cells are selected prior to sequencing, as well as what kind of additional information besides transcript sequencing can be obtained. The three most widely used options are microtitre-plate-based(that use FACS to sort the cells in the plate's wells), microfluidic-array-based(Fluidigm C1) and microfluidic-droplet-based(10X chromium) methods.

- **Microtitre-plate** methods rely on isolating cells into individual wells of the plate using, for example, pipetting, microdissection or fluorescent activated cell sorting (FACS). One advantage of well-based methods is that one can take pictures of the cells before library preparation, providing an additional data modality. For example, one can identify and discard damaged cells or find wells containing doublets (wells with two or more cells). When using automatic FACS sorting, it is also possible to associate information such as cell size and the intensity of any used labels with the good coordinates, and therefore with individual cell indices in downstream analysis. The main drawback of these methods is that they are often low-throughput and the amount of work required per cell may be considerable.
- **Microfluidic-array** platforms, provide a more integrated system for capturing cells and for carrying out the reactions necessary for the library preparations. Thus, they provide a higher throughput than microtitre-plate-based methods. Typically, only around 10% of cells are captured in a microfluidic platform and thus they are not appropriate if one is dealing with rare cell types or very small amounts of input. Care also has to be taken with the cell sizes captured by the arrays, as the nanowells are customized for particular sizes (this may therefore affect the unbiased sampling of cells in complex tissues). Moreover, the chip is relatively expensive, but since reactions can be carried out in a smaller volume, money can be saved on reagents.
- **Microfluidic-droplet** methods offer the highest throughput and are the most popular method used nowadays. They work by encapsulating individual cells inside a nanoliter-sized oil droplet, together with a bead. The bead is loaded with enzymes and other components required to construct the library. In particular, each bead contains a unique barcode that is attached to all of the sequencings reads originating from that cell. Thus, all of the droplets can be pooled, and sequenced together and the reads can subsequently be assigned to the cell of origin based on those barcodes. Droplet platforms have relatively cheap library preparation costs on the order of 0.05 USD/cell. Instead, sequencing costs often become the limiting factor and a typical experiment the coverage is low with only a few thousand different transcripts detected.

To obtain an unbiased view of the cellular composition of a sample, one must capture all cells during the isolation process. Here attention must be paid to very small or large cells that may be excluded during FACS isolation or captured in microfluidic systems, respectively. However, for many experiments, it may be necessary to enrich for or exclude some cell types to increase the total number of cells of interest in the final scRNA-seq libraries. Target populations can be selected by FACS and MACS with appropriate labeling (e.g., antibodies or transgenic systems). Microtiter plates and some nanowell capture systems allow index sorting, in which fluorescence intensity or cell size (FACS information) is associated with capture coordinates and subsequently with single-cell indices. The FACS device records the sorting position and intensity values of a given cell, thereby enabling the subsequent integration of transcriptome profiles with the recorded cell properties.

To define adequate cell numbers per experiment, the experiment's methodology must consider sample heterogeneity(very important for scRNA-seq) and subpopulation

frequency (the estimated abundance of the cell type of interest). In particular, larger cell numbers are required to resolve the structure of heterogeneous samples with many expected sub-populations. Also, the total number of cells required increases when rare cell types need to be identified. One can calculate the required cell numbers by estimating both subpopulation structure and low-frequency cell-type abundance and defining the desired cell number per group, a computational tool is available and accessible at https://satijalab.org/howmanycells. Because most experiments target poorly described systems, heterogeneity can only be estimated, so pilot experiments are recommended before large-scale data production. Similarly, seemingly homogeneous samples can be initially profiled using higher cell numbers and sequencing depth to reveal yet uncharted sample complexity. Note that higher cell numbers can also be beneficial for homogeneous samples, as this increases statistical power during analysis. A common strategy for determining heterogeneity in a sample is to analyze highly variable genes across datasets and identify how many markers can be observed in the data. A thorough feature-selection step to remove uninformative or noisy genes increases the signal-to-noise ratio but also reduces the computational complexity. Commonly used strategies for extracting variable genes in scRNA-seq tools exploit the relationship between the mean transcript abundance and a measure of dispersion such as the coefficient of variation, the dispersion parameter of the negative binomial distribution or the proportion of total variability.

Another thing to take into account on the methodology is **sample preservation**. All common scRNA-seq methods were initially designed to use freshly isolated cells. However, in research and clinical practice, immediate sample processing can be challenging because of a lack of the required infrastructure or specialized equipment, such as FACS devices.

Moreover, although samples may be collected at multiple time points, **simultaneous sample processing** may be preferred to **avoid technical batch effects**. Sample preservation is a viable solution because it disconnects the location and time of sampling from the downstream processing steps. In this context, **cryopreservation** has been established for single-cell transcriptome analysis. Similarly, methanol fixation has been established as an alternative for droplet-based single-cell methods, and could also be used to avoid technically induced variations in gene expression triggered by prolonged sample processing time. both methods allow the archiving and transport of samples and broaden the range of applications of scRNA-seq methods, for example, to the clinical context. However, both approaches have shown a potential bias in cell-type composition, and it is strongly recommended to thoroughly evaluate preservation methods for new cell types that have not been tested.

A consideration about FACS is that Fluorescence Activated Cell Sorting can be **used upstream of any of the capture methods**, to select a subpopulation of cells. A common way in which this is used is to stain the cells with a dye that distinguishes between live and dead cells (e.g. due to membrane rupture), thus enriching the cell suspension with viable cells and removing the dead cells from the sample that will be sequenced.

3.2.2 Transcript quantification

Transcriptome profiling of individual cells can be split into four major components: RNA molecule capture, RT and transcriptome amplification, sequencing library preparation, and sequencing. Various scRNA-seq methods exist, but they all apply the same underlying principles. Transcript quantification is the final objective of the process and it encompasses in itself the four major components described. Many of the methods that will be described in this section are confirmed in their generally high accuracy, although efficiency, scalability and costs vary considerably among them. This should be taken into account during the selection of methods for a given experiment.

The first steps to start with the true sequencing part of the scRNA-seq methodology are **RNA molecule capture, reverse transcription and transcriptome amplification for sequencing library preparation.** Most scRNA-seq methods, including those during this research, capture poly(A)-tailed RNA, although specific protocols are available for profiling total RNA[54], miRNA [37] and lncRNA [42] as already introduced previously. After cell lysis, poly(A)-tailed RNA is captured by poly(T) oligonucleotides, which exclude abundant RNA types such as rRNA and tRNA. After capture, the RNA is reverse-transcribed into stable cDNA, at which point most methods add single-cell-specific barcodes within the poly(T) oligonucleotides that allow cost-effective multiplexed processing of pooled samples. Moreover, random nucleotide-sequence stretches in the poly(T) oligonucleotide serve as unique molecule identifiers (UMIs) that allow the user to correct amplification biases and reduce technical noise. RT is a crucial step, and different protocols have been optimized in various ways with efficient enzymes and specific additives that maximize efficiency. Even though the protocols for RT are very similar to the one described in 2, the protocols for RT in single-cell are tested for low quantities of genetic material and are directly embedded in the protocols that are used within the experiment. That means that, for protocols that use microbeads, the RT and amplification steps are done on-bead while for protocols that do not use beads (aside from tagmentation protocols), these steps are done in solution.

cDNA can then be amplified by PCR or through in vitro transcription (IVT). For this, adaptor sequences or RNA polymerase promoter sequences are introduced during RT or second-strand synthesis. Although IVT is less prone to biases through linear amplification of molecules, it requires additional downstream steps to convert the amplified RNA into cDNA and sequencing-ready libraries. PCR-based protocols require less hands-on time, but the exponential amplification phase leads to biases in RNA composition in the final libraries. Both approaches were shown to provide interpretable results and were successfully implemented in several scRNA-seq methods.

Transcriptome profiling can be done through **full-length transcript** analysis or by **digital counting of 3' or 5' transcript ends**, also called **tag-based protocols**. The choice of the sequencing method should be dictated by the goal of the experiment. For example, to prioritize cost-effectiveness over retention of sequence information, or deeper sequencing and more reads over the final quality of the sequencing results. Digital RNA counting is a cost-effective quantification strategy, although sequence information of the transcripts is lost to a large extent. Full-length transcriptome sequencing allows the detection of splice variants and alternative transcripts, as well

as genetic alterations in the transcribed fraction, such as single-nucleotide variants and fusion transcripts.

Unlike 3'- and 5'-end methods, **full-length protocols do not allow the introduction of UMIs and impede early cellular barcoding and pooling**, which results in higher costs for library preparation. This limitation can be overcome through **the use of long-read sequencing technologies** that do not need library fragmentation. However, such technologies generate smaller quantities of sequencing reads, and transcriptome quantification is not yet possible.

In the following subsections, some examples of protocols are presented, these examples are taken from the most used methodologies that are present in the literature.

3.2.3 Microtiter-plate-based isolation approach with full-length transcript quantification

After isolation of single cells into microtiter plates by FACS, a full-length transcript protocol can be applied. Smart-seq2 [119] is a widely used method to reverse-transcribe and amplify full-length transcripts. After RT, the enzyme adds cytosines to the cDNA, providing the basis for a template- switching reaction. Here a Template Switching Oligos (TSO) binds to the extra cytosine and provides the template for the addition of PCR adaptor sequences for subsequent cDNA amplification. Compared with the original version, the updated protocol improves molecule-capture efficiency and yield by using locked nucleic acids in the TSO and adding betaine to the RT reaction. Sequencing libraries are prepared by tagmentation, which simultaneously fragments and indexes the cells. The Smart-seq2 protocol is highly efficient in capturing RNA molecules, although the late indexing step makes it more expensive than other methods. Furthermore, the absence of UMIs makes downstream data analysis more challenging and prone to errors generated during PCR. Nevertheless, the protocol provides an adequate solution if deep single-cell phenotyping is required while making a trade-off for cell quantity and the quality and accuracy of the reads obtained (e.g., for homogeneous samples or for analysis of weakly expressed genes). A visualization of the workflow for the Smart-seq2 protocol can be seen in figure 3.5

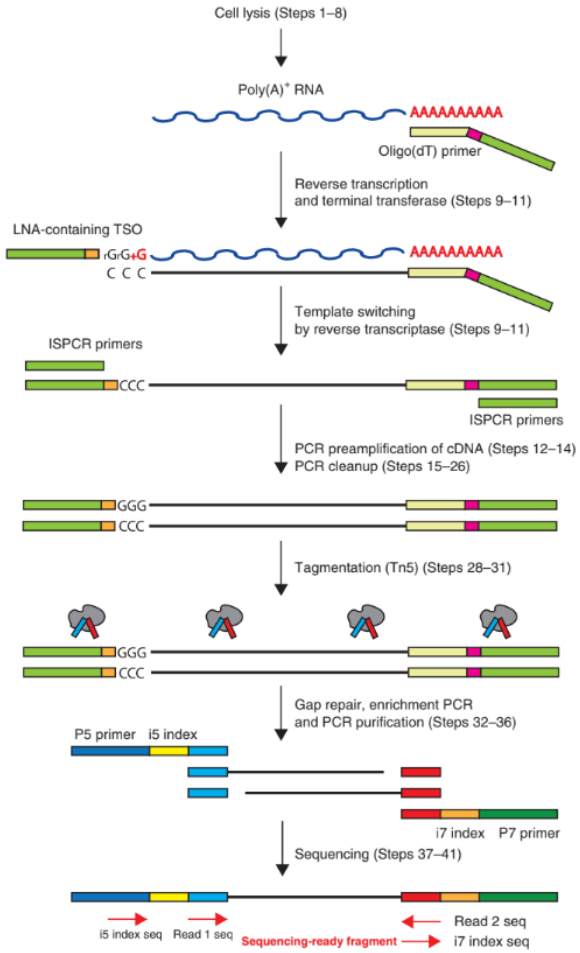


Figure 3.5: **Smart-seq2 protocol for library preparation**: **1)** Cell are disaggregated into a single-cell suspension and lysed with methods capable of lysing cells without interfering with or inhibiting the RT reaction. **2)** mRNA is hybridized with the Oligo(dT) primer to select the mRNA and start the reaction pipeline. **3)** RT is done and the TSO is added to the original sequence along the LNA to start the template switching. **4)** ISPCR primers are added to the sequence and will be used during the pre-amplification step. **5)** PCR preamplification and cleanup are done. **6)** Tagmentation is done on the fragments to attach library adapters (Nextera XT [63]) and divide the cDNA in fragments of a good length for the sequencing platform. **7)** post-tagmentation step are done to repair errors introduced, PCR enrichment and purification of the fragments after PCR is done. **8)** The sequencing library is ready with Illumina adapters(primers for sequencing and the i5-i7 indexes), indexes will be used for pooling (as barcodes)

3.2.4 Microtiter-plate-based isolation approach with 3' or 5' tag-based transcription quantification

STRT-seq [[66],[68]] uses a similar strategy for RT and template switching, but it incorporates single-cell barcodes into the TSO. This allows for early pooling of cells and cost-effective multiplex processing. STRT-seq enriches 5' transcript ends through the use of biotinylated purification and 5'-specific PCR primers. Analysis of the 5' transcripts has the advantage of providing information about transcription start sites.

Moreover, cell barcodes and transcripts are obtained in a single read, which allows for cost-effective single-end sequencing. Although the original STRT-seq protocol could not correct for amplification biases, later updates for the first time included UMIs in an scRNA-seq method.

The SCRB-seq [[15],[143]] protocol incorporates single-cell barcodes and UMIs in the poly(T) primer, thereby enabling 3' amplification of transcripts, and, as with STRT-seq, early indexing allows cell pooling to reduce costs. The RNA capture efficiency of the original protocol was improved by an increase in the RT mix density.

Other protocols exist for this kind of scRNA-seq protocol (Quartz-seq[130], Quartz-seq2[129], CEL-seq2[52], MARS-seq[71]) but they will not be seen here since the technicalities of the protocols are pure biology-related and this research should focus more on the data and the algorithms along with an introduction to the most used technologies and characteristics that will be encountered in the data.

3.2.5 microfluidic systems-based approaches.

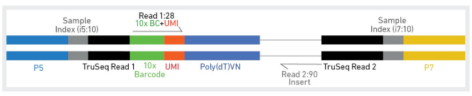
For full-length approaches, the preferred platform is the Fluidigm C1 introduced before and the preferred method is Smart-seq2. This methodology is not that used these days since the microfluidic process is more suited for high throughput with a high cell count protocol while Smart-seq2 and other protocols are limited by the reagents used (since the Illumina indexes seen for smart-seq have been tested for more or less 96 pooled libraries).

Notably, this high-throughput version switched from full-length to 3' RNA sequencing. Also, the array formats, which are restricted to specific cell sizes (small, medium and large arrays), affect unbiased sampling from complex sample types. To further increase cell numbers, microfluidics progressed to open nanowell systems that allow better scalability. In STRT-seq-2i[68], the original protocol was applied in a nanowell platform with 9,600 sites, with cells loaded by limiting dilution or direct addressable FACS sorting. Positioning cells by FACS allows for index sorting that assigns cell properties (e.g., fluorescence signal or size) to array coordinates and barcodes. Nanowells containing cells can be specifically utilized by targeted dispensing, which substantially reduces reagent costs and contamination by ambient RNA. Moreover, the array format allows imaging to exclude doublets. To guarantee high cell viability during the time-consuming loading into nanowells, FCS can be added to the buffer and sample aliquots can be kept on ice. Alternatively, Seq-Well[45] provides a nanowell-based method that captures cells in 86,000 sub-nanoliter reactions. The underlying principle is the preloading of nanowells with barcoded beads before cells enter the capture sites through limited dilution. Subsequently, the arrays are sealed for cell lysis and RNA molecule capture on beads before the immobilized molecules are pooled for 3'-end library production. The Seq-Well system is portable, and so allows sample processing at the sampling sites, as large equipment is not required. The fact that no major investments are required makes the Seq-Well system a flexible and cost-effective alternative. However, although **cells can be monitored by microscopy**, the random distribution of barcoded beads **does not allow the user to integrate imaging data**. Also, the method requires experienced users to obtain reproducible, high- quality results.

Although they are scalable to higher throughputs, the IFC and nanowell approaches are **intrinsically constrained by the number of reaction sites**. **Droplet-based** systems overcome this by encapsulating cells in **nanoliter microreactor droplets**. Here, cell numbers scale linearly with the emulsion volume, and large numbers of droplets are produced at high speed, which facilitates large-scale scRNA-seq experiments. Furthermore, droplet size can be adjusted to reduce potential biases during cell capture. Because barcodes are introduced into droplets randomly, this approach **does not allow the assignment of cell barcodes to images and so precludes the visual detection of doublets and the integrative analysis of cell properties** (e.g., fluorescent signals) with transcriptome profiles. Two droplet-based methods are available:

- **inDrops** [[81],[174]]: This protocol encapsulates cells by using hydrogel beads bearing poly(T) primers with defined barcodes, after which the photo-releasable primers are detached from the beads to improve molecule-capture efficiency and initiate in-drop RT reactions. The barcoded cDNAs are then pooled for linear amplification (IVT) and 3'-end sequencing-library preparation. The technique has extremely high cell-capture efficiency ($> 75\%$) owing to the synchronized delivery of deformable beads, allowing near-perfect loading of droplets. Therefore, the system is most suitable for experiments with limited total numbers of cells. The inDrops system is licensed to 1CellBio, and a variant protocol has been commercialized as the Chromium Single Cell 3' Solution (10x Genomics)[172]. The Chromium system is straightforward to implement and standardize, although library preparation costs are considerably higher than those of the original system. A visualization of the library preparation for the 10x Genomics Chromium platform can be seen in figure 3.6a. The difference between 3' and 5' libraries is that the two assays are similar but capture different ends of the polyadenylated transcript in the final library. Both solutions use polydT primer for reverse transcription, although in the 3' assay the polydT sequence is located on the gel bead oligo, while in the 5' assay the polydT is supplied as an RT primer. A template switching oligo (TSO) is used in both workflows to reverse transcribe the full-length transcript. The final libraries will have the barcode at the 3' end of the transcript (as the 10x Barcode is adjacent to the polyA tail on the 3' end of the transcript) or the 5' end of the transcript (as the the 10x Barcode is adjacent to the TSO and the 5' end of the transcript). A visualization of the two resulting libraries can be seen in figure 3.6b for the 3' libraries reads structure and figure 3.6c for the 5' libraries reads structure.
- Drop-seq[98] uses beads with random barcodes. After cell lysis and RNA capture, the drops are broken and pooled, covalent binding is carried out through cDNA synthesis, the cDNA is amplified by PCR, and 3'-end sequencing libraries are produced by tagmentation. Drop-seq has lower cell-capture efficiency than inDrops methods because beads and cells are delivered by double limiting dilution (double Poisson distribution), which results in 2-4% barcoded cells. The Drop-seq system is commercially available through Dolomite Bio, and a similar system is provided by Illumina (ddSEQ).

(a) **10X Genomics Chromium library preparation from beads to**



(b) **Single Cell 3' v3.1(Dual Index) Gene Expression Library**

(c) **Single Cell 5' v2 Gene Expression Library:**

Figure 3.6: **10X Genomics Chromium library preparation. Source** [2]: Cells are captured in individual oil droplets containing a bead (called GEMs). An individual bead contains adapters with a common barcode, but diverse and distinct Unique Molecular Identifier (UMI) sequences. A poly(dT) primer is used to reverse-transcribe mRNA with poly-A tails into cDNA. The GEMs are then broken and the pooled cDNA (from all barcoded cells) is amplified by PCR. Finally, the cDNA is fragmented and another Illumina adapter is ligated at the other end of the molecule.

3.2.6 split-pool barcoding-based approaches

Conceptually different from the previous techniques are methods based on combinatorial barcoding. Here, cells are not processed as individual units but isolated in pools. These pools are split and mixed, with each round integrating pool-specific barcodes. The combination of such pool indices results in unique barcode combinations for each cell through their random assignment during consecutive pooling processes. Both split-pooling methods, SPLiT-seq (split-pool ligation-based transcriptome sequencing)[127] and sci-RNA-seq (single-cell combinatorial-indexing RNA-seq)[22], were shown to reliably produce single-cell transcriptomes and to be scalable to hundreds of thousands of cells per experiment. SPLiT-seq includes four rounds of indexing, resulting in ¿20 million possible barcode combinations. After initial indexing during RT, two rounds of index ligation and a final PCR indexing step create cell-specific barcoded 3'-transcript

libraries. During the second ligation round, UMIs are incorporated for the subsequent correction of amplification biases. Additional rounds of barcoding or a switch from 96-well to 384-well microtiter formats could further scale up cell numbers. The original sci-RNA-seq protocol includes a two-step indexing workflow with the first index and UMI introduced during RT and a second index during PCR amplification (after tagmentation). The use of indexed tagmentation sequences could further scale up possible barcode combinations and increase cell numbers per experiment.

A complete workflow of scRNA-seq can be seen in figure 3.7

3.2.7 Sequencing

Sequencing is the same as for bulk since the protocols used attach the same primers that are used for bulk sequencing. The final results will be the reads for the fragments of cDNA that will contain barcodes (necessary for pooling of different fragments from different cells) and optionally UMI if the protocols permit it. High-throughput microfluidics-based experiments generally involve sequencing to lower depths (¡100,000 reads per cell), whereas higher read numbers (500,000 reads per cell) are optimal for many microtiter-plate formats. Nevertheless, single-cell libraries are usually not sequenced to saturation, and the phenotyping resolution (detection of more genes and of those expressed at lower levels) can benefit from further increases in the sequencing depth. Annotation of splice variants from full-length transcriptomes requires deeper sequencing to better resolve the expression levels of transcript variants.

After the sequencing, the output will be a **fasta or fastq** file containing the reads, exactly the same as 2.1. Additional information could be associated and written in other files (like the composition of the sample obtained via FACS or microscopy imaging).

Another thing to consider after having sequenced the fragments is the quality of the reads. Many tools exists to control the quality of the reads, in this research **FASTQC** (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) will be used.

3.2.8 Alignment

After the sequencing process, the alignment process is done after the preprocessing of the reads since they contain both barcodes and UMI (for tag-based protocols).

The process of alignment for full-length methods that do not use UMI is straightforward since, after the first step that will be demultiplexing of the reads, the resulting demultiplexed reads (that could be saved into different fastq files that do not have the barcode) can be aligned directly with the method seen during bulk alignment.

What was introduces previously in the 2.2.3 section for bulk RNA-seq and full-length scRNA-seq methods must be slightly modified for tag-based protocols since:

- Each sequenced fragment will have a cell barcode (CB) that identifies the cell where the RNA is expressed;
- Each sequenced fragment will have useful UMI for mutations during PCR and to control possible bias introduced by amplification
- The amount of sequencing data produced will be immensely less than with bulk RNA-seq methods, so the data obtained is the result of greater amplification

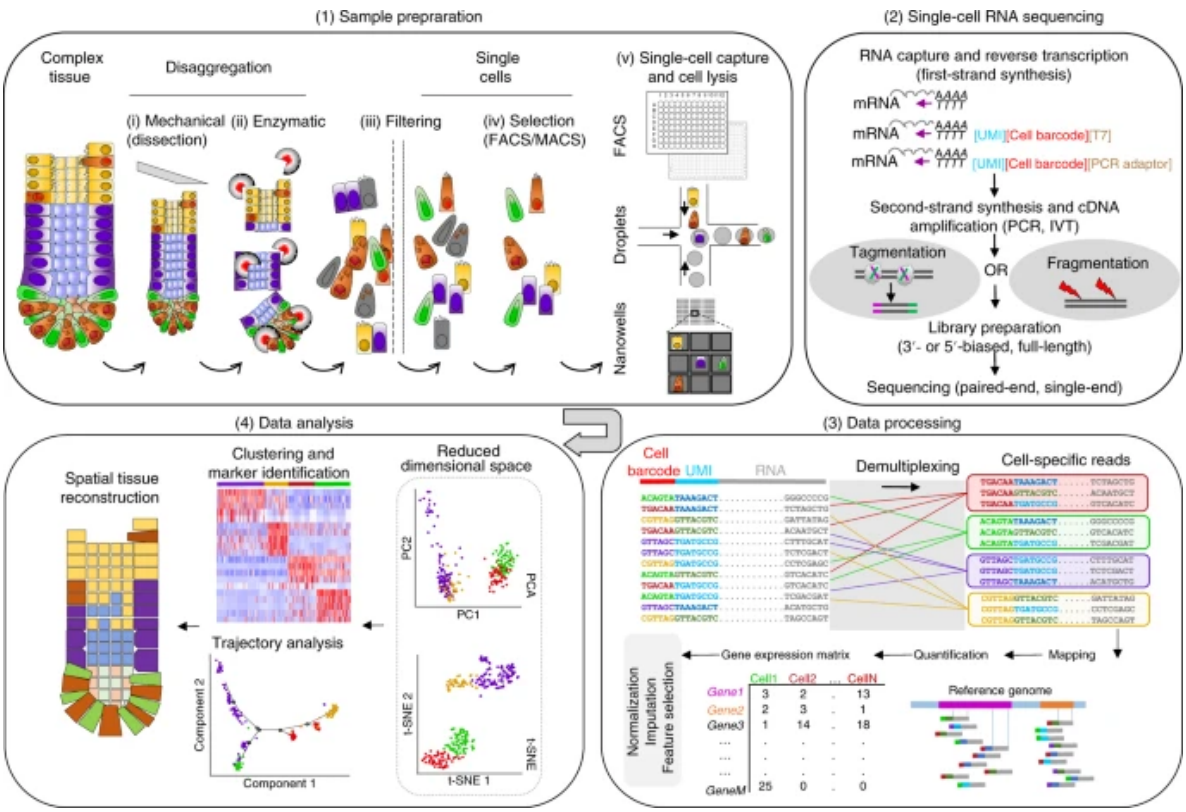


Figure 3.7: **The single-cell RNA sequencing process**. (1) During sample preparation, cells are physically separated into a single-cell solution from which specific cell types can be enriched or excluded (optional). After they have been captured in wells or droplets, single cells are lysed, and the RNA is released for subsequent processing. (2) To convert RNA into sequencing-ready libraries, poly(A)-tailed RNA molecules are captured on poly(T) oligonucleotides that can contain unique molecular identifier (UMI) sequences and single-cell-specific barcodes (5'- and 3'-biased methods). To allow for subsequent amplification of the RNA by PCR or IVT, adaptors or T7 polymerase promoter sequences, respectively, are included in the oligonucleotides. After RT into cDNA and second-strand synthesis (optional), the transcriptome is amplified (PCR or IVT). For conversion into sequencing libraries, the amplicons are fragmented by enzymatic (e.g., tagmentation) or mechanical (e.g., ultrasound) forces. Sequencing adaptors are attached during a final amplification step. Full-length sequencing can be carried out, or 5' or 3' transcript ends can be selected for sequencing using specific amplification primers (optional). For most applications, paired-end sequencing is required. (3) The sequencing reads are demultiplexed on the basis of cell-specific barcodes and mapped to the respective reference genome. UMI sequences are used for the digital counting of RNA molecules and for correction of amplification biases. The resulting gene-expression quantification matrix can subsequently be normalized, and missing values could be imputed. (4) Dimensionality reduction is done for reducing noise and visualization. Data analysis can then be tailored to the underlying dataset and objective(clustering, identifying cell types and states, or ordered along a predicted trajectory in pseudotime). Eventually, the spatial cellular organization can be reconstructed through the interrogation of marker genes (experimentally) or through marker-guided computational reconstruction (inference).

Therefore, tag-based protocols are often accompanied by paired-end sequencing strategies which on one side of the nucleotide sequence reads the CB + UMI pair, on the other the actual transcript(with the adapter and the index in case of Illumina sequencing). A possible pipeline for building expression data is as follows:

1. Assign reads to cells (via demultiplexing and using barcodes)
2. Counting the number of unique RNA molecules (UMI demultiplexing/counting and deduplication)
3. Map cDNA fragments to a reference database
4. Assigning reads to genes (counting step that will be seen in the next section 3.3)

The first important step is de-multiplexing of reads using cell barcodes. Whereas Smart-seq libraries can be directly de-multiplexed using the index reads, the 3'-end tag-based methods require a dedicated processing step to identify the single-cell indexes in the sequencing reads. De-multiplexed reads are then mapped to reference genomes with the alignment tools seen in 2.2.3 (UMIs need to be taken into account with tools that group reads from their UMI identifier like **UMI-tools**[141], but they are usually taken into consideration by the alignment tool if available). Proprietary software for alignment is available and will take into account the singularities of the library like UMI, barcodes and specific characteristics of the reads. One of this tools is **CellRanger** [1] from 10x Genomics that will take reads from the 10x Chromium platform and workflow. Cell Ranger is a set of analysis pipelines that process Chromium single-cell data to align reads, generate feature-barcode matrices, perform clustering and other secondary analysis like UMI counting and uses STAR[32] to align the reads. UMI counting consists of read counting followed by PCR duplicate collapsing based on UMI sequences. Before counting UMIs, Cell Ranger attempts to correct for sequencing errors in the UMI sequences. Reads that were confidently mapped to the transcriptome are placed into groups that share the same barcode, UMI, and gene annotation. If two groups of reads have the same barcode and gene, but their UMIs differ by a single base (i.e., are Hamming distance 1 apart), then one of the UMIs was likely introduced by a substitution error in sequencing. In this case, the UMI of the less-supported read group is corrected to the UMI with higher support. Cell Ranger again groups the reads by barcode, UMI (possibly corrected), and gene annotation. If two or more groups of reads have the same barcode and UMI, but different gene annotations, the gene annotation with the most supporting reads is kept for UMI counting, and the other read groups are discarded. In case of a tie for maximal read support, all read groups are discarded, as the gene cannot be confidently assigned. After these two filtering steps, each observed barcode, UMI, gene combination is recorded as a UMI count in the unfiltered feature-barcode (i.e. gene-cell) matrix. The number of reads supporting each counted UMI is also recorded in the molecule info file.

Another tool that can take into account the different characteristics of the reads without being proprietary software is **STARsolo** [74] that is also part of the STAR[32] project and is designed to be a drop-in replacement for 10X CellRanger gene quantification output. It follows CellRanger logic for cell barcode whitelisting and UMI deduplication and produces nearly identical gene counts in the same format while being several times faster. STARsolo is a standalone pipeline that is a part of STAR RNA-seq aligner mentioned in this chapter previously. It was developed with a goal to

generate results that are very similar to Cell Ranger, while remaining computationally efficient. Normally, STARsolo is several times faster than Cell Ranger on the same dataset. STARsolo methods for UMI collapsing, cell barcode demultiplexing, and cell filtering are purposefully re-implementing the algorithms used by Cell Ranger. In recent versions, STARsolo is also capable of quantifying multi-mapping read correctly, making it a very attractive option for fast and accurate scRNA-seq processing. Additional benefit of STARsolo is its flexible implementation of cellular barcode and UMI search: knowing a relative location within a read, and length of each sequence, it's possible to process the data generated by most scRNA-seq approaches.

Recent alignment tools not directly related to the ones previously seen and directly developed with the objective of scRNA-seq in mind were optimized for fast handling of large-scale datasets without loss of accuracy. For example, **Kallisto**[21] reduces the alignment time by two orders of magnitude through pseudo-alignment, as opposed to alignment of individual bases, this method is also really similar to **Salmon**[115] seen in the 2.2.3 section. In a final processing step, mapped reads are quantified to create a transcript expression matrix. RSEM[87], Cufflinks[154] and HTSeq[9] can be used for full-length transcript datasets, whereas special tools, such as UMI-tools[141], which accounts for sequencing errors in UMI sequences are available for counting UMI-tagged data types. Another module directly integrated in salmon to treat 3' libraries with barcodes and UMI is **Alevin** [128].

A useful resource to **find the perfect pipeline/workflow** to use to analyze the data is found at `https://www.scrna-tools.org/`, which is a search tool that queries a database that provides a comprehensive list of available computational tools for data processing and analysis. Methods are categorized by analysis task, and researchers can select tools according to the required analysis type.

After the alignment of the reads, the next step that was introduced in this section and is also part of the classic pipeline of bulk is the estimation of **gene expression** via counting/quantification for the single cells.

3.3 Gene expression for single cells

After the cells were demultiplexed (usually into single fastq files), and optionally aligned (since some quantification tool are alignment-free or integrate a step of pseudo-alignment directly in the quantification/counting step, as introduced in 2.2.3 and explained in 2.3), the next step is to map the reads to the genes and create the vectors of counts/quantification-score for each gene in every cell. In the case of full-length protocols, after demultiplexing the reads with the cell barcodes, the results should be single fastq files for the single cell. To generate the count/quantification vector for each cell (with the whole matrix as the genes as rows and the cells as columns), the tools seen in 2.3 could be used with no problem, e.g. salmon[115] or [9].

When dealing with 3' or 5' tag-based reads, the problem is different since the counting tools need to take into account UMI, especially quantification tools that take fastq files directly and integrate the alignment step directly in the tool itself (e.g. salmon or kallisto). The reads originating from different molecules of the same transcript would

have originated only from the 3' end of the transcripts, so would have a high likelihood of having the same sequence. However, the PCR step during library preparation could also generate read duplicates. To determine whether a read is a biological or technical duplicate, these methods use the unique molecular identifiers.

- Reads with different UMIs mapping to the same transcript were derived from different molecules and are biological duplicates - each read should be counted.
- Reads with the same UMI originated from the same molecule and are technical duplicates - the UMIs should be collapsed to be counted as a single read. This is a consensus collapse since the same reads with the same UMI that are in the majority will be the original read and will be collapsed into the single read while the other reads that have a lesser count will be discarded.
- If the reads originating from the same UMIs have a uniform distribution to the reads associated, that means that there is no consensus on the read associated with the original strand so all the reads are discarded and the UMI is unusable.

Once gene expression has been quantified it is summarized as an expression matrix where each row corresponds to a gene (or transcript) and each column corresponds to a single cell.

3.3.1 Quality control

After the quantification and the summarization of the expression matrix, the matrix should be examined to remove poor-quality cells. Failure to remove low-quality cells at this stage may add technical noise which has the potential to obscure the biological signals of interest in the downstream analysis. In the case of 3' or 5' tag-based protocols, unfiltered ("raw") feature-barcode matrix contains many columns that are in fact empty droplets. Gene expression counts in these droplets are not zero due to technical noise, e.g. the presence of ambient RNA from broken cells. However, they can usually be distinguished from captured real cells by the amount of RNA present. The process of filtering these cells is called **cell filtering** and it is implemented in Cell Ranger. There are two algorithms implemented for this cell filtering in Cell Ranger, which will be referred to as "Cell Ranger 2.2" and "Cell Ranger 3.0" filtering. The original algorithm (Cell Ranger 2.2) identified the first "knee point" in the "barcode count vs UMIs per barcode" plot. Cell Ranger 3.0 introduced an improved cell-calling algorithm that is better able to identify populations of low RNA content cells, especially when low RNA content cells are mixed into a population of high RNA content cells.

Another form of quality control is the removal of genes that seems to be dropouts or outliers. This type of quality control was already introduced with the UMI preprocessing, where lower quality reads and reads where the UMI were not associated principally with a certain read were discarded. This was a form of QC since the genes that would have been associated with the read will not be expressed and, consequently, filtered since the reads associated with that gene are too few or zero.

Other forms of quality control could be done if the scRNA-seq is associated with bulk RNA-seq from the same tissue and the same patient/environment since the deconvolution of bulk data could be used to control the quality of the cells found and also see if the overall(the sum of the expression for all the cells) expression of the single cells

is similar to the expression resulting from the bulk sequencing method. More methods that use bulk and single-cell protocols at the same time will be seen in 4.

For this research, during the experimental analysis done in 8 genes that do not exceed a threshold will be discarded, and cells that have not enough expressed genes will be discarded as well.

3.3.2 Imputation methods

In addition to having a high noise level, scRNA-seq datasets are also very sparse, or fraction of observed "zeros", where a zero refers to no unique molecular identifiers (UMIs) or reads mapping to a given gene in a cell. which poses further challenges to cellular phenotyping and data interpretation. Non-expressed genes and technical shortcomings, such as dropout events (unsequenced transcripts), result in many zeros in the expression matrix, and thus an incomplete description of a single cell's transcriptome.

In contrast to bulk RNA-sequencing (RNA-seq), the scRNA-seq problem of the increased sparsity can be due to biological (relevant or nuisance) fluctuations in the measured trait or technical limitations related to challenges in quantifying small numbers of molecules. Examples of the latter include mRNA degradation during cell lysis or variation by chance of sampling lowly expressed transcripts. The word **dropout** has been previously used to describe both biological and technical observed zeros, but the problem with using this catch-all term is it does not distinguish between the types of sparsity.

To reduce sparsity, missing transcript values can be computationally inferred by imputation, recent work has led to the development of "imputation" methods, in a similar spirit to imputing genotype data for genotypes that are missing or not observed. However, one major difference is that in scRNA-seq standard transcriptome reference maps such as the Human Cell Atlas or the Tabula Muris Consortium are not yet widely available for all species, tissue types, genders, and so on. Therefore, the majority of imputation methods developed to date do not rely on an external reference map.

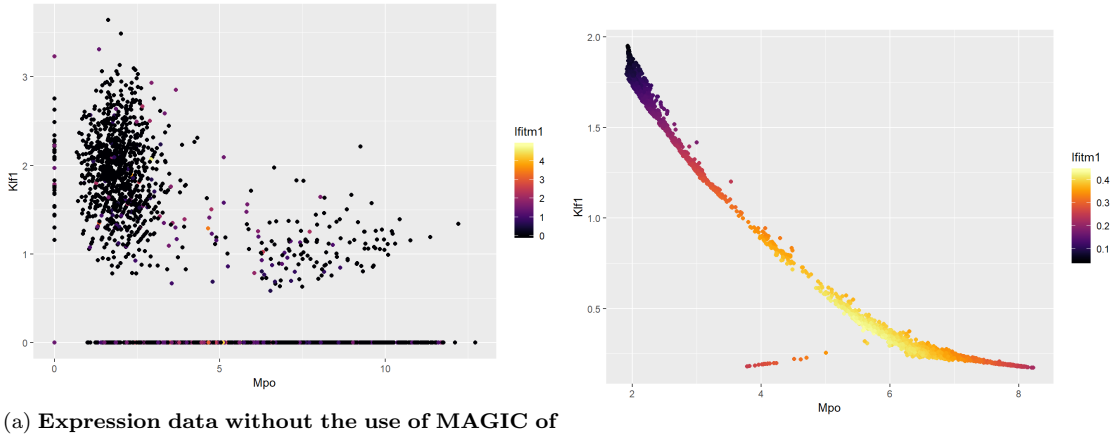
These imputation methods can be categorized into three broad approaches. The first group is imputation methods that directly model the sparsity using probabilistic models. These methods may or may not distinguish between biological and technical zeros, but if they do, they typically impute gene expression values for only the latter. A second approach adjusts (usually) all values (zero and non-zero) by smoothing or diffusing the gene expression values in cells with similar expression profiles identified, for example, using neighbors in the graph. The third approach first identifies a latent space representation of the cells, either through low-rank matrix-based methods (capturing linear relationships) or deep-learning methods (capturing non-linear relationships), and then reconstructs the observed expression matrix from the low-rank or estimated latent spaces, which will no longer be sparse. For the deep-learning approaches, such as variational autoencoders, both the estimated latent spaces and the "imputed" data (i.e., reconstructed expression matrix) can be used for downstream analyses, but otherwise only the imputed data is typically provided for downstream analyses.

Evaluations and comparisons between scRNA-seq imputation methods have been limited or restricted to a subset of imputation methods and downstream applications.

Imputation methods can require varying types of raw or processed data as input, may rely on different methodological assumptions, and may be appropriate for only certain scRNA-seq experimental protocols, such as UMI-based (tag-based) or full-length transcript methods previously introduced. Given these differences, the performance of these methods varies in the evaluations and in the final results or conclusions obtained. Therefore, the answer to the question of which methods can, let alone should, be used for a particular analysis is often unclear. Another key criterion to evaluate the imputation methods is their ability to recover transcriptome dynamics (gene-gene interaction, regulation, response, etc.) in real single-cell data set. The most important thing that an imputation method needs to have is the ability to discern true zero-expressed genes (or very low expressed genes) from false zeros and have similar values to the normal tissue and cells (for a population or a reference).

A study that compares the tools used for imputation and scores them based on expression values from bulk RNA-seq (by comparing results of various pipelines and the expression values of the single-cells after imputation compared to bulk RNA-seq) is [60]. A visualization that exposes the pipeline while using various imputation methods in the aforementioned research can be seen in figure 3.8. To get more information about these tools and a benchmark that compares these methods and the pros and cons for every method refer to the research [60]. The comparison of methods is done on bulk RNA-seq for homogeneous tissue or single cells of the same type grouped in a sample and sequenced. Genes for cells resulting from scRNA-seq are compared to the expression distribution of the bulk RNA-seq counterpart where the majority of the cells are of the same type as the one being analyzed.

A famous method that will be seen here in more detail is **MAGIC(Markov Affinity-Based Graph Imputation of Cells)**[156], which uses diffusion maps to find data structures and restore missing information. An example of the use of MAGIC can be seen in figure 3.9.



(a) **Expression data without the use of MAGIC of two marker genes**

(b) **Expression data after the use of MAGIC**

Figure 3.9: **MAGIC use, gene-gene relationship between two markers before and after MAGIC**: The final expression imputed after MAGIC match what is seen in the literature about the type of cells seen in the source [99]

Alternatively, scImpute[89] learns a gene's dropout probability by fitting a mixture

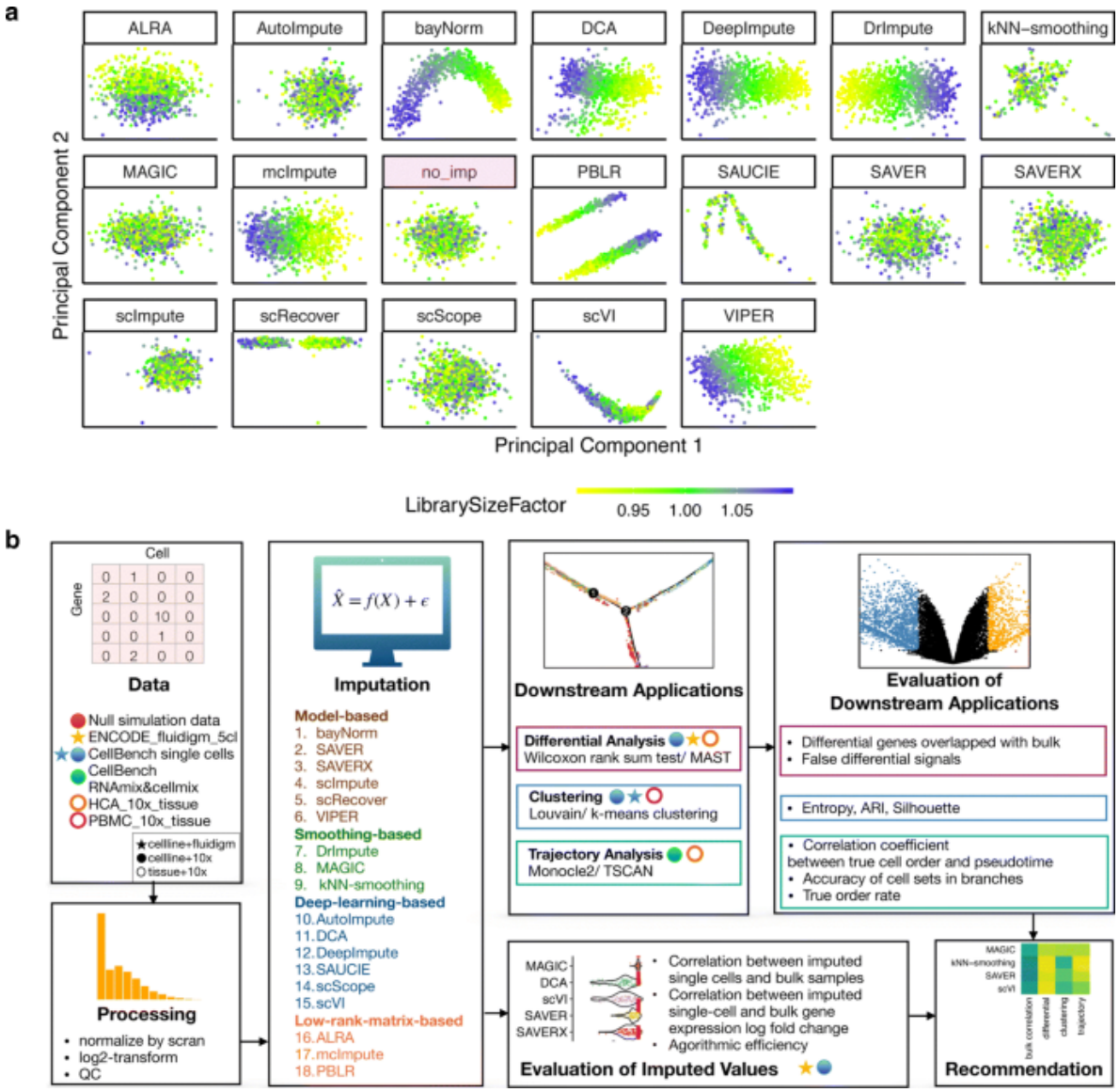


Figure 3.8: Single-cell workflow with imputation methods and benchmarking. Source [60]

model and then imputes probable dropout events by borrowing information from similar cells (selected on the basis of genes that are not severely affected).

Another method that uses Graphs and Graph neural networks (that will be introduced in 6.1) is GraphSCI (Graph Single Cell Imputation) [124]. The existing imputation methods for scRNA-seq aim at learning the similarity of cells or genes but do not consider gene-gene relationships and cell-cell correlations simultaneously, resulting in the fact that they cannot retain biological variation across cells or genes. Decades of molecular biology research have discovered much about the principles of gene interaction and their influence on gene expression. For example, the gene is truly not expressed due to gene regulation, but imputed by similar cells, which makes it difficult to study cell-cell variation and downstream analysis. This means that imputation methods not only need to take advantage of the information between similar cells but also gene-gene relationships (these relations and network regulation of genes will be at the center of chapter 6.1 and in particular when talking about biological networks in chapter 5). More importantly, as imputation proceeds, the imputed gene expression matrix could infer more accurate gene-gene relationships while the inferred gene-gene relationship helps improve the accuracy of imputation. The imputation method used by GraphSCI needs to be able to dynamically integrate the imputation of gene expressions and inference of the gene-gene relationships during the training process. GraphSCI is a Single-Cell Imputation method that combines Graph convolution network (GCN) and Autoencoder neural networks, to impute the dropout events in scRNA-seq by systematically integrating the gene expression with gene-gene relationships. GraphSCI uses gene-gene relationships as prior knowledge to recover gene expression in a single cell because gene-gene interactions are likely to affect gene expression sensitively. And the combination of GCN and autoencoder neural networks makes it possible to dynamically utilize the increasingly accurate gene-gene relationships to impute gene expressions. By stacking the GCN and autoencoder network, GraphSCI is capable of exploring the gene-gene relationships in an explicit way, so as to impute the sparsity events effectively. Furthermore, the deep generative model with gene-specific distribution could learn the true data distribution of scRNA-seq data and then impute the dropout events and avoid overfitting. The gene-gene relationships can be regarded as a gene graph, in which the gene is the node and the edge is the relationship. As a consequence, the imputation task of gene expression can be converted into the node recovering problem on graphs. GCN [78] is a very powerful neural network architecture for machine learning on graphs but it is limited due to the convolutional nature that is not always the best approach for graphs. GCN was designed to learn hidden layer representations that encode both local graph structure and features of nodes and edges. Some research describes applications of GCN such as node recovering problems. More recent research on graph neural networks [173] shows better approaches that could be used for future graph-based algorithms for imputation but are not used in GraphSCI. Based on the co-embedding attributed network, GraphSCI combines GCN and autoencoder neural network to systematically learn the low-dimensional embedded representations of genes and cells. GCN exploits the spatial feature of gene-gene relationships effectively while Autoencoder neural network learns the non-linear relationships of cells and counting structures of scRNA-seq data, and thus the deep learning framework reconstructs gene

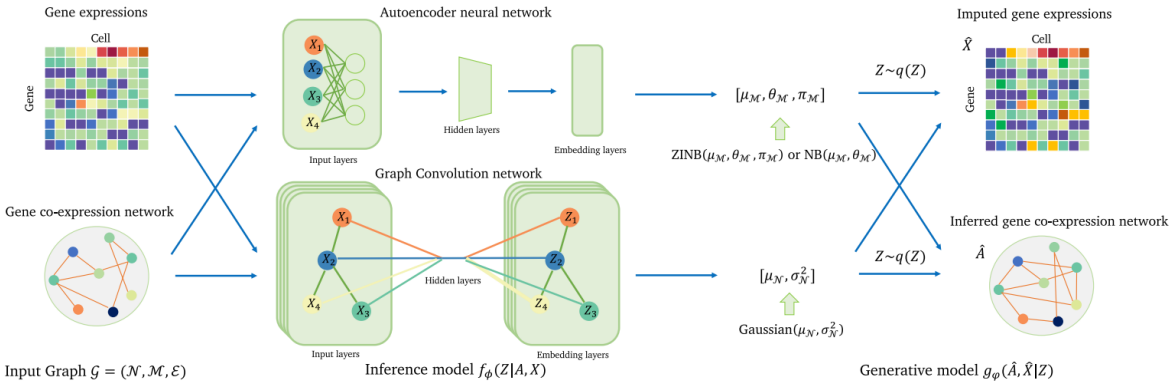


Figure 3.10: **The workflow of the GraphSCI model**: The input of GraphSCI is the gene expression matrix and the gene-gene relationships. The Inference model f_φ is to learn the low-dimensional representations of genes and cells based on a combination of graph convolution network and Autoencoder neural network. The Generative model g_Φ utilizes the posterior distributions to reconstruct gene expression and gene-gene relationships respectively.

expressions by integrating gene expressions and gene-gene relationships dynamically in the backward propagation of neural networks. GraphSCI proposed method is declared to outperform competing methods over both simulated and real data sets by diverse downstream analyses. From the results that they propose, it seems that GraphSCI is comparable to other methods but it models the results to maximize clustering metrics (silhouette) resulting in close cells and too much bias. The method is innovative nonetheless but can be improved a lot more.

The workflow that GraphSCI uses can be seen in figure 3.10

All the methods seen here do not use cell types as reference to generate the new expression values but impute the data based on statistical inference (MAGIC,knn-smooth), based on models(GraphSCI and scImpute, although scImpute make a distinction not as the type of cells but with cell clusters and by selecting some cells for the imputation of a group) and combined methods that use both bulk and single-cell data that will be seen in 4(SCRABBLE[117]).

It was also demonstrated that all the imputation method have an inherent problem related to the generation of false positives [10] and thus, statistical tests applied to imputed data should be treated with care.

In addition to the methods introduced here, another project called SCRABBLE that uses both **bulk** and **single-cell** data will be used for imputation [117] and this method will be seen in 4.

3.3.3 Normalization

Library sizes vary because scRNA-seq data is often sequenced on highly multiplexed platforms the total reads which are derived from each cell may differ substantially. Some quantification methods (eg. Cufflinks, RSEM) incorporated library size when determining gene expression estimates and thus do not require this normalization.

However, if another quantification method was used then library size must be corrected by multiplying or dividing each column of the expression matrix by a "normal-

ization factor" which is an estimate of the library size relative to the other cells. Many methods to correct for library size have been developed for bulk RNA-seq and can be equally applied to scRNA-seq like the ones seen in 2.3.1 (e.g. RPKM, FPKM, TPM).

Single-cell RNA-seq datasets show high levels of noise and variability related to non-biological technical effects, including dropout events due to stochastic RNA loss during sample preparation, biased amplification and incomplete library sequencing. Technical variation also results from batch effects on processing units (e.g., plates or arrays), time points, facilities and other sources. Moreover, natural variability complicates analysis because of, for example, variable cell size and RNA content, different cell cycle stages and gender differences. Therefore, dataset normalization becomes an important step for meaningful data analysis. This can be guided by the addition of artificial spike-in RNA, which is used to model technical noise. However, it is not clear whether artificial RNA sufficiently reflects the behavior of endogenous RNA, or whether cellular RNA influences spike-in detection. Recent high-throughput methods distribute cells by limiting dilution, which makes the use of spike-in RNA impracticable because of the high number of otherwise empty reaction volumes. Alternative normalization methods originally developed for bulk RNA sequencing, such as log-expression, trimmed mean M-values(TMM) and upper-quartiles can also be used in scRNA- seq, although more-specialized normalization methods are being developed that can better handle many aspects of this specific type of data. Recent single-cell approaches apply between-sample normalization or normalize on cell-based factors after pool-based size factor deconvolution (scran[83] that implements a variant on CPM specialized for single-cell data). However, for correction for large-scale sources of variation, a recommended and standard procedure is data modeling with the correct distribution. Here, confounding factors can be incorporated as covariates into the model and regressed out. Whereas batch effects are usually detected by visual inspection of reduced-space representations (e.g., principal components).

Digital normalization methodologies should take into consideration cell size and the protocol used to obtain the reads since some protocols are aimed at a deeper sequencing while others search for more coverage and throughput of cells.

Right now there is no state of the art for normalization methodology in scRNA-seq, so the suggestion is to try different methods and normalize the expression matrix based on the characteristics of the library and the objectives of the research.

The method for normalization used in **Seurat** is a simple **log-normalization** where feature counts for each cell are divided by the total counts for that cell and multiplied by a scale factor(10000 default, otherwise passed as an argument). Then the result is transformed via a natural logarithm. The Log-normalization will be used for this project during 8

3.4 Differential expression techniques in scRNA-seq

After identification of the cell type identities of the scRNA-seq clusters, the research methodology should often proceed to perform differential expression analysis between conditions within particular cell types. It was seen both in the introduction 1 and from

common biology that single cells within a sample are not independent of each other, since they are isolated from the same animal/sample from the same environment. If the cells are treated as samples, then the research is not truly investigating variation across a population and information obtained from the process, but variation among an individual. Therefore, the only conclusions that could be made are at the level of the individual, not the population. Usually, the research need to focus in inferring which genes might be important for a condition at the population level (not the individual level), so the samples are needed to be acquired from different organisms/samples, not different cells. To do this, the current best practice is using a pseudo-bulk approach, which involves the following steps:

- Subsetting to the cells for the cell type(s) of interest to perform the DE analysis.
- Extracting the raw counts after QC filtering of cells to be used for the DE analysis
- Aggregating the counts and metadata to the sample level.
- Performing the DE analysis (Need at least two biological replicates per condition to perform the analysis, but more replicates are recommended).

The reasons to use pseudo-bulk analysis lie in how scRNA-seq data is characterized by the following properties of single-cell biology and derived data:

- scRNA-seq data tend to exhibit an abundance of zero counts, a complicated distribution, high heterogeneity
- The heterogeneity within and between cell population pose a challenge to classical techniques for DE (especially for bulk RNA-seq methodologies)
- Single-cell methods to identify highly expressed genes as DE and exhibit low sensitivity for genes having low expression
- Single-cell methods often inflate the p-values as each cell is treated as an independent sample. If cells are treated as samples, then variation across a population is not truly investigated
- Additional analysis from different samples should not be done on different cells (as it will be seen as one of the possible paths to take to do DE analysis) but on different organisms with the same cell type.

Pseudo-bulk methods aggregate single-cell gene expression by summarizing the input measurement values for a given gene over all cells in each subpopulation and by sample. The resulting pseudo-bulk data matrix has dimensions G $\times$ S, where S denotes the number of samples(summarized samples after the aggregation), with one matrix obtained per subpopulation. Depending on the specific method, which includes both a type of data to operate on (e.g., counts, logcounts) and summary function (e.g., mean, sum), the varying number of cells between samples and subpopulations is accounted for prior to or following aggregation. Every column in the resulting matrix is a summarized replicate that could be used for Differential expression analysis or other types of analysis. The aggregation function could be seen as a convolution or another type of non-linear mapping that can map the expression matrix for the cell type with some certain features (control, treatment, disease, etc.) into a matrix with a lesser column dimension. Methods from artificial intelligence can be used to create the aggregated pseudobulk matrix like CNN or other algorithms that transform a matrix of G x M in a matrix of G x S.

A workflow to visualize the steps of pseudo-bulk differential analysis for single cells

can be seen in figure 3.11

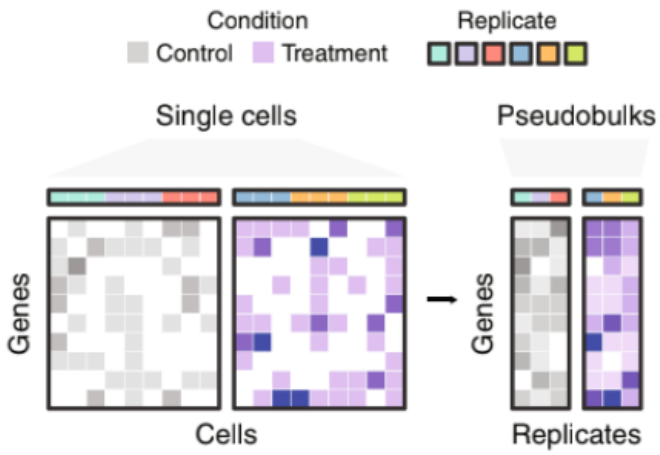


Figure 3.11: **Pseudo-bulk methodology**

After the pseudo-bulk, one can follow two (or more depending on the experiment) different paths to conduct DE analysis:

- Compare different cell groups associated with some types from the same sample, the comparison will be done between different types of cells. This approach need to aggregate the differentially expressed genes in some way, and in this project the ways presented will be two:
  - **Union** of all the differentially expressed genes between one group and all the others. This approach has a wider spectrum of differentially expressed genes while losing precision.
  - **Intersection** of all the differentially expressed genes between one group and all the others. This approach is more precise and will obtain very specific differentially expressed genes unique to the fixed one group.
- Compare different patient samples from the same tissue with the same type associated with the cell.

It is also possible to follow an **ensemble approach** that uses both of the two paths seen and does DE analysis between cells while also validating the differentially expressed genes among different samples.

To get the same type of cell among different patient's samples, one can think of clustering cells in different patients (of the same tissue if doing traditional differential expression analysis) with some of the methods defined in 3.5.1 and associate with every cluster a cell type as it will be described in 3.5.2. The information of the origin of the cell will be retained as metadata and will be used during aggregation to create the pseudo-bulk matrices (**one for every pair (sample/patient + cellType) treated as an identifier for the matrix**).

After having obtained the pseudo-bulk matrix for the pair (sample/patient + cellType), the analysis can be done with the classical Differential Expression tools seen in 2.4 between the matrices obtained with different cell types and without making a distinction from the source sample or between matrices with the same cell type associated

and different samples. Also as stated before, a hybrid method can also be used to find differentially expressed genes that are both differentially expressed cellType-cellType and sample-sample with the same cellType.

It is also possible to fix a single group as an "altered sample" and treat all the other cells as being part of the "controls samples" when comparing cells that have the same sample source, while also aggregating the controls sample to imitate bulk differential expression analysis. In this way, the resulting differentially expressed genes found should be more accurate overall and also resemble how DE is done in bulk since aggregated/pseudobulked cells should have similar values to what is obtained in bulk sequencing of the same tissue(without the "altered sample" since that group will not be part of the pseudobulk replicate). This approach also has a lot of drawbacks since single-cell information is lost, expression values are lost and the differentially expressed genes will have less precision when targeting differences for smaller groups.

A different differential expression package and methodology that is used for single-cell and is used, especially, by Seurat if set as an option is **MAST** [104]. MAST framework models single-cell gene expression using a two-part generalized linear model. One component of MAST models the discrete expression rate of each gene across cells, while the other component models the conditional continuous expression level (conditional on the gene being expressed). Details of the model can be found in the reference and will not be seen here since it is not used for this research.

For this research, during the experimental analysis in 8, a pseudo-bulk approach will be used(where cells of the same type will be summed into a more compact matrix) and **Deseq2** [97] will be used to search for the differentially expressed genes in one cell group/cluster in comparison with the other groups. These identified differentially expressed genes will be used for building the pathway embeddings with some of the methods seen in 6.

3.5 Additional methods

An additional and new method to take into account is Smart-seq3 which is the new version of Smart-seq and introduces UMI to full-length protocols. Most single-cell RNA-sequencing (scRNA-seq) methods that were seen in this chapter count RNAs by sequencing a unique molecular identifier (UMI) together with a short part of the RNA (from either the 5' or 3' end) . These RNA end counting strategies have been effective in estimating gene expression across large numbers of cells, while controlling for PCR amplification biases, yet RNA end sequencing provides limited coverage of transcribed genetic variation and transcript isoform expression. Moreover, many massively parallel methods suffer from rather low sensitivity (that is, capturing a small fraction of RNAs present in cells) . In contrast, Smart-seq2 has combined higher sensitivity with full-length coverage , which enabled allele-resolved expression analyses , but at the cost of lower cellular throughput, higher cost and without the incorporation of UMIs. Sequencing of full-length transcripts using long-read sequencing technologies can directly quantify allele/isoform-level expression, yet their current read depths hinder their broad application across cells, tissues and organisms. To overcome these

shortcomings, a new version of Smart-Seq wad developed to account for sensitive short-read sequencing method that would extend the RNA counting paradigm to directly assign individual RNA molecules to isoforms and establish their allelic origin in single cells. At the end of the protocol, the reads for 3' UMI and for the full-length part will be separated. The protocols will not be seen in detail here since it is quite complex and it was only mentioned since it is one of the most important protocols for scRNA-seq.

3.5.1 Clustering

Clustering is an important step that is used to group cells in clusters and these clusters should be composed of similar cells with the same biological function, and that means that the grouped cells should have similar expressions and also share the same **Gene Markers** to differentiate them from other groups.

Although prior assumptions and canonical population markers allow supervised clustering (e.g., Monocle[27] uses a supervised clustering approach first to evaluate and build the pseudo-time model), hypothesis-free unsupervised clustering is preferred in most cases. A commonly used unsupervised algorithm is hierarchical clustering, which provides consistent results without a pre-defined number of clusters. Hierarchical clustering can be conducted in an agglomerative (bottom-up) or divisive (top-down) manner, with consecutive merging or splitting of clusters, respectively. After the hierarchical algorithm has finished, the final result will be a **dendrogram**, that is a tree that illustrates the arrangement of the clusters produced by the corresponding analyses. To get the clusters, one can **cut the dendrogram** in a specific depth and get the clusters as labels associated with the nodes (the cells in case of scRNA-seq). Another suitable unsupervised clustering algorithm is k-means, which estimates k centroids (centers of the clusters), assigns cells to the nearest centroid, recomputes centroids on the basis of the mean of cells in the centroid clusters, and then reiterates these steps.

Other unsupervised approaches use **graph-based clustering** (such as Seurat[131]), which builds graphs with nodes representing cells and edges indicating similar expression, and then partitions the graphs into interconnected "quasi-cliques" or "communities". These approaches have been some of the most used since they provide consistency and use topological information that will be otherwise lost with other algorithms that prefer a more statistical/inferential approach.

An important consideration to take into account when computing the clusters is that clustering can be done directly on the basis of expression values or more processed data types, such as principal components, reduced dimensions from UMAP or T-SNE or similarity matrices, the latter of which shows improved yield in cluster separation. Cluster stability is measured via resampling methods (e.g., bootstrapping) or on the basis of cell similarities within assigned clusters (e.g., silhouette index), while measurements for dimensionality reduction that use these methods is not available directly aside from PCA(as it will be seen in 3.5.3). It is really important to understand that clustering should be done on **higher dimensions** than the one used for **visualization** (Section 3.5.4) to improve the retention of information from the genes expressed and their relations, that means that clusters visualized in 2D or 3D could be seen as sparse and overlapping since the clustering was done on higher dimensions.

3.5.2 Cell group identification

Cells possess an enormous "landscape" of potential states that they can adopt over the course of development and in disease progression. However, few reliable markers exist for any given cell type, and hidden diversity remains even with well-established markers (e.g., cluster of differentiation (CD) markers in immune cells).

Marker genes that discriminate subpopulations can be identified by differential gene expression analysis of clusters using. Individual genes can be evaluated to serve as binary classifiers for cell identity with, for example, ROC or LRT tests based on the zero-inflated data. Some methods use differential expression analysis methods like the ones seen in 3.4 as a preprocessing and filtering step for cell identification while using these differentially expressed genes in a subpopulation as a guide for understanding and estimating the group type.

For this research and the especially the part about experimentation of this project, both manual annotation (from marker genes found with Seurat[131] *FindMarkers* function for the clusters found using the methodology seen in the previous section 3.5.1) and automatic annotation with SingleR [137] will be done. SingleR does not consider the clusters found in the previous section in its default use but uses internal clustering and hypothesis testing to find groups of cells and identifies these groups from an annotation dataset that contains marker genes for specific cells.

The results of annotation and cell identification right now are not great in general, especially for small groups of cells, so the field is growing rapidly but it has not yet reached the point where methodology is clearly defined other than the manual annotation or the mining of datasets for markers (done as a final step in SingleR or with "manual" annotation).

3.5.3 Dimensionality reduction

To avoid the "the curse of dimensionality," dimensionality reduction is typically performed after read count normalization in scRNA-seq experiments. A common way to visually inspect cellular sub-population structures is to carry out dimensionality reduction (DR) and project cells into a two- or three-dimensional space. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are commonly used approaches for data representation. Diffusion components and uniform manifold approximation and projection (UMAP) are viable alternatives that overcome some limitations of PCA and t-SNE by preserving the global structures and pseudo-temporal ordering of cells, as well as being faster. Even though DR techniques can guide the initial data inspection, more-robust clustering algorithms are needed to define subpopulations among cells.

Principal component analysis (PCA) is a widely used unsupervised linear dimensionality reduction method. By projecting cells into 2D space, samples can be easily visualized with increased interpretability. Additional non-linear dimensionality reduction methods, such as t-distributed stochastic neighbor embedding (t-SNE), multidimensional scaling, locally linear embedding (LLE), UMAP(uniform manifold approximation and projection) and Isomap can also be utilized. The two non-linear methods

(UMAP and t-SNE) are really similar and differ mainly on the starting step of the algorithm(UMAP start computing the embedding from points obtained from a **spectral embedding** while t-SNE starts at random points) and on how the new embeddings for points are moved at every iteration(t-SNE moves all the embeddings at every iteration while UMAP moves only a subset of the embeddings).

For PCA there is the possibility of establishing the quality of the dimensionality reduction since it is, at its core, an algorithm that directly embeds the information retained inside the algorithm itself (by retaining a certain amount of variance with every principal component). For UMAP and t-SNE there is no common way of understanding how much information is retained, the only thing that could be done is a downstream analysis with some ground truth or try to get the original data from the dimensionally reduced data.

t-SNE is implemented in the popular Cell Ranger pipeline (10× Genomics) and in Seurat (http://satijalab.org/seurat/) in the R package while UMAP is used in Seurat as a default visualization tool for the cells. It is necessary and quite obvious to note that dimension reduction may result in the loss of important biological information so further caution is needed.

3.5.4 Visualization

The easiest way to overview the data in a visual way is by transforming it using the principal component analysis(PCA) previously introduced and then visualizing the first two(or three) principal components. Other methods are t-SNE (used by seurat) and UMAP. As already described in the previous section, these methods are non-linear and map the original data into a manifold.

For visualization, the method used should be the same as the one used for clustering since the final clusters will be dependent on the methodology and if the step of clustering uses a different type of dimensionality reduction than visualization(not in the number of dimensions but in the change of method), the resulting visualization will be useless.

3.5.5 Trajectory analysis

Single-cell genomics offers a means of precisely quantifying the state of individual cells and thus may enable the construction of explicit, genome-scale dynamical cellular models. Early single-cell transcriptomic studies lend support to the idea that cells are occupants of a vast, complex landscape of possible states and raise doubts that cell types are precisely defined, discrete entities. Time series experiments of differentiation have observed cells transitioning between a starting state and one or more end states, with many cells distributed along a "trajectory" between them. The Monocle[27] algorithm introduced the notion of pseudotime, a quantitative measure of biological progression through a process such as cell differentiation.

Many biological processes manifest as a continuum of dynamic changes in the cellular state. The most obvious example is that of differentiation into increasingly specialized cell subtypes, but it might also be considered phenomena like the cell cycle or immune cell activation that are accompanied by gradual changes in the cell's transcriptome.

These processes could be characterized from single-cell expression data by identifying a "trajectory", i.e., a path through the high-dimensional expression space that traverses the various cellular states associated with a continuous process like differentiation. In the simplest case, a trajectory will be a simple path from one point to another, but it can also be observed that more complex trajectories can branch to multiple endpoints.

The "pseudotime" is defined as the positioning of cells along the trajectory that quantifies the relative activity or progression of the underlying biological process. For example, the pseudotime for a differentiation trajectory might represent the degree of differentiation from a pluripotent cell to a terminal state where cells with larger pseudotime values are more differentiated. This metric allows us to tackle questions related to the global population structure in a more quantitative manner. The most common application is to fit models to gene expression against the pseudotime to identify the genes responsible for generating the trajectory in the first place, especially around interesting branch events.

The pseudotime is simply a number describing the relative position of a cell in the trajectory, where cells with larger values are consider to be "after" their counterparts with smaller values. Branched trajectories will typically be associated with multiple pseudotimes, one per path through the trajectory; these values are not usually comparable across paths. It is worth noting that "pseudotime" is a rather unfortunate term as it may not have much to do with real-life time. For example, one can imagine a continuum of stress states where cells move in either direction (or not) over time but the pseudotime simply describes the transition from one end of the continuum to the other. In trajectories describing time-dependent processes like differentiation, a cell's pseudotime value may be used as a proxy for its relative age, but only if directionality can be inferred.

The big question is how to identify the trajectory from high-dimensional expression data and map individual cells onto it. A lot of algorithms and methodologies try to tackle the problem of estimating the pseudotime for the cells in a sample. There is also the consideration that a trajectory could not be in the data. One can interpret a continuum of states as a series of closely related (but distinct) subpopulations, or two well-separated clusters as the endpoints of a trajectory with rare intermediates.

This research will not be focusing on trajectory analysis since it is not related directly to the objective of this project. For more information about trajectory analysis see the documentation and the related publications for Monocle [27].

3.6 Applications and impact

Precision medicine aims to improve patient outcomes by tailoring treatment to the unique genomic background of a disease or a treatment for a group of patients (or a single patient) and how patients will respond to it (with **resistance** or **effectiveness**). However, efforts to develop prognostic and drug response biomarkers largely rely on bulk 'omic' data, which fails to capture cellular information such as intratumor heterogeneity (ITH) for tumors and deconvolve signals from normal versus tumor cells. These shortcomings in measuring clinically relevant features are being addressed with

single-cell technologies, which provide a fine-resolution map of the genetic and phenotypic heterogeneity in tumors and their micro-environment, as well as an improved understanding of the patterns of subclonal tumor populations. Single-cell genomics is a rapidly developing field, and current technologies can assay a single cell's gene expression, DNA variation, epigenetic state, and nuclear structure.

Libraries for single-cell RNA sequencing are cell-specific towards investigating cellular functionalities of DNA and RNA in different cellular subsets. The advent of scRNA-seq have revolutionized the field of phenotyping and transcriptomics analysis since the resolution of the results obtained is incomparable to bulk RNA-seq. Advancing towards the next generation in cancer precision medicine will require incorporation of the effects of tumor heterogeneity, clonal evolution, and microenvironment, on drug resistance and patient outcomes and these objectives of integration of different information is only attainable with more resolution of sequence data to characterize transcript at **cellular level**, so scRNa-seq is necessary. Single-cell technologies are enabling separation and characterization of molecular signals from heterogeneous populations of malignant and non-malignant cells. New biomarkers and therapeutic strategies can be developed against resistant subclonal genotypes and phenotypes, determined at a high resolution with single-cell approaches. Novel mechanisms of drug resistance are being discovered by combining single-cell data with forward genetics and large-scale pharmacological screens. Single-cell strategies are being adopted in the clinic to predict response to immune therapy, targeted drugs, and the study of tumor evolution using non-invasive techniques.

The rapid expansion of microfluidic technology in recent years has transformed the research capabilities of both basic scientists and clinicians. Applications of this technology include long-term analysis of single bacterial cells in a microfluidic bioreactor and the quantification of single-cell gene expression profiles in a highly parallel manner.

With the development of advanced RNA-seq techniques to better understand the transcriptome of tissues and individual cells, RNA-seq has led to the discoveries of various non-coding RNAs, including long non-coding RNAs and circular RNAs and their role in regulating different genes for different functionalities. However, recent advances in single-cell RNA-seq (scRNA-seq) have led to discoveries of phenotypically diverse and complicated networks of cells (aka, the "**cellulome**") within cardiac tissue at the single-cell level. Usage of scRNA-seq has opened a new field of single-cell level precision to diagnostics and therapeutics to combat cardiovascular disease. Therefore, research into the transcriptomic alterations within the diverse "cellulome" of the tissue and pathology under research has tremendous translational potential for the field of personalized medicine and specific treatment/response.

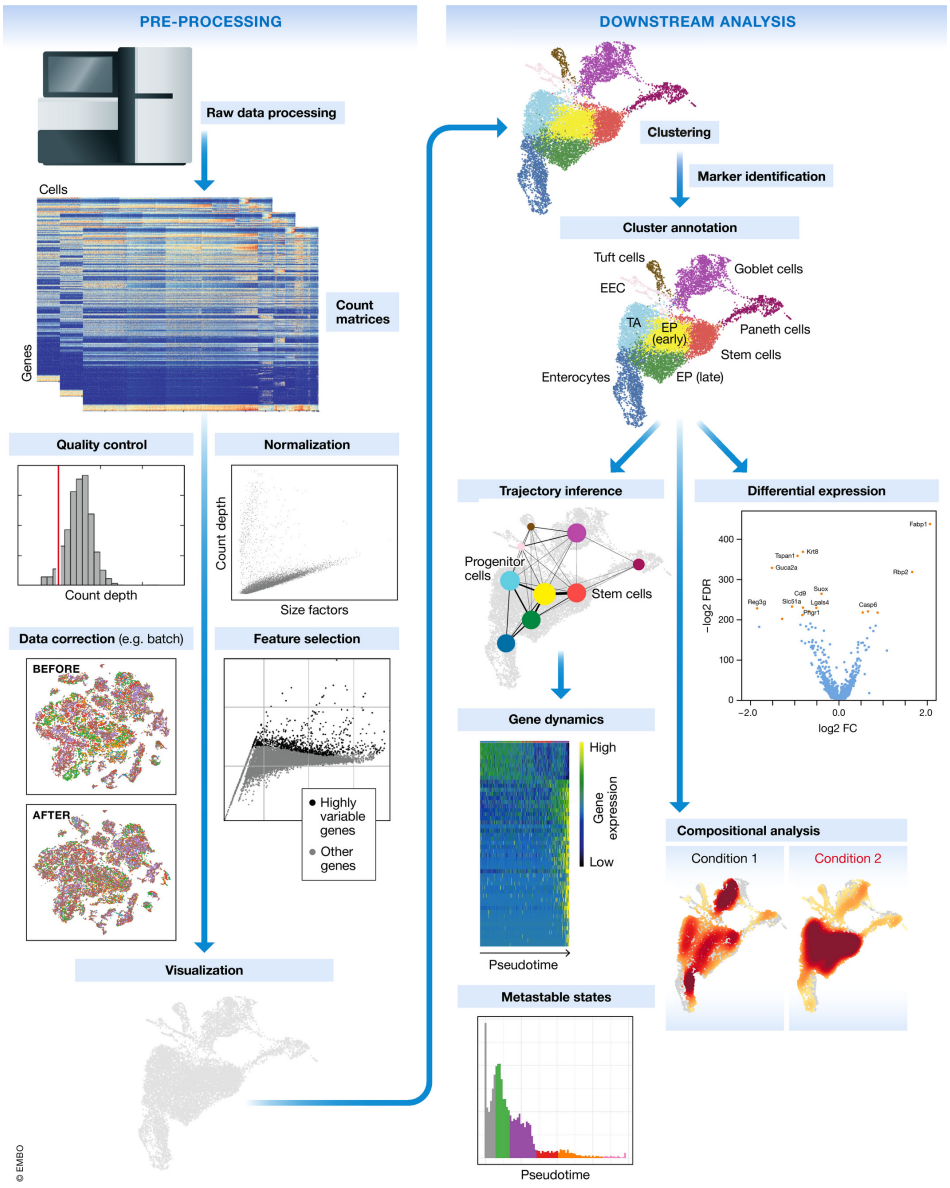


Figure 3.12: General workflow of scRNA-seq

Statistical and computational methods are central to single-cell genomics and allows for extraction of meaningful information. The translational application of single-cell sequencing in precision cancer therapy has the potential to improve cancer diagnostics, prognostics, targeted therapy, early detection, and noninvasive monitoring. Single-cell technologies should be used in collaboration with other tools since it was seen that isolating the cells will alter the cells, but the applications of single-cell methodologies with other sources can be the forefront of personalized medicine.

A visualization that summarises some of the methods, tools and protocols that were seen during this chapter can be seen in figure 3.12.

3.7 Challenges and considerations

Although experimental methods for scRNA-seq are increasingly accessible to many laboratories, computational pipelines for handling raw data files remain limited. Some commercial companies provide software tools, such as 10× Genomics and Fluidigm, but this area remains in its infancy, and gold-standard tools have yet to be developed. In the sections below, current bioinformatics tools available for the analysis of scRNA-seq data will be discussed.

During library preparation for scRNA-seq, a pool-specific index can be introduced that allows the multiplexed sequencing of multiple experiments. Full-length methods introduce the cell-specific barcodes only after fragmentation, thus **impeding pooled processing of cells at earlier stages of the protocol**. Apart from STRT-seq, scRNA-seq libraries require paired-end sequencing, in which one read provides information about the transcripts while the other reads the single-cell barcodes and UMI sequences. STRT-seq incorporates the cell barcode and UMI at the 5'-transcript end, which allows cell, molecule and transcript information to be captured in a single read, as no poly(T) stretch separates the respective sequences.

To address for full-length protocols limitations like no UMI no possibility of having high-throughput, new protocols have been under development and one of them is **Smart-seq3** already introduced in 3.5

Another big challenge in single-cell RNA-seq is that labs and researchers in general have a very low amount of starting material observed per cell. This results in very sparse data, where most of the genes remain undetected and so our data contains many zeros. These may either be due to the gene not being expressed in the cell (a "real" zero) or the gene was expressed but it is very difficult, if not impossible, to be able to detect it (a "dropout" or sparsity event). This leads to cell-cell variation that is not always biological but rather due to technical issues caused by uneven PCR amplification across cells and gene "dropouts" (where a gene is detected in one cell but absent from another). Improving the transcript capture efficiency and reducing the amplification bias are solutions for these problems and still active areas of technical research. However, as it shall be seen in this project, it is possible to alleviate some of these issues through proper data normalisation and imputation, even though the methods that will be seen are not perfect and introduce false information(imputation) or fail to normalize data across different samples and cells(normalization).

Another important aspect to take into account and control are batch effects. Batch effects are technical artifacts that are added to the samples during handling. For example, if two sets of samples were prepared in different labs or even on different days in the same lab, then it might be observed that greater similarities between the samples that were handled together. In the worst-case scenario, batch effects may be mistaken for true biological variation. These can be observed even when sequencing the same material using different technologies (figure 3.13), and if not properly normalized, can lead to incorrect conclusions. The processing of samples should also be done in a manner that avoids confounding between experimentally controlled variables (such as a treatment, a genotype or a disease state) and the time when the samples are prepared and sequenced. For example, if planning an experiment to compare healthy and diseased

tissues from 10 patients each, if only 10 samples can be processed per day, it is best to do 5 healthy + 5 diseased together each day, rather than prepare all healthy samples one day and all diseased samples in another (figure 3.13). Another consideration is to ensure that there is the replication of tissue samples. For example, when collecting tissue from an organ, it may be a good idea to take multiple samples from different parts of the organ. Or consider the time of day when samples/replicates are collected (due to possible circadian changes in gene expression). In summary, all the common best practices in experimental design should be taken into account when performing scRNA-seq.

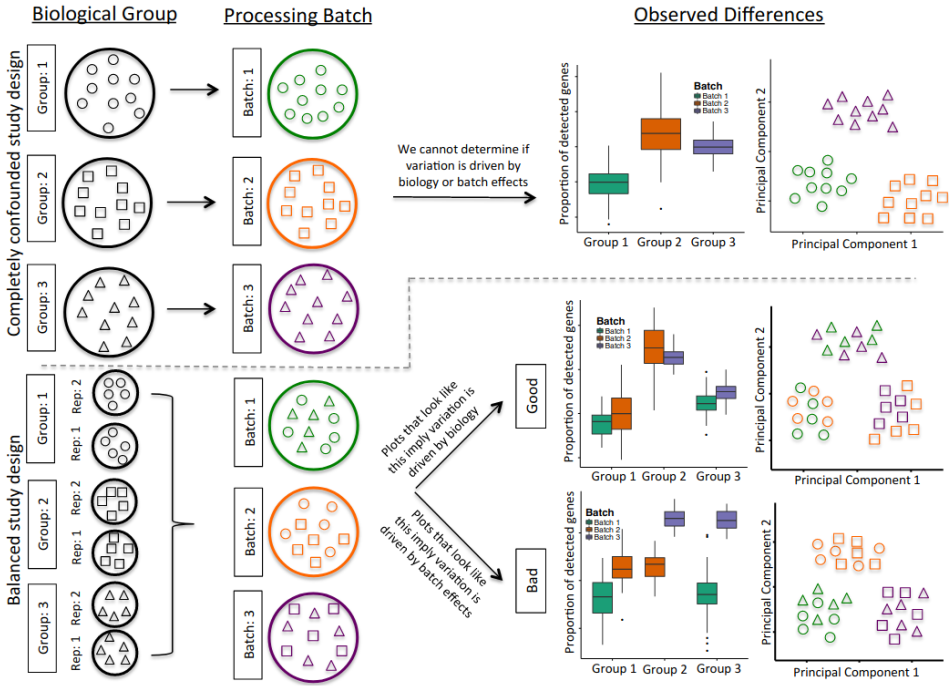


Figure 3.13: **Illustration of a confounded (top panels) and balanced (bottom panels) designs**. Shapes denote different sample types (e.g. tissues or patients) and colours processing batches. Source [55]

There is also another problem introduced previously to consider during data analysis and quality control of sequencing data produced by bead-based and nanowell-based micro-fluidics, and that is **cell doublets**. It is an intrinsic problem for most microfluidics-based methods in that two cells can be captured per reaction site (nanowell or droplet), both receiving identical barcodes. Doublet rates can be experimentally determined in species- mixture experiments, but otherwise can only be estimated. They occur when cells are positioned randomly in reaction sites by limiting dilution and can be controlled by the cell suspension concentration. The relationship between cell loading and doublet rate was systematically quantified for the Chromium system. The doublet rate decreases at higher dilutions, with a resulting increase in reagent costs per cell, as fewer total cells are captured per experiment. Researchers can partially overcome this handicap by jointly capturing samples from different individuals, such that genotype differences allow the user to distinguish between donors and thereby re-

liably identify doublets. Specifically, single-nucleotide polymorphisms identified from the RNA sequencing reads are used to determine the donor origin of the cells and to discriminate samples that were processed in a single batch. However, such a workflow is practicable only when the experimental design includes different human individuals or model organisms with distinct genetic backgrounds. Currently, there is no computational method for credibly identifying doublets, so doublet rates must be minimized by experimental design. Doublets can have dramatic consequences for data interpretation, as artifactual mixed transcriptomes can easily be mistaken for intermediate cell states in dynamic systems.

As mentioned during the 3.4 section, methods for pseudo-bulk use aggregation to compact the information of genes within a sample of cells and transform it into smaller and aggregated sample replicates that are proven to reduce false negatives for classic DE analysis. The creation of the pseudo-bulk matrix is done with some aggregation function that summarizes the cells into a smaller amount of samples (like bulk, in fact, cells gene expression vectors are usually summed to create the sample and, logically, that is supposed to be a bulk sample that has that type of cells in it). A future research challenge could be to find a methodology to build these pseudo-bulk matrices by using artificial intelligence and also using bulk data as well to get better results.

One of the most important limitations of Single cell RNA sequencing (scRNA-seq) is that cells are isolated and cannot influence each other as they usually do when they are in their original tissue, so cell signalling and interaction are completely cut off from all the methods seen in this chapter since after dilution and separation, the cells find themselves completely alone and incapable of transmitting and receiving stimuli from outside. This limitation consequences are devastating to the originality and the fairness of the expression data reported since the cells will maintain their usual markers that make them differentiated but, at the same time, will lose almost completely their network and interaction capabilities and activities that will be not reported in the sequencing. Bulk RNA-seq also has this problem but it is much less evident since the tissue is preserved as a whole and the majority of the cells can interact with each other like in the organism of origin, so the final results obtained will be quite an accurate representation of the average expression of the whole tissue activities. More about the network activities of cells/organisms in a tissue or environment will be seen in chapter 5 where biological networks will be seen and explained in their importance.

As a consequence of meager RNA capture rate, low starting materials, and challenging experimental protocols, the scRNA-seq faces computational and analytical challenges. The noise and sparsity due to the technical (dropout events) and biological factors make the downstream analysis of scRNA-seq data a complicated and demanding task in terms of resources. Additionally, the rapidity in the development of new and exciting experimental methods for scRNA-seq is paving the way for a large accumulation and flow of data that will result in a continuous stream of heterogeneous data that need to be treated in different ways that share the steps seen during this thesis at the start of this chapter. This large agglomeration of data is nothing but the genomic face of "big data." These two challenges together give rise to a new paradigm of Big Single-Cell Data Science. Although a plethora of algorithms and computational tools have already been developed, it is essential to address these challenges collectively and

produce a robust, accurate, parallel, and scalable framework.

Single-cell analysis should not be seen as the only thing to consider when doing personalized analysis, but it is an instruments like all the ones that were presented and that will be presented in the next chapters along the introduced -omics sources that could deepen the understanding of how tissues work by modeling them, and estimate the response to disease and drugs with even more accuracy(how much the estimated response is similar to the real response) and precision (if different estimated responses are similar to each other when they simulate the same system).

Chapter 4

Combining and comparing bulk RNA-seq and scRNA-seq

The main difference between bulk and single-cell RNA-seq is that each sequencing library represents a single cell, instead of a population of cells. Therefore, there is no way to have "biological replicates" at a single-cell level: each cell is unique and impossible to replicate. Instead, cells can be clustered by their similarity, and comparisons can then be done across groups of similar cells (as it shall be presented here and was introduced during the whole previous chapter).

RNA-seq allows profiling of the transcripts in a sample in an efficient and cost-effective way. It was a major breakthrough in the last two decades and has become ever more popular since largely replacing other transcriptome-profiling technologies such as microarrays. Part of its success is due to the fact that RNA-seq allows for an unbiased sampling of all transcripts in a sample, rather than being limited to a pre-determined set of transcripts (as in microarrays or RT-qPCR).

Typically, RNA-seq has been used in samples composed of a mixture of cells, referred to as bulk RNA-seq, and has many applications. For example, it can be used to characterize expression signatures between tissues in healthy/diseased, wild-type/mutant, or control/treated samples. Or in evolutionary studies, using comparative transcriptomics of tissue samples across different species. Besides its use in transcript quantification, it can also be used to find and annotate new genes, gene isoforms, and other transcripts, both in model and non-model organisms.

However, with bulk RNA-seq only estimates of the average expression level for each gene across a population of cells could be estimated, without regard for the heterogeneity in gene expression across individual cells of that sample. Therefore, it is insufficient for studying heterogeneous systems, e.g. early development studies or complex tissues such as the brain.

To overcome this limitation, new protocols were developed that allow applying RNA-seq at single-cell level (scRNA-seq), with its first publication in 2009 [149]. This technology became more popular around 2014 when new protocols and lower sequencing costs made it more accessible. Unlike with the bulk approach, scRNA-seq common workflows can estimate a distribution of expression levels for each gene across a population of cells.

Compared to the traditional technique of bulk-RNA-seq, the main improvement that scRNA-seq achieves is that while bulk-RNA-seq averages gene expression across all cells in a sample, scRNA-seq profiles the transcriptome of each individual cell in the tissue sample. Significantly, this means that scRNA-seq makes high throughput investigations

of tissue samples far more specific by visualizing the phenotypes at single-cell resolution.

This allows us to answer new biological questions where cell-specific changes in the transcriptome are important. For example discovering new or rare cell types, identifying differential cell composition between healthy/diseased tissues, or understanding cell differentiation during development. One of the most iconic uses of this technology is in building gene atlases for cell types, which provide a comprehensive compendium of the cell diversity in organisms, with many applications in health as well as fundamental research.

	Goal	Protocol	Quality control	Normalization	Analyses
bulk RNA-seq	• Measure the average gene expression across the population of cells in a sample • To identify differences between sample conditions	• RNA is extracted from all cells in the sample • Reverse transcription converts RNA to cDNA, facilitates ligation of sequencing adaptors • Amplification	• GC content, presence of adaptors, over-represented k-mers, duplicated reads • Percentage of reads that map to reference • Reproducibility between replicates	• Batch effect • Between-sample variability: sequencing depth Quantile normalization, spike-ins • Within-sample variability: feature length, library size effects RPKM, FPKM, TPM	• Estimate gene and transcript expression • Differential expression analysis • Alternative splicing
scRNA-seq	• Measure the gene expression of individual cells in a sample • To identify differences between cell types/states • To build cell-specific transcriptomes and characterization	• RNA is extracted from isolated cells, labeled with cell-specific identifier • UMIs, spike-ins often included, to account for higher levels of noise • Reverse transcription, amplification similar to bulk protocol	• Reads, number of genes per cell • Percentage of reads that map to spike-ins (if used), percentage of reads that map to mitochondria • QC metrics used in bulk RNA-seq are also examined	• Batch effect and within-sample variability are corrected for similarly to bulk RNA-seq • Between-sample variability methods must additionally account for capture efficiency and dropout sources of noise	• Dimensionality reduction • Identify cell subpopulations • Differential expression • Pseudotime/trajectory analysis

Table 4.1: Summarization of sequencing properties for bulk and single-cell RNA-seq

A summarization of the differences between bulk sequencing and single-cell sequencing can be seen in table 4.1

Bulk RNA-seq should be used in **homogeneous tissue** while scRNA-seq should be used for **heterogeneous tissue** where the cells network is complex and the interactions between the cells are very different and cannot be explained by a summarized value like it is done in bulk RNA-seq. On the other hand, there is the fact/problem of the isolation of the cells that take away the network interaction of the cells as introduced at the end of section 3.7, so bulk is advised also for heterogeneous tissue since it maintains the interactions of the cells and the sequencing starts as soon as possible starting from tissue collection (as seen in section 2.1) and maintaining the expression of the whole tissue accurate to the moment where the tissue was sampled, while the process of true sequencing of single-cell will start only after the isolation of the cells, and that will surely change the expression values obtained from the whole process.

4.1 Combining bulk and single-cell

The combination of single-cell analysis and bulk analysis can be thought as a type of ensemble method that uses **multi-omics** even though the two methods result in the same kind of data for two different but related settings:

- Bulk workflows result in a snapshot of the activity of the whole tissue, it does not make any distinction of the component of the tissue and treats it as a single component with its functions
- Single-cell workflows result in a snapshot of the activity of cells in the tissue, it makes a distinction between the components of the tissue and treats every cell as a single component with its own functions that contributes to the overall activity of the whole tissue.

By combining the two methodologies, the resulting analysis is a lot more complete and overcomes some of the limitations seen during this research. For example, the accuracy of the cells found is obtained via scRNA-seq and validated by deconvolution and bulk RNA-seq while the expression values obtained can be seen as a machine learning model built on top of both bulk data and single-cell data to capture the most accurate description as a snapshot of the original sample without losing the interactions and the functionalities of the cells and the whole tissue as a whole.

There exists a lot of new methods that use bulk RNA-seq along scRNA-seq, especially in the last year since the two methodologies are done at the same time to obtain additional information about the tissue that is being researched. In this research, some of them will be seen in broad terms.

4.1.1 Subclonal trees estimation

Many cancers have substantial genomic heterogeneity within a given tumor, and to fully understand that diversity requires a deep and accurate analysis of single cells in a tissue. Understanding the clonal architecture and evolutionary history of a tumor poses one of the key challenges of biology and tumor evolution and overcoming treatment failure due to resistant cell populations. Before the advent of single-cell analysis, studies on subclonal tumor evolution have been primarily based on bulk sequencing, in recent years single-cell was used for studying the composition of tumors but nothing substantial was done in the field of subclonal tree estimation. In this section, a tool that uses both bulk RNA-seq data and scRNA-seq data will be used, this tool is called B-SCITE(Bulk-Single Cell Inference of Tumor Evolution) [100]. B-SCITE is a tool that follows the same route as its predecessor SCITE [70] where a stochastic search algorithm is used to identify the evolutionary history of a tumor from noisy and incomplete mutation profiles of single cells but integrates bulk sequencing data to infer tumor phylogenies from combined single-cell and bulk sequencing data.

Cancer is a genetic disease that develops through a branched evolutionary process, resulting in uncontrollable growth and spread to other parts of the body and it is caused by changes to genes that control the way our cells function, especially how they grow and divide that occurs for various reasons, some of most common of them are:

- errors that occur as cells divide.

- damage to DNA caused by harmful substances in the environment, such as the chemicals in tobacco smoke and ultraviolet rays from the sun.
- inheritance from parent organisms.

The body normally eliminates cells with damaged DNA before they turn cancerous, but the body's ability to do so goes down with age and there is also a probabilistic chance that errors in the DNA are not repaired and will remain in the sequence. Cancer(tumor and mutations in general) is characterized by the emergence of genetically distinct subclones through the random acquisition of mutations at the level of single cells and shifting prevalences at the subclone level through selective advantages purveyed by driver mutations. This interplay creates complex mixtures of tumor-cell populations, which exhibit different susceptibilities to targeted cancer therapies and are suspected to be the cause of treatment failure. Therefore, it is of great interest to obtain a better understanding of the evolutionary histories of individual tumors and their subclonal composition.

By only using bulk tumor samples the type of information obtained are indirect measurements of the subclonal tumor composition in the form of aggregate total and variant read counts measured across hundreds of thousands or millions of cells. The methods that use bulk data only rely on estimation of the frequencies of SNV/SNP and deconvolution to understand the relationships of the tumor cell population. However, the underlying statistical problem is underdetermined, and single-nucleotide variants with similar VAF are automatically clustered into a single subclone. This inevitably leads to incorrect phylogenies for tumors with multiple distinct subclones of similar prevalences (figure 4.1). The aggregate sequencing data additionally pose a limitation to the achievable tree resolution, as mutational signals of smaller subclones cannot be distinguished from noise and therefore not be reliably represented in the tree. Sequencing multiple samples from the same tumor and increasing the coverage can to some extent mitigate these issues but is not always practicable.

Another solution is the use of single-cell sequencing data which provides mutation profiles of individual cells, such that the phylogeny can be directly inferred without any form of deconvolution. The main challenges here instead are the high levels of noise found (primarily introduced during DNA amplification and a necessary step to obtain sufficient DNA material for deep-sequencing and having enough data to work with) and the other problems seen during the previous chapter 3. False negatives are the most prevalent error type due to allelic sparsity, but also false positives occur when an error is introduced early in the amplification that results in some strands being excessively amplified (while other strands are not considered during sequencing and these will result in dropouts), so the final results are not clear at all. Further noise stems from doublet-mutation profiles (seen for the scRNA-seq protocols that use microfluidics as an isolation steps, the problem of doublet-detection was also introduced in 3.7). Classic approaches for phylogeny reconstruction are not suitable for dealing with these single cell-specific noise profiles, and a number of probabilistic approaches have been developed to specifically account for the error types found in data, as introduced in the previous chapter, but none of them is yet in the maturity stage since the field is quickly expanding and some methods are inherently prone to technical errors that are impossible to recover by using the protocol in question.

A major difference between the evolutionary histories of tumors inferred from bulk and SCS data is that the former typically are clonal trees where mutations with similar frequencies are clustered together and represented in a single tree node, while trees derived from Single-cell data are fully resolved trees that can be either cell lineage trees, binary trees where the cells form the leaves and mutations occur along tree branches or mutation trees that depict the partial temporal order in which mutations were acquired. For cell lineage trees, a heuristic has been proposed for clustering cells into clones in a post-processing step, which results in trees that are closer to bulk clonal trees.

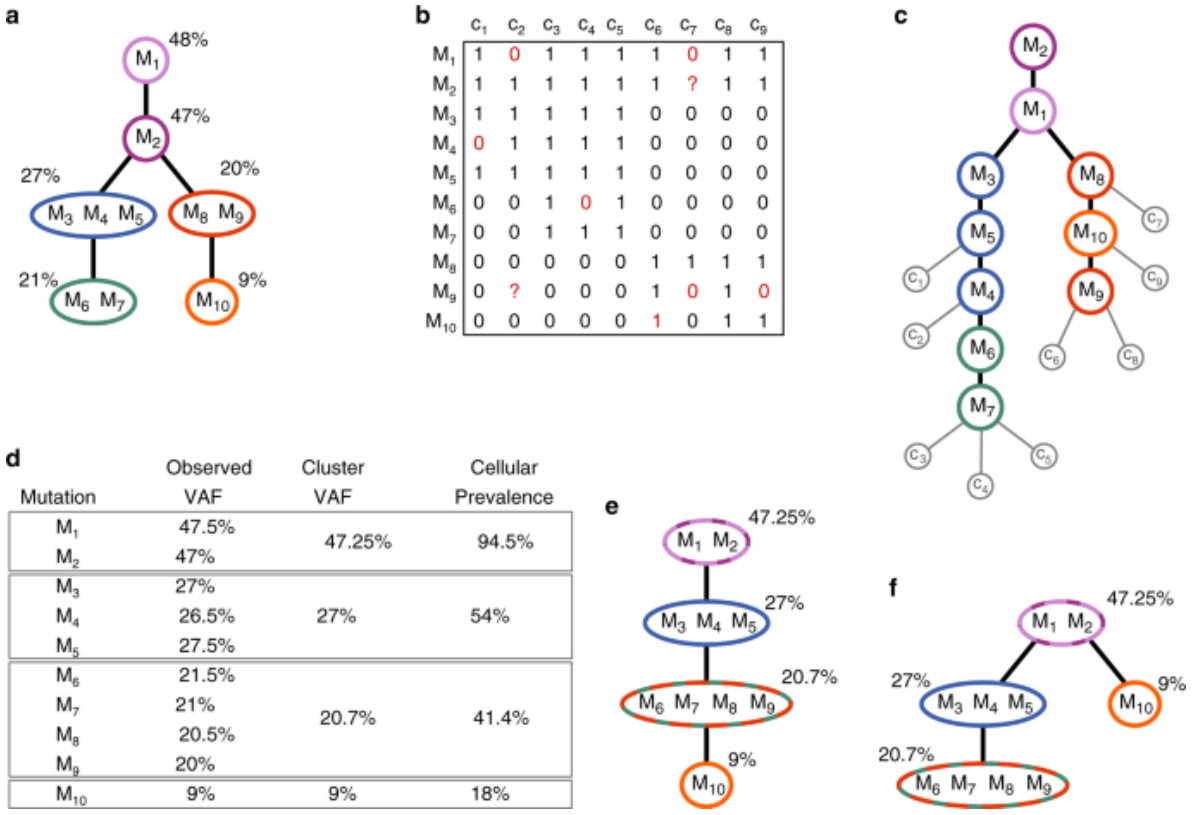


Figure 4.1: **Comparison of inferred mutation histories based on single-cell and bulk-sequencing data**. Source [100]

In figure 4.1 there is the comparison of the results of inferred clonal trees from bulk data and from single-cell data.

a) Ground truth clonal tree with mutations M_i: the colored nodes represent the subclones, and the tree structure indicates the partial temporal order in which the subclones emerged (from top to bottom). Each subclone contains the mutations it acquired in comparison with its parent and is annotated with the mean VAF of these mutations. (For a heterozygous mutation in a copy-number-neutral region, the VAF is half the mutation's cellular prevalence)

b) Cell mutation profiles obtained from the single-cell sequencing data for nine cells c_i: "1" indicates the observed presence of a mutation, and "0" absence. A "?" indicates a missing data point (NA), e.g., due to insufficient coverage. The red "0"s are false negatives (e.g., sparsity events that introduce dropouts), the red

"1" indicates a false positive(e.g., sparsity events that introduce overexpression of genes). Due to these errors, the mutation matrix defines no perfect phylogeny.

c) Inferred single-cell mutation tree annotated with single-cell placements. Not all cells can be placed, such that their observed mutation profile matches with the mutations acquired along the lineage from the root to their attachment point. The branching point of the ground truth tree is inferred correctly, due to the strong signal that the red/orange and blue/green mutations do not occur in the same cell. However, mutation order in linear segments is not reliably inferred from the data; especially in the right branch, a mutation with low prevalence (M10) is placed above a more prevalent mutation (M9) due to errors in the mutation profiles of cells c6, c7 and c9.

d) Variant allele frequencies were obtained from bulk sequencing. VAF-based clustering of mutations leads to the merging of subclones.

e), f) Both clonal trees are compatible with the VAFs and the clustering inferred in (d). Due to the clustering and incompatible VAFs, the correct branching between the blue and red subclone is not inferred, but in both trees, mutation ordering is consistent with their true prevalences

As the strengths and weaknesses of single-cell and bulk-sequencing data are to a large extent complementary with respect to phylogeny inference, using both data types for a joint inference should improve the overall results for subclonal tumor evolution and tree estimation. B-SCITE uses a probabilistic approach for the inference of tumor mutation histories by the use of SNV data obtained from single-cell and bulk DNA sequencing (figure 4.2).

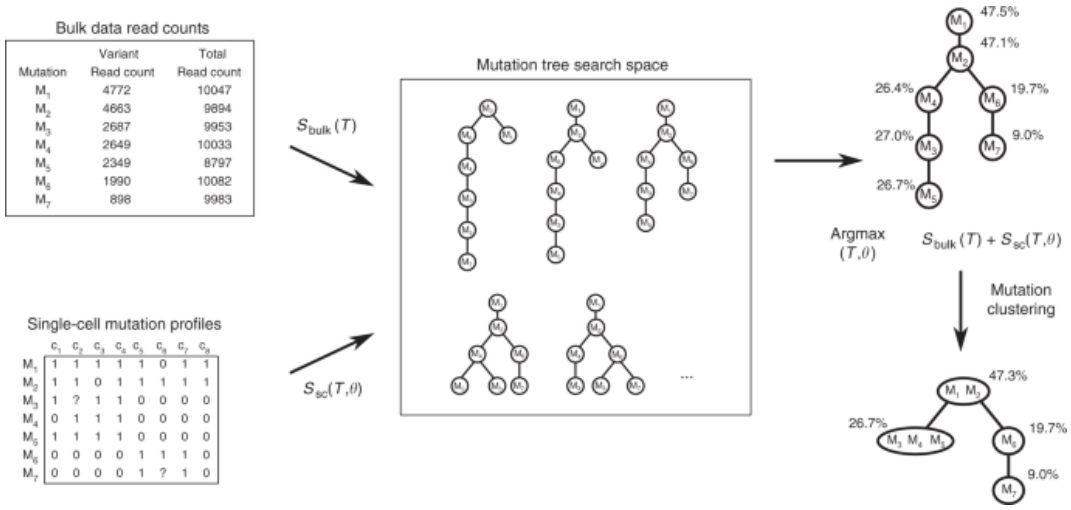


Figure 4.2: **B-SCITE workflow**: B-SCITE uses a Markov chain Monte Carlo approach to search the space of candidate mutation histories. Each candidate tree is scored based on its joint fit to the single-cell and bulk DNA sequencing data. The bulk data consist of a high-coverage variant and total read counts for the mutated loci. The single-cell data consist of the observed mutation profiles of the sequenced cells. These single-cell profiles are characterized by high noise rates θ that are learned from the single-cell sequencing data along with the tree topology T. B-SCITE reports the tree with the highest joint score. This tree is a fully resolved mutation tree. To obtain a clonal tree, the linear tree parts can be clustered based on the variant allele frequencies of the bulk data

More information about B-SCITE, the full methodology, the algorithm and its performance in the original publication [[100],[70]] and [41]. Also see more recent studies [105] and the tools related like TriSiCell (https://trisicell.readthedocs.io/).

4.1.2 Cell composition

A tool for estimating cell type proportions in bulk expression that uses scRNA-seq is **Bisque** [72]. Bisque implements a regression-based approach that utilizes scRNA-seq or snRNA-seq data to generate a reference expression profile and learn gene-specific bulk expression transformations to robustly decompose RNA-seq data. These transformations significantly improve decomposition performance compared to other methodologies when there is significant technical variation in the generation of the reference profile and observed bulk expression. Importantly, compared to existing methods, our approach is extremely efficient, making it suitable for the analysis of large genomic datasets that are becoming ubiquitous. When applied to various tissue data, Bisque manages to replicate previously reported associations between cell type proportions and measured phenotypes across abundant and rare cell types.

Traditional methods for determining cell-type composition, such as immunohistochemistry or flow cytometry, rely on a limited set of molecular markers and lack scalability relative to the current rate of data generation. Single-cell technologies provide a high-resolution view of and into cellular heterogeneity and cell-type-specific expression. However, these experiments remain costly and noisy compared to bulk RNA-seq. The collection of bulk expression data remains an attractive approach for identifying population-level associations, such as differential expression regardless of cell-type specificity. Moreover, many bulk RNA-seq studies that have been performed in recent years resulted in a large body of data that is available in public databases such as dbGAP and GEO. Given the wide availability of these bulk data, the estimation of cell-type proportions often termed decomposition, can be used to extract large-scale cell-type-specific information.

There exist a number of methods for decomposing bulk expression, many of which are regression-based and leverage cell-type-specific expression data as a reference profile. The distinct nature of the technologies used to generate bulk and single-cell sequencing data may present an issue for decomposition models that assume a directly proportional relationship between the single-cell-based reference and observed bulk mixture. For example, the capture of mRNA and the chemistry of library preparation can differ significantly between bulk tissue and single-cell RNA-seq methods, as well as between different single-cell technologies. Moreover, some technologies may be measuring different parts of the transcriptome, such as nuclear pre-mRNA in snRNA-seq experiments as opposed to cellular and extra-cellular mRNA observed in traditional bulk RNA-seq experiments. As it was introduced and shown, these differences may introduce gene-specific biases that break down the correlation between cell-type-specific and bulk tissue measurements. Thus, while single-cell RNA-seq technologies have provided unprecedented resolution in identifying expression profiles of cell types in heterogeneous tissues, these profiles generally may not follow the direct proportionality assumptions of regression-based methods.

Bisque is an efficient tool to measure cellular heterogeneity in bulk expression through robust integration of single-cell information, accounting for biases introduced in the single-cell sequencing protocols. The goal of Bisque is to integrate the different technologies of single-cell and bulk RNA-seq to estimate cell-type proportions from tissue-level gene expression measurements across a larger set of samples. Our reference-based model decomposes bulk samples using a single-cell-based reference profile and, while not required, can leverage single-cell and bulk measurements for the same samples for further improved decomposition accuracy. This approach employs gene-specific transformations of bulk expression to account for biases in sequencing technologies as described above. When a reference profile is not available, the authors propose a semi-supervised model that extracts trends in cellular composition from normalized bulk expression samples using only cell-specific marker genes that could be obtained using single-cell data.

Bisque reference-based decomposition model requires bulk RNA-seq counts data and a reference dataset with read counts from single-cell RNA-seq. This reference-based model is based on the assumption that cell populations are equally represented in single-cell and bulk RNA sequencing of the same tissue samples. In addition, the single-cell data should be labeled with cell types to be quantified. A reference profile is generated by averaging read count abundances within each cell type in the single-cell data. Given the reference profile and cell proportions observed in the single-cell data, our method learns gene-specific transformations of the bulk data to account for technical biases between the sequencing technologies. Bisque can then estimate cell proportions from the bulk RNA-seq data using the reference and the transformed bulk expression data using non-negative least-squares (NNLS) regression.

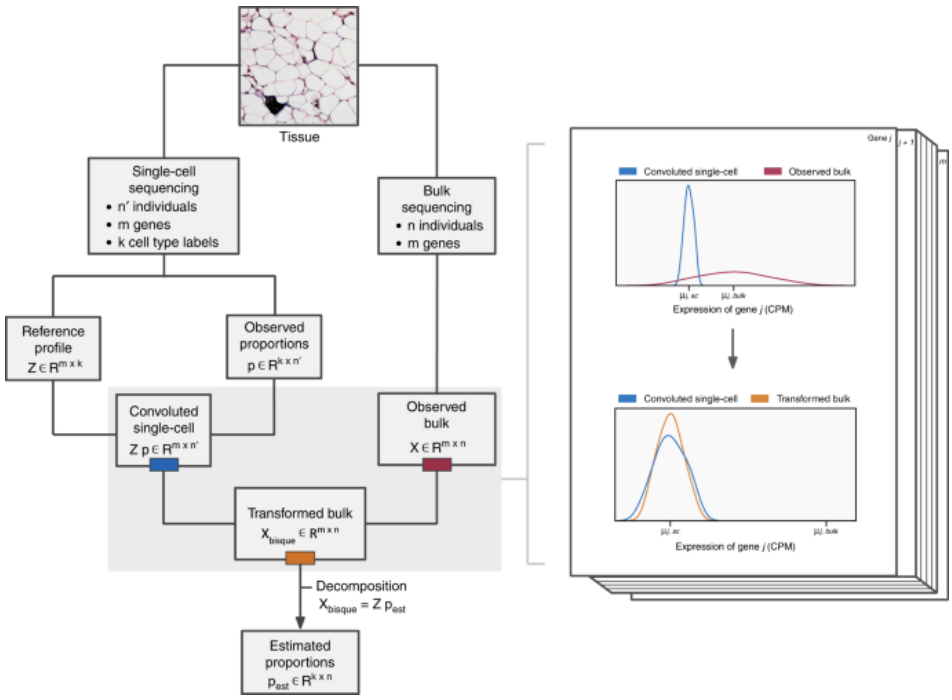


Figure 4.3: **Bisque methodology overview**: Source [72]

A graphical overview of Bisque is presented in Figure 4.3

4.1.3 Gene imputation

The SCRABBLE[117] project is an algorithm for imputing scRNA-seq data by using bulk RNA-seq as a constraint. SCRABBLE only requires consistent cell population between single-cell and bulk data. The bulk data represent the unfractionated composite mixture of all cell types without sorting them into individual types. For many scRNA-seq data, there are usually existing bulk data on the same cell/tissue. And it is becoming increasingly common to collect matched bulk data when a new scRNA-seq experiment is performed. Bulk RNA-seq data allows SCRABBLE to achieve a more accurate estimate of the gene expression distributions across cells than using single-cell data alone. SCRABBLE is based on the framework of matrix regularization that does not impose an assumption of specific statistical distributions for gene expression levels and dropout probabilities. It also does not force the imputation of genes that are not affected by dropout events.

SCRABBLE is based on the mathematical framework of matrix regularization [20]. It imputes dropout data by optimizing an objective function that consists of three terms (Fig. 1). The first term ensures that imputed values for genes with non-zero expression remain as close to their original values as possible, thus minimizing unwanted bias towards expressed genes. The second term ensures the rank of the imputed data matrix is as small as possible. The rationale is that it is only expected that a limited number of distinct cell types is present in a given tissue sample. The third term operates on the bulk RNA-seq data. It ensures consistency between the average gene expression of the aggregated imputed data and the average gene expression of the bulk RNA-seq data. SCRABBLE is composed of a convex optimization algorithm to minimize the objective function that will not be seen here but is available in the original documentation[117]. The existence of an optimal solution is guaranteed mathematically [20].

A major application of scRNA-seq is to better understand the gene-gene and cell-cell relationships in complex tissue and thus, a good imputation method should preserve the data structure that reflects the true gene-gene and cell-cell relationships. This attention to cell-cell and gene-gene interactions and mutual regulation was already introduced during the introduction of the GraphSCI[124] framework used also for Imputation with the use of Graph neural networks. SCRABBLE pays particular attention to gene-gene and cell-cell interaction by computing correlation matrices for gene-gene and cell-cell using the data simulated using down-sampled real bulk RNA-seq dataset as a simulation strategy, where the algorithm introduced dropout events using an exponential function to control dropout rate (parameter λ) and a Bernoulli process to introduce dropout events at the corresponding dropout rate. Using Pearson correlation, the algorithm then determines the similarity between the correlation matrices based on true data and dropout/imputed data. The algorithm was also tested for **pathways**, especially for gene-gene and cell-cell relationships. For more information about the algorithm and how it compares to other tools see the original published article about SCRABBLE [117].

4.1.4 Honorable mentions

A methodology and research to take into account as one of the most exciting for medicine especially is the estimation of **inflammatory cell states in rheumatoid arthritis joint synovial tissues**[167]. This research presents complex analysis and a diverse collection of methodologies that use a combination of bulk and single-cell along with other types of information obtained from flow cytometry, mass cytometry, target sequencing, longitudinal collections of data and many other things. This research is really interesting since it provides a clear description of what is done both from a lab perspective and a digital sight, by providing documentation for every step. This research, and the data that accompanies it in particular, will be used for future research and to provide a better methodology and results for this research, since the data that is being used right now is from [41] and it is not the best, but for a matter of time, it is the only available data during this project. More about the data and the pipeline definition will be seen in 7 and 8.

To conclude this chapter, and to remark a point that is needed and will be revisited in almost every chapter, combining and comparing different sources of data and information is necessary to model and estimate accurate results. For this reason, when multiomics data is available, it should be always considered together by combining the different sources of information with a model that can simulate the whole system by building the small components of it by using these different sources [147].

Chapter 5

Biological network overview

Networks are one of the most common and powerful ways to represent systems as complex sets of binary interactions or relations between different entities. In the context of biology, biological systems are represented by their entities (the nodes in the graph) and their interactions (the relations from one node to another node). These interactions are not only related to the type of relationship defined on the entities, but could also contain information about the whole graph state and other graphs (e.g. in temporal networks where a link is also a temporal interaction that defines the state of the graph on that time).

Biological networks are models that use a combination of various subjects to reach a good estimation of a complex system. Recent complex systems research has also suggested some far-reaching commonality in the organization of information in problems from biology, computer science, and physics. Problems especially in computer science and mathematics are applicable to biological complex systems to obtain a better representation capable of simulation of the system as a whole and its parts. Integrating the topological information in a definition of a model used, for example, for estimating the usefulness of a drug when released to the public can be a

Topological information is very important for all complex systems composed of more entities that are related and interact with each other. Networks are widely used in many branches of biology as a convenient representation of patterns of interaction between appropriate biological elements. These biological networks include:

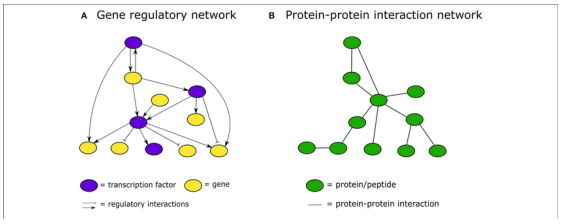
- **biochemical networks**: Biochemical networks represent the molecular-level patterns of interaction and mechanisms of control in the biological cell. The principal types of these networks are metabolic networks, protein-protein networks, genetic regulatory networks and signal transduction networks (transmissions of signals within the cell and outside of the cell with cell-cell interactions).
- **neural networks**: or neuronal network to not be confused with the Artificial Intelligence inference networks based on the perceptron model. Neuronal networks assemble the cellular components needed for sensory, motor and cognitive functions. Nowadays, technical advances such as multielectrode arrays and optical imaging techniques have rendered population recordings fairly common, opening the way for a more refined understanding of neuronal networks. The activity of populations of neurons is derived based on the intrinsic properties of "simplified" versions of neurons (with little to no chemical signals from neurotransmitters released by the neurons) and of their connectivity pattern. Neurons in the brain are deeply connected with one another in complex networks, these are present in the

structural and functional aspects of the brain. For example, small-world network properties have been demonstrated in connections between cortical regions of the primate brain or during basic and common actions for humans, like swallowing and breathing. This suggests that cortical areas of the brain are not directly interacting with each other, but most areas can be reached from all others through only a few interactions.

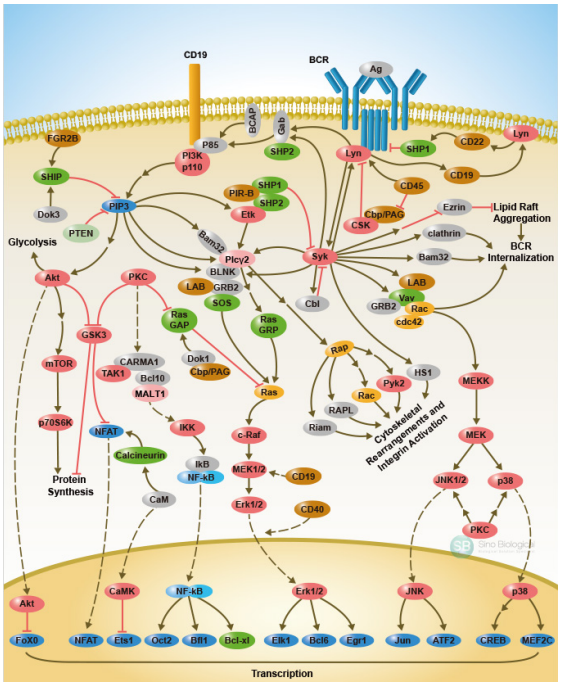
- **ecological networks**: representations of the interactions that occur between species within a community. The interactions include competition, mutualism and predation, and network properties of particular interest include stability and structure. For ecological graphs, species (nodes) are connected by pairwise interactions (links). These interactions can be trophic(food network or interconnection of many food-chains in an ecosystem) or symbiotic. Ecological networks are used to describe and compare the structures of real ecosystems, while network models are used to investigate the effects of network structure on properties such as ecosystem stability.

In biology, the most useful type of network is a pathway. A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell. It can trigger the assembly of new molecules, such as fat or protein, turn genes on and off, or spur a cell to move. In a pathway point of view, metabolic networks become metabolic pathways, gene regulatory networks become genetic pathways and signal transduction networks become signaling pathways.

Most biological networks can also be seen as a kind of state machine since the network of reactions can be seen as a state and this state changes(or stays the same) with every new stimulus/event from inside or outside of the cell.



(a) **Structure of graphs in GRN and PPI**



(b) **Signaling pathway of B cell receptor**

Figure 5.1: Examples of biological networks

In this research, the focus is on biochemical networks such as metabolic networks and, especially, gene regulatory networks and signaling pathways.

5.1 Metabolic networks

Metabolic networks describe the relationships between small biomolecules (metabolites) and the enzymes (proteins) that interact with them to catalyze a biochemical reaction. Metabolic networks, metabolic control and modeling of metabolic networks in genome-wide reconstructed models is a central area in systems biology. A metabolic network is the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. These networks comprise the chemical reactions and interactions of metabolism, the metabolic pathways, and the regulatory interactions that regulate these reactions. A metabolic pathway is a linked series of chemical reactions occurring within a cell that enables it to keep living, growing and dividing. The reactants, products, and intermediates of an enzymatic reaction are known as metabolites, which are modified by a sequence of chemical reactions catalyzed by enzymes. In most cases of a metabolic pathway, the product of one enzyme acts as the substrate for the next. However, side products are considered waste and removed from the cell and these enzymes require cofactors (non-protein chemical compound or metallic ion that is required for an enzyme's role as a catalyst) to function properly.

Metabolic networks/pathways give a complete vision of cellular metabolism and the number of chemical substances that go into and out of the cell. They represent information of how the cell uses and create substances for its growth and survival and generates energy to the very detail with a simple representation. Some applications of metabolic networks are:

- Cure disease or inflammation(altered state of a patient) by using the knowledge obtained from the metabolic processes modeled by metabolic networks. Identifying what genes, proteins and other molecules are involved in a biological pathway can provide clues about what goes wrong when a disease strikes. These types of analysis is similar to what is done to differential expression analysis but it is applied to differences in the metabolic pathways of two groups or patients or more (as a type of differential network analysis or a study of the difference between graphs/pathways, this type of analysis is very useful for personalized medicine since it relates expression information of the patient to the hidden structural information underlying in the known pathway).
- Control infections of pathogens by understanding the differences of metabolism between the host organism and the pathogen itself;
- Model the effect of a drug that is known to enhance or inhibit the creation of some metabolites and model the reaction in silico.

Metabolic pathways are one of the most important pathways for personalized medicine since finding out what pathway is involved in a disease and identifying which step of the pathway is affected in each patient may lead to more personalized strategies for diagnosing, treating and preventing disease.

There are two types of metabolic pathways that are characterized by their ability to

either synthesize molecules with the utilization of energy (anabolic pathway) or break down complex molecules and release energy in the process (catabolic pathway). The two pathways complement each other in that the energy released from one is used up by the other. The degradative process of a catabolic pathway provides the energy required to conduct the biosynthesis of an anabolic pathway. In addition to the two distinct metabolic pathways is the amphibolic pathway, which can be either catabolic or anabolic based on the need for or the availability of energy.

Each metabolic pathway consists of a series of biochemical reactions that are connected by their intermediates: the products of one reaction are the substrates for subsequent reactions, and so on. Metabolic pathways are often considered to flow in one direction. Although all chemical reactions are technically reversible, conditions in the cell are often such that it is thermodynamically more favorable for flux to proceed in one direction of a reaction. The nodes of the metabolic pathway are metabolite(intermediate and final products of metabolism) and genes(or the proteins that are related to genes to be more precise since the protein are the functional parts of the network that catalyze chemical reactions).

Metabolic networks and pathways play a central role in metabolism and the functioning of the whole system of a cell, tissue and the whole organism as well, as can be seen in figure 5.2.

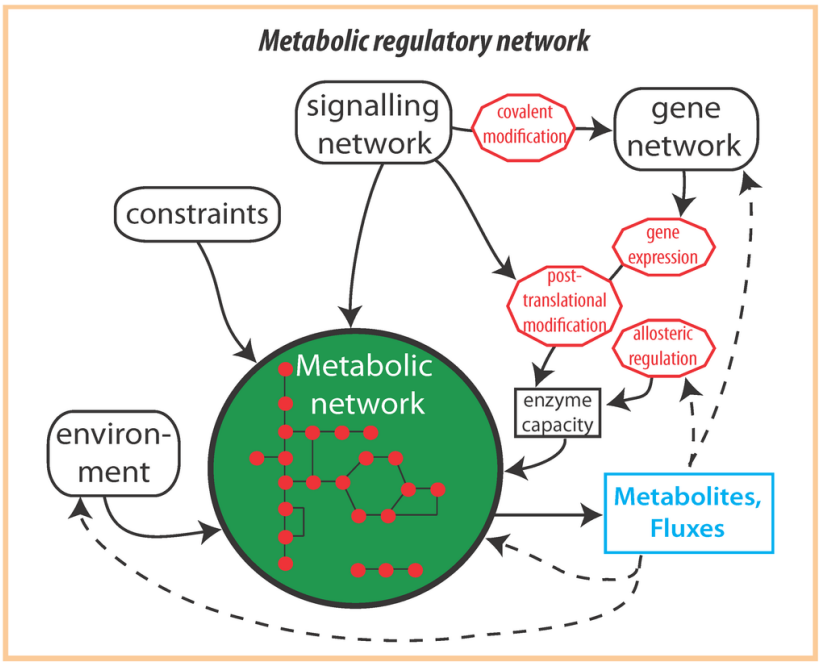


Figure 5.2: Overview of regulatory interactions involved in metabolic regulatory networks. The function of metabolic networks are governed by constraints. The regulation of a metabolic network involves a tight interplay between different cellular networks such as signaling and gene networks and by interactions with its environment. Source [19]

The study of metabolic networks has been increasing day by day in the field of systems biology and has become a dynamic complex interactive non-linear system. Numerous metabolic databases are available to the public. Important information obtained

from these databases can help in metabolic network reconstruction of the organism. Now, semi-automatic assembly of metabolic network reconstructions is crucial due to concerns of time and effort.

Many important biological pathways are discovered and researched through laboratory studies of cultured cells, bacteria, fruit flies, mice and other organisms. Many of the pathways identified in these model systems are the same as or are similar to counterparts in humans.

Still, many biological pathways remain to be discovered. It will take years of research to identify and understand the complex connections among all the molecules in all biological pathways, as well as to understand how these pathways work together.

KEGG [75] is one of the first pathway databases that were initiated to move from the existing gene catalogs to pathways catalogs. It is a bioinformatics database that contains information such as proteins, genes, pathways, and reactions. It is also a bioinformatics resource for linking genomes to life and the environment. In the KEGG organism area, it has been divided into two parts, eukaryotes and prokaryotes, which comprise much data such as gene and DNA information that can be simply searched by typing in the enzyme of choice. KEGG is commonly applied in the area of in silico modeling of metabolic networks/pathways and signaling pathways that will be seen shortly in section 5.3.

5.2 Gene regulatory network

A gene regulatory network (GRN) is a collection of regulatory relationships between transcription factors (TFs) and TF-binding sites of specific mRNA to govern certain expression levels of mRNA and their resulting proteins. It is also called **genetic pathway**.

The regulator can be DNA, RNA, protein and complexes of these. The interaction can be direct or indirect (through transcribed RNA or translated protein). In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases, this protein will be structural and will accumulate at the cell membrane or within the cell to give it particular structural properties. In other cases, the protein will be an enzyme(micro-structure that catalyzes a certain reaction), such as the breakdown of a food source or toxin. Some proteins though serve only to activate other genes, and these are the transcription factors that are the main players in regulatory networks or cascades. By binding to the promoter region at the start of other genes they turn them on, initiating the production of another protein, and so on. Some transcription factors are inhibitory.

The elucidation of gene regulatory networks (GRNs) can enhance the information that is hidden in complex cellular processes in living cells, and these networks generally reveal regulatory interactions between genes and proteins as can be seen in figure 5.3. It should be noted that GRN determination is not the final outcome of a biological study, but rather an intermediate bridge connecting genotypes and phenotypes. Previously, microarray-based bulk RNA-seq was utilized to uncover these networks, although scRNA-seq has been more recently applied for this purpose. Single-cell genomics has

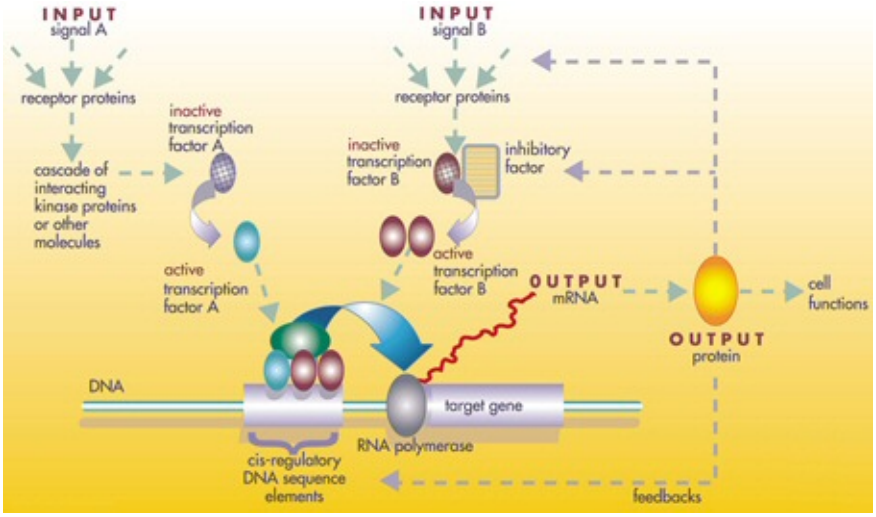


Figure 5.3: Structure of a gene regulatory network.[162]

made it easier to infer GRNs, as typical experiments allow the capture of thousands of cells in one condition, which increases statistical power. However, GRN determination remains challenging due to intracellular heterogeneity and the vast number of gene-gene interactions.

Biological cells can be thought of as chaotic bags of biological chemicals that have a structure and relations between the components that compose them. In the discussion of gene regulatory networks, these chemicals are mostly the messenger RNAs, proteins that arise from gene expression and environmental properties(pH inside and outside of the cell that are also influenced by the proteins and the other chemicals in the environment). These mRNA and proteins interact with each other with various degrees of specificity. Some diffuse around the cell. Others are bound to cell membranes, interacting with molecules in the environment, pumping them, transforming them or transporting them around the cell. Still others pass through cell membranes and mediate long range signals to other cells in a multi-cellular organism. These molecules and their interactions comprise a gene regulatory network. A typical gene regulatory network can be seen in figure 5.4.

The nodes of this network can represent genes, proteins, mRNAs, protein/protein complexes or cellular processes. Nodes that are depicted as lying along vertical lines are associated with the cell/environment interfaces, while the others are free-floating and can diffuse. Edges between nodes represent interactions between the nodes, that can correspond to individual molecular reactions between DNA, mRNA, miRNA, proteins or molecular processes through which the products of one gene affect those of another, though the lack of experimentally obtained information often implies that some reactions are not modeled at such a fine level of detail and nowadays, instruments that can measure a high level of specificity are not yet available or are limited. These interactions can be inductive/regulative (usually represented by arrowheads or the + sign), with an increase in the concentration of one leading to an increase in the other, inhibitory (represented with filled circles, blunt arrows or the minus sign), with an increase in one leading to a decrease in the other, or dual, when depending on the circumstances the

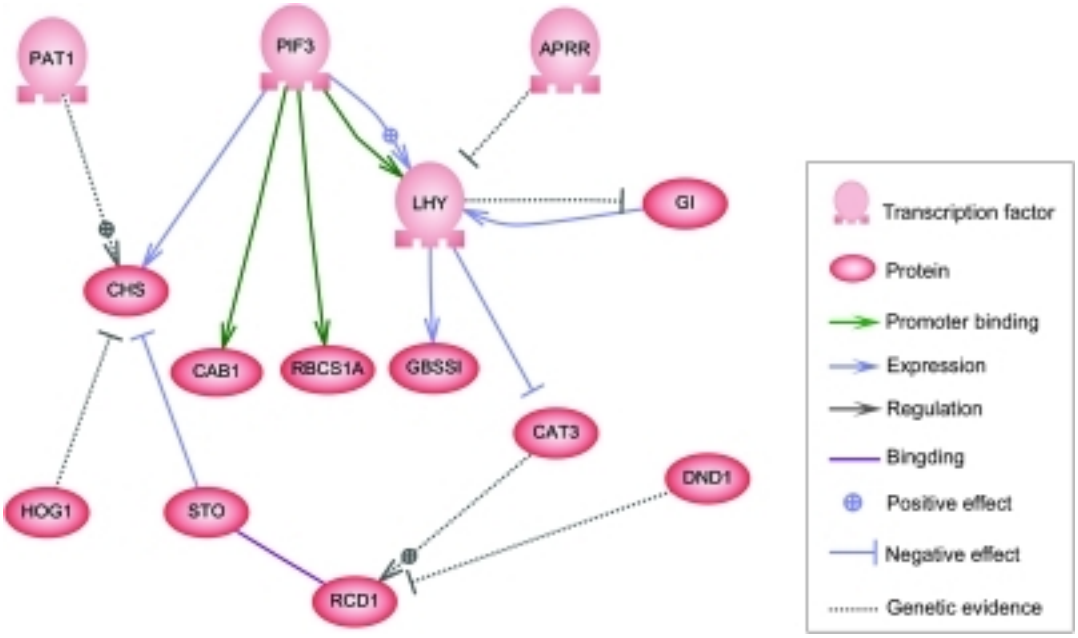


Figure 5.4: Example of gene regulatory network.[162]

regulator can activate or inhibit the target node. The nodes can regulate themselves directly or indirectly, creating feedback loops(that are, most of the times, not even considered in most of the tools since cycles need particular care, instruments that will be seen in the next chapter, MITHrIL [5] and PHENSIM[6] cannot handle cycles in regulatory networks/pathways), which form cyclic chains of dependencies in the topological network. The network structure is an abstraction of the system's molecular or chemical dynamics and complex interactions, describing the manifold ways in which one substance affects all the others to which it is connected and the activity of the cell itself. In practice, such GRNs are inferred from the biological literature on a given system and represent a distillation of the collective knowledge about a set of related biochemical reactions. To speed up the manual curation of GRNs, some recent efforts try to use text mining, curated databases, network inference from massive data, model checking and other information extraction technologies for this purpose, but the research is quite young and the field of building GRNs is thriving.

Gene regulatory networks are predicted from RNA-seq data (from bulk and single cell). Predicted regulatory networks are formed from transcription factors (TFs) and downstream target genes (termed regulons) based on large-scale RNA-seq data as well as the known TF-target relationships for various human and animal(mouse, mus musculus,etc.) conditions. The prediction is done by a pipeline that is compromised of an RNA-seq input, deploys a clustering of the expressed genes, a motif search to enrich the graph and uses a co-expression analysis to build the resulting graph [[36],[90]] with Bayesian networks and inference models(regression-based). The networks created are not permitted to have any cycle even though a lot of biological cases where TF are influenced by cycles can be observed. The approaches used to model gene regulatory networks have been constrained to be interpretable and, as a result, are generally sim-

plified versions of the network. In past years, Boolean networks have been used due to their simplicity and ability to handle noisy data but lose data information by having a binary representation of the genes. Also, artificial neural networks cannot be using a hidden layer so that they can be interpreted, losing the ability to model complex and non-linear correlations in the data. Using a model that is not burdened with interpretability can produce more accurate models. In recent years and with the advent of better interpretability for new AI models, the possibility of getting better results is growing rapidly and the GRN are even more accurate. This growing accuracy is also due to scRNA-seq that permits a higher level of accuracy for the GRN of a specific cell.

Numerous computational algorithms have been developed to address the massive amount of gene expression data generated from bulk population analysis and uncover GRNs. These methods can be categorized into machine learning-based, co-expression-based, model-based, and information theory-based approaches. Co-expression-based approaches are perhaps the simplest method for identifying putative relationships, but these approaches are unable to model the precise dynamics of cellular systems. Model-based inference, such as Bayesian networks, uses many parameters and is time-consuming. Additionally, probabilistic graphical models require searching for all possible paths for many genes, which is an NP-hard problem. More recently, information theory-based methods utilizing mutual information and conditional mutual information have gained popularity because they are assumption-free and can measure non-linear associations between genes.

From a single-cell view, the stochastic features of a single cell must be properly integrated into GRN models. As noted above, technical noise is difficult to distinguish from true biological variability, and the remaining variability is still poorly understood. However, the asynchronous nature of single-cell data, as well as the presence of multiple cell subtypes and the sparsity of current and most used sequencing technologies, may provide the inherent statistical variability required to detect putative regulatory relationships.

It is worth emphasizing that the detection of regulatory relationships should be possible in a reasonable timescale, as transcriptional changes do not persist forever. Further, the directionality between genes in identified networks must be validated and refined with perturbation studies or temporal data in order to infer causality.

A common database to use to browse gene regulatory networks is **Gene Regulatory Network database** (GNRdb) [47]. All the regulations in GRNdb are predicted from the omics data rather than being experimentally determined. Users can easily search, browse, and download the TF-target pairs and corresponding motifs of a variety of conditions at the single-cell or bulk level, as well as investigate the expression profile of a list of genes simultaneously and analyze the association between gene expression level and the patients' survival of diverse TCGA cancers.

5.3 Signalling pathway

Signal transduction pathways (or signaling pathways for short) are events that carry signals and the transmissions of these signals inside and outside of the cell with response

events to stimuli. These events establish the changes in the gene transcription and in the final functions of the pathway itself by using expression control and regulation. An event in a pathway has a "global" effect on the cell since a reaction is usually followed by one or more other reactions that build up the pathway originating from the stimuli and that have some specific attributes and functions. The structure is very similar to Genetic pathways but GRNs are processes that regulate the cell activity and are usually used to represent a state machine of the cell-state with the interactions of TFs and genes while signaling pathways result in some products and are seen as an input-output program, but the difference between the two is unclear and ambiguous, at least for me right now.

Signal transduction pathways move a signal from a cell's exterior to its interior. Different cells are able to receive specific signals through structures on their surface called receptors. After interacting with these receptors, the signal travels into the cell, where its message is transmitted by specialized proteins that trigger a specific reaction in the cell. For example, a chemical signal from outside the cell might direct the cell to produce a particular protein inside the cell. In turn, that protein may be a signal that prompts the cell to move. Signal transduction is the process by which a chemical or physical signal is transmitted through a cell as a series of molecular events, most commonly protein phosphorylation catalyzed by protein kinases, which ultimately results in a cellular response. Proteins responsible for detecting stimuli are generally termed receptors, although in some cases the term sensor is used. The changes by signal sensing in a receptor give rise to a biochemical chain reaction, which is a series of biochemical events that build the signaling pathway. When signaling pathways interact with one another they form networks, and these networks can be modeled by **metapathways** that are the union(as the operation of graphs union, where vertices are collapsed and edges are added if one edge exists in at least one of the two or more graphs in the union) of pathways graphs to form a complete network that incorporates. At the molecular level, such responses include changes in the transcription or translation of genes, post-translational and conformational changes in proteins, as well as changes in their location.

Each node of a signaling pathway is classified according to the role it plays with respect to the initial stimulus. Ligands are termed first messengers, while receptors are the signal transducers, which then activate primary effectors. Such effectors are typically proteins and are often linked to second messengers, which can activate secondary effectors, and so on. Depending on the efficiency of the nodes, a signal can be amplified (a concept known as signal gain), so that one signaling molecule can generate a response involving hundreds to millions of molecules. Usually and especially for this project, nodes in signaling pathways are genes (with other signals embedded in the genes that express them).

The question of how to understand if a specific pathway is activated is very important and the subject of research since their discovery and the advancement of sequencing technology and methodology. One way of knowing if a pathway is active is to see if the **targets of the pathway are activated**, but with this approach, other pathways with the same targets could also be influencing the results. Another method is to see if the gene in the pathway is activated and expressed (or not expressed enough in case

of signal inhibition), but with this approach, there are a lot of cases when some genes are expressed even when the pathway has not yet been activated. Finally, probably the best way to see if the pathway is activated is to see if the genes are **differentially expressed**. The last approach will be used in 6.2 and 6.3 to generate the embedding where every value represents the activity of the genes and the pathway as a whole.

As already seen during section 5.1, signaling pathways are available in a lot of databases but the most used and famous is KEGG [75], where signal transduction pathways are published to the public.

5.4 Protein interaction

Protein interactions are one of the most important topics of biology since they regulate the functions and reactions of a cell (and a system as a whole).

This type of network holds information about how different proteins operate with each other to enable a biological process within a cell. The interactions in a PPI network can be physical or predicted. Notably, a whole interactome can capture all PPIs happening in a cell or an organism. In vivo and in vitro methods for detecting PPIs physical interactions include:

- **X-ray crystallography**: a technique used to obtain the three-dimensional structure of a particular protein by x-ray diffraction of its crystallized form. This three-dimensional structure is crucial to determining a protein's functionality
- **NMR(Nuclear magnetic resonance spectroscopy)**: a spectroscopic technique to observe local magnetic fields around atomic nuclei. In the case of proteins (**protein NMR**), it is used to obtain information about the structure and dynamics of proteins, and also nucleic acids, and their complexes.
- **tandem affinity purification (TAP)**: is a methodology for the isolation of protein complexes from endogenous sources. It involves the incorporation of a dual-affinity tag into the protein of interest and the introduction of the construct into desired cell lines or organisms. Using the two affinity handles, the protein complex assembled under physiological conditions, which contains the tagged target protein and its interacting partners, can be isolated by a sequential purification scheme.
- **affinity chromatography**: called affinity purification in the case of protein study of the interactions. It is used similarly as TAP, it involves the separation of molecules in solution (mobile phase) based on differences in binding interaction with a ligand that is immobilized to a stationary material (solid phase)
- **Co-immunoprecipitation**: a technique to identify physiologically relevant protein-protein interactions by using target protein-specific antibodies to indirectly capture proteins that are bound to a specific target protein. These protein complexes can then be analyzed to identify new binding partners, binding affinities, the kinetics of binding and the function of the target protein. Co-immunoprecipitation is an extension of IP that is based on the potential of ImmunoPrecipitation reactions to capture and purify the primary target (i.e., the antigen) as well as other macromolecules that are bound to the target by native interactions in the

sample solution. Therefore, whether or not an experiment is called an IP or co-immunoprecipitation depends on whether the focus of the experiment is the primary target (antigen) or secondary targets (interacting proteins).

- **protein arrays**: is a high-throughput method used to track the interactions and activities of proteins and to determine their function in a local setting and on a large scale. Its main advantage lies in the fact that large numbers of proteins can be tracked in parallel like high-throughput sequencing. The technology is a direct descendant of the DNA microarrays introduced in 2.1.
- **protein fragment complementation**: is a method for the identification and quantification of protein-protein interactions where the proteins of interest ("bait" and "prey") are each covalently linked to fragments of a third protein("reporter"). Interaction between the bait and the prey proteins brings the fragments of the reporter protein in close proximity to allow them to form a functional reporter protein whose activity can be measured. This principle can be applied to many different reporter proteins and is also the basis for the yeast two-hybrid system that will be introduced soon
- **phage display**: an in vitro screening technique for identifying ligands for proteins and other macromolecules. In this technique, a gene encoding a protein of interest is inserted into a phage coat protein gene, causing the phage to "display" the protein on its outside while containing the gene for the protein on its inside, resulting in a connection between genotype and phenotype. These displaying phages can then be screened against other proteins, peptides or DNA sequences, in order to detect the interaction between the displayed protein and those other molecules.
- **mass spectrometry**: used to identify a group of proteins that could interact among themselves in a network. It is a method for the accurate mass determination and characterization of proteins, it is used for the identification of proteins and their post-translational modifications, the elucidation of protein complexes, their subunits and functional interactions, as well as the global measurement of proteins in proteomics.
- **yeast two-hybrid**: used to locate binary interactions by testing for physical interactions (such as binding) between two proteins, the test is the activation of downstream reporter gene(s) by the binding of a transcription factor onto an upstream activating sequence (UAS).

Other methods that are not experimental but computational for identifying PPI's functional interactions are:

- correlation of expression profiles from mRNA sequencing;
- Genetic interactions;
- Methods in-silico, where a simulation of the PPI is performed on a computer or via computer simulation. For example, gene fusion is where two genes present in a group are fused together in another group.
- Machine learning techniques that use transcriptomics to find protein structure and interactions with other proteins to build a network.

Another field related to protein-protein interaction but also in the topology and shape of the protein molecules that strands of aminoacid produce in 3D space is **contact**

maps. A protein contact map represents the distance between all possible amino acid residue pairs of a three-dimensional protein structure using a binary two-dimensional matrix. Contact maps provide a more reduced representation of a protein structure than its full 3D atomic coordinates. The advantage is that contact maps are invariant to rotations and translations. They are more easily predicted by machine learning methods. Contact maps are also used for protein superimposition and to describe similarities between protein structures. They are either predicted from protein sequence, calculated from a given structure or inferred from transcriptomics.

At the root of the structure of proteins is protein folding and the process of folding is not instantaneous, so contact maps could also incorporate a temporal feature in their representation (as temporal or multi-layer graphs)

5.5 Cell interaction

Cell–cell interactions orchestrate organismal development, homeostasis and single-cell functions. When cells do not properly interact or improperly decode molecular messages, disease ensues. Thus, the identification and quantification of intercellular signaling pathways have become a common analysis performed across diverse disciplines. The expansion of protein–protein interaction databases and recent advances in RNA sequencing technologies have enabled routine analyses of intercellular signaling from gene expression measurements of bulk and single-cell data sets. In particular, ligand–receptor pairs can be used to infer intercellular communication from the coordinated expression of their cognate genes. Direct interactions between cells, as well as between cells and the extracellular matrix, are critical to the development and function of multicellular organisms. Some cell-cell interactions are transient, such as the interactions between cells of the immune system and the interactions that direct white blood cells to sites of tissue inflammation. In other cases, stable cell-cell junctions play a key role in the organization of cells in tissues. For example, several different types of stable cell-cell junctions are critical to the maintenance and function of epithelial cell sheets. Plant cells also associate with their neighbors not only by interactions between their cell walls but also by specialized junctions between their plasma membranes.

Many cell–cell interactions have not been well characterized, partially because, until recently, much work was devoted to examining cell–matrix interactions, and partially because studying cell–cell associations comes with a long list of challenges. There are three key challenges to the latter. First, such associations are generally more geometrically complex. This is not to say that cell–cell adhesion plaques are necessarily smaller, but cell–substrate interactions can be examined over the area of some spread, adhered cell, whereas cell–cell junctions are limited to a slightly thickened perimeter of the cell, with 3D orientation even when cells are plated on flat surfaces. Second, various extracellular matrices are readily isolated or purchased, while cell–cell junctions rely more on specific receptor–receptor interactions, which in turn rely on expensive antibodies or purified receptors. Finally, the effects of cell–cell interactions are very difficult to isolate from cell–substrate interactions in adherent cells. The reverse is a bit simpler—some cell models can work with sparsely plated cells. But generating consistent cell culture

conditions where cell–cell interactions dominate is not always a simple matter.

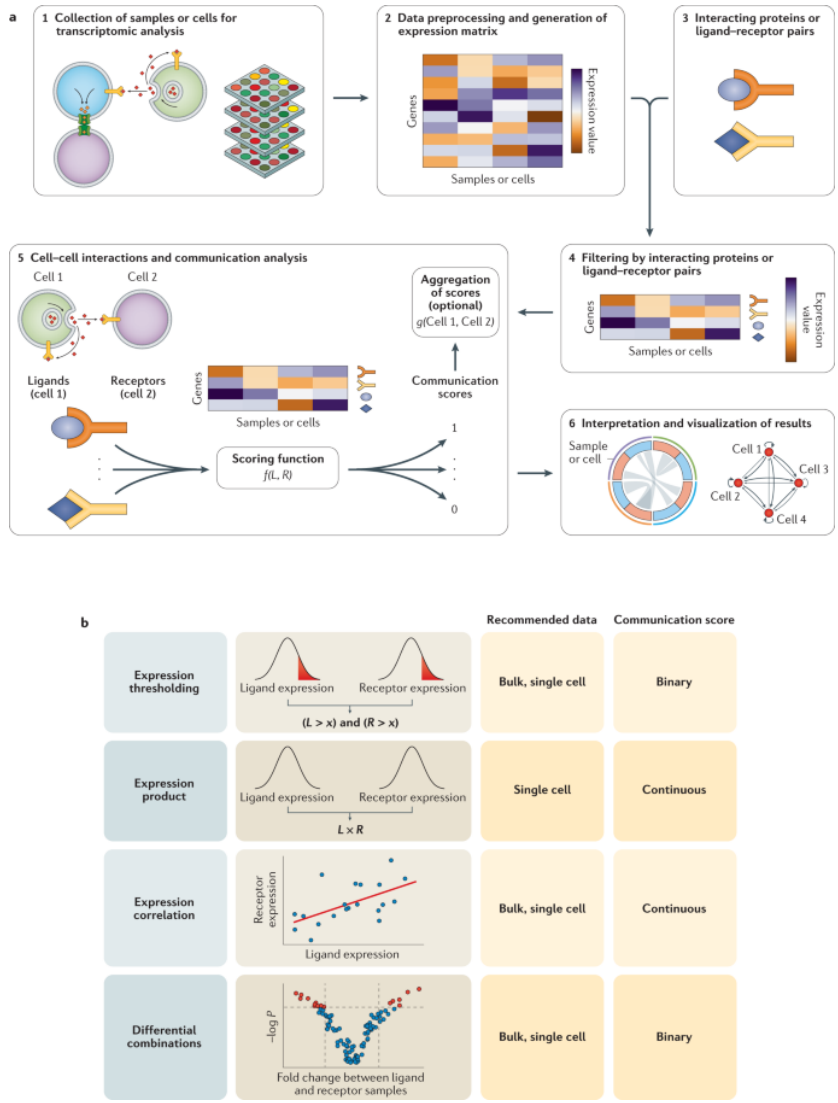


Figure 5.5: **Analysis workflow for inferring cell–cell interactions and communication from gene expression.** Source [12]

Aside from a biological standpoint, cell-cell interactions are modeled easily by graphs and, especially in the case of spatial transcriptomics, **spatial networks**(or geometric networks, a graph in which the vertices or edges are spatial elements associated with geometric objects, cells in the case of CCI). These types of graphs can be a snapshot of a particular point in time (for spatial transcriptomics and spatial identification of the cells) or can be a general model to find relations and interactions between cells. In the case of spatial networks, a temporal feature can be associated with both nodes and edges, making the complex network representation even more detailed. Spatial transcriptomics and the reconstruction of 3-dimensional structure from data is a very useful and thriving field in recent years. In the setting of spatial transcriptomics, temporal graph analysis can be done if a temporal feature is associated. For example, a motif

finding (isomorphism identification) can be done on these spatial/temporal graphs to find some sub-structures of interaction that are common in the graph and that incorporate topological information that could be useful to understand some applications for drugs and stimuli response. Algorithms for motif finding in graphs are one of my specialties, one of the algorithms that I have developed for temporal graphs is TemporalRI [[96],[94],[106]]. Algorithms that build spatial graphs use different techniques from graph building/inference(Voronoi graphs, fixed graph representation in case of already defined position, space/cell decomposition) and data obtained from scRNA-seq, spatial transcriptomics(with platforms that sequence reads and assign a position to the sequence data) and other -omics. To build CCI networks, on the other hand, the process is similar to the data used to build the network but the spatial information is lost/not considered most of the time since the interactions are the main focus of CCI network creation and interaction identification. No further information about CCI will be seen here since it is not the focus of this project, but it could be useful especially for scRNA-seq since interactions between cells can be used to make even more accurate and enriched data from scRNA-seq. For more information about CCI and cell-cell graph creation and inference see [[150],[33],[12]].

An image that represents the workflow to infer CCI can be seen in figure 5.5.

5.6 Other networks

Aside from networks built mainly from experiments in a direct way, there are also networks that can be inferred from sequencing data and -omics data.

Another very useful type of graph that is used in biology and that uses RNA-seq data is a **Correlation network** that uses gene co-expression and differential expression analysis to build a network of genes related by their relations. Weighted correlation network(or weighted gene co-expression network, abbreviated WGCN) in particular, is a widely used data mining method especially for studying biological networks based on pairwise correlations between variables(genes). It can also be applied to most high-dimensional data sets and is not limited to gene expression (as long as a differential analysis can be done on the data). It allows to define clusters, network nodes with regard to groups and to comparison of the network topology of different networks (as differential network analysis as introduced before when talking about metabolic networks/pathways). WGCNA can be used as a data reduction technique, as a clustering method (fuzzy clustering), as a feature selection method, as an embedding and latent space analysis tool to find hidden variables and summarize information about the system, as a framework for integrating complementary (genomic) data (based on weighted correlations between quantitative variables and on other types of -omics data), and as a data exploratory technique. Gene co-expression networks are not the direct results of experiments but are built from data.

The intuition behind weighted correlation networks is that genes with similar expression patterns among samples(co-expressed) are functionally associated, suggesting that these genes share some of the same functions, are involved and activate possibly the same pathways and GRN and influence each other in some way, for example, they

influence each other or the same/similar set of genes that they are influenced to (the entering nodes of the nodes/genes in the graph) or a similar set of genes influenced by these genes(the outcoming edges from these genes end up at a similar set of targets).

The building of correlation networks is straightforward in the classic setting but can get really difficult when involving other attributes and data (other -omics and also bulk/single-cell data at the same time). In the simplest case, a correlation matrix is built with pairwise correlation on genes samples, and the entries that are over a certain threshold (as absolute values) become edges in the network. Other than that, the network can be pruned by seeing differentially expressed genes and by locating modules of genes with clustering analysis. A representation of Weighted gene co-expression network analysis can be seen in figure 5.6

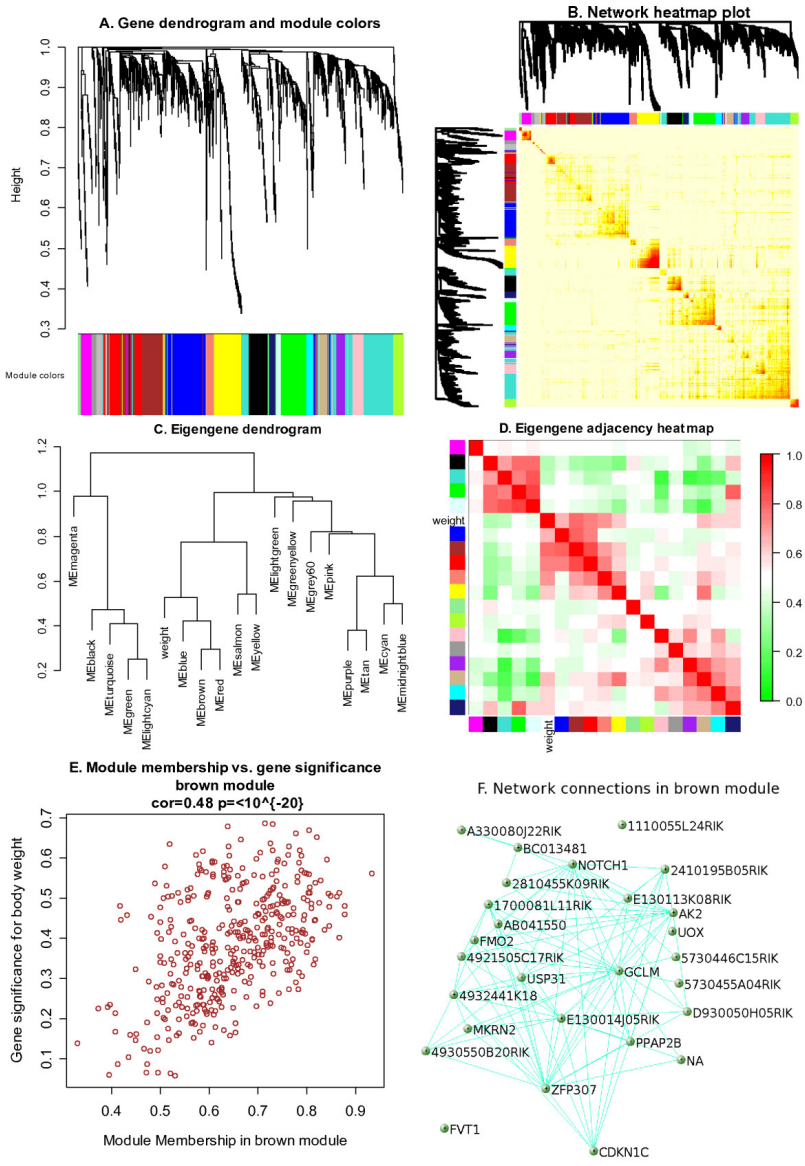


Figure 5.6: **Example WGCNA analysis of liver expression data in female mice.** Source [84]

Another type of network representation used for biology and, especially, for personalized medicine is an association network. These networks link two types of conditions (disease-gene or drug target) with the genes that cause the two conditions or target them in the same way, the resulting graph will be a bipartite graph.

Knowledge graphs are also used in biology to model ontologies with topological relations and build a model that groups several categories and characteristics of multi-omics and attributes of the data. Some types of relations present in a biology knowledge graph are:

- **Gene expression** — expression, promotor binding, miRNA effect
- **Proteomics/physical interaction** — direct regulation, protein modification, binding
- **Biomarkers** — biomarker, genetic change, quantitative change, state change
- **Metabolomic transport and modification** — molecular synthesis, molecular transport, chemical reaction
- **Functional association** (between a disease and a cellular process or another disease)
- **Regulation**, which is the least specific type of relationship and is used if no more specific information is available

Knowledge graphs often serve as a backbone for data integration within organizations, providing a common representation structure that enables querying across data sources. The use of knowledge graphs as a data source for machine learning methods to solve complex problems in life sciences has rapidly become popular in recent years and, with the advancement of AI methodologies, knowledge graphs gained one more important purpose: to serve as training data for ML models, and graph machine learning models in particular. Their newfound use as training data has to be taken into account when constructing knowledge graphs and influences core design choices.

The life sciences domain has been an early adopter of ontologies and linked data technologies. The primary motivation, given the vast amount of knowledge to capture, was the need for standardization of the vocabularies and taxonomies. This was essential for the integration of data within and across large organizations and for enabling data access. For this reason, the design of graph datasets focused on high granularity of the models and the ability to capture the complexity of the domain in the most precise way. Such ontologies included, for example, the Gene Ontology (GO) [13] for gene functions, the Human Disease Ontology (DO) [133] for capturing different disease classifications, or BioPax [31] for pathway-related information.

Constructing knowledge graphs to support graph analytics and machine learning has its own set of requirements that do not completely match the scenarios where the primary use of the graph is to enable complex cross-source querying. For example, a very precise and detailed data model can make it hard for an algorithm to learn essential relations between nodes, if they are not connected directly, but separated by several hops. While the abilities to search and formulate expressive queries are still important, the possibility to customize and manipulate the graph for different use cases becomes particularly valuable. Depending on the task (e.g., drug repurposing or target selection) or the domain (e.g., specific disease like asthma), the users should be able to extract a custom relevant subgraph to apply statistical methods and train models. In

general, it is important to combine for analysis both public reference datasets capturing state-of-the-art knowledge.

To build a knowledge graph there is only the need to have **semantic or RDF triples** that describe the relation of a subject (drug, group of patients, cell type, etc.) and an object (a disease, an environment, some conditions, etc.). This type of representation of knowledge permits the complexity of the structure of a problem and a representation of information (as "unstructured" information, as opposed to structured information found in schemas and SQL databases). To build the knowledge graph and the relation, NLP and semantic web mining is used to query the biological database and build the RDF (if they are not available).

Extracting a large amount of structured information from biomedical literature remains a very complex task. The language used to describe biological entities and relationships varies across temporal and geographic dimensions and is subject to various phenomena in language evolution (for instance, neologisms, synonyms, multi-entity constructs, etc.).

Progress in developing a generalized knowledge base population solution is hampered by:

- Lack of sufficiently diverse training data covering the breadth and depth of biomedical knowledge
- Biases in existing training data sets (such as only covering a subset of diseases)
- Uncertainty/fluidity in our understanding of disease mechanisms, resulting in inconsistent language usage over time
- Contradictory evidence and or irreproducible results
- Uncertainly in the interpretation of data, often resulting in authors hedging the assertions they make
- Biases in literature to only report positive findings
- Poor generalisability of ML models, research findings are generally overfitted to small datasets and not suitable for production offerings

Another approach to building and modelling a complex system of interactions that integrates both cell-cell interaction and gene-gene regulation is to consider a **multi-layer metapathway** where every layer represents a cell-cluster (that could be typed or not) and interactions between layers represent communication between cells. The procedure to create this network can be quite difficult but, from a logical standpoint, the process should be almost the same as the one used for differential analysis for different clusters (seen in 3.4), but the genes used to see where two clusters communicate and establish the edges from layer to layer should be genes used for the communication between cells. These cell-cell communication related genes could be lygand-receptors as seen in [82], and the edges layers' genes could be established if and only if the genes are differentially expressed significantly.

Chapter 6

Pathway analysis

Biological pathways are a way to represent sequences of interactions among molecules in a cell that leads to a certain product or a change in a cell or system in an organism as already seen in the previous chapter 5. This project is focused on understanding the applications of single-cell analysis in personalized medicine. By having this objective, the focus of the project also shifts and is divided into different settings and analyses that could be done with the methodologies defined previously in section 3 and section 4. The methods previously seen do not consider topological data to do this analysis and are, most of the time, only based on expression values and references to see if there is any significant difference between groups of samples.

To obtain and **embed** more information in the models that could be used to implement a personalized therapy for a group of patients or some study related to using sequencing data from both bulk, single-cell and other technologies, graphs can be used to encode topological information about the interaction of the entities that are studied (genes in this case).

Graph embeddings is a graph learning model used to build a function that maps the nodes of the network to some tensors that describe the nodes and can be used to find the similarity of two different nodes. Classic embedding techniques try to find an embedding such that, given a simple graph(directed or undirected) (V, E), two nodes $u, v \in V$, a similarity function of the nodes $sim : V^2 \longrightarrow \mathbb{R}$ and a mapping function $map : V \longrightarrow \mathbb{E}$ where $\mathbb{E}$ is the embedding space that could be a subspace of $\mathbb{R}^n$ or whatever is needed for the problem(as long as it can be a vector space with inner product and an ordered set of values):

$$sim(u, v) \approx map(u)^T map(v) \tag{6.1}$$

The similarity function is at the core of how embeddings are generated since it guides how two nodes can be seen as similar or not. Common properties to consider for defining a similarity function between two nodes u, v are:

- are u and v connected
- do they have a similar neighborhood (the neighborhoods share a good amount of nodes)
- do they have similar topological structure
- do they have similar properties (aside from structure and topology in general)
- do they have the same significance for the graph (as a measure of significance both for the graph and for the underlying data represented as a graph)

There are many ways of embedding nodes in a graph, for example via random walks [118], where the similarity function is the probability of encountering the nodes

in random walks in the graph. The embedding uses a random walk with fixed length and log-likelihood optimization to find the optimal parameters used to compute the embedding.

These methods were very popular at the start of the research about graph embeddings and nowadays, Machine learning methods are more popular and offer better performances than inference-based methods overall.

There is also the possibility of embedding the whole graph or a subset of nodes for the graph to compare different graphs directly. Some of the most famous methods use: a summarization function to find a summarized embedding over the nodes chosen in the graph; an additional node that has the whole set of chosen nodes as a neighborhood, this node is then embedded and the result of the embedding is reported as the embedding of the whole graph[109]; Another very popular method uses anonymous random walks [69] that builds a distribution of the anonymous random walks (where the nodes in the walk are identified in the order that they appear during the walk). More information about anonymous random walk embedding in [69].

Aside from the methods introduced previously and only seen to understand the concept of graph embeddings in the context of algorithms and methodologies, the three methods that will be seen in this section are the following:

- **GNN (Graph neural Network)**[78]: These algorithms use the topology of the graph to build neural networks used to estimate embeddings useful for inference and further analysis.
- **MITHrIL** [5]: not properly a graph embedding algorithm, but it generates a signature/vector for a meta-pathway where the single values associated can be a score for the activity of the pathway or the vector where the genes are mapped to scores that measures how much they are perturbated in the pathway.
- **PHENSIM** [6]: not a graph embedding algorithm by definition as well, it generates a signature for the data passed as input (up-regulated, down-regulated and not expressed genes) that encodes the simulation results.

Some of the methods and tools that will be seen during this chapter do not generate proper "embeddings" since MITHrIL[5] generates a score for genes(**perturbation factor**) and for the single pathway(as an **impact factor**) and these scores could be used to generate embeddings for a sample or a group of samples by using different pathways. In this way, a signature will be associated with a sample (or group of samples) and this signature can be seen as an embedding since a similar signature means similar activities in pathways for different groups. The same can be said about PHENSIM[6].

More about MITHrIL and PHENSIM will be seen in this chapter in later sections.

6.1 Pathway embedding techniques

6.1.1 Graph neural networks

In this section, Graph Neural network models will be seen, this part of the project is a direct descendant of one of my research finished in the past months[95]. This part of the project will use meta-pathways built from common pathways and the integration of

genes coding for microRNA molecules, the resulting meta-pathways will be graphs of interactions between genes(expression or inhibition) but will also retain the membership of subgroups to pathways. These networks of gene interactions hide some information that is not currently used in most applications of clinical analysis and that information could be used as new **Biomarkers**.

The objectives reached by this part of the project

Most of the most famous tools used for GNN are in python. The most famous and used libraries and APIs used to create the GNN are StellarGraph(for the implementation of the GNN, which uses Tensorflow)[30] and Tensorflow (for the training of models and the definition of the core definitions used to get the models). Another alternative to stellargraph is **spektral**[144] but the core algorithms implemented are more or less the same (because the code for GNN and GraphSage are taken from the original code[77][51])

The main objectives of the research done in [95] were:

- to understand the hidden links and pathways related to some illness and confront the results with data that is from a known normal case (sane patients) to get some knowledge and obtain new and unknown facts from the generated data.
- to get some meaningful genes and pathways embeddings that could lead to trustworthy and explainable classifiers (with explainable knowledge, consequences and decisions) capable of identifying the categories and the differences to a group of people, given their gene expressions, by generating some feature that will be used for inference.

and both of them were reached.

All Graph Neural Network algorithms are conceptually related to node embedding approaches, general supervised approaches to learning over graphs, and classification of nodes or graphs. In this research, there will not be an implementation of matrix factorization methods(related to spectral clustering, multidimensional scaling, or approaches related to PageRank with random walks) because these methods were tested on the data and extremely unsatisfactory results were obtained from the models obtained, especially for the task at hand, because these type of models do not take into account any(or very little) graph structural information during training.

Graphs are a kind of data structure that models a set of objects(nodes) and their relationships (edges). As a unique non-Euclidean data structure for machine learning, graph analysis focuses on tasks such as node classification, link prediction, and clustering. Graph neural networks (GNNs) are deep learning-based methods that operate on the graph domain. Due to its convincing performance, GNN has become a widely applied graph analysis method recently. Almost all of the work on GNN is born from the transposition of the CNN models to graph data, but the passage from Euclidean data to non-Euclidean data rises a lot of concerns and problems on the definitions from CNN and the theory behind it. Extending deep neural models to non-Euclidean domains, which is generally referred to as geometric deep learning, has been an emerging research area.

Another useful matter linked to Graph learning is **graph representation learning** which learns to represent graph nodes, edges or subgraphs by low-dimensional vectors. In the field of graph analysis, traditional machine learning approaches usually rely on

hand-engineered features and are limited by their inflexibility and high cost. Following the idea of representation learning and the success of word embedding, a lot of ways to do Graph Embeddings were born (Node2Vec is one of these embedding algorithms that uses random walks and anonymous walks to create embeddings for nodes and graphs [49], it was not presented at the start of the chapter since it is similar to the techniques presented).

Based on CNNs and graph embedding, variants of graph neural networks (GNNs) are proposed to collectively aggregate information from graph structure. Thus they can model input and/or output consisting of elements and their dependency (usually local dependency in the graph, especially for inductive models that build models based upon local neighborhoods of a defined depth, sampling is also really useful in these models.

One of the most important limitations of GNN is that the graphs used need to be static(no dynamic edges during runtime or entering nodes). More recent algorithms for GNN also have dynamic compatibility for entering nodes in the graph and changing settings/topology for the graph [[138],[116],[168]] but they are not the focus of this project, even though they could have a major impact in the case of dynamic and temporal pathways.

Graph Convolutional Networks are a transposition of Convolutional Neural networks on images and bear resemblance to spectral approaches for graph clustering and classification. GCN are called convolutional because filter parameters are typically shared over all locations in the graph(or a subset).

For these models, the goal is to learn a function of signals/features on a graph $G = (V, E)$ which takes as input:

- A feature description x_i for every node i summarized in a $N \times D$ feature matrix X (N: number of nodes, D: number of input features)
- A representative description of the graph structure in matrix form; typically in the form of an adjacency matrix A (or some function thereof)

The model produces a node-level output Z (an $N \times F$ feature matrix, where F is the number of output features per node). Every neural network layer can then be written as a non-linear function as defined in the following pattern formula:

$$H^{(l+1)} = f(H^{(l)}, A) \tag{6.2}$$

with $H^{(0)} = X$ and $H^{(L)} = Z$ (or z for graph-level outputs), L being the number of layers. The specific models then differ only in how $f(\_, \_)$ is chosen and parameterized.

An example of a very simple form of a layer-wise propagation rule is the following:

$$f(H^{(l)}, A) = \sigma(AH^{(l)}W^{(l)}) \tag{6.3}$$

where $W^{(l)}$ is a weight matrix for the l-th neural network layer and $\sigma(\_)$ is a non-linear activation function like the ReLU.

There are two limitations to the model previously described:

1. Multiplication with A means that, for every node, sum up all the feature vectors of all neighboring nodes but not the node itself (unless there are self-loops in the graph). This can be fixed by enforcing self-loops in the graph: the identity matrix is added (matrix addition) to A.

2. A is typically not normalized and therefore the multiplication with A will completely change the scale of the feature vectors (understandable by looking at the eigenvalues of A). Normalizing A such that all rows sum to one, i.e. $D^{-1}A$, where D is the diagonal node degree matrix, gets rid of this problem. Multiplying with $D^{-1}A$ now corresponds to taking the average of neighboring node features. In practice, dynamics get more interesting when using a symmetric normalization, i.e. $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ (as this no longer amounts to mere averaging of neighboring nodes).

The final formula after these considerations is :

$$f(H^{(l)}, A) = \sigma(D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{6.4}$$

This example is a way of using graph convolutional network to predict and compute features, the details about GCN and comparisons of various will not be seen in detail but they are seen in [95].

The original GCN algorithm [78] is designed for semi-supervised learning in a transductive setting, and the exact algorithm requires that the full graph Laplacian is known during training. A variant of the GraphSage algorithm can be viewed as an extension of the GCN framework to the inductive setting.

The approach followed by GraphSage(SAmple and aggreGatE) is an **inductive** approach that takes into consideration unseen nodes during the training, in contrast of **trasductive** approaches that need all nodes to be present during the training of the model to predict embeddings. This is done by taking a subset of the neighborhood during the training of the model and aggregate the embeddings in this subset to find the node embedding. Unlike embedding approaches that are based on matrix factorization, by incorporating node features in the learning algorithm, the algorithm simultaneously learn the topological structure of each node's neighborhood as well as the distribution of node features in the neighborhood.

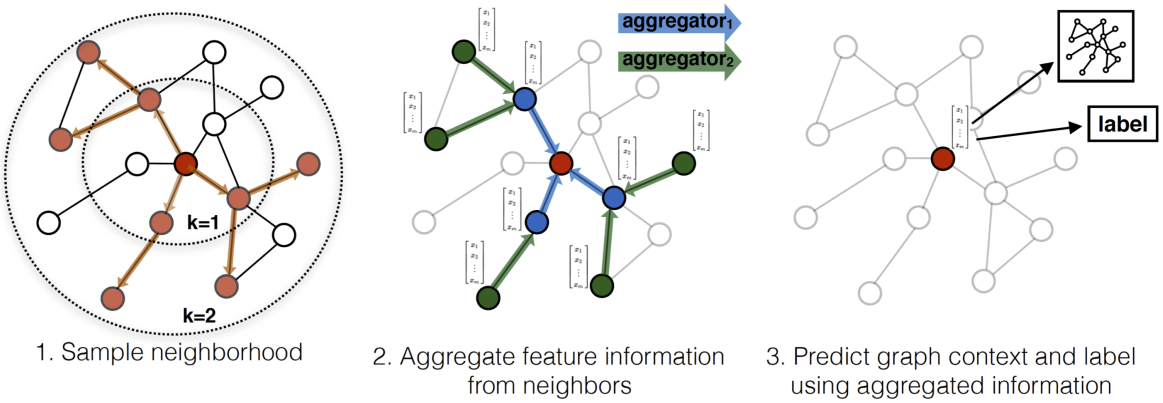


Figure 6.1: Sample and aggregation visualization of Graph Sage

However, in this research, a transductive approach is considered where all nodes need to be present for the model to work, this is because the loss function needs all the edges to be present to find the minimum(or local minimum in a certain range).

Instead of training a distinct embedding vector for each node, the model first trains a set of aggregator functions that learn to aggregate feature information from a node's local neighborhood. Each aggregator function aggregates information from a different number of hops, or search depth, away from a given node.

The key idea behind this model is that it learns how to aggregate feature information from a node's local neighborhood. The first part of the GraphSAGE model is the training of the model with a loss function(known loss function or user-defined) to find the parameters that will be used to generate the embeddings, while the second part is the embedding generation (forward propagation) algorithm, which generates embeddings for nodes assuming that the GraphSAGE model parameters are learned by the first step.

The embedding generation, or forward propagation algorithm(algorithm 3), assumes that the model has already been trained and that the parameters are fixed. In particular, the model assumes that the parameters of K aggregator functions are available(denoted $\text{AGGREGATE}_k, \forall k \in 1, ..., K$, which aggregate information from node neighbors, as well as a set of weight matrices $W_k, \forall k \in 1, ..., K$) are learned, which are used to propagate information between different layers of the model or "search depths".

Algorithm 3: Embedding generation (forward propagation)

Input: Graph $G(V, E)$; input features $x_v, \forall v \in V$; depth K; weight matrices $W_k, \forall k \in 1, ..., K$; non-linearity σ; differentiable aggregator functions $\text{AGGREGATE}_k, \forall k \in 1, ..., K$; neighborhood function $N : v \longrightarrow 2^V$

Output: Vector representations $z_v \forall v \in V$

$h_v^0 \longleftarrow x_v, \forall v \in V$;

for *k = 1,...,K* **do**

  for $v \in V$ **do**

    $h_{N(v)}^k \longleftarrow \text{AGGREGATE}_k(h_u^{k-1}, \forall u \in N(v))$;

    $h_v^k \longleftarrow \sigma(W^k \cdot \text{CONCAT}(h_v^{k-1}, h_{N(v)}^k))$;

  end

  $h_v^k \longleftarrow \frac{h_v^k}{||h_v^k||_2}, \forall v \in V$

end

The intuition behind the algorithm is that at each iteration, or search depth, nodes aggregate information from their local neighbors, and as this process iterates, nodes incrementally gain more and more information from further reaches of the graph. Further considerations on the algorithm could be seen in [51]

The neighborhood is taken by a sample of k-distant nodes from the node considered.

Ideally, an aggregator function would be symmetric (i.e., invariant to permutations of its inputs) while still being trainable and maintaining high representational capacity. The symmetry property of the aggregation function ensures that the model neural network model can be trained and applied to arbitrarily ordered node neighborhood feature sets. The aggregation function used could be user-defined or taken from different aggregation architectures, the one that will be used for this research is the mean aggregator that takes the element-wise mean of the vectors in the neighborhood(It is also called convolutional because it is the generalization of the filters seen previously for

GCN, with an inductive approach, so it is a linear approximation of a localized spectral convolution), that is:

$$h_v^k \longleftarrow \sigma(W \cdot \text{MEAN}(h_v^{k-1} \cup h_u^{k-1}, \forall u \in N(v))) \tag{6.5}$$

An important distinction between this convolutional aggregator and the other aggregators(that will not be used in this research) is that it does not perform the concatenation operation in line 5 of Algorithm 3 i.e., the convolutional aggregator does concatenate the node's previous layer representation h_v^{k-1} with the aggregated neighborhood vector $h_{N(v)}^k$. This concatenation can be viewed as a simple form of a "skip connection" [173](this paper is really good and a must-read because it is a review of most of the strategies used for GNN and the differences/similarities between them) between the different "search depths", or "layers" of the GraphSAGE algorithm, and it leads to significant gains in performance.

The model learns a function that generates embeddings by sampling and aggregating features from a node's local neighborhood. In the main paper[51] the algorithm could also be used to predict embeddings in dynamic graphs, but in the **stellargraph** implementation, the graphs used needs to be static and defined at runtime (no nodes or edges could be added to a graph at runtime to be embedded) or this is what the documentation seems to be implying(for GraphSage, the given demo takes a static graph and train the model on the graph).

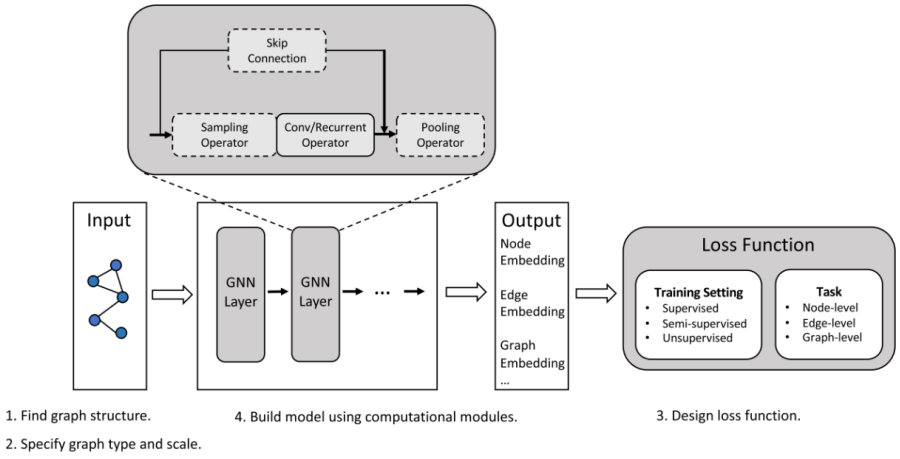


Figure 6.2: Pipeline for a GNN model

More details about graph embeddings for pathways and genes, statistical analysis of the results and the problems related to embedding genes and pathways with static GNNs can be seen in my published research [95].

6.2 MITHrIL

MITHrIL (Mirna enrIched paTHway Impact anaLysis) is a methodology and a tool for the analysis of signaling pathways. MITHrIL extends pathways by adding missing

regulatory elements, such as miRNA, and their interactions with genes. The method takes as input the expression values of genes and/or microRNAs and returns a list of pathways sorted according to their deregulation degree, together with the corresponding statistical significance (p-values). The innovative view of MITHrIL is that it considers post-transcriptional regulatory interactions enacted by miRNAs, so the possibility of having some genes down-regulated(or up-regulated if the target gene of the miRNA is a gene that inhibits some part of the transcription in the pathway)genes if miRNA genes are expressed. Another innovative part of MITHrIL is that, with newer versions, it can consider the whole meta-pathway(pathways glued/united between each other) to represent the whole route of a possible perturbation or understanding of how the whole meta-pathway behaves.

MITHrIL takes as an input expression values of genes and/or microRNAs, returns a list of pathways sorted according to the degree of their de-regulation, together with the corresponding statistical significance (p-values), and a predicted degree of alteration for each endpoint (a pathway node whose alteration, based on current knowledge, affects the phenotype in a specific way).

MITHrIL uses some scores to embed information about the activity of genes in the pathway and the dysregulation of the whole pathway. A score for the whole pathway called impact factor (IF) is a pathway-level score that takes into account biological factors such as the magnitude of change in gene expression, the type of interactions between genes, and their location in the pathway. Each pathway is modeled as a graph in which nodes represent genes, while edges represent interactions between them. Another score was introduced for gene-level statistics called perturbation factor or PF as a linear function of the change in gene expression and the perturbation of its neighborhood. Such a statistic is then combined for each element in a pathway, and a p-value is computed by means of an exponential distribution.

MITHrIL requires a case/control expression data set from which statistically differentially expressed features have been extracted (genes, miRNAs, or both). For such elements, their Log-Fold-Change(seen in 2.4) is computed. Starting from such information, MITHrIL computes, for each gene in a pathway, a Perturbation Factor (PF), which is an estimate of how much its activity is altered considering its expression and neighborhood. Positive (negative) values of PF indicate that the gene is likely activated (inhibited). By appropriately combining each PF of a pathway the algorithm is able to calculate an Impact Factor (IF) and an Accumulator (Acc). The IF of a pathway is a metric expressing how important are the changes detected in the pathway, the greater the value, the most significant are the changes. The *Acc* indicates the total level of perturbation in the pathway and the general tendency of its genes: positive Acc values indicate a majority of activated genes (or inhibited miRNAs), while negative ones correspond to an abundance of inhibited genes (or activated miRNAs). To the Acc is also assigned a p-value which is an estimate of the probability of getting such accumulator by chance.

Formally:

Definition 6.2.1. (Perturbation factor). Given a pathway $P_i = (V, E)$ where the nodes are the genes and edges are the interactions between genes, a node n in pathway P_i, a

function $\Delta E : V \longrightarrow \mathbb{R}$ that computes the log-fold-change, two sets of nodes that are the predecessors(upstream) of n $U(n, P_i)$ and the successors(downstream) of n $D(n, P_i)$ in pathway P_i and a weight function $\beta : E \longrightarrow \mathbb{R}$ associated with an edge of the graph P_i(that is the power and effect of the interaction in the pathway, negative values of indicate an inhibitory effect, while positive values an activating one.) the perturbation factor is defined as:

$$PF(n, P_i) = \Delta E(n) + \sum_{u \in N(n,P_i)} \frac{\beta(u, n)}{\sum_{d \in D(u,P_i)} |\beta(u, d)|)} PF(u, P_i)$$

The process of computing the perturbation factor seen in definition 6.2.1 is recursive, but can be iterative if starting the computation from the nodes that have no predecessor (where $PF(n, P_i) = \Delta E(n)$) and resolving the nodes that have all the already computed predecessors iteratively. This approach cannot permit the presence of cycles, so they should be filtered properly.

The impact factor reflects the importance of the changes observed in a pathway.

Definition 6.2.2. (Impact factor). Given a pathway $P_i = (V, E)$ where the nodes are the genes and edges are the interactions between genes, a probability function $prob : P \longrightarrow \mathbb{R}$ that computes the probability, calculated using a hypergeometric distribution, of obtaining a number of differentially expressed nodes at least equal to the observed one in the pathway, a function $\overline{\Delta E} : P \longrightarrow \mathbb{R}$ that computes the log-fold-change of the pathway as the mean log-fold-change of the nodes in the pathway and N_{de} are the number of differentially expressed genes in the pathway, the impact factor for a pathway is defined as:

$$IF(P_i) = log(\frac{1}{prob(P_i)}) + \frac{\sum_{n \in P_i} |PF(n, P_i)|}{|\overline{\Delta E}|(P_i) N_{de}(P_i)}$$

The computation of the impact factor uses the perturbation factor defined in 6.2.1 and can be seen as a summarization of the perturbation of the nodes in the pathway and how much the pathway is "activated".

Definition 6.2.3. (Accumulation of pathway). The accumulation measures the total perturbation accumulated from genes and miRNAs of the pathway,

$$Acc(P_i) = \sum_{g \in genes(P_i)} [PF(g, P_i) - \Delta E(g)] - \sum_{m \in miRNAs(P_i)} [PF(m, P_i) - \Delta E(m)]$$

The accumulation function takes into account exclusively the contribution to the perturbations of genes and miRNAs given by the perturbations of the downstream nodes (excluding $DeltaE$) The minus sign for miRNAs is linked to the inhibitory action of miRNAs on the expression of genes which results in a lowering of the regulatory action of the genes and of the total perturbation of the pathway.

There is also a p-value associated with the results. P-value estimation is performed by combining the Z-scores, computed through an inverse Standardized Normal distribution, associated with two probabilistic terms: the first is the probability of obtaining

by chance a number of differentially expressed genes in the pathway at least equal to the observed one, while the second consists of the probability of observing by chance an accumulator higher than the computed one. The first term corresponds to $prob(P_i)$. introduced during the definition of the impact factor 6.2.2. The second term$(prob_1(P_i))$, instead, has to be estimated through a permutation test. A Log-Fold-Change is assigned to a random group of genes in the pathway selected randomly from the input ones, so as to compute a random accumulator. The procedure is repeated several times and the final probability is estimated as the ratio between the number of random accumulators greater than $Acc(P_i)$. and the number of repetitions performed. The final p-value will be a correction with Benjamini-Hochberg [18] and a combination of the two p-values with Stouffer method[146]. The final function is:

$$pvalue(P_i) = \Phi \frac{\Phi^{-1}(prob(P_i)) + \Phi^{-1}(prob_1(P_i))}{\sqrt{2}}$$

6.3 PHENotypes SIMulator

Aside from direct experiments with *in vitro* measurements, there is a need for simulations that can be done directly *in silico* and estimate the behaviour of a biological system.

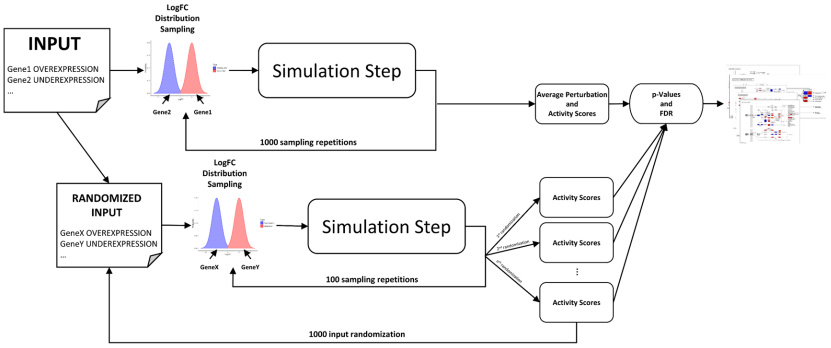


Figure 6.3: PHENSIM workflow

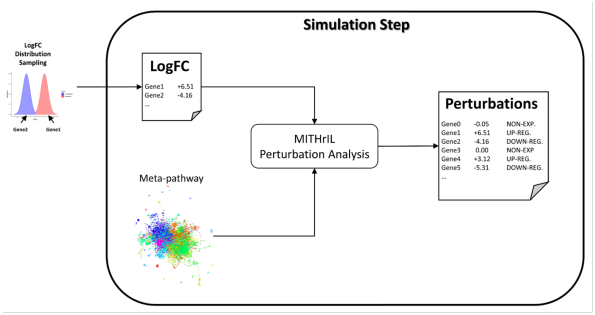


Figure 6.4: PHENSIM simulation step

PHENSIM[6] is a web tool available at https://phensim.tech/ that can simulate the effects of activation/inhibition of one or multiple biomolecules on cell phenotypes by

exploiting signaling pathways. Unlike MITHrIL, the method do not start from quantitative data (log-fold change) but qualitative (over-expressed genes and under-expressed genes with no values), therefore the knowledge required is lower. The dysregulation of each gene of the pathway is measured in terms of the Activity Score. The system is useful to simulate the effect of a drug that targets a gene on one or more pathways. The workflow of PHENSIM starts from a list of genes with the notation of $state \in OVEREXPRESSION, UNDEREXPRESSION, NOTEXPRESSED$ taken as an input and ends up with Activity Scores for all the genes in a pathway, considering the whole metapathway, the final output will be a matrix of Activity scores. The whole workflow and the simulation step of PHENSIM are represented in figure 6.3 and figure 6.4

As it can be seen from 6.4, the simulation step of PHENSIM uses MITHrIL at its core to generate the perturbation factors and impact factors for the generated data. PHENSIM supports three pathway maps for three different organisms(mus musculus, rattus norvegicus, homo sapiens) at the core but it has been expanded in recent years. The basic hypothesis that PHENSIM uses is that phenotypes can be described through changes in pathway activity. PHENSIM is a randomized algorithm for computing the effect of (up/down) de-regulated genes, metabolites, or microRNAs on KEGG pathways. The results are synthesized as an Activity Score estimated for each element in a pathway. The sign of this score indicates the predicted effect on the node: positive for activation, negative for inhibition. Its value represents the log-likelihood of the effect compared to a null model. PHENSIM performs all calculations using the meta-pathway built from all KEGG pathways. These pathways are merged to build a single directed network using a two steps. First, all disease pathways are removed and then all nodes and edges are merged, removing duplicates. The meta-pathway is then annotated with the experimentally validated miRNA-target interactions.

A discrete random variable V_i is associated to each node in the pathway that represents whether a node is activated(1), inhibited (-1) or not altered (0).

PHENSIM summarizes the activity of a node V_i, given the input E, in an activity score, $AS_E(V_i)$. Such a value has a dual function: the sign indicates predicted activity (positive means activation, negative inhibition), the value represents the log-likelihood of such an observation with respect to the null model. Therefore, to compute this value, PHENSIM needs to determine a log-likelihood ratio, $\mathcal{L}(E|V_i = v_i)$, for each possible outcome V_i of a node. Let $Prob(V_i = v_i|E)$ and $Prob(V_i = v_i|E_{null})$, be the probabilities that V_i state is V_i in the input and null model, respectively. The log-likelihood ratio is defined as:

$$\mathcal{L}(E|V_i = v_i) = log(\frac{Prob(V_i = v_i|E)}{Prob(V_i = v_i|E_{null})}) \tag{6.6}$$

The most important part of the algorithm is the simulation step that simulates how the pathways should behave in the case of underexpressed genes/miRNA and overexpressed genes/miRNA(passed as input). Given the set of up/down-regulated genes $E = \{v_{j_1}, v_{j_2}, ..., v_{j_n}\}$, and a function to compute the log-likelihood as defined in 6.6, the simulation step works as follows:

1. Compute the random expression values for all genes that are in the simulation by sampling values of overexpression from a Gaussian distribution N(5,2) and

underexpression from a Gaussian distribution N(-5,2).

2. Execute MITHrIL to compute PF for all genes considered
3. repeat step 1 and 2 T times
4. Compute the probability $Prob(V_i = 1|E)$ as the ratio of the number of iterations where the gene i has resulted positively perturbed and the total number of iterations.
5. Compute the probability $Prob(V_i = -1|E)$ as the ratio of the number of iterations where the gene i has resulted negatively perturbed and the total number of iterations.

The activity score for node V_i can be determined as:

$$AS_E(V_i) = \begin{cases} \mathcal{L}(E|V_i = 1) & \text{if } \mathcal{L}(E|V_i = 1) > max(\mathcal{L}(E|V_i = -1), \mathcal{L}(E|V_i = 0)) \\ -\mathcal{L}(E|V_i = -1) & \text{if } \mathcal{L}(E|V_i = -1) > max(\mathcal{L}(E|V_i = 1), \mathcal{L}(E|V_i = 0)) \\ 0 & \text{if } \mathcal{L}(E|V_i = 0) > max(\mathcal{L}(E|V_i = 1), \mathcal{L}(E|V_i = -1)) \end{cases}$$

There is also a p-value associated with the single AS. The simulation step described previously is also applied in the random model, to derive a p-value of the measured activity score to measure its significance. The p-value is obtained through a permutation test. Let k be the number of genes in the input list of altered genes E. The random model consists of a set of K random variants of the input (default K = 1000). Each random variant R_i is represented by the pathway P in which log-fold changes are calculated on a E_i list of k genes selected at random. In the end, K activation probability, K inhibition probability and consequently K activity score are obtained for each gene i, that is $AS_{E_1}(i), AS_{E_2}(i), ..., AS_{E_k}(i)$. The computation will not be seen in detail.

To see additional information about the methodology of PHENSIM see the original article [6] and visit the site to read the user manual and try to run some simulations.

To account for multi-layer networks (like the one described at the end of the previous chapter 5), PHENSIM should be modified to take into account the interactions from layer to layer while also understanding and personalizing how the layers interact with each other (using a different methodology and activity score used here, and also using different perturbation factors for genes in the MITHrIL simulations).

For this thesis, the differentially expressed genes for every group/type of cells identified will be used as the input for PHENSIM to see what is the behavior of pathways and the metapathway for single cells compared to bulk, where differentially expressed genes are also given as an input to other simulations of PHENSIM. The expected behavior is that activity score should be divided among the cells while they are treated as a whole for bulk data, but this expectation is only idealistic and almost certainly not true due to technical errors, the fact that RNA outside of the cells is not sequenced and the fact that isolation of the cells will change expression values and these problems are not easy to overcome, especially for the last two(consequences of the isolation of the cells). More about the procedure will be seen in 7 and 8

Chapter 7

Methodology and pipeline definition

In this section, the methodology to treat scRNA-seq data will be seen along some pipelines to get some useful results out of single-cell sequencing. Most of the methods seen here are only presented as methodology and will not be seen during experimental analysis 8 since the project is at its end and the deadline is too close to continue experimenting with all the pipelines. Nonetheless, the pipelines presented here remain valid and usable to understand the overall significance of experiments and to reach the objective of one project. Also, the pipelines that will not be seen during experimentation in 8 will be validated after the deadline of this project as well with more and better data at hand(problems with the data used for this research will be seen at the start of the next chapter about experimentation), additional updates about the project will be seen in the repository [73] and will be part of the SCAPE project outside of this thesis.

In this chapter, there will be a division in sections and the type of data used will be **bulk** sequencing data(RNA-seq) and single-cell sequencing data(scRNA-seq) as inputs. There will not be any reference to real data since they will be seen directly during experiments in 8, also references genomes/transcriptomes along with annotation data will be referenced but not used directly In this section for the same reason.

The methodology seen in this section has the purpose of treating scRNA-seq data and computing useful results that can be used for additional analysis. The final objective of the methodology (and of experimentation especially) is to build a pipeline that integrates bulk and single-cell, along with the topology of pathways and meta-pathways.

7.1 Preliminary steps

As preliminary steps, the first things to consider are:

- **Obtaining the data**: This is one of the most important steps. Obtaining the data can follow two main routes that offer different pros and cons:
  - Obtaining the sequencing data from databases like GEO [107], SRA [108] or other databases that contain experiments and research studies(all related to each other if possible, since different datasets are obtained from different platforms and methods, so the data will be inconsistent overall, even if what is declared seems similar between experiments of different research). This approach is best suited for labs that do not have the equipment, personnel or time to do the sequencing experiments but comes with a lot of downsides related to the data and to the time spent on searching and establishing the quality of the data needed for the project, also hidden factors about

the experiments that are not presented in the research and documentation where the data is taken will arise most of the time(these are the problems encountered for this project as well and a lot of time was spent to search for datasets and for accounting for problems with the data that have arisen during computational analysis and quality control);
  - Using the instruments available locally and organizing the experiments and building of the sequence libraries. This option is the best to get better-suited data for the objective of the experiment since one can do whatever is possible to get the best results and the documentation is directly available and created by the researcher. No problems with underlying details that have been hidden or not documented will arise, since the data is created from the experiments needed and no details will be hidden(at least locally). This option is obviously the best since data can be built when needed, but the problems are all related to time and funds related to the project. Platforms for scRNA-seq need to be available or bought along reagents, researchers and lab employees need to be paid to get the data, time needs to be spent for every single sequence library obtained and, when organizing more experiments for the same tissue, the quantity of personnel need to be enough to perform the experiments at the same time with the same conditions and environment to not introduce confounding factors.

  In this research, there was no possibility to get privately produced data since the faculty here in Catania lacks the instruments and platforms to do single-cell sequencing, so the only possibility was to search for sequencing libraries and datasets to use.
- **Control the quality of the data** not from a sequencing perspective(by seeing sequencing depths and the quality reported in FASTQ), but by reading the source of the data used(the documentation of the platform used for sequencing or the research and documentation related to the data obtained from databases). This step is necessary since the quality of the data is dependent on the origin of the data and the methods used to obtain it. Low-quality data is categorized with reference to the objectives of the project. If a project is aimed at getting more depth with sequencing by sacrificing the number of cells sequenced, one needs to use full-length protocols, on the other hand, tag-based protocols are suited for a large number of cells with better quality of sequence(for the presence of UMI, less sequence depth though) libraries.
- **Control the completeness of the data** by seeing if the data obtained has all the things needed for the project itself. For example, a project about the use of differential expression analysis in single cells should have both **controls** and **altered** datasets at its disposal, protocols used should also be the same. The issue of completeness depends on the data used and on the objectives since the data need to have certain characteristics. This step also takes into consideration the **soundness** of the data(having correct data that will result in correct results), if these requirements about the data cannot be satisfied, the project will be meaningless.
- **Control the documentation related to the data** especially when searching

about how to deal with the data, the metadata related, the description of the source of the data (patients, organisms, sequence method used) and confront the results if the research will have some junctions in the methodology used.

The sequencing data used in this project is taken from the research of clonal trees [41], and this data will be from 4 patients of acute lymphoblastic leukemia of the 6 seen in the research(the first 4 patients), the samples are taken from bone marrow tissue of humans. The sequencing data related to that research is obtained from the Fluidigm C1 [39] platform with full-length methods. The data available that will be used for this project are a bulk sequencing library and a certain number of cells for single-cell. The accession number for the data is **SRP044380**. Data for controls was taken from this accession numbers **SRP304033** related also to bone marrow samples.

Before starting with the analysis, other preliminary steps need to be taken into account, for example, the building of the indexes for some methods used for alignment(salmon [115] or BowTie2[85]) and the acquisition of a reference genome-transcriptome along some annotation data. All the data used as reference is directly taken from the Ensembl database [34] and some references will be downloaded directly from here `http://www.ensembl.org/info/data/ftp/index.html` and here for the indexes [35]. For this research, Homo Sapiens reference transcriptome/genome, annotation data and indexes built from this type of organism will be used.

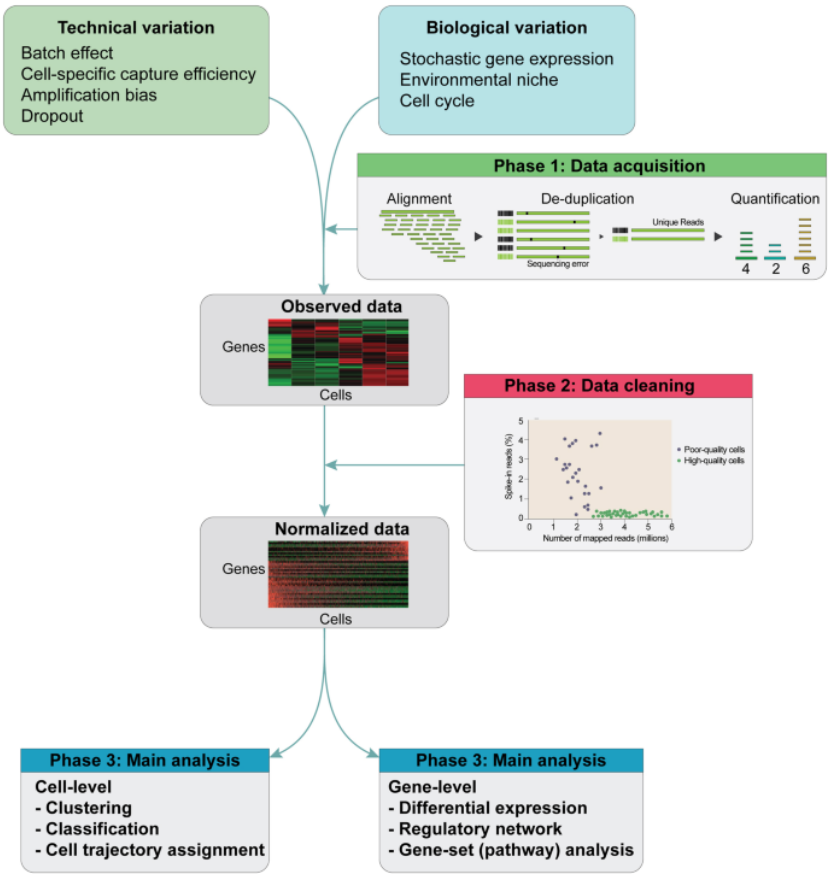


Figure 7.1: scRNA-seq analysis pipelines.

After obtaining the data(with personal sequencing libraries building or with databases), the next step is to use the data and meta-data obtained during the preliminary step and start the computational analysis itself.

7.2 Preprocessing the data

Once reads are obtained from well-designed scRNA-seq experiments, quality control (QC) is performed for both single-cell data and bulk data. Of the existing QC tools available, FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) is a popular tool for inspecting quality distributions across entire reads.

Adapter sequences can be removed at this pre-processing step, although it is not necessary during the alignment or pseudo-alignment, it is advised to do so to maintain a high quality of the alignment by also evading useless noise in the data. Common tools used are in the Biostrings package in R and are also integrated into some alignment tools(not in the ones seen here).

7.2.1 bulk RNA-seq data

For bulk RNA-seq, the process is straightforward and well-documented. The pipelines for pre-processing available are to align the reads to the transcriptome or the genome. The case of alignment to the genome is considered when dealing with DNA-sequencing (not only RNA-seq) and when there is a need to find a variant of a gene(with variant calling as already introduced in 2) for tumor/polymorphisms analysis or to find variant alleles to study phenotype. Tools used to align to the genome are BowTie2[85] Burrows-Wheeler Aligner (BWA). The pipeline that uses the alignment to transcriptome uses data obtained from RNA-seq and aligns the reads to a reference transcriptome

7.2.2 scRNA-seq data

Read alignment is the next step of the scRNA-seq analysis, and the tools available for this procedure, including the HISAT2 and STAR, are the same as those used in the bulk RNA-seq analysis pipeline. When UMIs are implemented, these sequences should be trimmed prior to alignment(also part of the adapter removal but it is another kind of removal since UMI are not common to all reads but only a very small subset of reads). A common tool used to do UMI removal is UMI-tools [141]. The RNA-seQC program provides post-alignment summary stats, such as uniquely mapped reads, reads mapped to annotated exonic regions, and coverage patterns associated with specific library preparation protocols. When adding transcripts of known quantity and sequence (external spike-ins) for calibration and QC, a low-mapping ratio of endogenous RNA to spike-ins would be an indication of a low-quality library caused by RNA degradation or inefficiently lysed cells. A schematic overview of the single-cell analysis pipeline is described in figure 7.1

Also, the pipeline previously seen is already implemented in Cell Ranger[1], along barcode demultiplexing, UMI identification, alignment and read counting.

scRNA-seq data are inherently noisy with confounding factors, such as technical and biological variables. After sequencing, alignment and de-duplication are performed to quantify an initial gene expression profile matrix. Next, normalization is performed with raw expression data using various statistical methods. Additional QC can be performed when using spike-ins by inspecting the mapping ratio to discard low-quality cells. Finally, the normalized matrix is then subjected to main analysis through clustering of cells to identify subtypes. Cell trajectories can be inferred based on these data and by detecting differentially expressed genes between clusters.

After alignment, reads are allocated to exonic, intronic, or intergenic features using transcript annotation in General Transcript Format. Only reads that map to exonic loci with high mapping quality are considered for the generation of the gene expression matrix (N (cells)×m (genes)). A distinctive feature of scRNA-seq data is the presence of zero-inflated counts due to reasons such as dropout or transient gene expression. To account for this feature, normalization must be performed; normalization is necessary to remove cell-specific bias, which can affect downstream applications (e.g., determination of differential gene expression).

Normalization is also done on the data depending on what are the objectives. For simple analyses that do not rely to much on differential analysis, a classic approach should be enough(TPM,RPKM or FPKM), but when considering groups of cells with different sizes and transcript sizes, new methods should be used. The read count for a gene in each cell is expected to be proportional to the gene-specific expression level and cell-specific scaling factors (random). These nuisance variables, including capture and reverse transcription efficiency and cell-intrinsic factors, are usually difficult to estimate and are thus typically modeled as fixed factors. Although nuisance variables can be jointly estimated with expression counts for normalization, fits are made to only a particular statistical model, and the procedure is computationally demanding. In practice, raw expression counts are normalized using scaling factor estimates by standardizing across cells, assuming that most genes are not differentially expressed.

Another thing to consider during preliminary data preparation is the filtering of the genes that are not expressed or that are lesser than a threshold. For this project, every single gene that is not expressed will be discarded and marked as *Not-expressed* during downstream analysis (for PHENSIM). This is done for both bulk and single-cell.

The final results of this step are count matrices for every patient, where every row is a gene and every row is an isolated cell.

For this project, alignment and read counting were done directly on the data with salmon [115], every sample for bulk was quantified using salmon with the index, reference transcriptome and annotation data specified in the previous section. For single-cell, the data was demultiplexed(At the start of the project, after some time it has been noticed that there was already demultiplexed data so there was no need to demultiplex as the data was already available in single fastq files isolated by the single-cell), and salmon was applied to get the quantification of the transcripts.

The final results for the single-cell pipeline were joined together in a matrix while the bulk quantification results remained vectors of values(since one of the objectives of this project is to use single patients to get some useful data without using other patients in the same group).

As already seen previously the acquisition of the count matrices, normalization needs to be done on the data before the true analysis starts. The method used in this research is RPKM.

Another thing to take into account is the use of single-cell imputation for estimating the expression values of dropout and sparse genes. The use of imputation is not considered much in this research since the methods for imputation seen in previous chapters were not convincing enough and should be treated properly when used since they introduce mock information in the data that could be not consistent to the cell itself.

7.3 Clustering and grouping of cells

Clustering and grouping for scRNA-seq is useful to find relations and association rules for groups of cells and the groups found can be assigned a type or some characteristics accordingly. The clustering and grouping of cells is not a necessary step (aside from annotation of the cells group which is a necessary step to get useful information about the cells themselves) since the other steps could not need any clusters to get information or to do additional analysis.

This step and the annotation of the cell groups could use **dimensionality reduction** to get rid of the noise and compact the data. Dimensionality reduction is used both for the visualization of the cells and also for clustering and, optionally, annotation of cell groups. As already said previously, computational steps should use a higher dimensionality than what is used for visualization, so clustering and annotation steps could use higher dimensions for dimensionality reduction.

For this project, UMAP for dimensionality reduction and a graph-based clustering approach [134] is used to create clusters that will be used for annotation in the next step. All the steps mentioned here are already implemented in Seurat[131]. The code used for Seurat is a modification of the vignette available here `https://satijalab.org/seurat/articles/integration_introduction.html`. More about the function used and the results obtained will be seen in 8

7.4 Annotation of the cell groups

Cell mapping and annotation is not implemented directly in Seurat (although methods exist to map and annotate the cells based upon their markers they are mainly reference-based and done manually, an example vignette is `https://satijalab.org/seurat/articles/integration_mapping.html`). To annotate the data with a more precise and complex methodology, singleR[137] is used. The Bioconductor package SingleR implements an automatic annotation method that, given a reference dataset of samples (single-cell or bulk) with known labels, it assigns those labels to new cells from a test dataset based on similarities in their expression profiles. SingleR is a very good tool since it also implements **pseudobulk**

After the annotation of cell groups, the information obtained could be validated with bulk deconvolution to see if the cell population estimated is somewhat similar to

the annotation of the single-cell groups.

7.5 Differential analysis

Differential analysis covers the position of a middle step between single-cell analysis and biological network analysis in the pipeline that will be defined here. Differential expression analysis is usually used to understand the differences between two or more groups of samples, but in single-cell, the sample is complex and composed of different cells.

For this project, differential analysis is done on the controls and the ALL samples for bulk while, since the research for single-cell controls was not successful at all (no bone marrow single-cell samples were found in 2 months of searching for the data), the differential expression for single-cells will be done between groups of identified cell types.

7.5.1 Differential expression analysis with bulk data

Differential expression analysis follows what was already seen in 2.4, nothing much was added since this type of analysis is already well documented and usable in many ways.

For this project, the data used is the same defined at the start of this chapter, so expression values from bone marrow tissue for controls and ALL patients, differential expression was done for single patients without any biological replicates(by using the other patients of ALL) or technical replicates at first. After some preliminary tests, technical replicates were introduced to be more consistent with the methodology. The packages used were DESeq2 [97] and edgeR [126], but the results were very similar so only DESeq2 results will be presented in the next chapter.

7.5.2 Differential expression analysis with single-cell data

As already introduced in 3.4, differential expression can be done in two ways for scRNA-seq workflows:

- Identify the differentially expressed genes within a cell type/group by the comparison of two or more sample sources(altered samples and controls, or something similar) that have the same cell group/type associated with them.
- Identify the differentially expressed genes between cell groups by the computation of pairwise differentially expressed genes between groups. This approach was discussed during 3.4 and some methodologies were seen that take into account the union of all the differentially expressed genes between groups(that results in wider results) or the intersection of the differentially expressed genes between groups for a single group treated as a control. Union will be used for this project.
- Use a mixed/hybrid approach and combine both differential analysis cell-cell with the same patient and between different groups, while also taking into account differential analysis for different patients and cells within the same group(treating

the cell in the same group or with the same type as single samples if not using aggregation/pseudobulk).

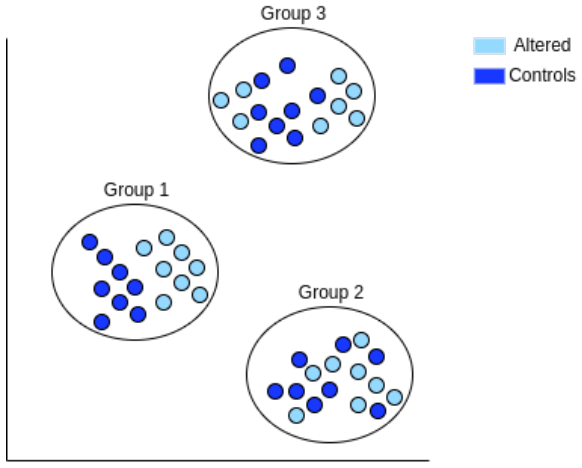


Figure 7.2: **Visualization of the data used for differential analysis**

An image to visualize an example of data where different cells with different sources (controls and altered) belong to the same group can be seen in figure 7.2, where the cells are plotted in a 2-dimensional graph(suppose that there are only 2 cells with no dimensionality reduction, so the final types associated with the cells will respect the clusteroids formed in the graph). In that image, if there is no availability for control cells(as it was with this project), the only possibility is to do differential expression analysis between groups of cells to identify the differentially expressed genes with a fixed group in reference to the other cells.

7.6 Pathway embedding

For pathway embedding, the path that could be taken was the one introduced in 6. Since research for GNN applied to expression values(and not differentially expressed genes) and meta-pathways was already done in the past[95], this research will focus only on pathway embedding and analysis with instruments such as MITHrIL[5] and PHENSIM[6] especially.

MITHrIL can be used to evaluate the dysregulation of a pathway by feeding to the algorithm the differentially expressed genes along their log-fold-change and the algorithm will return **Perturbation factors** for every gene in the pathway and an overall score (**Impact factor**) for the dysregulation of the pathway (along the **accumulation** factor that measures the strength for the total perturbation and a p-value for it to validate the possibility of getting a similar accumulation factor with a random model). The results obtained can be used to analyze different responses for different patients and groups and cluster these responses(vectors of genes Perturbation factors when considering single pathways or vectors of Impact factor/accumulation when considering a metapathway and its sub-graphs that are pathways).

On the other hand, **PHENSIM** takes only the list of differentially expressed genes as defined in 6.3. The tool will simulate the effect of the perturbation of these genes in the pathways/metapathway and will return a list of **Activity scores** for every single gene in every pathway.

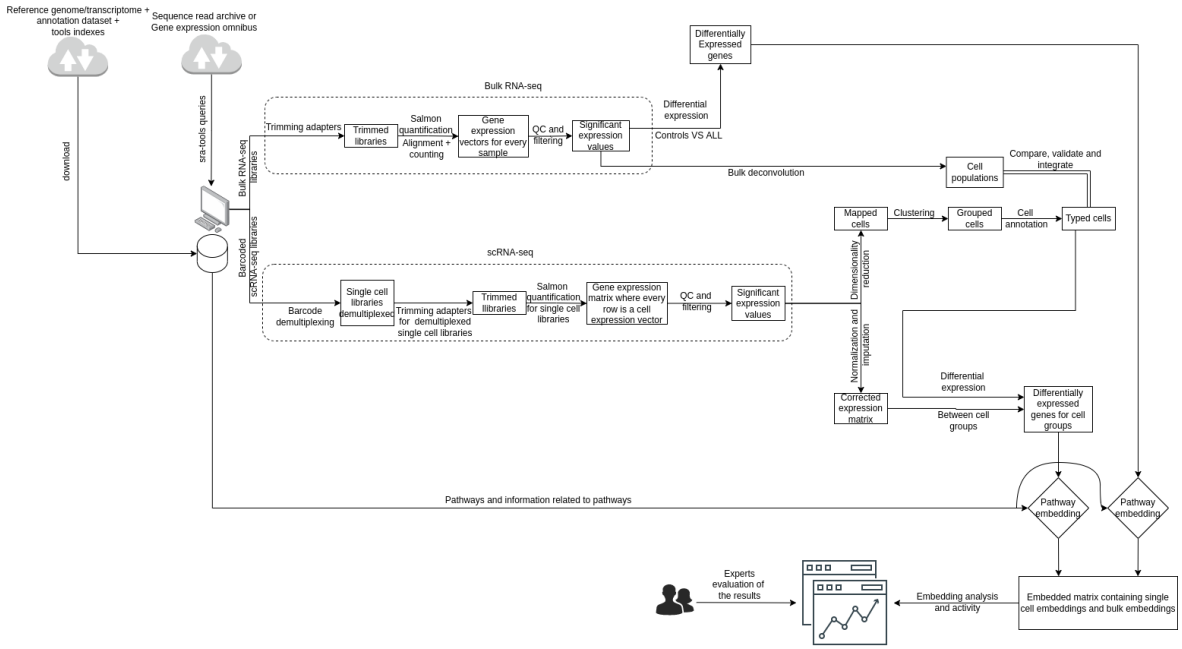


Figure 7.3: Generalized pipeline of SCAPE

For this project, only PHENSIM will be used and a matrix of activity scores (*genes* X *pathways*) will be returned. This matrix can be seen by experts in the sector of medicine and biology to see if the activity scores obtained in a pathway are abnormal or present some useful characteristics. The data used as the input for PHENSIM will be differentially expressed genes identified for single groups/types of cells and differentially expressed genes identified for bulk as defined in 7.5.

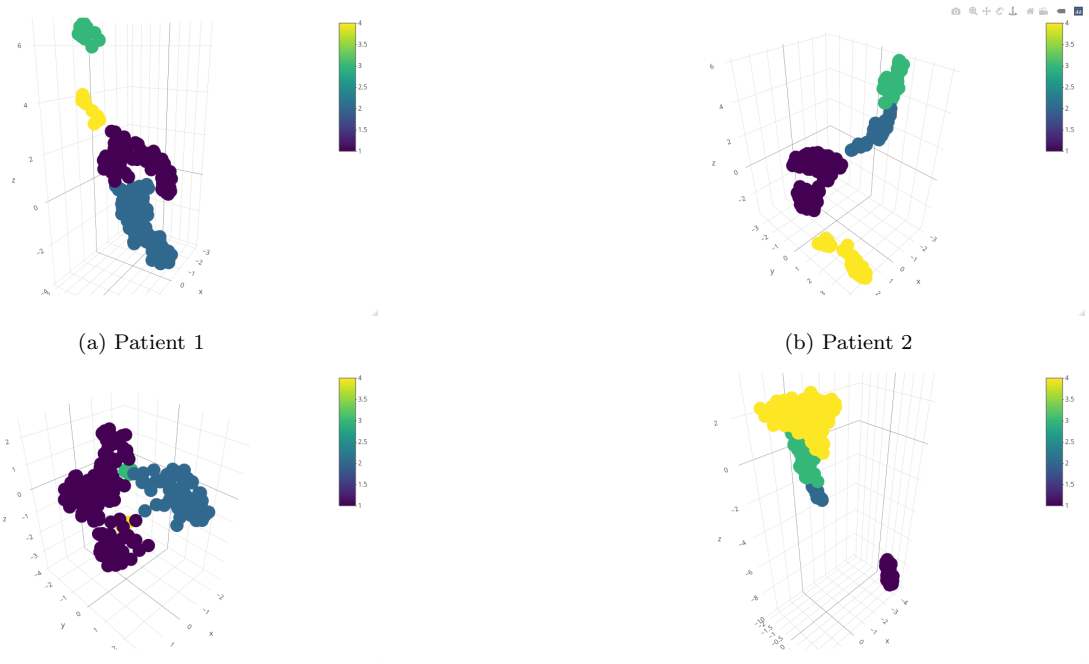
The full generalized pipeline for this project(without the mentions of the methods used to maintain flexibility in workflow definition) can be seen in figure 7.3

Chapter 8

Experimental analysis

Following the methodology described in the previous chapter, 4 acute lymphoblastic leukemia patients will be analyzed and compared. No controls will be used for scRNA-seq.

The analysis done for this project was organized first without using tools like Seurat[131], and the results of dimensionality reduction and clustering are seen in 8.1. To obtain these umapped cells and groups for every patient, normalization and filtering steps were done prior to dimensionality reduction. After that, dimensionality reduction was made with UMAP and the dimensions of the cell vectors(of expression values) were reduced to 3 to facilitate visualization of the results. The results presented here will be not that different from what will be shown in figure 8.2 since the results of Seurat are very similar to the workflow followed here. Aside from that, the methodology presented in 7 is applicable to both cases (for user-defined pipelines or Seurat workflows).



(a) Patient 1

(b) Patient 2

(c) Patient 3

(d) Patient 4

Figure 8.1: **Cell groups plotted with UMAP dimensionality reduction**: Groups are built from UMAP features(from all the expression values of genes available) in 5 dimensions and hierarchical clustering. This visualization is only a visualization of the data, these groups are similar but not the same as the one obtained from Seurat

As it can be seen in 8.1, almost all of the patients' cells seem to be clustered together with no common structures. The only patient that shows some variability in the cell expression values is patient 4 where two clusters are significantly distant from each other and have enough density. These results will be the same for the methods used in Seurat.

The clustered groups identified in Seurat can be seen in figure 8.2 for all the patients chosen for this project. All the groups seem to be sparse and do not form any visible cluster in the 2D UMAP space, but for patient 4, some visible clusters are evident

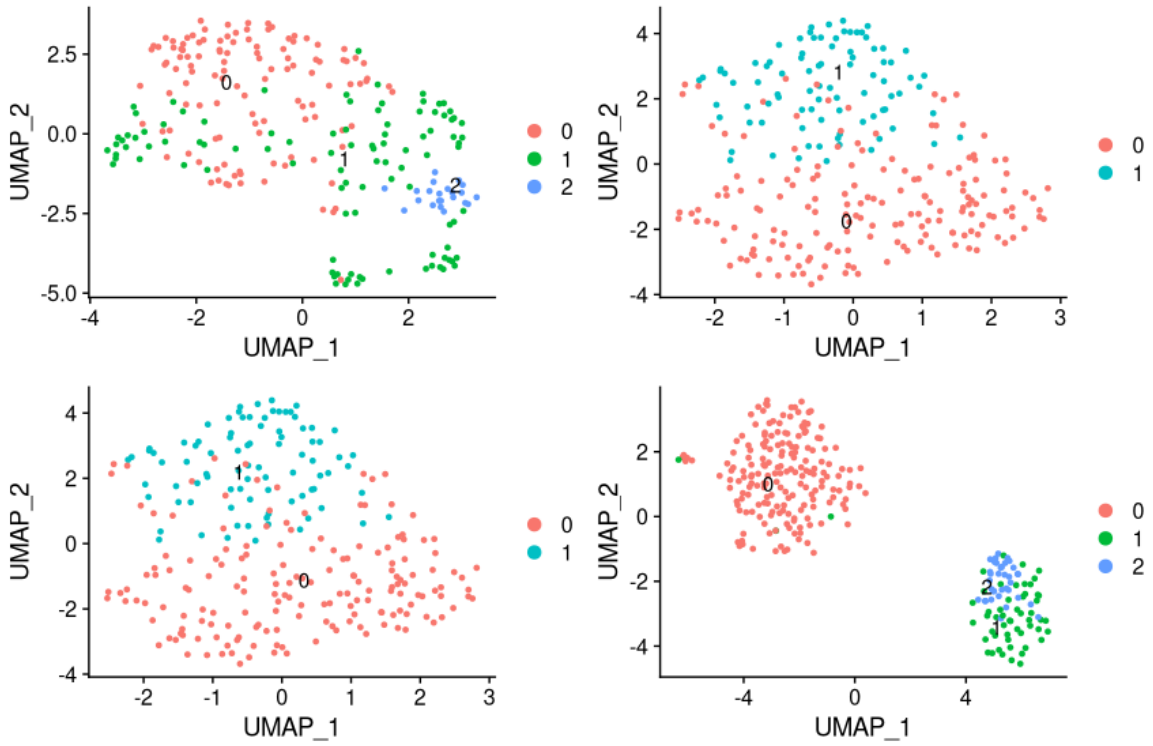


Figure 8.2: UMAP mapped cells clustered in more dimensions and plotted in 2D

Cell identification is done via SingleR [137] with the use of aggregation(pseudobulk) and the clusters found previously with Seurat since the default run of SingleR was returning too many cell groups that seemed really close together. The final results can be seen in figure 8.3. The cell types assigned to the clusters are:

- iPS(Induced pluripotent stem) cells, a type of pluripotent stem cell.
- fibroblasts, a type of cell that synthesizes the extracellular matrix and collagen. These types of cells produce the structural framework (stroma) for animal tissues and play a critical role in wound healing.
- Monocyte-derived Dendritic cells(DC monocytes) are a distinct DC subset, involved in inflammation and infection, they originate from monocytes upon stimulation in the circulation and their activation and function may vary in autoimmune diseases.

All the cells identified seem to be related to the tissue where they were taken (bone

marrow, where fibroblasts especially and iPS cells are present [160]) and the condition of ALL(for the dendritic cells, as shown in various studies [155]), the results obtained remain unsure and the data is not enough to get some real information about the cells.

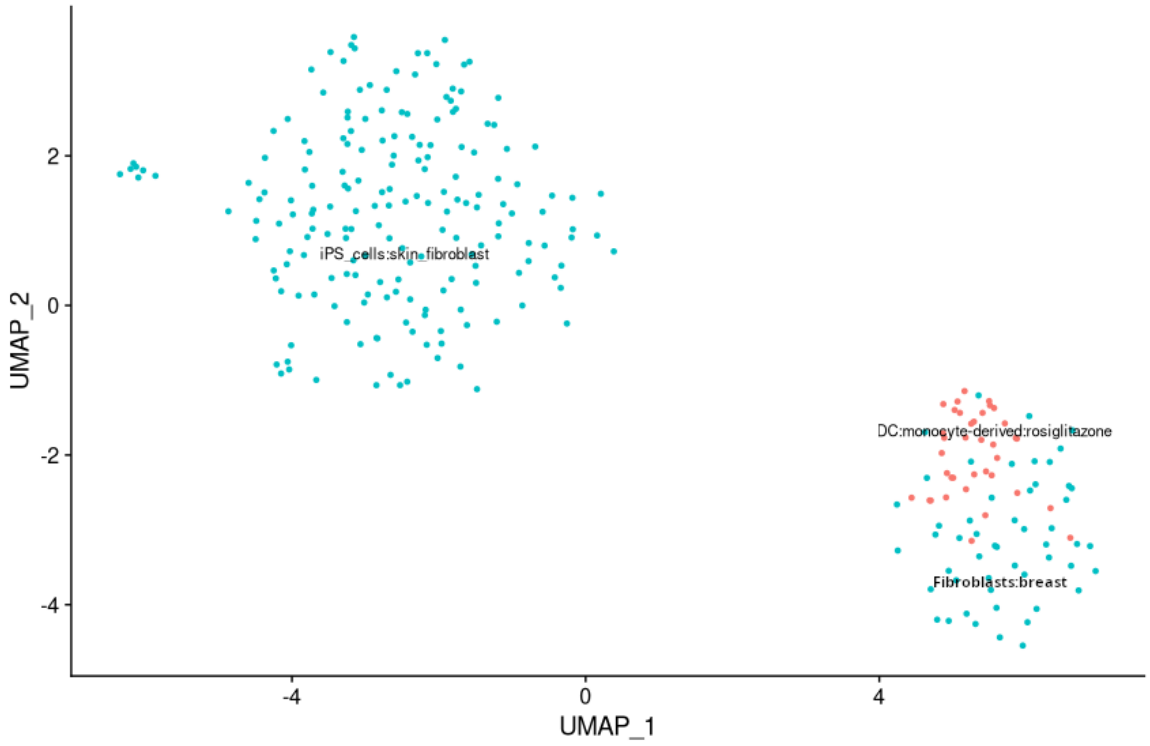


Figure 8.3: Cell types associated to the cell groups in patient 4

To validate the results obtained, the bulk sequencing data associated with the patient was deconvoluted with **granulator** that is a sub-module of Bioconductor in R to estimate the population of cells. The results obtained show that the majority of cell frequencies found are not DC monocytes and iPS cells(although fibroblasts are estimated in enough quantities). This means that the single-cell data used is not consistent with the bulk sample associated with the patient.

Experimental results for PHENSIM are not significant since the cell types found were not too many (considering the fact that this project used the clusters to maintain consistency in the project) and the results were not enough to prove anything, so the results were not reported here. In the future, with some additional and more complete datasets and experiments, there could be a conclusion to this project but, for now, the final results of the pipeline defined in 7 are inconclusive and cannot be used.

In the case when the analysis is done on complete and detailed data that will result in better results, the outputs can be analyzed and seen by experts in the sector of biology and medicine to see if the activity of a cell in an altered tissue is somewhat significant to a reference or is extremely unique. On the other hand, the activity scores generated can be used as embeddings for pathways and could be compared with other embeddings of other tissues or of other cells.

Chapter 9

Conclusion

Since the sequencing data obtained from scRNA-seq is a snapshot of a single moment in time for the state of a tissue and its cells, the pathways can be modeled as **Temporal networks** and the algorithms that were seen in 6 need to be adapted and also discarded for new algorithms that take into account the temporal features of the graphs resulting from pathways (and the meta-pathway as well).

The nomenclature used in biological experiments and in biological experiments and, in particular, medical analysis is somewhat ambiguous and difficult to find sometimes. For example, the "control" samples that are used for differential analysis in 2.4 and 3.4, while in literature and publications these controls are not called controls but non-"disease" where *disease* is the object of the study, or normal. I have tried to standardize the concept for differential expression analysis with the word "target" since during differential expression analysis, these samples used are used as **target** samples that describe how the "disease" sample is different from a normal case study. It is really important to standardize a way of using these terms since they are used for queries and finding other datasets and research about these topics. I had a lot of problems finding datasets for reads, bulk datasets, single-cell datasets and a lot more since the nomenclature is sometimes interchangeable while some people use a non-conventional way of naming things in their research. Another example of this is the "deconvolution" seen as an additional method for bulk RNA-seq in 2.5, the word is also used to reach some conclusions and transform the data of single-cell RNA-seq as well, but the correct term, when used for single-cell data, should be "convolution" since the results depend on the cells that will be convoluted into a piece of single information or latent feature, while bulk deconvolution is a way of estimating the population of cells, obtaining more information about the composition of the sample and deconvolving the bulk results into its components.

One of the problems here is a direct consequence of one of the objectives of this project, which is to personalize a treatment from unique information for a single patient. The analysis is centered around the single patient as a single entity so the final results obtained will be insignificant compared to group analysis and comparison since groups with the same characteristic (e.g. tumor patients) can be analyzed statistically and significant conclusions can be extracted.

As it was seen during chapter 7 and chapter 8, the data used for this project was not well-formed and inconsistent with the objectives of the project. After the compilation of this thesis, more recent, well-formed datasets suited for the objective of the project will be used and expanded to obtain significant results aside from the initial presentation

done in this thesis. Two sequencing libraries and datasets will be used to continue the project:

- [167] already mentioned previously since the data is tightly coupled with multi-omics and bulk/single-cell datasets are well documented.
- [44] is also well documented and a different field of study since bulk samples are homogeneous and the methodologies used are up to date.

Integration of spatial transcriptomics will be considered to expand the projects with additional information that will enrich the network of pathways or co-expression, in particular, the multi-dimensional meta-pathway described at the end of chapter 5 can be used in addition to spatial transcriptomics to model the whole tissue and simulate behaviour with methods similar to the ones seen at the end of chapter 6.

In conclusion, scRNA-seq is the frontier of analysis and it is quickly becoming one of the most useful tool in phenotyping and transcriptome analysis, so it should be considered as another very useful instrument to use (integrated with other instruments like spatial transcriptomics and bulk RNA-seq) for getting more information and accuracy in the transcriptome and for the cell population in a sample. Cell atlas (`https://data.humancellatlas.org/`) are becoming more and more important than ever and are slowly entering the state where they can be used as trustworthy sources of data since the data that they provide can be used as direct reference when doing experimentation, while for this project there were a lot of problems regarding the data.

The future of single-cell sequencing is bright and new technologies are rising day by day along methodologies and algorithms used to tackle big questions. The research related to this project will not finish here and will continue, all the future updates of the project will be listed in the repository of **SCAPE** [73] while future articles will reference this project since it was my first look at single-cell sequencing (and in particular for the data analysis done on scRNA-seq databases) and I am grateful for having discovered this field along all the associated fields, all for a better future.

Acronyms

ALL acute lymphoblastic leukemia. 123, 127, 130, 132

ddNTPs di-deoxynucleotide triphosphates. 13

LNA locked nucleic acid. 54
lncRNA long non-coding RNA. 41, 52

miRNA micro RNA. 11, 52, 116

NGS Next Generation Sequencing. 5

scRNA-seq Single cell RNA sequencing. 4, 6, 7, 9, 11, 28, 31, 33, 39–42, 47, 49, 50, 52, 55, 58, 61–64, 66, 68, 69, 72, 77, 78, 80, 83–85, 88, 99, 105, 121, 122, 126, 127, 130, 133, 134
siRNA small interfering RNA. 11
SNP single nucleotide polimorphism. 85
snRNA small nuclear RNA. 11
snRNA-seq Single nucleus RNA sequencing. 88
SNV single nucleotide variant. 85

TSO Template Switching Oligos. 9, 44, 53, 54

UMI Unique molecular identifier. 10, 41–43, 59

VAF variant allele frequencies. 85, 86

Bibliography

[1] 10x genomics. 10x genomics cell ranger software. https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger, 2022. [Online; accessed 10-August-2022].

[2] 10x genomics. 10x genomics chromium gem user guide. https://assets.ctfassets.net/an68im79xiti/1C16trEdzy1Folq5xbOijE/7e6fb1f504e130bd561d898384da99d9/CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex__RevB.pdf, 2022. [Online; accessed 10-August-2022].

[3] 10x genomics. 10x genomics chromium platform. https://www.10xgenomics.com/instruments/chromium-controller, 2022. [Online; accessed 10-August-2022].

[4] 10x genomics. 10x genomics site. https://www.10xgenomics.com/, 2022. [Online; accessed 10-August-2022].

[5] Salvatore Alaimo, Rosalba Giugno, Mario Acunzo, Dario Veneziano, Alfredo Ferro, and Alfredo Pulvirenti. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget*, 7(34):54572, 2016.

[6] Salvatore Alaimo, Rosaria Valentina Rapicavoli, Gioacchino P Marceca, Alessandro La Ferlita, Oksana B Serebrennikova, Philip N Tsichlis, Bud Mishra, Alfredo Pulvirenti, and Alfredo Ferro. Phensim: phenotype simulator. *PLoS computational biology*, 17(6):e1009069, 2021.

[7] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[8] Akram Alyass, Michelle Turcotte, and David Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics*, 8(1):1–12, 2015.

[9] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *bioinformatics*, 31(2):166–169, 2015.

[10] Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.

[11] Wilhelm J Ansorge. Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203, 2009.

[12] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[13] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[14] Jaber Aslanzadeh. Preventing pcr amplification carryover contamination in a clinical laboratory. *Annals of Clinical & Laboratory Science*, 34(4):389–396, 2004.

[15] Johannes W Bagnoli, Christoph Ziegenhain, Aleksandar Janjic, Lucas E Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann, and Wolfgang Enard. mcscrb-seq: sensitive and powerful single-cell rna sequencing. *BioRxiv*, page 188367, 2017.

[16] Lars Barquist and Joerg Vogel. Accelerating discovery and functional analysis of small rnas with new technologies. *Annual review of genetics*, 49:367–394, 2015.

[17] Sarka Benesova, Mikael Kubista, and Lukas Valihrach. Small rna-sequencing: Approaches and considerations for mirna analysis. *Diagnostics*, 11(6):964, 2021.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[19] Jan Berkhout, Frank J Bruggeman, and Bas Teusink. Optimality principles in the regulation of metabolic networks. *Metabolites*, 2(3):529–552, 2012.

[20] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

[21] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[22] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

[23] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, 17(6):1009–1023, 2016.

[24] Jiayi Chen, Yuqin Yao, Xiaolan Su, Ying Shi, Xuejiao Song, Linshen Xie, Jia You, Liantian Tian, Luo Yang, Aiping Fang, et al. Comparative rna-seq transcriptome analysis on silica induced pulmonary inflammation and fibrosis in mice silicosis model. *Journal of Applied Toxicology*, 38(5):773–782, 2018.

[25] Xi Chen, Sarah A Teichmann, and Kerstin B Meyer. From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science*, 1:29–51, 2018.

[26] Biswanath Chowdhury and Gautam Garai. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5-6):419–431, 2017.

[27] cole trapnell lab. monocle 3 repository and documentation. https://cole-trapnell-lab.github.io/monocle3/, 2022. [Online; accessed 10-August-2022].

[28] Cytiva. Sequencing library quantitation for normalization. https://www.cytivalifesciences.com/en/us/news-center/dna-library-normalization-in-ngs-10001, 2022. [Online; accessed 6-August-2022].

[29] A Danilov, Yu Ivanov, R Pryamonosov, and Yu Vassilevski. Methods of graph network reconstruction in personalized medicine. *International journal for numerical methods in biomedical engineering*, 32(8):e02754, 2016.

[30] CSIRO's Data61. Stellargraph machine learning library. https://github.com/stellargraph/stellargraph, 2018.

[31] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'eustachio, Carl Schaefer, Joanne Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

[32] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[33] Holger K Eltzschig, Thomas Weissmüller, Alice Mager, and Tobias Eckle. Nucleotide metabolism and cell-cell interactions. *Cell-Cell Interactions*, pages 73–87, 2006.

[34] Ensembl. Ensembl site. "https://www.ensembl.org/index.html".

[35] ensembl. Indexes for alignment and quantifying tools. "http://refgenomes.databio.org/v3/genomes/splash/2230c535660fb4774114bfa966a62f823fdb6d21acf138d4".

[36] Li Fang, Yunjin Li, Lu Ma, Qiyue Xu, Fei Tan, and Geng Chen. Grndb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic acids research*, 49(D1):D97–D103, 2021.

[37] Omid R Faridani, Ilgar Abdullayev, Michael Hagemann-Jensen, John P Schell, Fredrik Lanner, and Rickard Sandberg. Single-cell sequencing of the small-rna transcriptome. *Nature biotechnology*, 34(12):1264–1266, 2016.

[38] Da-Fei Feng and Russell F Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of molecular evolution*, 25(4):351–360, 1987.

[39] Fluidigm. Fluidigm c1 platform. https://www.fluidigm.com/products-services/instruments/c1, 2022. [Online; accessed 10-August-2022].

[40] Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications*, 1, 2014.

[41] Charles Gawad, Winston Koh, and Stephen R Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, 111(50):17947–17952, 2014.

[42] Katerina AB Gawronski and Junhyong Kim. Single cell transcriptomics of non-coding rnas and their cell-specificity. *Wiley Interdisciplinary Reviews: RNA*, 8(6):e1433, 2017.

[43] Genome.gov. Point mutation glossary. https://www.genome.gov/genetics-glossary/Point-Mutation, 2022. [Online; accessed 10-August-2022].

[44] GEO. Gene expression profiling of sars-cov-1/2 infected human cell lines at bulk and single-cell level dataset. "https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148729".

[45] Todd M Gierahn, Marc H Wadsworth, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nature methods*, 14(4):395–398, 2017.

[46] Rosalba Giugno, Alfredo Pulvirenti, Luciano Cascione, Giuseppe Pigola, and Alfredo Ferro. Midclass: Microarray data classification by association rules and gene expression intervals. *PloS one*, 8(8):e69873, 2013.

[47] GNRdb. Gene regulatory network database. http://www.grndb.com/, 2022. [Online; accessed 10-August-2022].

[48] Laura H Goetz and Nicholas J Schork. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963, 2018.

[49] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[50] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton JM Larsson, Omid R Faridani, and Rickard Sandberg. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology*, 38(6):708–714, 2020.

[51] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.

[52] Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron De Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, et al. Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome biology*, 17(1):1–7, 2016.

[53] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[54] Tetsutaro Hayashi, Haruka Ozaki, Yohei Sasagawa, Mana Umeda, Hiroki Danno, and Itoshi Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, 9(1):1–16, 2018.

[55] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

[56] Desmond G Higgins and Paul M Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, 1988.

[57] Wieteke AM Hoeijmakers, Richárd Bártfai, Kees-Jan Françoijs, and Hendrik G Stunnenberg. Linear amplification for deep sequencing. *Nature protocols*, 6(7):1026–1036, 2011.

[58] Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. Rna sequencing: new technologies and applications in cancer research. *Journal of hematology & oncology*, 13(1):1–16, 2020.

[59] Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. Rna sequencing: new technologies and applications in cancer research. *Journal of hematology & oncology*, 13(1):1–16, 2020.

[60] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks. A systematic evaluation of single-cell rna-sequencing imputation methods. *Genome biology*, 21(1):1–30, 2020.

[61] htseq. Htseq documentation. https://htseq.readthedocs.io/en/release_0.11.1/count.html, 2022. [Online; accessed 10-August-2022].

[62] Che-Lun Hung, Yu-Shiang Lin, Chun-Yuan Lin, Yeh-Ching Chung, and Yi-Fang Chung. Cuda clustalw: An efficient parallel algorithm for progressive multiple sequence alignment on multi-gpus. *Computational biology and chemistry*, 58:62–68, 2015.

[63] Illumina. Illumina tagmentation. https://emea.illumina.com/techniques/sequencing/ngs-library-prep/tagmentation.html, 2022. [Online; accessed 6-August-2022].

[64] Illumina. Sequencing platform illumina miseq. https://emea.illumina.com/systems/sequencing-platforms/miseq.html, 2022. [Online; accessed 6-August-2022].

[65] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15, 2003.

[66] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Highly multiplexed and strand-specific single-cell rna 5' end sequencing. *Nature protocols*, 7(5):813–828, 2012.

[67] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

[68] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

[69] Sergey Ivanov and Evgeny Burnaev. Anonymous walk embeddings. In *International conference on machine learning*, pages 2186–2195. PMLR, 2018.

[70] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):1–17, 2016.

[71] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[72] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1):1–11, 2020.

[73] Josura. Scape repository. https://github.com/josura/SCAPE, 2022. [Online; accessed 10-August-2022].

[74] Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. Starsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus rna-seq data. *Biorxiv*, 2021.

[75] KEGG. Kyoto encyclopedia of genes and genomes. https://www.genome.jp/kegg/, 2022. [Online; accessed 10-August-2022].

[76] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[77] Thomas Kipf. Graph convolutional neural network. "https://tkipf.github.io/graph-convolutional-networks/".

[78] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[79] Martin Kircher and Janet Kelso. High-throughput dna sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

[80] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[81] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[82] Manu P Kumar, Jinyan Du, Georgia Lagoudas, Yang Jiao, Andrew Sawyer, Daryl C Drummond, Douglas A Lauffenburger, and Andreas Raue. Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics. *Cell reports*, 25(6):1458–1468, 2018.

[83] Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14, 2016.

[84] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

[85] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[86] Lexogen. Lexogen rna selection. https://www.lexogen.com/polya-rna-selection-kit/, 2022. [Online; accessed 6-August-2022].

[87] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1–16, 2011.

[88] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.

[89] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):1–9, 2018.

[90] Yunjin Li, Jiwei Chen, Qiyue Xu, Zebei Han, Fei Tan, Tieliu Shi, and Geng Chen. Single-cell transcriptomic analysis reveals dynamic alternative splicing and gene regulatory networks among pancreatic islets. *Science China. Life Sciences*, 64(1):174–176, 2021.

[91] Wen-Ling Liao and Fuu-Jen Tsai. Personalized medicine: A paradigm shift in healthcare. *BioMedicine*, 3(2):66–72, 2013.

[92] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[93] Shiyi Liu, Zitao Wang, Ronghui Zhu, Feiyan Wang, Yanxiang Cheng, and Yeqiang Liu. Three differential expression analysis methods for rna sequencing: limma, edger, deseq2. *Journal of Visualized Experiments*, 175, 2021.

[94] Giorgio Locicero. Temporal networks and applications.

[95] Giorgio Locicero. Bioinformatics: Genes and pathway embedding with graph neural networks, 2022.

[96] Giorgio Locicero, Giovanni Micale, Alfredo Pulvirenti, and Alfredo Ferro. Temporalri: a subgraph isomorphism algorithm for temporal networks. In *International Conference on Complex Networks and Their Applications*, pages 675–687. Springer, 2020.

[97] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[98] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[99] MAGIC. Magic example. http://htmlpreview.github.io/?https://github.com/KrishnaswamyLab/MAGIC/blob/master/Rmagic/inst/examples/bonemarrow_tutorial.html, 2022. [Online; accessed 10-August-2022].

[100] Salem Malikic, Katharina Jahn, Jack Kuipers, S Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):1–12, 2019.

[101] S Manfredo Vieira, M Hiltensperger, V Kumar, D Zegarra-Ruiz, C Dehner, N Khan, FRC Costa, E Tiniakou, T Greiling, W Ruff, et al. Translocation of a gut pathobiont drives autoimmunity in mice and humans. *Science*, 359(6380):1156–1161, 2018.

[102] Elaine R Mardis. Next-generation dna sequencing methods. *Annual review of genomics and human genetics*, 9(1):387–402, 2008.

[103] Kevin P McCormick, Matthew R Willmann, and Blake C Meyers. Experimental design, preprocessing, normalization and differential expression analysis of small rna sequencing experiments. *Silence*, 2(1):1–19, 2011.

[104] Andrew McDavid, Greg Finak, and Masanao Yajima. Mast: model-based analysis of single-cell transcriptomics. *Genome Biol*, 16:278, 2015.

[105] Farid Rashidi Mehrabadi, Kerrie L Marie, Eva Pérez-Guijarro, Salem Malikić, Erfan Sadeqi Azer, Howard H Yang, Can Kızılkale, Charli Gruen, Welles Robinson, Huaitian Liu, et al. Profiles of expressed mutations in single cells reveal subclonal expansion patterns and therapeutic impact of intratumor heterogeneity. *bioRxiv*, 2021.

[106] Giovanni Micale, Giorgio Locicero, Alfredo Pulvirenti, and Alfredo Ferro. Temporalri: subgraph isomorphism in temporal networks with multiple contacts. *Applied Network Science*, 6(1):1–22, 2021.

[107] NCBI. Gene expression omnibus. "https://www.ncbi.nlm.nih.gov/geo/".

[108] NCBI. Sequence read archive. "https://www.ncbi.nlm.nih.gov/sra/".

[109] Feiping Nie, Jing Li, Xuelong Li, et al. Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification. In *IJCAI*, pages 1881–1887, 2016.

[110] Matthew T Noakes, Henry Brinkerhoff, Andrew H Laszlo, Ian M Derrington, Kyle W Langford, Jonathan W Mount, Jasmine L Bowman, Katherine S Baker, Kenji M Doering, Benjamin I Tickman, et al. Increasing the accuracy of nanopore dna sequencing using a time-varying cross membrane voltage. *Nature biotechnology*, 37(6):651–656, 2019.

[111] Vigdis Nygaard and Eivind Hovig. Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling. *Nucleic acids research*, 34(3):996–1014, 2006.

[112] Christopher A Odhams, Deborah S Cunninghame Graham, and Timothy J Vyse. Profiling rna-seq at multiple resolutions markedly increases the number of causal eqtls in autoimmune disease. *PLoS genetics*, 13(10):e1007071, 2017.

[113] OmegaBiotek. Rna-seq isolation and purification kit. https://www.omegabiotek.com/product/total-cellular-rna-mag-bind-total-rna-96, 2022. [Online; accessed 6-August-2022].

[114] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011.

[115] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

[116] Hao Peng, Hongfei Wang, Bowen Du, Md Zakirul Alam Bhuiyan, Hongyuan Ma, Jianwei Liu, Lihong Wang, Zeyu Yang, Linfeng Du, Senzhang Wang, et al. Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting. *Information Sciences*, 521:277–290, 2020.

[117] Tao Peng, Qin Zhu, Penghang Yin, and Kai Tan. Scrabble: single-cell rna-seq imputation constrained by bulk rna-seq data. *Genome biology*, 20(1):1–12, 2019.

[118] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[119] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[120] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature methods*, 14(7):687–690, 2017.

[121] Qiagen. Qiagen rna selection. https://www.qiagen.com/kr/qec/psg-rna-selection-tools/, 2022. [Online; accessed 6-August-2022].

[122] Qiagen. Rna-seq isolation and purification kit. https://www.qiagen.com/us/product-categories/discovery-and-translational-research/dna-rna-purification/rna-purification/, 2022. [Online; accessed 6-August-2022].

[123] Anto P Rajkumar, Per Qvist, Ross Lazarus, Francesco Lescai, Jia Ju, Mette Nyegaard, Ole Mors, Anders D Børglum, Qibin Li, and Jane H Christensen. Experimental validation of methods for differential gene expression analysis and sample pooling in rna-seq. *BMC genomics*, 16(1):1–8, 2015.

[124] Jiahua Rao, Xiang Zhou, Yutong Lu, Huiying Zhao, and Yuedong Yang. Imputing single-cell rna-seq data by combining graph convolution and autoencoder neural networks. *Iscience*, 24(5):102393, 2021.

[125] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[126] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

[127] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Sumit Mukherjee, Wei Chen, David J Peeler, Zizhen Yao, Bosiljka Tasic, Drew L Sellers, et al. Scaling single cell transcriptomics through split pool barcoding. *BioRxiv*, page 105163, 2017.

[128] Salmon. Alevin module for salmon, documentation site. https://salmon.readthedocs.io/en/latest/alevin.html, 2022. [Online; accessed 10-August-2022].

[129] Yohei Sasagawa, Hiroki Danno, Hitomi Takada, Masashi Ebisawa, Kaori Tanaka, Tetsutaro Hayashi, Akira Kurisaki, and Itoshi Nikaido. Quartz-seq2: a high-throughput single-cell rna-sequencing method that effectively uses limited sequence reads. *Genome biology*, 19(1):1–24, 2018.

[130] Yohei Sasagawa, Itoshi Nikaido, Tetsutaro Hayashi, Hiroki Danno, Kenichiro D Uno, Takeshi Imai, and Hiroki R Ueda. Quartz-seq: a highly reproducible and sensitive single-cell rna sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, 14(4):1–17, 2013.

[131] satijalab. Seurat repository and documentation. https://satijalab.org/seurat/, 2022. [Online; accessed 10-August-2022].

[132] Amanda N Scholes and Jeffrey A Lewis. Comparison of rna isolation methods on rna-seq: implications for differential expression and meta-analyses. *BMC genomics*, 21(1):1–9, 2020.

[133] Lynn M Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, et al. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962, 2019.

[134] Seurat. Seurat clustering approach. "https://satijalab.org/seurat/articles/pbmc3k_tutorial.html#cluster-the-cells-1".

[135] Hibah Shaath, Radhakrishnan Vishnubalaji, Ramesh Elango, Shahryar Khattak, and Nehad M Alajez. Single-cell long noncoding rna (lncrna) transcriptome implicates malat1 in triple-negative breast cancer (tnbc) resistance to neoadjuvant chemotherapy. *Cell Death Discovery*, 7(1):1–14, 2021.

[136] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[137] SingleR. Singler documentation. https://bioconductor.org/books/release/SingleRBook/, 2022. [Online; accessed 10-August-2022].

[138] Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168, 2021.

[139] Michal Slyper, Caroline Porter, Orr Ashenberg, Julia Waldman, Eugene Drokhlyansky, Isaac Wakiro, Christopher Smillie, Gabriela Smith-Rosario, Jingyi Wu, Danielle Dionne, et al. A single-cell and single-nucleus rna-seq toolbox for fresh and frozen human tumors. *Nature medicine*, 26(5):792–802, 2020.

[140] Tom Smith, Andreas Heger, and Ian Sudbery. Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome research*, 27(3):491–499, 2017.

[141] Tom Smith, Andreas Heger, and Ian Sudbery. Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome research*, 27(3):491–499, 2017.

[142] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[143] Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell rna-seq. *BioRxiv*, page 003236, 2014.

[144] Spektral. Spectral documentation. "https://graphneural.network".

[145] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[146] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1, 1949.

[147] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[148] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[149] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[150] Rui Tang, Christopher W Murray, Ian L Linde, Nicholas J Kramer, Zhonglin Lyu, Min K Tsai, Leo C Chen, Hongchen Cai, Aaron D Gitler, Edgar Engleman, et al. A versatile system to record cell-cell interactions. *Elife*, 9:e61080, 2020.

[151] ThermoFisher. Rna-seq extraction kit. https://www.thermofisher.com/it/en/home/life-science/dna-rna-purification-analysis/rna-extraction.html, 2022. [Online; accessed 6-August-2022].

[152] ThermoFisher. Sequencing platform iontorrent. https://www.thermofisher.com/it/en/home/brands/ion-torrent.html, 2022. [Online; accessed 6-August-2022].

[153] Julie D Thompson, Toby J Gibson, and Des G Higgins. Multiple sequence alignment using clustalw and clustalx. *Current protocols in bioinformatics*, pages 2–3, 2003.

[154] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[155] Jaris Valencia, Lidia M. Fernández-Sevilla, Alberto Fraile-Ramos, Rosa Sacedón, Eva Jiménez, Angeles Vicente, and Alberto Varas. Acute lymphoblastic leukaemia cells impair dendritic cell and macrophage differentiation: role of bmp4. *Cells*, 8(7):722, 2019.

[156] David van Dijk, Juozas Nainys, Roshan Sharma, Pooja Kaithail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *BioRxiv*, page 111591, 2017.

[157] Panagiotis D Vouzis and Nikolaos V Sahinidis. Gpu-blast: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, 27(2):182–188, 2011.

[158] Jinglu Wang, Dylan C Dean, Francis J Hornicek, Huirong Shi, and Zhenfeng Duan. Rna sequencing (rna-seq) and its application in ovarian cancer. *Gynecologic oncology*, 152(1):194–201, 2019.

[159] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348, 1994.

[160] Wikipedia. Bone marrow composition. "https://en.wikipedia.org/wiki/Bone_marrow#Hematopoietic_components".

[161] Wikipedia contributors. Blosum — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=BLOSUM&oldid=1094129527, 2022. [Online; accessed 10-August-2022].

[162] Wikipedia contributors. Gene regulatory network — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Gene_regulatory_network&oldid=1106222025, 2022. [Online; accessed 26-August-2022].

[163] Wikipedia contributors. Rna-seq — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=RNA-Seq&oldid=1100026702, 2022. [Online; accessed 6-August-2022].

[164] Jane AC Wilson, Natalie A Prow, Wayne A Schroder, Jonathan J Ellis, Helen E Cumming, Linden J Gearing, Yee Suan Poo, Adam Taylor, Paul J Hertzog, Francesca Di Giallonardo, et al. Rna-seq analysis of chikungunya virus infection and identification of granzyme a as a major promoter of arthritic inflammation. *PLoS pathogens*, 13(2):e1006155, 2017.

[165] Nina Witt, Gillian Rodger, Jo Vandesompele, Vladimir Benes, Alimuddin Zumla, Graham A Rook, and Jim F Huggett. An assessment of air as a source of dna contamination encountered when performing pcr. *Journal of biomolecular techniques: JBT*, 20(5):236, 2009.

[166] S Samuel Yang, Zheng Jin Tu, Foo Cheung, Wayne Wenzhong Xu, JoAnn FS Lamb, Hans-Joachim G Jung, Carroll P Vance, and John W Gronwald. Using rna-seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC genomics*, 12(1):1–19, 2011.

[167] Fan Zhang, Kevin Wei, Kamil Slowikowski, Chamith Y Fonseka, Deepak A Rao, Stephen Kelly, Susan M Goodman, Darren Tabechian, Laura B Hughes, Karen Salomon-Escoto, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nature immunology*, 20(7):928–942, 2019.

[168] Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[169] Sushen Zhang, Seyed Mojtaba Hosseini Bamakan, Qiang Qu, and Sha Li. Learning for personalized medicine: A comprehensive review from a deep learning perspective. *IEEE Reviews in Biomedical Engineering*, 12:194–208, 2019.

[170] Yingdong Zhao, Ming-Chung Li, Mariam M Konaté, Li Chen, Biswajit Das, Chris Karlovich, P Mickey Williams, Yvonne A Evrard, James H Doroshow, and Lisa M McShane. Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. *Journal of translational medicine*, 19(1):1–15, 2021.

[171] Yingqi Zhao and Donglin Zeng. Recent development on statistical methods for personalized medicine discovery. *Frontiers of medicine*, 7(1):102–110, 2013.

[172] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

[173] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[174] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature protocols*, 12(1):44–73, 2017.