UNIVERSITÀ DEGLI STUDI DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL' INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN

BIOINGEGNERIA

"Comparative analysis of three computational approaches for investigating cell-cell communication in single cell RNA-seq data of non-small cell lung cancer"

Supervisor:

Prof. Giacomo Baruzzo

Master Candidate:

Brian De Marchi

Co-Supervisors:

Dr. Giulia Cesaro

Prof. Zlatko Trajanoski

Academic Year 2023/2024

Graduation Date 16/04/2024

ABSTRACT

Cell-cell communication is a fundamental process that enables cells to respond to changes in their environment. Understanding these interactions may be an important achievement for comprehending biological complexity.
This project aims to observe whether bioinformatics can have an important role in increasing the knowledge about these biological processes through computational methods applied to single cell RNA-seq data. Additionally the project addresses the absence in the field of a benchmark that may allow for better comparisons between the different tools to assess relative performances.

For the execution of this project an high-resolution single-cell atlas is used. This atlas comes from the dissection of the non-small cell lung cancer (NSCLC) tumor microenvironment by integrating 1,283,972 single cells and originates from the Centre for Chemistry and Biomedicine (CCB) of Innsbruck. Three tools, namely scSeqComm, CellphoneDB, and NicheNet, are tested on this data, each employing distinct statistical frameworks for cell-cell communication analysis. They utilise a count matrix representing RNA molecule counts for genes in individual cells within a sample. The focus is on intercellular communication, specifically ligand-receptor pairing.

Each method works with a specific proper framework and gives different outputs, therefore the comparison among them considers the time required for calculations and output generation, the number of ligand-receptor pairs identified, and the influence of using different databases for providing prior knowledge.
In conclusion, this project demonstrates that bioinformatics and computational tools for the analysis of cell-cell communication may play an important role in advancing knowledge in this field. However, it acknowledges the ongoing need for substantial efforts to achieve reliable results applicable in clinical practice.

OVERVIEW

List of figures

List of tables

List of abbreviations

CCB – Centre for Chemistry and Biomedicine of Innsbruck

CCC – Cell-Cell communication

CCIs – Cell-Cell interactions

cDNA – complementary DNA

DB – Database

DNA – Deoxyribonucleic acid

GE – Gene expression

HPC – High-Performance Computing

HPCC – High-Performance Computing Cluster

LR – Ligand-Receptor

LUAD – Lung adenocarcinoma

LUSC – Lung squamous cell carcinoma

NSCLC – Non Small Cell Lung Cancer

PCA – Principal component analysis

PPIs – Protein-protein interactions

RNA – Ribonucleic acid

RNA-seq – RNA sequencing

scRNA-seq – single-cell RNA sequencing

SGE – Sun Grid Engine

TME – Tumor microenvironment

t-SNE – t-distributed stochastic neighbour embedding

UMAP – Uniform manifold approximation and projection

Chapter 1: Introduction

1.1. Genomics

An organism's genome may be thought of as the collection of all the instructions that code for the proteins required for reproduction, interaction with the environment, and survival [1]. However, protein expression, that is, the repertoire of proteins produced or expressible in response to a stimulus, allows distinct cells to take on particular and differentiated traits and activities [2]. In fact, cells not only contain instructions for protein coding but also information regarding the conditions under which proteins should be synthesized. The manifestation of this information is made possible by extremely intricate regulatory and control systems [3].

The integrated analysis of this vast amount of data (DNA, mRNA, and proteins) along with the quantification of metabolites and other substrates of interest, allows for the foundation of systems biology [4] on new and solid experimental grounds. Gene sequences, their expression and translation into proteins, the characteristics of these proteins and their functions collectively encode all the information required for a cell to function; as such, they represent the basic processes that occur prior to the systemic events that take place as a result of a disease or in response to an external stimulus [5].

The purpose of studies in genetics, genomics, and proteomics is therefore to reveal the mechanisms underlying cellular processes with the aim of translating basic knowledge into increasingly sophisticated tools to diagnose diseases early, predict their progression, and, prospectively, develop personalized therapies targeted not to a population but to the needs of the individual. An important aid for all these scientific and technological advances has been the development of computational procedures and infrastructures for the management, analysis, and exploitation of the abundance of data generated by various molecular and cellular biology techniques [6].

1.1.1. Gene expression

Gene expression (GE) refers to the process of the synthesis of a functional gene product, such as proteins, using the information provided by deoxyribonucleic acid (DNA) [7]. Ribonucleic acid (RNA) is synthesized from DNA through the process of transcription, which is part of the process of GE. Cells have the ability to modify the kind and quantity of GE as an organism grows or adapts to changes in its environment. Hence, studies of GE provide insights into cellular responses at a given point in time. Minor modifications in the regulatory pathways linked to changes in GE may account for significant phenotypic variations amongst organisms. Numerous factors have led to a rise in research into GE alterations. First, there is an increasing amount of clinical samples available from tissue repositories, as well as novel techniques for measuring GE from diverse tissues [8]. Second, a sizable collection of experimental GE data is openly accessible via databases. Third, the most advanced GE measurement tools (such as RNA sequencing [RNA-seq]) are getting more widely applicable and less expensive [9].

Gene expression is the result of the transcription of DNA into RNA and the translation of RNA into proteins. These processes combine to produce unique molecules that carry out different tasks within the cell. Gene expression is strictly controlled and is impacted by a number of variables, such as cellular signals, developmental phases, and environmental stimuli. Both genes that are transcribed into messenger RNA (mRNA) and subsequently translated into proteins, as well as genes that are transcribed into RNA but not translated into proteins, such transfer and ribosomal RNAs, are considered expressed genes [10].

The steps involved in the expression of genes are transcription, RNA splicing, RNA export, and translation, as shown in Figure 1.

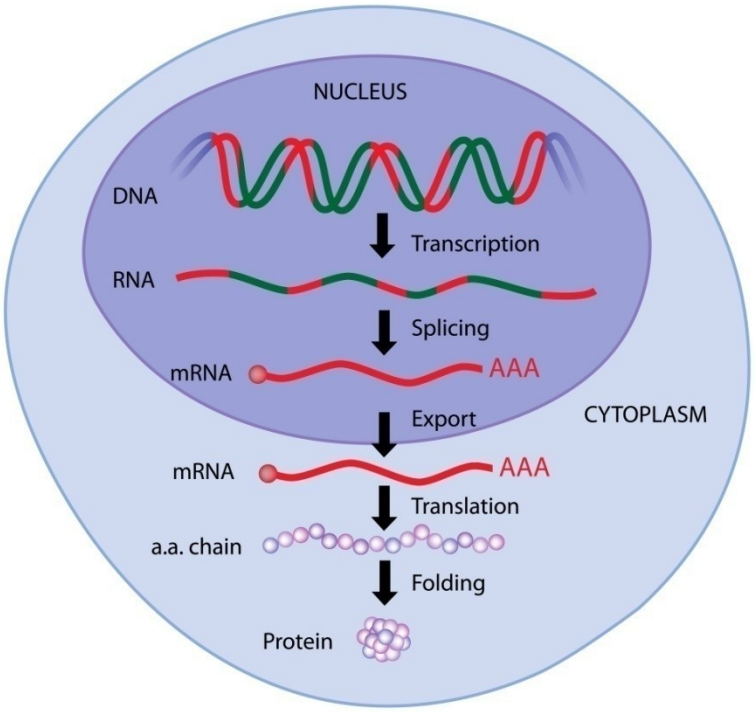

Figure 1. *Steps of the process of gene expression: transcription, RNA splicing, RNA export and translation. Image taken from [11].*

During **transcription**, an RNA molecule is created by copying a gene's DNA sequence. The reason this phase is named transcription is that it entails rewriting, or transcribing, the DNA sequence in an analogous RNA "alphabet." The RNA molecule in eukaryotes needs to be processed in order to mature into a mRNA.

During **translation**, a polypeptide's amino acid sequence is determined by decoding the mRNA sequence. The term "translation" refers to the process of translating the nucleotide sequence of mRNA into the entirely distinct "language" of amino acids.

Because the balance between a protein's synthetic and degradative biochemical pathways determines how much of a given protein is present in a cell at any given time, control over these two processes is essential to understanding which proteins are present in cells and in what quantities. It's crucial to remember that, from a synthetic perspective, protein synthesis starts with transcription (DNA to RNA) and advances through translation (RNA to protein).

Furthermore, a cell's processing of freshly synthesised proteins and RNA transcripts has a significant impact on the amount of protein present.

A cell's function is reflected in the types and quantities of mRNA molecules present in that cell. In fact, each cell produces thousands of transcripts per second. This statistic makes it not surprising that the main regulator of gene expression is often transcription, which starts at the very beginning of the process of making proteins.

In a cell, only a portion of its genes are expressed at any given moment. Different cell types exhibit varying gene expression profiles, which can be attributed to the various transcription regulator sets present in these cells. While some of these regulators block or reduce transcription, others function to promote it.
Normally, transcription starts when an RNA polymerase attaches itself to a DNA molecule's promoter sequence. This sequence is almost always located just upstream from the starting point for transcription (the 5' end of the DNA), though it can be located downstream of the mRNA (3' end).
A few regulatory proteins have an impact on several genes' transcription. This happens because a cell's genome has several copies of the regulatory protein binding sites. As a result, regulatory proteins may have distinct roles for various genes. This is one way that cells can simultaneously coordinate the regulation of numerous genes.
In eukaryotes, the regulation of gene expression is extremely intricate. Generally speaking, there are more regulatory proteins at play, and transcription promoter sites may be fairly distant from regulatory binding sites. Furthermore, the regulation of eukaryotic gene expression is typically mediated by many regulatory proteins working in concert, providing for increased control over gene expression.

Cells need to be able to react to changes in their surroundings in order to survive. This plasticity depends on the control of transcription and translation, the two primary processes in the synthesis of proteins. In addition to being able to regulate which genes are translated, cells are also able to modify the activity of proteins and transcripts through biochemical processing.

The processes that control GE include transcription, many epigenetic pathways, and posttranscriptional modifications. Regulation of GE, which is cell-specific and involves a

number of complex biochemical processes, is essential for the growth of the organism as well as its ability to respond to changes in its environment. Understanding the mechanisms that control GE is essential to being able to analyse GE data.

One of the most significant mechanisms is related to transcription factors. Gene expression is controlled by fewer than 2,000 of these proteins [12], known as transcription factors, which have the ability to initiate and regulate transcription. Protein kinases activate induced transcription factors so they can attach to specific response elements. For instance, elevated blood hormone levels can trigger the activation of particular transcription factors by activating cell-surface receptors that in turn trigger a series of protein kinase-activated cell-signalling pathways [13]. The levels of GE can be altered by modifications in hormone levels, cell-surface receptor expression, or transcription factor expression. By selectively transcriptionally regulating a subset of genes, transcription factors regulate the amounts of GE in a cell. Morphological alterations may result from transcription factor overexpression.

1.2. Gene expression measure

The most widely utilised laboratory techniques for determining GE levels are quantitative polymerase chain reaction (qPCR), DNA microarray, and RNASeq. This section explains different approaches and contrasts their advantages and disadvantages.

1.2.1. qPCR

qPCR, or quantitative polymerase chain reaction, is a real-time method used to measure gene expression [14]. It quantifies the amount of genetic material using a spectrophotometer, often using mRNA as the template. During qPCR, mRNA is converted to complementary DNA (cDNA) and then synthesized into double-stranded DNA, which undergoes exponential amplification. Fluorescent labels are added to track the amplification, with the fluorescent signal proportional to the amount of amplified DNA. The threshold cycle (Ct) value indicates when amplification is first detected above a baseline threshold, providing an estimate of gene expression level. A standard curve generated from a control DNA dilution allows for absolute quantification. Multiplex qPCR can detect multiple transcripts simultaneously. Advantages of qPCR include ease of use, relatively short quantification time (8–12 hours), and the ability to

detect multiple transcripts. However, limitations include the need for prior knowledge of target transcript sequence and the limited number of transcripts that can be quantified in each reaction, reducing throughput.

1.2.2. Microarray

For more than 15 years, microarrays have been a standard method for assessing GE [15]. This method measures many transcripts at once by using nucleic acid hybridization of cDNA strands [16]. In situ synthesized oligonucleotide microarrays and cDNA microarrays are the two primary forms of microarrays. In the first, short oligonucleotides are adhered to a chip surface, whereas in the latter, reverse-transcribed cDNA from mRNA is used. RNA extraction, reverse transcription to cDNA, labelling, hybridization with probes on a chip, washing, fluorescence signal detection, and data analysis are the various stages involved in microarray research. Benefits of microarrays include their ability to quantify several transcripts at once, their affordability, and the fact that they don't require prior knowledge of transcript sequences. However, they have limitations such as inability to test multiple tissue samples at once, time-consuming sample preparation, and dependence on specialized equipment and software for data processing.

1.2.3. RNAsequencing

RNA-seq is a method used to quantify the levels of different types of RNA in a sample by directly sequencing the RNA and counting the sequences. Unlike the other techniques such as qPCR, which quantifies RNA through amplification and dye intensity, and microarray, which quantifies RNA through template hybridization and dye intensity; RNA-seq offers several advantages. While several methods for sequencing RNA exist [17], they all share a similar overall process. The type of RNA sequenced depends on the objective of the study. For example, total RNA-seq attempts to measure all of the expressed RNA. Coding RNA can be enriched by poly(A) capture techniques, and small RNAs can be enriched through size selection and gel electrophoresis The RNA can be fragmented and adapters are ligated to the fragments before amplification and sequencing. RNA-seq allows for massive parallel sequencing of transcripts, enabling the detection of genomic alterations at single-nucleotide

resolution and offering a greater dynamic range for quantifying transcripts compared to microarray technology. However, RNA-seq has limitations such as higher cost per sample and greater computational and data-storage requirements for downstream analyses compared to microarray. Nonetheless, advancements in cheaper assays with increased sensitivity and easier-to-use data-storage procedures and analysis tools are continuously evolving [18].

1.2.4. Single cell RNA sequencing

The study of RNA allows for the understanding of the functional components of cells and how their genes are used. The standard analysis, referred to as bulk RNA-seq, examines RNA expression in large populations of cells yielding an aggregate expression value. Due to the measurement's inability to reveal the variations among individual cells within a group, single-cell RNA sequencing (scRNA-seq) has been developed in recent years and has allowed for modifications and improvements. This new approach allows us to study individual cells and learn more about the many types of cells [19].

Single-cell RNA sequencing is the gene expression profiling of singlet cells. It can show a large variety of cell types and subpopulation that were unseen with traditional experimental techniques, and it also led to the discovery of new information in regard to the cell composition. scRNA-seq describes RNA molecule with high resolution and on the genomic level allowing the comparison of single cell transcriptome.

The essential steps in the scRNA-seq process (Figure 2) are single-cell isolation and capture, cell lysis, reverse transcription (conversion of their RNA into cDNA), cDNA amplification and library preparation. Among the stages involved in preparing libraries, single-cell capture, reverse transcription, and cDNA amplification present the greatest challenges. The rapid and diverse development of RNA-seq library preparation technologies has coincided with the emergence of numerous sequencing platforms. To better apply these techniques to clinical applications and make informed decisions in scientific research, it is crucial to understand the characteristics and uses of various single-cell RNA sequencing library preparation methods.

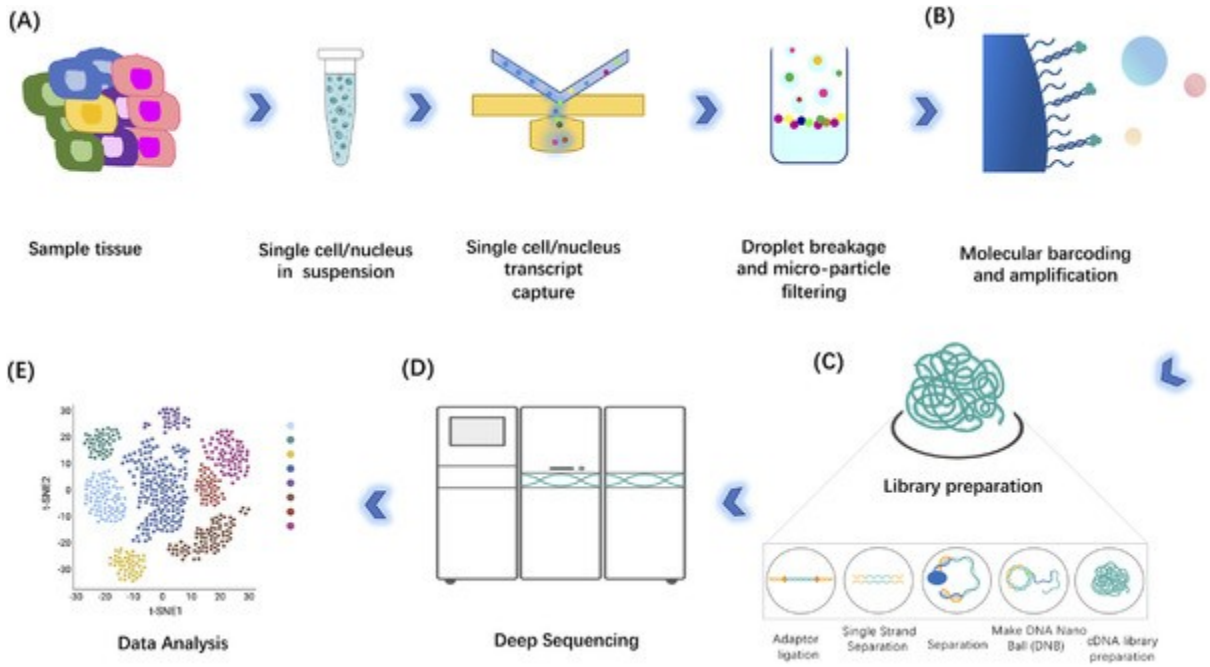

Figure 2. *An overview of the single-cell RNA-sequencing procedures. (A) Isolation of the cells from tissue samples and capturing of the single cells, wrapping of each individual cell with a bead inside a nanoscale droplet (each bead contains unique molecular identifiers), (B) barcoding and amplification of complementary DNA (cDNA) and (C) library preparation procedure. After single-cell RNA sequencing (D), the snapshot data would be analyzed to present and classify the landscape of gene expression in cells of a heterogeneous population (E). Illustrative figure in (E) is generated with BioRender with license for publication. Image taken from [23]*

Single-cell isolation and capture is the process of capturing high-quality individual cells from a tissue, in order to extract precise genetic and biochemical information and enable the investigation of distinct genetic and molecular mechanisms [20]. Conventional transcriptome, epigenome, or proteome analyses performed on bulk RNA/DNA samples are limited to capturing the overall signal level from tissues/organs; they are unable to discriminate between distinct cell variants. Depending on the organisms, tissues, or cell characteristics, there are significant differences in the single-cell isolation and capture techniques [21].

Cell isolation can be accomplished by isolating whole cells, cell-specific nuclei or cell-specific organelles, and even by separating the desired cells expressing specific marker proteins [22]. The most common techniques of single-cell isolation and capture include limiting dilution, fluorescence-activated cell sorting (FACS), magnetic-activated cell sorting, microfluidic system and laser microdissection. The key outcome of single capture, and particularly in high throughput, is that each single cell is captured in an isolated reaction mixture, of which all transcripts from one single cell will be uniquely barcoded after converted into complementary DNAs [23].

When evaluating the single-cell RNA sequencing data, every cell is regarded as an individual sample, much like in the study of conventional bulk RNA-seq data. Because of systemic faults or technological disturbances, such as variations in sequencing depth and transcriptome capture rate for each cell, the expression levels between cells are not comparable, and the original expression matrix cannot be used directly for downstream analysis. The goal of normalisation is to guarantee cross-cell comparability and mitigate any bias or technological noise [24].

With hundreds of genes expressed in each of the tens of thousands of cells that make up a sample, the single-cell RNA sequencing data set is high-dimensional. The majority of genes in every cell are housekeeping genes because their existence tends to mask the true biological signals and they are characterised by no discernible changes in expression levels between cells. Genes with substantial cell-to-cell variance within the data set are referred to as highly variable genes.

Apart from feature selection, one of the primary approaches for handling high-dimensional data is dimensionality reduction. Two rounds of dimension reduction are typically needed for single-cell RNA sequencing data: first, principal component analysis (PCA) dimension reduction, followed by visualization-related dimension reduction using either Uniform Manifold Approximation and Projection (UMAP) or t-distributed stochastic neighbour embedding (t-SNE).

1.3. Cell atlases

Single-cell RNA sequencing technology has become the state-of-the art method for understanding the heterogeneity and complexity of RNA transcripts within individual cells. It helps us see the various kinds of cells and what they do in complex structures like tissues and organs. One important application of the scRNA-seq technology is to build a better and high-resolution catalogue of cells in all living organism, commonly known as atlas, which is key resource to better understand and provide a solution in treating diseases.

Researchers are producing enormous volumes of data as scRNA-seq technology develops, which calls for robust methods for analysing, interpreting, and storing this data. The usage of

single-cell data atlases is one such method that is becoming more and more popular. By developing these atlases, we will be able to better comprehend the cellular landscape and create therapies that are specific to the needs of each patient.

Atlases offer an in-depth investigation of the patterns of gene expression in the cells that collectively make up a certain organ or tissue. These atlases are typically the result of cooperative efforts between several institutions and research organisations. They are useful in the identification of novel cell types, the study of gene expression patterns specific to particular cell types, and the organisation and functionality of cells in various tissues.

A single-cell data atlas is a broad collection of single-cell transcriptomic data from various tissues or organisms, where each cell's gene expression profile is recorded. These atlases offer a big amount of resources for comprehending cellular variety and dynamics by enabling researchers to navigate and investigate the variability of cellular populations within certain tissues or developmental stages.

These extensive resources enable previously unthinkable discoveries and are invaluable references for researchers globally. We will gain even more understanding of the complex world of cells as we develop and expand these atlases, opening the door to innovative treatment approaches and personalized medicine.

An example of atlas is the The Human Cell Atlas [25]. Established in 2016, the Human Cell Atlas (HCA) is an international initiative. Fundamentally, the goal of the HCA is to offer an in-depth description of the human body's cellular constitution. The HCA provides a complete comprehension of the location, function, and patterns of gene expression of every cell in the human body by utilising state-of-the-art single-cell and spatial analytic techniques. For this reason, the HCA is a great resource for researching human biology in both health and illness. Researchers may learn what controls the development and activity of various cell types, how these cells interact with one another, and where these cells are distributed within tissues and the body thanks to the cellular reference maps produced by HCA. Then, researchers can study the biological changes that occur in disease and the therapeutic potential of these changes.

1.3.1. Atlas building

The cells in a multicellular organism all have the same genetic code, but distinct transcriptome programmes found in tissues and organs are the result of nongenetic cellular heterogeneity, which is manifested through various combinations and patterns of expression. Cell types can be identified based on a globally similar transcriptome. The easiest method is to gather a lot of cells, sequence their transcriptomes, and use computer analysis to identify cell types that are comparable. Moreover, the analysis of transcriptional programmes throughout cellular differentiation is made possible by scRNA-seq during the course of development, which captures the variability of cellular development.

While unsupervised clustering in a single scRNA-seq experiment can be used to characterise cell types and states, computational challenges arise due to the growing number of cells and associated batch effects. Large-scale single-cell transcriptome datasets produced with various technologies have systematic differences that are unique to a batch, known as batch effects. Although batch effects make it difficult to easily combine differently generated datasets, a single technology cannot sample the complexity of an entire organism at once. Therefore, a compilation of several studies integrated in a way that minimises technical error will be required to create complete single-cell datasets applying batch effect correction methods [26].

In Figure 3 we can see the steps for the construction of a whole-organism cell atlas. The figure shows that every cell in the body has the same genome but remarkably different functions, mostly because the associated epigenetic layer is different. Mammalian tissues and organs are composed of many different cell types that can vary in abundance and cell state, and this heterogeneity is not captured by bulk analyses. In contrast, single-cell analysis can uncover biological differences among cell populations, leading to a complete understanding of their function in the physiology of the organism. To achieve a nearly complete navigable map of all the cell types and states, one needs to combine gene expression information with information on proteins and cellular location, among other parameters.

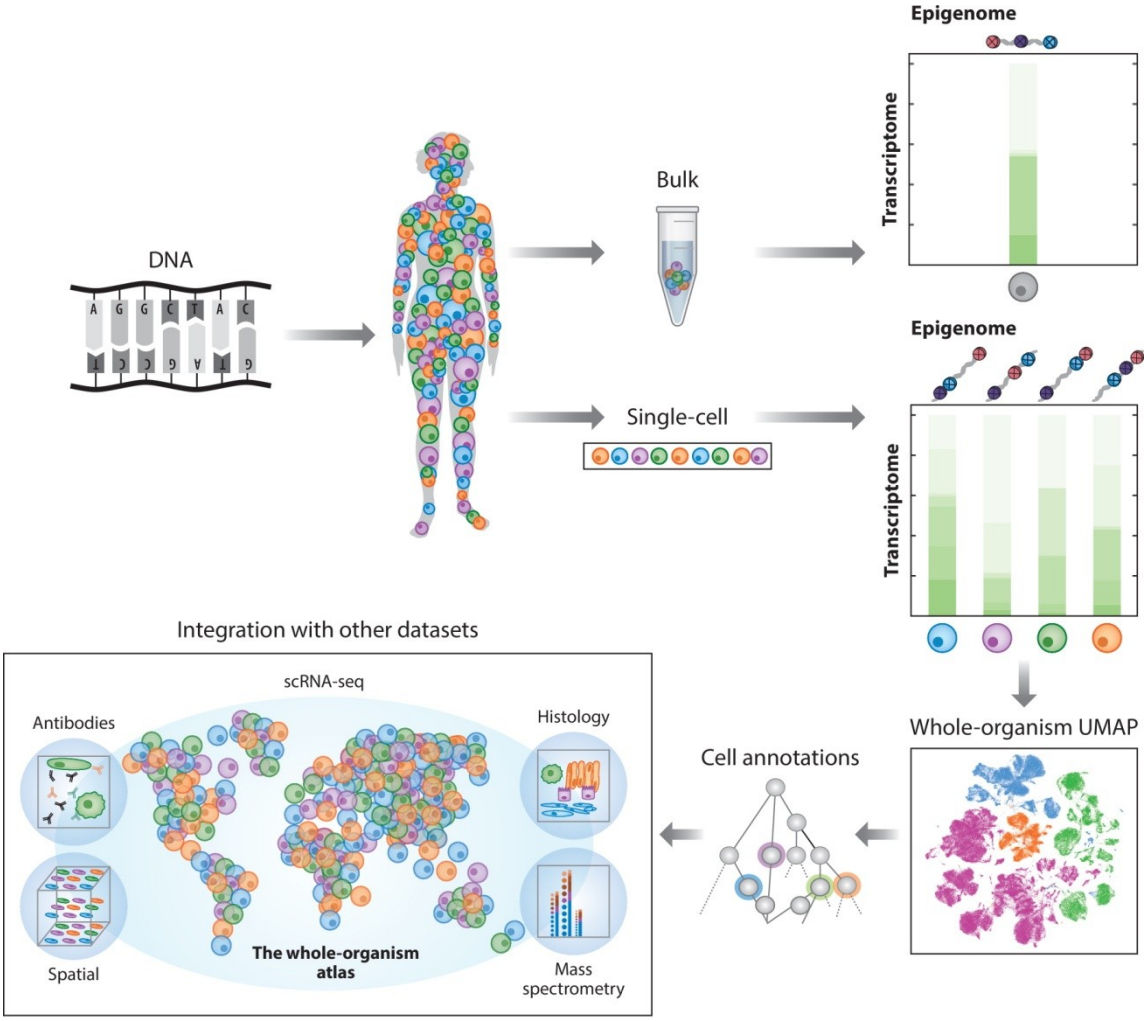

Figure 3. *Multistep construction of a whole-organism cell atlas. Image taken from [26].*

1.4. Cell-cell communication

To better understand the contents of this thesis it is necessary to delve deeper into the concept of cell-cell communication, a fundamental aspect of cellular biology with significant implications for various physiological processes. By examining the intricate mechanisms and dynamics of how cells interact and exchange information, we seek to unravel the complexities underlying tissue balance, development, and disease progression.

Intercellular communication represents the basis of systemic responses to environmental cues. One-to-one interaction between similar cells (homotypic interaction) or cells of different origins (heterotypic interaction) can regulate collective behavior, such as migration, or mount a coordinated response, such as antipathogenic activity [27].

Multicellular organisms rely on cell-cell interactions (CCIs) to regulate individual cell functions and develop tissue structure [28]. They also help to coordinate a variety of biological processes, including development, differentiation, and inflammation, as well as to maintain intercellular interactions. As shown in figure 4 a CCI happens when information is sent from one cell, known as the *sender cell*, to another, known as the *target cell* (or *receiver cell*), using signalling molecules. CCIs are caused by a variety of signalling molecules, including ions, metabolites, integrins, receptors, junction proteins, structural proteins, ligands, and extracellular matrix-secreted proteins.
Lastly, it influences the expression of *target genes* and the function of transcription factors in the receiver cell. In recent years, the most prevalent scenario for the computational investigation of CCIs has been CCIs mediated by LR (Ligand-Receptor) interactions.

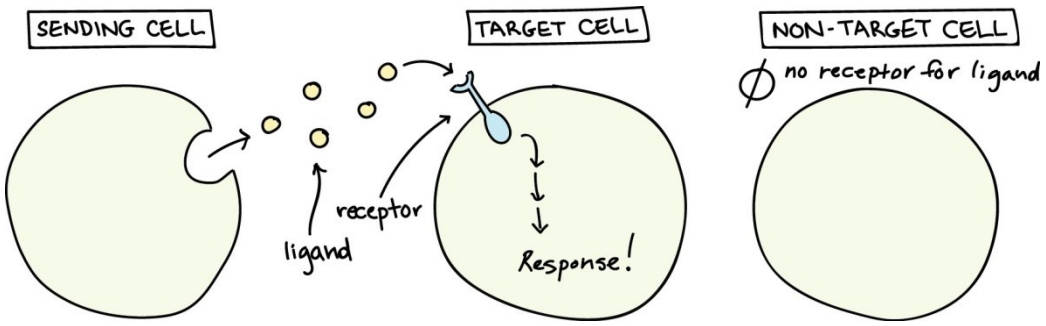

Figure 4. *Intercellular communication between sender cell and target cell.*
Image taken from [29].

Signaling molecules are often called *ligands*, a general term for molecules that bind specifically to other molecules (such as *receptors*).

The message delivered by a ligand is frequently passed along via an intracellular network of chemical messengers (figure 5). In the end, it results in a modification of the cell, like a change in a gene's activity. As a result, the initial signal that was *intercellular* (between cells) gets transformed into an *intracellular* (within cell) signal that initiates a reaction.

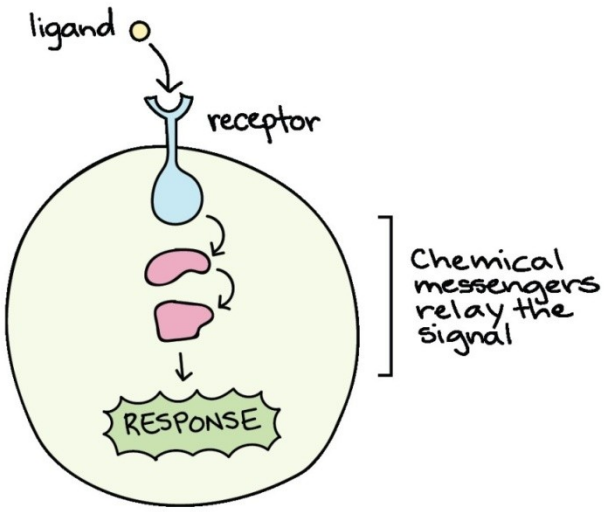

Figure 5. *Schematization of information transfer in intracellular communication. Image taken from [29].*

Structural CCIs are supported by some molecules, for example cell adhesion molecules, whereas ligands such as hormones, growth factors, chemokines, cytokines and neurotransmitters mediate cell–cell communication (CCC).

The signalling events behind CCC are often mediated by interactions of various types of protein, encompassing ligand–receptor, receptor–receptor and extracellular matrix–receptor interactions.

Through appropriate receptors, receiver cells initiate downstream signalling that typically results in changed transcription factor activity and gene expression. These expression-altered cells participate in additional interactions with their surroundings. It is necessary to recognise the protein signals transmitted between cells in order to comprehend each cell's function within its immediate environment; determining the expression of messenger molecules and the related pathways is vital for knowing the direction, size, and biological significance of CCC. Furthermore, these proteins can't always be investigated in their natural environment, necessitating the use of specialised biochemical assays and in-depth domain expertise for the

direct measurement of the proteins driving CCC. Traditional assays of the underlying protein–protein interactions (PPIs) include yeast two-hybrid screening, co-immunoprecipitation, proximity labelling proteomics, fluorescence resonance energy transfer imaging and X-ray crystallography [30-31]. Many connections between proteins that are secreted or expressed extracellularly to facilitate intercellular communication have been uncovered using these strategies. Studies like this can be strengthened further by proteomics and transcriptomics, since PPIs are supported by expression data.

Although direct measurement of protein abundances makes proteomic technologies preferable for these analyses, RNA sequencing data sets are more abundant, simpler to obtain, and need less analysis. When it comes to determining the cell type of origin of proteins causing CCIs and measuring expression in uncommon cell types, single-cell RNA-seq is better than bulk analysis. Transcriptomics results need to be carefully analysed and verified to prevent false assumptions, but because of its widespread use and simplicity of analysis, several recent research have been able to deduce CCC from gene expression, producing plausible hypotheses in a variety of academic fields. For example, intercellular communication can be inferred from the coordinated gene expression of ligands and receptors [32].

1.4.1. Types of cell–cell interactions and communication

Not all cell pairs exchange signals in the same way, nor are all sending and receiving cells next-door neighbours.

Within multicellular organisms, chemical signalling falls into four fundamental categories (Figure 6): paracrine, autocrine, endocrine, and juxtacrine [32]. The distance a signal travels through an organism to reach its target cell is the primary distinction between the various signalling categories.

Paracrine signalling

Chemical messengers, or ligands that can diffuse through the space between the cells, are often released by nearby cells to facilitate communication. Paracrine signalling is the name for this kind of signalling in which cells exchange signals over comparatively short distances.

Cells and their neighbours can coordinate activity locally thanks to paracrine signalling. Paracrine signals are utilised in a wide range of tissues and situations, but they are particularly

significant during development because they enable one group of cells to inform a nearby group of cells about which cellular identity to adopt.

Autocrine signaling

In autocrine signaling, a cell signals to itself, releasing a ligand that attaches to receptors on the cell's surface (or, depending on the signal type, to receptors inside the cell). Although it might seem strange for a cell to do this, autocrine signalling is crucial to a number of different processes.

For example, autocrine signalling plays a crucial role in development by assisting cells in assuming and maintaining their correct identities. From a medical perspective, autocrine signalling is significant in cancer and is believed to be crucial for metastasis, or the process by which cancer spreads to other parts of the body. In many cases, a signal may have both autocrine and paracrine effects, binding to the sending cell as well as other similar cells in the area.

Endocrine signaling

Cells frequently employ the circulatory system as a message delivery network when they need to transfer signals over great distances. Long-distance endocrine signalling involves the production of signals by specific cells that are then transported by the circulation to target cells located in different regions of the body. Hormones are signals that originate in one area of the body and move via the bloodstream to distant locations.

The thyroid, pituitary, hypothalamus, gonads (testes and ovaries), and pancreas are examples of endocrine glands in humans that release hormones. One or more hormones, many of which are master regulators of physiology and development, are released by each endocrine gland.

Juxtacrine signaling

Tiny channels called gap junctions in animals and plasmodesmata in plants are used to directly join adjacent cells. These water-filled channels allow small signaling molecules, called intracellular mediators, to diffuse between the two cells. Large molecules like proteins and DNA cannot fit through the channels without additional aid, but small molecules and ions can flow between cells.

The current condition of one cell is communicated to its neighbour through the passage of signalling molecules. As a result, a cluster of cells can synchronise their reaction to a signal that might have reached only one of them.

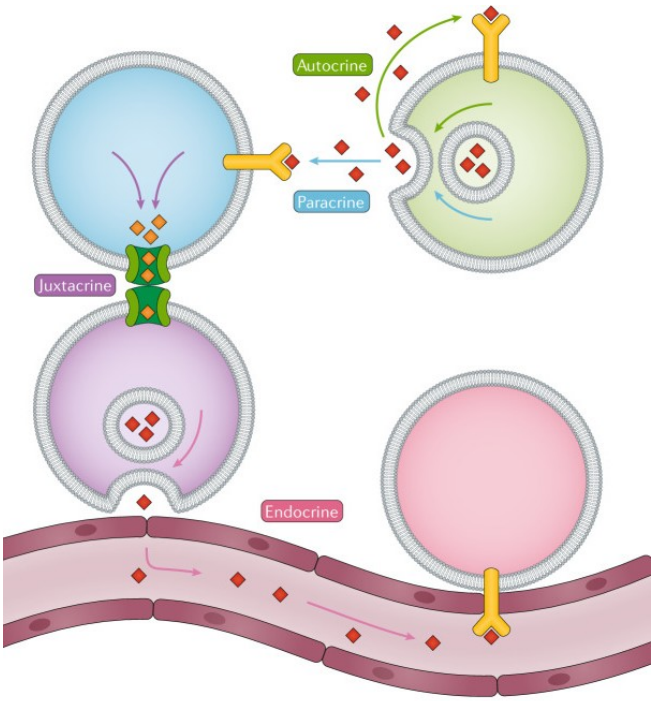

***Figure 6**. Types of cell–cell interactions and communication. Image adapted from [32].*

1.5. Bioinformatics analysis of cell-cell communication

Inferring CCC from transcriptomics relies on gene co-expression, whereby one gene in a given pair comes from one interacting cell and the other gene comes from the second interacting cell. Several studies focused on intercellular signalling using co-expression of all genes or specific cell markers [33], the similarity between expression profiles [34] or the properties of regulatory networks [35]. However, most studies rely on literature-curated lists of interacting proteins, which facilitates the biological interpretation of results (Figure 7). Although several studies have used interactions between any class of cell-surface protein and secreted protein [36] the predominant class of interactions used for studying CCC are known ligands and their associated receptors.

Complex extracellular responses start with the binding of a ligand to its cognate receptor and the activation of specific cell signaling pathways. Mapping ligand–receptor interactions is fundamental to understanding cellular behavior and response to neighboring cells. With the exponential growth of single-cell RNA sequencing, it is now possible to measure the

expression of ligands and receptors in multiple cell types and systematically decode intercellular communication networks that will ultimately explain tissue function in homeostasis and their alterations in disease. Identifying ligand–receptor interactions from scRNA-seq requires both the annotation of complex ligand–receptor relationships from the literature and a statistical method that integrates the resource with scRNA-seq data and selects relevant interactions from the dataset.

Many computational tools have been developed to identify CCIs through scRNA-seq data integration under specific cellular and physiological conditions [37]. These CCI prediction tools, in general, follow a common pipeline, including cell-type classification, LR interaction inference, CCI network construction and CCI visualization. However, each tool has its specific emphasis and algorithmic details.

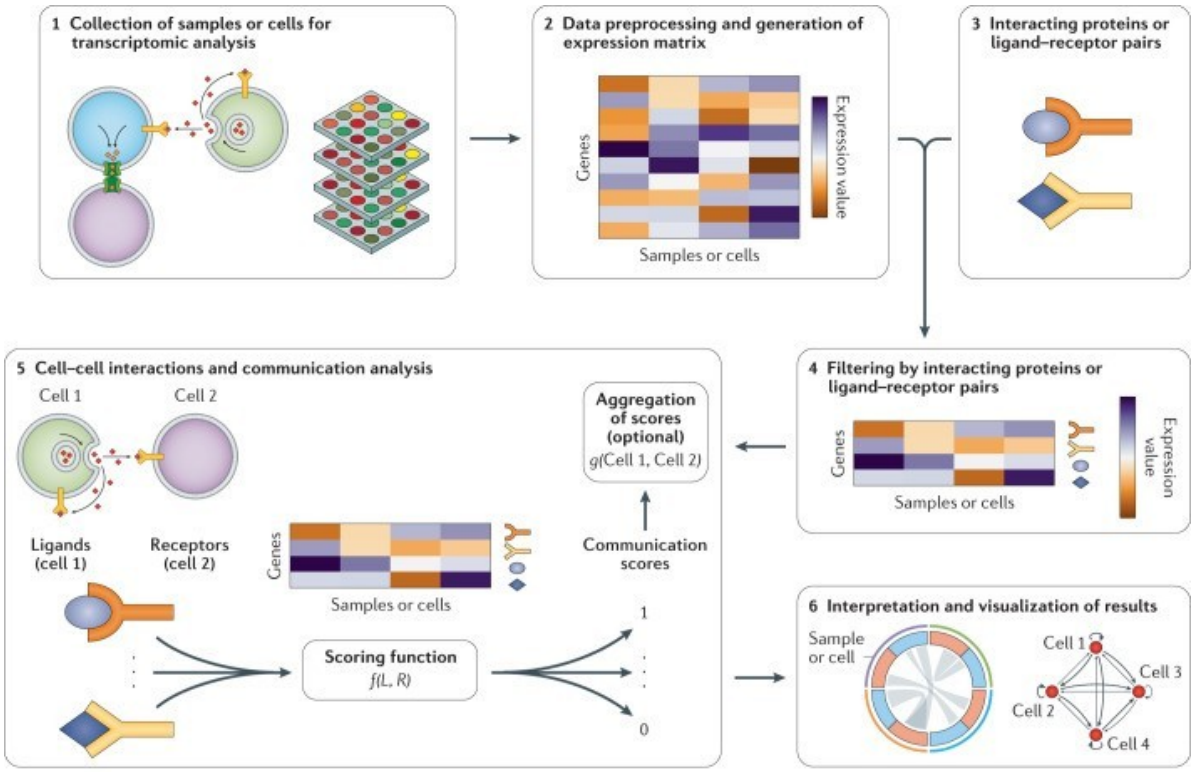

Figure 7. *General analysis workflow for inferring cell–cell interactions and communication. Image adapted from [32].*

In Figure 7 a general analysis workflow for inferring cell–cell interactions and communication is shown. First, samples or cells are analysed by transcriptomics to measure the expression of genes (step 1). Then the data generated are preprocessed to build a gene expression matrix, which contains the transcript levels of each gene across different samples or cells (step 2). The term "count matrix for gene expression" refers to a fundamental data structure used in the analysis of gene expression data, particularly in the context of RNA sequencing experiments. In RNA-seq, the count matrix represents the raw data obtained from sequencing the RNA molecules present in a biological sample. Each row of the count matrix corresponds to a gene, while each column represents a sample (e.g., different cells, tissues, or experimental conditions). The entries of the matrix indicate the number of RNA sequencing reads that align to each gene in each sample. This count matrix serves as the input for the CCC analysis methods.

Then the other input required by the methods is a database or a list of interacting proteins that are involved in intercellular communication that may be generated or obtained from other sources (step 3), often including interactions between secreted and membrane-bound proteins (commonly ligands and receptors, respectively).

Only the genes associated with the interacting proteins are held in the gene expression matrix (step 4). Their expression levels are used as inputs to compute a communication score for each ligand–receptor pair using a scoring function (function f(L, R), where L and R are the expression values of the ligand and the receptor, respectively). Communication scores can be binary or continuous, each providing different insights into the signalling pathways that cells use. Binary scores are simpler, whereas continuous scores enable more precise quantification of intercellular signalling. In binary scoring, expression thresholding is widely used because of its easy implementation and interpretation. Through the process of thresholding expression values of both interacting partners in every ligand-receptor pair, it is possible to quantify all intercellular communication processes. If both genes are expressed above a threshold, the ligand–receptor pair is considered ‘active’; otherwise it is ‘inactive’ (assigning ones and zeros, respectively).

Measurement of individual communication scores facilitates the study of CCC, exposing the roles of specific signalling mechanisms; however, it does not reveal the entire interaction state between cells. Thus, it may be desirable to use an aggregate score to define the interactions between pairs of cells [37]. To compute an overall state of interaction between the respective samples or cells an aggregation function is used (function g(Cell 1, Cell 2), where Cell 1 and Cell 2 are all communication scores of those cells or corresponding samples) (step 5). The

most common approach quantifies the number of active ligand–receptor pairs between cells (that is, the sum of binary communication scores). This score suggests which cells interact more strongly and enables the building of CCC networks to perform graph-based analyses. Finally, communication and aggregated scores can be represented to facilitate the interpretation of the results (step 6). In this sense, the tools also include powerful visualization features that facilitate the interpretation of results. Several of the more common visualization methods, such as heatmaps, dot plots, circos plots and interaction networks display data by directly plotting ligand–receptor co-expression patterns and communication scores providing higher-level intuition concerning overall CCI levels and the directionality of these effects between cell types. Thus, several tools not only quantify CCIs and CCC but also facilitate their analysis and interpretation.

It is important to note that a number of studies have used different approaches to catalogue known ligand-receptor pairs. Despite the diversity of approaches employed, the most suitable metric for capturing the underlying biological phenomena remains uncertain. Furthermore, any method that depends on gene expression data is limited by the ligand-receptor list's exhaustiveness, which may result in the exclusion of other cellular communication channels. As a result, there is still a need for methods that can evaluate the whole potential for cell-to-cell communication.

In addition several computational approaches and methods have been developed to predict cell-cell interactions using ligand-receptor interaction detection. Although many of these techniques have been tested theoretically, little research has been done on how well modern LR-based CCI prediction tools function in practice and what kind of results they provide when used with publicly available single-cell RNA sequencing datasets. Existing comparative studies of CCI tools mainly report their advantages and disadvantages based on the theoretical analysis [37]. There is a lack of running assessments to understand the performance and effectiveness of the most recent CCI inference tools in real application scenarios [38].

1.5.1. LR databases

The process of inferring intercellular communication from transcriptomics data relies on the concept of gene co-expression, in which one gene within a pair originates from one cell interacting with another gene from a second cell. This approach has been extensively explored in several studies, employing diverse methodologies such as examining co-expression patterns of all genes [39], assessing similarity between gene expression profiles [40], and analyzing regulatory networks [41].

A common strategy in these studies involves utilizing literature-curated lists of interacting proteins, with a particular emphasis on ligands and their corresponding receptors. These curated lists (Ligand-Receptor databases) serve as valuable resources for interpreting the biological significance of the observed interactions. Over time, numerous databases have been established, housing an extensive collection of ligand-receptor pairs. This wealth of data has greatly facilitated the comprehensive investigation of communication processes between cells. However, integrating data from multiple sources poses its challenges and validating predicted protein-protein interactions (PPIs) is crucial to ensure the reliability of the findings and to minimize the risk of false positives.

Recent advancements in computational tools have addressed some of these challenges by incorporating information about multimeric proteins and interactions between ligand-receptor complexes. By considering subunit co-expression, these tools offer a more accurate representation of functional interactions, particularly for proteins that require multisubunit assembly for proper functioning.

Furthermore, efforts have expanded beyond only focusing on ligand-receptor pairs to incorporate other aspects of CCC, including metabolite interchange and the activation of intracellular signaling pathways [42]. However, incorporating downstream signaling gene products and regulatory networks into the analysis requires additional information on ligand-receptor pairs, which can be a labor-intensive process and may be sensitive to the quality of available databases.

Despite these challenges, protein-protein interactions, especially ligand-receptor pairs, remain essential in deciphering CCC in various biological contexts. They serve as basic elements in understanding the intricate communication networks that govern cellular behavior and response.

Multi subunit database

When a ligand-receptor pairs database is described as "multi subunit," it typically refers to the fact that the receptor complexes involved in cell signaling are composed of multiple protein subunits.

In cellular signaling, ligands bind to receptors on the cell surface, initiating a cascade of intracellular events. Many receptors are not single proteins but rather complexes made up of multiple subunits. Each subunit plays a specific role in the receptor's function, such as ligand binding, signal transduction, or regulation of downstream pathways.

Describing a ligand-receptor pairs database as "multi subunit" indicates that it includes information about such complexes, including the ligands that bind to them and the specific subunits involved in the receptor complexes. This information is crucial for understanding the intricacies of cell signaling pathways and how they regulate various cellular processes.

1.6. Non-small cell lung cancer

The Italian Association of Medical Oncology (AIOM) and the Italian Association of Tumor Registries estimated about 41,500 new cases and 33,836 deaths from lung cancer in Italy in 2018, with a 5-year survival rate of 16% and a 10-year survival of 12% (11% for men and 15% for women). Currently, lung cancer represents the third most common neoplasm in the overall Italian population, and it is the first cause of cancer death in men and the third in females, with significant differences observed across the different Italian regions [43]. Although cigarette smoking is the main cause, anyone can develop lung cancer. Lung cancer is highly treatable, no matter the size, location, whether the cancer has spread, and how far it has spread.

The lungs contain many different types of cells. Most cells in the lung are epithelial cells. Epithelial cells line the airways and make mucus, which lubricates and protects the lung. The lung also contains nerve cells, hormone-producing cells, blood cells, and structural or supporting cells.

There are 2 main classifications of lung cancer: small cell lung cancer and non-small cell lung cancer (NSCLC). These 2 types are treated differently.

NSCLC begins when healthy cells in the lung change and grow out of control, forming a mass called a tumor, a lesion, or a nodule. This can begin anywhere in the lung. The tumor can be

cancerous or benign. When a cancerous lung tumor grows, it may shed cancer cells. These cells can be carried away in blood or float away in the fluid, called lymph, that surrounds lung tissue. Lymph flows through tubes called lymphatic vessels that drain into collecting stations called lymph nodes. Lymph nodes are the small, bean-shaped organs that help fight infection. They are located in the lungs, the center of the chest, and elsewhere in the body. The natural flow of lymph out of the lungs is toward the center of the chest, which explains why lung cancer often spreads there first. When a cancer cell moves into a lymph node or to a distant part of the body through the bloodstream, it is called metastasis.

There are different types of NSCLC. It is important to know the type of NSCLC because it can change treatment options. Doctors determine which type of NSCLC a person has based on the way the cancer looks under a microscope and the kind of cells the cancer starts in.

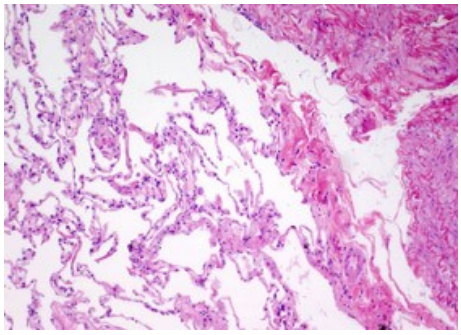

Figure 8. *Histology of a normal lung tissue. Image taken from [44].*

The different types of NSCLC are:

Adenocarcinoma (LUAD)

This type of NSCLC begins in the epithelial cells that line the outside of the lungs. These cells make mucus. It is the most common type of lung cancer at about 40% of all NSCLC cases.

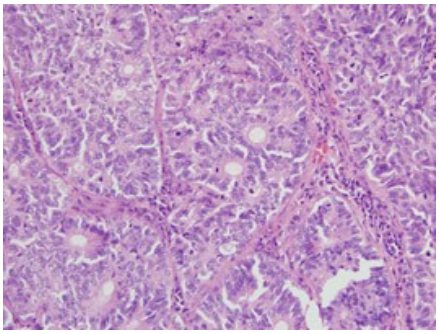

Figure 9. *Histology of lung adenocarcinoma. Image taken from [44].*

Squamous cell carcinoma (LUSC)

This type of cancer starts in the squamous cells, which are flat cells that line the inside of the lungs. About 30% of all NSCLC cases are squamous cell carcinoma.

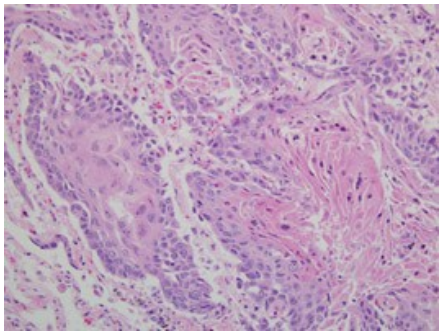

Figure 10. *Histology of squamous cell carcinoma. Image taken from [44].*

Large cell carcinoma

The cells in large cell carcinoma do not look like adenocarcinoma or squamous cell carcinoma, instead they look like large cells. This is the least common type of NSCLC and as diagnostic tools get better, more large cell carcinomas are being classified as adenocarcinoma or squamous cell carcinoma.

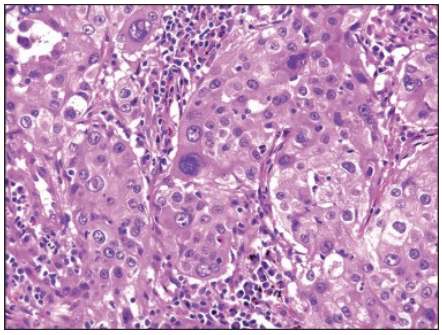

Figure 11. *Histology of large cell carcinoma. Image taken from [44].*

NSCLC-NOS (not otherwise specified) or NSCLC undifferentiated

It is a non-small cell lung cancer subtype that is difficult to classify. Due to overlapping features or insufficient data, doctors may find it difficult to definitively classify the tumor into particular histological subtypes, even after thorough testing, including histological inspection and molecular profiling. Consequently, tumors that are not easily classified into recognised subtypes like squamous cell carcinoma or adenocarcinoma are included in NSCLC-NOS.

Doctors can often classify NSCLC with a “stage” when giving a person their diagnosis. Staging is a way of describing where the cancer is located, if or where it has spread, and whether it is affecting other parts of the body. Doctors use diagnostic tests to find out the cancer’s stage, so staging may not be complete until they finish all the tests. Knowing the stage helps your doctor decide what kind of treatment is best. The stage can also help predict your prognosis. The stage of NSCLC is described by a number, from 0 through 4 (Roman numerals I through IV).

1.7. Aim of the project

Gene expression profiling has found its natural placement and development in cancer: the disease most closely associated with progressive alteration of the cell's genome, through complex processes ranging from mutations to the loss or acquisition of different genes to alterations in transcription and translation processes. Given that a considerable proportion of genomic and proteomic research has focused on cancer patients, it becomes imperative, upon identifying patients at higher risk, to determine the most suitable therapy for each individual. In this regard, gene expression analysis plays a pivotal role in providing essential insights into how a patient may respond to a particular treatment. Moreover, genomic analysis enables the discovery of novel treatment targets by identifying a list of genes differentially expressed in tumor tissue compared to normal tissue, or in individuals with a better prognosis compared to those with a worse prognosis. This makes it possible to design new medications and evaluate how well they work as well as any possible negative effects while treating a particular pathology.

Due to the recent advancements in single-cell technologies, a significant amount of single-cell RNA sequencing data has been publicly available. The investigation of CCIs at single-cell resolution, especially LR-based CCIs, was driven by the availability of these data. Dozens of computational techniques and tools have been developed to forecast CCIs. A large number of these tools have undergone theoretical reviews. Nevertheless, limited research has been conducted on the performances of existing LR-based CCI prediction algorithms and their outcomes when applied to publicly available scRNA-seq datasets. These approaches lack a comprehensive validation of predicted ligand-receptor interactions. Since there is no ground

truth, the challenge lies in validating and measuring the performances in terms of accuracy, sensitivity, and precision of the prediction made by these tools. Currently, there is no gold standard for the validation of these tools, as many of them differ in the number of ligand-receptor pairs found and in the overlap of their results. Therefore, establishing a robust validation framework becomes essential to assess the reliability and consistency of the predictions generated by these tools.

This highlights the absence in the field of a reference benchmark to compare the results obtained by individual algorithms; the only possibility is to compare the computationally obtained result with what has been derived from laboratory analyses and that is present in the literature.

To close this gap, in this work, the aim is to test three computational methods for the analysis of intercellular communication on data from a scRNA-seq atlas created at the Centre for Chemistry and Biomedicine (CCB) in Innsbruck, containing gene expressions of non-small cell lung cancer cells.

The three tools to be compared are scSeqComm [45], CellphoneDB and NicheNet [47].

Initially, the atlas and the characteristics of the ligand-receptor pairs databases that will be used for comparison will be presented. Subsequently, the fundamental aspects of the CCI analysis structure of each individual method and the main differences between them will be shown.

Importance is then given to the procedural choices and requirements adopted for this specific objective, focusing on the implementation aspects that differentiate each method. It is crucial to specify that the focus is on the results of intercellular communication, not intracellular communication, limiting observations to the ligand-receptor pairs that the methods are capable of providing at the end of the analysis.

The specific communication of interest for the study is that which occurs in the tumor microenvironment and follows the direction from tumor cells, considered in this case as sender cells, towards immune system cells, considered as receiver cells. The methods were initially run on the entire available cellular atlas; subsequently, however, two subcategories of non-small cell lung cancer, LUAD and LUSC, were selected to verify if these two different types of tumor cells interacted differently with nearby immune system cells.

These same operations were repeated for three different ligand-receptor pairs databases: Eferemova [46], Browaeys [47], and Jin [48]. To evaluate the performances of these three

methods, execution time, RAM memory space occupied, and the number of LR pairs obtained from each method were tracked.
The results are presented at the end of the thesis in the form of Venn diagrams to evaluate the number of LR pairs calculated by each method and the overlap of the pairs found in common. Finally, only the pairs common to all three methods are visualized through circos plots, and to provide a clinical and functional context to the findings, a biological explanation of the most important LR pairs is provided.

At the end, it will be recognized how the methods exhibit significant differences in their CCC analysis framework and how they have different goals. As a result, different strategies and techniques will need to be used to execute each of them, leading to different execution times and outcomes. However, the methods will still show a 10% agreement in the detected cell-cell communication, and scSeqComm will prove to be the method with the simplest and most intuitive application and implementability.

Chapter 2: Cell-cell communication analysis of non-small cell lung cancer cell atlas

2.1. NSCLC atlas

The technical advances in single-cell RNA sequencing technologies enabled the dissection of the complex NSCLC tumor microenvironment (TME) in different stages, and numerous scRNA-seq NSCLC studies have identified a so far underestimated TME heterogeneity in early and advanced disease. Furthermore, these studies highlighted the importance of small cell populations in governing essential biological pathways such as immune cell activation or trafficking by tumor endothelial cells [49]. However, a major limitation of these studies is the limited number of analyzed patient samples per study. Moreover, the lack of genomic data as well as long-term follow-up data prevents comprehensive dissection of the biological heterogeneity and its potential contribution to therapy resistance and survival outcome. Technical and methodological variations between the different studies result in significant inconsistencies and knowledge gaps. As such, not all cell types (e.g., neutrophilic granulocytes) have been portrayed in the same depth and extension yet, posing an unmet need to characterize these populations as well. In NSCLC, it is well accepted that next to cancer cells, leukocytes compose the majority of cells within the TME. Particularly since immunotherapy is routinely used in clinical practice, in-depth characterization of the cancer immune cell compartment has been intensively pushed forward, and diverse cellular subsets have been profiled [50]. Previous compositional analyses by flow cytometry as well as histological work ups suggested that neutrophils compose a significant proportion of all tumor-resident leukocytes, with an estimated abundance ranging from 8% to 20%. Intriguingly, when looking at the scRNA-seq studies in NSCLC published over the last years, neutrophils are clearly under-represented. This discrepancy is most likely based on technical issues rather than on biological phenomena, but its clarification is of immense importance for the fundamental immunological understanding of NSCLC and for potential translational clinical investigations.

To overcome the above-mentioned hurdles, Salcher et al. [51] compiled major publicly available datasets into a comprehensive NSCLC scRNA-seq atlas covering 232 patients with NSCLC and 86 non-cancer controls. Additionally, given the scarcity of neutrophil single-cell data, they complemented the atlas by analyzing samples from 17 patients with NSCLC using a platform that captures cells with very low transcript count and carried out deep characterization of tissue-resident neutrophils (TRNs) including both tumor-associated neutrophils (TANs) and normal adjacent tissue-associated neutrophils (NANs).

The results of the work done by the researchers is the generation of a core large-scale NSCLC single-cell atlas.
First they developed a core NSCLC atlas by compiling scRNA-seq data from 19 studies and 21 datasets comprising 505 samples from 298 patients. This comprehensive NSCLC single-cell atlas integrates expert-curated, quality-assured, and pre-analyzed transcriptomic data from publicly available studies as well as their own dataset (UKIM-V) in early and advanced stage NSCLC of any histology.

In total, the core atlas includes transcriptomic data from 212 patients with NSCLC and 86 control individuals, comprising 196 tumor samples and 168 non-tumor control samples. Of the 212 patients with NSCLC, 156 were histopathologically annotated as lung adenocarcinoma (LUAD), 41 as lung squamous cell carcinoma (LUSC), and 15 were not otherwise specified (NSCLC NOS). NSCLC samples include tissue of the primary tumor (n = 176) or metastasis (n = 45) that were obtained either by surgical resection or by computed tomography- and bronchoscopy-guided biopsies. They clustered the disease stages of the patients with NSCLC as early (UICC stage I–II) versus advanced (UICC III–IV) diseases, as not all studies provided sufficient information on tumor stages.
Among the control samples, 89 were derived from distant non-malignant tissue of patients with lung tumors (annotated as normal_adjacent), of which 65 have a patient-matched tumor sample. Further, 10 samples were derived from non-tumor-affected lymph nodes of patients with NSCLC (annotated as normal) and 79 samples from patients without evident lung cancer history (annotated as normal). Of the control patients, 18 had a history of chronic obstructive pulmonary disease (COPD). Overall, the core atlas integrates 898,422 single cells, which are annotated to 12 coarse cell-type identities and 44 major cell subtypes or cell states (e.g., dividing cells) based on previously established canonical single-cell signatures including

169,223 epithelial cells, 670,409 immune cells, and 58,790 stromal and endothelial cells (Figure 12). They also annotated the cell-type composition for each dataset, the tissue of origin, and the patients within the core atlas. Previous scRNA-seq studies discriminated the clinically relevant types of LUSC and LUAD.

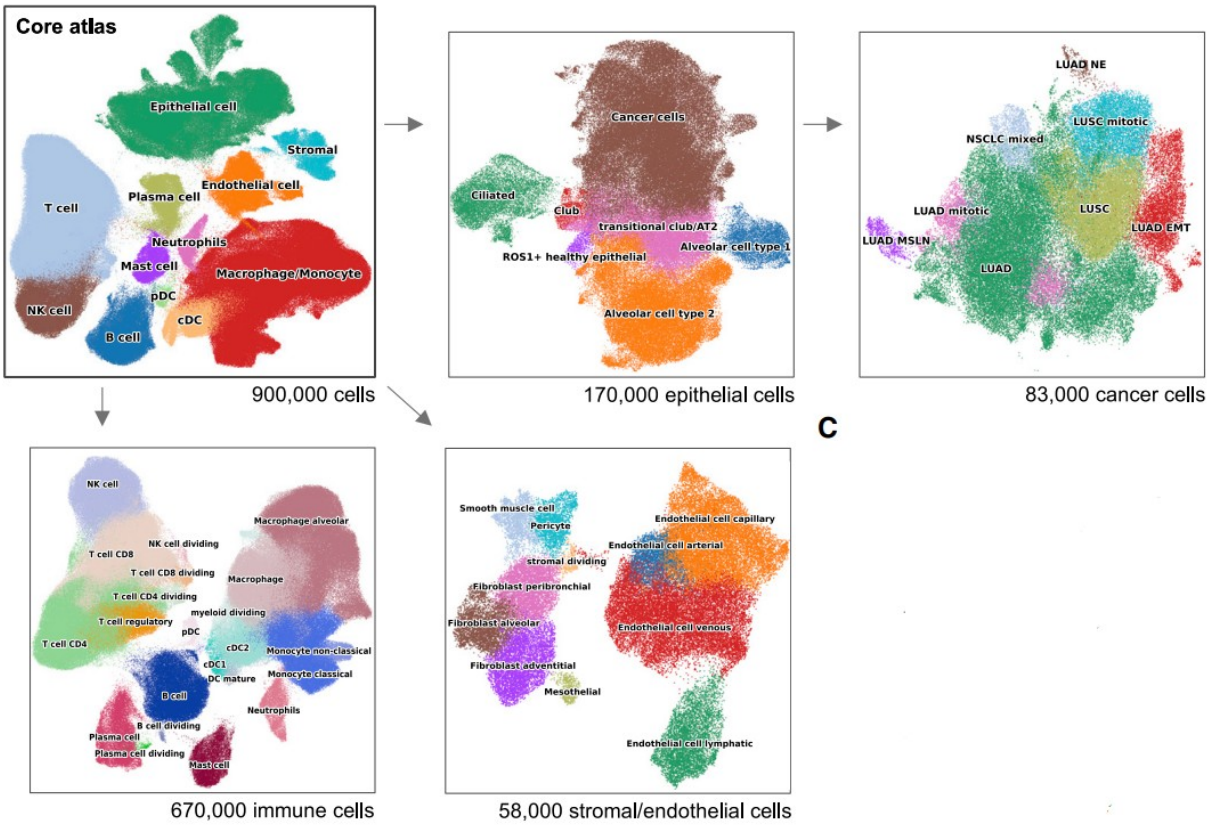

Figure 12. *Overview of the core NSCLC atlas and the epithelial, immune, and stromal/endothelial components depicted as uniform manifold approximation and projection (UMAP) plots.*
Image adapted from [51].

From this available cell atlas, it was decided to proceed with the analysis of cell-cell communication by examining the communication occurring from tumor cells, considered as sender cells, to immune system cells, considered as receiver cells. Therefore, the immune system cells included in the analysis from the atlas are as follows.

'B cell': also known as B lymphocytes, are a type of white blood cell that plays a critical role in the immune system. They are responsible for producing antibodies, which are proteins that recognize and neutralize foreign invaders such as bacteria, viruses, and other pathogens.

'DC mature': Mature dendritic cells (DCs) are a specialized type of immune cell that plays a crucial role in initiating and regulating immune responses. These cells are considered professional antigen-presenting cells (APCs), meaning they are adept at capturing, processing, and presenting antigens to other immune cells, such as T cells.

'Macrophage': Macrophages are a type of white blood cell that plays a pivotal role in the immune system's defense against pathogens, tissue repair, and regulation of inflammation. They are highly versatile cells found throughout the body in various tissues, where they act as phagocytes, engulfing and digesting foreign particles, dead cells, and cellular debris.

'Mast cell': Mast cells are a type of white blood cell that is primarily known for its role in allergic reactions and inflammation. They are found in connective tissue throughout the body, particularly near blood vessels and nerves, where they play a crucial role in the body's immune response. When activated, mast cells release a variety of mediators, including histamine, cytokines, and chemotactic factors, which play a central role in initiating and regulating inflammatory and allergic responses.

'Monocyte': Monocytes are a type of white blood cell that is produced in the bone marrow and circulates in the bloodstream. They are considered part of the innate immune system and play several important roles in immune responses and tissue repair. When monocytes leave the bloodstream and enter tissues, they differentiate into macrophages.

'NK cell', Natural killer (NK) cells are a type of white blood cell that plays a critical role in the innate immune system's defense against viral infections and certain types of cancer. They are called "natural killers" because they have the ability to recognize and kill infected or abnormal cells without the need for prior activation or the presence of antibodies.

'Neutrophils': Neutrophils are the most abundant type of white blood cell and play a fundamental role in the innate immune system's defense against infections. They are part of the body's first line of defense and are among the first immune cells to migrate to sites of infection or tissue damage. They perform phagocytosis, production of Reactive Oxygen Species (ROS), release of granules, formation of neutrophil extracellular traps (NETs).

'Plasma cell': also known as effector B cells, are a specialized type of white blood cell that plays a central role in the adaptive immune response. They are derived from activated B cells

and are responsible for producing large quantities of antibodies, also known as immunoglobulins (Ig), that circulate in the bloodstream and target specific pathogens or foreign substances.

'T cell': also known as T lymphocytes, are a type of white blood cell that plays a central role in the adaptive immune response. They are produced in the bone marrow and mature in the thymus gland, which is where they derive their name. T cells are critical for coordinating immune responses against specific pathogens and abnormal cells. They are characterized by the presence of T cell receptors (TCRs) on their cell surface, which allow them to recognize specific antigens presented by other cells. T cells can recognize a wide range of antigens, including those derived from pathogens, cancer cells, and even self-antigens in cases of autoimmune diseases.

'cDC': Conventional dendritic cells (cDCs), they act as professional antigen-presenting cells (APCs) that capture, process, and present antigens to T cells, thereby initiating specific immune responses.

'pDC': Plasmacytoid dendritic cells (pDCs) are a specialized subset of dendritic cells that play a crucial role in the immune system's defense against viral infections and regulation of immune responses. They are primarily known for their ability to produce large amounts of type I interferons (IFNs) in response to viral infections, which are essential for antiviral defense and immune activation.

2.2. Ligand-Receptor databases used in this study

Let's now examine the features of the databases that were used for the analyses related to this thesis project (Table 1).

2.2.1. Database Efremova 2020

A public database of ligands, receptors, and their interactions was established by Efremova et al. [46] to facilitate an in-depth, systematic investigation of the molecules involved in cell-cell communication. Their database is based on manual curation of particular groups of proteins involved in cell-cell communication, as well as annotation of receptors and ligands using publicly available resources. They incorporate subunit architecture to accurately describe heteromeric complexes, both for ligands and receptors. This is crucial because, contrary to what most databases and research utilise, cell-to-cell communication depends on multi-subunit protein complexes. They created a SQLite relational database to combine all the data in a modifiable, distributable, and adaptable setting.

The database stores a total of 978 proteins: 501 are secreted proteins and 585 are membrane proteins. These proteins are involved in 1,396 interactions; out of all proteins stored in CellPhoneDB, 466 are heteromers. There are 474 interactions that involve secreted proteins and 490 interactions that involve only membrane proteins. There are a total of 250 interactions that involve integrins.

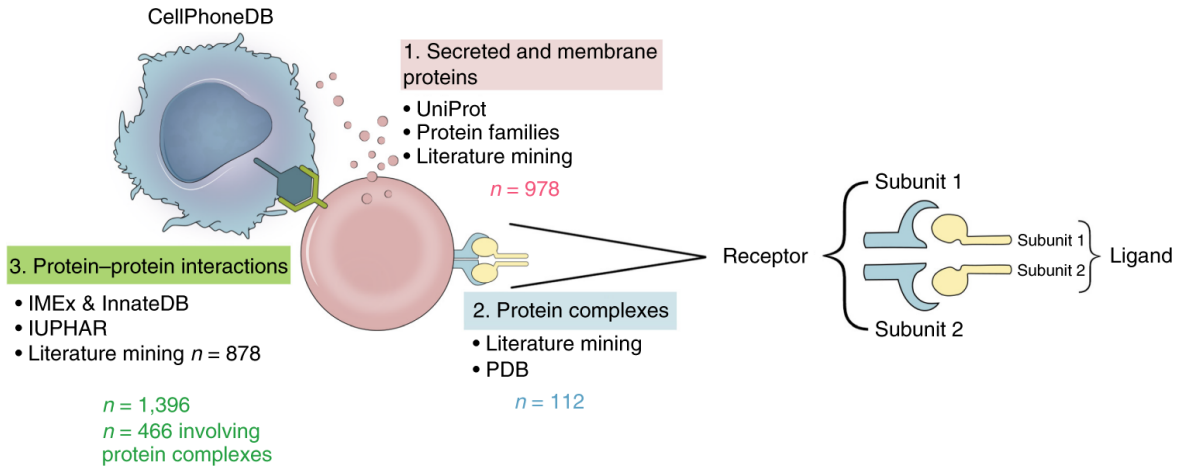

Figure 13. *Overview of Efremova database. Image taken from [46].*

2.2.2. Database Browaeys 2019

Browaeys et al. [47] for their method called NicheNet created integrated networks considering a collection of ligand–receptor, intracellular signaling and gene regulatory interactions.
Ligand–receptor interactions were collected from KEGG [52] (Kyoto Encyclopedia of Genes and Genomes), Ramilowski et al. [53] and IUPHAR Guide to Pharmacology [54] (via Harmonizome [55]). In addition to this, they predicted ligand–receptor interactions by searching in protein–protein interaction databases for interactions between genes annotated as ligands and receptors.
The result is a database containing 12,659 LR pairs with annotation of sources and databases but it doesn't support protein multi-subunit structure.

2.2.3. Database Jin 2020

To construct a database of ligand-receptor interactions that comprehensively represents the current state of knowledge, Jin et al. [48] manually reviewed other publicly available signaling pathway databases as well as peer-reviewed literature. The majority of ligand–receptor interactions were manually curated on the basis of KEGG [52] signaling pathway database. Additional signaling molecular interactions were gathered from recent peer-reviewed experimental studies. They took into account not only the structural composition of ligand-receptor interactions, which often involve multimeric receptors, but also cofactor molecules, including soluble agonists and antagonists, as well as co-stimulatory and co-inhibitory membrane bound receptors that can prominently modulate ligand-receptor mediated signaling events. To further analyze cell-cell communication in a more biologically meaningful way, they grouped all of the interactions into 229 signaling pathway families, such as WNT, ncWNT, TGFβ, BMP, Nodal, Activin, EGF, NRG, TGFα, FGF, PDGF, VEGF, IGF, chemokine and cytokine signaling pathways (CCL, CXCL, CX3C, XC, IL, IFN), Notch, and TNF. The supportive evidences for each signaling interaction is included within the database.
The result is a human database containing 2,005 LR pairs that take into account protein complexes and functional annotations.

Database	Organism	Number of LR pairs	Ligand/Receptor multi-subunit
Efremova et al. (2020)	Human	878	Yes
Browaeys et al. (2019)	Human	12,659	No
Jin et al.(2020)	Human	2,005	Yes

Table 1. *Overview of available lists of ligand-receptor pairs in literature used in this thesis project.*

2.3. Computational tools compared in this study

In the next paragraphs an overview of the three methods compared in this thesis is provided.

2.3.1. scSeqComm

scSeqComm [45] is a computational approach introduced to deduce, quantify, and delineate both intercellular and associated intracellular signaling pathways from single-cell RNA sequencing data.

The primary innovation of this tool lies in the development of a novel methodology to identify and quantify intercellular signaling patterns derived from scRNA-seq datasets. This recently suggested scoring system supports the intricate multi-subunit structure of ligands and receptors present in modern ligand-receptor databases. It adopts a more conservative approach compared to conventional methods, with the specific aim of reducing and prioritizing experimental targets effectively.

Furthermore, the computational framework is able to measure the impact of ongoing intracellular signaling within the recipient cells, focusing on the activation of established biological signaling pathways. By quantifying evidence of both intercellular and intracellular signaling, the method allows for the integration of these two components, resulting in a more robust inference of cellular communication from scRNA-seq data.

Additionally, genes associated with the identified cellular communication effects are leveraged to conduct integrated Gene Ontology (GO) enrichment analyses, facilitating the functional characterization of cell-cell communication effects.
The proposed methodology is implemented in the R package called scSeqComm, accessible at the provided URL https://gitlab.com/sysbiobig/scseqcomm.

In Figure 14 a graphical overview of the scSeqComm R package is provided, illustrating its functionalities and workflow.

Intercellular signaling

The method requires as input a normalized scRNA-seq gene expression matrix with indication about cell clusters.
To identify intercellular crosstalk across different clusters of cells, the method first assigns a score to each ligand and each receptor expressed in a specific cluster of cells. Then, for each known ligand–receptor pair between two groups of cells or within the same group, it infers an ongoing intercellular communication as a function of the ligand score and the receptor score.
For each ligand or receptor gene (g) in a specific cluster (k), it calculates a score to measure how much the gene's expression level is higher compared to what would be expected by chance. It determines the expected distribution of average gene expression levels, using a permutation approach, by randomly shuffling the genes in the cluster and computing the average expression levels multiple times. This distribution follows a normal distribution, even if the original data is not normally distributed, thanks to the central limit theorem. The score is then calculated as the probability of observing values lower than the gene's average expression level when sampling from this normal distribution. This approach also considers factors like gene expression variability and the number of cells in the cluster. Additionally, if a ligand or receptor consists of multiple subunits, the score is computed as the geometric mean of the scores of its subunits. This ensures that if any subunit is inactive, the overall score will be zero.
This scoring method helps understand the activity of signaling molecules like ligands and receptors in different groups of cells. But to consider ongoing cell communication, both the ligand and its matching receptor need to be active. So, a new score called the intercellular signaling score is introduced. This score, denoted as Sinter, measures the likelihood of ongoing communication between two cell clusters, k1 and k2, through a specific ligand-receptor pair (l, r), where l is expressed in k1 and r is expressed in k2. The method calculates

the intercellular signaling score as the minimum score between the ligand and receptor scores thus implementing a 'fuzzy logical AND' operator, reflecting that communication is ongoing only when both are active. This method ensures that even if one gene's expression is much higher than the other, it doesn't overshadow the overall signaling intensity. However, different combinations of ligand and receptor scores might give the same intercellular signaling score. To address this, another score is introduced, the intracellular signaling score.

Intracellular signaling

As second step, scSeqComm quantify the intracellular signaling associated to cell–cell communication measuring the evidence of a transcriptional response in target genes regulated by known TFs and weighting the association between TFs and upstream receptors using available biological knowledge from signaling pathway databases and regulatory network databases.

As third step, scSeqComm performs a GO enrichment analysis on target genes associated with the detected intracellular signaling to functionally characterize the effect of the detected cell–cell communication. The tool provides as output, in both a tabular and graphical form, the inferred evidence of ongoing intercellular and intracellular signaling, as well as the functional characterization from GO analysis.

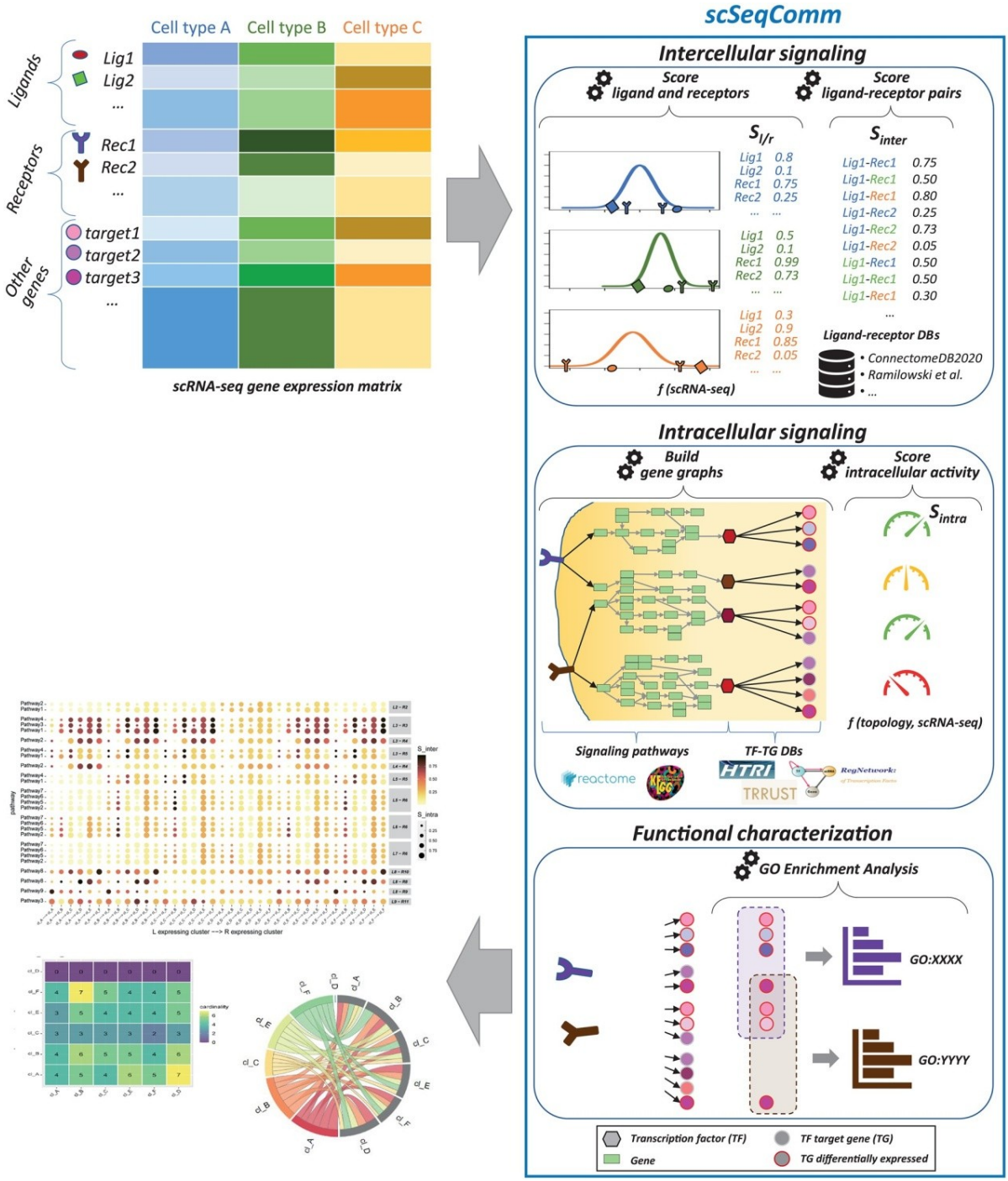

Figure 14. *Schematic overview of the scSeqComm pipeline.*
Image taken from [45].

2.3.2. CellphoneDB

In CellPhoneDB v2.0 [46] multiple subunit architecture is taken into account for both ligands and receptors. Additionally, CellPhoneDB has assembled a LR database which contains multiple LR subunits. Significant LR interaction pairs are identified based on their respective cell-type enrichment's likelihood estimation. The CCIs unique to two cell types are ranked according to the number of meaningful LR interaction pairs. It is possible to build CCI-based networks to evaluate cellular crosstalk between various cell types. Runtime and memory consumption are decreased by using the cell subsampling technique.

Using scRNA-seq data, this computational method finds interacting ligand-receptor couples that are biologically significant. Following the upload of the scRNA-seq data and the use of geometric sketching subsampling [56] (Figure 15a), cells that have the same cluster annotation are combined into a single cell state. Based on the expression of a ligand by one cell state and a receptor by another, it determines enhanced ligand–receptor interactions between two cell states. The mean gene expression and the proportion of cells expressing each gene in the cluster are computed (Figure 15b). It determines which ligand-receptor combinations exhibit considerable cell-state specificity by taking into account the expression levels of ligands and receptors inside each cell state and applying empirical shuffling (Figure 15c,d). This produces possible cell-cell communication networks and predicts molecular interactions between cell populations via certain protein complexes, which may be visualised with the use of simple tables and charts (Figure 15e). Because some ligand-receptor pairings are widely expressed by the cells in a tissue, they are not indicative of specific communication between various cell states, which makes the specificity of the ligand-receptor interaction crucial. The computational code requires Python v.3.5 or higher and it can be found on GitHub at https://github.com/Teichlab/cellphonedb, and www.CellPhoneDB.org hosts an easy-to-use web interface. It is advised to use the first option for datasets larger than 10 GB.

Statistical inference of ligand–receptor specificity

To assess cellular crosstalk between different cell types, it has been used a statistical framework for inferring cell–cell communication networks from scRNA-seq data. The method predicts enriched receptor–ligand interactions between two cell types on the basis of expression of a receptor by one cell type and a ligand by another cell type. To identify

biologically relevant interactions, it looks for cell-type-enriched ligand–receptor interactions. Only receptors and ligands expressed in more than a user-specified threshold percentage of the cells in the specific cluster are considered for the analysis (default is 10% and it is the value that has been used also in this analysis). It then performs pairwise comparisons between all cell types in the dataset. First, it randomly permutes the cluster labels of all cells (1,000 times by default) and determines the mean of the average ligand expression level in a cluster and the average receptor expression level in the interacting cluster. In this way it generates a null distribution for each ligand–receptor pair in each pairwise comparison between two cell types. It obtains a P value for the likelihood of cell-type enrichment of each ligand–receptor complex by calculating the proportion of the means that are as high as or higher than the actual mean. On the basis of the number of significant pairs, it then prioritizes interactions that are highly specific between cell types, so that the user can manually select biologically relevant ones. For multi-subunit heteromeric complexes, it requires that all subunits of the complex be expressed (using a user-specified threshold), and it uses the member of the complex with the minimum average expression for random shuffling.

The results of the analysis are compiled in an output folder comprising four .txt files: *significant_means.txt*, *pvalues.txt*, *means.txt*, and *deconvoluted.txt*. Additionally, the visualization of the results can be facilitated through intuitive tables, plots, and network files provided by the package.

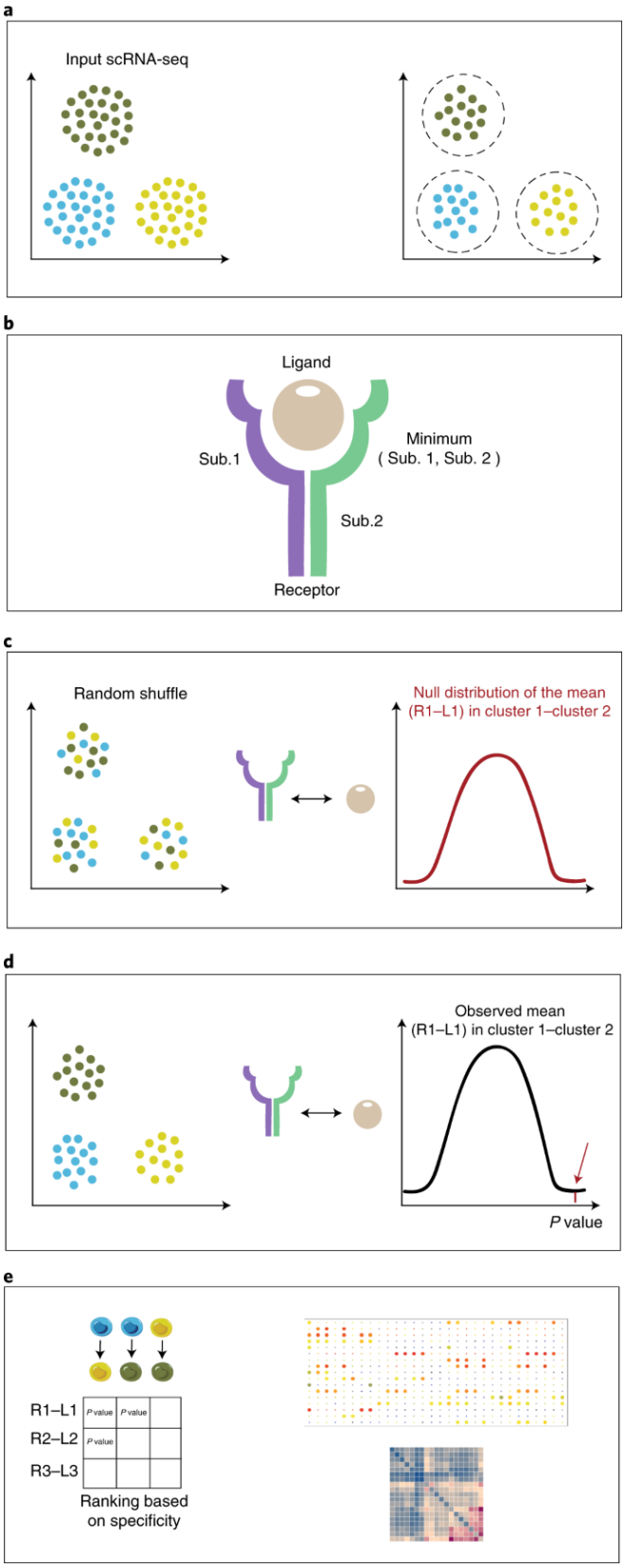

Figure 15. *Overview of the statistical method framework used in CellphoneDB to infer ligand–receptor complexes specific to two cell types from single-cell transcriptomics data.*
Image taken from [46].

2.3.3. NicheNet

NicheNet [47] is a computational approach, available on an R package, that aims to clarify the functional understanding of CCIs by assuming the functional impact of ligands in the sender cells on the expression of genes in the recipient cells, known as target genes. NicheNet combines data on gene regulatory interaction, signal transduction, and LR interaction to achieve that. The IUPHAR/BPS Guide to Pharmacology, Reactome, KEGG, and PPI databases were used to compile the LR interactions [52-55]. Individual interactions were arranged as weighted networks, and gene regulatory interactions were transformed into a weighted gene regulatory network. LR and signalling networks were joined to form a weighted ligand-signaling network. To combine several data sources, a weighted sum of the distinct networks was subsequently calculated. Network operations, such as PageRank, were applied to this integrated network to derive a prior model of ligand-target regulatory potential.

When applying NicheNet to investigate communication between cells, this general prior model of ligand–target regulatory potential is combined with the gene expression data given as input. NicheNet infers active ligand–target links between interacting cells by combining their expression data with this prior knowledge model on ligand–target links (Figure 16).

The gene expression data from the atlas, once loaded into R, is structured as a Seurat Object [57], this means that to utilize NicheNet effectively, familiarity with the functions used to modify a Seurat object is essential. A Seurat object is a type of data structure employed in single-cell analysis. Seurat, widely used in R, offers a range of functionalities for exploring, visualizing, and analyzing data generated by techniques like single-cell RNA sequencing. In essence, a Seurat object represents a collection of single-cell data that has been preprocessed, normalized, and analyzed using Seurat's capabilities. It contains information about the cells, their genomic characteristics, and their relationships within a specific single-cell experiment [58].

After combining the expression profiles of interacting cells, NicheNet can prioritize the regulatory potential of ligands on the target genes.

First, NicheNet prioritises which sender cell ligands are most likely to have affected the gene expression in interacting receiver cells. This procedure, called ligand activity prediction, ranks ligands according to how well their prior target gene predictions correspond to the observed gene expression changes resulting from communication with sender cells. To then predict active ligand-target links, NicheNet searches for genes that are affected in receiver cells and are possibly regulated by these prioritized ligands as indicated with a high regulatory potential score. Finally, users can visualize possible signaling paths between ligands and target genes of interest to analyze why the model infers specific ligand–target links.

As observed, the primary output obtained by applying NicheNet is the interaction between ligands and target genes. Only subsequently it is possible to obtain the ligand-receptor interaction, which is the one required for comparing the methods in this thesis.

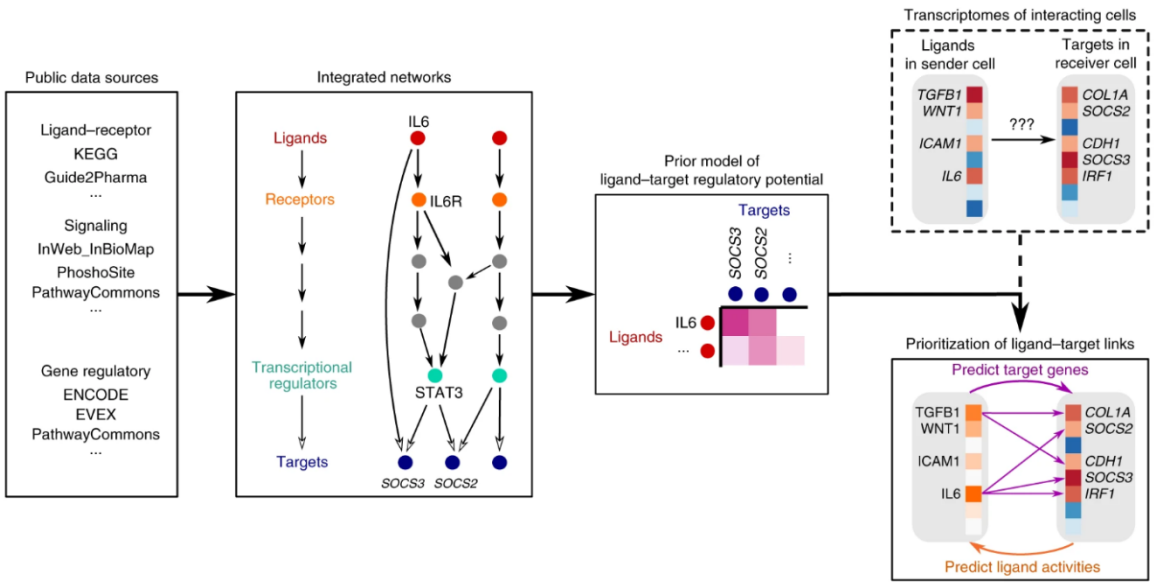

Figure 16. *General workflow of NicheNet.*
Image adapted from [47].

Chapter 3: Execution of cell-cell communication analysis methods

3.1. Materials, tools and computing environment

The CCB of Innsbruck is equipped with a computing infrastructure based on HPC (High-Performance Computing).
HPC, or High-Performance Computing, refers to the utilization of powerful computing systems capable of delivering significantly higher computational power and speed compared to traditional desktop or workstation computers. These systems typically consist of multiple interconnected processing units, such as CPUs or GPUs, working in parallel to perform complex calculations and data processing tasks. HPC systems are widely used in scientific research, engineering simulations, data analysis, and other computationally intensive applications where large volumes of data need to be processed or complex mathematical models need to be solved. They offer advantages such as faster processing times, the ability to handle massive datasets, and scalability to accommodate increasing computational demands.

More specifically, what was utilized for this thesis work was the High Performance Computing Cluster (HPCC), which is a type of computing infrastructure composed of multiple interconnected computers, or nodes, working together to perform complex computational tasks at high speeds. This cluster is designed to handle massive amounts of data and perform computations much faster than traditional computers.

This cluster is based on Linux, and its hardware characteristics are as follows:

- 1 x Head Node: zeus.icbi.local
  - 64 CPU cores / 3.0 TB RAM
  - 2 x 10 GBit ceph storage network, 1 x 1 Gbit cluster network
  - 2 x 480 GB SSD RAID for system
  - 2 x 1.6 TB SSD RAID for local scratch, OS/Tools mirror, backup

- 10 x Compute Nodes: apollo-01 … apollo-10
  - 44 CPU cores / 1.0 TB RAM
  - 2 x 10 Gbit ceph storage network, 1 x 1 Gbit cluster network
  - 2 x 480 GB SSD RAID for system
  - 2 x 800 GB SSD RAID for local scratch

The R and Python computing environments were already available on the cluster. The analyses conducted using R packages such as scSeqComm and NicheNet were performed entirely on the Head Node named Zeus, as well as the preparation of input data for CellphoneDB, which was carried out in Python.
Regarding the CellphoneDB analyses instead, they were launched utilizing a scheduler organizing the work on the compute nodes Apollo.

Specifically, the job scheduling system used is SGE (Sun Grid Engine). In practice, the Sun Grid Engine queuing system is valuable when there is a large number of tasks to execute and one wants to distribute them across a cluster of machines. For instance, it might be necessary to run hundreds of simulations/experiments with varying parameters.
SGE operates by breaking down computational tasks into smaller units known as jobs and then assigning these jobs to available computing resources within the cluster or grid. It enhances resource utilization by scheduling jobs based on factors such as priority, resource requirements, and system load.
Utilizing a queuing system in such scenarios offers several advantages:

- *Job Scheduling*: It enables scheduling a virtually unlimited amount of work to be performed as resources become available. This allows submitting as many tasks (or jobs) as needed and letting the queuing system handle their execution.
- *Load Balancing*: It automatically distributes tasks across the cluster to prevent any single node from becoming overloaded compared to others.
- *Queue Management*: Jobs are submitted to queues, and SGE manages their execution based on queue configurations and policies. Queues can have different priorities, access controls, and scheduling policies to accommodate various user requirements and system constraints.

- *Monitoring/Accounting*: It provides the ability to monitor all submitted jobs and query which cluster nodes they're running on, their completion status, error encounters, etc. It also facilitates querying job history to track tasks executed on a given date, by a specific user, etc.

The use of this job scheduling program, specifically for the analyses conducted with CellphoneDB, means that the computational times and RAM usage for this method cannot be directly compared with those of the other two methods. Indeed, since it was run on a different node of the cluster, it exhibits different CPU and RAM characteristics compared to the head node. However, as it will be shown in Chapter 4 which is dedicated to the results, CellphoneDB turns out to be a computationally demanding method, and its execution time will be much greater compared to that required by the other two methods, scSeqComm and NicheNet.

When running R on an HPC cluster, the memory usage reported typically includes two main components: memory used by the session and memory used by the system.

***Memory used by the Session*:**
This refers to the amount of memory consumed by the R session itself, including all objects, variables, functions, and data loaded into memory during the current R session. It reflects the memory footprint of the specific R process or instance running on the cluster.
This memory usage is specific to the R session and is managed by R's memory management system. It includes all objects created and loaded into memory during the execution of R scripts or commands.

Memory used by the System:
This refers to the total amount of memory utilized by the operating system to run all processes and applications on the HPC cluster node, including the R session and any other concurrent processes. It includes memory used by the R session, as well as memory used by other system processes, background tasks, and system-level caching mechanisms. This memory usage is controlled and managed by the operating system's memory management system, which allocates and deallocates memory resources among different processes running on the node.

In summary, memory used by the session specifically pertains to the memory consumption of the R process or instance, while memory used by the system refers to the overall memory usage of the entire system, including the R session and other processes running on the cluster node so it provides an indication of the overall system workload and resource utilization, including the impact of the R session alongside other concurrent activities on the cluster node.

3.2. Launching scSeqComm

3.2.1. Input data

Firstly, to conduct the analyses with scSeqComm, it was necessary to install the package in the R environment. Following this, after loading the .rds file containing the NSCLC atlas, a dataframe was created with the first column containing cell names ("Cell_ID") and the second column containing their assigned clusters.
In particular, the cell clusters chosen for the analysis are those of the immune system previously described in Chapter 2. More precisely the intercellular communication of interest for the study involves tumor cells as sender cells and immune system cells as receiver cells.

Subsequently, the relevant gene expression matrix was extracted from the file, and to handle the large size of the variable, it was converted to a more memory-efficient data.structure using 'sparsetoBigMemory' function as suggested by the author.

3.2.2. Main function

The main scSeqComm function that was used is called 'scSeqComm_analyze'. It takes as input the gene expression matrix, the previously created dataframe, the LR pairs database, the Transcriptional Regulatory Networks, and the Receptor-Transcription factor a-priori association, which are already included in the installed package. The function was launched using 20 cores and choosing "Wilcoxon" as the method for calculating the differential expression. The function performs the intercellular and/or intracellular communication analysis for the given scRNA-seq dataset.

3.2.3. Output

At this stage, the tool has provided as output the inferred evidence of ongoing intercellular and intracellular signaling in the form of intercellular and intracellular scores for the respective ligand-receptor pairs, and the result can be visualized, for example, in a table of scores as shown in Figure 17. The results of the analysis can be expressed both in tabular form, consisting of tables containing the ligand, the receptor, the LR pair, the ligand-associated score, the receptor-associated score, the intercellular score, the intracellular score within the cluster expressing the receptor, and the respective cluster memberships of the ligand and the receptor. Additionally, the results can be represented graphically through heatmap visualization, aiding in the visualization of the scores.

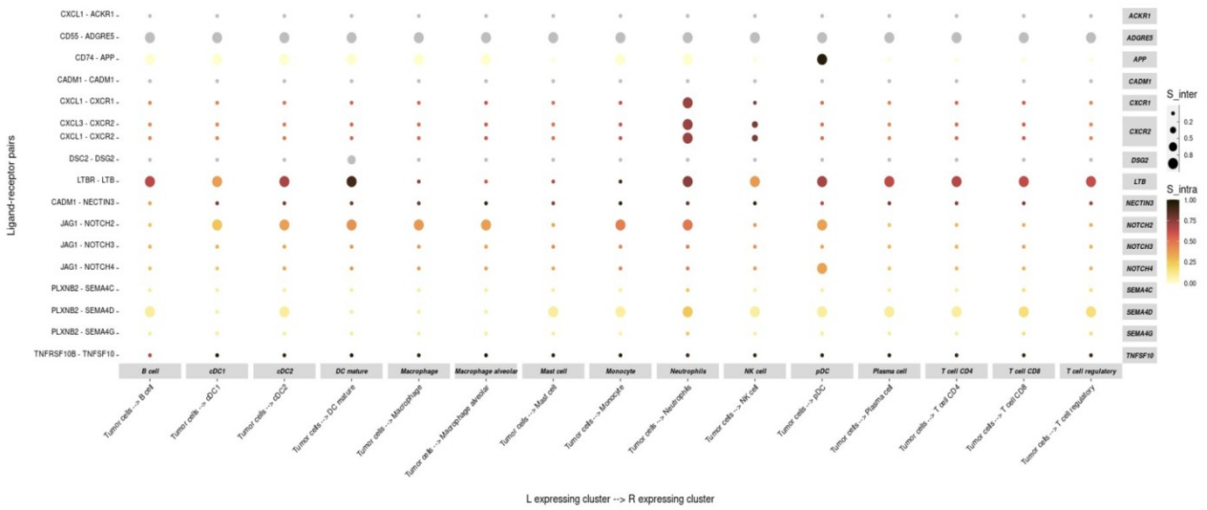

Figure 17*. Subset of LR pairs to visualize their scores.*

Subsequently, in order to select the ligand-receptor pairs of greatest interest for the study, the focus was placed on the intercellular score, which is a number between 0 and 1 returned for each ligand-receptor pair, where a value of zero indicates minimal interaction, while a value of 1 indicates maximum interaction.

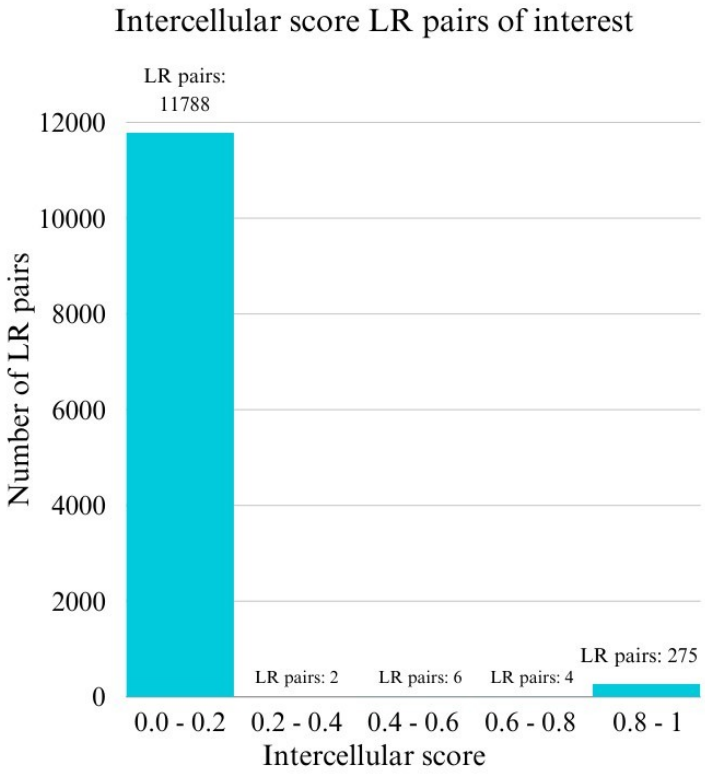

Figure 18. *Histogram of intercellular scores for scSeqComm in the case of Efremova database and running on the entire atlas.*

After observing the distribution of scores in the histogram (Figure 18), it was decided to consider only the LR pairs with a score greater than or equal to 0.8 in order to select communications with the highest interactions. Furthermore, the pairs that would be missed by using a lower threshold, such as 0.5 (default), were not many compared to the pairs identified with a higher threshold.

3.3. Launching CellphoneDB

3.3.1. Input data

CellphoneDB requires an input *metadata file* that must be generated by the user, associating individual cells with their respective clusters identified by scRNA-seq data (e.g., using packages such as Seurat [57] and SCANPY [59]). This file consists of two columns: 'Cell', indicating the name of the cell; and 'cell_type', indicating the name of the considered cluster. Additionally, another file called *Counts file* is required as input, which contains scRNA-seq count data containing gene expression values. In this file, rows represent genes presented with gene name identifiers (Ensembl IDs, gene names, or hgnc_symbol annotation), and columns represent cells.

These two files were created using the SCANPY Package installed on JupyterHub. *Metadata file* was produced in .txt format while *Counts file* called *local.h5ad* was produced as an anndata object from anndata Python package [60].

3.3.2. Creating SGE file

The method requires being run on Python 3.5 or higher. Due to its computational complexity and the large number of cells present in the count matrix, it has been decided to execute the method using the facility's job scheduling system named SGE, as previously described in paragraph 3.1.

Specifically, to launch the method, I first accessed the personal account @zeus.icbi.local within the cluster. Then, I installed and activated a Conda virtual environment in which I was able to install the CellphoneDB package. Subsequently, it was necessary to create an .sge file (reported in Figure 19), which is the document containing all the necessary information to properly execute the method using the SGE scheduler.

```
#! /bin/sh
#$ -S /bin/sh

#$ -pe smp 20
#$ -cwd
#$ -V

#### Jobdescription at qstat
#$ -N cpdb_stat

#### Error Outputfile
#$ -e /home/demarchi/myScratch/projects/Zlatko/cellphonedb-master/cellphonedb-Brian/LOGS/$JOB_NAME-$JOB_ID.err
#$ -o /home/demarchi/myScratch/projects/Zlatko/cellphonedb-master/cellphonedb-Brian/LOGS/$JOB_NAME-$JOB_ID.log

#### Resubmit
#$ -r y

hostname

conda activate cpdb
cellphonedb method statistical_analysis cellphonedb_meta.txt local.h5ad --threads 20
conda deactivate

# qsub /path/to/run_cpdb.sge
```

Figure 19. *.sge file for CellphoneDB.*

This .sge file, in its SGE-specific tag lines, specifies the UNIX shell for running locally and specifies the UNIX shell for running in the queuing system. Additionally, it reserves 20 CPU/cores in the SGE and executes the job from the current working directory. It also specifies the path for the output and finally activates the Conda virtual environment named *cpdb* previously created, runs the *statistical_analysis* method of CellphoneDB, and deactivates the virtual environment.

The job script is submitted to the queuing system using the *qsub* function from the command window.

3.3.3. Output

CellphoneDB outputs 4 files in .txt format: *significant_means.txt*, *pvalues.txt*, *means.txt*, and *deconvoluted.txt*. From the returned output file *pvalues.txt*, only the ligand-receptor pairs with a p-value equal to or less than 0.05 were retained, and only the pairs annotated in a priori database as true ligands were kept.

3.4. Launching NicheNet

3.4.1. Input data

NicheNet requires a pre-established model based on prior knowledge of ligand-to-target signaling pathways. This is why the first step in running NicheNet involves creating a 'ligand_target_matrix', which is specific to each individual database. This matrix is generated using a built-in function provided by the developers.

The other input required by NicheNet is the Seurat [57] object containing the gene expression data from the atlas, which needs to be combined with this prior model. Once the Seurat package is installed in R, it is sufficient to load the *local.rds* file containing the atlas, which will be read as a Seurat object.

3.4.2. Running analysis on Seurat object

To enable the NicheNet package's functions to identify the cellular clusters as mentioned in section 2.1, it was first necessary to modify the 'Idents' of the Seurat object, which refers to the cell identification information assigned to each cell, setting the ‘cell_type_major’ identity which was one of the identities already available in the Seurat object.
Next, the calculation of the most highly expressed genes in the sender cell clusters ('Tumor cells') and in the receiver cells (the same clusters of immune cells previously described in Chapther 2) was performed using NicheNet package functions that also remove genes not expressed in at least 10% of the cells in each specific cluster.
As a subsequent step, the method requires defining a gene set of interest: these are the genes in the receiver/target cell population that may be influenced by ligands expressed by interacting cells (e.g., genes differentially expressed upon cell-cell interaction). In order to accomplish this, genes that were differentially expressed in immune cells were analysed, taking into account the tissue sample, tumor vs healthy, as a condition.
Once the target genes are identified, it is necessary to identify a set of potential ligands: these are ligands expressed by the sender cell population and bind to receptors expressed by the receiver/target population. Subsequently, NicheNet ligand activity analysis is performed (assessing how well a ligand can predict the observed differentially expressed genes

compared to the background of expressed genes), ranking the potential ligands based on the presence of their target genes in the gene set of interest. Among different measures for ligand activity (AUROC, AUPR, Pearson correlation coefficient), developers indicate Pearson correlation coefficient between a ligand's target predictions and the observed transcriptional response as the most informative measure to define ligand activity. Then, ligand-receptor pairs are inferred using the prior model.

In NicheNet, it is also necessary, if one wants to maintain the clustering membership of LR pairs, to execute the code for each receiver cell cluster separately, and only afterwards combine the outputs, as otherwise NicheNet loses the clustering membership of LR pairs.

3.4.3. Output

Finally, it is possible to visualize signaling pathways between ligands and target genes of interest and analyze how the model deduces specific ligand-target connections. Two visualizations of NicheNet outputs are the heatmaps shown in the next two figures: the first (Figure 20) depicts the interaction between potential ligands and predicted target genes that is the main output from NicheNet analysis, while the second (Figure 21) illustrates the interaction between ligands and receptors. The strength of the interaction is determined by the previous model generated in the first phase, as demonstrated, for instance, in these two heatmaps, and it is computed for tumor-immune cells considered collectively without taking into account the various subpopulations of immune cells.

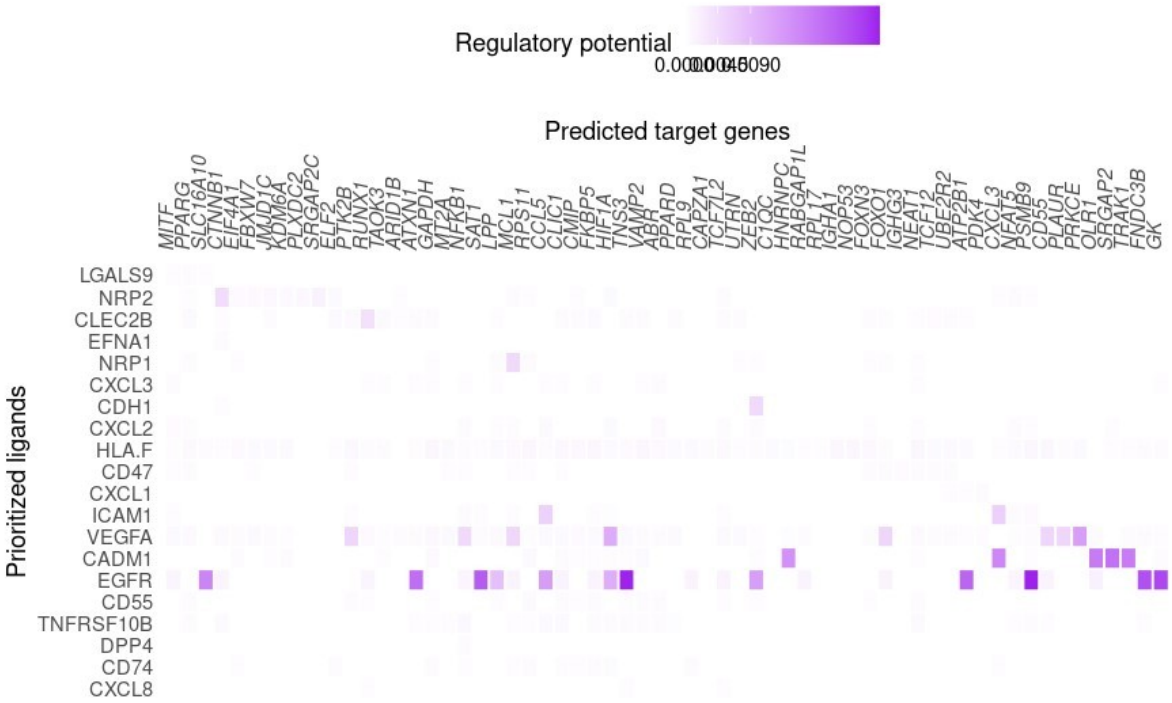

Figure 20. *NicheNet heatmap of prioritized ligands vs predicted target genes.*

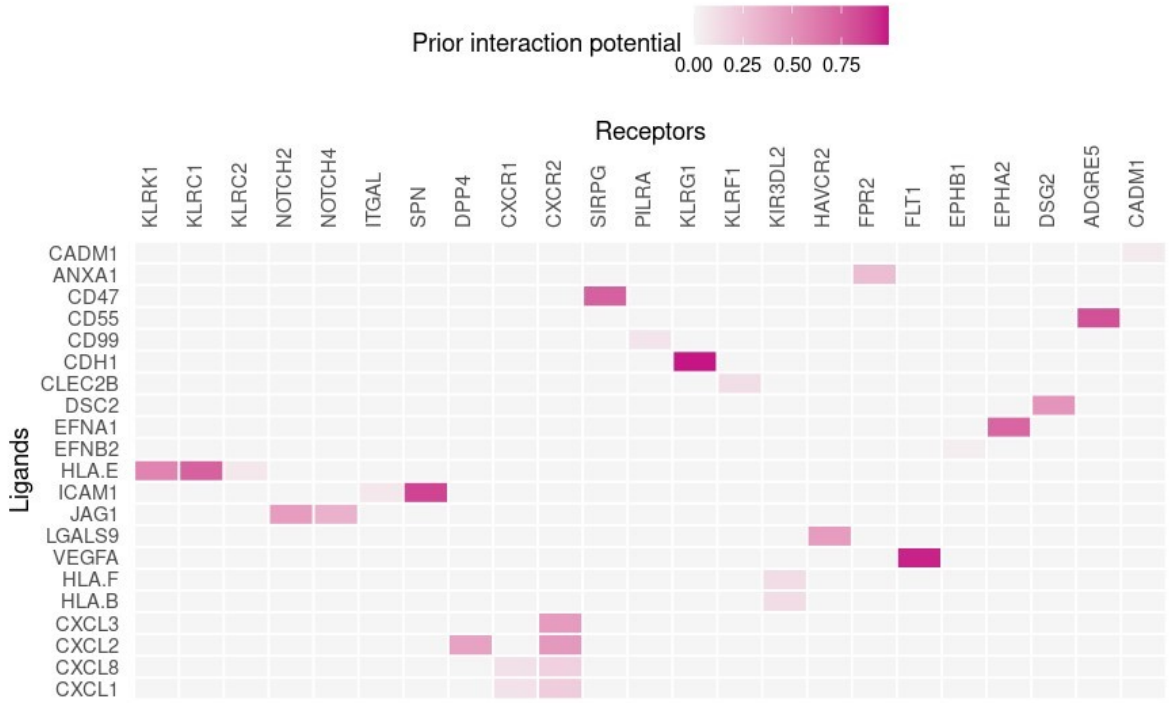

Figure 21. *NicheNet heatmap of ligands and receptors interaction potential.*

3.5. Comparative analysis: workflow and configuration

For a more personalized and tailored representation of the information being presented, I proceeded with the analysis by collecting the results from each individual method and organizing them in a way that allows for easier visualization and comparison of the information derived from the respective analyses.

All the tools require a curated LR interaction database in addition to gene expression data as input. scSeqComm and CellphoneDB also need the cell-type annotation as input. NicheNet performs cell type annotation by embedding certain cell-clustering procedures, such as Seurat, in its pipeline, and then assumes cluster-corresponding cell types. All of these tools output the predicted LR interaction pairs between cell types. Such LR pairs can then be used to construct CCI networks, suggesting the potential communication between cells. Additionally, all of them can provide visualization of CCIs.

The initial analysis was conducted using gene expression data related to the entire atlas and the LR database Efremova as inputs. Subsequently, two subgroups of the entire atlas were created by dividing it into a subgroup containing LUAD cells and another subgroup containing LUSC cells. These subgroups were created using the functions of the Seurat object in R for scSeqComm and NicheNet, while for CellphoneDB, the SCANPY package in Python was utilized, working on the anndata object. The same analyses described in the previous paragraphs were then executed first on the LUAD subgroup and subsequently on the LUSC subgroup.
Furthermore, these CCI analyses were repeated for the entire atlas, the LUAD subgroup, and the LUSC subgroup, also attempting to change the LR database given as input to the methods. Initially, the Browaeys database was used, and finally, the Jin database.

Table 2 summarizes the analysis scenarios just described, indicating in the first two columns the input data used by the methods, and in the last column the analysis scenario, which specifies the analyzed dataset and, in parentheses, the ligand-receptor database.

INPUT DATA		Analysis Scenario
Gene expression data	**LR database**	
Entire Atlas	Efremova et al. (2020)	Entire Atlas (Efremova DB)
LUAD	Efremova et al. (2020)	LUAD (Efremova DB)
LUSC	Efremova et al. (2020)	LUSC (Efremova DB)
Entire Atlas	Browaeys et al. (2019)	Entire Atlas (Browaeys DB)
LUAD	Browaeys et al. (2019)	LUAD (Browaeys DB)
LUSC	Browaeys et al. (2019)	LUSC (Browaeys DB)
Entire Atlas	Jin et al. (2020)	Entire Atlas (Jin DB)
LUAD	Jin et al. (2020)	LUAD (Jin DB)
LUSC	Jin et al. (2020)	LUSC (Jin DB)

Table 2. *List of analysis scenarios conducted in this thesis.*

To compare the methods, computational times of scSeqComm and NicheNet were tracked using the *system.time()* function from R to measure the execution time of an expression, while in CellphoneDB, the elapsed time was visible from the command window. In addition to computational times, RAM usage for various operations required by the methods was tracked looking at the memory usage reports.

However, the most important results concern the LR pairs obtained as output from the methods. To facilitate effective comparison, scripts were developed to collect the results and organize them into dataframes, which were then saved as .rds variables containing the necessary information. The dataframe consists of 9 columns:

- *int_pair* : ligand-receptor pair.

- *geneA*: first gene of the pair.
- *geneB*: second gene of the pair.
- *typeA*: the type of the first gene, whether ligand or receptor.
- *typeB*: the type of the second gene, whether ligand or receptor.
- *clustA*: the cellular cluster membership of the first gene.
- *clustB*: the cellular cluster membership of the second gene.
- *value*: the intensity of the interaction.
- *p_value*: the associated p-value for that LR pair.

Using these dataframes, Eulero-Venn diagrams were generated, plotted with specific scripts that take the dataframes as input. Since a ground truth was not available, Eulero-Venn diagrams are essential to observe the difference in the quantity of LR pairs obtained from each of the three methods, also noting that LR pairs found in the intersection of the Eulero-Venn diagrams are those LR pairs common to all three methods.
Instead, the difference in using different databases is observable in the quantity of ligand-receptor pairs obtained as output, and the next chapter shows this through a histogram.

At the end of the analyses, it seemed appropriate to examine in more detail the LR pairs found and analyze their significance. To do this, only the ligand-receptor pairs common to all three methods were visualized using a visualization method via circos links. Another dataframe was constructed containing only the LR pairs common to all three methods for the various analysis scenarios, and a package in R called *"intercellar"* [61] was used. This package, when given the properly constructed dataframe as input, can perform a Gene Ontology in the function-verse section and plot circos plots that connect the ligand-receptor pairs, highlighting their cluster membership.

In section 4.3, the biological significance of some of these LR pairs will also be shown.

Chapter 4: Interpretation of the results

4.1. Execution times and RAM monitoring

Starting with scSeqComm, for this method, the main function *'scSeqComm_analyze'* is the one that consumes the majority of the time the method takes to perform the analysis. From the image below (Figure 22), we can see that in this phase, the function first checks the input data, then conducts the analysis of intercellular communication, reporting the time it takes to calculate individual ligand and receptor scores as well as the S_inter scores. Subsequently, it also analyzes intracellular communication and reports the time taken for it.

```
Check input data...
**** scRNA-seq gene expression matrix having
 - 17811 genes
 - 892296 cells
 - 16 cell groups/clusters (  B cell  cDC1  cDC2  DC mature  Macrophage  Macrophage alve
atory  Tumor cells  )
...already big memory
Working with ... big.matrix

**** Ligand-receptor pair database having:
 - 688 known ligands
 - 857 known receptors
 - 12659 known ligand-receptor pairs
**** Considering only ligands and receptors present in the input scRNA-seq data, the lig
 - 576 known ligands
 - 733 known receptors
 - 8147 known ligand-receptor pairs

Analyzing intercellular communications...
**** Compute ligands and receptors scores
943.381 sec elapsed
**** Compute intercellular signaling evidence (i.e. ligand-receptor pair score S_inter)
769.642 sec elapsed

**** Trascriptional regulatory networks database having 942 transcription factors and 1
**** Considering only transcription factors regulating genes present in the input scRNA
s

Analyzing intracellular communications...
**** Compute TF activity
 - Identify realiable target genes
... using 20 cores with doParallelMC ...
817.66 sec elapsed
 - Identify differentially expressed target genes
... using 20 cores with doParallelMC ...
2216.421 sec elapsed
 - Compute TF activity score
17.466 sec elapsed
**** Load TF PPR scores
**** Compute intracellular signaling evidence (i.e. score S_intra)
... using 20 cores with doParallelMC ...
314.792 sec elapsed

Combine intercellular and intracellular evidence, and prepare output data...
```

Figure 22. *R console visualization during scSeqComm Entire Atlas (Efremova DB) analysis scenario.*

Time taken in analysis scenario Entire Atlas (Efremova DB) with the *'scSeqComm_analyze'* function being run with 20 cores on the head cluster Zeus, is 5528 seconds.

It has been observed that the time required for the analysis is consistently higher when analyzing the entire atlas because the larger the number of cells to consider, the longer the

analysis takes. Indeed, when the file of the entire atlas is loaded into R, it occupies 30.9 GB of RAM (892,296 cells and 17,811 genes), while the files of the LUAD subgroup (410,927 cells and 17,811 genes) and LUSC subgroup (92,430 cells and 17,811 genes) occupy 14 GB and 3.1 GB, respectively.

To give an example of the amount of RAM memory required to perform our scSeqComm analyses, let's consider the memory usage report example in the entire atlas scenario with the Efremova database:

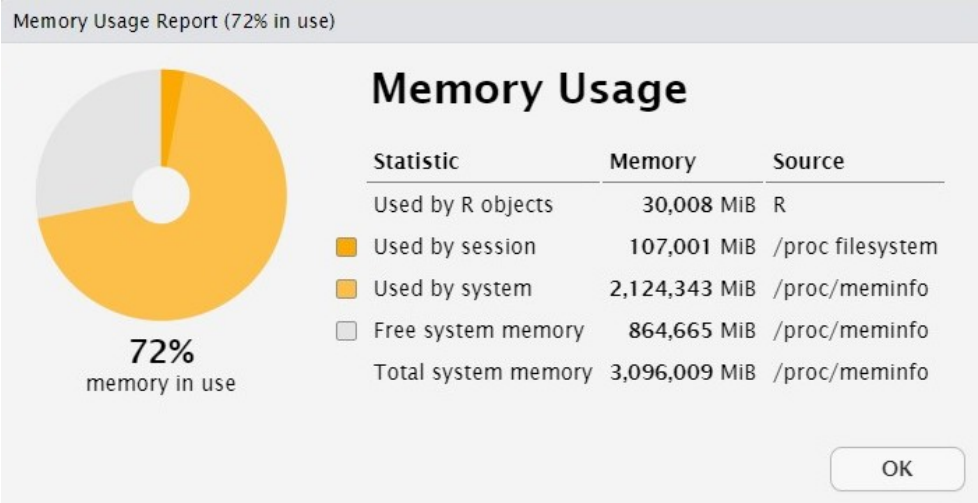

Figure 23. *Memory usage report for scSeqComm analysis of entire atlas (Efremova DB) scenario.*

From this report in Figure 23, we can observe that the analysis required an approximate usage of 112 GB in terms of memory used by the session and 2227 GB of memory used by the system.

Moving to CellphoneDB, this tool has proven to be the most computationally demanding. In fact, its statistical analysis method is expected to require approximately 1.5 hours for a dataset of around 10 GB, comprising 10,000 cells, according to the authors.
For this project, the tool was executed on the HPC cluster, as explained earlier, leveraging the SGE scheduling system and setting 20 CPUs/cores. In the entire atlas (Efremova DB) scenario, the size of the dataset is approximately 900,000 cells.

What the method returned in the command window is shown below:

```
[demarchi@zeus CellphoneDB]$ cat LOGS/cpdb_stat-3691086. log
apollo-05. local
[ ][CORE] [18/07/23-13:43:42] [INFO] Initializing SqlAlchemy CellPhoneDB Core
[ ][CORE] [18/07/23-13:43:42] [INFO] Using custom database at /home/demarchi/
.cpdb/releases/v2.0.0/cellphone.db
[ ] [APP] [18/07/23-13:43:42] [INFO] Launching Method cpdb_statistical_analysis_local_method_launcher
[ ] [APP] [18/07/23-13:43:42] [INFO] Launching Method_set_paths
[ ] [APP] [18/07/23-13:43:42] [INFO] Launching Method _load_meta_counts
[ ] [APP] [18/07/23-13:46:23] [INFO] Launching Method _check_counts_data
[ ][CORE] [18/07/23-13:46:23] [INFO] Launching Method cpdb_statistical_analysis_launcher
[ ][CORE] [18/07/23-13:46:23] [INFO] Launching Method _counts_validations
[ ][CORE] [18/07/23-13:47:01] [INFO] Launching Method get_interactions_genes_complex
[ ][CORE] [18/07/23-13:47:34] [INFO] [Cluster Statistical Analysis] Threshold:0.1 Iterations:1000 Debug-
seed:-1 Threads:20 Precision: 3
[ ][CORE] [18/07/23-13:49:41] [INFO] Running Real Analysis
[ ][CORE] [18/07/23-13:49:42] [INFO] Running Statistical Analysis
[ ][CORE] [19/07/23-00:12:44] [INFO] Building Pvalues result
[ ][CORE] [19/07/23-00:12:46] [INFO] Building results
[demarchi@zeus CellphoneDB]$
```

The output was obtained after around 10 hours and 30 minutes, as expected.
As mentioned, this execution time cannot be directly compared to that required by scSeqComm and NicheNet since the method was run on a different cluster of the HPCC. However, there still remains a considerable difference in the times required by the methods. In fact, CellphoneDB requires approximately five times longer than the other two tools to obtain results.

Lastly, regarding the NicheNet tool, since it requires the complete loading of the Seurat object contained in the .rds file to work on it, and given that the method is articulated in several sections, it has been decided to analyze the computational times for each section and then calculate the total time required by the method to obtain the outputs.

The various steps of the method in the analysis scenario of the entire atlas (Efremova DB) take the following computational times:

- Loading time for the local.rds file: 202 seconds
- Seurat object modification time: 0.8 seconds
- Ligand-target matrix creation time: 1610 seconds
- Calculation time for most expressed genes: 2362 seconds
- Calculation time for geneset of interest: 1528 seconds

- Calculation time for ligand activities: 4 seconds
- Calculation time for each cluster: 1036 seconds

By summing the times of the various implemented steps, it can be observed that the method requires approximately just under 2 hours to obtain the outputs.
Regarding the other analysis scenarios, typically the method requires less time when using subsets of the Seurat object for LUAD and LUSC. Even when changing the databases, similar times are observed for the entire atlas and its subsets.

Regarding the RAM memory requirement, let's examine the memory usage report of NicheNet in the analysis scenario of the entire atlas with the Efremova database (Figure 24). It can be observed that when running the method on the head cluster Zeus, it requires approximately 30 GB of memory to store the Seurat object on which to operate. Meanwhile, the memory usage by the system is similar to that required by scSeqComm.

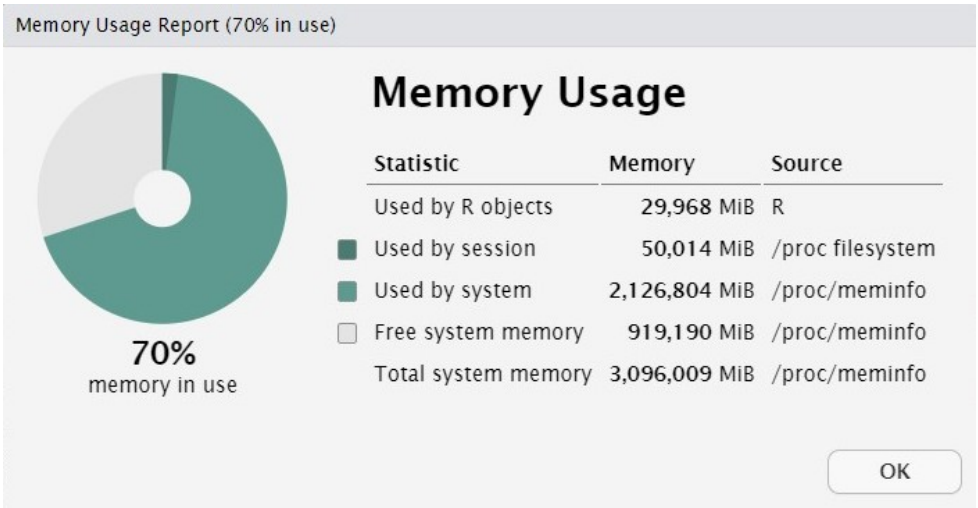

Figure 24. *Memory usage report for NicheNet analysis of entire atlas (Efremova DB) scenario.*

4.2. Results

4.2.1. Agreement across methods

It is now necessary to graphically represent, in a clear and effective way to understand format, the quantity of ligand-receptor pairs identified by each method and observe the overlap of the output results among the three different methods. To achieve this, Eulero-Venn diagrams have been constructed, which allow to easily visualize the number and percentage of common and non-common LR pairs in different scenarios (Figures 25-26-27). In these Eulero-Venn diagrams, the circles contain the LR pairs found by the methods, and each circle is represented with a different colour based on the method. The overlaps between the circles indicate the LR pairs shared by the methods, and the numbers observed in the figures indicate the number of LR pairs actually present in that area of the diagram; the percentage number indicates the percentage corresponding to that number of ligand-receptor pairs out of the total LR pairs found collectively by all three methods.

The first observation is that the overlap is relatively light. The limited overlap occurs because all three tools have different objectives and therefore produce different measures, leading to their agreement falling only on a small number of ligand-receptor pairs. In the analysis scenario of the complete atlas with the Efremova database (Figure 25), both scSeqComm and CellPhoneDB manage to discover the highest number of LR pairs. Specifically, scSeqComm identifies a total of 62 pairs: this number includes 44 LR pairs detected exclusively by scSeqComm (depicted in the purple circle), in addition to 4 + 10 + 4 pairs overlapping with those obtained by the other 3 methods. Meanwhile, CellphoneDB identifies 50 LR pairs, consisting of the 35 pairs detected solely by CellphoneDB, along with 4 + 10 + 1 pairs shared with the other methods.
Their overlap between scSeqComm and CellphoneDB is 14 pairs. This is where the biggest overlap occurs.
NicheNet returns fewer pairs compared to CellPhoneDB and scSeqComm because this method has the primary objective, unlike the other two, of finding ligand-target gene connections.

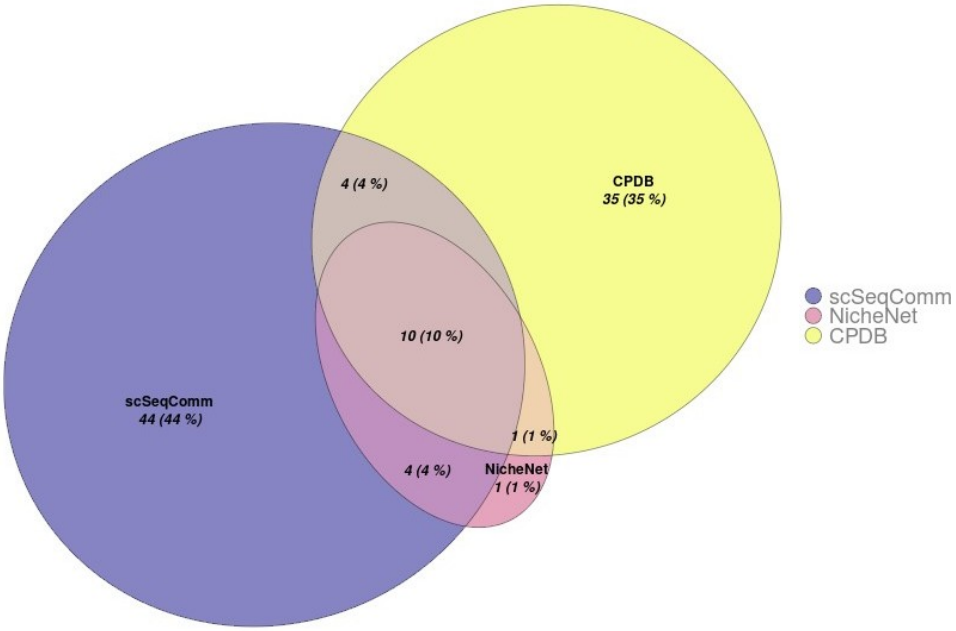

Figure 25. *Eulero-Venn diagram for the entire atlas (Efremova DB). The overlaps between the circles indicate the LR pairs shared by the methods, and the numbers observed in the figures indicate the number of LR pairs actually present in that area of the diagram; the percentage number indicates the percentage corresponding to that number of ligand-receptor pairs out of the total LR pairs found collectively by all three methods.*

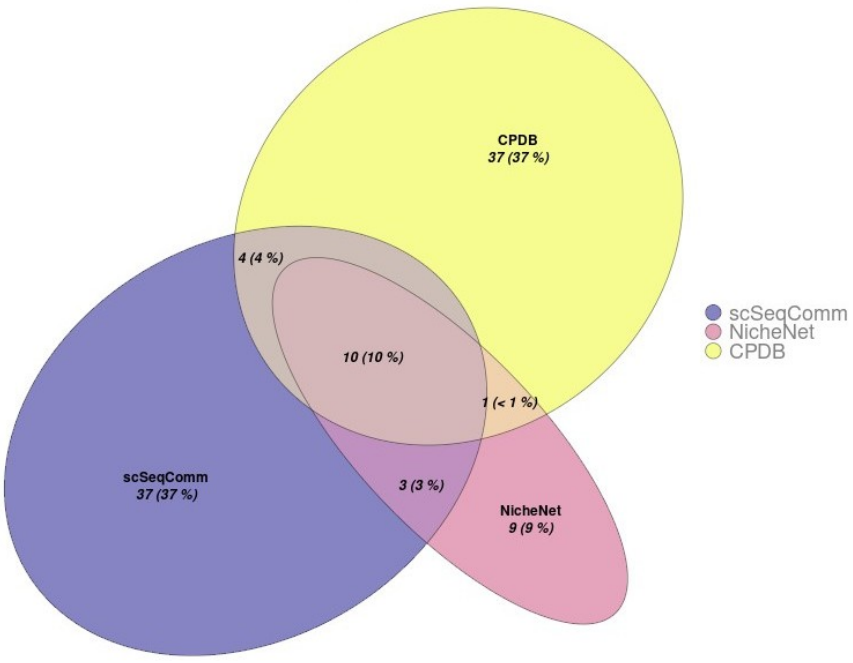

Figure 26. *Eulero-Venn diagram for the LUAD subset (Efremova DB). The overlaps between the circles indicate the LR pairs shared by the methods, and the numbers observed in the figures indicate the number of LR pairs actually present in that area of the diagram; the percentage number indicates the percentage corresponding to that number of ligand-receptor pairs out of the total LR pairs found collectively by all three methods.*

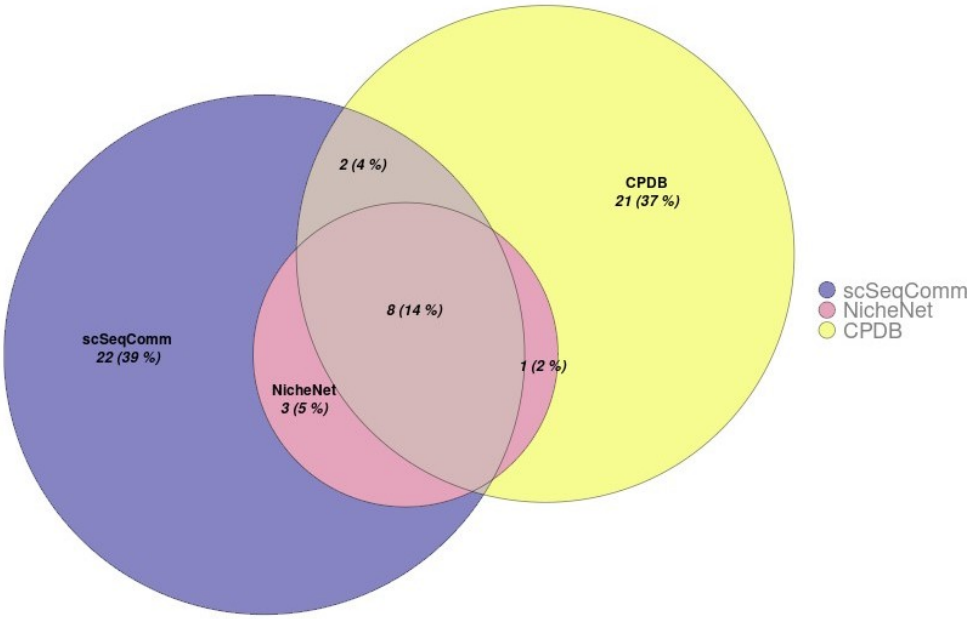

Figure 27. *Eulero-Venn diagram for the LUSC subset (Efremova DB). The overlaps between the circles indicate the LR pairs shared by the methods, and the numbers observed in the figures indicate the number of LR pairs actually present in that area of the diagram; the percentage number indicates the percentage corresponding to that number of ligand-receptor pairs out of the total LR pairs found collectively by all three methods.*

Whether considering the entire atlas or the LUAD and LUSC subsets, as observed in Figures 26 and 27, the overlap among the three methods remains quite similar in terms of the number of LR pairs found in common among the three methods showing an agreement of almost 10%, this means that the number of cells considered in the input does not influence the analyses.

What highlights the difference between the LUAD and LUSC subsets compared to the analysis performed for the entire atlas is seen in the number of pairs individually found by the methods. Specifically, scSeqComm and CellphoneDB find a lower number of LR pairs (54 and 52 for LUAD, and 35 and 32 for LUSC, respectively).

4.2.2. Role of LR databases

To appreciate the differences in using different databases, the following histograms (Figure 28) were constructed for the analysis conducted on the entire atlas. These histograms allow us to understand the actual number of interactions returned as output.

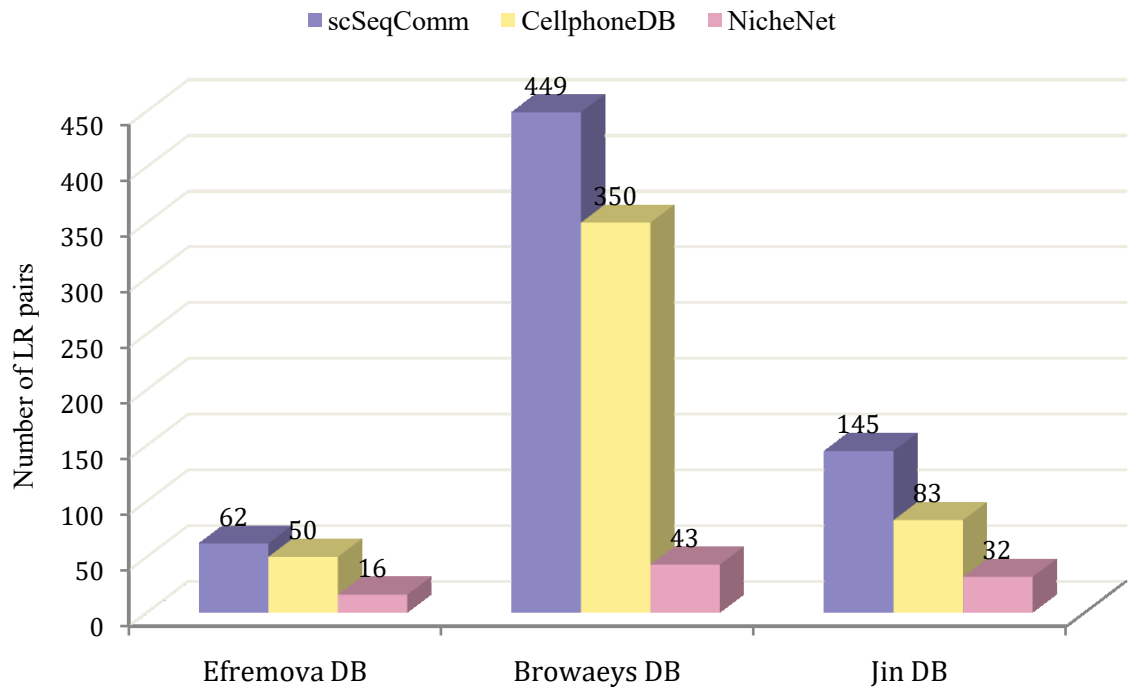

Figure 28. *Histograms representing the number of LR pairs output by each method with the different databases using the entire atlas.*

It is easily identifiable that CellPhoneDB and scSeqComm are consistently the methods capable of finding the highest number of pairs. As expected, the highest number of pairs returned as output is obtained for all three methods with the Browaeys database, which initially contains the largest number of LR pairs. This indicates that the number of interactions returned as output depends on the ligand-receptor database chosen as input and the number of LR pairs it comprises. The greater the number of LR pairs provided as input, the greater the number of pairs obtained as output. However, in all cases, NicheNet is the method that returns the fewest LR interactions.

4.2.3. Detected cell-cell communication

The quickest and most intuitive visualization of the ligand-receptor pairs obtained is provided by these circos plots, as previously mentioned, created using the intercellar package. In these plots, it can be observed how the ligand identified with HGNC nomenclature [62] is connected via an arrow to its corresponding receptor. Additionally, in a second outer circle, the ligand/receptor's cellular cluster membership is indicated with different colours. For this thesis purpose, the ligand belongs to the tumor cells, which represent sender cluster, while the receptor belongs to the immune cells, which represent receiver cell clusters.

These circos plots depicting the analysis scenarios utilizing the Efremova database are worth examining. To facilitate visualization and avoid creating an excessively large circos plot, only the LR pairs that are found in common among all three methods, as seen from the Eulero-Venn diagrams, are plotted. These pairs are also the ones with higher reliability and interesting biological significance. In Figure 29, we can observe the circos plot obtained for the analysis scenario of the entire atlas, while in Figures 30 and 31, respectively, for the LUAD and LUSC scenarios.

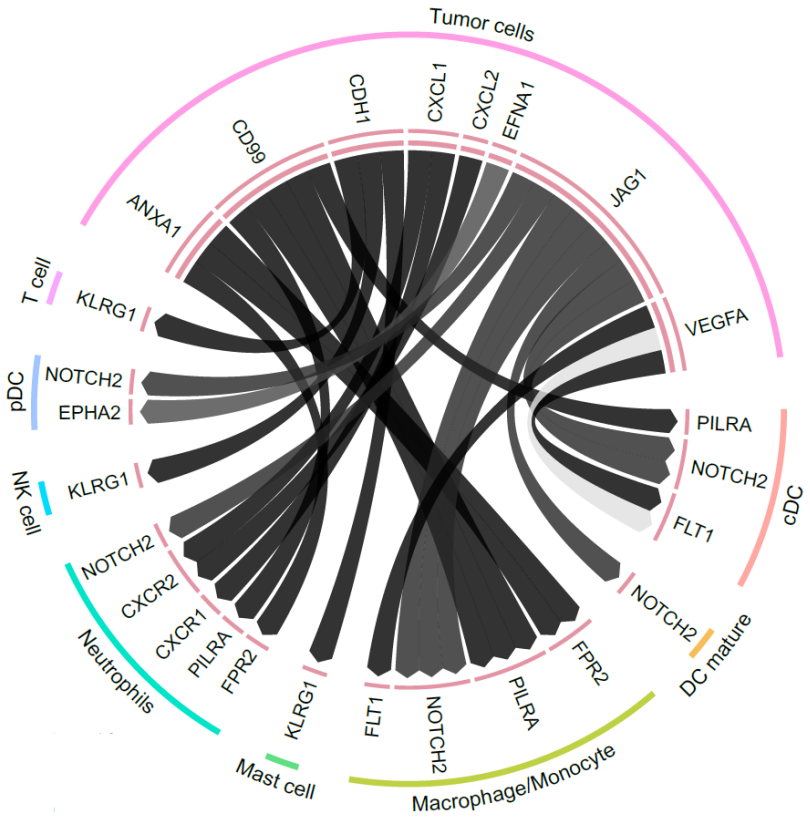

Figure 29. *Circos plot entire atlas analysis scenario (Efremova DB).*

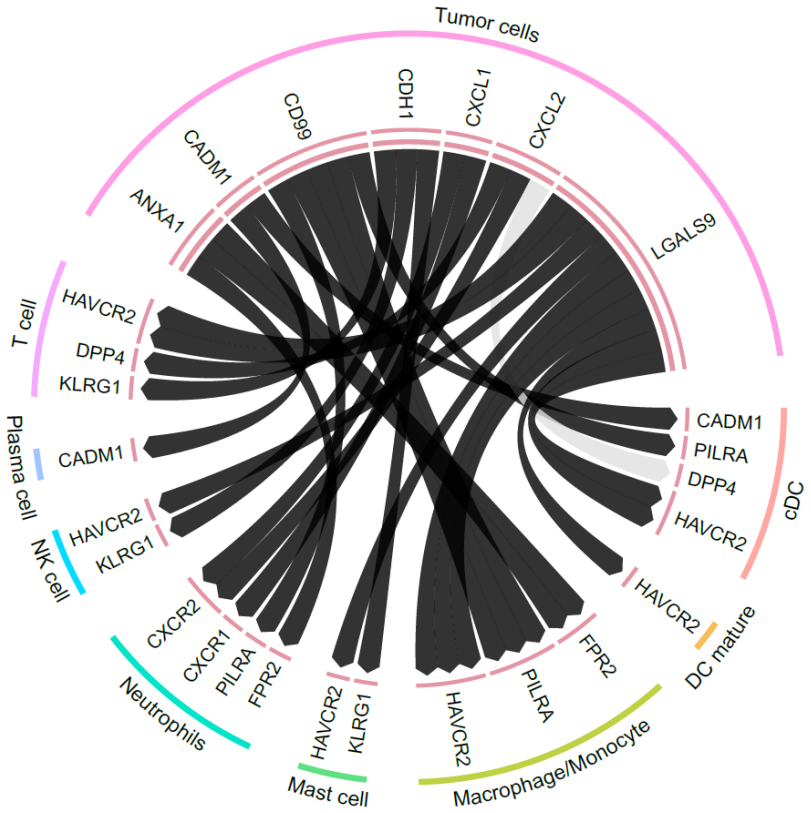

Figure 30. *Circos plot LUAD analysis scenario (Efremova DB).*

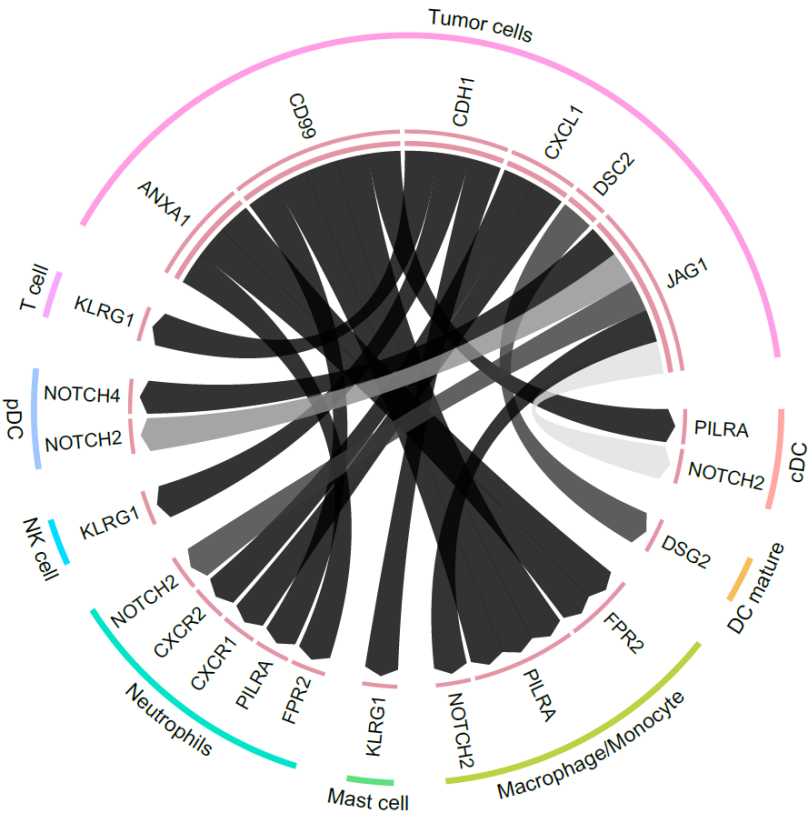

Figure 31. *Circos plot LUSC analysis scenario (Efremova DB).*

4.3. Biological meaning

These circos plots shown in the previous paragraph have enabled the identification of some ligand-receptor pairs of interest. In this paragraph, we will see how some of these pairs play a significant role in the organization of the tumor microenvironment.

By conducting a research in the scientific literature, it has been observed, for example, that in non-small cell lung cancer, the interaction we can see in figure 29, between VEGFA and FLT1 plays a significant role in tumor growth and progression. VEGFA promotes the formation of new blood vessels (angiogenesis) in tumors, supplying them with nutrients and oxygen, and facilitating their growth and metastasis. FLT1, as a receptor for VEGFA, mediates the signaling pathways that contribute to angiogenesis and tumor vascularization. In NSCLC, overexpression of VEGFA and increased activity of FLT1 are commonly observed [63]. This dysregulated VEGFA-FLT1 axis promotes angiogenesis within the tumor microenvironment, leading to enhanced tumor growth, invasion, and metastasis. Additionally, VEGFA-FLT1 signaling can contribute to treatment resistance by facilitating the development of blood vessels that provide a route for tumor cells to escape therapy and metastasize to other organs. Targeting the VEGFA-FLT1 axis has been a therapeutic strategy in NSCLC. Drugs that inhibit VEGFA or FLT1, such as bevacizumab (a VEGFA antibody) or tyrosine kinase inhibitors targeting FLT1, have been used in combination with standard chemotherapy or other targeted therapies to disrupt tumor angiogenesis and improve treatment outcomes in NSCLC patients.

The members of the VEGF family perform their functions by binding with their receptors. VEGF receptors are categorized into two types: tyrosine kinase receptors (VEGF receptors, VEGFR), which include VEGFR-1, VEGFR-2, and VEGFR-3, and neuropilin receptors (NRPs), which include NRP-1 and NRP-2. Evidence suggests that VEGF acts in tumors not only by promoting angiogenesis but also by directly working on cancer cells. VEGF can promote tumor development and progression by interacting with receptors expressed on tumor cells through autocrine and/or paracrine mechanisms. In addition to tyrosine kinases, NRPs can regulate the function and transportation of growth factor receptors and integrins, thus playing a crucial role in mediating VEGF action on tumor cells. However, studies showed that blocking endogenous VEGF with bevacizumab, the VEGF antibody, did not inhibit NSCLC cell line growth, suggesting that VEGF alone does not maintain lung cancer cell proliferation in vitro. The authors Zhao et al. [64] believe that the development of NSCLC is caused by a

combination of factors in the TME. Therefore, in a single cell line without tumor angiogenesis or TME, blocking VEGF alone does not effectively inhibit tumor cell growth. It is probable that therapy targeting VEGF-VEGFR also acts on the TME to counteract the immunosuppression present there, thereby preventing the growth of tumours. The authors conclude that immunotherapy and VEGF-VEGFR-targeted therapy together may have more effective therapeutic effects on NSCLC.

Now, considering another LR pair that has been found, CXCL1 and its receptor CXCR1, let's analyse the role of chemokines in NSCLC.

Chemokines are a family of soluble proteins that direct the migration of leukocytes under physiological conditions and during inflammation. They are important in embryonic development, activation of the immune response, and in driving both physiological and pathological angiogenesis.

For the past 20 years, there have been studies aimed at understanding the role of chemokines in the pathophysiology of cancer. It is currently accepted that the system of chemokines and their receptors has direct and indirect effects on the pathophysiology of cancer and that these molecules are important in the development and progression of the disease. Chemokines and their receptors are regulators of angiogenesis, which allows tumor growth and metastasis. Furthermore, chemokines and their receptors mediate the recruitment of cells of the immune system such as neutrophils to the tumor microenvironment. These cells actively modify the microenvironment; for example, macrophages are recruited by a pro-inflammatory environment and contribute to perpetuate inflammation.

Aberrant angiogenesis occurs in cancer as a result of alterations in the expression of molecules controlling the process, such as the chemokines.

Angiogenesis is important to support tumor growth, while infiltrating cells contribute to the tumor microenvironment through the secretion of growth factors, cytokines and chemokines, important molecules in the progression of the disease. Chemokines are important in development, activation of the immune response, and physiological angiogenesis. In addition, chemokines promote tumor cell survival, as well as the directing and establishment of tumor cells to metastasis sites. High levels of CXCL1 and CXCR1 expression correlate with advanced disease stage, increased tumor aggressiveness, and resistance to therapy [65].

Chapter 5: Conclusion

5.1. Discussion

It is evident from the second and third chapters that the three approaches have different analysis strategies, which in turn lead to different implementation procedures needed to run the methods.

For instance, scSeqComm performs an analysis of both intercellular and intracellular communication, employing a scheme focused on calculating scores for individual ligands and receptors, followed by computing an aggregate score to provide an indication of both intercellular and intracellular communication. On the other hand, CellphoneDB relies on a strategy that computes scores for statistically significant ligand-receptor specificity, utilizing a permutation test by randomly permuting cell labels, and it only provides an indication of potential ongoing intercellular communication. In contrast, NicheNet differs from the other two methods as its main goal is to search for links between ligands and target genes, with the analysis shifting to ligand-receptor pair identification only afterwards.

In terms of implementation, scSeqComm proves to be a relatively intuitive computational method, as the input data is straightforward to create, and the method relies on a main function that takes variables such as LR databases already loaded in the R package, allowing users to obtain results for both intercellular and intracellular communication with a single function. CellphoneDB, in turn, requires input data that is relatively simple to create but it demands the availability of powerful computing machines and the use of dedicated servers, which means learning how to use new tools and technologies like job scheduling systems and conda environments.
In contrast, NicheNet necessitates the creation of a prior model as input, complicating its implementation. Moreover, its analysis procedure, as seen, involves various steps and separate functions operating on a Seurat object, requiring proficiency in its manipulation.

From the previous chapter focused on interpreting the results, it is evident that in terms of execution times, the fastest methods are scSeqComm and NicheNet, taking approximately 2 hours to provide outputs when analyzing an atlas comprising around 900,000 cells with 20 CPU/cores on the head node. In terms of RAM usage, they are comparable. On the other hand, because CellphoneDB was started on a different HPCC cluster, its execution timings are not exactly comparable with those of the other two approaches, but they are significantly longer, taking about 10 hours and 30 minutes to analyse the atlas of about 900,000 cells.

Regarding the results in terms of the quantity of LR pairs obtained as output, as mentioned, scSeqComm and CellphoneDB have proven to be the methods that find the greatest number of ligand-receptor pairs, while NicheNet, having a different primary purpose, is limited in finding a restricted number of ligand-receptor pairs as output. The three methods also exhibit a 10% agreement, as shown by the Eulero-Venn diagrams, meaning that 10% of the total LR pairs found by the methods are common to all three. This is valid for both altering the input LR database and analysing NSCLC subgroups. Moreover, changing the input LR database allows for finding a greater number of LR pairs if the LR database input was constructed with a higher number of LR pairs.

The main challenges encountered during this thesis project pertain to adapting such a large-scale atlas to the methods because many operations performed on the atlas require time, and attention must be paid to every characteristic and piece of information present in the atlas. Waiting for the required computation periods to view the findings of the analysis proved to be difficult as well, particularly as it was challenging to simply restart an operation after making a mistake. It was also necessary to gain familiarity with the three tools, comprehending the techniques they employ for processing data, their analysis frameworks through an understanding of the statistics supporting them, their outputs, and the requirements for proper implementation. This involved learning two distinct programming languages, R and Python, and knowing how to use SGE as well as the properties of a Seurat object.

In accordance with the foregoing considerations, it might be safe to employ scSeqComm as a technique to carry out a preliminary analysis of cell-cell interaction, taking advantage of its simplicity of use and ability to produce results for both intracellular and intercellular communication analyses in a relatively brief period of time. With the 10% agreement, however, one possible recommendation would be to use all three tools and take into account

only the LR pairs that are found to be common to all three methods, as this thesis has shown and is supported by literature to be the most robust result. That is, if time permits for conducting the analyses.

5.2. Conclusion

This thesis work has highlighted the lack of running assessments to understand the performance and effectiveness of the most recent cell-cell interaction inference tools in real application scenarios, particularly in applying these tools to large scRNA-seq datasets. To address this gap, the objective was set to apply and compare three computational methods for cell-cell communication analysis on a large dataset originating from NSCLC.

It was pointed that there is no benchmark reference in the literature and it was highlighted that each tool operates, based on the interaction strengths of ligands and receptors, with different analysis structure and each has its advantages and weaknesses.
Results that allow for the comparison of the three methods in terms of execution times and RAM usage were presented. Then the focus was shifted on the most important findings, including the agreement across methods, the role of LR databases, and the detected cell-cell communication.

The discussion in Section 5.1 has illustrated how the aim of the project of effectively comparing the computational aspects and the results obtained from the application of these three methods for analyzing cell-cell communication on a large-scale atlas has been achieved. Furthermore, an agreement of 10% among the results obtained from the methods has been observed, and it has been suggested to prefer scSeqComm for conducting an initial simple analysis of cell-cell communication.

Moreover, it is worth noting that incredible advances are currently emerging in inferring cell-cell interactions and communication from gene expression data. There is significant potential for future applications, particularly in biomedicine and biotherapeutics. However, it's important to acknowledge that each approach for inferring CCIs and CCC has its own assumptions and limitations. Therefore, when utilizing such strategies, it is crucial to be aware of their strengths and weaknesses and to select the most appropriate one for analyses.

Methodological and technological challenges remain, but many opportunities exist to increase our understanding of intercellular interactions.

References

[1] Gruppo Nazionale di Bioingegneria. *Genomica e proteomica computazionale.* A cura di Riccardo Bellazzi, Silvio Bicciato, Silvio Cavalcanti, Claudio Cobelli, Gianna Maria Toffolo. Patron Editore.

[2] Lodish, H. F. (2000). *Molecular cell biology*. 4th ed. New York, W.H. Freeman.

[3] Strachan, T., & Read, A. (2018). *Human Molecular Genetics* (5th ed.).

[4] SERVICE R.F. (1999) *Exploring Systems of Life*. Science 284: 80-83 (the entire issue 5411 of the magazine is dedicated to complex systems).

[5] Chaurand P, Sanders ME, Jensen RA, Caprioli RM. *Proteomics in diagnostic pathology: profiling and imaging proteins directly in tissue sections.* Am J Pathol. 2004 Oct;165(4):1057-68. doi: 10.1016/S0002-9440(10)63367-6. PMID: 15466373; PMCID: PMC3118833.

[6] Dan E. Krane , Michael L. Raymer. *Fundamental Concepts of Bioinformatics.* Benjamin-Cummings Pub Co; 1st edition (January 1, 2002).

[7] Perdew, G. H., Vanden-Heuvel,J. P., & Peters, J. M. (2006). *Regulation of gene expression—Molecular mechanisms.* Totowa, NJ: Humana Press.

[8] Shabihkhani, M., Lucey, G. M.,Wei, B., Mareninov, S., Lou, J. J., Vinters, H. V., ... Yong,W. H. (2014). *The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings.* Clinical Biochemistry, 47, 258–266. doi:10.1016/j.clinbiochem.2014.01.002

[9] Hou, Z., Jiang, P., Swanson, S. A., Elwell, A. L., Nguyen, B. K., Bolin, J. M., ... Thomson, J. A. (2015). *A cost-effective RNA sequencing protocol for large-scale gene expression studies.* Scientific Reports, 5, 9570. doi:10.1038/srep09570

[10] Singh KP, Miaskowski C, Dhruva AA, Flowers E, Kober KM. *Mechanisms and Measurement of Changes in Gene Expression.* Biol Res Nurs. 2018 Jul;20(4):369-382. doi: 10.1177/1099800418772161. Epub 2018 Apr 29. PMID: 29706088; PMCID: PMC6346310.

[11] https://www.nature.com/scitable/topicpage/gene-expression-14121669/

[12] Holdt, L. M., Stahringer, A., Sass, K., Pichler, G., Kulak, N. A., Wilfert, W., ... Teupser, D. (2016). *Circular non-coding RNA ANRIL modulates ribosomal RNA maturation and atherosclerosis in humans.* Nature Communications, 7, 12429. doi:10.1038/ ncomms12429

[13] Gibson, G., & Muse, S. V. (2009). *A primer of genome science* (pp. 191–258). Sunderland, MA: Sinauer Associates.

[14] VanGuilder, H. D., Vrana, K. E., & Freeman, W. M. (2008). *Twentyfive years of quantitative PCR for gene expression analysis*. Biotechniques, 44, 619–626. doi:10.2144/000112776

[15] Schulze, A., & Downward, J. (2001). *Navigating gene expression using microarrays—A technology review.* Nature Cell Biology, 3, E190–E195. doi:10.1038/35087138

[16] Sinicropi, D., Cronin, M., & Liu, M. (2006). *Gene expression profiling utilizing microarray technology and RT-PCR.* In M. Ferrari, M. Ozkan, & M. J. Heller (Eds.), BioMEMS and biomedical nanotechnology (pp. 23–46). Boston, MA: Springer.

[17] Hrdlickova, R., Toloue, M., & Tian, B. (2017). *RNA-Seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews.* RNA, 8. doi:10.1002/wrna.1364

[18] Byron, S. A., Keuren-Jensen, K. R. V., Engelthaler, D. M., Carpten, J. D., & Craig, D. W. (2016). *Translating RNA sequencing into clinical diagnostics: Opportunities and challenges.* Nature Review Genetics, 17, 257–271.

[19] Ziegenhain C, Vieth B, Parekh S, et al. *Comparative analysis of single-cell RNA sequencing methods.* Mol Cell. 2017;65:631- 643.e4.

[20] Macaulay IC, Voet T. *Single cell genomics: advances and future perspectives*. PLoS Genet. 2014;10:e1004126.

[21] Sanz E, Yang L, Su T, et al. Cell-*type-specific isolation of ribosome-associated mRNA from complex tissues.* Proc Natl Acad Sci. 2009;106:13939-13944.

[22] Zeb Q, Wang C, Shafiq S, et al. *An Overview of Single-Cell Isolation Techniques.* Elsevier; 2019.

[23] Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-*cell RNA sequencing technologies and applications: A brief overview.* Clin Transl Med. 2022 Mar;12(3):e694. doi: 10.1002/ctm2.694. PMID: 35352511; PMCID: PMC8964935.

[24] Goh WWB, Wang W, Wong L. *Why batch effects matter in omics data, and how to avoid them.* Trends Biotechnol. 2017;35:498-507.

[25] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N; Human Cell Atlas Meeting Participants. *The Human Cell Atlas*. Elife. 2017 Dec 5;6:e27041. doi: 10.7554/eLife.27041. PMID: 29206104; PMCID: PMC5762154.

[26] Pisco AO, Tojo B, McGeever A. *Single-Cell Analysis for Whole-Organism Datasets.* Annu Rev Biomed Data Sci. 2021 Jul 20;4:207-226. doi: 10.1146/annurev-biodatasci-092820-031008. Epub 2021 May 11. PMID: 34465173.

[27] Konry T, Sarkar S, Sabhachandani P, Cohen N. *Innovative Tools and Technology for Analysis of Single Cells and Cell-Cell Interaction.* Annu Rev Biomed Eng. 2016 Jul

11;18:259-84. doi: 10.1146/annurev-bioeng-090215-112735. Epub 2016 Feb 24. PMID: 26928209.

[28] Rouault, H. & Hakim, V. *Different cell fates from cell-cell interactions: core architectures of two-cell bistable networks.* Biophys. J. 102, 417–426 (2012).

[29] https://www.khanacademy.org/science/ap-biology/cell-communication-and-cell-cycle/cell-communication/a/introduction-to-cell-signaling

[30] Rao, V. S., Srinivasa Rao, V., Srinivas, K., Sujini, G. N. & Sunand Kumar, G. N. *Protein-protein interaction detection: methods and analysis.* Int. J. Proteom. 2014, 1–12 (2014).

[31] Zhou, Y. et al. *Evaluation of single-cell cytokine secretion and cell-cell interactions with a hierarchical loading microwell chip.* Cell Rep. 31, 107574 (2020)

[32] Armingol, E., Officer, A., Harismendy, O. et al. *Deciphering cell–cell interactions and communication from gene expression.* Nat Rev Genet 22, 71–88 (2021). https://doi.org/10.1038/s41576-020-00292-x

[33] Oh, E.-Y. et al. *Extensive rewiring of epithelial-stromal co-expression networks in breast cancer.* Genome Biol. 16, 128 (2015).

[34] Han, X. et al. *Mapping the mouse cell atlas by microwell-seq.* Cell 172, 1091–1107.e17 (2018)

[35] Krämer, A., Green, J., Pollard, J. Jr & Tugendreich, S. *Causal analysis approaches in ingenuity pathway analysis.* Bioinformatics 30, 523–530 (2014).

[36] Baccin, C. et al. *Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization.* Nat. Cell Biol. 22, 38–48 (2020).

[37] Noël F, Massenet-Regad L, Carmi-Levy I, et al. *Dissection of intercellular communication using the transcriptome-based framework ICELLNET.* Nat Commun 2021;12:1089.

[38] Saidi Wang, Hansi Zheng, James S Choi, Jae K Lee, Xiaoman Li, Haiyan Hu, *A systematic evaluation of the computational tools for ligand-receptor-based cell–cell interaction inference*, Briefings in Functional Genomics, Volume 21, Issue 5, September 2022, Pages 339–356, https://doi.org/10.1093/bfgp/elac019

[39] Oh, E.-Y. et al. *Extensive rewiring of epithelial-stromal co-expression networks in breast cancer.* Genome Biol. 16, 128 (2015).

[40] Han, X. et al. *Mapping the mouse cell atlas by microwell-seq.* Cell 172, 1091–1107.e17 (2018).

[41] Krämer, A., Green, J., Pollard, J. Jr & Tugendreich, S. *Causal analysis approaches in ingenuity pathway analysis.* Bioinformatics 30, 523–530 (2014).

[42] Wang, S., Karikomi, M., MacLean, A. L. & Nie, Q. *Cell lineage and communication network inference via optimization for single-cell transcriptomics.* Nucleic Acids Res. 47, e66 (2019).

[43] Mangone L, Marinelli F, Bisceglia I, Zambelli A, Zanelli F, Pagano M, Alberti G, Morabito F, Pinto C. *Changes in the Histology of Lung Cancer in Northern Italy: Impact on Incidence and Mortality.* Cancers (Basel). 2023 Jun 14;15(12):3187. doi: 10.3390/cancers15123187. PMID: 37370797; PMCID: PMC10296491.

[44] https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/introduction

[45] Baruzzo G, Cesaro G, Di Camillo B. *Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm.* Bioinformatics. 2022 Mar 28;38(7):1920-1929. doi: 10.1093/bioinformatics/btac036. PMID: 35043939.

[46] Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. *CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.* Nat Protoc. 2020 Apr;15(4):1484-1506. doi: 10.1038/s41596-020-0292-x. Epub 2020 Feb 26. PMID: 32103204.

[47] Browaeys R, Saelens W, Saeys Y. *NicheNet: modeling intercellular communication by linking ligands to target genes.* Nat Methods. 2020 Feb;17(2):159-162. doi: 10.1038/s41592-019-0667-5. Epub 2019 Dec 9. PMID: 31819264.

[48] Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. *Inference and analysis of cell-cell communication using CellChat*. Nat Commun. 2021 Feb 17;12(1):1088. doi: 10.1038/s41467-021-21246-9. PMID: 33597522; PMCID: PMC7889871.

[49] Goveia, J., Rohlenova, K., Taverna, F., Treps, L., Conradi, L.C., Pircher, A., Geldhof, V., de Rooij, L.P.M.H., Kalucka, J., Sokol, L., et al. (2020). *An integrated gene expression landscape profiling approach to identify lung tumor endothelial cell heterogeneity and angiogenic candidates.* Cancer Cell 37, 21–36.e13. https://doi.org/10.1016/j.ccell.2019.12.001.

[50] Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., Kang, B., Liu, Z., Jin, L., Xing, R., et al. (2018). *Global characterization of T cells non-small-cell lung cancer by single-cell sequencing.* Nat. Med. 24, 978–985. https://doi.org/10.1038/s41591-018-0045-3.

[51] Salcher S, Sturm G, Horvath L, Untergasser G, Kuempers C, Fotakis G, Panizzolo E, Martowicz A, Trebo M, Pall G, Gamerith G, Sykora M, Augustin F, Schmitz K, Finotello F, Rieder D, Perner S, Sopper S, Wolf D, Pircher A, Trajanoski Z. *High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer.* Cancer Cell. 2022 Dec 12;40(12):1503-1520.e8. doi: 10.1016/j.ccell.2022.10.008. Epub 2022 Nov 10. PMID: 36368318; PMCID: PMC9767679.

[52] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, *K. KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res. 45, D353–D361 (2016).

[53] Ramilowski, J. A. et al. *A draft network of ligand–receptor-mediated multicellular signalling in human*. Nat. Commun. 6, 7866 (2015).

[54] Pawson, A. J. et al. *The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands*. Nucleic Acids Res. 42, D1098–D1106 (2014).

[55] Rouillard, A. D. et al. *The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins*. Database 2016, baw100 (2016).

[56] Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. *Geometric sketching compactly summarizes the singlecell transcriptomic landscape*. Cell Syst. 8, 483–493.e7 (2018).

[57] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol. 2015 May;33(5):495-502. doi: 10.1038/nbt.3192. Epub 2015 Apr 13. PMID: 25867923; PMCID: PMC4430369.

[58] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. *Comprehensive Integration of Single-Cell Data*. Cell. 2019 Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6. PMID: 31178118; PMCID: PMC6687398.

[59] Wolf FA, Angerer P, Theis FJ. *SCANPY: large-scale single-cell gene expression data analysis.* Genome Biol. 2018 Feb 6;19(1):15. doi: 10.1186/s13059-017-1382-0. PMID: 29409532; PMCID: PMC5802054.

[60] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf. *anndata: Annotated data*, bioRxiv 2021 Dec 19. doi: 10.1101/2021.12.16.473007.

[61] Interlandi, M., Kerl, K. & Dugas, M. *InterCellar enables interactive analysis and exploration of cell−cell communication in single-cell transcriptomic data*. Commun Biol 5, 21 (2022). https://doi.org/10.1038/s42003-021-02986-2

[62] Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. *Guidelines for human gene nomenclature.* Nat Genet. 2020 Aug;52(8):754-758. doi: 10.1038/s41588-020-0669-3. PMID: 32747822; PMCID: PMC7494048.

[63] Zhang SD, McCrudden CM, Kwok HF. *Prognostic significance of combining VEGFA, FLT1 and KDR mRNA expression in lung cancer.* Oncol Lett. 2015 Sep;10(3):1893-1901. doi: 10.3892/ol.2015.3415. Epub 2015 Jun 24. PMID: 26622771; PMCID: PMC4533253.

[64] Zhao Y, Guo S, Deng J, Shen J, Du F, Wu X, Chen Y, Li M, Chen M, Li X, Li W, Gu L, Sun Y, Wen Q, Li J, Xiao Z. *VEGF/VEGFR-Targeted Therapy and Immunotherapy in Non-small Cell Lung Cancer: Targeting the Tumor Microenvironment.* Int J Biol Sci. 2022 May 29;18(9):3845-3858. doi: 10.7150/ijbs.70958. PMID: 35813484; PMCID: PMC9254480.

[65] Rivas-Fuentes S, Salgado-Aguayo A, Pertuz Belloso S, Gorocica Rosete P, Alvarado-Vásquez N, Aquino-Jarquin G. *Role of Chemokines in Non-Small Cell Lung Cancer: Angiogenesis and Inflammation.* J Cancer. 2015 Aug 7;6(10):938-52. doi: 10.7150/jca.12286. PMID: 26316890; PMCID: PMC4543754.