UNIVERSITÀ DEGLI STUDI DI CATANIA

DOTTORATO DI RICERCA IN BASIC AND APPLIED BIOMEDICAL SCIENCES

XXXIII CICLO

Dipartimento di Scienze Biomediche e Biotecnologiche (BIOMETEC)

Dott.ssa Valentina Di Salvatore

Gene expression, data analysis and computational modeling methodologies in cancer and neurodegenerative disorders

Tesi di dottorato

Coordinatore:	Relatore:
Ch.ma Prof.ssa Stefania Stefani	Ch.mo Prof. Francesco Pappalardo

---

ANNO ACCADEMICO 2019/2020

Gene expression, data analysis and computational modeling methodologies in cancer and neurodegenerative disorders

Valentina Di Salvatore

SOTTOMESSA IN PARZIALE ADEMPIMENTO AI REQUISITI PER
IL CONSEGUIMENTO DEL TITOLO DI
"DOTTORE DI RICERCA"
ALL' UNIVERSITÀ DEGLI STUDI DI CATANIA
DATA, ITALIA, CATANIA

UNIVERSITÀ DEGLI STUDI DI CATANIA

DOTTORATO DI RICERCA IN BASIC AND APPLIED BIOMEDICAL SCIENCES

Dipartimento di Scienze Biomediche e Biotecnologiche (BIOMETEC)

I sottoscritti certificano che hanno letto e raccomandato alla Facoltà per gli Studi di Dottorato l'accettazione della tesi intitolata **"Gene expression, data analysis and computational modeling methodologies in cancer and neurodegenerative disorders"** di Valentina Di Salvatore, a parziale adempimento ai requisiti per il conseguimento del titolo di **"Dottore di Ricerca"**.

Tutor (Prof. Francesco Pappalardo): _____

Coordinatore (Prof.ssa Stefania Stefani): _____

UNIVERSITÀ DEGLI STUDI DI CATANIA

Autore:	**Valentina Di Salvatore**
Titolo:	**Gene expression, data analysis and computational modeling methodologies in cancer and neurodegenerative disorders**
Dipartimento:	**Scienze Biomediche e Biotecnologiche (BIOMETEC)**
Grado:	**Dottore di Ricerca**

Firma dell'autore: _____

INDEX

SUMMARY

With the advent of new technologies in sequencing and the resulting production of huge amounts of data, the need for a new way to analyze such a vast amount of data arises: bioinformatics is tasked with taking on this burden, mining data, storing data, and ensuring valid biological conclusions can be drawn from data. Computer knowledge is not the only requirement a bioinformatic researcher should have: being able to give a correct interpretation of analyzed data is fundamental in order to avoid the danger of overinterpreting data, so a good biological understanding of the system under study is still essential. Bioinformatics, indeed, represents a multidisciplinary field of increasing interest in medicine, biology, and genetics, and finds its application in fields such as proteomics, transcriptomics, genomics, systems biology, structural bioinformatics, evolutionary bioinformatics, modeling, imaging, biophysics, population genetics, and clinical bioinformatics.

In this thesis, three different research topics have been presented, but in all three bioinformatics played an essential role both in the data analysis and interpretation of the results obtained, proving once again to be such a powerful and versatile tool as to facilitate and accelerate the various and complex phases of a traditional analysis.

In the first topic, bioinformatic tools have been used in order to investigate the role of MAP3K8, a serine/threonine kinase expressed in thyroid cancer stem cells (CSCs), in mediating drug resistance in human thyroid cancer and its relationship to tumor behavior. Through the use of specific informatic tools, we performed a complete analysis of data downloaded from NCBI Gene Expression Omnibus data repository, which comprises survival, gene expression, pathway and stemness index evaluation analysis. As a result of this analysis, we demonstrated that high values of MAP3K8 expression are related to a particularly aggressive and lethal tumor type, Anaplastic Thyroid Cancer (ATC), thus highlighting the role of MAP3K8 as potential prognostic biomarker.

In the second topic, we dealt with the design of oncolytic viruses. Oncolytic viruses are a form of immunotherapy that employs attenuated viruses to restore the immune system response in order to infect and destroy cancer cells. In collaboration with Etna Biotech company of Catania, we investigated the antitumor efficacy of *cetuximab*, a widely used anti-epidermal growth factor receptor (EGFR) monoclonal antibody, combined with measles virus (MV). To evaluate the functionality of the four protein models representing H protein-cetuximab fusion structure provided

by the company, we used tools as SWISSMODEL and ProQ thus obtaining a complete score-based assessment for each model which has been subsequently confirmed by laboratory tests.

The third and last topic of this thesis aims to provide a complete workflow for RNA-seq data analysis. Starting from the sequencing data of four *Mus musculus* samples, we build a framework upon which any future RNA-seq data analysis could be structured, combining both informatic and statistical tools. Even if this study is still in progress as new data with more biological replicates are needed, yet it represents an excellent starting point for future and more in depth analysis.

SOMMARIO

Con l'avvento di nuove tecnologie nel sequenziamento e la conseguente produzione di enormi quantità di dati, sorge la necessità di un nuovo modo di analizzare una quantità così vasta di dati: la bioinformatica ha il compito di assumersi questo onere, estraendo dati, archiviando dati e garantendo che si possano trarre conclusioni biologiche valide dai dati. La conoscenza dell'informatica non è l'unico requisito che un ricercatore bioinformatico dovrebbe avere: saper dare una corretta interpretazione dei dati analizzati è fondamentale per evitare il pericolo di una errata interpretazione dei dati, quindi una buona comprensione biologica del sistema in esame è comunque essenziale. La bioinformatica, infatti, rappresenta un campo multidisciplinare di crescente interesse per la medicina, la biologia e la genetica, e trova la sua applicazione in campi come proteomica, trascrittomica, genomica, biologia dei sistemi, bioinformatica strutturale, bioinformatica evolutiva, modellistica, imaging, biofisica, genetica di popolazione e bioinformatica clinica.

In questo lavoro di tesi sono stati presentati tre diversi temi di ricerca, ma in tutti e tre la bioinformatica ha giocato un ruolo essenziale sia nell'analisi dei dati che nell'interpretazione dei risultati ottenuti, dimostrandosi ancora una volta uno strumento così potente e versatile da facilitare e accelerare le varie e complesse fasi di un'analisi tradizionale.

Nel primo argomento, sono stati utilizzati strumenti bioinformatici per indagare il ruolo di MAP3K8, una serina/treonina protein-chinasi espressa nelle cellule staminali del cancro della tiroide (CSC), nel mediare la resistenza ai farmaci nel cancro della tiroide umano e la sua relazione con il comportamento del tumore. Attraverso l'uso di strumenti informatici specifici abbiamo eseguito un'analisi completa dei dati scaricati dal repository di dati NCBI Gene Expression Omnibus, che comprende analisi di valutazione di sopravvivenza, espressione genica, pathway e indice di staminalità. Come risultato di questa analisi, abbiamo evidenziato che valori elevati di espressione di MAP3K8 sono correlati a un tipo di tumore particolarmente aggressivo e letale, il cancro alla tiroide anaplastico (ATC), evidenziando così il ruolo di MAP3K8 come potenziale biomarcatore prognostico.

Nel secondo argomento, ci siamo occupati della progettazione dei virus oncolitici. I virus oncolitici sono una forma di immunoterapia che impiega virus attenuati per ripristinare la risposta del sistema immunitario al fine di infettare e distruggere le cellule tumorali. In collaborazione con la società Etna Biotech di Catania, abbiamo studiato l'efficacia antitumorale del *cetuximab*, un anticorpo monoclonale anti-epidermico recettore del fattore di crescita (EGFR), combinato con il virus del

morbillo (MV). Per valutare la funzionalità dei quattro modelli proteici rappresentativi della struttura di fusione proteina H-cetuximab forniti dall'azienda, abbiamo utilizzato strumenti come SWISSMODEL e ProQ ottenendo così una valutazione completa basata su punteggi per ogni modello che è stata successivamente confermata da test di laboratorio.

Il terzo e ultimo argomento di questa tesi mira a fornire un workflow completo per l'analisi dei dati RNA-seq. A partire dai dati di sequencing di quattro campioni di *Mus musculus*, abbiamo costruito un framework su cui strutturare qualsiasi futura analisi dei dati RNA-seq, combinando strumenti informatici e statistici. Anche se questo studio è ancora in corso, in quanto sono necessari nuovi dati con più replicati biologici, rappresenta tuttavia un ottimo punto di partenza per analisi future e più approfondite.

AKNOWLEDGEMENTS

From the day I chose my thesis in computer engineering, which concerned a dielectrophoretic system for the separation of microparticles in plasma, I have always known that bioinformatics, which in those days did not even exist as a real discipline, would be my way. I have always thought that the perfect blend of computer science and medicine would revolutionize the world of scientific research, and even if immediately after graduation I was unable to carry on this goal of mine, I never stopped dreaming of pursuing a career in this sector.

My first approach to the world of scientific research was to enter the CNR Institute of Neurological Sciences, thanks to a call which was found almost by chance. Since then I have never stopped studying and learning as much as possible about the tools and techniques that form the heart of Bioinformatics.

But it was only thanks to this PhD that I can say that I really started my journey in bioinformatics. In these three years of PhD I have had the opportunity to approach wonderful people both from a human and professional point of view.

The first person I particularly want to thank is my tutor and mentor, Prof. Francesco Pappalardo, who welcomed me into his team with enthusiasm and kindness. I am flattered for the trust he has always shown in me, and I am grateful to him for all the opportunities he has given to me, allowing me to gain ever greater security and confidence in myself and in my abilities. My professor never treated me like a student, but immediately made me feel like a real researcher by assigning me tasks that I would not even have imagined being able to complete, showing that he had more faith in me than I did me in myself.

And it is precisely this sense of challenge in doing things that many times I do not feel up to that makes this job different and stimulating every day.

In the COMBINE group of which I can now proudly say that I am also a part, there is the second person to whom I would like to say thank you with all my heart: Dr. Giulia Russo. Her constant support, her kindness, her patience even in my moments of panic, have been of enormous help to me in these three years. Her complete and profound preparation from a professional point of view is comparable only to the kindness which she always tries to help me with.

A heartfelt thanks also goes to all the professionals whom I have been lucky enough to collaborate with, who have allowed me to expand my knowledge and my skills to work in a team.

Last but not least, I sincerely thank my wonderful family for the constant support and encouragement they give me every day. My parents, my sisters, my parents in-laws, my uncles and cousins, are all fundamental elements of my life, they are my fixed points and my most precious allies.

And finally, a special thanks goes to my husband Dario, who has always been a huge support for me also from a technical-practical point of view (and I hope he won't give up now!) and my son Francesco: both of them fill my life with love and games and laughter and without them today I would have nothing.

1 INTRODUCTION

1.1 Bioinformatics background

Bioinformatics, along with genomics and proteomics, represents one of the three new disciplines that have emerged and rapidly developed over the past decade. While genomics and proteomics respectively generate information on genomic sequences and protein properties, the aim of bioinformatics is to provide methods and tools for genomic and proteomic data analysis. Essentially, bioinformatics is the application of computer science and information technology to the fields of biology [1]. The word "bioinformatics" was first used in 1968 and its definition was first given in 1978. Bioinformatics has also been referred to as "computational biology", even if, more exactly, computational biology deals mainly with modeling of biological systems. Bioinformatics made its first appearance in Switzerland in the early 1980s, and then, no one could have imagined the importance that this discipline would assume in the following years. Swiss bioinformaticians created software for the comparison of DNA sequences, developed programs for the analysis of experimental protein data, invented tools for modeling the three-dimensional structures of proteins, and built databases containing key information about protein sequences. Some of these resources became world leaders in their respective fields and played a central role, not only in bioinformatics but also in life science research in general. In 1998, the five existing bioinformatics groups, representing 30 scientists from five academic institutions, organized themselves in a federation of academic researchers to create the Swiss Institute of Bioinformatics (SIB), a nonprofit foundation (www.isb-sib.ch) [2]. With the advent of new technologies in sequencing and the resulting production of huge amounts of data, it was immediately clear that it was not possible to continue to analyze such a vast amount of "big data" by hand: bioinformatics is tasked with taking on this burden, mining data, storing data, and ensuring valid biological conclusions can be drawn from data. Obviously, computer knowledge alone is not sufficient to be able to give a correct interpretation of analyzed data: there is a danger of overinterpreting data, so a good biological understanding of the system under study is still essential. Today, bioinformatics represents a multidisciplinary field of increasing interest in medicine, biology, and genetics, and finds its application in fields such as proteomics, transcriptomics, genomics, systems biology, structural bioinformatics, evolutionary bioinformatics, modeling, imaging, biophysics, population genetics, and clinical bioinformatics.

1.2 Bioinformatics Tools

Bioinformatics can be seen as a combination of various other disciplines like biology, mathematics, computer science, and statistics, to develop methods for storage, retrieval and analyses of biological data. There is a large variety of bioinformatic tools, depending on the specific requirements of each particular project. Basically, most of bioinformatics applications are shown in fig. 1 below:

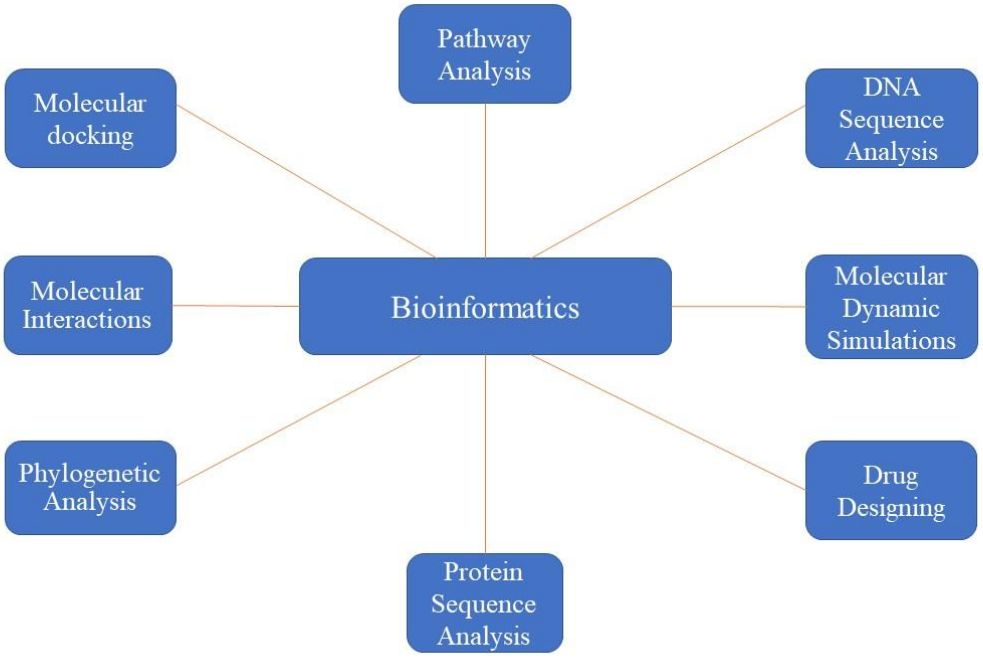

Figure 1 - Schematic representation of bionformatics fields of application

- Molecular docking: is a computational method that aims to predict the affinity in a ligand – protein bond when they form a stable molecular complex [3].
- Molecular Interactions: are attractive or repulsive forces between molecules and between non-bonded atoms. They are also known as noncovalent interactions or intermolecular interactions [4].
- Phylogenetic Analysis: is the study of evolutionary development of a species or a group of organisms or a particular characteristic of an organism.
- Protein Sequence Analysis: consists of several experimental methods aimed to determine the amino acid sequence of a protein or of parts of a protein, that can be used in bioinformatic analyses to predict protein structure and possible functions.

- DNA Sequence Analysis: is the process of determining the exact order of the four bases (adenine, A; thymine, T; cytosine, C; guanine, G) in a given DNA template.
- Molecular Dynamic Simulations: is a set of computational simulation techniques which, through the integration of the equations of motion, allows to study the evolution dynamics of a physical and chemical system at the atomic and molecular level.
- Drug Designing: is the inventive process of finding new medications based on the knowledge of a biological target. It involves the design of molecules that are complementary in shape and charge to the molecular target with which they interact and bind [5].
- Pathway Analysis: is the study of all the interactions between biochemical compounds in a particular biological process.

Among all the useful available tools, special attention should be given to:

- Sequence Databases

Biological databases basically aim to make updated high-quality data collection available and easily searchable for the entire scientific community. Virtual computerized archives are used to store and prepare records in such a way that the data can be retrieved effortlessly through a system of ramified searching criteria. Basically, sequence analyss is about the understanding of all the different features of a biomolecule like nucleic acid or protein, and of their functions. The first step in this type of analyses, is the search of the sequences of corresponding molecules from public databases. Databases contain a variety of information: for this reason, they can be classified into Primary, Secondary, or Composite databases, depending upon the information stored in them. The data in a primary database is obtained through experimentation. Some examples of primary databases are:

- UniProt (https://www.uniprot.org/) - is a collection of sequences derived from various other databases PIRPSD, Swiss-Prot, and TrEMBL.
- PIR (https://proteininformationresource.org/) - maintains the Protein Sequence Database (PSD), an annotated protein database containing over 283000 sequences covering the entire taxonomic range.
- GenBank (https://www.ncbi.nlm.nih.gov/genbank/) - is a comprehensive database that contains publicly available nucleotide sequences for more than 300 000 organisms named at the genus level or lower.
- EMBL (https://www.embl.de/) - is a molecular biology research institution supported by 27 member states.

- DDBJ (https://www.ddbj.nig.ac.jp/) - is a biological database that collects DNA sequences. It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC.
- Protein Databank PDB (https://www.rcsb.org/pdb/) - is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids.

A secondary database contains informations derived from the analysis of primary databases data. Some examples of secondary databases include:

- SCOP (http://scop.mrc-lmb.cam.ac.uk/) - aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- CATH (http://www.cathdb.info/) - is a classification of protein structures downloaded from the Protein Data Bank.
- PROSITE (https://prosite.expasy.org/) - consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.
- MOTIF (https://www.genome.jp/tools/motif/) – is a database containing protein sequence motifs.

A composite database contains information derived from different primary sources. Some examples of composite databases include:

- NRDB (nonredundant database, https://pubmlst.org/analysis/nrdb.shtml) contains data obtained from GenBank (CDS (CoDing Sequence) translations), PDB, SWISS-PROT, PIR (Protein Information Resource), and PRF (Protein Research Foundation).
- INSD (International Nucleotide Sequence Database, http://www.insdc.org/) is a collection of nucleic acid sequences from EMBL, GenBank, and DDBJ.
- The UniProt (universal protein sequence database, https://www.uniprot.org/) is a collection of sequences derived from various other databases PIRPSD, Swiss-Prot, and TrEMBL.
- wwPDB (worldwide PDB, http://www.wwpdb.org/) is a composite of 3D structures in the RCSB (Research Collaboratory for Structural Bioinformatics), PDB, MSD (Macromolecular Structure *Database*), and PDBj.

- Pathway Databases

A biological pathway is a sequence of interactions between biochemical compounds that leads to a certain product or a change in a cell in response to specific stimuli. The main purpose of pathway analysis is to analyze data obtained from high-throughput technologies, detecting relevant groups of related genes that are altered in case samples in comparison to a control [6]. The most relevant available pathway databases are the following:

- Reactome (https://reactome.org/) is a free, open-source database of biological pathways. Data stored in Reactome has been extracted from the experimental literature and maintained by researchers, curators, editors and reviewers. In Reactome references to other public databases such as UniProt, Ensembl, KEGG and many others can be found.
- Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/) is a database connecting genomic, biochemical and phenotypical information from multiple individual databases. It contains information about metabolic pathways and the genomes, genes, proteins and enzymes that contribute to those pathways, as well as details about genetic and environmental processes, diseases and drugs pathways. There are many links to external databases such as NCBI Entrez Gene, OMIM and UniProt.
- WikiPathways (https://www.wikipathways.org/index.php/WikiPathways) is an open-source project different to the other pathway databases. It is part of the MediaWiki software and relies on creation, curation and editing of various biochemical pathways by any user with a WikiPathways account. WikiPathways contains many different signaling pathways involved in different biological processes across many species. WikiPathways is a new paradigm for storing and organizing large amounts of biological data, relying on community commitment to maintain and curate the data, contributing to the overall success [7].

- Programming language

R (https://www.r-project.org/) is an open-source language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, it is highly extensible and it is free. Due to the huge variety of highly specific tasks involved in bioinformatic analyses, R surely represents one of the most useful and powerful resources for a researcher to know. R runs on all popular platform: its cross-platform interoperability represents one of the main features of this language.

R is not the only available language for statistical computing and graphics: Python (https://www.python.org/) is a powerful object-oriented programming language with an easy-to-use and simple syntax. Python is extremely popular among data scientists and researchers. Most of the packages in R have equivalent libraries in Python as well. While R is the first choice of statisticians and mathematicians, professional programmers prefer implementing new algorithms in Python. The choice between R and Python also depends on what the requirements of a certain project are. If it aims to analyze a dataset and present the findings in a research paper, then R is probably the best choice.

1.3 RNA-sequencing techniques

RNA sequencing (RNA-seq) aims to identify most RNA species inside a cell, providing huge amounts of sequence "reads" and information on a large number of individual bases. To extract valid insights from such a massive quantity of data, a high understanding of bioinformatics and statistics is required. RNA-seq consists in the application of a variety of next-generation sequencing techniques to study RNA. The first step in RNA-seq experiments, is the choice of the sequencing platform: data obtained from the different RNA sequencing platforms may vary, and this variation can affect the interpretation of an experiment. Protocols for sample preparation depend on the sequencing platform, hence the choice of the right platform is a key requirement for the experiment to be successfully completed. Several Next Generation Sequencing (NGS) platforms are commercially available: most are based on sequencing-by-synthesis technology, with a DNA polymerase or ligase as key component. The sequencing platforms can be further categorized as either single molecule-based (sequencing a single molecule) or ensemble-based (sequencing of multiple identical copies of a DNA molecule). Transcriptome assembly is necessary to transform individual reads into sequences of entire mRNAs or noncoding transcripts. The longer the individual reads, the simpler it is to assemble transcripts unambiguously [8]. After the sequencing, data must be processed to not only identify matches to the transcriptome, but also for assembly into transcripts before any biological interpretation. Data is most often supplied in FASTQ format. This format contains an ID number for each read, the read sequence, and a quality score. There are two main steps for sequencing data analysis:

1. removal of sequencing artifacts and errors from the data set. Artifacts may include the ligation adaptors and low-complexity reads. There are publicly available tools that can be used for this purpose;

2. alignment of the processed data to a reference genome using an appropriate aligner and downstream data analysis. Although most of the commercial sequencing platforms have developed their own data analysis pipelines, there are some publicly available programs that can be freely downloaded and efficiently run by individual laboratories to carry out total RNA-seq data analysis.

Further in-depth analysis, may include differential expression analysis, consisting in the use of a variety of statistical models to assess the significance of the data. The final data can then be viewed in a visualization program.

1.4 Gene expression profiling and microarray technology

Microarray-based assay technology was born to allow investigators to measure the expression profile of thousands of genes in a single experiment. Traditional gene expression analyses, such as Southern and Northern blotting, and microarray technology deal with one basic principle, known as hybridization. Complementary nucleic acids will hybridize, and this hybridization provides great selectivity of complementary stranded nucleic acids, with high sensitivity and specificity. In microarray-based technologies, the solid surface, such as glass, contains hundreds to thousands of immobilized DNA spots (targets) which can be simultaneously hybridized with two samples (probes) labeled with different fluorescent dyes, while in the traditional techniques, such as Southern and Northern blotting, this simultaneous hybridization of test and reference samples becomes complicated. Today, two different formats of microarray-based technologies are available, depending on the target nucleic acid components: the oligonucleotide array and the cDNA microarray. The oligonucleotide type of array consists of oligonucleotide targets which are generated *in situ* on a solid surface by light-directed synthesis. Synthetic linkers modified with photochemically removable protecting groups are attached to the glass substrate. Light is then directed through a photolithographic mask to specific areas on the surface to produce localized photodeprotection. Hydroxyl-protected deoxynucleotides are incubated with the surface so that chemical coupling occurs at the sites that have been illuminated in the previous step. By repetition of these procedures with new masks, hundreds of thousands of oligonucleotides can be synthesized in a very small area. In contrast, the cDNA microarray is fabricated by the printing of cloned and amplified cDNAs onto the solid surface. The advantages of the cDNA microarray compared with the oligonucleotide array have been thought to include less susceptibility and higher specificity due to the longer sequences of the targets. However, cDNA may contain repetitive sequences that are

often observed in various genes, or similar sequences that are found in family member genes. These non-specific sequences may affect the sensitivity of the cDNA microarray [9]. After being hybridized and washed, the microarray is scanned by means of a dual-wavelength confocal laser scanner. For fluorescent signals to be detected, wavelengths of 532 nm and 635 nm are required for Cy3 and Cy5, respectively. Scanning of the hybridized microarray should be carried out immediately after the washing, because the fluorescent dyes lose signal intensity with time. For an accurate comparison of two samples, the scanned signal intensities of Cy3 and Cy5 should be at the same level.

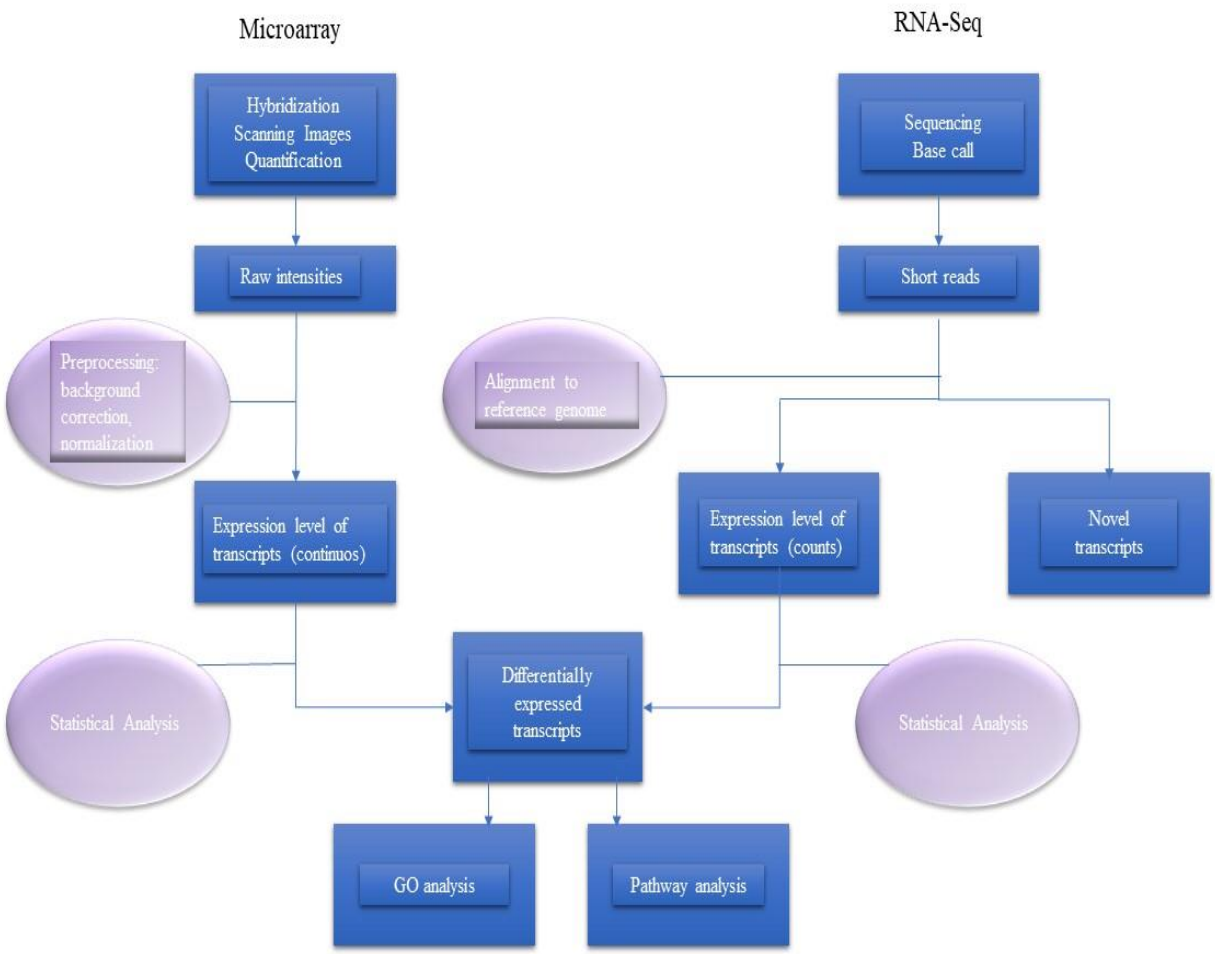

Figure 2 – An overview of analysis workflow for microarray and RNA-seq transcriptional profiling

Figure 2 represents how both methodologies, microarray and RNA-Seq, work, showing an overall of each pipeline.

1.5 Cancer and neurodegenerative diseases

Bioinformatics finds interesting applications in the study of several diseases, making easier the data mining process and facilitating the comparison with publicly available datasets, the statistical analysis and graphical visualization of the results: this allows researchers to form a general framework of a specific pathology as broad and comprehensive as possible in order to draw observations and connections that might otherwise escape analysis. It is only thanks to bioinformatics that a link between cancer and neurodegenerative diseases has recently been discovered: although they are two distinct pathological disorders, emerging evidence indicates that these two types of disease share common mechanisms of genetic and molecular abnormalities. Mutations in a variety of genes involved in regulation of the cell cycle, DNA repair pathways, protein turnover, oxidative stress, and autophagy have been implicated in both of these apparently dichotomous diseases [10]. Genes included in some specific pathways, including amyloid precursor protein (APP), ataxia-telangiectasia mutated (ATM), phosphatase and tensin homolog (PTEN), parkinson protein 2 E3 ubiquitin protein ligase (PARK2), protein tyrosine phosphatase delta, β-secretase (BACE1), and mammalian target of rapamycin (mTOR), are involved in the pathogenesis of both cancer and neurodegenerative diseases. Some of these genes have been demonstrated to be targets of miRNA regulation, suggesting that these miRNAs are also involved in the pathogenesis of both diseases. Researchers have discovered an inverse correlation between the risk of getting cancer and of getting a neurodegenerative disease. Recently, has been highlighted, through a genomic-type study on ageing, that aging and neurodegeneration are marked by induction of inflammatory genes and suppression of cell-cycle genes, while the opposite happens in cancer. For example, let us consider Alzheimer's disease (AD): it is surely one of the most devastating neurodegenerative pathologies that especially afflicts the elderly, and despite decades of intense research, there is still no effective treatment. Interesting studies suggest that cancer survivors have a lower risk of AD, and that people with AD have a lower risk of cancer. It is quite plausible that anti-cancer treatments could decrease AD risk, since some common drugs are known to improve outcomes in mouse models of AD through multiple mechanisms including stabilization of microtubules and dissolution of tangles. Other cancer drugs interrupt the cell cycle in its early stages and may thus prevent neuronal cell death. Furthermore, cancer and AD share many key

pathophysiologic features, including oxidative stress, metabolic dysregulation, DNA damage, and inflammation. Agents that suppress these pathways might be used as chemoprevention for both diseases. Another example is the diabetes drug metformin, for which there is emerging evidence of both anti-neoplastic and neuroprotective effects [11]. Both Alzheimer's and Parkinson's diseases (PD) are less frequent in survivors of many types of cancer (and vice versa), suggesting that a propensity towards one family of diseases may decrease the risk of the other. However, the inverse correlation does not seem to apply to any type of cancer; the increased risk of malignant melanoma in patients with PD seems to prove this fact. More and more literature on the subject, describes genes, proteins, and pathways dysregulated in both cancer and neurodegenerative disease— often in opposite directions. For example, the expression of p53, a known tumor-suppressor gene, is upregulated in AD, PD and Huntington's disease (HD), but downregulated in the large majority of cancers [12].

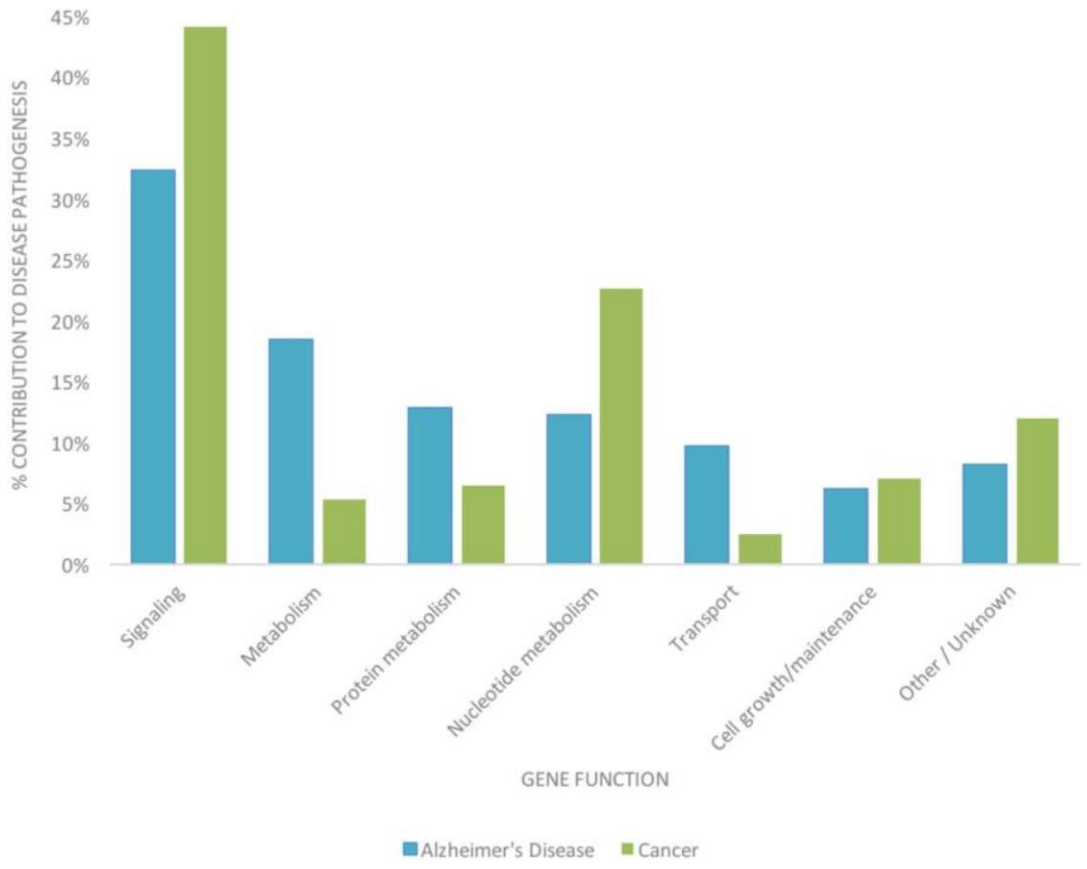

Figure 3 - NetAge gene categorization of AD and cancer [12]

The NetAge database (http://antiaging.org.ua/projects/356-netage-project-org) is an online biogerontological database for the study of aging, longevity, and age-related diseases. It contains

lists of genes involved both in cancer and AD where each gene is classified within a categorical function creating a topologic map of other genetic interactions, as shown in Figure 3: this figure shows clearly how contributions to disease pathogenesis from each relevant categorical function show an inverse trend both in AD and cancer: for example, genes related to nucleotide metabolism are more prominent in cancer than in AD, reflecting the mitotic activity of these cells. General metabolism, protein metabolism, and transport are elevated in AD when compared with cancer, a reflection of the importance of waste clearance and maintenance of cellular structure and function in these long-lived cells [12].

1.6 Neurodegenerative diseases

Neurodegenerative diseases (NDDs) represent specific disturbs dealing with a progressive loss of functionality of synapses, neuron and glial cells, mostly caused by a deposition of physiochemically altered variants of physiological proteins in the nervous system, both on neurons and glial cells. These altered proteins, or misfolded proteins, seem to play a central role in the pathogenesis of NDDs, and have been considered as potential biomarker in recent new studies [13]. NDDs are classified on the basis of their clinical presentation, anatomical regions and cell types affected, and misfolded proteins involved in the pathogenetic process. Clinical manifestations include cognitive decline, dementia and alterations in high-order brain functions, or movement disorders, like hyperkinetic, hypokinetic, cerebellar, or upper and lower motor neuron dysfunction or early combinations of these. The *molecular pathological classification* focuses on the distinction of synaptic, intracellular and extracellular protein accumulations. The proteins associated with the majority of sporadic and genetic adult-onset NDDs are:

- amyloid-beta (Aβ), which is cleaved from the transmembrane amyloid precursor protein (APP), a 770-aa protein—the APP gene has been mapped to chromosome 21q21.3;
- α-synuclein, a 140-aa protein encoded by a gene (SNCA) on chromosome 4;
- prion protein (PrP), which is a 253-aa protein encoded by the gene of PrP (PRNP) located on chromosome 20;
- the microtubule-associated protein tau is represented by different isoforms and encoded by a single gene (MAPT) on chromosome 17q21;
- transactive response DNA-binding protein 43 (TDP-43), a highly conserved nuclear 414-aa protein encoded by the TARDBP gene on chromosome 1;

- FET (abbreviation of the following three proteins) proteins, which include the fused in sarcoma (FUS), Ewing's sarcoma RNA-binding protein 1 (EWSR1) and TATA-binding protein-associated factor 15 (TAF15).

Consequently, NDDs are classified in

- Tauopathies - are classified as primary (when tau pathology is the driving force in the pathogenesis) or secondary. The mixed 3R and 4R tauopathy affecting the medial temporal lobe, termed also as NFT-predominant dementia, represents the severe end of primary age-related tauopathy (PART).
- α-synucleinopathies - two major groups are distinguished: the neuron-predominant α-synucleinopathies showing Lewy body pathology, and multiple system atrophy (MSA) which is dominated by glial cytoplasmic inclusions (PappLantos bodies).
- TDP-43 proteinopathies - TDP-43 is a major component of the ubiquitin-positive inclusions that characterise amyotrophic lateral sclerosis (ALS) and a common form of frontotemporal lobar degeneration (FTLD).
- FUS/FET proteinopathies - Rare sporadic disorders associated with FTLD, such as basophilic inclusion body disease, atypical FTLD-U and neuronal intermediate filament inclusion disease, show neuronal (cytoplasmic/ nuclear) and glial cytoplasmic inclusions immunoreactive for FET proteins.
- Prion diseases - are classified based on the aetiology as idiopathic/sporadic, acquired and genetic forms. A clinicopathological grouping is based on historical descriptions, which define phenotypes as Creutzfeldt-Jakob disease (CJD: spongiform encephalopathy; sporadic, iatrogenic, variant or genetic), kuru, Gerstmann-Sträussler-Scheinker disease (PrP-amyloidosis), and familial or sporadic fatal insomnia (selective thalamic degeneration without prominent spongiform change).
- Trinucleotide repeat diseases.
- Neuroserpinopathy, ferritinopathy.
- Cerebral amyloidosis.

Aβ, as one of the most frequently detected NDD-associated proteins, accumulates in Alzheimer's disease (AD) together with tau. AD is characterised by the extracellular deposition of Aβ fibrils and by the intraneuronal accumulation of abnormally phosphorylated tau protein.

Currently, there is no cure for any neurodegenerative disease and the treatments available only manage the symptoms or slow down the progression of the disease. New therapies can arise from

three main sources: synthesis, natural products, and existing drugs. This last source is known as *drug repurposing*, which is the most advantageous, since the drug's pharmacokinetic and pharmacodynamic profiles are already established, and the investment put into this strategy is not as significant as for the classic development of new drugs. There have been several studies on the potential of old drugs for the most relevant neurodegenerative diseases, including Alzheimer's disease, Parkinson's disease, Huntington's disease, Multiple Sclerosis and Amyotrophic Lateral Sclerosis [14].

1.6.1 Neuropathological Classification and overview of features

1. Alzheimer's Disease

Alzheimer's disease (AD) is one of the most devastating brain disorders of elderly humans. The clinical manifestations of AD include disturbances in the areas of memory and language, visuospatial orientation, and higher executive function. Noncognitive changes include personality changes, decreased judgment ability, wandering, psychosis, mood disturbance, agitation, and sleep abnormalities. The main neuropathological features of AD appear to be senile plaques and neurofibrillary tangles. The senile plaques seem to develop first in brain areas associated with cognition and spread to other cortical areas as the disease progresses. The senile plaques consist, among other components, of insoluble deposits of amyloid p-peptide (Aβ), a fragment of the amyloid precursor protein (APP). Aβ peptide is generated from APP by two consecutive cleavage events: proteolytic activity by β-secretase generates one end of the Ap peptide, while γ-secretase generates the other end, also by proteolysis. There appear to be two types of Aβ: a longer species, Aβ42, and a shorter species, Aβ40. Aβ42 seems to be deposited initially and may have a role in initiating the events that ultimately lead to amyloid deposition. It is still not clear if the senile plaques are the cause or a by-product of AD, although there are increasing data that dysfunction in the metabolism of APP with subsequent increase in the insoluble Aβ is responsible for AD. Aβ seems toxic to the neuron either directly, or indirectly by causing inflammation or increasing the production of free radicals.

The accumulation of neurofibrillary tangles in neurons is a second distinguishing feature of AD. Neurofibrillary tangles are mostly formed by chemically altered (abnormally folded and phosphorylat3. ed) tau protein, a protein involved in microtubule formation. Tangle formation is

related to the severity of disease; the more advanced the stage of disease, the more tau tangles in the brain [15].

2. Prion Diseases

Prion (pree-ahn) diseases are a group of neurodegenerative diseases caused by the conversion of the normal prion protein (PrP^{C}, prion-related protein, in which C stands for the cellular form of the protein) with a primarily α-helical structure into an abnormal form of the protein called the prion (PrP^{Sc}, in which Sc stands for scrapie, the prion disease of sheep and goats), which stands for *proteinaceous infectious particle*, and has a primarily β-pleated sheet structure. Prion diseases can occur by three mechanisms: spontaneous (sporadic), genetic (familial), and acquired (infectious/ transmitted). The model of prion disease is that the pathologic disease-causing misfolded form of the prion protein, PrP^{Sc} (in which "Sc" stands for scrapie, the prion disease of sheep and goats) acts as a template, such that when it comes into contact with a prion protein, PrP^{C} (in which "C" stands for the normal, cellular form of the protein), it transforms PrP^{C} into PrP^{Sc}, resulting in two prions. These two prions, in turn, transform two more PrP^{C} into PrP^{Sc}, which then transform four more, and so forth, leading to an exponential transformation and accumulation of prions. Histopathologic changes in human prion diseases include nerve cell loss, gliosis, vacuolation (formerly called spongiform change), and PrP^{Sc} deposition [16].

3. Tauopathies

Tauopathies are neurodegenerative disorders characterized by the deposition of abnormal tau protein in the brain. The spectrum of tau pathologies expands beyond the traditionally discussed disease forms like Pick disease, progressive supranuclear palsy, corticobasal degeneration, and argyrophilic grain disease. Emerging entities and pathologies include globular glial tauopathies, primary age-related tauopathy, which includes neurofibrillary tangle dementia, chronic traumatic encephalopathy (CTE), and aging-related tau astrogliopathy. Clinical symptoms include frontotemporal dementia, corticobasal syndrome, Richardson syndrome, parkinsonism, pure akinesia with gait freezing and, rarely, motor neuron symptoms or cerebellar ataxia. Some disorders show specific neuroimaging features, while examination of the cerebrospinal fluid awaits markers for in vivo stratification of cases. The possibility of cell-to-cell propagation is a novel aspect of the pathogenesis of tauopathies, which is partly reflected by the hierarchic involvement of anatomic regions [17].

4. α-synucleinopathies

The term alpha-synucleinopathy is used to name a group of disorders having in common the abnormal deposition of alpha-synuclein in the cytoplasm of neurons or glial cells, as well as in extracellular deposits of amyloid. In Parkinson's disease and Lewy body dementia, alpha-synuclein is the main component of Lewy bodies and dystrophic neurites; alpha-synuclein also accumulates in the cytoplasm of glial cells. In multiple system atrophy, alpha-synuclein conforms the cytoplasmic oligodendroglial inclusions and the neuronal inclusions which are the hallmark of this disease. Finally, the amyloidogenic fragment 61-95 amino acids of alpha-synuclein is the non-Abeta component of senile plaque amyloid in Alzheimer disease. Accumulations of alpha-synuclein in all these disorders have in common a fibrillar configuration, but they differ in the binding of alpha-synuclein to distinct proteins with the exception of ubiquitin whose binding to alpha-synuclein is common to all alpha-synuclein inclusions [18].

5. TDP-43 proteinopathies

The pathological sequestration of TAR DNA-binding protein 43 (TDP-43, encoded by TARDBP) into cytoplasmic pathological inclusions characterizes the distinct clinical syndromes of amyotrophic lateral sclerosis and behavioral variant frontotemporal dementia, while also co-occurring in a proportion of patients with Alzheimer's disease, suggesting that the regional concentration of TDP-43 pathology has most relevance to specific clinical phenotypes [19]. Trans-activation response DNA-binding protein of 43 kDa (TDP-43), encoded by the *TARDBP* gene on chromosome 1, is a major component of tau-negative and ubiquitin-positive inclusions that characterize amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration (FTLD) linked to TDP-43 pathology (FTLD-TDP). TDP-43 aggregation and neuropathology have been observed in a spectrum of distinct neurodegenerative disorders collectively known as the TDP-43 proteinopathies, suggesting a central role for TDP-43 in neurodegenerative disease pathogenesis. Indeed, the identification of more than 35 missense mutations in the *TARDBP* gene has further implicated abnormal TDP-43 function as a cause, rather than a consequence, of neurodegeneration in ALS and FTLD-TDP [20].

6. FUS-Proteinopathies

FUS pathology characterizes familial ALS cases and a rare group of diseases with frontotemporal lobar degeneration (FTLD). Sporadic disorders of FTLD, such as basophilic inclusion body disease (BIBD), atypical FTLD-U (aFTLD-U, where U represent ubiquitin-only, later designated as FTLD-TDP), and neuronal intermediate filament inclusion disease (NIFID), feature neuronal (cytoplasmic/nuclear) and glial cytoplasmic inclusions immunoreactive for FUS, thus establishing a

new category (FUS/FET proteinopathies). Immunoreactivity for further FET proteins is helpful to distinguish sporadic and FUS-gene mutation related cases [13].

1.7 Conclusions

Apparently, cancer and neurodegenerative diseases seem to have very little in common. Indeed, on closer inspection, they seem to be the opposite of each other: one is characterized by an improved resistance to cell death, the others by a tendency to an early cell death [21]. However, many epidemiological studies have found connections between cancer and neurodegeneration that seem to suggest the existence of an inverse correlation between the odds of contracting one of the two diseases. The origins of the association between cancer and neurodegeneration are still unclear, but increasing evidence suggests that new discoveries in genetics of these two conditions could help scientists find some answers about the cancer–neurodegeneration relationship in the next future. Several studies, indeed, show that the genes causing neurodegeneration are often mutated or abnormally expressed in cancer, thus suggesting the idea that genes that predispose to cancer also cause neurodegeneration and vice versa. In this context, a great help may come from Bioinformatics: in the hand of skilled researcher, it may become a powerful tool to accelerate and facilitate all the studies finalized to the discovery of all the common genetic causes underlaying both pathologies, making the possibility of finding effective treatments over the next ten years more concrete.

2 AIM OF THE THESIS

Bioinformatics finds several and interesting applications in the study of a great number of different diseases: it allows researchers to observe the most complexes pathologies from a new perspective, giving a complete and connected general vision, being capable of integrating the most diverse and apparently distant branches of scientific research into a single discipline. It is only thanks to Bioinformatics that today researchers all over the world have the possibility to easily compare their results thus finding biological connections which otherwise would go unnoticed.

Aim of this thesis is to show the power of bioinformatic tools in making data analysis easier and faster and in inferring biological interpretations and connections from the results as accurate and precise as possible. This goal is achieved and described through three different research projects, which embrace different aspects of what can be considered the main fields of application of bioinformatics: proteomics, transcriptomics, genomics, systems biology, modeling, imaging and clinical bioinformatics.

The issues addressed in this thesis work and the developed methodologies mainly concern tumors and neurodegenerative diseases but can easily be extended and adapted to other diseases too. Chapter three deals with the analysis of the role of MAP3K8, a serine/threonine kinase expressed in thyroid cancer stem cells (CSCs), in mediating drug resistance in human thyroid cancer and its relationship to tumor behavior. Our results can pave the way for future more in-depth analyzes of all the cascade of reactions involved in the development and progression of thyroid cancer with particular attention to the causes of non-response to conventional therapies.

In chapter four we describe the process of modeling, simulation and prediction of protein structures for the design of oncolytic viruses (OVs). OVs can be considered as a new powerful weapon in the fight against cancer thanks to their ability to restore the immune system response, impaired by tumor progression, and to selectively infect and destroy cancer cells without harming healthy cells. Laboratory tests confirmed our predictions, thus proving that bioinformatic analysis is effective in directing laboratory tests to subjects it deems promising, saving time and resources.

Chapter five deals with RNA-sequencing data analysis, providing a complete workflow which can be easily adapted to the analysis of any type of samples or tissue in any context.

3 ANALYSIS OF MAP3K8 EXPRESSION IN THYROID CANCER

3.1 Thyroid Cancer Background

Thyroid cancer is the most common type of endocrine tumor showing an increasing incidence over the last three decades. Malignant carcinoma of the thyroid arises from two different cell types, follicular and parafollicular. Follicular cells are involved in the production of thyroid hormones and they may give rise to well-differentiated and anaplastic thyroid carcinomas [22].

Well-differentiated thyroid carcinoma is the most common endocrine neoplasia: it can often remain clinically silent for many years, and half of all cases come to medical attention as incidental findings on physical examination or ultrasonography, or as a previously unsuspected histological finding after surgery for benign thyroid disease [23]. It usually begins in adulthood as an asymptomatic thyroid mass. Papillary carcinoma (PTC) is the most common form of well-differentiated thyroid cancer, and the most common form of thyroid cancer to result from exposure to radiation. Papillary carcinoma appears as an irregular solid or cystic mass or nodule in a normal thyroid parenchyma [22].

Anaplastic carcinoma (ATC) is the most aggressive form of thyroid cancer. It is associated with an extremely poor prognosis, whose mortality rates up to 95%, in contrast to the usual behavior of the more common well-differentiated thyroid carcinoma. It is characterized by a high rate of proliferation and displays highly undifferentiated, spindle-shaped, epithelioid, giant cells that tend to replace the other cells in the thyroid and rapidly extend to the neck soft tissues [22] [24].

Poorly differentiated thyroid carcinomas (PDTCs) are a rare subtype of thyroid carcinomas that are biologically situated between well-differentiated thyroid carcinomas and anaplastic thyroid carcinomas. It can be considered as an independent thyroid cancer histotype. Despite its rarity, it represents the main cause of death from non-anaplastic follicular cell-derived thyroid cancer [24].

The parafollicular C cells are responsible for the calcitonin production and they may give rise to medullary thyroid cancer (MTC) [25]. In the hereditary form of medullary thyroid cancer, the growth of these cells is due to a mutation in the RET gene which was inherited. This mutated gene may first produce a premalignant condition called C cell hyperplasia. The parafollicular C cells of the thyroid begin to have unregulated growth and may form a bump or nodule in any portion of the thyroid gland. With this type of cancer, patients may not be diagnosed unless the cancer has spread to the lymph nodes of the neck and presented with a “lump in the neck [24].

Mutations in the *BRAF* serine/threonine kinase represent the most common genetic cause in thyroid cancer, occurring in approximately 45% of papillary thyroid cancer (PTC) and in a lower proportion of poorly differentiated thyroid cancer (PDTC) and anaplastic thyroid cancer (ATC) [22].

The *BRAF* gene has a key role in the creation process of a protein that helps transmit chemical signals from outside the cell to the nucleus. This protein is part of a signaling pathway known as the *RAS/MAPK* pathway, which controls several important cell functions. Specifically, the *RAS/MAPK* pathway regulates cell proliferation, differentiation, migration and the programmed death (apoptosis) [25].

The *BRAF* gene belongs to a class of genes known as oncogenes. When mutated, oncogenes have the potential to change normal cells to cancerous. Among the various Raf kinase isoforms, the *B-type RAF V600E* (*BRAFV600E*) mutation is the most commonly observed inducing excessive proliferation and differentiation of tumor cells at the initial tumor stage [25]. Moreover, *BRAF* gene mutations may be considered as a predictive factor for lymph node metastasis, extrathyroidal extension, advanced disease stages III and IV, and disease recurrence [26].

Identification of *BRAF* mutations in various tumor types has led to development of various *ATP*-competitive *RAF* inhibitors: *vemurafenib* is one of the most extensively evaluated agent in this class.

Our previous studies have shown that thyroid cancer stem cells derived from 8505 cell line are resistant to the *BRAF* inhibitor *vemurafenib*, despite harboring *BRAFV600E* mutation. In these cancer stem cells the resistance to *vemurafenib* was mediated by a paradoxical over-activation of *ERK* and *AKT* pathways. By our ordinary differential equations based computational model coupled with algorithmic approaches developed in previous study [22], we found a fundamental role of mitogen-activated protein kinase 8 (*MAP3K8*), a serine/threonine kinase expressed in thyroid CSCs, in mediating this drug resistance.

The computational model, developed with the purpose of simulate biological pathways and their interactions, allowed to reveal biochemical and genetic mechanisms underlaying the non-response to traditional treatments of thyroid cancer. This means that being able to highlight all the genetic alterations and all the activations of particular nodes within specific pathway, may lead to the identification of new potential biomarkers or drug target.

3.2 Materials and method

In order to analyze the trend of MAP3K8 expression values across all types of thyroid cancers (ATC, PTC, PDTC), the following datasets have been selected and downloaded from Gene Expression Omnibus data repository (https://www.ncbi.nlm.nih.gov/geo/):

- GSE33630 (11 ATC, 49 PTC and 45 normal samples, expression profiling by array)

- GSE76039 (17 PDTC and 20 ATC samples, expression profiling by array)

- GSE58545 (27 PTC and 18 normal samples, expression profiling by array)

- GSE3678 (7 PTC and 7 normal samples, expression profiling by array).

Using Bioconductor statistical programming language R [27], we performed a complete analysis on these data:

1. Survival analysis
2. Gene expression analysis
3. Pathway analysis
4. Stemness index evaluation

- Survival Analysis

Survival analysis is used to investigate the time it takes for an event of interest to occur. In cancer studies, the two most important measures are: i) the *time to death*; and ii) the *relapse-free survival time*, which corresponds to the time between response to treatment and recurrence of the disease. Two related probabilities are used to describe survival data: the *survival probability* and the *hazard probability*. The *survival probability*, also known as the survivor function $S(t)S(t)$, is the probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified future time t. The *hazard*, denoted by $h(t)h(t)$, is the probability that an individual who is under observation at a time t has an event at that time.

The Kaplan-Meier (KM) method is a non-parametric method used to estimate the survival probability from observed survival times [28].

The survival probability at time t_i, $S(t_i)$, is calculated as follow:

$$S(t_i) = S(t_i - 1)(1 - \frac{d_i}{n_i})$$

Where:

- $S(t_i - 1)$ = the probability of being alive at $(t_i - 1)$
- n_i = the number of patients alive just before t_i
- d_i = the number of events at t_i
- t_0= 0, $S(0)$ = 1

The estimated probability $S(t)$ is a step function that changes value only at the time of each event. It is also possible to compute confidence intervals for the survival probability.

The KM survival curve, a plot of the KM survival probability against time, provides a useful summary of the data that can be used to estimate measures such as median survival time.

Bioconductor survival package [29] has been used to obtain Kaplan-Meier survival curves. Kaplan-Meier curves show the survival probability over time of specific subjects under specific conditions, as shown in Figure 1:

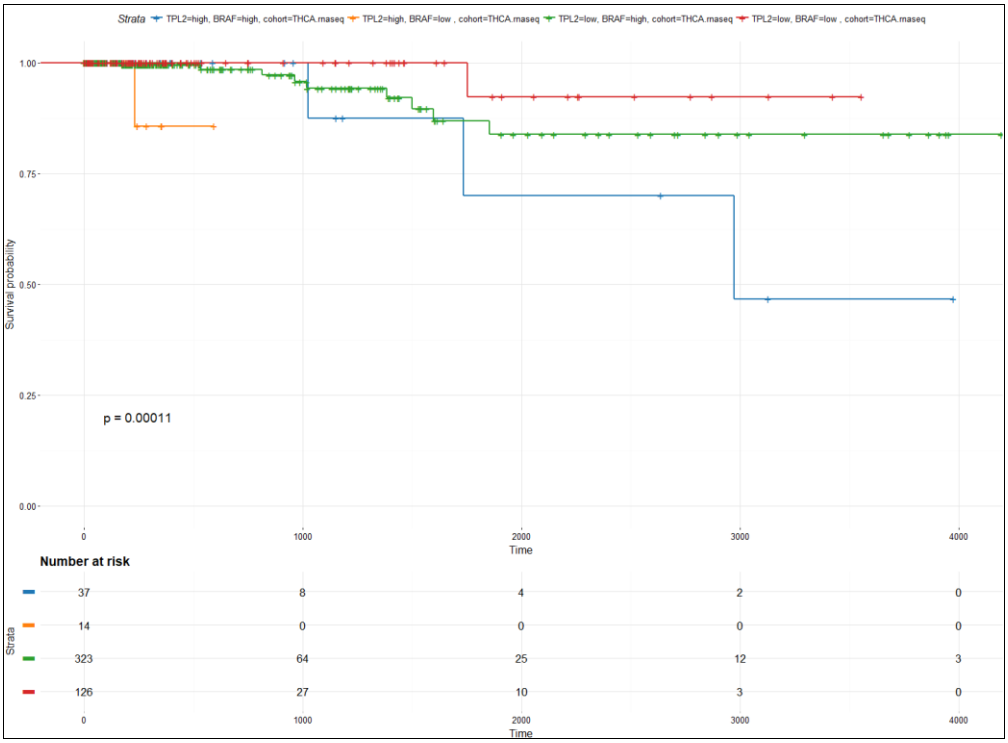

Figure 4. Kaplan-Meier curves show the survival probability of groups of patients with different BRAF and TPL2 expression levels.

In order to produce Kaplan-Meier curves, both clinical and expression values data were used.

- Gene expression Analysis

Bioconductor GEOquery package [30] were used for analyses of microarray data. GC-RMA algorithm was used for normalization process: GC-RMA algorithm used probe sequence information to estimate probe affinity to non-specific binding [31]. GCRMA incorporates probe sequence composition into background adjustment, following the physical model of Naef and Magnasco [32]. The model describes a probe affinity that is dependent on its base composition and the position of each base along the probe and suggests that probe sequence can significantly affect the intensity of the signal generated from that probe, independent of the concentration of its target.

$$\ln\left(B \middle/ N\right) = \sum_{k=1}^{25} \sum_{l \in (A,T,G,C)} S_{lk} A_{lk}$$

where B is the raw probe intensity, M is the median intensity of the array, l is the nucleotide index (A, C, G or T), k is the position of l along the probe (note that k has a range of 1 to sequence length, that is 25 for GeneChip probes), S is a Boolean variable equal to 1 if the probe sequence has l at k and zero otherwise, and A is the per-site-per- nucleotide affinity.

The first step of analysis consisted in filtering out uninformative data such as control probesets and other internal controls and removing genes with low variance, that could have compromised the results of statistical tests for differential expression. Once datasets were filtered, data were sent to limma package [33] for differential gene expression analysis.

- Pathway Analysis

Bioconductor package SPIA [34] has been used to implement a Signaling Pathway Impact Analysis (SPIA) on our list of differentially expressed genes along with their log fold changes and KEGG signaling pathway topologies in order to identify most relevant pathways under our conditions of study. Among all the 136 pathways we obtained, we selected only those where our target gene MAP3K8 was involved in:

i) MAPK signaling pathway (map04010)

ii) Toll-like receptor signaling pathway (map04620)

iii) T cell receptor signaling pathway (map04660).

All the mentioned above pathways are respectively depicted in Figure 5, 6 and 7.

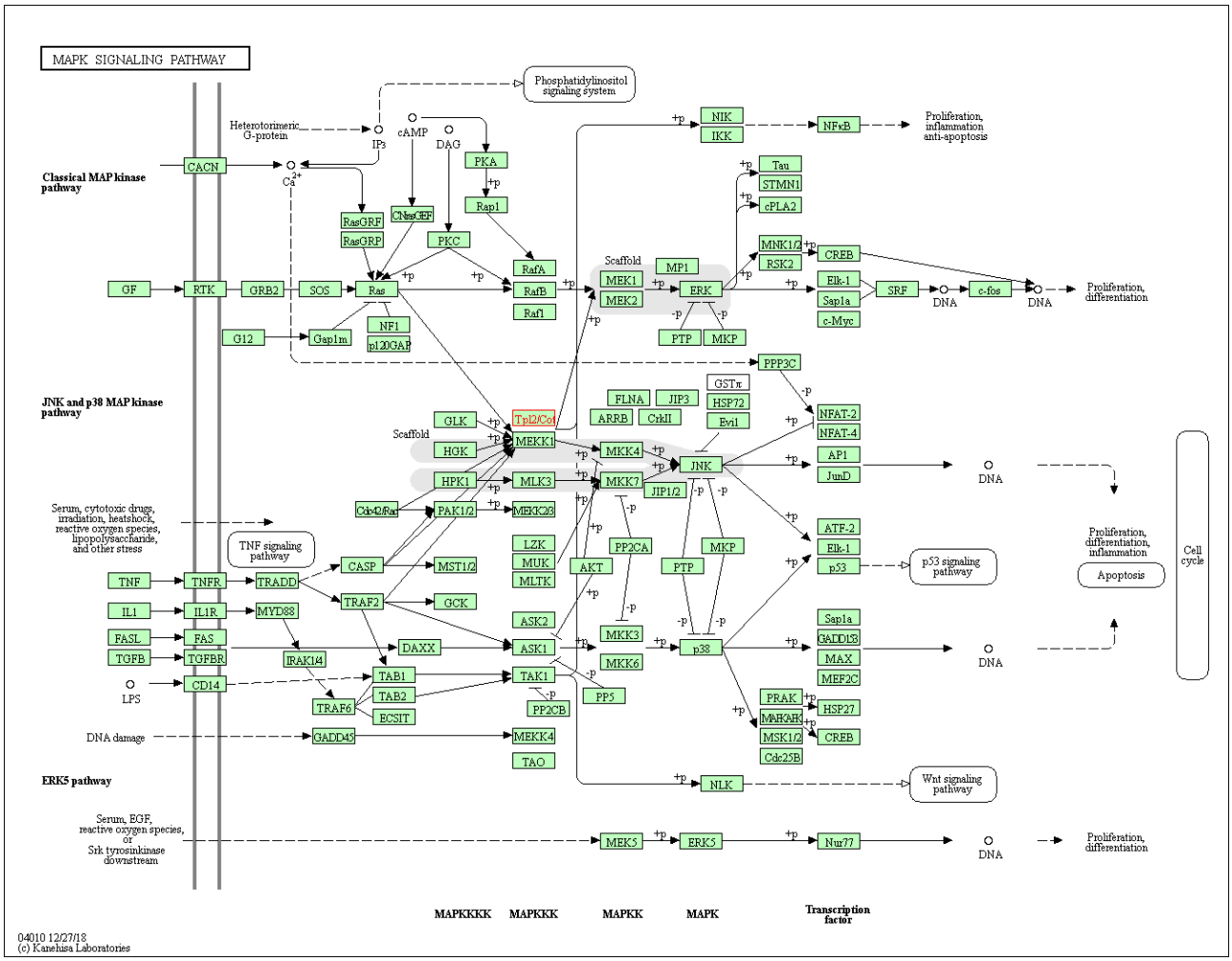

Figure 5. MAPK Signaling Pathway, MAP3K8 is highlighted in red (alias ID Tpl2/Cot).

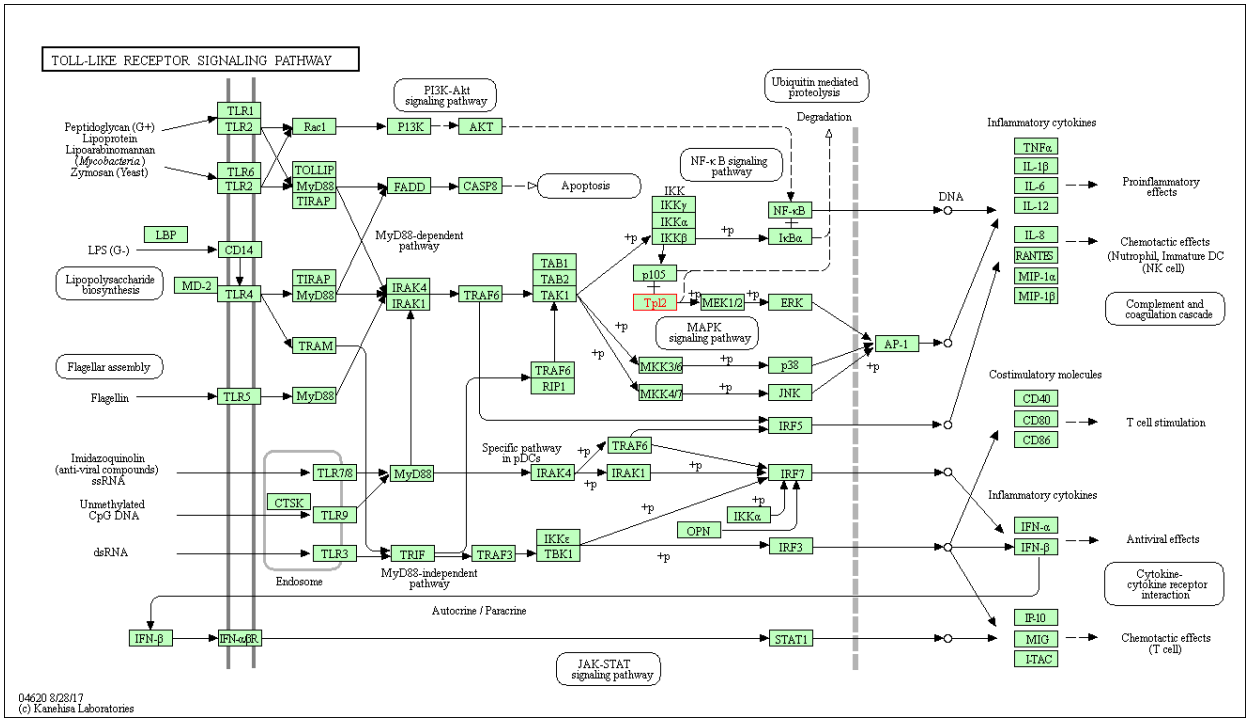

Figure 6. Toll-like receptor signaling pathway: Tpl2 target gene is highlighted in red.

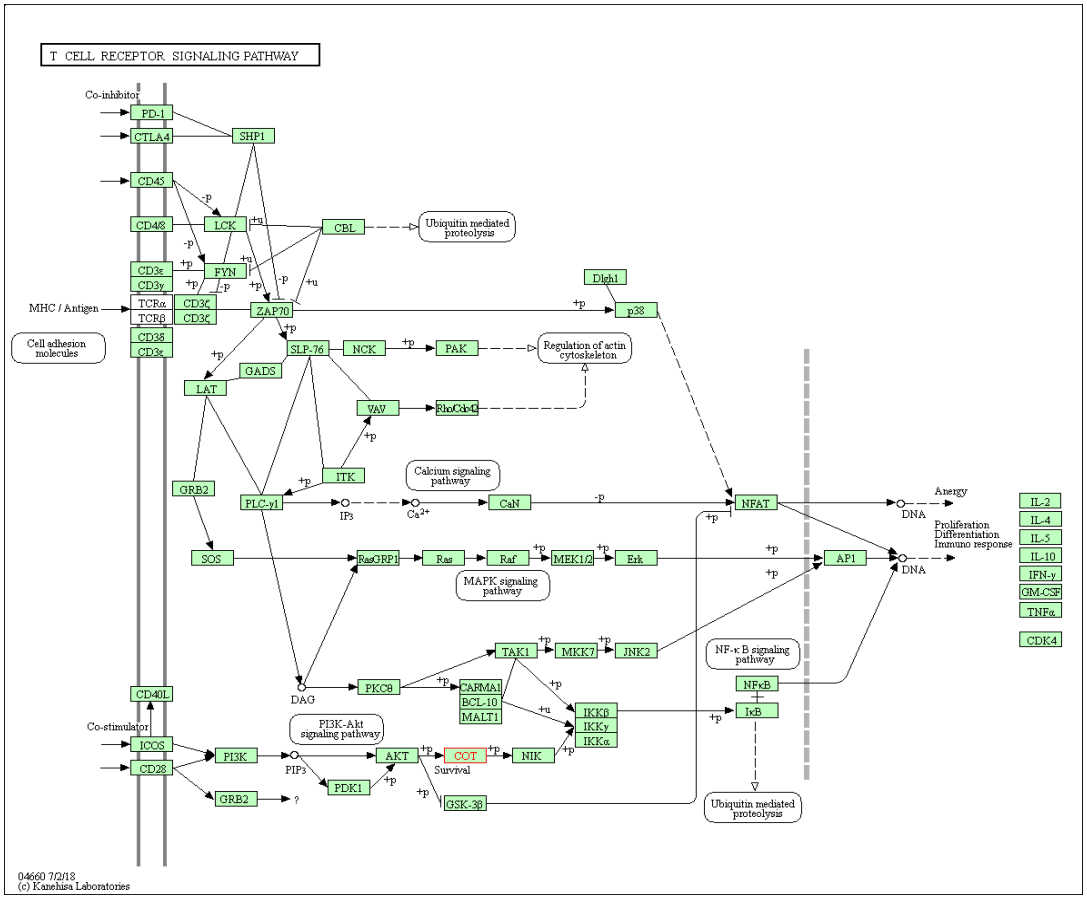

Figure 7. T cell receptor signaling pathway: COT (alias ID Tpl2/MAP3K8) is highlighted in red.

- Stemness index evaluation

As cancer progression is accompanied by a gradual loss of a differentiated phenotype and the acquisition of stem-cell like phenotype, we assessed whether the stemness degree of a particular thyroid cancer type could be a predictive parameter of tumor outcome. Some of the characteristics of stem like cancer cells are:

- extensive self-renewal ability [35]
- cancer-initiating ability on orthotopic implantation [35]
- karyotypic or genetic alteration [35]
- aberrant differentiation properties [35]
- capacity to generate non-tumorigenic end cells [35]
- multilineage differentiation capacity [35]

Using a one-class logistic regression (OCLR) algorithm trained on stem cell, we derived mRNA expression-based stemness index (mRNAsi) on 37 samples of our selected dataset GSE76039 [36].

Formally, given a set of *n* samples $\mathcal{X} = \{\boldsymbol{x_i}\}$, a *one-class logistic regression* model can be defined by a weight vector w that maximizes the log-likelihood

$$l(w \mid \mathcal{X}) = \sum_{i=1}^{n} \log p(x_i|w)$$

where the likelihood is modeled with the logistic function:

$$p(x_i|w) = \frac{e^{w^T x_i}}{1+e^{w^T x_i}}$$ [37]

Stemness index values range from low (0) to high (1): after calculating mRNAsi for each sample of our reference dataset, we ranked our data depending on mRNAsi value thus obtaining two groups, one showing high index values and the other one showing low index values, using a 0.5 value as cut-off.

3.3 Results and discussion

Differentially expressed genes (DEGs) provided a list with all genes displaying any changes in expression levels between normal and cancer samples. This list includes many genes whose involvement in many types of tumor formation is widely discussed in literature, as shown in Figure 8.

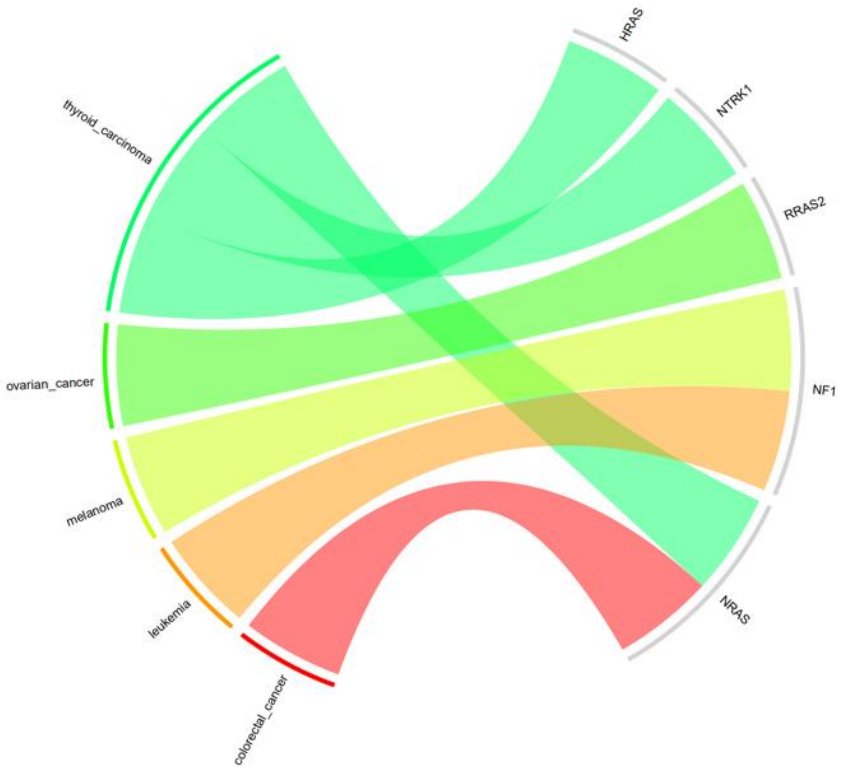

Figure 8. This graphic shows the correlation between different cancer types and some of the genes of our DEGs list.

Survival curves showed that death probability is higher in case of both BRAF mutation and MAP3K8 high expression levels, thus confirming MAP3K8 role as antagonist to BRAF inhibitor drugs. DEGs analysis conducted simultaneously on all thyroid cancer types across all samples showed a MAP3K8 over-expression in case of ATCs, with almost doubled expression values in ATCs compared to normal samples (Table 1).

Gene Symbol	**Normal**	**ATC**	**PDTC**
MAP3K8	2.745913	4.481513	3.450524

Table 1. Degs analysis shows MAP3K8 expression values across all types of thyroid cancer.

Stemness index evaluation showed a correlation between higher index values and MAP3K8 higher expression levels but this aspect needs to be further investigated in future.

MAP3K8 up-regulation in ATC samples could suggest possible correlations with some clinico-pathological parameters which are typical features of this tumor type. For example, as ATC is a rare and aggressive form of thyroid cancer that affects mainly elderlies displaying high prevalence of BRAF-V600E mutation, it is clear that a strong correlation between age and BRAF mutation exists. The possibility to detect this MAP3K8 over expression in previous tumor stages or in earlier ages could be a very powerful strategy for early diagnosis and prevention.

4 ONCOLYTIC VIRUSES ENGINEERING

4.1 Oncolytic viruses background

Immunotherapy represents the ultimate frontier in modern oncologic care: it enhances the immune system's ability to recognize, target, and eliminate cancer cells, wherever they are in the body, making it a potential universal weapon to cancer [38]. Although the progression of cancer leads to the inexorable weakening of the immune system, cancer immunotherapy can potentially reactivate the patient's suppressed immune system, ideally resulting in the eradication of the disease [38].

Immune system is a primitive and innate mechanism that is typically able to detect both internal threats (e.g., malignant cells) and external threats (e.g., viruses, bacteria, fungi, parasites) within a very short response time [38]. It becomes stronger during adulthood because the exposition to different pathogens provides more immunity. The innate immune response is not capable of immunological memory and is not directed against any particular target or organism. Conversely, the innate immune response has recognition properties with low specificity that are based on the molecular patterns displayed by membrane proteins called Toll-like receptors [38].

The T and B lymphocytes (T and B cells) are the only cells in the organism able to recognize and respond specifically to each antigenic epitope. The B cells have the ability to transform into plasmocytes and are responsible for producing antibodies (Abs) [39]. Antibodies are special proteins that lock on to specific antigens. Each B cell makes one specific antibody.

Humoral immunity depends on B cells activity while cell immunity depends on T cells one. There are distinct types of T lymphocytes:

- Helper T cells (Th cells) — they coordinate the immune response. Some of them communicate with other cells, and some stimulate B cells to produce more antibodies. Others attract more T cells or cell-eating phagocytes.
- Killer T cells (cytotoxic T lymphocytes) —these T cells attack other cells. They are particularly useful for fighting viruses. They work by recognizing small parts of the virus on the outside of infected cells and destroy the infected cells.

Cancers induce immune and inflammatory responses as they invade healthy tissue and metastasize, and even if immune system mechanisms are functional, tumors can escape from immune attack through a wide range of mechanisms. Major mechanisms by which tumors suppress the immune system and evade destruction include upregulation of checkpoint receptor ligands that

downmodulate tumor-infiltrating lymphocyte (TIL) activity; recruitment of suppressor immune cells, such as T regulatory cells (Tregs), tumor-associated macrophages, and myeloid-derived suppressor cells and the production of soluble factors associated with immunosuppression, such as IL-10 and transforming growth factor-beta (TGF- β) [38].

Immunotherapy can:

- educate the immune system to recognize and attack specific cancer cells;
- boost immune cells to help them in fighting and trying to eliminate cancer cells;
- provide the body with additional components to enhance the immune response.

Oncolytic viruses are a form of immunotherapy that employs attenuated viruses to restore the immune system response in order to infect and destroy cancer cells. Most available oncolytic viruses are genetically modified to enhance tumor tropism and reduce virulence for non-neoplastic host cells. Therefore, they can stimulate a proinflammatory environment to counteract the immune resistance of malignant cells. Oncolytic viruses also aim to take advantage from the tumor's tolerogenic mechanisms, which can facilitate viral infection and killing of cells that are not protected by the immune system. This generates a theoretical domino effect including chained viral transference between neoplastic cells and further immune activation [40].

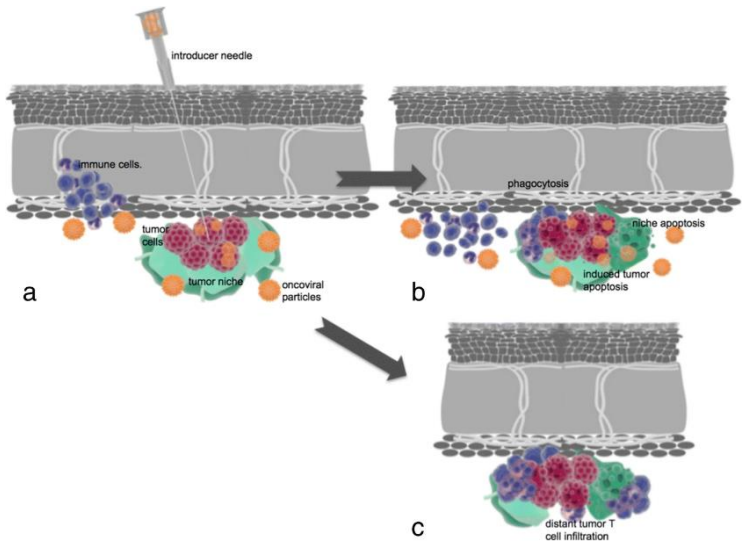

Figura 9. ***a*** *Intratumoral inoculation of an oncolytic virus with transfection and early immune cell recruitment.* ***b*** *Advanced transfection of an oncolytic virus into tumor and niche cells with induction of immune cells resulting in apoptosis, direct cell lysis, niche disruption, and phagocytosis.* ***c*** *Distant tumor immune infiltration induced by local immune conditioning. Blue: immune cells. Red: tumor cells. Orange: oncoviral particles. Green: tumor niche [40]*

Just like other immunotherapies types, the action of oncolytic viruses manifests itself in a multi-active mechanism with both direct and indirect toxic effects on tumor cells such as autolysis, immune cell honing, destruction of vascular supply and potentiation of other adjunctive anti-cancer therapies [40].

Oncolytic virus platforms under evaluation in clinical trials include:

- Adenovirus: a family of common viruses that can cause a wide range of typically mild effects including sore throat, fatigue, and cold-like symptoms [41].
- Herpes simplex virus: a virus that can cause the formation of sores on or near the mouth [41].
- Maraba virus: a virus found exclusively in insects [41].
- Measles virus: a highly contagious virus that infects the respiratory tract and can cause measles [41].
- Newcastle Virus: a virus primarily found in birds; can cause mild conjunctivitis and flu-like symptoms in humans [41].
- Picornavirus: a family of viruses that can cause a range of diseases in mammals and birds; the coxsackie virus is an example from this family that is being clinically tested [41].
- Reovirus: a family of viruses that can affect the gastrointestinal and respiratory tracts in a range of animal species [41].
- Vaccinia virus**:** the virus that was used to help vaccinate against and eliminate smallpox; rarely causes illness in humans and is associated with a rash covering the body [41].
- Vesicular stomatitis virus: a virus that belongs to the same family as the Maraba virus; can cause flu-like symptoms in humans [41].
- Rubeola virus: a highly contagious virus that infects the respiratory tract and can cause measles [41].

We focused in particular on measles virus (MV): attenuated MV are potent and selective oncolytic agents showing impressive antitumor activity. To minimize risk to the patient, an ideal oncolytic virus should have two fundamental requirements:

1- be selective for the tumor, nonpathogenic for normal host tissues, non-persistent and genetically stable;

2- be non-transmissible and preferably derived from a virus to which the population is generally immune.

Attenuated MV fulfill the above requirements [42].

4.2 Materials and methods

Three protein molecules have been identified as MV receptors in facilitating MV entry: the signaling lymphocyte activation molecule (SLAM), CD46, and nectin-4. Wild-type MV do not, in general, use CD46 as a cell entry receptor, but acquire the CD46 tropism during tissue culture adaption via a mutation in the H-attachment protein coding sequence that changes the amino acid at position 481 in the H-protein, from asparagine to tyrosine. Attenuated MV strains carrying this mutation are typically selected. CD46, also known as membrane cofactor protein, is expressed by all human cells except erythrocytes and acts as a cofactor for complement factor I, a serine protease which protects autologous cells against complement-mediated injury by cleaving C3b and C4b deposited on host tissue. C4b and C3b are opsonins, which are able to bind covalently to glycoproteins on the target cell surface. Opsonization by C3b or C4b leads to engagement of complement receptors (CR1, CR3 or Cr4) on acceptor phagocytic cells, targeting foreign particles for phagocytosis. C3b and C4b have been reported to facilitate the clearance of immune complexes (IC) which begins with covalent attachment of C3b or C4b to IC [43]. CD46 also acts as a costimulatory factor for T-cells which induces the differentiation of CD4+ into T-regulatory 1 cells. T-regulatory 1 cells suppress immune responses by secreting interleukin-10, and therefore are thought to prevent autoimmunity. CD46 is frequently overexpressed on human cancer cells compared with their normal non-transformed counterparts, and this could possibly be as a sort of protection mechanism of the cancer cells from complement mediated lysis. Overexpression of CD46 has been documented in gastrointestinal, hepatocellular, colorectal, endometrial, cervical, ovarian, breast, renal, and lung carcinomas, also in leukemias and multiple myeloma, and has been found to limit the therapeutic potential of monoclonal antibody therapy. CD46 mediates not only the attachment and entry of attenuated measles viruses, but also drives the process of virus induced cell-to-cell fusion between a virus infected cell and its neighboring cells [42].

In collaboration with the Etna Biotech company of Catania, we investigated the antitumor efficacy of *cetuximab*, a widely used anti-epidermal growth factor receptor (EGFR) monoclonal antibody, combined with MV.

MV enters cells through membrane fusion in a pH-independent manner, like other paramyxoviruses. MV possesses two glycoproteins on its envelope, an attachment protein hemagglutinin (H) (MV-H) and a fusion (F) protein (MV-F). MV-H and MV-F form hetero-oligomers required to induce membrane fusion. Upon receptor binding, MV-H is thought to undergo a conformational change,

which in turn would trigger a structural rearrangement of MV-F from the metastable pre-fusion forms to the intermediate and post-fusion ones. This conformational change of MV-F would drive the fusion between the viral envelope and the host cell membrane [44].

The paramyxovirus attachment proteins are classified into three groups based on their functions: the hemagglutinin–neuraminidase (HN) protein, the H protein, and the G protein [44].

Epidermal growth factor receptor (EGFR) signaling is involved in apoptosis, angiogenesis, cell proliferation, migration, and invasion. Cetuximab binds to the extracellular domain of EGFR and has been applied widely to suppress tumor growth. Inhibition of EGFR activation leads to downregulation of the RAS/RAF/MEK/ERK and PI3K/AKT pathways, which decrease vascular endothelial growth factor (VEGF) promoter activity. Inhibition of EGFR activity by cetuximab induces an antiangiogenic effect by decreasing VEGF production. Moreover, cetuximab can induce antibody-dependent cellular cytotoxicity (ADCC) through Fcγ receptors on immune effector cells, such as natural killer (NK) cells and macrophages [45].

To evaluate the functionality of the 4 protein models representing H protein-cetuximab fusion structure provided by the company, SWISS-MODEL (http://swissmodel.expasy.org) server for automated comparative modeling of three-dimensional (3D) protein structures was used. 3D protein structures allow an effective design of experiments, such as site-directed mutagenesis, studies of disease-related mutations or the structure based design of specific inhibitors, thus providing valuable insights into the molecular basis of protein function [46]. The SWISS-MODEL workspace integrates programs and databases required for protein structure modelling in a web-based workspace. SWISS-MODEL workspace is a web-based integrated service dedicated to protein structure homology modelling: homology models of proteins are of great interest for planning and analysing biological experiments when no experimental three-dimensional structures are available [47]. Homology (or comparative) modelling methods make use of experimental protein structures ("templates") to build models for evolutionary related proteins ("targets"). Among all current theoretical approaches, comparative modeling is the only method that can reliably generate a 3D model of a target protein from its amino acid sequence.

SWISS-MODEL pipeline comprises the following four main steps:

1. Identification of structural template(s).

   Basic local alignment search tool (BLAST) and HH-suite (HHblits) are used to identify templates. HHblits is an open-source software package for sensitive protein sequence and containing programs that can search for similar protein sequences in protein sequence databases.

   The templates are stored in the SWISS-MODEL Template Library (SMTL), which is derived from Protein Data Bank (PDB).

2. Alignment of target sequence and template structure(s).
3. Model building and energy minimization.
4. Assessment of the model's quality using QMEAN, a statistical potential of mean force.

These steps can be iteratively repeated, until a satisfying model structure is achieved. Possible applications of protein models depend mostly on the quality of the models. The accuracy of a model can vary significantly, even within different regions of the same protein: usually highly-conserved core regions can be modeled in a more reliable way than variable loop regions or surface residues. SWISS-MODEL offers several tools to evaluate the reliability of the model: similar to B-factors in crystal structures, the corresponding column in SWISS-MODEL result files consists of a C-score, which gives an estimate of the variability of the template structures at this position. Briefly, SWISS-MODEL analyzes the model by breaking it down into its parts, recognizes the sequences and associates them with known sequences. Based on the degree of similarity of these sequences with the known ones, and on the basis of other parameters that are GMQE (estimate of the overall quality of the model) and the QMEAN (linked to different geometric properties of the model) returns a first evaluation of the goodness of the model.

The four most satisfying structures obtained from SWISS-MODEL have been subsequently subdued to ProQ – Protein Quality Predictor (https://proq.bioinfo.se/ProQ/ProQ.html). ProQ is a neural network-based predictor that based on a number of structural features predicts the quality of a protein model. It use two quality indexes to assess the quality score of a model: LGscore and MaxSub.

LGscore is -log of a P-value and MaxSub ranges from 0-1, were 0 is insignificant and 1 very significant.

Different ranges of quality:

Correct	Good	Very good
LGscore > 1.5	LGscore > 3	LGscore > 5
MaxSub > 0.1	MaxSub > 0.5	MaxSub > 0.8

4.3 Results and discussion

Combining SWISS-MODEL and ProQ evaluations, we were able to rank all four models, thus giving a prediction about the best structure to optimize the subsequent laboratory test phases. These are the four models along with their average evaluation (green=linker, red=heavy chain, blue=light chain):

- **Model 1**:

>H-scFv (VL-VH)
MSPQRDRINAFYKDNPHPKGSRIVINREHLMIDRPYVLLAVLFVMFLSLI
GLLAIAGIRLHRAAIYTAEIHKSLSTNLDVTNSIEHQVKDVLTPLFKIIG
DEVGLRTPQRFTDLVKFISDKIKFLNPDREYDFRDLTWCINPPERIKLDY
DQYCADVAAEELMNALVNSTLLETRTTNQFLAVSKGNCSGPTTIRGQFSN
MSLSLLDLYLGRGYNVSSIVTMTSQGMYGGTYLVEKPNLSSKRSELSQLS
MYRVFEVGVIRNPGLGAPVFHMTNYLEQPVSNDLSNCMVALGELKLAALC
HGEDSITIPYQGSGKGVSFQLVKLGVWKSPTDMQSWVPLSTDDPVIDRLY
LSSHRGVIADNQAKWAVPTTRTDDKLRMETCFQQACKGKIQALCENPEWA
PLKDNRIPSYGVLSVDLSLTVELKIKIASGFGPLITHGSGMDLYKSNHNN
VYWLTIPPMKNLALGVINTLEWIPRFKVSPYLFNVPIKEAGEDCHAPTYL
PAEVDGDVKLSSNLVILPGQDLQYVLATYDTSRVEHAVVYYVYSPSRSFS
YFYPFRLPIKGVPIELQVECFTWDQKLWCRHFCVLADSESGGHITHSGMV
GMGVSCTVTREDGTNRRGGGGSGGGGSGGGGSGGGGSDILLTQSPVILSV
SPGERVSFSCRASQSIGTNIHWYQQRTNGSPRLLIKYASESISGIPSRFS
GSGSGTDFTLSINSVESEDIADYYCQQNNNWPTTFGAGTKLELKRGGSSR
SSSSGGGGSGGGGQVQLKQSGPGLVQPSQSLSITCTVSGFSLTNYGVHWV
RQSPGKGLEWLGVIWSGGNTDYNTPFTSRLSINKDNSKSQVFFKMNSLQS
NDTAIYYCARALTYYDYEFAYWGQGTLVTVSA*

ProQ - Results model 1
LGscore: 3,026
MaxSub: 0,467

- **Model 2**:

>H-scFv (VH-VL)
MSPQRDRINAFYKDNPHPKGSRIVINREHLMIDRPYVLLAVLFVMFLSLI
GLLAIAGIRLHRAAIYTAEIHKSLSTNLDVTNSIEHQVKDVLTPLFKIIG
DEVGLRTPQRFTDLVKFISDKIKFLNPDREYDFRDLTWCINPPERIKLDY
DQYCADVAAEELMNALVNSTLLETRTTNQFLAVSKGNCSGPTTIRGQFSN
MSLSLLDLYLGRGYNVSSIVTMTSQGMYGGTYLVEKPNLSSKRSELSQLS
MYRVFEVGVIRNPGLGAPVFHMTNYLEQPVSNDLSNCMVALGELKLAALC
HGEDSITIPYQGSGKGVSFQLVKLGVWKSPTDMQSWVPLSTDDPVIDRLY
LSSHRGVIADNQAKWAVPTTRTDDKLRMETCFQQACKGKIQALCENPEWA
PLKDNRIPSYGVLSVDLSLTVELKIKIASGFGPLITHGSGMDLYKSNHNN
VYWLTIPPMKNLALGVINTLEWIPRFKVSPYLFNVPIKEAGEDCHAPTYL
PAEVDGDVKLSSNLVILPGQDLQYVLATYDTSRVEHAVVYYVYSPSRSFS
YFYPFRLPIKGVPIELQVECFTWDQKLWCRHFCVLADSESGGHITHSGMV
GMGVSCTVTREDGTNRRGGGGSGGGGSGGGGSGGGGSQVQLKQSGPGLVQ
PSQSLSITCTVSGFSLTNYGVHWVRQSPGKGLEWLGVIWSGGNTDYNTPF
TSRLSINKDNSKSQVFFKMNSLQSNDTAIYYCARALTYYDYEFAYWGQGT
LVTVSAGGSSRSSSSGGGGSGGGGDILLTQSPVILSVSPGERVSFSCRAS
QSIGTNIHWYQQRTNGSPRLLIKYASESISGIPSRFSGSGSGTDFTLSIN
SVESEDIADYYCQQNNNWPTTFGAGTKLELKR*

ProQ - Results model 2
LGscore : 5.316
MaxSub : 0.488

- **Model 3**:

>H-scFab (VL-VH)
MSPQRDRINAFYKDNPHPKGSRIVINREHLMIDRPYVLLAVLFVMFLSLI
GLLAIAGIRLHRAAIYTAEIHKSLSTNLDVTNSIEHQVKDVLTPLFKIIG
DEVGLRTPQRFTDLVKFISDKIKFLNPDREYDFRDLTWCINPPERIKLDY
DQYCADVAAEELMNALVNSTLLETRTTNQFLAVSKGNCSGPTTIRGQFSN
MSLSLLDLYLGRGYNVSSIVTMTSQGMYGGTYLVEKPNLSSKRSELSQLS
MYRVFEVGVIRNPGLGAPVFHMTNYLEQPVSNDLSNCMVALGELKLAALC
HGEDSITIPYQGSGKGVSFQLVKLGVWKSPTDMQSWVPLSTDDPVIDRLY
LSSHRGVIADNQAKWAVPTTRTDDKLRMETCFQQACKGKIQALCENPEWA
PLKDNRIPSYGVLSVDLSLTVELKIKIASGFGPLITHGSGMDLYKSNHNN
VYWLTIPPMKNLALGVINTLEWIPRFKVSPYLFNVPIKEAGEDCHAPTYL
PAEVDGDVKLSSNLVILPGQDLQYVLATYDTSRVEHAVVYYVYSPSRSFS
YFYPFRLPIKGVPIELQVECFTWDQKLWCRHFCVLADSESGGHITHSGMV
GMGVSCTVTREDGTNRRGGGGSGGGGSGGGGSGGGGSDILLTQSPVILSV
SPGERVSFSCRASQSIGTNIHWYQQRTNGSPRLLIKYASESISGIPSRFS
GSGSGTDFTLSINSVESEDIADYYCQQNNNWPTTFGAGTKLELKRTVAAP
SVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESV
TEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGE
CGGSSGSGSGSTGTSSSGTGTSAGTTGTSASTSGSGSGGGGSGGGGSAG
GQVQLKQSGPGLVQPSQSLSITCTVSGFSLTNYGVHWVRQSPGKGLEWLG
VIWSGGNTDYNTPFTSRLSINKDNSKSQVFFKMNSLQSNDTAIYYCARAL

TYYDYEFAYWGQGTLVTVSAASTKGPSVFPLAPSSKSTSGGTAALGCLVK
DYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQT
YICNVNHKPSNTKVDKKVE*

ProQ - Results model 3 (SwissModel QMEAN -4,4)

Lgscore: 2.332
MaxSub: 0.118

- **Model 4**:

>H-scFab (VH-VL)
MSPQRDRINAFYKDNPHPKGSRIVINREHLMIDRPYVLLAVLFVMFLSLI
GLLAIAGIRLHRAAIYTAEIHKSLSTNLDVTNSIEHQVKDVLTPLFKIIG
DEVGLRTPQRFTDLVKFISDKIKFLNPDREYDFRDLTWCINPPERIKLDY
DQYCADVAAEELMNALVNSTLLETRTTNQFLAVSKGNCSGPTTIRGQFSN
MSLSLLDLYLGRGYNVSSIVTMTSQGMYGGTYLVEKPNLSSKRSELSQLS
MYRVFEVGVIRNPGLGAPVFHMTNYLEQPVSNDLSNCMVALGELKLAALC
HGEDSITIPYQGSGKGVSFQLVKLGVWKSPTDMQSWVPLSTDDPVIDRLY
LSSHRGVIADNQAKWAVPTTRTDDKLRMETCFQQACKGKIQALCENPEWA
PLKDNRIPSYGVLSVDLSLTVELKIKIASGFGPLITHGSGMDLYKSNHNN
VYWLTIPPMKNLALGVINTLEWIPRFKVSPYLFNVPIKEAGEDCHAPTYL
PAEVDGDVKLSSNLVILPGQDLQYVLATYDTSRVEHAVVYYVYSPSRSFS
YFYPFRLPIKGVPIELQVECFTWDQKLWCRHFCVLADSESGGHITHSGMV
GMGVSCTVTREDGTNRRGGGGSGGGGSGGGGSGGGGSQVQLKQSGPGLVQ
PSQSLSITCTVSGFSLTNYGVHWVRQSPGKGLEWLGVIWSGGNTDYNTPF
TSRLSINKDNSKSQVFFKMNSLQSNDTAIYYCARALTYYDYEFAYWGQGT
LVTVSAASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSG
ALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKV
DKKVEGGSSGSGSGSTGTSSSGTGTSAGTTGTSASTSGSGSGGGGGSGGG
GSAGGDILLTQSPVILSVSPGERVSFSCRASQSIGTNIHWYQQRTNGSPR
LLIKYASESISGIPSRFSGSGSGTDFTLSINSVESEDIADYYCQQNNNWP
TTFGAGTKLELKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAK
VQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACE
VTHQGLSSPVTKSFNRGEC*

ProQ - Results model 4
LGscore : 1.037
MaxSub : -0.256

For what concerns the 3D models structure, the most representative are shown as follows:

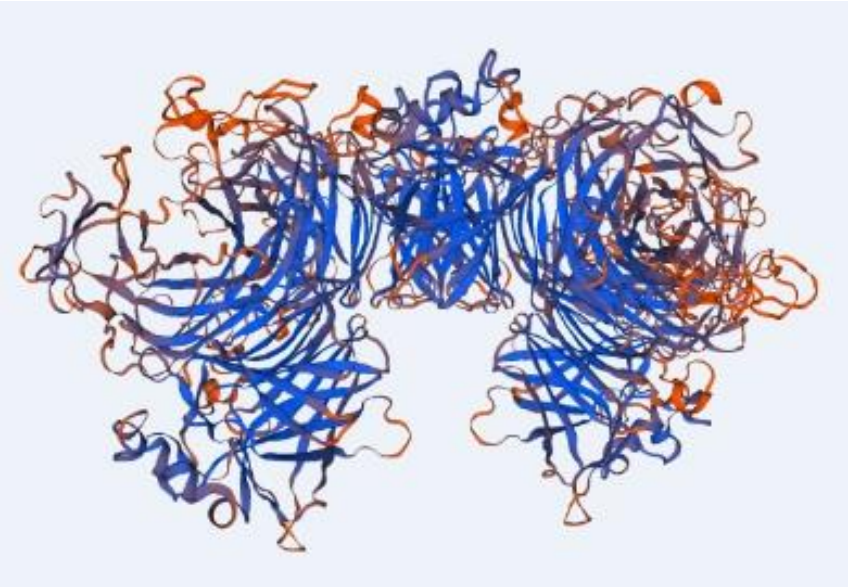

Figure 10 - Model 1 H-scFv (VL-VH) 3D structure

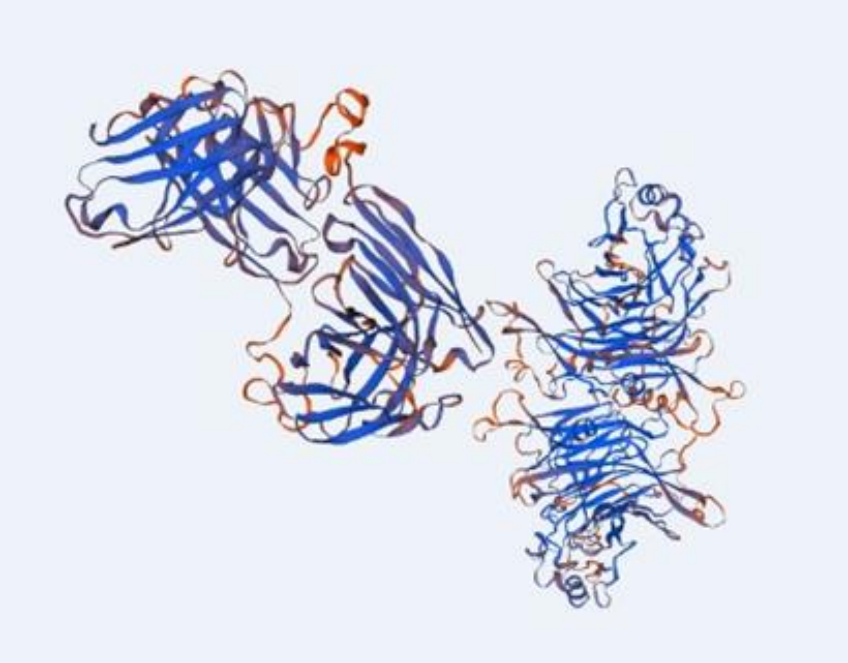

Figure 11 - Model 2 H-scFv (VH-VL) 3D structure

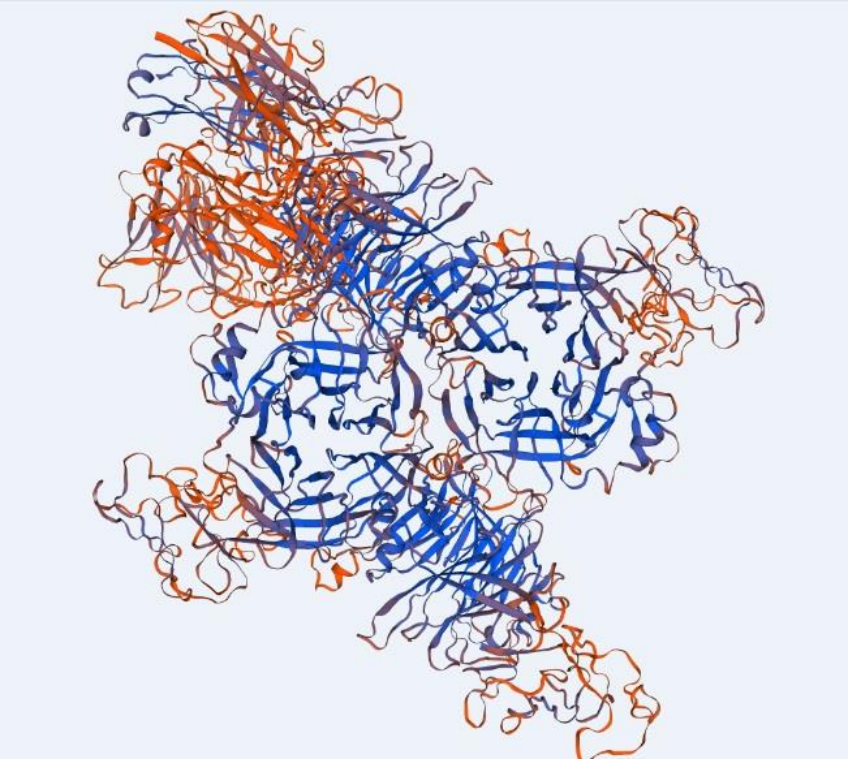

Figure 12 - Model 3 H-scFab (VL-VH) 3D structure

Figure 13 - Model 4 H-scFab (VH-VL) 3D structure

Transfection experiments are still ongoing in Etna Biotech laboratories, but early tests showed the rescue of the oncolytic virus related to construct 4 (H protein fused with scFabVH-VL): using the plasmid vector with the recombinant gene inserted, the permissive cells have been transfected to the measles infection thus obtaining the recombinant alive virus. The next steps will be the propagation of the virus on cancer and normal cells, the viral titration and the verification of the viral genome sequence after extraction of RNA from infected cells.

5 LONG AND SMALL RNA ANALYSIS

5.1 RNA-sequencing background

The transcriptome is the complete set of all RNA molecules in a cell, a population of cells or an organism. Not all RNAs are translated into proteins: the transcriptome has a high degree of complexity and contains multiple types of coding and noncoding RNA species. Even if transcriptomics is most commonly applied to the messenger RNAs (mRNAs), which represent the coding transcripts, transcriptomics also provides important data regarding noncoding RNAs, including rRNA, tRNA, lncRNA, siRNA, and others [48]. Transcriptome analysis represents the study of the transcriptome of the complete set of RNA transcripts produced by the genome for a specific developmental stage or physiological condition, through the usage of high-throughput methods. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding in detail the developmental processes and related diseases. The main purposes of transcriptomics are:

- catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs;
- determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications;
- quantify the changing expression levels of each transcript during development and under different conditions [49].

Several technologies have been developed to deduce and quantify the transcriptome, including hybridization-based or sequence-based approaches. Specialized microarrays have also been designed. Genomic tiling microarrays that represent the genome at high density have been constructed, allowing the mapping of transcribed regions to a very high resolution. Hybridization-based approaches are high throughput and relatively inexpensive, except for high-resolution tiling arrays that interrogate large genomes. However, these methods have several limitations: i) reliance upon existing knowledge about genome sequence; ii) high background levels owing to cross-hybridization; iii) a limited dynamic range of detection owing to both background and saturation of signals. Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods.

Recently, the development of high-throughput next-generation sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis through the sequencing of complementary DNA (cDNA). This method is RNA-Seq (RNA sequencing), and it has clear advantages over existing approaches [49]. RNA-Seq uses recently developed deep-sequencing technologies. In general, a population of RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput way to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). The reads are typically 30–400 bp, depending on the DNA-sequencing technology used.

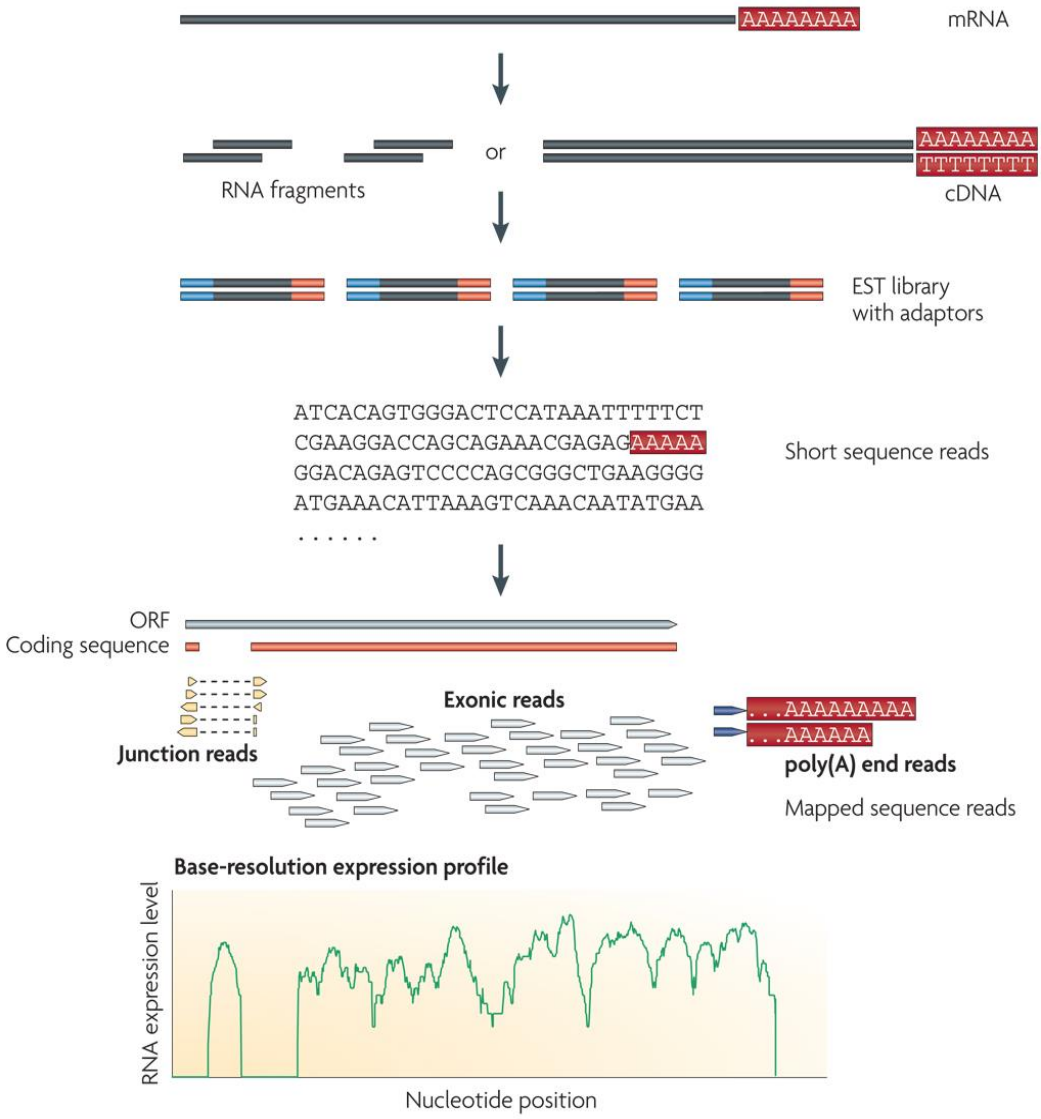

Figure 14 - A typical RNA-Seq experiment: long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene [49].

RNA-Seq offers several advantages over existing technologies:

- RNA-Seq is not limited to detect transcripts that correspond to existing genomic sequence;
- RNA-Seq can reveal the precise location of transcription boundaries, to a single-base resolution;
- 30-bp short reads from RNA-Seq give information about how two exons are connected, whereas longer reads or pair-end short reads should reveal connectivity between multiple exons;
- RNA-Seq can reveal sequence variations (for example, single nucleotide polymorphisms or SNPs) in the transcribed regions;
- relative to DNA microarrays, RNA-Seq has very low background signal because DNA sequences can be unambiguously mapped to unique regions of the genome;
- RNA-Seq does not have an upper limit for quantification;
- RNA-Seq is highly accurate for quantifying expression levels;
- RNA-Seq requires less RNA sample.

The first step in transcriptome sequencing is the isolation of RNA from a biological sample. To ensure a successful RNA-Seq experiment, the RNA should be of sufficient quality to produce a library for sequencing. The quality of RNA is typically measured using an Agilent Bioanalyzer, which produces an RNA Integrity Number (RIN) between 1 and 10 with 10 being the highest quality samples showing the least degradation. The RIN estimates sample integrity using gel electrophoresis and analysis of the ratios of 28S to 18S ribosomal bands. Following RNA isolation, the next step in transcriptome sequencing is the creation of an RNA-Seq library, which can vary by the selection of RNA species and between NGS platforms. The construction of sequencing libraries principally involves isolating the desired RNA molecules, reverse-transcribing the RNA to cDNA, fragmenting or amplifying randomly primed cDNA molecules, and ligating sequencing adaptors. Before constructing RNA-Seq libraries, it is important to choose an appropriate library preparation protocol that will enrich or deplete a “total” RNA sample for particular RNA species. The total RNA pool includes ribosomal RNA (rRNA), precursor messenger RNA (pre-mRNA), mRNA, and various classes of noncoding RNA (ncRNA). In most cell types, the majority of RNA molecules are rRNA, typically accounting for over 95% of the total cellular RNA. If the rRNA transcripts are not removed before library construction, they will consume the bulk of the sequencing reads, reducing the overall depth of sequence coverage and thus limiting the detection of other less-abundant RNAs. When designing an RNA-Seq experiment, the selection of a sequencing platform is important and dependent on the experimental goals. Currently, several NGS platforms are commercially available

and other platforms are under active technological development. The majority of high-throughput sequencing platforms use a sequencing-by-synthesis method to sequence tens of millions of sequence clusters in parallel. The NGS platforms can often be categorized as either ensemble-based (i.e. sequencing many identical copies of a DNA molecule) or single-molecule-based (i.e. sequencing a single DNA molecule). The conventional pipeline for RNA-Seq data includes generating FASTQ-format files contains reads sequenced from an NGS platform, aligning these reads to an annotated reference genome, and quantifying expression of genes [50].

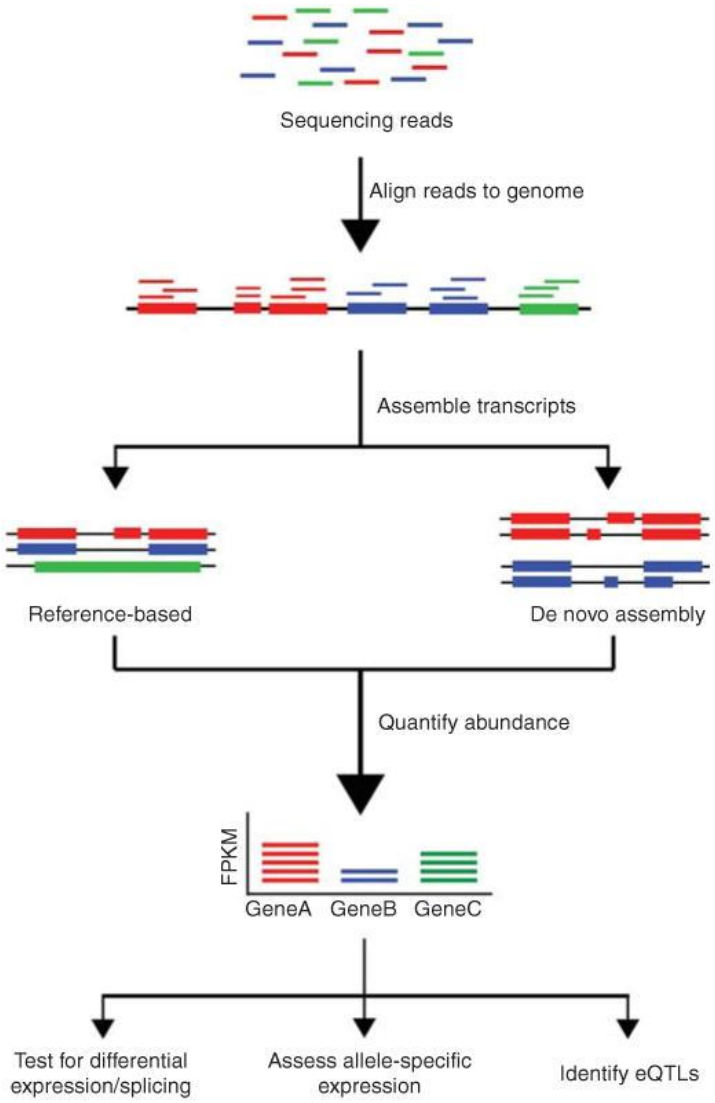

Figure 14- Overview of RNA-Seq data analysis. Following typical RNA-Seq experiments, reads are first aligned to a reference genome. Second, the reads may be assembled into transcripts using reference transcript annotations or de novo assembly approaches. Next, the expression level of each gene is estimated by counting the number of reads that align to each exon or full-length transcript. Downstream analyses with RNA-Seq data include testing for differential expression between samples, detecting allele-specific expression, and identifying expression quantitative trait loci (eQTLs) [50].

5.2 Long and small RNA background

Transcription mechanism allows human genome not only to code for several types of proteins but also to generate thousands of non-coding RNA molecules (ncRNA). As revealed by ENCODE consortium, around 70% of the genome is transcribed for RNA molecules that have no protein coding capacity. Instead of simple transcriptional noise, they directly function as structural, catalytic and regulatory RNAs.

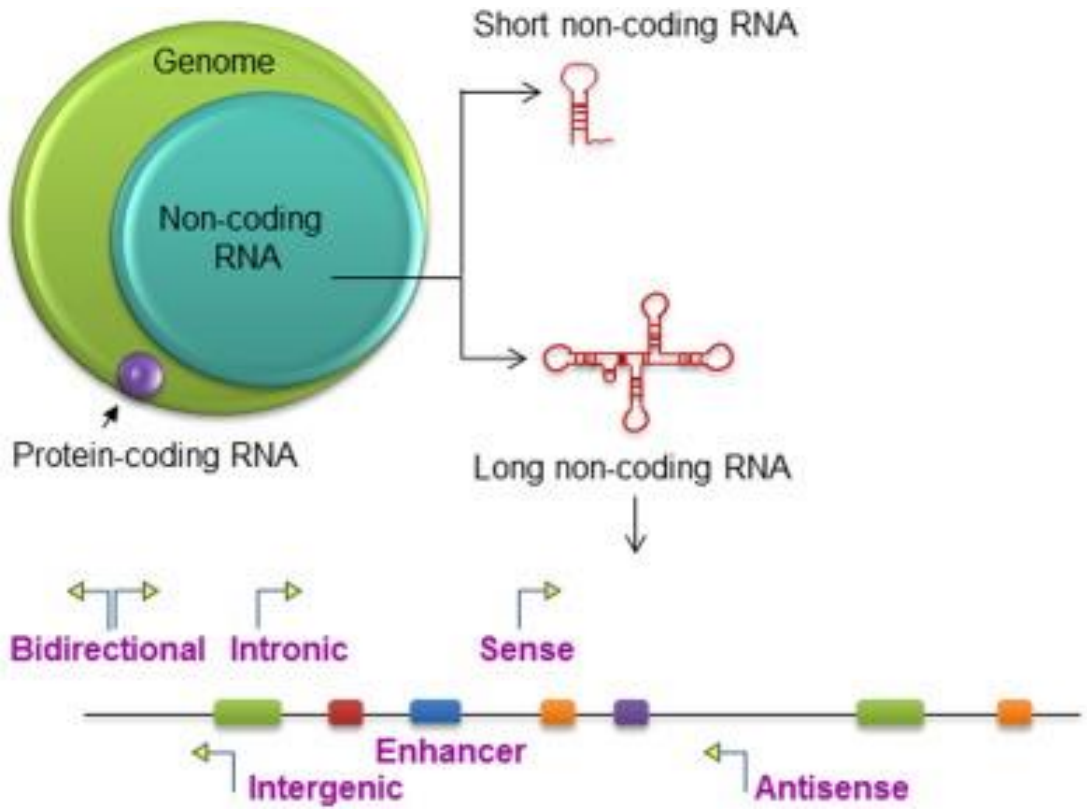

Figure 15- Non-coding RNAs comprise a much larger portion of the human genome than protein-coding RNA, which comprise < 3% of the genome. Non-coding RNAs are arbitrarily classified into short and based on transcript size. LncRNA can be designated as Intergenic, Intronic, Enhancer, Sense, Antisense, or Bidirectional based on their genomic location relative to that of nearby protein-coding genes [51].

NcRNAs smaller than 200nt are categorized into small non-coding RNA (sncRNA) which includes small interfering RNA (siRNA), microRNA (miRNA) and piwi-interacting RNA (piRNA). The rest of the ncRNA larger than 200nt are categorized into long non-coding RNA (lncRNA) which includes long intergenic non-coding RNA (lincRNA), natural antisense transcript (NAT), transcribed ultraconserved region (T-UCR) and non-coding pseudogene [52].

- Long RNA

Long RNAs are molecules greater than 200 nucleotides, and include long intergenic non-coding RNAs (lincRNAs), natural antisense transcripts (NATs), transcribed ultra-conserved regions (T-UCRs), long enhancer ncRNAs, non-coding repeat sequences, and pseudogenes. Long RNAs are involved in a variety of functions such as gene transcription, epigenomic regulation, translation of protein-coding genes, RNA turnover, chromatin organization, and genome defense. Depending on genomic location, they can be divided into several categories: intronic lncRNAs (located within an intron of a coding gene); intergenic lncRNAs (lincRNAs, present between two protein-coding genes); bidirectional lncRNAs (expressed within 1 kb of a coding transcript of the opposite strand); enhancer lncRNAs (e-lncRNAs, present in the enhancer regions close to the promoter); sense or antisense lncRNAs (those that overlap with one or more introns and exons of a different transcript on the same or opposite strand, respectively) [51]. While the majority of lncRNA are present within the cytoplasm, some lncRNAs are primarily confined within the nucleus. Trans-acting nuclear lncRNAs can act in a tissue-specific manner in conjunction with chromatin modifiers such as histone-modifying complexes and DNA methyltransferases, to epigenetically regulate or to modulate transcription. However, the mechanism by which these lncRNAs can recognize specific genomic loci is not well understood. The ability of lncRNA to alter cellular physiology by altering gene expression raises the possibility that deregulated lncRNA expression may contribute to disease pathophysiology. Altered expression of lncRNA has been reported in many cancers and in other disease settings.

- Small RNA

Small RNAs are defined as short (~ 18 to 30 nucleotides [nt]), non-coding RNA molecules that can inhibit the expression of target genes via post-transcriptional gene silencing (PTGS) and chromatin-dependent gene silencing (CDGS), in both cytoplasm and nucleus. Small RNAs can be classified into three main categories: microRNAs (miRNAs), siRNAs and Piwi-interacting RNAs (piRNAs). miRNAs are ssRNA molecules with a length of approximately 21 or 22 nt. miRNAs are initially transcribed in the nucleus to form large pri-miRNA transcripts, which are subsequently processed in the nucleus into approximately 70-nt pre-miRNAs. These pre-miRNAs are then transported into the cytoplasm to become mature miRNAs. Today more than 850 mature human miRNA sequences have been identified, many of which are highly conserved among many species. siRNAs can be produced from RNA transcribed in the nucleus (endogenous siRNAs); they can be virally derived or can be introduced experimentally as chemically synthesized dsRNAs (exogenous siRNAs). piRNAs are the most recently discovered class of siRNAs and, as their name suggests, are germ-cell-specific small RNAs that bind to the Piwi clade of Argonaute proteins. Notably, small RNAs are involved in

several processes, such as the regulation of all major cellular functions, including cell differentiation, growth/proliferation, migration, apoptosis/death, metabolism and defense. Given these diverse roles, small RNAs could act as regulators in development, physiology and disease: increasing evidence indicates that small RNAs are involved in the pathogenesis of different diseases including cancer, cardiovascular disease, stroke, neurodegenerative disease, diabetes, liver disease, kidney disease and infectious disease [53].

5.3 Material and methods

For this study, 4 Mus musculus species samples have been sequenced both for small and long analysis.

Samples for long RNA analysis are the following:

1 - astrocytes from mouse striatum in basal conditions (STR-)

2 - astrocytes from mouse striatum treated with Ccl3 (STR+)

1b - astrocytes from mouse midbrain in basal conditions (MID-)

2b - astrocytes from mouse midbrain treated with Ccl3 (MID+)

And the following one represent the samples for small RNA analysis:

1b - astrocytes from mouse midbrain in basal conditions (MID-)

2b - astrocytes from mouse midbrain treated with Ccl3 (MID+)

3b – exosomes from 1b (MID_EX-)

4b – exosomes from 2b (MID_EX+)

In order to analyze both long and small RNAs, we used two different approaches, specific for each task. Both of them include the following steps:

1. Alignment to reference genome and indexes creation
2. Exploratory analysis and visualization
3. Differential expressione analysis
4. GSEA analysis
5. Pathway analysis and visualization

5.4 Results and discussion

- Long RNAs analysis

The first step in long RNAs analysis pipeline is the alignment, which includes the mapping to the reference genome and subsequently the indexes creation.

To align our RNA-seq datasets, the Spliced Transcripts Alignment to a Reference (STAR) [54] was used. STAR software was designed to align high-throughput long and short RNA-seq data directly to the reference genome. STAR algorithm works as follows:

1. Seed serching

STAR is based on the concept of Maxible Mappable Prefix (*MMP*): given a read sequence *R*, read location *i* and a reference genome sequence *G*, the *MMP(R,i,G)* is defined as the longest substring (R_i, R_{i+1}, … , $R_{i+MML-1}$) that matches exactly one or more substrings of *G*. Basically, the algorithm finds the *MMP* starting from the first base of the read, and then the search is repeated for the unmapped portion of the read, in both forward and reverse directions throughout the read sequence.

2. Clustering, stitching and scoring

The algorithm builds alignments of the entire read sequence by stitching together all the seeds that were aligned to the genome in the seed searching phase. Basically, all the seeds in close proximity to a selected seed are clustered together assuming a local linear transcription model.

Bowtie [55] package was used for the alignment of sequencing reads to reference mouse genome, and for indexes creation. The package contains tools which enable ultrafast and memory-efficient alignment of large sets of sequencing reads to a reference using the index as a guide. Bowtie protocol uses the *bowtie-build* tool to take a collection of FASTA files for a reference genome and generate a collection of index files. Index files can then be used by Bowtie to align reads to the reference genome. The purpose of alignment is to identify origin point from the reads and then use that information, for example, to characterize differences between the subject and reference genome (SNPs), or to link the reads to annotations.

In order to summarize the main features of our datasets, an initial investigation called *Exploratory analysis* was performed by using summary statistics and graphical representations provided by DESeq2 package [56] for Bioconductor R language.

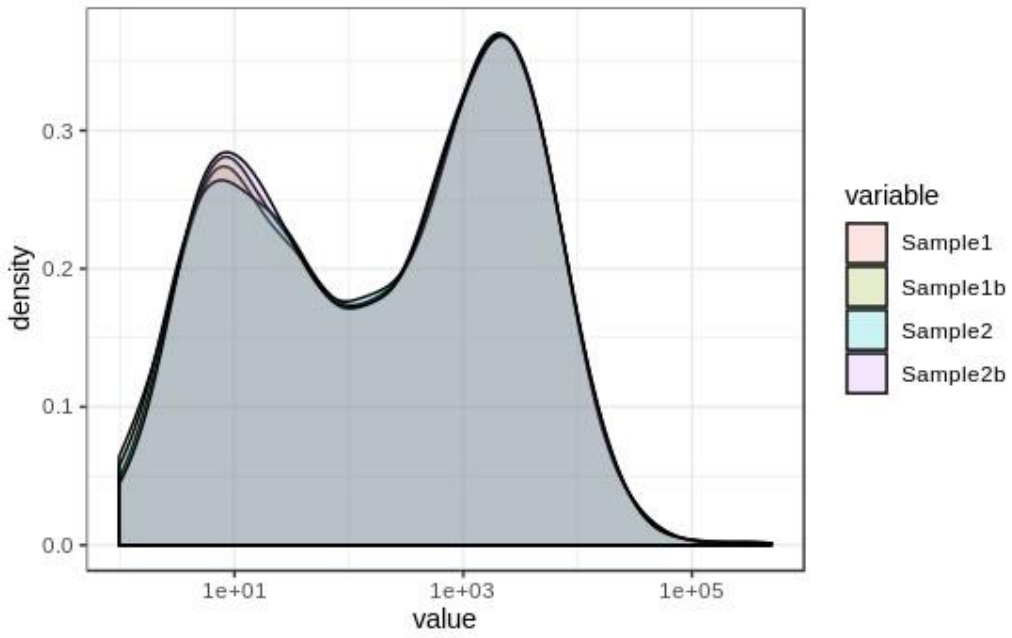

Figure 16 - Density plot of normalized counts. This figure shows how the trend of the distribution density in relation to the average of the normalized counts of the four samples is homogeneous

In Fig. 17, the distribution density of the variables is related to the average of the normalized counts. The variables are the following ones:

1 - astrocytes from mouse striatum in basal conditions

2 - astrocytes from mouse striatum treated with Ccl3

1b - astrocytes from mouse midbrain in basal conditions

2b - astrocytes from mouse midbrain treated with Ccl3.

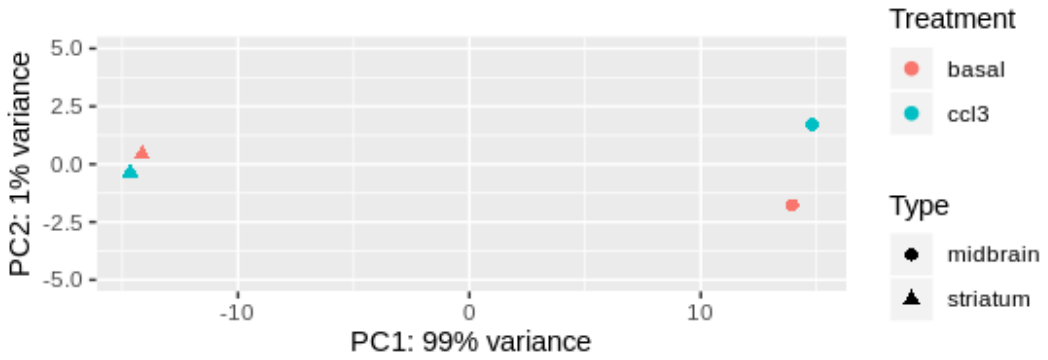

Figure 17 - Principal Component Analysis (PCA). This picture shows how data tend to cluster depending on the tissue, grouping on each side of the 2D plane: striatum samples are on left side and midbrain on the right.

Fig. 18 shows *Principal Component Analysis* performed on our dataset. *PCA* tries to make data structure clearer by reducing the dimensionality of multivariate data, calculating the linear combination of the variable accounting for as much of the total variation of the data as possible [57]. In this method, the samples are projected onto the 2D plane such that they spread out in two directions depending on the major differences between them. Both axes represent the directions where data separate for the most part. The values of the samples in x-axis direction are written *PC1*. The values of the samples in y-axis direction are written *PC2*. The percentage of the total variance that is contained in each direction is shown in the axis label [58].

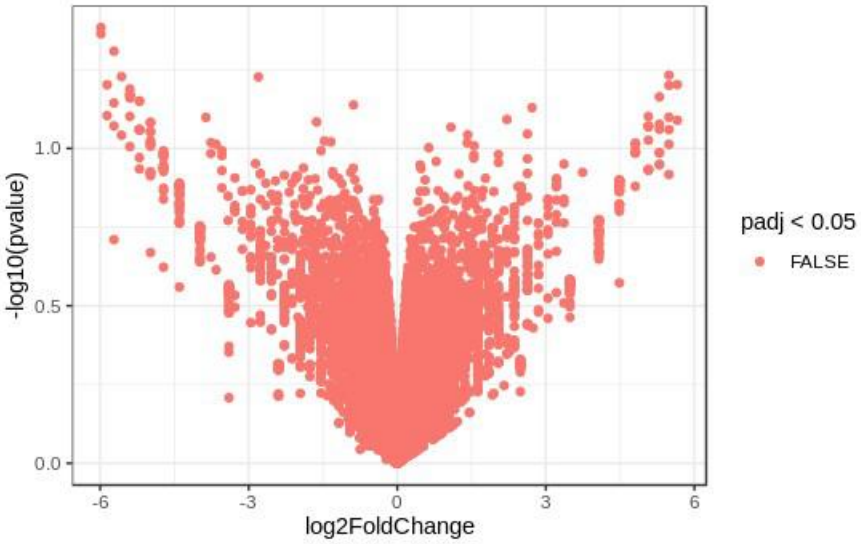

Figure 18 - Volcano Plot (condition: midbrain Ccl3 vs midbrain basal)

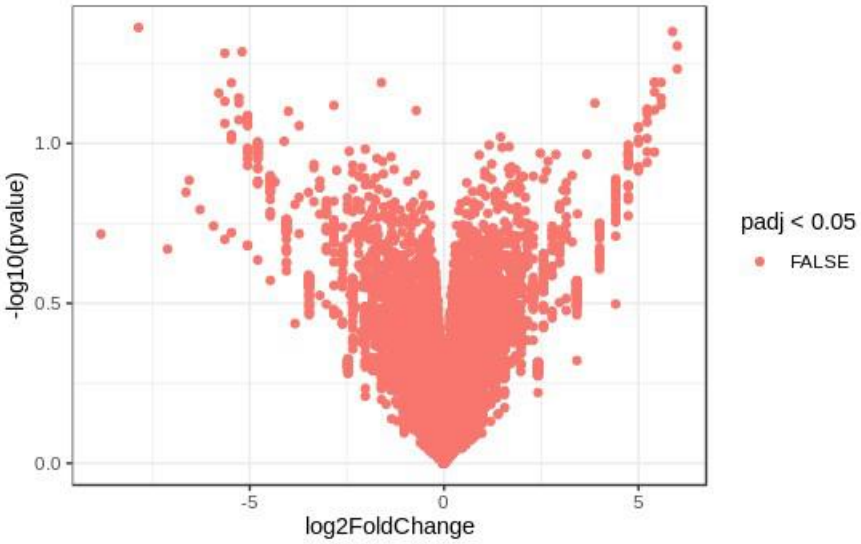

Figure 9 - Volcano Plot (condition: striatum Ccl3 vs striatum basal)

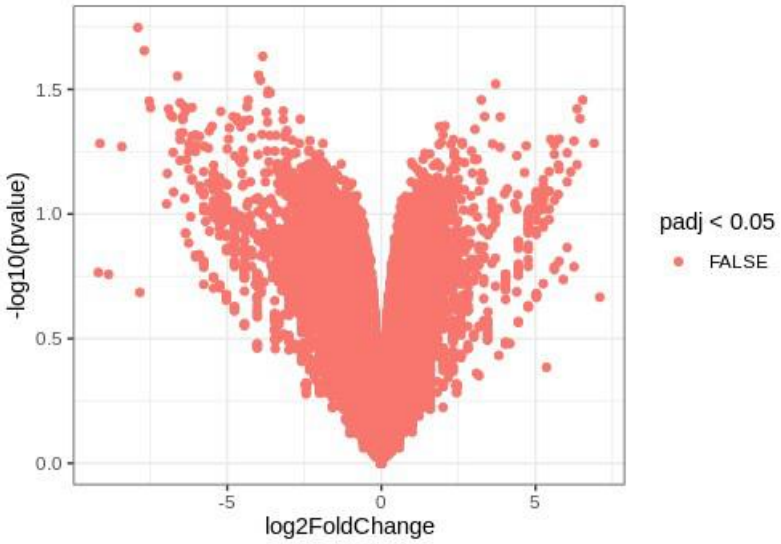

Figure 21 - Volcano Plot (condition: striatum basal vs midbrain basal)

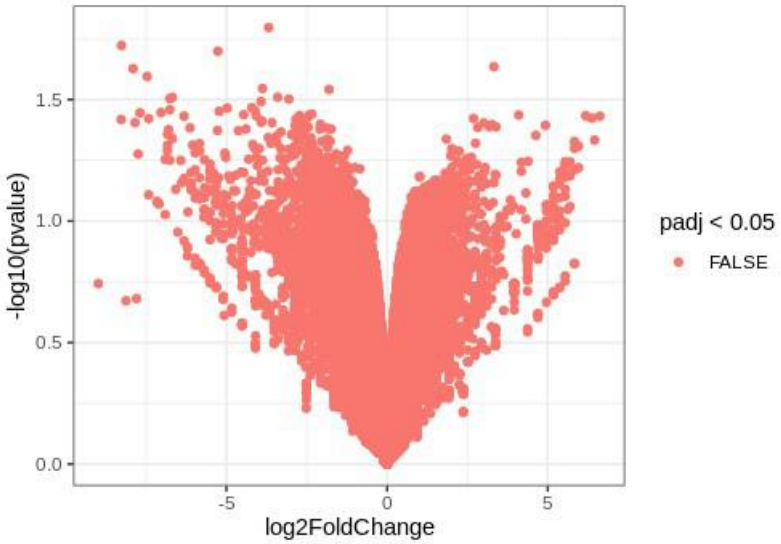

Figure 22 - Volcano Plot (condition: striatum Ccl3 vs midbrain Ccl3)

Figures 19, 20, 21 and 22 represent Volcano Plot: these plots are commonly used to display the results of RNA-seq analysis. It is a type of scatterplot that shows statistical significance (*p-value*) versus magnitude of change (*fold change*). It allows rapid identification of genes with high fold change values that are also statistically significant. In a Volcano plot, the most upregulated genes are towards the right, the most downregulated genes are towards the left, and the most statistically significant genes are towards the top.

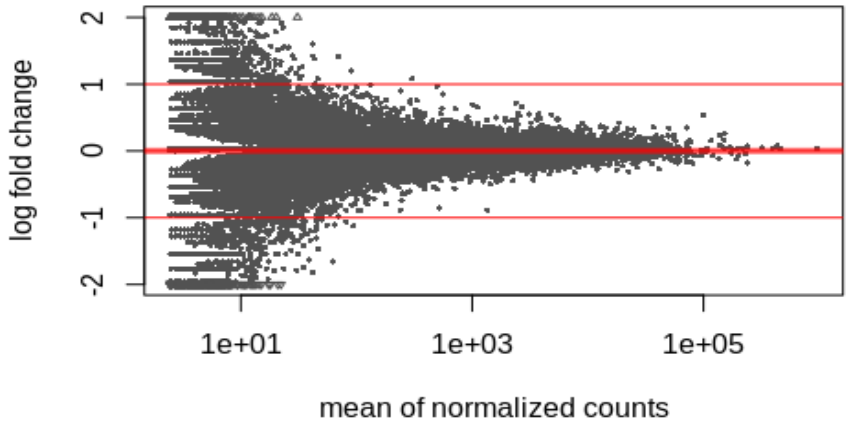

Figure 23 - MA Plot (condition: midbrain Ccl3 vs midbrain). MA plot highlights the distribution of the Cy5/Cy3 intensity ratio ('M') versus the average intensity ('A')—to identify possible errors introduced during the normalization procedure.

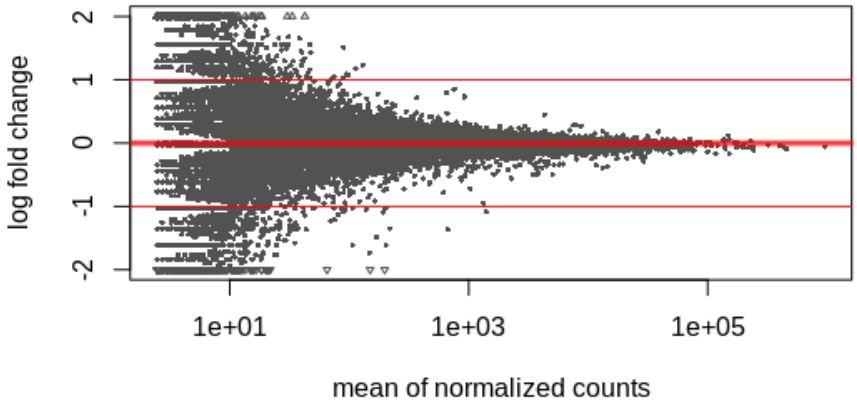

Figure 10 - MA Plot (condition: striatum Ccl3 vs striatum)

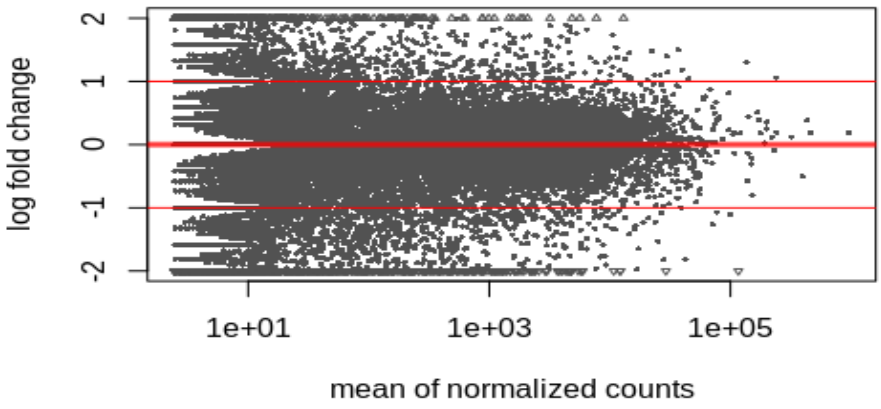

Figure 11 - MA Plot (condition: striatum basal vs midbrain)

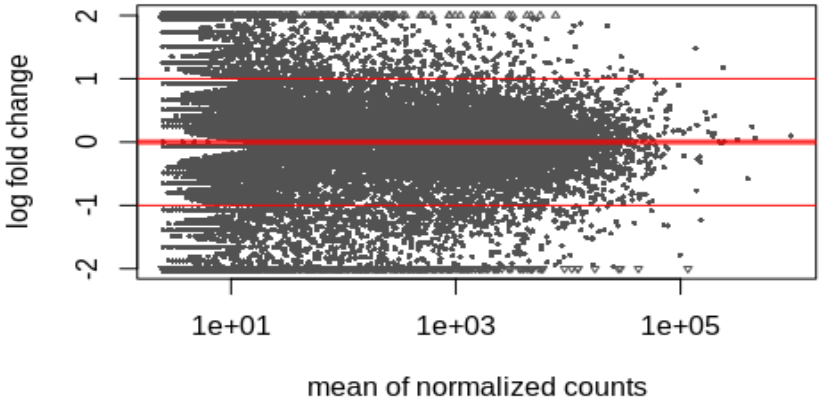

Figure 12 - MA Plot (condition: striatum Ccl3 vs midbrain)

Figures 23, 24, 25 and 26 represent MA Plot: this is a plot of log-intensity ratios (M-values) versus log-intensity averages (A-values). It shows the log2 fold changes attributable to a given variable over the mean of normalized counts for all the samples in the *DESeqDataSet*. Points will be colored in red if the adjusted *p* value is less than 0.1. Points which fall out of the window are plotted as open triangles pointing either up or down MA plot shows the $\log_2$ fold change for each gene against the average $\log_2$ counts per million, allowing for an overall visualization of the distribution of genes for each pairwise comparison. An MA plot is similar to a volcano plot in that it displays the $\log_2$ fold change against the $-\log_{10} P$ value [59].

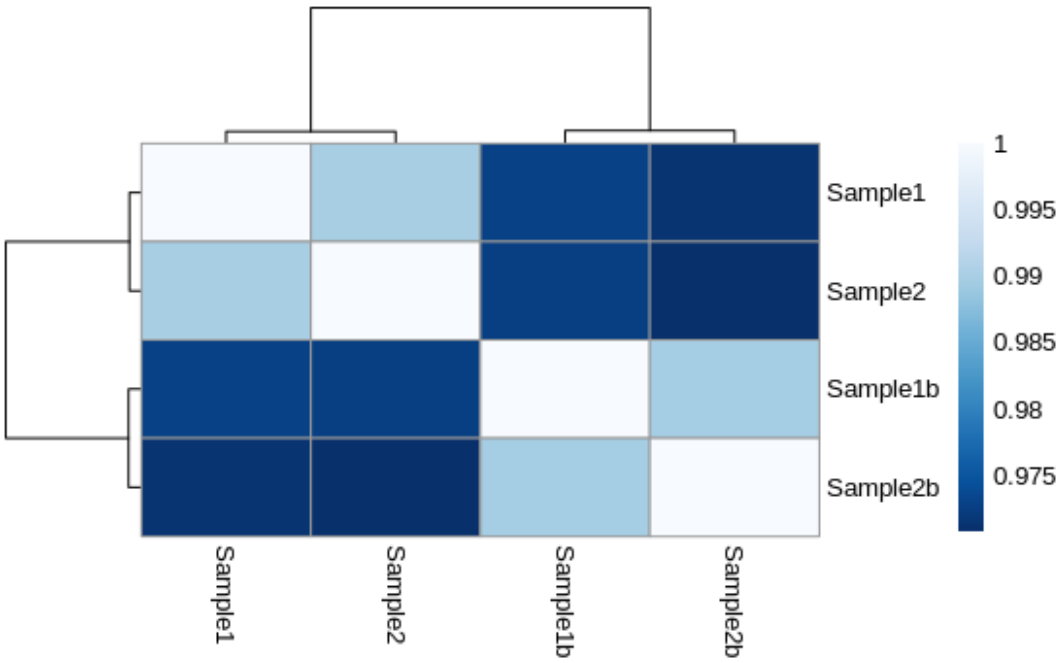

Figure 27 - Samples Correlation Heatmap

Fig.27 shows the samples correlation heatmap, where color bar represents Spearman's correlation coefficients. Spearman's correlation coefficient, (ρ, also signified by r_s) measures the strength and direction of association between two ranked variables: 1 means perfect correlation (identity) and represents the highest possible value. There are two methods to calculate Spearman's correlation depending on whether data do or do not have tied ranks. If there are no tied ranks the formula is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference in paired ranks and n = number of cases. The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where i = paired score.

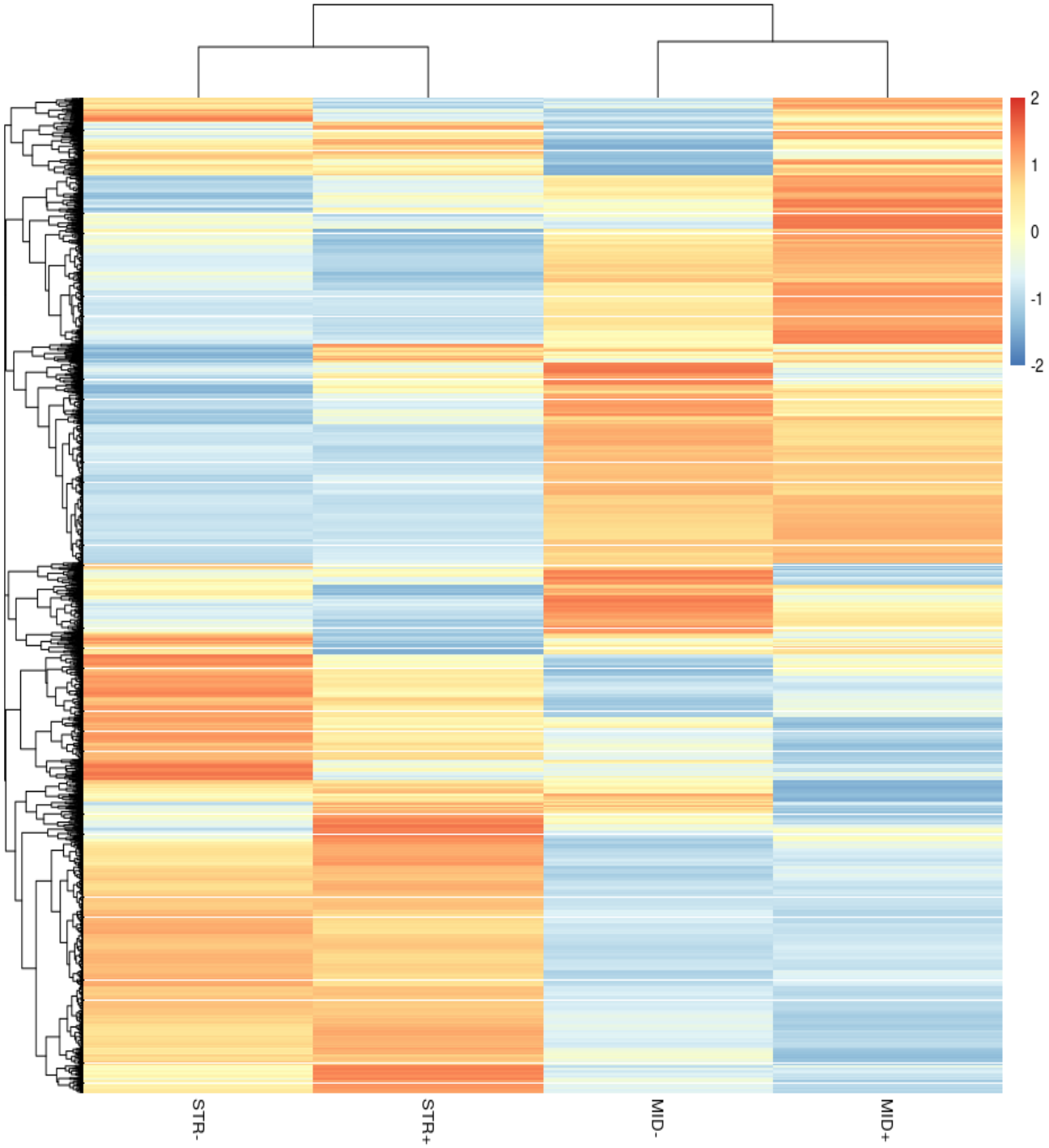

Figure 28 - Heatmap of z-normalized counts

Fig. 28 shows another heatmap with clustering of all genes having reads count > 10 across the four conditions. Notice that counts have been z-normalized. Normalization is used to scale the values of data so that they fall into a smaller range and to bring all the attributes on the same scale. In z-score normalization, values are normalized based on mean and standard deviation of the data.

Differential expression analysis has been conducted on normalized data using DESeq2 package [56] for Bioconductor R language. DESeq2 is an R package used to analyze count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression. The goal of differential expression analysis is the identification of all those genes whose expression differs under different conditions. This means taking the normalized read count data as input and performing statistical analysis on them to highlight quantitative changes in expression levels among experimental groups.

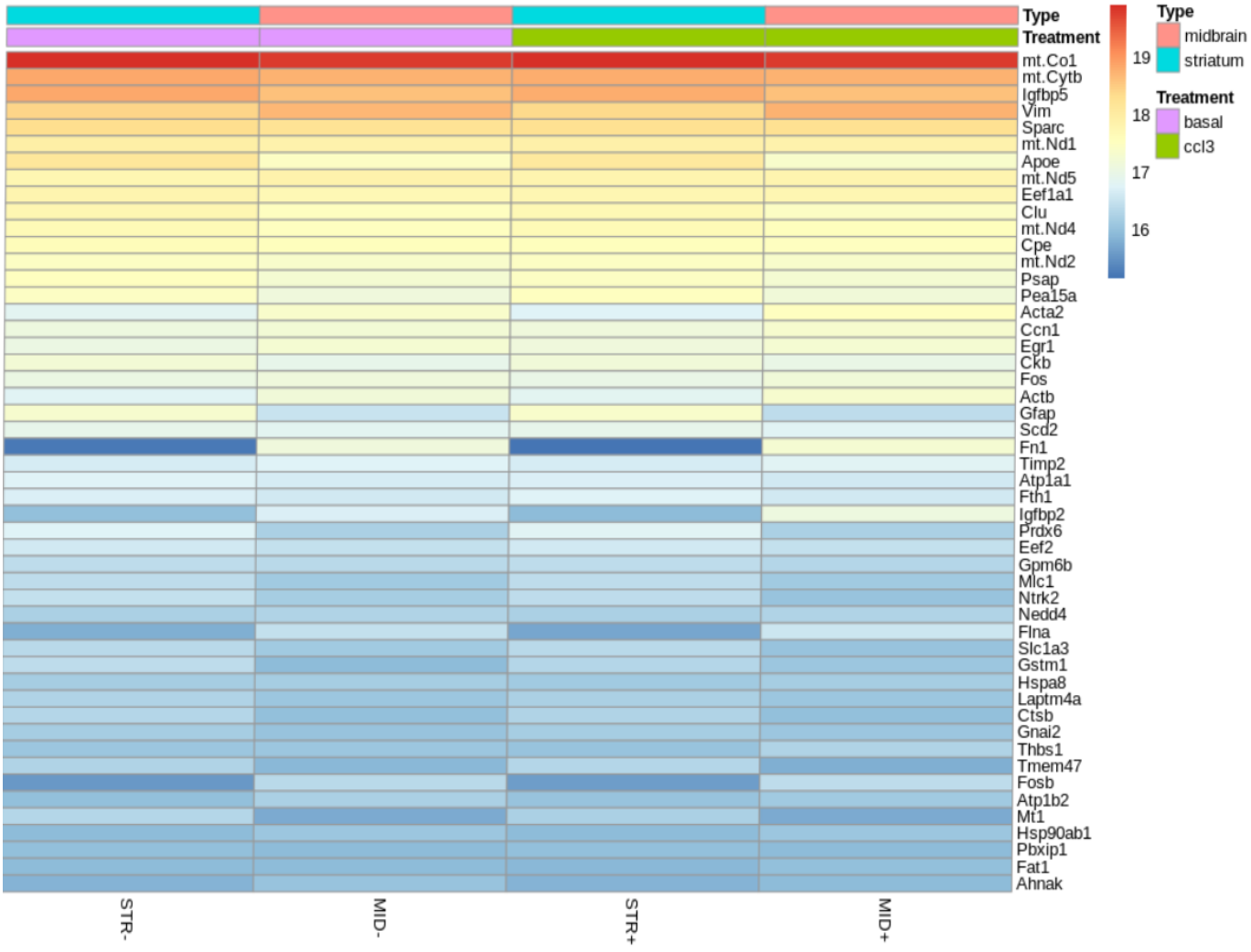

Figure 29 - Heatmap of counts

Fig. 29 shows the heatmap of the top 50 genes in terms of read counts across the four experimental conditions. Heatmaps are commonly used to visualize the expression of genes across the samples in a RNA-Seq dataset. In this case, we have the 50 most expressed genes on rows and samples on columns.

After exploratory and differential expression analysis, GSEA analysis was performed, followed by pathways visualization. GSEA means Gene Set Enrichment Analysis: it's a method that uses a statistical approach to identify groups of genes that are significantly over-represented compared to a large gene set of reference, which may show particular correlations with pathological phenotypes. The standard GSEA method includes the following steps:

1. Calculation of the enrichment score (ES) which represents the threshold above which the genes result over-represented at the beginning or end of the input list.

2. Estimation of the statistical significance of the ES.

3. Correction with multiple choice tests in the case of many genes analyzed at the same time.

In order to perform enrichment analysis, the following experimental conditions were considered:

RES1: MID+ vs MID

RES2: STR+ vs STR

RES3: STR- vs MID

RES4: STR+ vs MID+

Where:

STR- = astrocytes from mouse striatum in basal conditions

STR+ = astrocytes from mouse striatum treated with Ccl3

MID- = astrocytes from mouse midbrain in basal conditions

MID+ = astrocytes from mouse midbrain treated with Ccl3

For each condition, small subgroups of genes were extracted from the differentially expressed (DE) genes table, resulting from DESeq2 analysis, depending on the total number of reads in specific condition and fold change (FC) values, thus obtaining the following gene sets:

RES1:

res1_gene_name: 15823 genes (DE genes table)

genelist1: 15109 genes (all genes having total reads > 10 in both conditions)

de1: 55 genes (genes of genelist1 with log2FC>2)

de1bis: 361 (genes of genelist1 with log2FC>1)

RES2:

res2_gene_name: 15823 genes (DE genes table)

genelist2: 14796 genes (all genes having total reads > 10 in both conditions)

de2: 33 genes (genes of genelist2 with log2FC>2)

de2bis: 298 (genes of genelist2 with log2FC>1)

RES3:

res3_gene_name: 15823 genes (DE genes table)

genelist3: 14951 genes (all genes having total reads > 10 in both conditions)

de3: 598 genes (genes of genelist3 with log2FC>2)

de3bis: 1928 (genes of genelist3 with log2FC>1)

RES4:

res4_gene_name: 15823 genes (DE genes table)

genelist4: 15018 genes (all genes having total reads > 10 in both conditions)

de4: 727 genes (genes of genelist4 with log2FC>2)

de4bis: 2223 (genes of genelist4 with log2FC>1)

ClusterProfiler [60] package for R Bioconductor programming language was used to perform enrichment analysis. Gene lists obtained from differential expression analysis (this means from

DESeq2 package) represent the input vectors for clusterProfiler: each gene list, one for each experimental condition, should have these three features:

1. It must be a numeric vector: fold change or other type of numerical variable are required.
2. It must be a named vector: every number must be named by the corresponding gene ID.
3. It must be a sorted vector: number should be sorted in decreasing order.

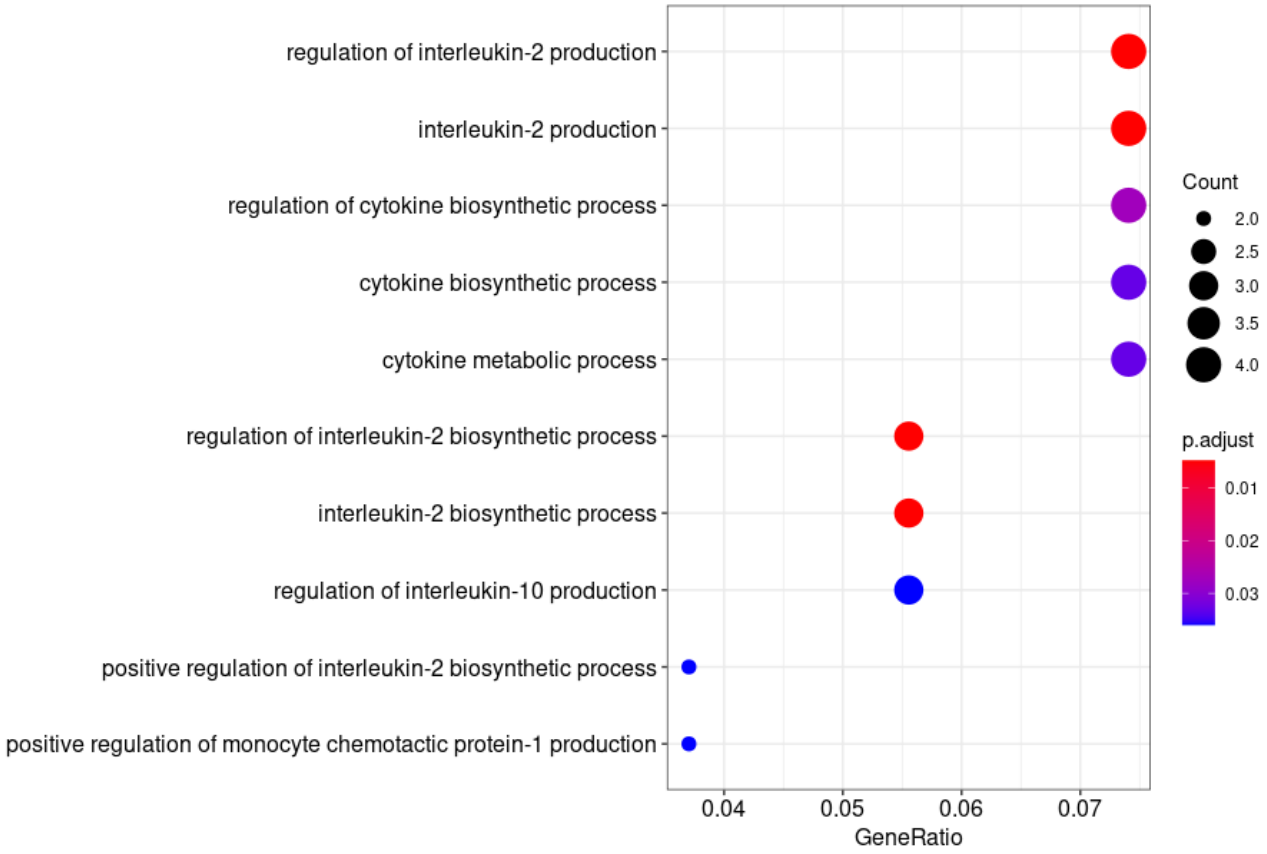

Figure 30 - Dotplot showing results of GO analysis (universe = GeneList1)

Figure 30 shows the results of Gene Ontologies (GO) analysis for RES1 condition (universe = genelist1, gene set = de1). Gene Ontology is a bioinformatic project that provides a collection of defined terms representing the properties of genes products. This collection is divided into three areas:

• cellular component: the parts of a cell or its extracellular environment;

• molecular function: the elementary activities of a gene product at the molecular level, such as binder or catalysis;

• biological process: the operations or complexes of molecular events with a beginning and an end defined, relevant to the functioning of integrated living units: cells, tissues, organs e organisms.

Dot plot is one of the most widely used methods to visualize enriched terms. Dot size represents gene counts or gene Ratio (number of genes on the input list associated to a specific GO term / total number of genes on the list), while dot color represents significance of enrichment in terms of *p.adjust*: adjusted pvalue is a correction of the p-value in the case of multiple tests; the most common are the False Discovery Rate (FDR) which estimates false positives on the enrichment analysis output taking into account the pvalue of each single test and the number of tests, and the Bonferroni correction, which modifies the threshold of significance by dividing it by the number of performed tests.

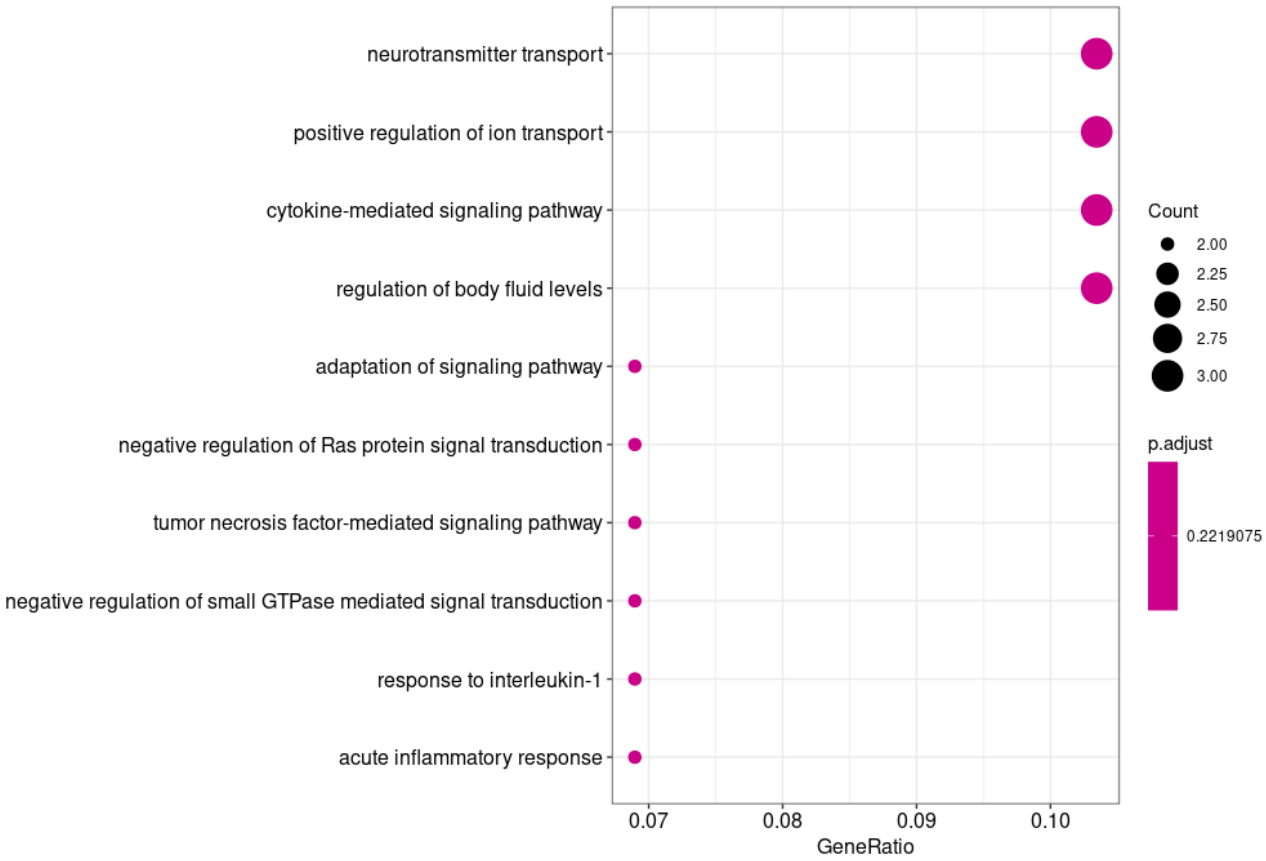

Figure 31 - Dotplot showing results of GO analysis (universe = GeneList2)

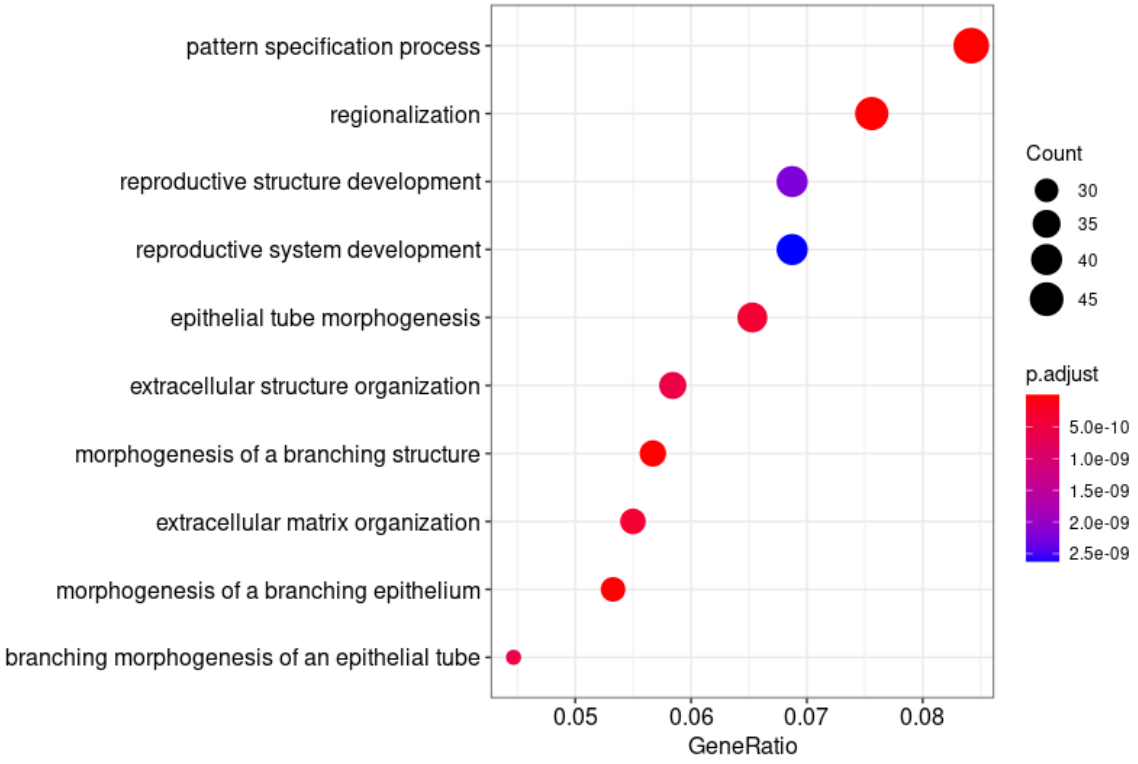

Figure 32 - Dotplot showing results of GO analysis (universe = GeneList3)

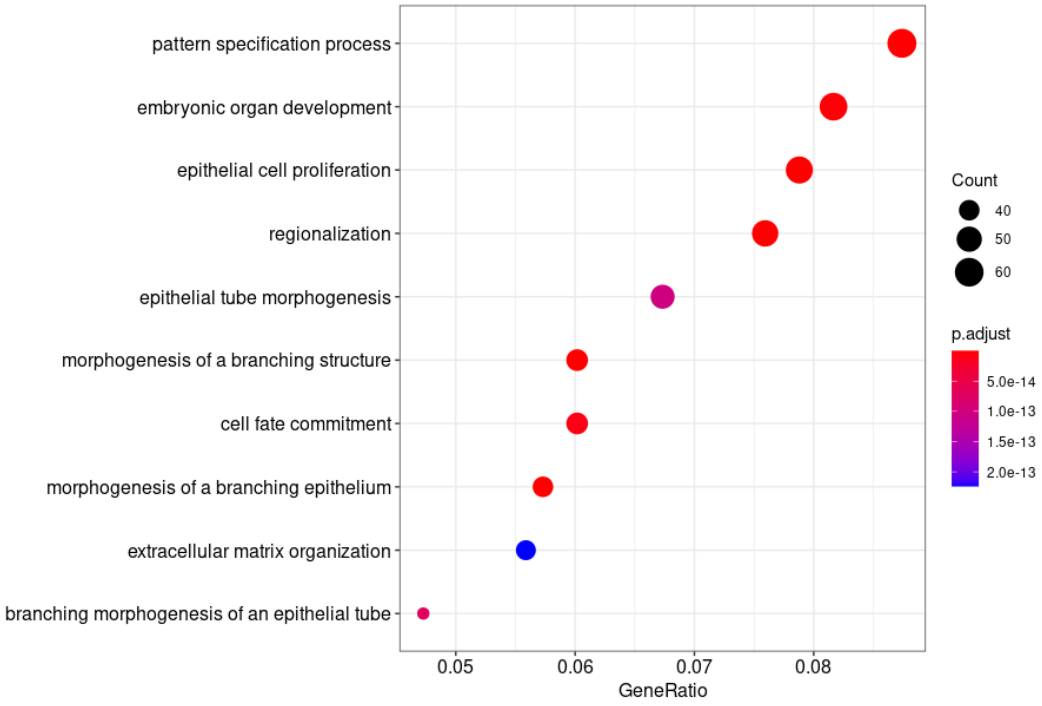

Figure 33 - Dotplot showing results of GO analysis (universe = GeneList4)

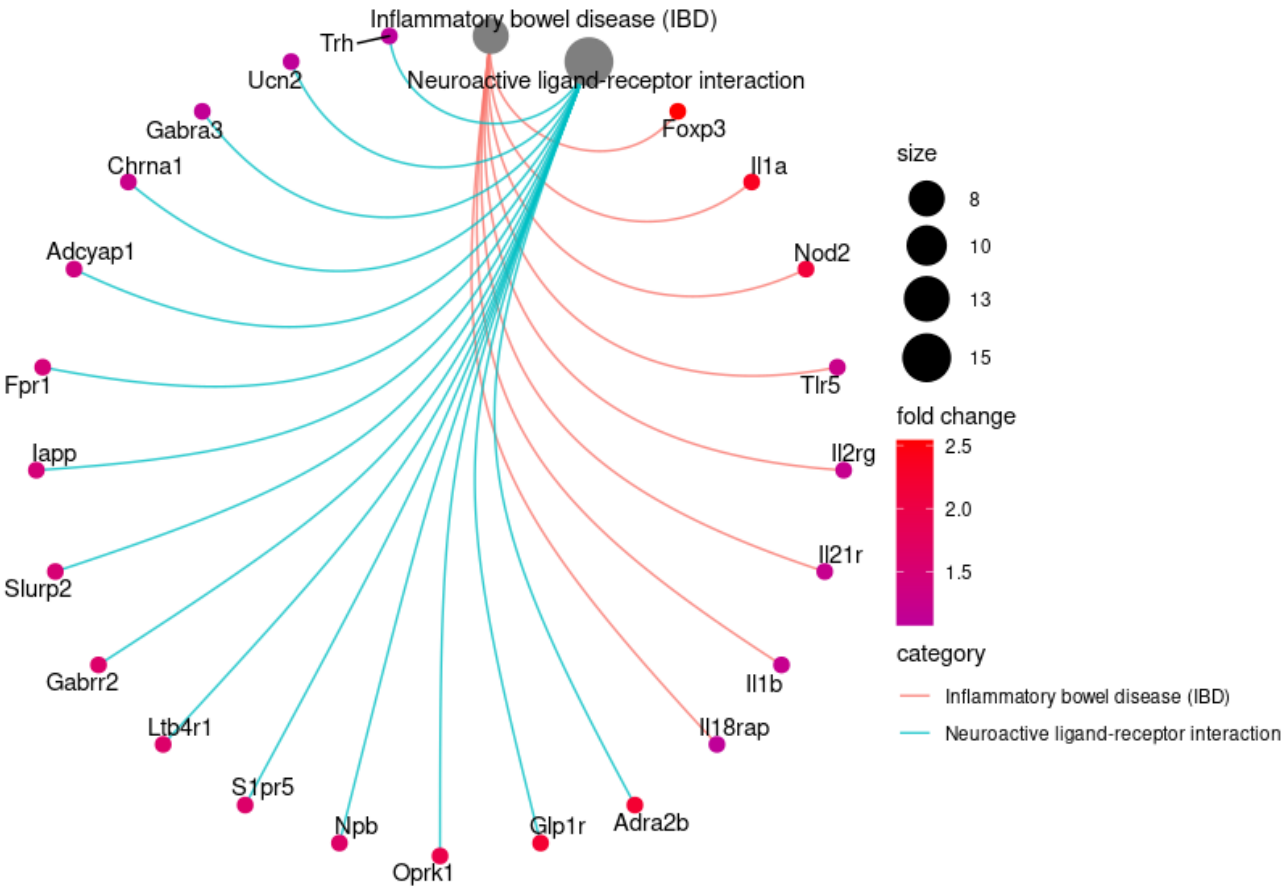

Figure 34 - Cnet plot for de1 bis gene list (pvalue=0.05)

Figure 34 shows Cnet plot related to de 1bis gene list. Cnet plot is a circular network plot widely used for the visualization of the complex associations between genes and gene sets: it depicts the linkage between genes and biological concepts (*e.g.* GO terms or KEGG pathways) as a circular network, with grey big dots representing the number of genes on that pathway and colored little dots indicating the fold change value.

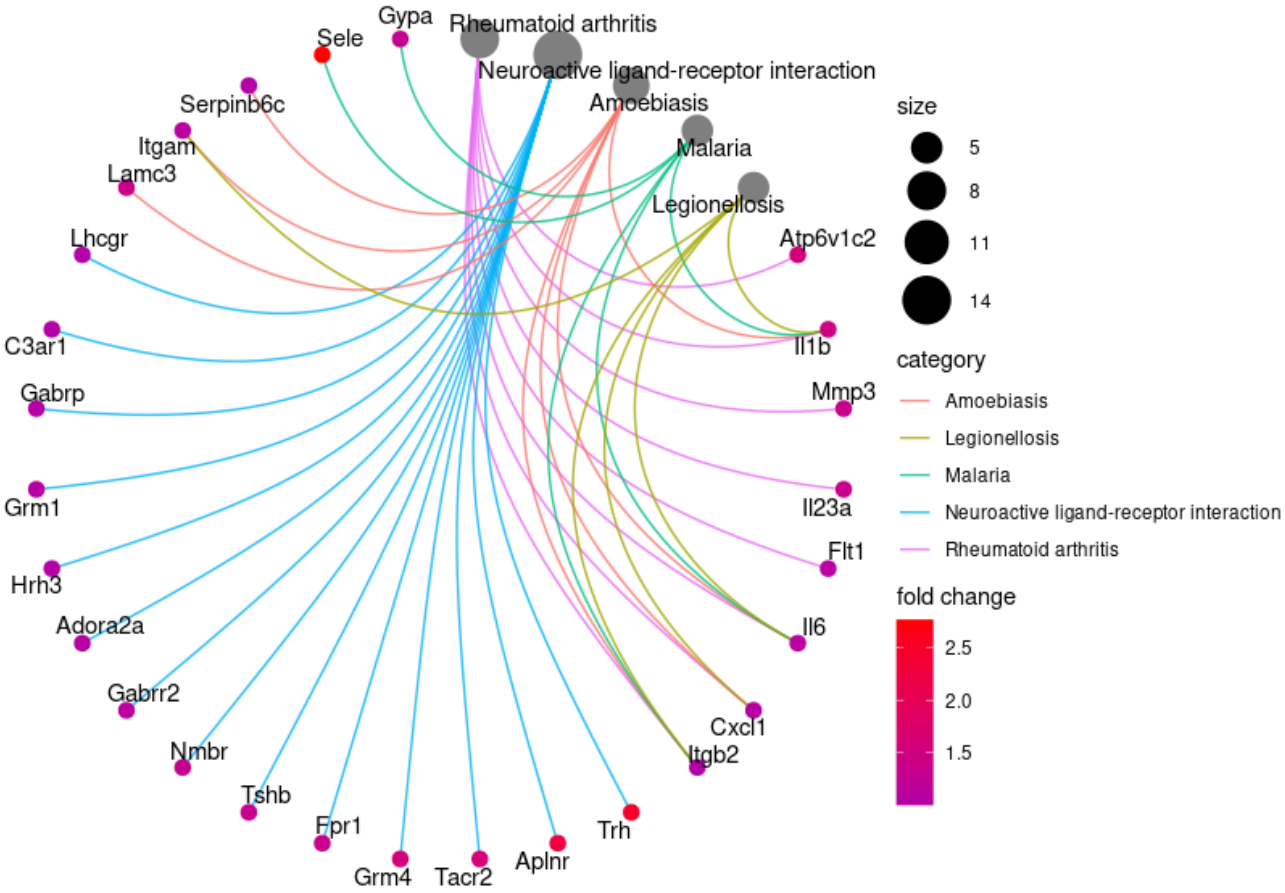

Figure 35 - Cnet plot for de2 bis gene list (pvalue=0.05)

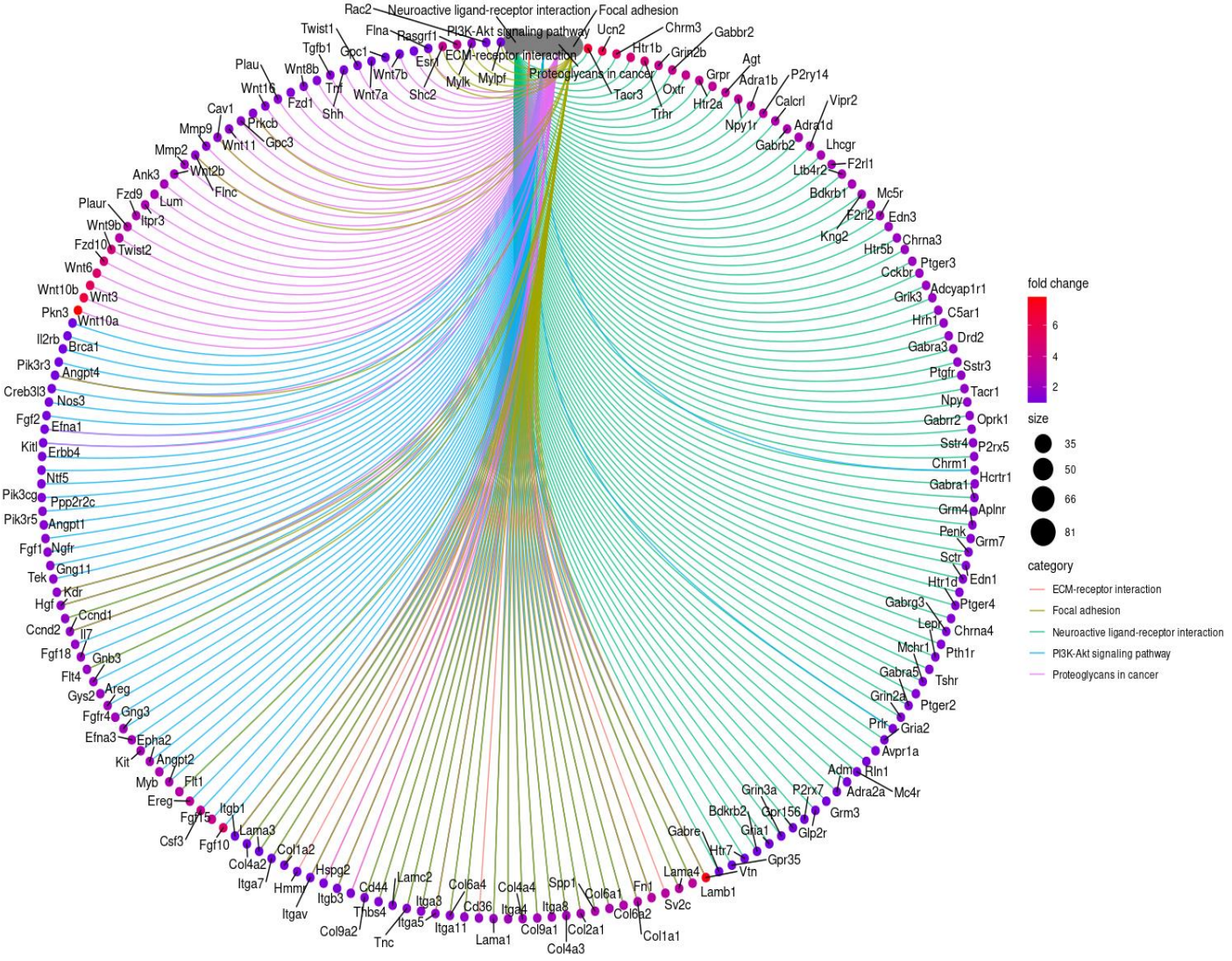

Figure 36 - Cnet plot for de3 bis gene list (pvalue=0.05)

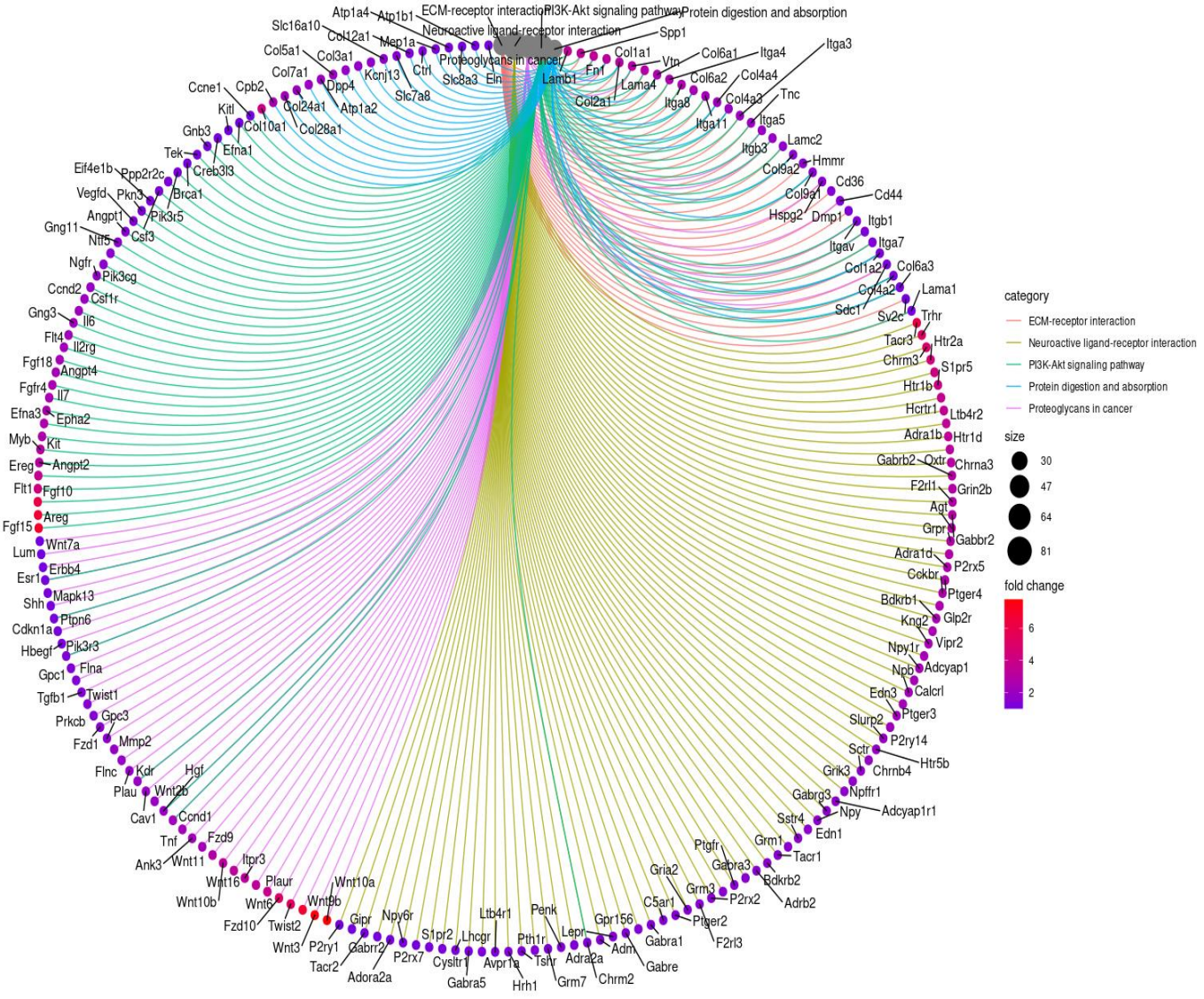

Figure 13 - Cnet plot for de4 bis gene list (pvalue=0.05)

- - Small RNAs analysis

Similarly to the long RNAs, the first step in small RNAs analysis consists in the alignment to the reference genome and subsequent indexes creation.

For this purpose, Multi Genome Alignment (MGA) [61] was used: MGA screens for contaminants by aligning sequence reads in FASTQ format against a series of reference genomes through the usage of Bowtie and against a set of adapter sequences using Exonerate. Two differing alignment approaches are taken for:

- identifying sequences originated from a different species to that being sequenced;
- detecting the presence of adapter sequences ligated to the ends of sequence fragments.

The alignment results for each species are ranked based on the number of reads aligned to each of them. Each read may be aligned to multiple species as a result of sequence homology between species. To distinguish contamination from sequence homology, each read is assigned to a single species based on the above ranking.

After alignment, high throughput data analysis can begin: in this step, Kraken software [62] was used. This software is actually a package consisting of 3 tools (Reaper, Tally and Sequence Imp) designed to streamline the analysis of next-generation sequencing data.

- Reaper is a C-program for demultiplexing, trimming and filtering short read sequencing data.
- Tally removes redundancy from sequence files by collapsing identical reads to a single entry while recording their number of instances.
- Sequence Imp is a pipeline incorporated in Reaper and Tally tools, designed to streamline RNA sequence analysis for multiple FASTQ files simultaneously.

The output data of these initial steps are the input for the next phase of the analysis: as for long RNAs, also in this case, an exploratory analysis is required before any differential expression analysis can be done. This exploratory analysis, and the subsequent differential expression analysis, have been conducted using DESeq2 package [56] for Bioconductor R language. The following plots show the results of the exploratory analysis for small RNAs.

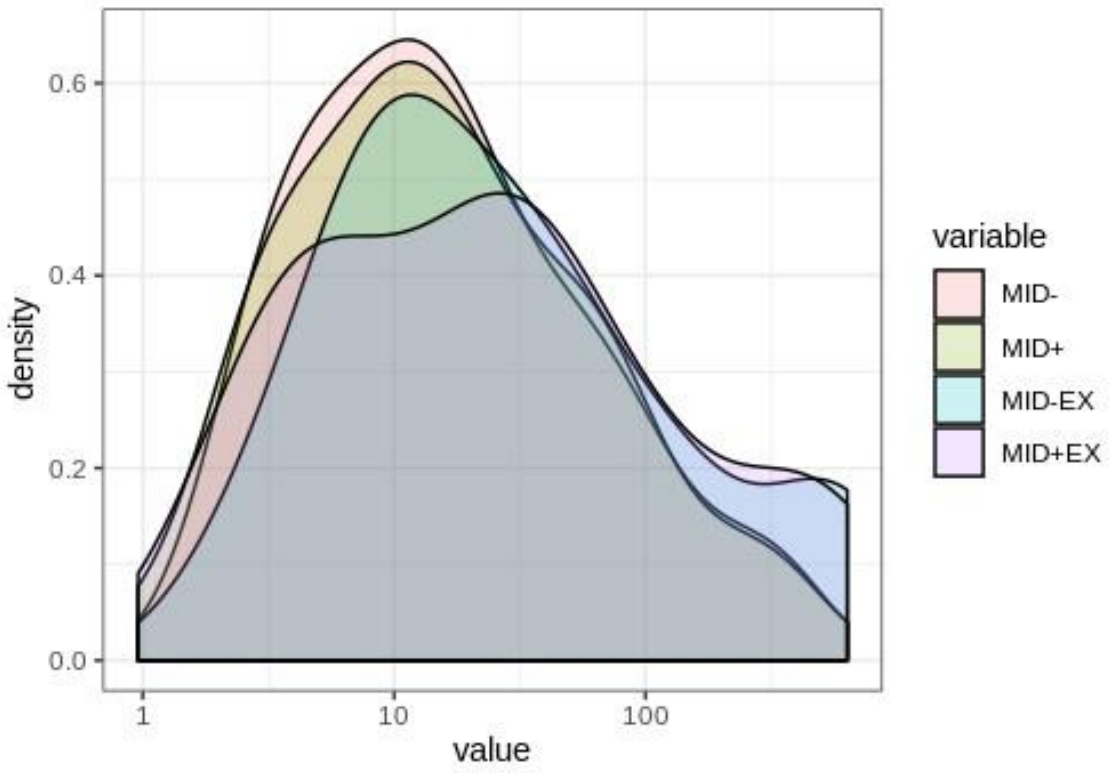

Figure 38 - Density plot of counts shows how the trend of normalized counts is not homogeneous among all samples.

Fig. 38 shows the distribution density of the variables related to the average of the normalized counts.

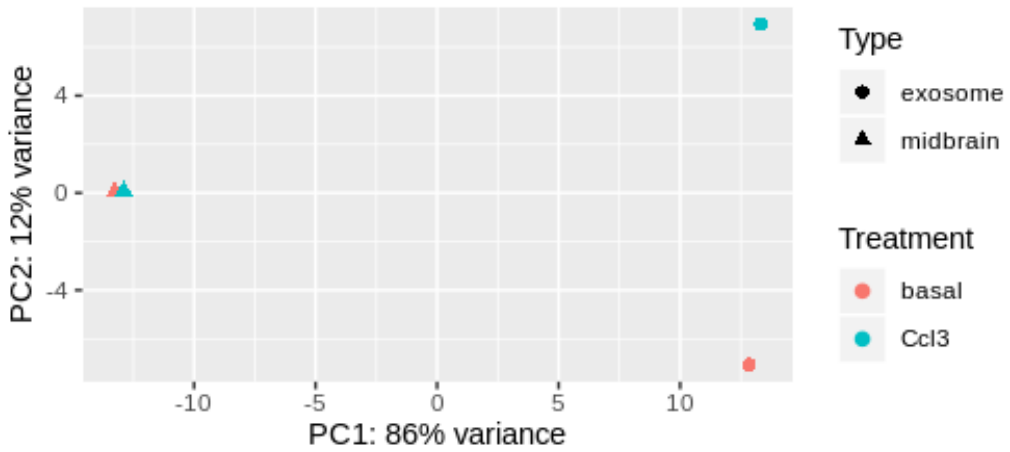

Figure 14 - Principal Component Analysis (PCA) highlights how the treatments can affect different samples in different ways.

In PCA plot shown in fig. 26, the data are represented in the X-Y coordinate system. PCA takes a large dataset as input and reduces the number of gene "dimensions" to a minimal set of linearly transformed dimensions reflecting the total variation of the dataset. The results are usually presented as a two-dimensional plot in which data are visualized along axes describing the variation within the dataset, known as the principal components (PCs). PC1 describes the most variation within the data, PC2 the second most, and so on. The variation represented by each PC can be calculated as a percentage of the total variance and visualized on the plot: a PCA plot may help to visualize grouping among replicates and aid in identifying technical or biological outliers [59].

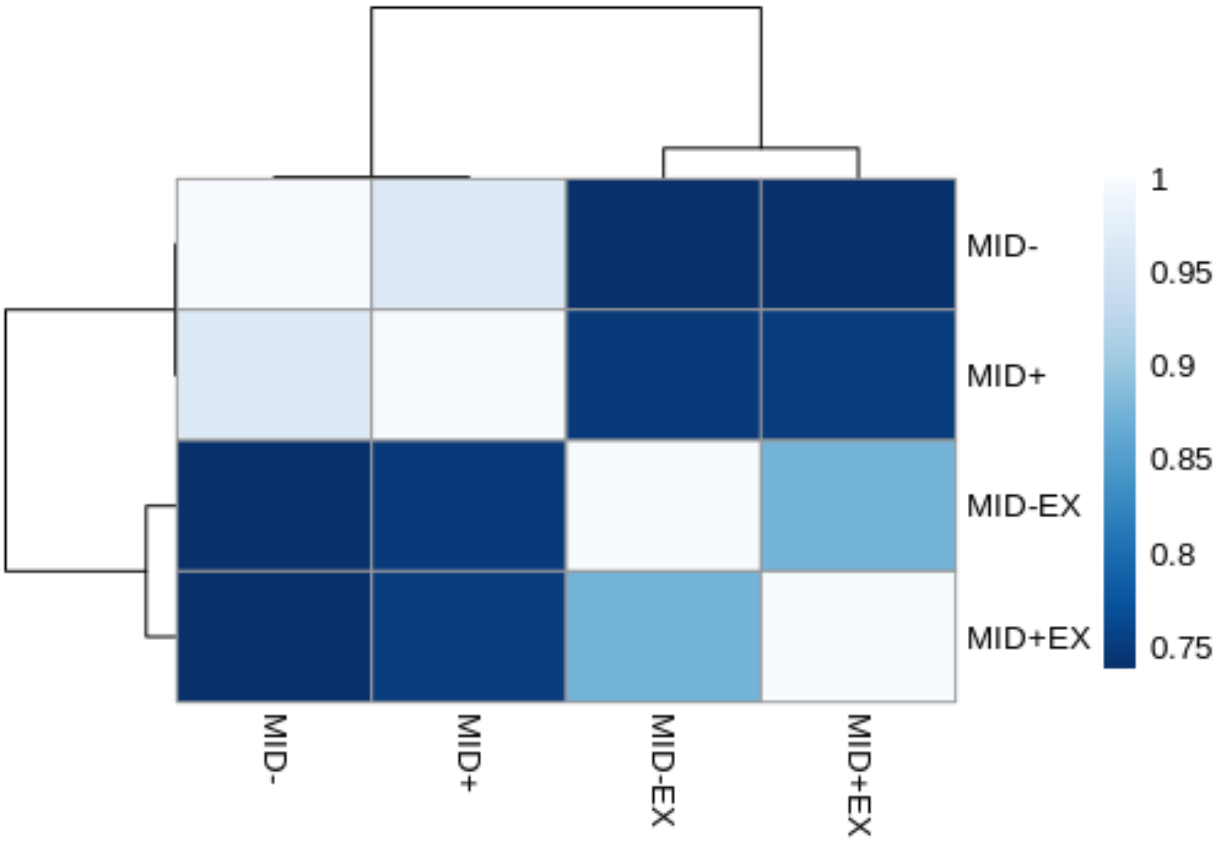

Figure 40 - Samples Correlation Heatmap

Color bar in fig. 40 represents Spearman's correlation coefficients; each color is indicative of the strength and direction of the association between two ranked variables: 1 value stands for perfect correlation (identity) and represents the highest possible value. The more similar the expression profiles for all transcripts are between two samples, the higher the correlation coefficient will be.

The following figures (28, 29, 30, 31) show Volcano plots for each of 4 experimental conditions:

1. midbrain Ccl3 vs midbrain basal
2. exosome Ccl3 vs exosome basal
3. midbrain basal vs exosome basal
4. midbrain Ccl3 vs exosome Ccl3

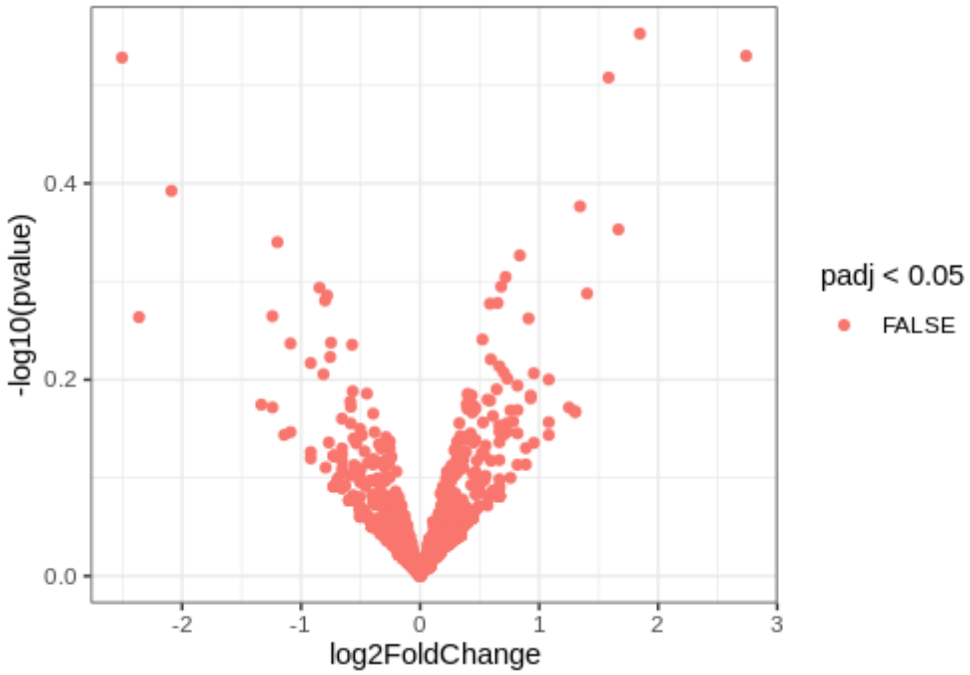

Figure 41 - Volcano Plot (condition: midbrain Ccl3 vs midbrain basal)

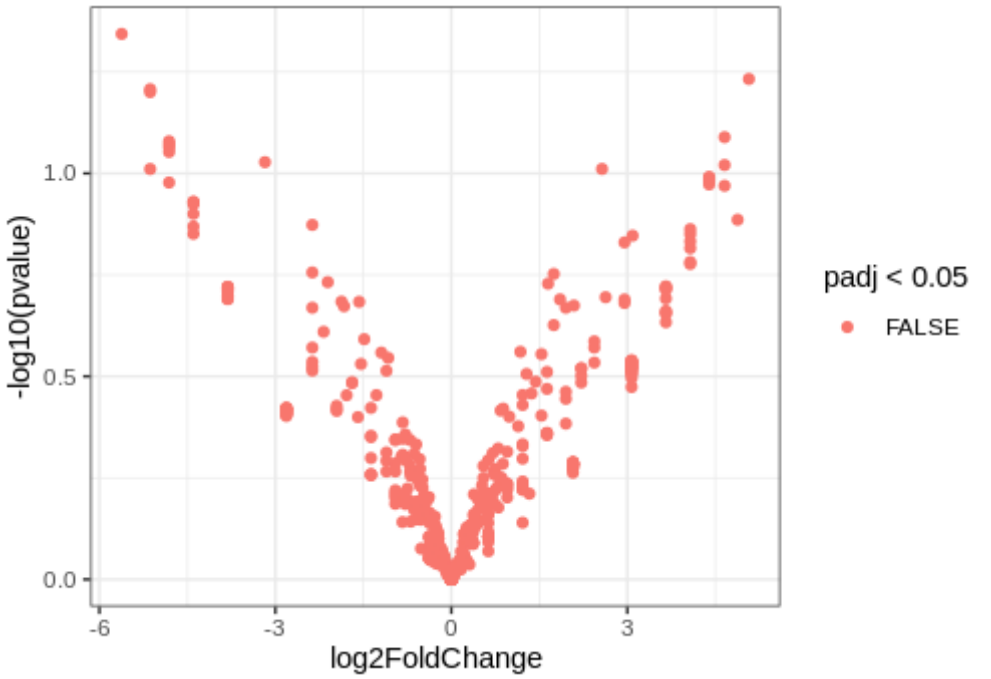

Figure 42 - Volcano Plot (condition: exosome Ccl3 vs exosome basal)

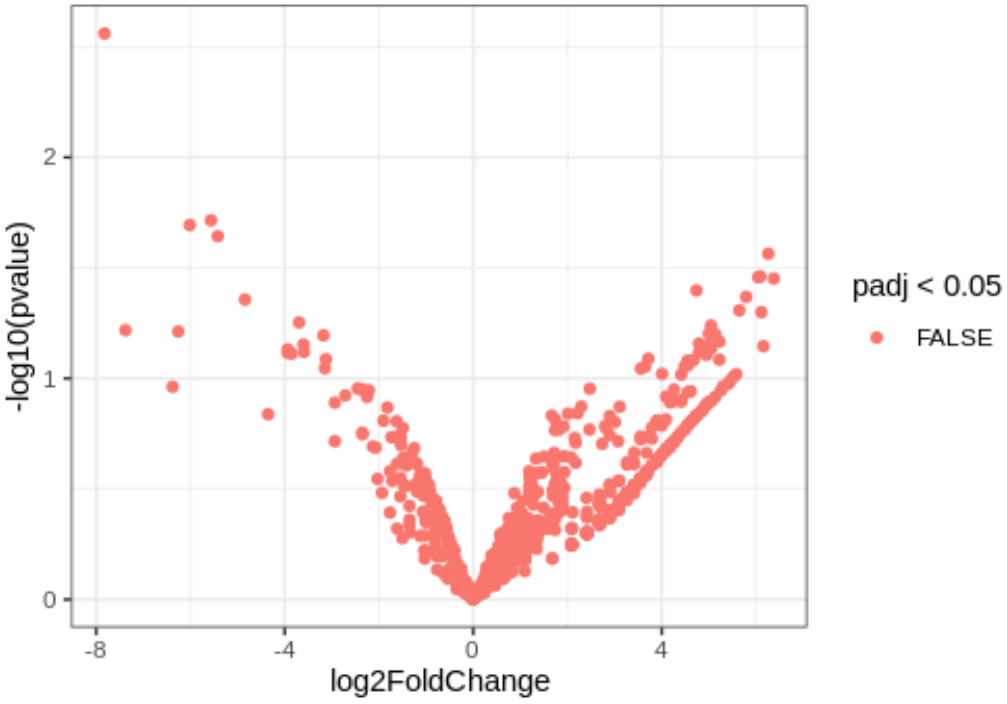

Figure 43 - Volcano Plot (condition: midbrain basal vs exosome basal)

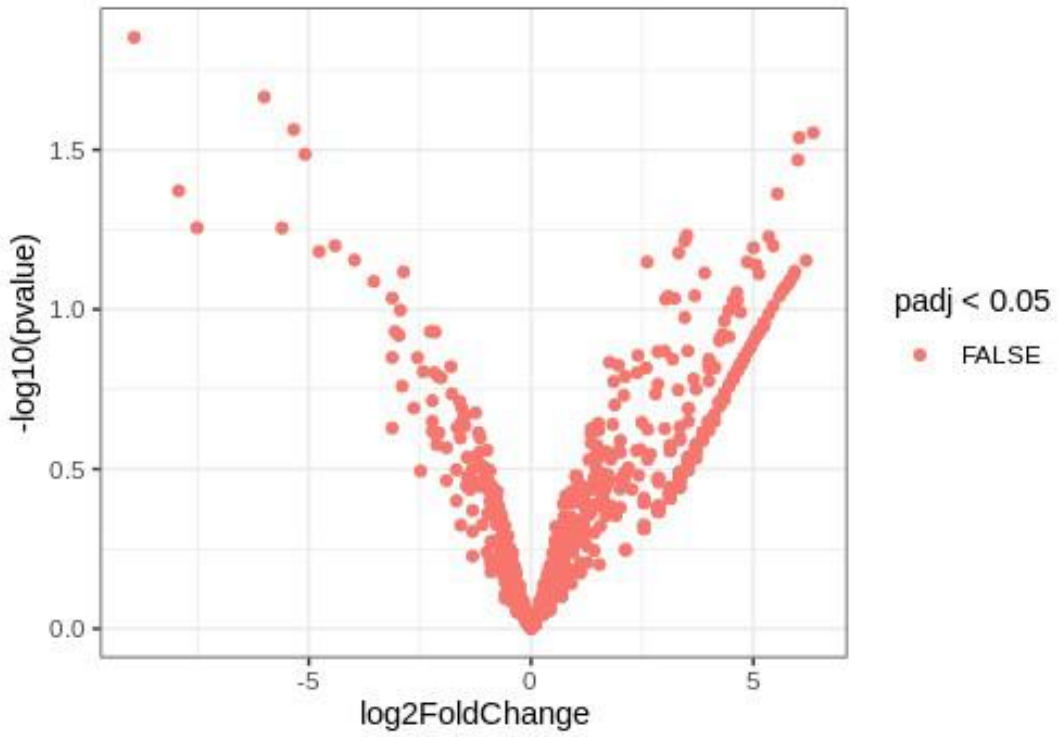

Figure 44 - Volcano Plot (condition: midbrain Ccl3 vs exosome Ccl3)

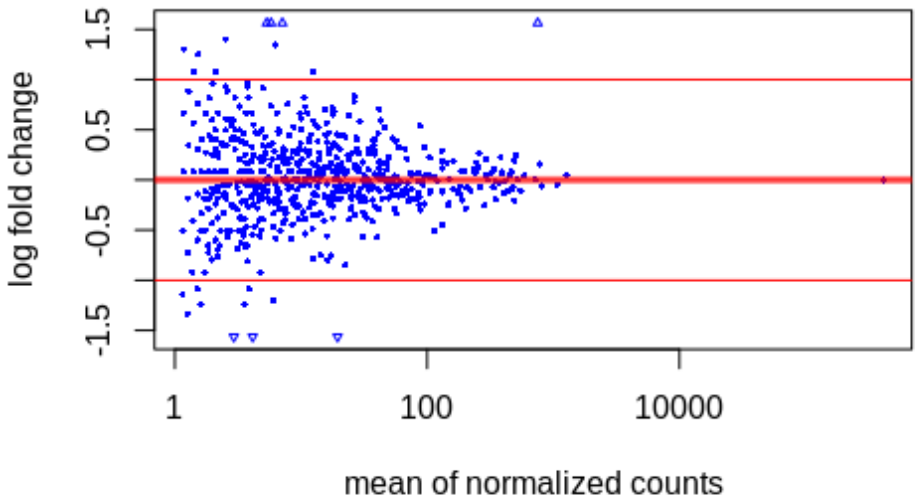

Figure 45 - MA Plot (condition: midbrain Ccl3 vs midbrain basal)

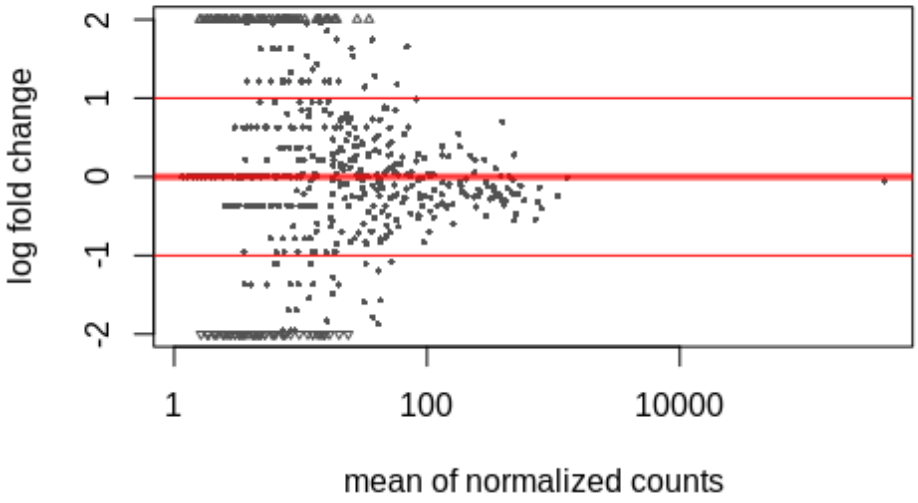

Figure 46 - MA Plot (condition: exosome Ccl3 vs exosome basal)

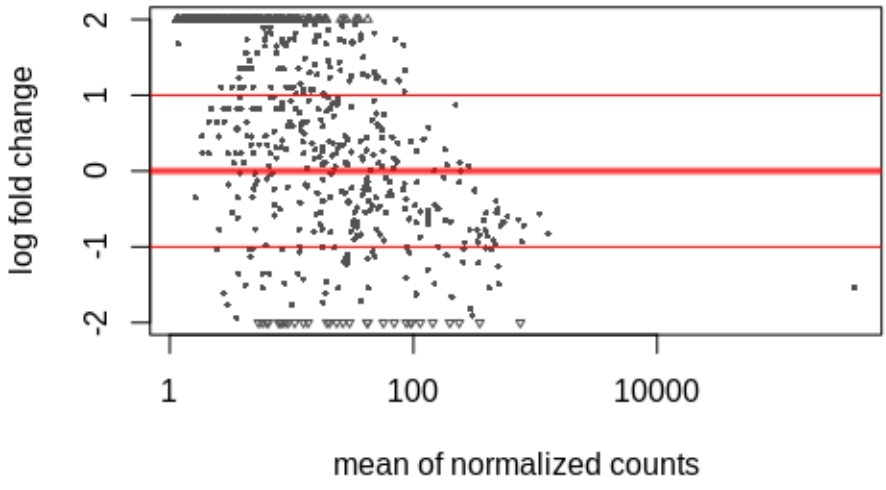

Figure 47 - MA Plot (condition: midbrain basal vs exosome basal)

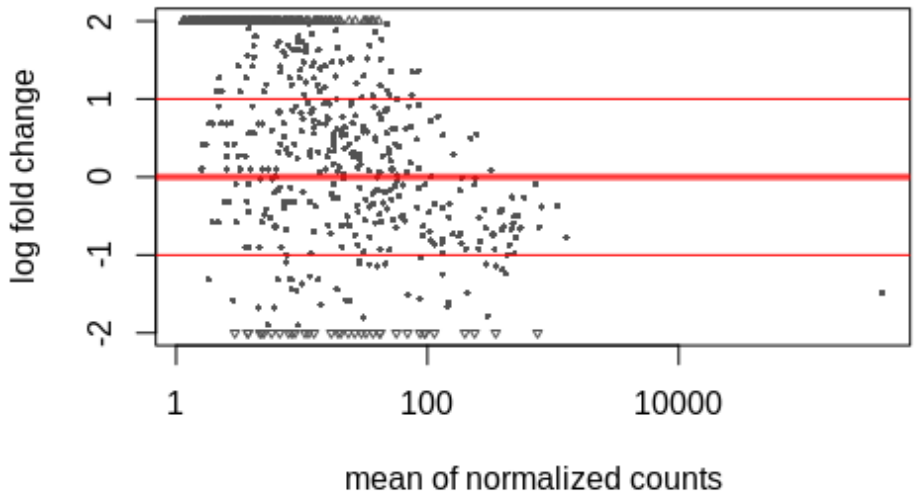

Figure 48 - MA Plot (condition: midbrain Ccl3 vs exosome Ccl3)

Figures 45, 46, 47, 48 show MA plots for each experimental condition: the abscissa indicates the mean of normalized counts, the ordinate the log2 (fold-change). The MA-plot represents each gene with a dot. The *x* axis is the average expression over all samples, while the *y* axis represents the log2 fold change of normalized counts (i.e., the average of counts normalized by the size factor) between treatment and control. Genes with an adjusted *p* value below a threshold 0.1, (the default level) are shown in red.

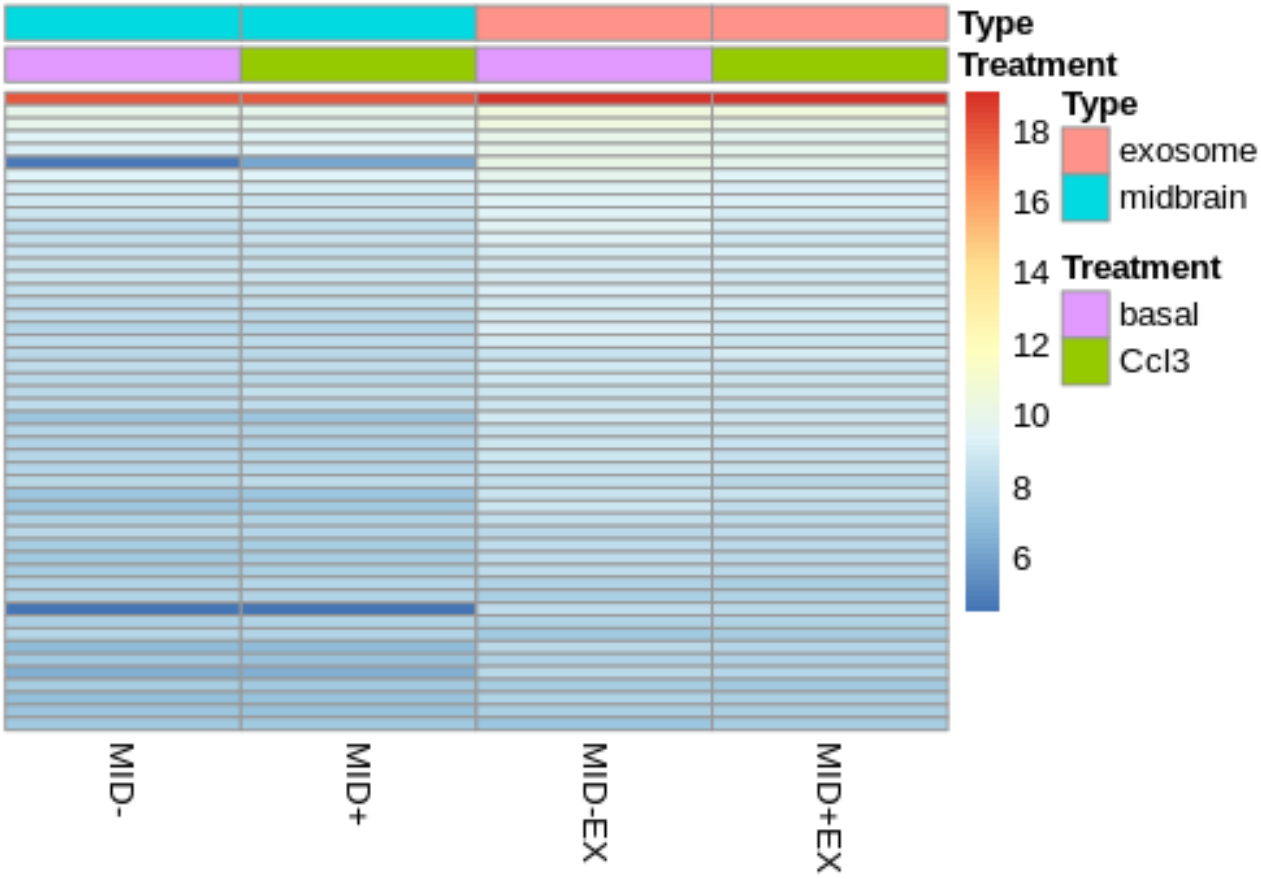

Figure 49 - Heatmap of counts (top 50 counts)

Differential expression analysis has been conducted on normalized data using DESeq2 package [56] for Bioconductor R language. Fig. 49 shows the heatmap of the top 50 genes in terms of read counts across the four experimental conditions.

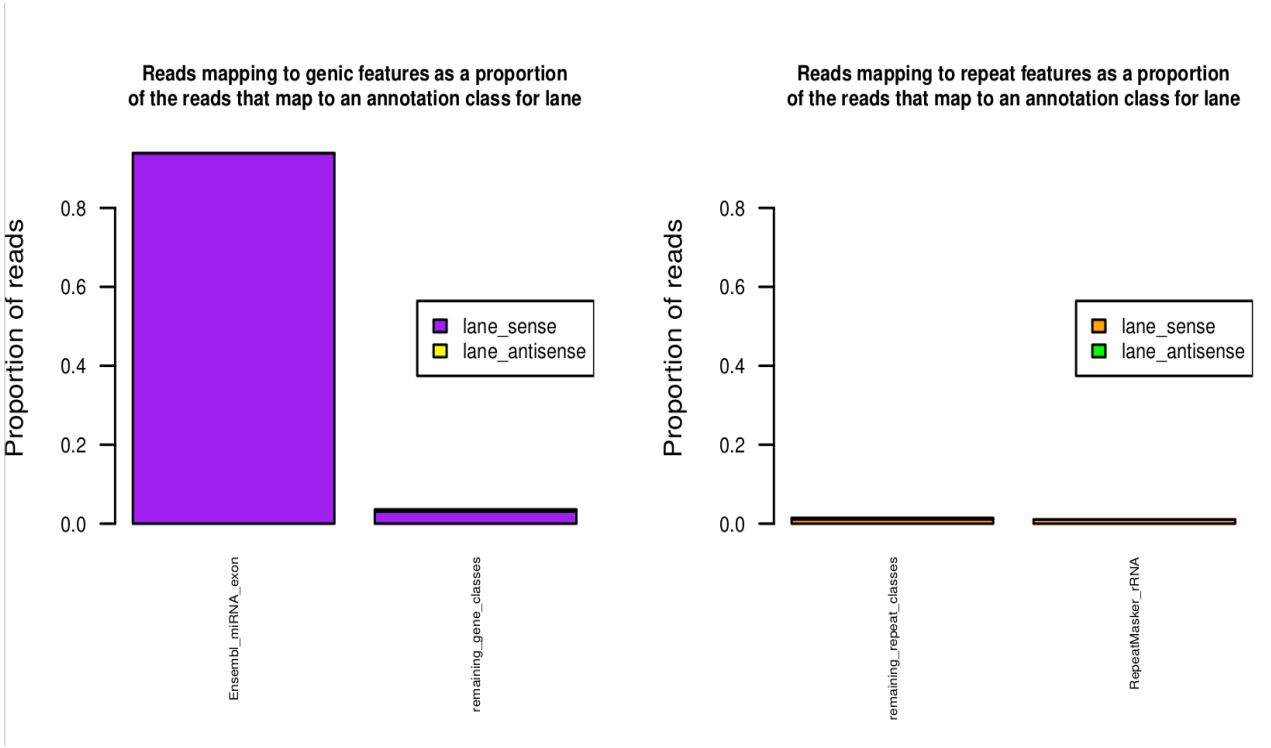

Figure 50 - miRNAs expression across samples (MID-)

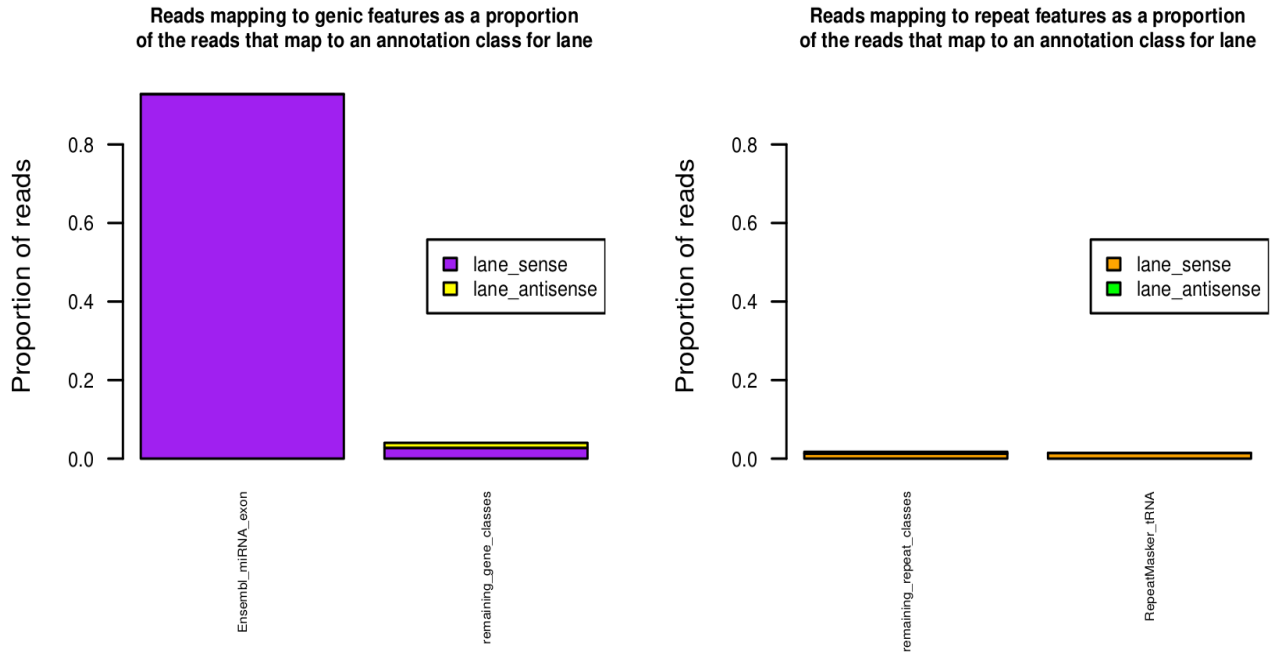

Figure 51 - miRNAs expression across samples (MID+)

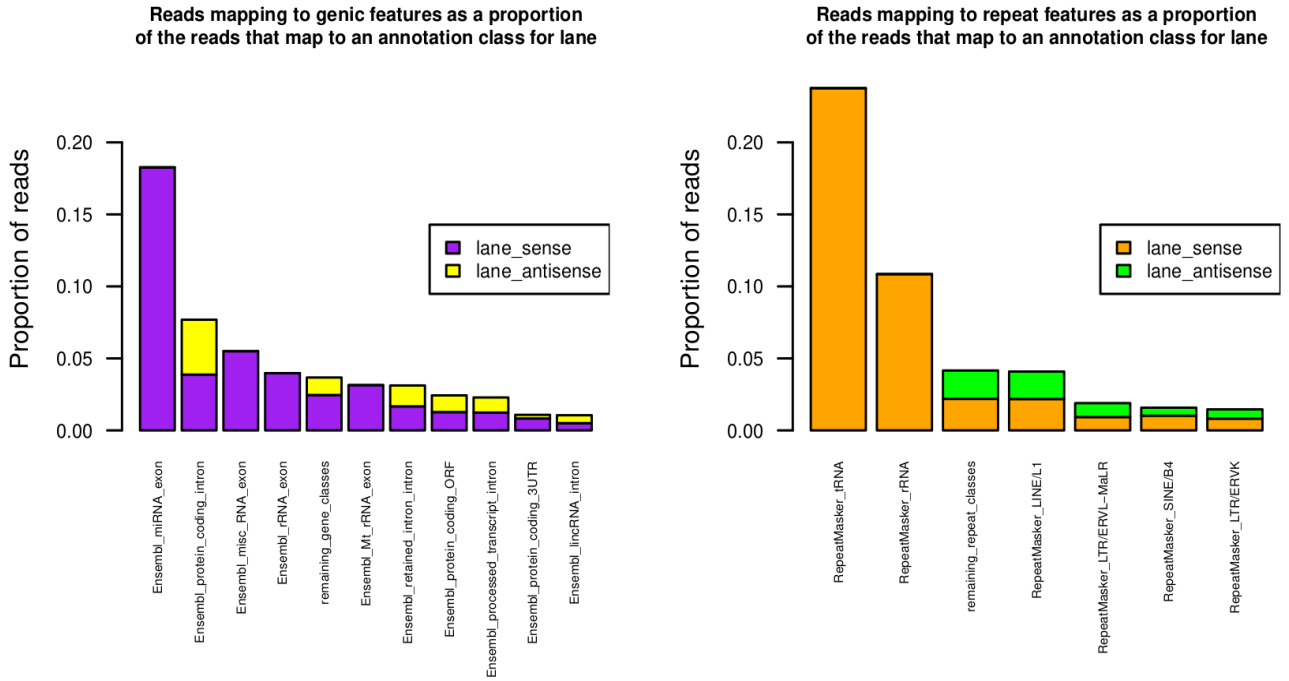

Figure 15 - miRNAs expression across samples (MID- EX)

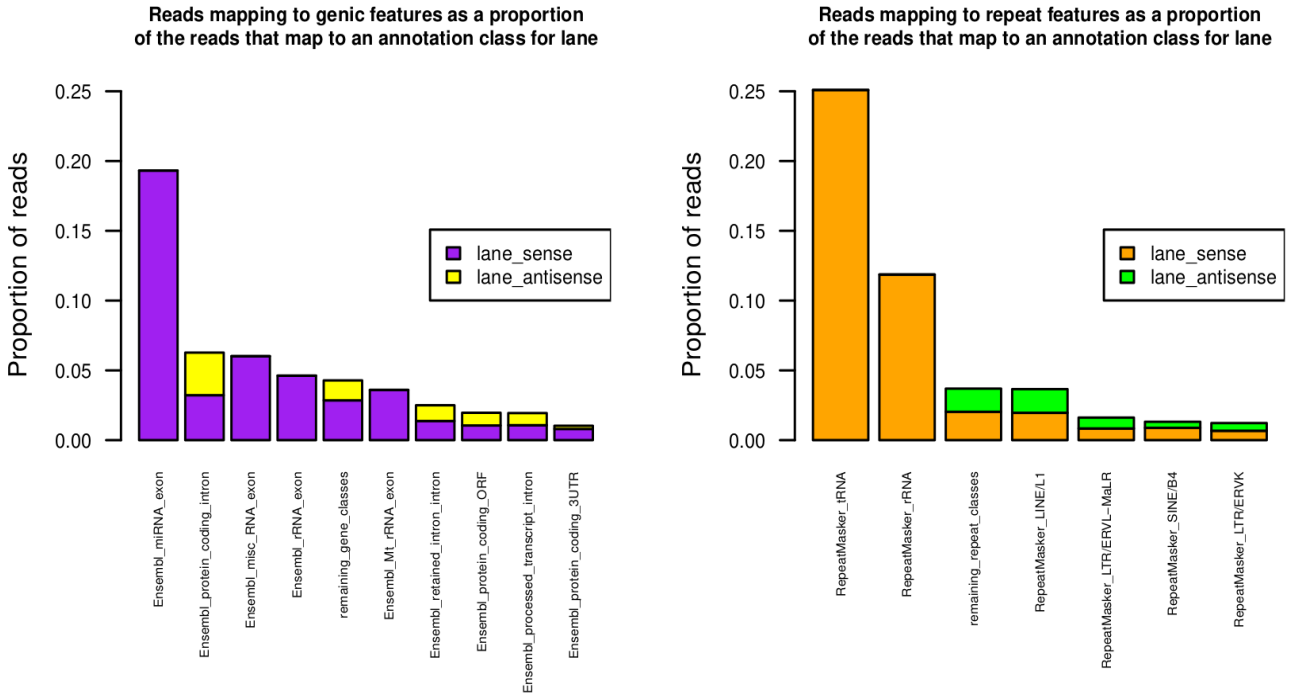

Figure 5316 - miRNAs expression across samples (MID+ EX)

Fig. 50, 51, 52, and 53 represent the result of a more detailed seqimp analysis on our four samples, conducted to obtain a more in-depth knowledge of the composition of the reads and the miRNAs distribution across all samples. The right side of each picture shows the seqimp analysis performed by using Repeat Masker annotations. Repetitive DNA sequences are frequent in a wide range of large genomes and in humans, covering nearly half of the genome. Human genome repeats are usually separated into two classes: short tandem repeats (also called microsatellites) and longer interspersed repeats (called short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs)) [63]. Repeats need to be masked before gene annotation, as they will cause non-specific gene hits, thus producing errors when interpreting results of alignment or in genome assembly.

In the past decade, RNA sequencing (RNA-seq) has become more widely used thanks to the decreasing costs and the disclosure of shared resource sequencing cores at many research institutes. The increasing popularity of RNA-seq has led to a growing need for bioinformatics expertise and computational resources. Basically, RNA-seq analysis aims to identify differentially expressed and coregulated genes and to infer biological meaning for further studies [59].

This work is the result of the collaboration between the University of Catania, in the people of prof. Nunzio Iraci and prof. Bianca Marchetti, and the Italian Institute of Technology (IIT) of Milan, in the person of Tommaso Leonardi, researcher in computational RNA biology; although it is still to be considered in a preliminary phase, and more samples with a greater number of replicates have yet to be analyzed in order to guarantee greater validity from a statistical point of view, it certainly represents an excellent starting point for a more in-depth and complete analysis: this workflow represents the framework upon which any future Rna-seq data analysis will be structured, as it can be easily adapted to any new dataset simply by adjusting thresholds and method anew.

6 CONCLUSIONS

With the advent of new technologies in sequencing and the resulting production of huge amounts of data, the need for a new way to analyze such a vast amount of data arises: bioinformatics was born with this very purpose, to mine data, store data, and ensure valid biological conclusions can be inferred from data. It is only thanks to Bioinformatics that today researchers worldwide can easily compare their results and then find biological connections that otherwise would go unnoticed.

In this PhD project, three different research topics have been presented. In each of them, bioinformatics played an essential role both in the data analysis and interpretation of the results obtained, showing once again what a powerful and versatile tool it is in facilitating and accelerating the various and complex phases of traditional analysis.

In the first topic, bioinformatic tools have been used to investigate the role of MAP3K8, a serine/threonine kinase expressed in thyroid cancer stem cells (CSCs), in mediating drug resistance in human thyroid cancer and its relationship to tumor behavior. As a result of this analysis, we demonstrated that high values of MAP3K8 expression are related to a particularly aggressive and lethal tumor type, Anaplastic Thyroid Cancer (ATC), thus highlighting the role of MAP3K8 as a potential prognostic biomarker. The possibility of considering MAP3K8 as a tumoral biomarker would be a great chance both to predict response to therapy and the progression of cancer itself but needs to be confirmed by future in vivo experiments.

The second topic deals with oncolytic viruses (OVs). Oncolytic viruses are a form of immunotherapy that employs attenuated viruses to restore the immune system response to infect and destroy cancer cells. In collaboration with Etna Biotech company of Catania, we investigated the antitumor efficacy of *cetuximab*, a widely used anti-epidermal growth factor receptor (EGFR) monoclonal antibody, combined with measles virus (MV). Although today experiments to create new constructs are still underway, we now know that with the bioinformatics tools at our disposal, we will be able to estimate their efficiency and evaluate their performance even before laboratory tests start, allowing this considerable saving of time and resources.

The third and last topic of this thesis aims to provide a complete workflow for RNA-seq data analysis. Starting from four *Mus musculus* samples sequencing data, we build a framework upon which any future RNA-seq data analysis could be structured, combining both informatics and statistical tools. Even if this study is still in progress as new data with more biological replicates are needed. Yet, it represents an excellent starting point for future and more in-depth analysis, which

could lead to a deeper understanding of the role of long and small non-coding RNA in both cancer and neurodegenerative diseases development, progression, and pathology and eventually to non-coding RNA-based therapies.

BIBLIOGRAPHY

[1] S. Y. Zhang and S. L. Liu, "Bioinformatics," in *Brenner's Encyclopedia of Genetics: Second Edition*, Elsevier Inc., 2013, pp. 338–340.

[2] A. Bridge, L. Lane, H. Stockinger, R. Appel, and I. Xenarios, "Databases and Datasources at SIB, Swiss Institute of Bioinformatics," in *Comprehensive Biomedical Physics*, vol. 6, Elsevier, 2014, pp. 191–204.

[3] K. Roy, S. Kar, and R. N. Das, "Other Related Techniques," in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Elsevier, 2015, pp. 357–425.

[4] Loren Dean Williams, "Molecular Interactions (Noncovalent Interactions)." [Online]. Available: https://ww2.chemistry.gatech.edu/~lw26/structure/molecular_interactions/mol_int.html. [Accessed: 03-Dec-2020].

[5] S. F. Zhou and W. Z. Zhong, "Drug design and discovery: Principles and applications," *Molecules*, vol. 22, no. 2. MDPI AG, 01-Feb-2017.

[6] V. K. Ramanan, L. Shen, J. H. Moore, and A. J. Saykin, "Pathway analysis of genomic data: Concepts, methods, and prospects for future development," *Trends in Genetics*, vol. 28, no. 7. Trends Genet, pp. 323–332, Jul-2012.

[7] D. D. Roumpeka, R. J. Wallace, F. Escalettes, I. Fotheringham, and M. Watson, "A review of bioinformatics tools for bio-prospecting from metagenomic sequence data," *Frontiers in Genetics*, vol. 8, no. MAR. Frontiers Research Foundation, 06-Mar-2017.

[8] Y. Chu and D. R. Corey, "RNA sequencing: Platform selection, experimental design, and data interpretation," *Nucleic Acid Ther.*, vol. 22, no. 4, pp. 271–274, Aug. 2012.

[9] K. Iida and I. Nishimura, "Gene expression profiling by DNA microarray technology," *Critical Reviews in Oral Biology and Medicine*, vol. 13, no. 1. Intern. and American Associations for Dental Research, pp. 35–50, 01-Jan-2002.

[10] L. Du and A. Pertsemlidis, "Cancer and neurodegenerative disorders: Pathogenic convergence through microRNA regulation," *Journal of Molecular Cell Biology*, vol. 3, no. 3. Oxford University Press, pp. 176–180, Jun-2011.

[11] J. Driver, "Investigating the link between cancer and neurodegenerative disease."

[12] A. L. Houck, S. Seddighi, and J. A. Driver, "At the Crossroads Between Neurodegeneration and Cancer: A Review of Overlapping Biology and Its Implications," *Curr. Aging Sci.*, vol. 11, no. 2, pp. 77–89, Feb. 2018.

[13] G. G. Kovacs, "Molecular pathological classification of neurodegenerative diseases: Turning towards precision medicine," *International Journal of Molecular Sciences*, vol. 17, no. 2. MDPI AG, 02-Feb-2016.

[14] F. Durães, M. Pinto, and E. Sousa, "Old drugs as new treatments for neurodegenerative diseases," *Pharmaceuticals*, vol. 11, no. 2. MDPI AG, 01-Jun-2018.

[15] A. S. Schachter and K. L. Davis, “Alzheimer’s disease,” *Dialogues Clin. Neurosci.*, vol. 2, no. 2, pp. 91–100, Jun. 2000.

[16] M. D. Geschwind, “Prion Diseases,” *CONTINUUM Lifelong Learning in Neurology*, vol. 21, no. 6. Lippincott Williams and Wilkins, pp. 1612–1638, 01-Dec-2015.

[17] G. G. Kovacs, “Tauopathies,” in *Handbook of Clinical Neurology*, vol. 145, Elsevier B.V., 2018, pp. 355–368.

[18] I Ferrer, “[Alpha-synucleinopathies] - PubMed,” *Neurologia*, vol. 16, no. 4, pp. 163–170, 2001.

[19] R. H. Tan *et al.*, “TDP-43 proteinopathies: Pathological identification of brain regions differentiating clinical phenotypes,” *Brain*, vol. 138, no. 10, pp. 3110–3122, Oct. 2015.

[20] T. J. Cohen, V. M. Y. Lee, and J. Q. Trojanowski, “TDP-43 functions and pathogenic mechanisms implicated in TDP-43 proteinopathies,” *Trends in Molecular Medicine*, vol. 17, no. 11. Trends Mol Med, pp. 659–667, Nov-2011.

[21] H. Plun-Favreau, P. A. Lewis, J. Hardy, L. M. Martins, and N. W. Wood, “Cancer and Neurodegeneration: Between the Devil and the Deep Blue Sea,” *PLoS Genet.*, vol. 6, no. 12, p. e1001257, Dec. 2010.

[22] F. Gianì, G. Russo, M. Pennisi, L. Sciacca, F. Frasca, and F. Pappalardo, “Computational modeling reveals MAP3K8 as mediator of resistance to vemurafenib in thyroid cancer stem cells,” *Bioinformatics*, vol. 35, no. 13, pp. 2267–2275, Jul. 2019.

[23] P. D. med. . T. L. D. med. . S. P. M. D. med. . M. C. K. P. D. med. . H. D. P. D. med. D. h. c. ,an. M. F. P. D. Ralf Paschke, “The Treatment of Well-Differentiated Thyroid Carcinoma,” *Deutches Ärzteblatt Int.*, vol. 112, no. 26, pp. 452–458, 2015.

[24] M. Ragazzi *et al.*, “Coexisting well-differentiated and anaplastic thyroid carcinoma in the same primary resection specimen: immunophenotypic and genetic comparison of the two components in a consecutive series of 13 cases and a review of the literature,” *Virchows Arch.*, 2020.

[25] V. Di Salvatore *et al.*, “Gene expression and pathway bioinformatics analysis detect a potential predictive value of MAP3K8 in thyroid cancer progression,” in *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, 2019, pp. 2202–2206.

[26] C. Yan, M. Huang, X. Li, T. Wang, and R. Ling, “Relationship between braf v600e and clinical features in papillary thyroid carcinoma,” *Endocr. Connect.*, vol. 8, no. 7, pp. 988–996, Jul. 2019.

[27] “R: A Language and Environment for Statistical Computing.”

[28] E. L. Kaplan and P. Meier, “Nonparametric Estimation from Incomplete Observations,” 1958.

[29] H. Lin and D. Zelterman, “Modeling Survival Data: Extending the Cox Model,” *Technometrics*, vol. 44, no. 1, pp. 85–86, Feb. 2002.

[30] D. Sean and P. S. Meltzer, “GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor,” *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, Jul. 2007.

[31] Z. Wu and R. Irizarry, “Description of gcrma package,” 2018.

[32] F. Naef and M. O. Magnasco, “Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays,” *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 68, no. 1, p. 4, 2003.

[33] M. E. Ritchie *et al.*, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Jan. 2015.

[34] A. L. Tarca *et al.*, “A novel signaling pathway impact analysis,” *Bioinformatics*, vol. 25, no. 1, pp. 75–82, Jan. 2009.

[35] M. H. Cruz, Å. Sidén, G. M. Calaf, Z. M. Delwar, and J. S. Yakisich, “The Stemness Phenotype Model,” *ISRN Oncol.*, vol. 2012, pp. 1–10, 2012.

[36] T. M. Malta *et al.*, “Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation,” *Cell*, vol. 173, no. 2, pp. 338-354.e15, Apr. 2018.

[37] A. Sokolov, E. O. Paull, and J. M. Stuart, “One-class detection of cell states in tumor subtypes,” in *Pacific Symposium on Biocomputing*, 2016, vol. 21, pp. 405–416.

[38] C. Lee Ventola, “Cancer immunotherapy, part 1: Current strategies and agents,” *P T*, vol. 42, no. 6, pp. 375–383, Jun. 2017.

[39] Y. S. A. R.-V. R. A. L. R. C. Juan-Manuel Anaya, “Autoimmunity: From Bench to Bedside [Internet] - PubMed,” 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29087650/. [Accessed: 03-Dec-2020].

[40] J. Raja, J. M. Ludwig, S. N. Gettinger, K. A. Schalper, and H. S. Kim, “Oncolytic virus immunotherapy: Future prospects for oncology 11 Medical and Health Sciences 1107 Immunology 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis,” *Journal for ImmunoTherapy of Cancer*, vol. 6, no. 1. BioMed Central Ltd., 04-Dec-2018.

[41] P. D. John C. Bell, “Oncolytic Virus Therapy – Cancer Research Institute (CRI).” [Online]. Available: https://www.cancerresearch.org/immunotherapy/treatment-types/oncolytic-virus-therapy. [Accessed: 17-Dec-2020].

[42] S. J. Russell and K. W. Peng, “Measles virus for cancer therapy,” *Current Topics in Microbiology and Immunology*, vol. 330. Curr Top Microbiol Immunol, pp. 213–241, 2009.

[43] J. F. Bohnsack, J. J. O’Shea, T. Takahashi, and E. J. Brown, “Fibronectin-enhanced phagocytosis of an alternative pathway activator by human culture-derived macrophages is mediated by the C4b/C3b complement receptor (CR1).,” *J. Immunol.*, vol. 135, no. 4, 1985.

[44] T. Hashiguchi, K. Maenaka, and Y. Yanagi, “Measles virus hemagglutinin: Structural insights into cell entry and measles vaccine,” *Frontiers in Microbiology*, vol. 2, no. DEC. Frontiers Research Foundation, 2011.

[45] Z. Wu *et al.*, “Combination of Cetuximab and Oncolytic Virus Canerpaturev Synergistically Inhibits Human Colorectal Cancer Growth,” *Mol. Ther. - Oncolytics*, vol. 13, pp. 107–115, Jun. 2019.

[46] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, “SWISS-MODEL: An automated protein homology-modeling server,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3381–3385, Jul. 2003.

[47] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, “The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling,” *Bioinformatics*, vol. 22, no. 2, pp. 195–201, Jan. 2006.

[48] M. Blumenberg, “Introductory Chapter: Transcriptome Analysis,” in *Transcriptome Analysis*, IntechOpen, 2019.

[49] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: A revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, vol. 10, no. 1. Nat Rev Genet, pp. 57–63, Jan-2009.

[50] K. R. Kukurba and S. B. Montgomery, “RNA sequencing and analysis,” *Cold Spring Harb. Protoc.*, vol. 2015, no. 11, pp. 951–969, Nov. 2015.

[51] M. A. Parasramka, S. Maji, A. Matsuda, I. K. Yan, and T. Patel, “Long non-coding RNAs as novel targets for therapy in hepatocellular carcinoma,” *Pharmacology and Therapeutics*, vol. 161. Elsevier Inc., pp. 67–78, 01-May-2016.

[52] C. Li and Y. Chen, “Small and Long Non-Coding RNAs: Novel Targets in Perspective Cancer Therapy,” *Curr. Genomics*, vol. 16, no. 5, pp. 319–326, Jul. 2015.

[53] C. Zhang, “Novel functions for small RNA molecules,” *Current Opinion in Molecular Therapeutics*, vol. 11, no. 6. NIH Public Access, pp. 641–651, Dec-2009.

[54] A. Dobin *et al.*, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.

[55] B. Langmead, “Aligning short sequencing reads with Bowtie,” *Curr. Protoc. Bioinforma.*, vol. Chapter 11, no. SUPP.32, 2010.

[56] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol.*, vol. 15, no. 12, Dec. 2014.

[57] D. J. Bartholomew, “Principal components analysis,” in *International Encyclopedia of Education*, Elsevier Ltd, 2010, pp. 374–377.

[58] M. I. Love, S. Anders, V. Kim, and W. Huber, “RNA-Seq workflow: gene-level exploratory analysis and differential expression,” *F1000Research*, vol. 4, p. 1070, Oct. 2015.

[59] C. M. Koch *et al.*, “A beginner’s guide to analysis of RNA sequencing data,” *American Journal of Respiratory Cell and Molecular Biology*, vol. 59, no. 2. American Thoracic Society, pp. 145–157, 01-Aug-2018.

[60] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, “ClusterProfiler: An R package for comparing biological themes among gene clusters,” *Omi. A J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, May 2012.

[61] J. Hadfield and M. D. Eldridge, “Multi-genome alignment for quality control and contamination screening of next-generation sequencing data,” *Front. Genet.*, vol. 5, no. FEB, 2014.

[62] M. P. A. Davis, S. van Dongen, C. Abreu-Goodger, N. Bartonicek, and A. J. Enright, “Kraken: A set of tools for quality control and analysis of high-throughput sequence data,” *Methods*, vol. 63, no. 1, pp. 41–49, Sep. 2013.

[63] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: Computational challenges and solutions,” *Nature Reviews Genetics*, vol. 13, no. 1. Nat Rev Genet, pp. 36–46, Jan-2012.