RESYS Project : Project Report

DELOUIS Maëlys, REY Soraya

October 2023

1 Introduction

Lung cancer is the leading source of cancer death worldwide, in 2020, there were an estimated 1.8 million deaths from lung cancer, accounting for 18% of all cancer deaths. It is an intricate disease that is driven by a diversity of genetic and environmental factors. One of the key characteristics of cancer is the disruption of gene regulatory networks. Gene regulatory networks are responsible for controlling the expression of genes, in the case of disrupted functionning of said networks, chaotic cell growth and division arise. [FSS+21]

Single-cell RNA sequencing (scRNAseq) is a potent tool for investigating gene regulatory networks in cancer. scRNAseq allows to measure the gene expression profiles of individual cells. This information can be employed to distinguish cell types, define cell states, and reconstruct gene regulatory networks.

In a recent study, Kim et al. (2020) used scRNAseq to examine the molecular and cellular reprogramming of metastatic lung adenocarcinoma (LUAD) cells. They singled out a number of key genes and pathways that are related to LUAD metastasis. Additionally they demonstrated that the gene regulatory networks of metastatic LUAD cells are different from the gene regulatory networks of normal lung cells [KKL+20]. Another study by Huang et al. (2017) [Hua17] used scRNAseq to determine the key regulators of lung cancer immune response. They reconstructed gene regulatory networks for tumor cells, T cells, and myeloid cells with a Bayesian inference algorithm. They pinpointed a number of key genes and pathways that play a role in the lung cancer immune response. Finally, a study by Lee et al. (2013) [LYK+13] used scRNAseq to reconstruct gene regulatory networks for healthy lung cells and lung cancer cells. They employed a correlation-based network reconstruction algorithm. They identified numerous key genes and pathways that are involved in lung cancer progression.

In this study, we will handle scRNAseq data from LUAD patients to reconstruct gene regulatory networks for normal lung cells and metastatic LUAD cells. We will then mainly seek to highlight the differences between these networks in order to identify the changes in gene regulation that occur as LUAD progresses.

As the previously quoted works illustrates, there are a variety of methods that can be employed to reconstruct gene regulatory networks using scRNAseq data. One common approach would be correlation-based network reconstruction. Correlation-based network reconstruction defines pairs of genes that are highly correlated with each other. These pairs of genes are then utilized to build a network of gene interactions [Ngu21]. Another common approach to network reconstruction is to make use of Bayesian inference. Bayesian inference is a statistical method that can be applied to model the relationships between genes using a probability distribution. This probabilistic method can be thus be used to infer the most likely gene regulatory networks, given the scRNAseq data [Ngu21]. For this approach, we will use MIIC, a web server that reconstruct causal networks from non-perturbative data. Finally, Machine Learning algorithms can be used as well to reconstruct gene regulatory networks from scR-

NAseq data. Machine learning algorithms learn from a dataset of known gene regulatory networks to predict new gene regulatory networks [Ngu21]. Huynh-Thu et al. (2010) [HTIWG10] used a tree-based method called GENIE3 to predict regulatory networks from gene expression data. GENIE3 works acording to the following process : firstly, it constructs a random forest, which is an ensemble of decision trees. Each decision tree in the random forest is trained on a distinct subset of the data, and the predictions of the individual trees are then averaged to produce a final prediction. To infer a regulatory network, Huynh-Thu et al. made use of the variable importance measures of the random forest to determine the most saillant genes for predicting the expression of other genes. The genes with the highest variable importance measures are then aknowledged to be the regulators of the other genes. GENIE3 counts a number of advantages over other methods for reconstructing regulatory networks from inference on gene expression data. Firstly, it is capable of integrating the predictions of multiple machine learning methods to gain accuracy on its results. Secondly, it is comparatively robust to noise in the data. Thirdly, it is comparably fast and efficient to train. GENIE3 is an encouraging strategy for the inference of regulatory networks from scRNAseq data, which is the type of data that we will be using in our study. GENIE3 has been shown to be efficient at inferring regulatory networks from scRNAseq data in other studies. Its robustness is a suplementary appeal to this apporach since scRNAseq data can be noisy, we also value the fact that this algorithm is fast and efficient to train, because scRNAseq data can be very large and computationally expensive to analyze.

We make the assumption that GENIE3 is a promising approach for inferring regulatory networks from scRNAseq data. We shall evaluate the performance of GENIE3 on our dataset and contrast it to the other network reconstruction methods that we are considering.

Our study will provide new insights into the gene regulatory networks that drive LUAD progression and metastasis. This knowledge could be used to develop new diagnostic and therapeutic strategies for lung cancer.

2 Method

2.1 Dataset

2.1.1 Introduction

We will use the scRNAseq dataset from the study by Kim et al. (2020) [KKL$^+$20] implanted by Ahn M, Lee H (2020). This dataset contains scRNAseq data from 58 samples from 44 patients with LUAD, from several types of cells (normal, tumorous and metastatic).
Available at:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907 (Accessed: 13 October 2023)

Description of available data :

- $GSE131907\_Lung\_Cancer\_Feature\_Summary.xlsx$: This file contains a summary of the features in the dataset, including the gene name, gene symbol, and chromosome location.
- $GSE131907\_Lung\_Cancer\_cell\_annotation.txt.gz$: This file contains the cell annotation for the dataset, including the cell type, cell subtype, and sample ID.
- $GSE131907\_Lung\_Cancer\_normalized\_log2TPM\_matrix.rds.gz$: This file contains the normalized log2TPM matrix for the dataset, which is a measure of the gene expression level for each gene in each cell.
- $GSE131907\_Lung\_Cancer\_normalized\_log2TPM\_matrix.txt.gz$: This file contains the normalized log2TPM matrix for the dataset in text format.
- $GSE131907\_Lung\_Cancer\_raw\_UMI\_matrix.rds.gz$: This file contains the raw UMI matrix for the dataset, which is a measure of the number of transcripts for each gene in each cell.
- $GSE131907\_Lung\_Cancer\_raw\_UMI\_matrix.txt.gz$: This file contains the raw UMI matrix for the dataset in text format.

2.1.2 RNA sequencing

RNA sequencing (RNA-seq) is a next-generation sequencing method used to measure the expression of genes in a sample. It is a powerful tool for understanding the molecular basis of biological processes, and it has been used to make significant advances in a wide range of fields, including cancer research, immunology, and neuroscience. Single-cell RNA sequencing (scRNA-seq) is a type of RNA-seq that allows for the measurement of gene expression in individual cells. This is in contrast to bulk RNA-seq, which measures the average gene expression across all cells in a sample. scRNA-seq has revolutionized the field of cell biology by allowing researchers to identify and characterize new cell types, to understand how cell types develop and differentiate, and to study the dynamics of gene expression in response to stimuli. RNA-seq can be used to measure the expression of all genes in a sample. This information can be used to identify differentially expressed genes, which may be involved in disease or other biological processes. RNA-seq can be used to assemble the transcriptome, which is the complete set of RNA transcripts expressed in a sample. This information can be used to identify new genes and to study the alternative splicing of genes. RNA-seq can be used to detect genetic variants, such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels). This information can be used to study the genetic basis of disease and to develop personalized treatments. scRNA-seq can be used to identify and characterize new cell types. This information can be used to study the development and differentiation of cells, and to understand the role of different cell types in disease and other biological processes. In our study we are using single cell RNA-seq to identify differentially

expressed genes between the different stages of lung cancer. By doing so, we aim to highlight the genes that play a role in lung cancer progression.

2.1.3 Preprocessing of the Data

The dataset we use is a count matrix. A count matrix in the context of single-cell RNA sequencing (scRNA-seq) is a tabular representation of the number of reads mapped to each gene in each cell. It is the primary input for most scRNA-seq analysis tools. The count matrix is typically generated by aligning the sequencing reads to a reference genome and then counting the number of reads that map to each gene. The rows of the count matrix represent the cells, and the columns represent the genes. Each cell-gene intersection contains the number of reads that map to that gene in that cell. Preprocessing the count matrix is an essential step in scRNA-seq data analysis. It helps to improve the quality of the data and make it more accurate and reliable for downstream analysis.

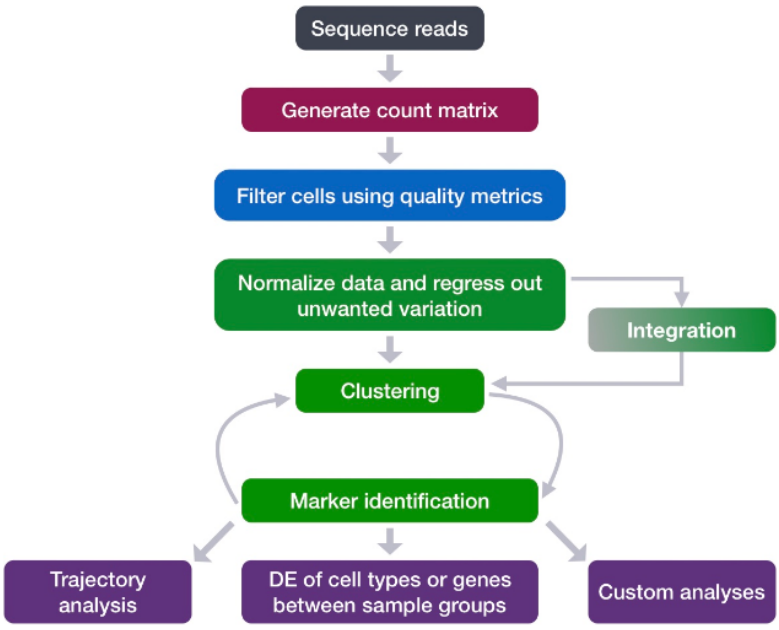


Figure 1: The Steps of Single-cell RNA-seq Analysis, figure extracted from the [Hbc23] tutorial

We based our preprocessing steps on the tutorial developped by [Hbc23], we applied the following protocol :

- Filtering low-quality cells: Cells with a low number of reads or cells with

a high percentage of mitochondrial reads can be filtered out.

- Normalizing the data: The count matrix can be normalized to account for differences in sequencing depth and other technical factors.
- Log-transforming the data: Log-transforming the data can help to normalize the distribution of the data and make it more suitable for statistical analysis.
- Identifying and removing batch effects: Batch effects are technical artifacts that can occur when samples are processed in different batches. Preprocessing can be used to identify and remove batch effects from the data.
- Selecting informative genes: Genes that are not expressed or are expressed at very low levels can be filtered out. Genes that are highly correlated with each other can also be filtered out, as this can reduce the dimensionality of the data without losing much information.

That same tutorial [Hbc23] provided us with insight regarding the evaluation of the quality and reliability of our data. We used the following usual quality criteria :

- Number of reads per cell: The number of reads per cell should be high enough to ensure that the gene expression profile of the cell is accurately captured. A typical threshold is 200 reads per cell.
- Percentage of mitochondrial reads: The percentage of mitochondrial reads should be low, as this indicates that the cell is healthy and that the cytoplasm has been captured effectively. A typical threshold is 10 % mitochondrial reads.
- Number of genes expressed: The number of genes expressed per cell should be high enough to ensure that the cell is a healthy cell and that the gene expression profile of the cell is informative. A typical threshold is 500 genes expressed per cell.
- Cell cycle score: The cell cycle score should be low, as this indicates that the cell is not in the process of dividing. A typical threshold is 0.2.

We disregarded the cell cylce score since our data doesn't include the required information.

To adress the differential expression of genes depending on the stage of lung cancer througout the network reconstruction methods we previously mentionned we must separate the data according to the stage of cancer. The raw file of the count matrix we used in this study has a size of 16 GB. We first separated the data into several datasets according to the stage of cancer however it wasn't sufficiant to obtain datasets of reasonable size to be manipulated in RStudio. Consequently we further splitted the data according to the origin of the sample and the cell type.

2.2 Network reconstruction

The state-of-the-art methods to infer gene regulatory networks (GRNs) from single-cell RNA-seq (scRNA-seq) data can be broadly divided into two categories: [KTK+23]

- Correlation-based methods: These methods identify regulatory relationships between genes based on the correlation between their expression patterns. Correlation-based methods are relatively simple to implement and can be used to infer GRNs from a variety of scRNA-seq datasets. However, they can be sensitive to noise in the data and may not be able to identify all of the regulatory relationships in a GRN.
- Machine learning-based methods: These methods use machine learning algorithms to infer GRNs from scRNA-seq data. Machine learning-based methods can be more accurate than correlation-based methods at identifying regulatory relationships in GRNs. However, they can be more computationally expensive to train and may require more data to train effectively.

GENIE3 figures among the state of the art machine learning algorithms to imput GRN [KTK+23], it is specifically designed to reconstruct GRNs from gene expression data. GENIE3 uses a variety of machine learning techniques, such as support vector machines and random forests, to learn the patterns of regulatory relationships in the gene expression data. This algorithm has the advantage of beeing robust to noise in the data and is able to identify both direct and indirect regulatory relationships. We chose to use this specific algorithm as it has been shown to produce accurate results on a variety of datasets, including scRNAseq data from lung cancer patients [HTIWG10]. I addition to this method we decided to implemant two supplementary algorithms based on different reconstruction methods; a probabilistic method and a constraint based method. The most common probabilistic approach to reconstruct GRN are Bayesian networks [KTK+23], they can model complex regulatory relationships and account for uncertainty in the data. However they can be computationally expensive and may not be as accurate as other methods for small datasets. Bayesian networks can be inferred from gene expression data using a variety of algorithms, such as Markov chain Monte Carlo (MCMC) and variational inference. Bayesian inference has been shown to produce accurate results on diverse datasets, including scRNAseq data [Ngu21]. Finally, we implemented the algorithm ARACNE, a constraint-based algorithm that is specifically designed to reconstruct GRNs from gene expression data. ARACNE uses a set of constraints, such as known regulatory interactions and signaling pathways, to infer the GRN. ARACNE works by iteratively removing edges from the GRN until the constraints are satisfied. ARACNE has been shown to produce accurate results on a variety of datasets, including scRNAseq data [Mar06]. ARACNE is able to reconstruct large and complex GRNs and identify both positive and negative regulatory relationships,it is less computationally expensive ($O(n^3)$)

than Bayesian inference ($O(n^4)$) and GENIE3 ($O(n^6)$) (where n is the number of genes).

We chose to implemant those three methods to compare their performances and dicern within the specific context of our data how well can the different approachs approximate the regulatory relationships between the genes. We wondered if compared to state of the art methods, simpler approaches as ARACNE which are less computationally expensive remind relevant, the question arises since RNA seq data is often quite dense. As we started to consult the state of the art around the identification of the genes responsible for the LUAD progression and metastasis , we observed that, first of all, as far as the RGN reconstruction methods are concerned, although some appear more reccurent than others there is a diversity amongst different papers that displayed good results. The state of the art algorithms are mostly machine learning or correlation-based methods such as SCODE, GENIE3 and SCGenRN [KTK+23]. SCODE is a correlation-based method that uses a variety of correlation metrics to identify regulatory relationships between genes. SCODE is able to infer GRNs from scRNA-seq datasets with as few as 100 cells, it has a complexity of $O(n^5)$. GENIE3 is a machine learning-based method that uses a support vector machine (SVM) classifier to identify regulatory relationships between genes. GENIE3 is able to infer GRNs from scRNA-seq datasets with a variety of different cell types it has a complexity of $O(n^6)$. SCGenRN is a machine learning-based method that uses a graph neural network (GNN) to infer GRNs from scRNA-seq data. SCGenRN is able to infer GRNs from scRNA-seq datasets with a variety of different cell types and conditions, it has a complexity of $O(n^7)$. Those methods have a good accuracy but are computionally expansive to train on large sets of data. The reconstruction of GRN from sRNA-seq usually involves a dense set of genes, consequently we aspired to evaluate if the use of methods, less computationally expansive can be a viable option in comparison with state of the art algorithms.

We conducted a comparative study of the results obtained from previously mentionned methods on our dataset. One common approach is to compare the network topologies. The network topology is the structure of the network, and it can be characterized by a variety of metrics, such as the network density, the network clustering, and the network centrality. Another approach to comparing the results from different network reconstruction methods is to compare the predicted gene regulatory relationships. This can be done by comparing the correlation coefficients between genes, or by comparing the mutual information between genes [Ngu21]. We applied both approaches, on the one side we performed an analysis to compare the performances of the differents approaches we used, on the other side we compared our resulting networks between the different stages of cancer for several cell type and cell origin to identify the differences and consequently the genes we can suppose to be instigating the LUAD progression and metastasis. We additionnally compared the genes identified by our different networks with the genes we identified as differentially expressed across cancer stages via statistical testing.

2.2.1 MIIC : Bayesian networks

MIIC.curie is a web server that reconstructs causal or non-causal networks from non-perturbative data. It is freely accessible at https://miic.curie.fr. MIIC works by first removing dispensable edges from a fully connected network. It then filters the remaining edges based on their confidence assessment or orients them based on the signature of causality in observational data. MIIC can be used for a broad range of biological data, including possible unobserved (latent) variables, from single-cell gene expression data to protein sequence evolution.

2.2.2 GENIE3

GENIE3 (Gene Network Inference with Ensemble of Trees 3) is a machine learning algorithm used for infering the gene regulatory network from gene expression data, like scRNAseq. It is then hightly appropriate to use on our dataset.

Its process is quite simple [HTIWG10]: first it takes an expression matrix, then this data is fitted in Random Forests, an ensemble of random decision trees that are combined to obtain the best accuracy. GENIE3 can then extimate each variable importance scores derived from Random forests to identify the regulators of each target gene.Based on this feature, it can construct a network where each edge is the predicted regulatory interaction between genes. We used in our code the GENIE3 R package directly.

2.2.3 ARACNE

ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [Mar06] is another algorithm used for the inference of gene regulatory networks from gene expression data. Like GENIE3, ARACNE aims to identify regulatory relationships between genes based on their expression patterns. However, it uses a very different approach.

Where GENIE3 will rather determine an importance for each gene which will be used for infering regulatory interaction, ARACNE will first need to calculate an Mutual Information Matrix (MIM), which can be done using several approaches. Mutual information measures the statistical dependence between two variables, in this case expression levels between two genes. In a first computation, we tried using for this step the Jaccard distance, as it is done in our reference for pre-processing [KKL$^+$20], but as our results were not satisfying (see discussion), we decided to switch to Spearman distance to compute the MIM.

Once the MIM computed, the algorithm will use a statistical test, the Data Processing Inequality (DPI), to filter indirect relationships. Finally, the processed matrix will be used to construct a network with the remaining direct regulatory relationships.

Use of Spearman Correlation We chose to use Spearman's rank correlation coefficient for distances in the MIM as it has been proven more accurate and precise than other methods when used with ARACNE [SAAM16]. Spearman correlation is a non-parametric measure of statistical dependence between

two variables. Unlike other approaches, it captures both linear and non-linear associations, and as it is non-parametric, it doesn't assume a distribution for the data, which explain its good performance. Spearman measures the strength and direction of the relationship between the variables, here the genes, according to their rank, e.g. their position in the sorted list of all values. It is particularly usefull when dealing with non-linear relationships in data, which is the case of scRNAseq.

3 Results

In order to conduct our analysis, within our preprocessing steps, we included the selection of the most significatively differently expressed genes between normal and tumoral cells for each cancer stage. We ended up with around 50 genes for each stage, using the t-test.

3.1 Network analysis

Gene	Degree
DNAAF1	18
RSPH1	18
C5orf49	24
CFAP126	26
C1orf194	28
HOPX	28

(a) Stage 1

Gene	Degree
CLDN3	6
LMO7	4
C11orf96	6
CFI	4
AQP1	18
DRAM1	22

(b) Stage 2

Gene	Degree
FAM183A	18
RP11.356K23.1	8
C2orf40	6
DYNLRB2	6
C12orf75	6
LRRC46	6

(c) Stage 3

Table 1: Genes with the best degree centrality at each stage

Gene	Closeness
C20orf85	0.006451613
HOPX	0.006493506
SFTPC	0.006622517
C1orf194	0.006756757
FABP6	0.006802721
SFTA2	0.006849315

(a) Stage 1

Gene	Closeness
CAV2	0.004132231
C11orf96	0.004149378
ETV1	0.004273504
GPRC5A	0.004291845
FABP5	0.004329004
LAMP3	0.004975124

(b) Stage 2

Gene	Closeness
FAM183A	0.007142857
MALL	0.007633588
RSPH1	0.007692308
HOPX	0.008064516
C9orf24	0.008928571
AGR3	NaN

(c) Stage 3

Table 2: Genes with the best closeness centrality at each stage

Gene	Betweenness
TFPI	155.0070
C20orf85	163.1470
CAV2	166.1375
AGER	175.4422
C1orf194	248.0922
CCDC170	255.4450

(a) Stage 1

Gene	Betweenness
LRRK2	277.2741
C16orf89	288.8477
GPRC5A	313.6095
LPCAT1	373.2273
ABCA3	413.9590
LAMP3	1404.1471

(b) Stage 2

Gene	Betweenness
MALL	317.5216
FXYD3	413.7867
FAM183A	451.7953
SFTPC	467.2135
HOPX	805.9024
C9orf24	1856.6789

(c) Stage 3

Table 3: Genes with the best betweeness centrality at each stage

3.1.1 Degree centrality

Degree centrality is a measure of how many connections a node has to other nodes in the network. A node with a high degree centrality is well-connected and has a lot of influence over other nodes in the network. In the context of gene expression networks, a gene with a high degree centrality is likely to be involved in many different biological processes and may play a key role in the regulation of other genes.

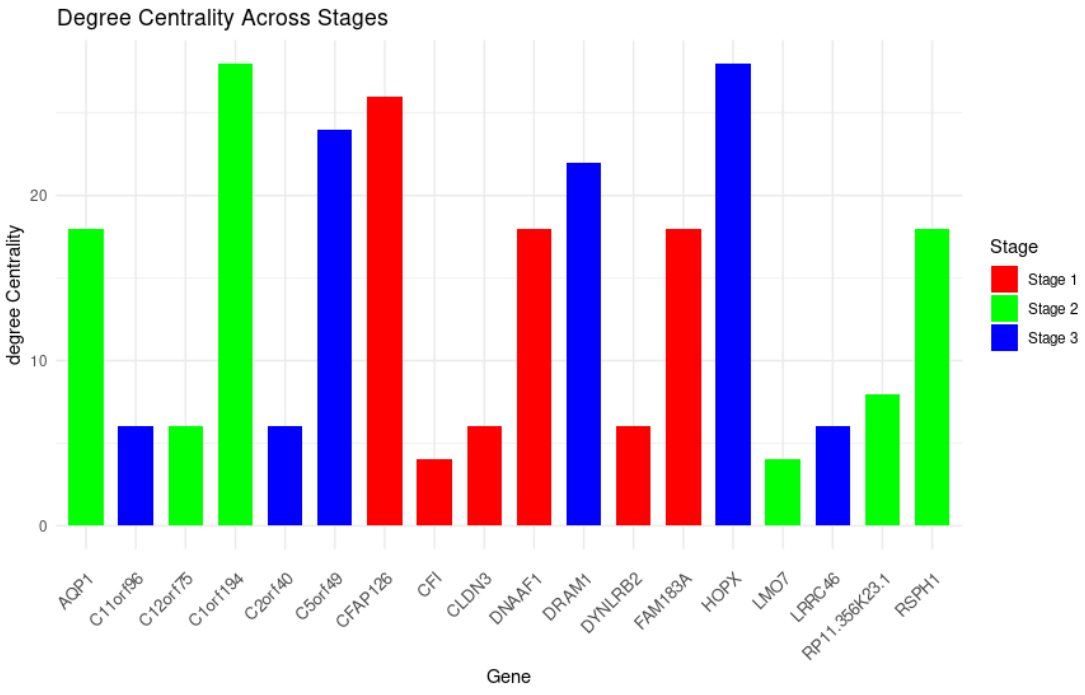


Figure 2: highest degree centrality across stages

3.1.2 Closeness centrality

Closeness centrality is a measure of how close a node is to all other nodes in the network. A node with a high closeness centrality is close to all other nodes in the network and can easily exchange information with them. In the context of gene expression networks, a gene with a high closeness centrality is likely to be involved in many different biological processes and is well-positioned to coordinate the activity of other genes.

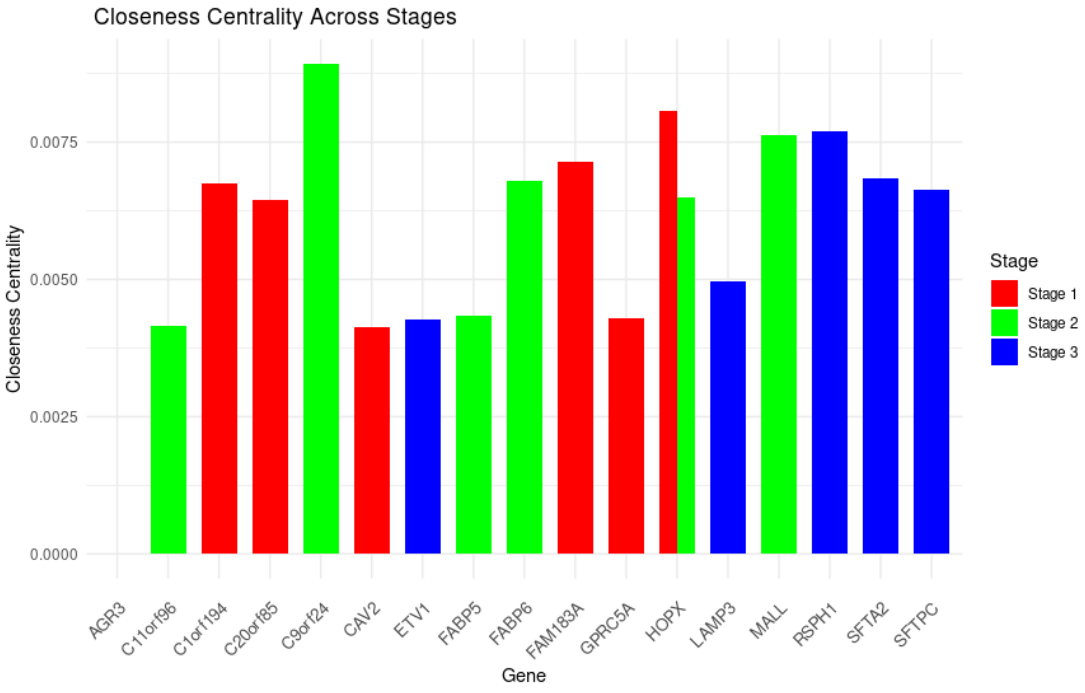


Figure 3: highest closeness centrality across stages

3.1.3 Betweenness centrality

Betweenness centrality is a measure of how often a node acts as a bridge between two other nodes in the network. A node with a high betweenness centrality is a key node in the network and controls the flow of information between other nodes. In the context of gene expression networks, a gene with a high betweenness centrality is likely to be involved in many different biological processes and is well-positioned to regulate the activity of other genes.

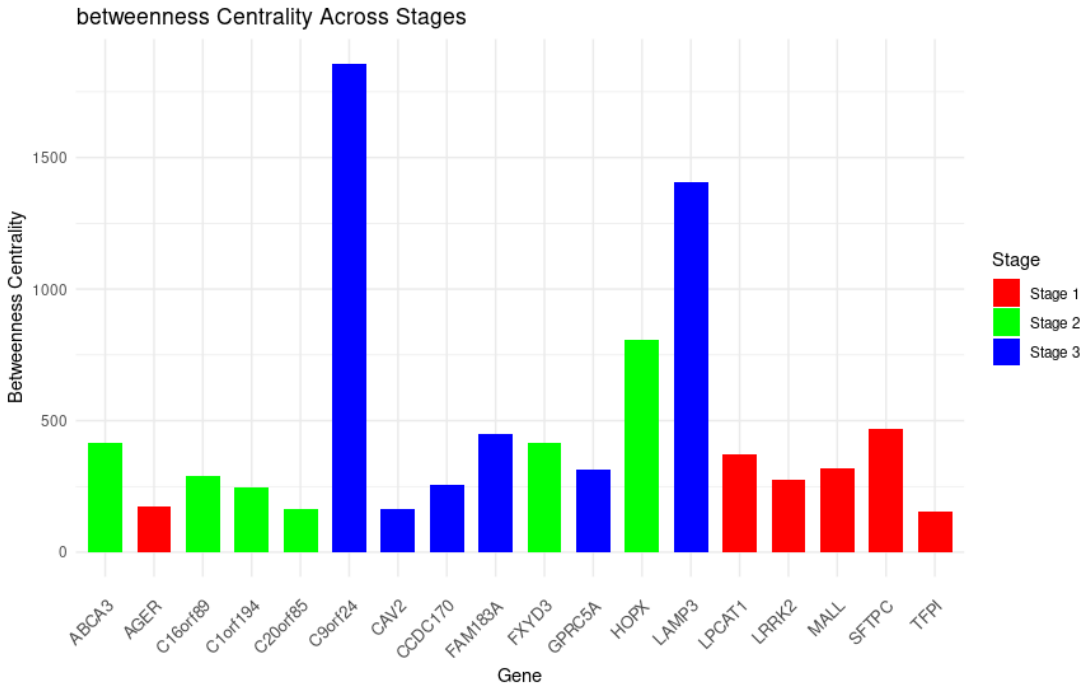
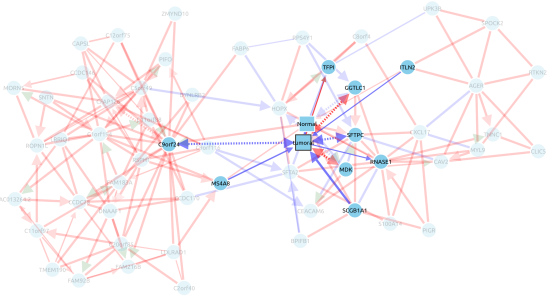


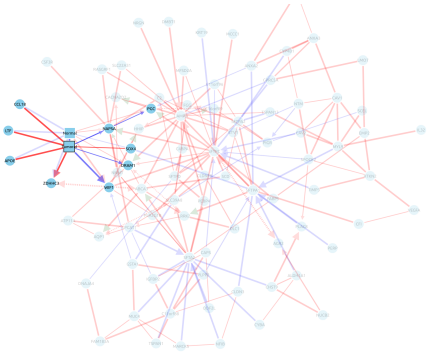
Figure 4: highest betweenness centrality across stages

3.1.4 Overall interpretation

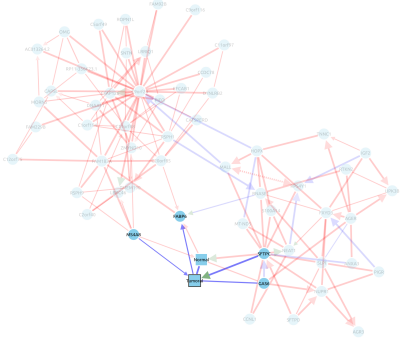
The genes that have the highe st degree centrality, closeness centrality, and betweenness centrality in our networks are likely to be key genes that are involved in the regulation of other genes and play a central role and so it is interesting to annalyse wether they correspond to genes that have a role in the tumorigenesis process and progression.



(a) Stage 1



(b) Stage 2



(c) Stage 3

Figure 5: MIIC Networks at each stage highlighting genes related to tumoral state

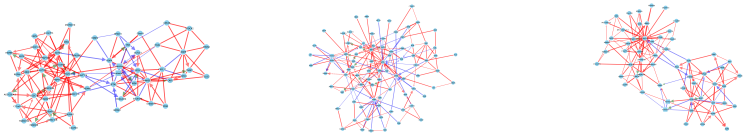
Table 4: Genes that are central and related to the tumoral state

Stage	**Genes**
Stage 1	SFTPC, TFPI
Stage 2	DRAM1, WIF1
Stage 3	MS4A8, FABP6, GAS6

We observe that a few of the most central genes in our networks for each stage are indeed related the tumoral state, the fact that, in each stage, the genes that appear related to cancer differ was a surprising result. We expected some form of continuity but the evolution of the influence of cancer on gene expression across the different stages appears to be chaotique. We cannot make the affirmation that the fact that the centrality of genes varies a lot across different stages of cancer is due to cancer from our networks, however the fact that some of the genes we found to be most likely related to cancer are central genes is an invitation to consider that question in futur work.

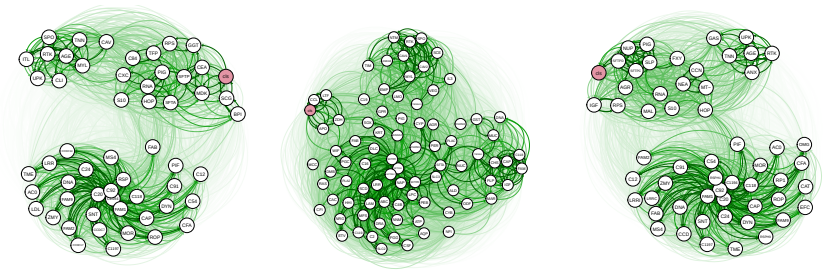
Another important observation that came as a surprise is that we distinguish on our graphs two main clusters for the stages 1 and 3 (see figures 6, 7 and 8), however for the stage two there appear to be only one cluster. This observation was made across all three of the network reconstruction algorithms we used. We also computed the distributions of our different measures of centrality of the genes to have a better idea of what goes on within our networks 9.

3.2 Comparing several approaches



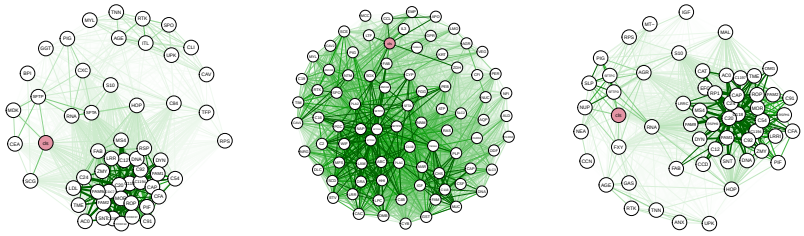
(a) Stage 1 (b) Stage 2 (c) Stage 3

Figure 6: MIIC Networks at each stage



(a) Stage 1 (b) Stage 2 (c) Stage 3

Figure 7: GENIE3 Networks at each stage



(a) Stage 1 (b) Stage 2 (c) Stage 3

Figure 8: ARACNE Networks at each stage

3.2.1 Identifying clusters

We were aiming to asses if algorithms like ARACNE and bayesian network, which are less computationally costly than GENIE3, perform well enough to promote their use for sRNA-seq data. From our results it seems that ARACNE doesn't perform very well and fails to infer clusters within the genes. In contrast MIIC appears like a good candidate, however it does seem less rafined than

GENIE3, indeed GENIE3 allows for a better visualization of existing cluster of regulation between the genes.

Comparing the networks we have computed, we can see that a global structure seems to be preserved throughout the different algorithms.

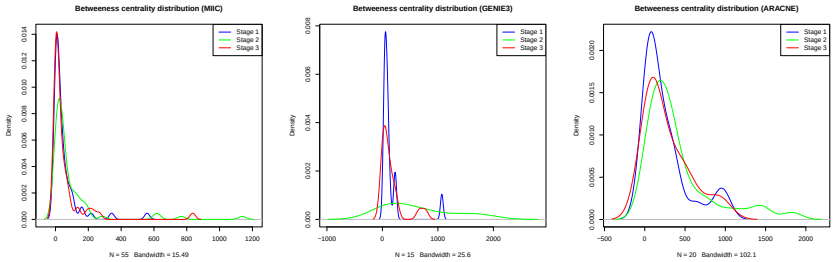
In stage 1, we can identify one cluster with hightly connected vertexes, and two second small ones, with less connectivity, but closer together. This can be very well observed in GENIE3 network, with MIIC as well, but less with ARACNE, where a single hightly connected cluster can be seen.

In stage 2, we obtained for any approach a very concentrated network. Similarly to stage 1, GENIE3 offers a supplementary insight of stage 2 where although the overall network has a spherical form we can actually distinguish four clusters.

Finally in stage 3, we found figures close to the first stage.

From a visual analysis, we can say that GENIE3 performs better, as it allows us to apprehend the global shape of the network. This is due to the fact that most edges are kept by the algorithm, resulting in very low-weight edges that give contrast to the figure. This effect would probably disapear if GENIE3 was given a threshold, or a more specific list of genes to consider. However, we point out the fact that this abundance of edges is not working in favor of precision, for it is less likely to see which interactions are important to consider than with the two other approaches.

3.3 Centrality characteristics



(a) Betweeness distribution from MIIC (b) Betweeness distribution from GENIE3 (c) Betweeness distribution from ARACNE

(d) Closeness distribution from MIIC (e) Closeness distribution from GENIE3 (f) Closeness distribution from ARACNE

(g) Degrees distribution from MIIC (h) Degrees distribution from GENIE3 (i) Degrees distribution from ARACNE

Figure 9: Betweeness centrality, closeness and degree centrality distributions for each stage and each algortithm

As we said earlier, in order to understand better the differences between the approaches, we computed the distribution of the different parameters. For MIIC and GENIE3 the Betweenness distribution is similar for the stages 1 and 3 and quite different for stage 2. For MIIC, GENIE3 and ARACNE the Degrees distribution is similar for the stages 1 and 3 and quite different for stage 2. For MIIC and GENIE3 the Closseness distribution is similar for the stages 1 and 3 and quite different for stage 2.

Overall it seems like ARACNE strugles more to identify the clusters, its results are quite different from GENIE3 and MIIC it was apparent on the networks as well but plotting the centrality measures distribution allowed us to refine our analysis. The pattern of the two disctinct clusters for stages 1 and 3 are most evident when we look at closeness distribution and observe two peaks.

3.3.1 Genes identified as central in regulation network

Stage	Genes
Stage 1	**SFTPC***, **SCGB1A1**, **MDK**, GGTLC1, CEACAM6
Stage 2	**CCL18**, **ZDHHC3**, **LTF**, **APOE**, **WIF1***
Stage 3	**SFTPC**, **SLPI**, **NUPR1**, **PIGR**, SFTPD

(a) GENIE3

Stage	Genes
Stage 1	**SCGB1A1**, **SFTPC***, **MDK**, MS4A8, C20orf85
Stage 2	**CCL18**,**LTF**, **ZDHHC3**, **APOE**, **WIF1***
Stage 3	**SFTPC**, **SLPI**, **NUPR1**, **PIGR**, FXYD3

(b) ARACNE

Table 5: Genes that are hightly related to the class (tumoral or normal) in ARACNE and GENIE3 networks. Genes noted with * are also found in MIIC as hightly related to tumoral lung. Genes in bold are common to GENIE3 and ARACNE.

We were abble to produce list of the five gene per stage most related to class vertex in networks from ARACNE and GENIE3, in order to compare them to those identified in MIIC. As they are not cross-validated with centrality, they are mainly a way to see the proximity between GENIE3 and ARACNE. We have seen before that ARACNE seemed way off considering centrality distribution and global structure. However, here we see that it doesn't perform pourly when identifying mutual information between the class and the genes. We can see in those tables that most genes from GENIE3 are also found in ARACNE, in closely similar order. Morever, both those algorithms were able to identify genes found in MIIC networks analysis, like SFTPC, even though we did not process their centrality.

We can then say that, even though ARACNE characteristics seem to be laking precision, we can suppose that an analysis of precise interection would give results close to what we found with MIIC.

3.4 Comparing to biological data

The second aim of this project was to identify the genes that are likely to play a role in lung cancer progression. We identified different genes for each stage keeping in mind that "relation isn't cause" we inferred genes that appear to be related to lung cancer progression we can't affirm that they are the cause for it. In order to evaluate if our findings are coherent we compared them to previous works on gene regulation networks for lung cancer. Here is a list of genes that have been linked to the progress of lung cancer [MRBP23]:

- EGFR (Epidermal Growth Factor Receptor): EGFR mutations are common in non-small cell lung cancer (NSCLC).
- KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog): KRAS mutations are found in various cancers, including lung adenocarcinoma.
- ALK (Anaplastic Lymphoma Kinase): ALK rearrangements are often seen in non-small cell lung cancer.
- ROS1 (ROS proto-oncogene 1): ROS1 rearrangements are another targetable alteration in lung cancer.
- BRAF (B-Raf Proto-Oncogene): BRAF mutations are associated with some cases of lung cancer.
- MET (Mesenchymal-Epithelial Transition Factor): MET amplification and overexpression are observed in lung cancer.
- LKB1 (Liver Kinase B1): LKB1 alterations are found in a subset of lung adenocarcinomas.
- PIK3CA (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha): PIK3CA mutations can be present in lung cancer.
- MYC (MYC Proto-Oncogene, BHLH Transcription Factor): MYC amplification is associated with aggressive forms of lung cancer.
- CCND1 (Cyclin D1): Overexpression of CCND1 is observed in some lung cancers.
- SOX2 (SRY-Box Transcription Factor 2): SOX2 amplification or overexpression is seen in lung squamous cell carcinoma.
- BCL2 (B-Cell CLL/Lymphoma 2): BCL2 expression may play a role in lung cancer survival.
- CDKN2A (Cyclin-Dependent Kinase Inhibitor 2A): CDKN2A alterations are associated with increased lung cancer risk.
- TP53 (Tumor Protein P53): TP53 mutations are common in various cancers, including lung cancer.

- RB1 (Retinoblastoma 1): RB1 alterations are associated with small cell lung cancer.
- ERCC1 (Excision Repair Cross-Complementation Group 1): ERCC1 expression is related to response to chemotherapy in lung cancer.
- DHFR (Dihydrofolate Reductase): DHFR alterations may be implicated in lung cancer.
- RASSF1A (Ras Association Domain Family Member 1A): RASSF1A is a tumor suppressor gene frequently silenced in lung cancer.
- MGMT (O-6-Methylguanine-DNA Methyltransferase): MGMT methylation status is relevant in lung cancer prognosis.

Here is the list of genes we infered to be related to lung cancer :

1. **C9orf24:** Limited information is available on its direct association with lung cancer.
2. **MS4A8:** The MS4A gene family has been implicated in various cancers, and its role in lung cancer might require further investigation.
3. **SCGB1A1:** This gene is associated with the production of secretoglobins, and alterations in its expression have been observed in lung cancer. It is often used as a marker for lung adenocarcinoma.
4. **RNASE1:** Limited information is available on its direct association with lung cancer.
5. **SFTPC:** Surfactant Protein C is primarily associated with lung function and surfactant production. Mutations in this gene can be linked to interstitial lung diseases.
6. **ITLN2:** Limited information is available on its direct association with lung cancer.
7. **TFPI:** Mainly known for its role in regulating blood coagulation. While it may have some implications in cancer-related thrombosis, its direct link to lung cancer is not as prominent.
8. **CGTLC1:** Limited information is available on its direct association with lung cancer.
9. **MDK (Midkine):** This gene has been reported to be overexpressed in lung cancer and is associated with tumor progression.
10. **NAPSA:** Limited information is available on its direct association with lung cancer.
11. **PGC (Progastricsin):** Overexpression of PGC has been associated with lung cancer.

12. **DRAM1:** Involved in autophagy regulation. Dysregulated autophagy is associated with cancer, and DRAM1's role in lung cancer may require further investigation.

13. **WIF1 (WNT Inhibitory Factor 1):** Acts as an antagonist of the Wnt signaling pathway. Aberrant Wnt signaling is associated with various cancers, including lung cancer.

14. **CL18:** Limited information is available on its direct association with lung cancer.

15. **LTF (Lactotransferrin):** Altered expression of LTF has been observed in lung cancer.

16. **APOE:** Limited information is available on its direct association with lung cancer.

17. **ZDHHC3:** Limited information is available on its direct association with lung cancer.

18. **SOX4:** Overexpression of SOX4 has been associated with lung cancer progression.

19. **FABP6:** Involved in fatty acid transport and metabolism. Dysregulated lipid metabolism is associated with cancer.

20. **GAS6 (Growth Arrest-Specific 6):** Implicated in promoting cell survival and proliferation. Dysregulation could contribute to cancer progression, including lung cancer.

A part of our findings are coherent with existing litterature, the rest has yet to be verified and could potentially constitute new discoveries.

4 Discussion

4.1 Use of Jaccard distance

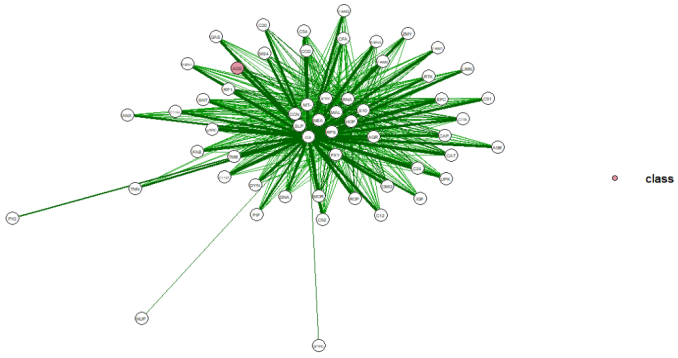


Figure 10: Result from mutual information matrix computed with Jaccard distance in ARACNE at stage 3

We tried during our project to comply with our reference article [KKL+20] for pre-processing as best as we could. However, as we tried to use Jaccard distance, it seemed that this method did not perform well with our datas, as the result network was nowhere close to the other algorithms results. There are a lot of explanation of why this mutual information calculation method did not work, but it is most likely that our process of pre-selection for the genes composing the network was too different from the one followed by Kim et al. [KKL+20] to use this tool.

4.2 Generating networks with target genes

During the process of selecting the genes, we could have singled-out a few genes to construct networks around of, which is a classical operation performed by GENIE3 for instance, instead of including our entire processed dataset in each network. Maybe this would have helped us to single out more easily the central genes in cancer regulation.

4.3 Data cutoff

The process of selection of a restricted number of genes throughout the project did at some point lacked a reference. When selecting the genes after the pre-processing with a Student test adjusted with Bonferroni adjustement, we applied the same thresholds as Kim et al. [KKL+20], but ended up with a bigger dataset than expected. We had over at least 2000 genes per stage, when they found less than a hundred. This is possibly due to our spliting of the data matrix upstream.

We chose to select the 100 genes with the hightest absolute log2FC, when they only recomanded to keep data with absolute log2FC ¿ 1, in order to get a list of the most significative genes for each stage.

4.4 Why is stage 2 so different ?

We observed throughout our results that a stage 2 network was very different from the other stages. We though of several explanations. Maybe this comes from the fact that this is a stage at which cancer starts to be localy impacting, which means that a lot of interaction are happening and explains a very hight centrality, before that at stage 3, when the metastasis start to spread to ganglions, the initial tumour might slow its generative activity. Maybe the simple explanation is that we have selected more genes for this stage (87, which is more than the other two stages, bith around 50), which explains the abundance of interaction. It also means that more various genes were statisticaly important at this stage. This question would require further analysis than what we could do in this project.

4.5 Why does ARACNE seems to performs pourly ?

We noticed that ARACNE's networks were a little off when looking at the parameters distributions. This maybe due to several facts.

First, it is possible that the quality of our datas was not optimal, which might have been corrected by algorithms using inference like GENIE3 and MIIC, but not by ARACNE, which is rather sensitive to noise and variability. Its use is maybe not optimal. We've maybe not evaluated enough the quality of our datas.

Second, it is possible that, even though we chose Spearman correlation for its non-linearity, ARACNE could not outperform GENIE3 and MIIC on this point as it is firstly a linear algorithm.

Moreover, ARACNE concentrates mainly on relationships between a couple of genes, and does not apprehend multivariate relationship, which is very likely to occur in a gene interaction network.

Finally, we put this algorithm which is remarkable for its simplicity in competition with two inference methods more fitted for this kind of data, which explains why, when compared, it performs more pourly.

5 Conclusion

Throughout this project, we have tried to answer two questions.

First, we asked which algorithm would perform best with our dataset. It appeared that GENIE3 was the most appropriate, even though we couldn't counduct further analysis of the genes regulating lung cancer with it as we did for MIIC. We noticed that without precise cutoffs in the resulting networks, it is hard to visualy analyse precise interactions. The main advantage of MIIC over GENIE3 is its maniability, because of its online interface, and its ability

to concentrate on the main interactions. This should also have been the main advantage of ARACNE, but we did not get convicing results with it as we detailed earlier. We did get close results when computing the main genes in interaction whith all the approaches. We would recommand using MIIC or GENIE3 with our dataset. An approach with ARACNE might be possible but would require further analysis of the mutual information estimation.

The second question was to identify genes with a key role in lung cancer at different stages with the hope to point out an evolution. We did sort out most of the genes based on their statistical significance in the dataset upstream, and were then able to find key genes in the regulation networks between normal and tumoral lung calculated with MIIC. We could afterward put this genes in relation with their function and did find in the literature that most of those identified genes were in fact known to be linked to cancer, like WIF1 (which we did also find with GENIE3 and ARACNE networks). But some genes are not directly linked to cancer; they would require further investigation that we could not proceed here.

We found surprising that throughout the different stages of cancer, we did not observe the same genes at the center of our networks, for we supposed that the evolution would appear as a more intense interactions between a maintained set of genes. Although this is what we can observe between stage 1 and 3, we could not investigate further stage 2 as we discused earlier. This question would be interesting to investigate further.

References

[FSS+21] Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Laversanne Torre, and Ahmedin Jemal. Cancer statistics for the year 2020: An overview. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[Hbc23] Hbctraining. Single cell rna-seq data analysis workshop, 2023.

[HTIWG10] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 5(9):1–10, 09 2010.

[Hua17] et al. Huang, Hsin-Yi. Network-based gene expression analysis identifies key regulators of lung cancer immune response. *Oncotarget*, 8(13):22095, 2017.

[KKL+20] N. Kim, H. K. Kim, K. Lee, Y. Hong, J. H. Cho, J. W. Choi, J. I. Lee, Y. L. Suh, B. M. Ku, H. H. Eum, S. Choi, Y. L. Choi, J. G. Joung, W. Y. Park, H. A. Jung, J. M. Sun, S. H. Lee, J. S. Ahn, K. Park, M. J. Ahn, and H. O. Lee. Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature communications*, 11(1):2285, 2020.

[KTK+23] Daniel Kim, Andy Tran, Hani Jieun Kim, Yingxin Lin, Jean Yee Hwa Yang, and Pengyi Yang. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *npj Systems Biology and Applications*, 9(1):51, 2023.

[LYK+13] Sang Won Lee, Hyo-Jeong Yoo, Jung-Eun Kim, Jae Ho Jang, Hyun-Jin Kim, Myung-Jin Kim, Woo-Jung Choi, Eun-Kyoung Park, Mi-Jeong Kim, Ji-Yeon Kim, Hyun-Ju Park, Sang-Won Kim, Jong-Won Lee, Young-Hoon Kim, Seung-Woo Jung, Seong-Wook Kim, Byung-Soo Kwon, and Hyuck-Jin Kwon. A network analysis of gene expression data reveals key regulators of lung cancer progression. *PLoS One*, 8(4):e62553, 2013.

[Mar06] et al. Margolin, Aaron A. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7:S7, 2006.

[MRBP23] Carlos Martínez-Ruiz, James R. M. Black, and Puttick. Genomic–transcriptomic evolution in lung cancer and metastasis. *Nature*, 616(7957):543–552, 04 2023.

[Ngu21] et al. Nguyen, Hung. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in bioinformatics*, 22(3):bbaa190, 2021.

[SAAM16] Ali Gorji Sefidmazgi, Fatemeh Ahmadi-Abkenari, and Seid Abolghasem Mirroshandel. Correlation analysis as a dependency measures for inferring of time-lagged gene regulatory network. pages 6–11, September 2016.