Statistical Impact of Gene Expression Data Generated by Three Measurement Platforms on Gene Regulatory Network Reconstruction Methods

Supervised by Professor Andy Lynch

University of
St Andrews

Qing Zhang

Master of Science in Statistics

School of Mathematics and Statistics

University of St Andrews

Aug 2022

Declaration

I certify that this dissertation report has been written by me, is a record of work carried out by me, and is essentially different from work undertaken for any other purpose or assessment.

Qing Zhang

Acknowledgement

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to thank my supervisor Professor Andy Lynch, for his many constructive and heuristic suggestions for my research. Your profound and professional feedback made my thinking more acute and brought my work to a higher level.

Additionally, I would like to thank my parents for their understanding and encouragement. You were always there for me.

Finally, a big word of thanks to Jean and Lu for being my amazing friends who provided happy distractions to rest my mind outside of my research.

Table of Contents

List of Figures

List of Tables

Glossary

DNA – Deoxyribonucleic acid

RNA – Ribonucleic acid

TF – Transcription Factor

TG – Target Gene

PCa – Prostate Cancer

AR – Androgen Receptor

NGS – Next-generation Sequencing

RNA-seq – RNA sequencing

DEG – Differentially Expressed Gene

DAVID – DAVID Bioinformatics Resource

ROC – Receiver Operating Characteristic

AUC – Area Under ROC Curve

GEP – Gene Expression Profile

GEM – Gene Expression Matrix

CRPC – Castration-resistant Prostate Cancer

Camcap – Data generated by microarrays

TCGA – Data generated by RNA-seq

ProExp – Protein Expression Data

FDR – False Discovery Rate

FC – Fold Change

MI – Mutual Information

ARACNE – Algorithm for the Rereconstruction of Accurate Cellular Networks

MRNET – Minimum Redundancy Networks

CLR – Context Likelihood or Relatedness network

C3NET – Conservative Causal Core Network

TP – True Positive

FP – False Positive

FN – False Negative

FP – False Positive

Abstract

Background. Gene expression data are the main input for reconstructing gene regulatory networks. However, depending on the type and quality of the input data, reconstructing coexpression-based gene regulatory networks (GRNs) from gene expression data can remain a difficult task.

Objectives. Our main goal was to summarise the statistical impact of different types of gene expression data for the process of GRN reconstructions in prostate cancer (PCa) studies.

Research Questions. (1) Can some of the investigated AR-interacting transcription factors be enhancers or inhibitors of PCa based on an examination of GRNs from our data? (2) How different are the pre-processing methods for our data? (3) How different are GRN analyses for the different data types? (4) Which GRN algorithms are the best performing models for different types of data?

Methods. We used three types of gene expression data: RNA microarray data, RNA-seq data and protein expression data. Based on the mutual information tools of coexpression-based GRN algorithms, we discussed differences throughout the entire transcriptomic analysis process.

Conclusions. RNA-seq data offer certain advantages over RNA microarray data and protein expression data in GRN reconstruction methods.

Keywords. Gene Regulatory Networks, Prostate Cancer, Microarray, RNA-seq, Protein

1 Introduction

DNA is a hereditary material, whilst genes are DNA sequences containing genetic information that control the traits of an organism. Genes enable the expression of genetic information in an organism through transcription and translation through a process known as gene expression. In living cells, genes produce a particular type of protein—known as a transcription factor (TF)—through the process of gene expression. During the transcription of a gene, a group of TFs act on the gene's promoter region to control the transcription of the target gene (TG), with these transcription factors being the products of other genes. In other words, the gene products produced during the expression of one gene affect the expression of other genes. Therefore, genes interact with each other, whilst genes and gene products also interact with each other. Throughout the entire process of gene expression, this interplay is a complex mechanism of gene regulation involving the regulation of DNA replication, transcriptional regulation and translational regulation.

Gene regulatory networks (GRNs) are a mapping of gene regulatory mechanisms that can graphically describe the complex relationships between genes and genes or between genes and gene products. In a GRN, nodes usually represent genes, TFs or miRNAs, whilst edges represent regulatory relationships. In this study, we built TF-gene GRNs and gene-to-gene GRNs to understand how TF-TG pairs and gene-to-gene pairs interact.

Human disease is linked to human genes. For example, cancers are caused by genetic mutations. Throughout a person's life, DNA is constantly subjected to cancer-inducing substances and at risk of self-replication errors. However, in extreme cases, changes in DNA sequences may lead to changes in the function of some genes, thereby causing uncontrolled cell proliferation. Therefore, GRNs may help to uncover pathogenic genes in cancers and design novel treatments[1].

Prostate cancer (PCa) is the most commonly diagnosed cancer among men in the Western world[2]. Notably, androgen receptor (AR) plays an important role in the development of PCa[3]. AR recruits various co-regulatory factors to activate the target genes (TGs) and then regulate gene expression. Additionally, several studies have noted that AR also interacts with several TFs, such as FOX-family[4], OCT1[5], GATA2[6], p53[7], NF1[8], AP1[9], IRF1[10] and STAT3[11] TFs. These AR-interacting TFs can be prognostic factors in PCa and might lead to the development of a new therapeutic intervention for PCa. Thus, our first research question is presented as follows: based on our examination of GRNs from our three types of data, can some of the investigated AR-interacting TFs be enhancers or inhibitors of PCa?

With the successful completion of the Human Genome Project and the increasing development of next-generation sequencing (NGS) technologies, we can access 99% of human genes and gene expression data using high-throughput technologies. The existing technologies are based on two assumptions, the first of which is that the level of gene transcription reflects the outcomes of gene regulation. This hypothesis implies that gene regulation primarily occurs at gene transcription. The second assumption is that only those genes associated with a trait will be expressed, whilst unrelated genes will not be expressed. Hence, based on these assumptions regarding NGS technologies, gene expression data can serve as an input for GRNs.

The gene expression data generated by high-throughput transcriptomic platforms include RNA microarray and RNA sequencing (RNA-seq) data. The microarray platform uses a set of short expressed sequence tags generated from the DNAs to hybridise target RNAs and then generate raw intensity, which is used to measure the RNA expression level of RNA[12]. The RNA-seq platform sequences the DNA generated from RNA and then directly generates raw counts to represent the RNA expression level[13]. Thus, the principles and experiments of the two technologies differ. Hence, the types of inputs, the number of detected genes, the number of

experiments and statistical pre-processing methods for RNA microarray data and RNA-seq data vary drastically. Thus, the second research question is presented as follows: How different are the pre-processing methods for our data?

Over the last few decades, many methods for reconstructing GRNs from gene expression data generated by high-throughput technologies have been proposed. However, reconstructing GRNs from expression data based on co-expression is a very difficult task. Each GRN reconstruction method has its own unique advantages and challenges depending on the type and quality of input data[14]. In the present study, we compared three types of gene expression data: RNA microarray data, RNA-seq data and protein expression data. Also, we used the GRN algorithms based on the mutual information tools for each data set. As such, the third research question is presented as follows: How does network analysis differ when using different types of gene expression data (i.e., RNA microarray data, RNA-seq data and protein expression data)? We mainly focused on statistical differences in data input, the workflow of algorithms and performance assessment of algorithms for these three data types.

Additionally, obtaining reliable gene networks from gene expression data remains an unsolved problem. The fourth research question is presented as follows: Which GRN algorithm and data type can provide the best performance model to make our inferences and analyses more reliable?

Depending on the aforementioned differences in the entire analysis, our main goal is proposed as follows: to summarise the statistical impacts of three types of gene expression data (i.e., RNA microarray data, RNA-seq data and protein expression data) for the process of GRN reconstructions.

2 Brief Discussion of the Research Questions

2.1 List of Research Questions

Our research questions are listed in Table 1. The first research question relates to a biological inference, whilst the remaining three research questions are statistical comparisons.

Table 1. List of Research Questions

1. Based on our examination of GRNs from our three types of data, can some of the investigated AR-interacting TFs be enhancers or inhibitors of PCa?
2. How different are the pre-processing methods for our data?
3. How does network analysis differ when using different types of gene expression data?
4. Which GRN algorithm and data type can provide the best performance model to make our inferences and analyses more reliable?

2.2 Plans to Address the Research Questions

For our first research question, we first analysed the differentially expressed genes (DEGs) of our three data types. The goal of our DEG analysis was to extract genes with significantly different expression levels under tumour/normal conditions and to identify whether the DEGs were up- or down-regulated. After we obtained the sets of DEGs, we performed functional

enrichment analysis. The analysis of DEGs cannot capture the functional implications of DEGs, which is why we then performed the analysis of functional enrichment. There are many methods to achieve functional enrichment. For this study, we used DAVID[15] with DEG lists to predict which TFs the DEG lists were enriched in. The workflow for TF-TG GRN reconstruction is presented in Chapter 4.2.1. Using this workflow, we could determine the TF-TG pairs that are nodes for GRNs. To address our research question, we selected the AR-interacting TFs mentioned in Chapter 1 as source nodes for GRN, the DEGs as target nodes and the interactions between TFs and DEGs as edges. Thus, we built unweighted TF-DEG GRNs in Cytoscape software to answer our first research question. To further address our first research question, we also checked whether one type of data offers advantages over the other types in the investigation of unweighted GRNs. The entire discussion of answers is in Chapter 5.1.

For our second research question, we will answer three parts of the differences: differences in data input, data pre-processing, and the analysis of DEGs (since DEG analysis is a preparation step for GRN reconstruction). The differences in data (inputs and pre-processing) are discussed in Chapter 3.5. The tools and results of DEG analyses can differ by data type, which is discussed in Chapter 4.1.3. The entire discussion of answers is in Chapter 5.2.

For our third research question, we mainly focused on the statistical impacts described in Chapter 1. Regarding data input, some GRN algorithms may require different inputs, which is discussed in Chapter 4.2.2. For performance assessments, we used the ROC curve and area under the curve (AUC) values to demonstrate how well models performed, which is discussed in Chapters 4.2.3 and 5.3.1.

For our fourth research question, we aimed to choose the GRN with the highest AUC value (determined by addressing the third research question) and attempted to determine which GRN

algorithms can provide the best performance model for the three data types. This process is discussed in Chapter 5.3.2.

The processes of data input, data pre-processing, DEG analysis, and performance measurement for different weighted GRNs can be achieved using R. The analysis of functional enrichment can be achieved using DAVID. Moreover, the drawing of unweighted TF-TG GRNs can be achieved using Cytoscape.

2.3 Summary of Main Goal

By answering all of the research questions, most of the discussion relates to differences based on data, which supports achieving our main goal of summarising the statistical impacts of three types of gene expression data on GRN reconstruction analyses. We divided this summary into four parts, as shown in Table 2.

Table 2. Summary of Comparisons on data

Summary	*Comparison*
Data Input	Measurement of gene expression level
	Number of detected genes with/without duplicate genes
	Number of experiments
Data Pre-processing	Workflow
	Number of pre-processed genes
	Number of pre-processed experiments
Differentially Expressed Gene	Tools
	Number of differentially expressed genes
	Concordance of gene lists
Gene Regulatory Networks	Number of nodes and edges
	Algorithm performance in one data
	Algorithm performance in three data

3 Data

The input in any analysis of gene expression data is a matrix with rows representing genes and columns representing samples. This type of matrix is known as a gene expression matrix (GEM)[16]:

$$X_{N\times M} = \begin{pmatrix} X_1^1 & \cdots & X_1^M \\ \vdots & \ddots & \vdots \\ X_N^1 & \cdots & X_N^M \end{pmatrix}$$

where each X_i^j denotes the expression level of the gene i in sample j. The expression level of gene i across all samples can be represented as a gene expression profile (GEP)[16]:

$$X_i = (X_i^1, X_i^2, \cdots, X_i^M)$$

3.1 RNA Microarray Data

3.1.1 Data Source

For RNA microarray data, we loaded data as an ExpressionSet R object from an R package called prostateCancerCamcap[17], which contained three data sets from Cambridge cohorts: assay data set, feature data set and phenotypic data set. The assay data set, called assayData in the R object, is a gene expression data frame with row names being protein IDs and column names being sample IDs. Although this is the appropriate data frame, we had to change the row and column names for further analyses. The feature data set, called featureData in the R object, describes experiment-specific information about 47,323 probes. Moreover, the phenotypic

data set, called phenoData in the R object, describes information about 199 samples. These 199 samples include 13 CRPC samples, 74 normal samples and 112 tumour samples. Since we performed transcriptomic analyses of data from PCa studies, we also considered CRPC samples as tumour samples. Thus, we obtained $112 + 13 = 125$ tumour samples and 74 normal samples. Here, we first had to extract $GEM_{47,323\times199}$ (called camcap_GEM from the assay data set) for the microarray data. Notably, the assay data set includes row names with probe IDs; thus, we first had to map probe IDs to genes.

3.1.2 Data Pre-processing

We used the 'AmpliSeq for Illumina Transcriptome Human Gene Expression Panel' with 20,800 genes and corresponding probe IDs as a reference. The mapping results were exported as a data frame $GEM_{47,323\times201}$ with probe IDs and gene names. The first four rows of the data frame are presented in Table 3.

Table 3. Mapping Results

	Probe ILMN_ID	*Symbol*	*ALCaP01_T*	*ALCaP04_T*	⋯	TB12.2315_23_B	TB12.2336_6_B
1	*1343291*	*EEF1A1*	14.2497	14.3627	⋯	14.6375	14.5501
2	*1343295*	*GAPDH*	11.3993	12.9675	⋯	11.2210	11.5159
3	*1651209*	*SLC35E2*	6.4105	6.6710	⋯	6.6929	6.5510
4	*1651210*	*DUSP22*	6.8542	6.4612	⋯	6.4262	6.4751

Microarrays are often printed with probes in duplicates[18] and we detected there are at least 6,000 duplicate genes in our data frame $GEM_{47,323\times201}$. The most common strategy to solve this problem is to take the mean value of duplicate gene expression before the analysis of DEGs; however, it may lose important information about gene expression variability[18]. Firstly, we present one duplicate gene 'EEF1A1' as an example in Table 4. From Table 4, some probes

have different gene expressions in this duplicate gene. Thus, we retained all duplicate genes in our data frame $GEM_{47,323\times201}$ for further analyses.

Table 4. Example of One Duplicate Gene

	Probe	*ILMN_ID*	*Symbol*	*ALCaP01_T*	*ALCaP04_T*	⋯	TB12.2315_23_B	TB12.2336_6_B
1	*1343291*	*EEF1A1*	14.2497	14.3627	⋯	14.6375	14.5501
27,957	*1810810*	*EEF1A1*	13.3797	14.2552	⋯	14.2857	14.0536
32,221	*2038774*	*EEF1A1*	13.2736	14.2136	⋯	14.2323	14.0776
44,339	*3251737*	*EEF1A1*	9.7879	7.3801	⋯	9.6488	9.4488

Then, we deleted the rows with missing values since most of the analyses for microarray data require complete datasets[19]. We got the camcap_GEM with 27,933 genes and 199 samples without probe IDs. Thirdly, we filtered out low-expressed genes with the value of 0, which improved the sensitivity and precision of DEGs[20]. We got the same camcap_GEM with 27,933 genes and 199 samples without probe IDs. Finally, we detected whether there are possible outliers in 199 samples based on the spearman correlation coefficients among all samples. We were inspired by an R function called TCGAanalyze_Preprocessing from the R package TCGAbiolinks, which performs an Array Array Intensity correlation matrix samples by samples to detect the samples with low correlation (usually less than 0.5), which can be identified as possible outliers[21]. We created a 199×199 correlation matrix among samples, and the matrix found no sample with a low correlation (spearman correlation coefficient lower than 0.5) that can be identified as a possible outlier. Since the gene expressions in camcap_GEM all range from 5 to 15, we identified this data as normalised and scaled data. Thus, there was no need for normalisation and log transformation for this matrix. The matrix $camcap\_GEM_{27,933\times199}$ with duplicate genes was our final GEM for microarray data analysis.

The first four rows of the results of microarray data pre-processing are presented in Table 5.

Table 5. Results of Microarray Data Pre-processing

	Symbol	*ALCaP01_T*	*ALCaP04_T*	⋯	TB12.2315_23_B	TB12.2336_6_B
1	*EEF1A1*	14.2497	14.3627	⋯	14.6375	14.5501
2	*GAPDH*	11.3993	12.9675	⋯	11.2210	11.5159
3	*SLC35E2*	6.4105	6.6710	⋯	6.6929	6.5510
4	*DUSP22*	6.8542	6.4612	⋯	6.4262	6.4751

The workflow for pre-processing microarray data is shown in Fig 1.

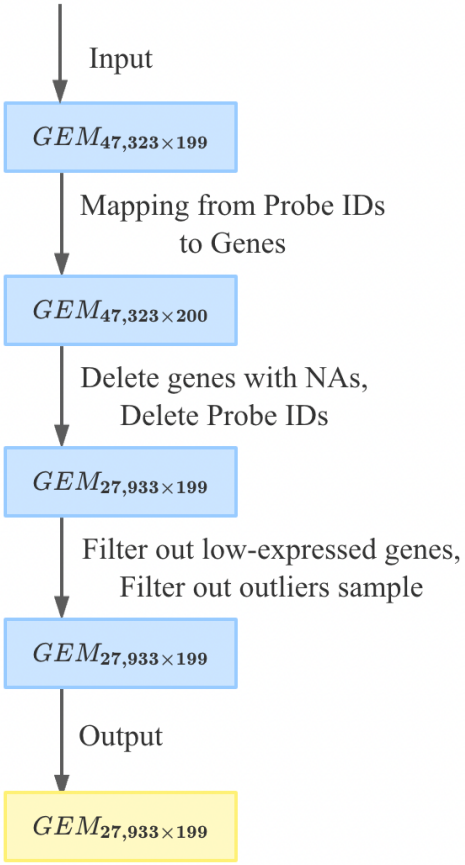

Figure 1. Workflow for Microarray Data Pre-processing

3.2 RNA Sequencing Data

3.2.1 Data Source

For RNA sequencing data, we downloaded data from an R package called TCGAbiolinks[22], which contains three main functions GDCquery, GDCdownload and GDCprepare that was used to search, download and load the data, respectively. The loaded data should be as a summarizedExperiment R object, also containing three data sets: assay data set, feature data set and phenotypic data set. The assay data set called assays in the R object is a gene expression data frame with rows names being gene names and columns names being sample IDs. This is the data frame we can use directly as the GEM for RNA-seq data. Here, the gene expression level is provided as read counts. The feature data set called rowRanges describes experiment-specific information about 19,947 genes, whilst the phenotypic data set called colData describes information about 549 samples. The 549 samples include 52 normal samples and 497 tumour samples. Here, we first extracted the $GEM_{19,947\times549}$ called TCGA_GEM for the RNA-seq data.

3.2.2 Data Pre-processing

Initially, we detected whether there were duplicate gene names. Notably, there were no duplicates in the matrix. Then, we deleted the rows with missing values since most of the analyses for RNA-seq data require a complete matrix[23]. There were no missing values in the matrix. Thirdly, we filtered out low-expressed genes within values of 0 and 10, which are considered 'noise' in RNA-seq analyses[24]. From the statistical aspect of filtering out low-expressed genes, this process facilitates a more reliable estimation of the mean-variance relationship[24]. Finally, we detected whether there were outliers in 549 samples using a correlation matrix (the inspiration and reason for this are provided in Chapter 3.1.2). We created

a 549×549 correlation matrix among samples, and the matrix found no samples with a low correlation that could be identified as a possible outlier. Since the expression levels in TCGA_GEM range from 10 and up to millions (maximum value: 3,521,698), there is a massive demand for the normalisation and log transformation in this matrix.

The matrix $TCGA\_GEM_{13,956\times549}$ was our final GEM for RNA-seq data analysis.

The first four rows of the RNA-seq data pre-processing results are presented in Table 6.

Table 6. Results of RNA-seq Data Pre-processing

Symbol	*TCGA-G9-7525-01A-31R-2263-07*	*TCGA-EJ-7789-01A-11R-2118-07*	⋯	*TCGA-KK-A6E5-01A-11R-A311-07*	*TCGA-EJ-5496-01A-01R-1580-07*
SERPINA5	45.6476	5.6378	⋯	17.1503	18.6947
MFSD2A	18.6915	35.9258	⋯	15.0115	18.6878
ACSL6	8.3188	21.7564	⋯	25.6407	18.6121
MCF2	3.1344	2.8035	⋯	7.5107	19.6508

The workflow for RNA-seq data pre-processing is presented in Fig 2.

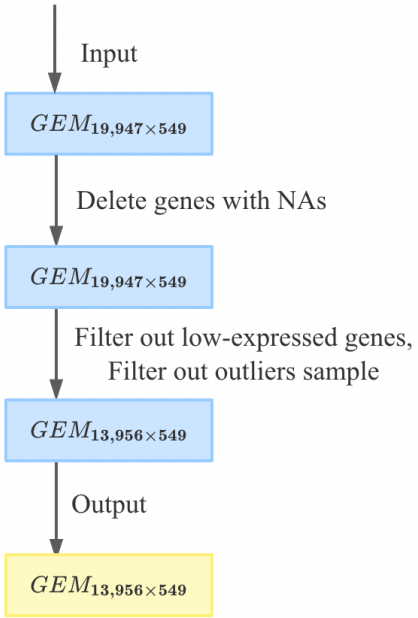

Figure 2. Workflow for RNA-seq Data Pre-processing

3.3 Protein Expression Data

3.3.1 Data Source

For protein expression data, we imported data from Supplementary Table 3 in the paper of Diego et al.[25]. This data set describes protein expression variation between tumour and normal samples. The expression is given as a log transformation of the normalised SILAC ratio. The normalised SILAC ratios of proteins identified in prostate tumours and normal tissue can be given as the ratio of the intensity of the tumour and the intensity of the SILAC standard. Thus, there is no need for normalisation and log transformation for this data set. The data set contains 1,224 genes, protein IDs, 1,224 unique protein IDs, 1,224 protein families and protein expression variation for 36 samples. The 36 samples include 28 tumour samples and 8 normal samples. According to the paper of Diego et.al, the p-values of the t-test between the two conditions are all less than 0.05[25], which indicates that the 1,224 genes in the data set are all DEGs. Since many elements in the data set can be GEM row names, we extracted two GEMs according to the requirements of the source nodes of GRNs. In one GEM, rows represent genes and columns represent samples. While in the other GEM, rows represent proteins using unique protein IDs and columns represent samples. Here, two GEMs—called ProExp_gene_GEM and ProExp_Uniprot_GEM—were extracted for the analyses of protein expression dat.

3.3.2 Data Pre-processing

3.3.2.1 First Matrix

Initially, we deleted the rows with missing values. Also, there were duplicate genes in the data frame. There was no need to filter out low-expressed genes since the data set already describes the expression variation in proteins (i.e., this is the data set after the analysis of DEGs). Finally, we detected whether there were outliers in 36 samples using a correlation matrix (the inspiration

and reason for this are described in Chapter 3.1.2). We created a 36×36 correlation matrix among samples. This matrix found no samples with low correlations that could be identified as possible outliers. The matrix $ProExp\_gene\_GEM_{880\times36}$ with duplicate genes was our final first matrix for protein data analysis.

3.3.2.2 Second Matrix

Initially, we deleted the rows with the missing values. We obtained the ProExp_Uniprot_GEM with 880 unique protein IDs and 36 samples. Since the protein IDs were unique, there were no duplicates. There was also no need to filter out low-expressed genes. Finally, based on the results of outlier detection for the samples in Chapter 3.3.2.1, we did not delete any samples. The matrix $ProExp\_Uniprot\_GEM_{880\times36}$ was our final second matrix for protein data analysis.

The workflow for protein data pre-processing is presented in Fig 3.

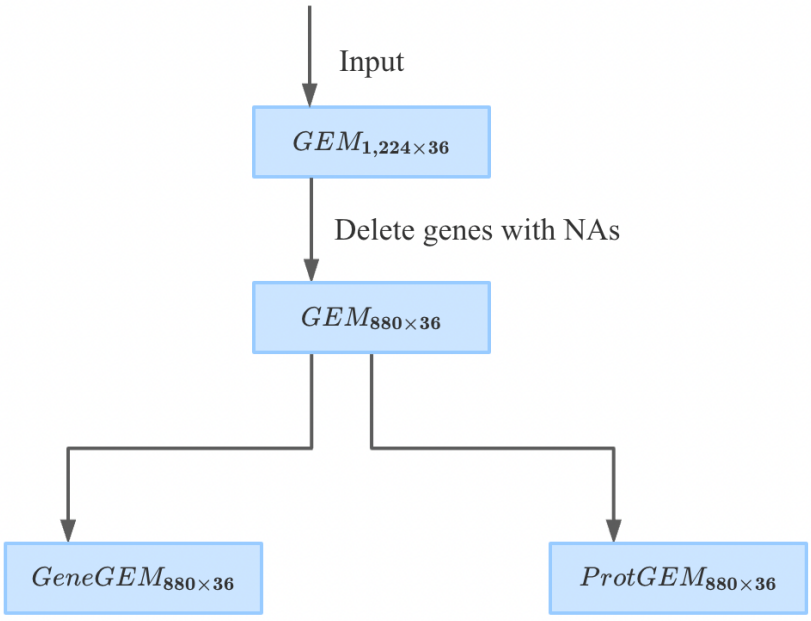

Figure 3. Workflow for Protein Expression Data Pre-processing

3.4 Validation Data

Complicated GRN reconstruction models can be considered binary classification models in machine learning, where the existence of edges in a predicted regulatory network is determined by predicting the existence of regulatory relationships between genes and genes or between TFs and genes. Thus, GRN reconstruction models regarded as binary classification models can be assessed by the performance assessment index in machine learning.

Additionally, several regulatory interactions have been experimentally retrieved in recent years. We regard these interactions as ‘known’ interactions. Although there remains a lack of knowledge regarding the ‘real’ interactions, we can only focus on the interactions that are ‘known’.

In a gene regulatory mechanism, if there is a ‘known’ interaction between a TF-gene or gene-to-gene pair, the ‘known’ interaction between this pair will be represented as a positive. In contrast, if there is no known interaction between a TF-gene or gene-to-gene pair, the relationship between this pair will be represented as a negative.

These ’known’ interactions can be accessed using by different databases, and we used the database BioGrid[21] as our validation data.

By comparing the ‘known’ interactions and the predicted interactions in our GRNs, we can know which algorithm performs best. Thus, the performance assessment aspects of our third research question can be answered.

3.5 Summary of Data

3.5.1 Summary of Data Input

We first discuss the differences in data input, including the measurement of gene expression level, the number of detected genes and the number of experiments.

For differences in the measurement of gene expression levels, RNA microarray data used intensity, RNA-seq data used raw counts, and protein expression data used protein expression variation. Additionally, RNA microarray data and protein expression data are scaled data, while RNA-seq data are raw data without normalisation and log transformation.

Regarding the number of detected genes, RNA microarray detected 47,323 probes with duplicates, RNA-seq detected 19,947 genes without duplicates, and protein expression data detected 1,224 genes with duplicates.

Regarding the number of experiments, RNA microarray data included 199 samples, RNA-seq included 549 samples, and protein expression data included 36 samples.

We summarise the aforementioned discussion in Table 7.

Table 7. Summary of Data Input Comparison

	Measurement of Gene Expression Level	*Number of Detected Genes*	*Number of Experiments (Tumour : Normal)*
RNA Microarray	Scaled intensity data	47,323 (with duplicates)	125: 74
RNA Sequencing	Raw count data	19,947 (without duplicates)	497: 52
Protein Expression	Scaled expression variation	1,224 (with duplicates)	28: 8

3.5.2 Summary of Data Pre-processing

Here, we discuss the differences in data pre-processing, including workflow, the number of pre-processed genes and the number of pre-processed experiments.

For differences in workflow, protein expression data have obvious differences from the other two data types. That is, protein expression data have two workflows for two GEMs. The workflows for RNA microarray data and RNA-seq data are similar. Additionally, the workflow for protein expression data does not include the process of filtering out low-expressed genes, whereas the other two workflows do.

Regarding the number of pre-processed genes, RNA microarray data has 27,933 genes (including several duplicate genes), RNA-seq data has 13,956 genes (without duplicate genes) and protein expression data has 880 genes (including several duplicate genes) and 880 unique protein IDs (without duplicates).

In terms of the number of pre-processed experiments, the results are the same as presented in Chapter 3.4.1.

The aforementioned discussion is summarised in Table 8.

Table 8. Summary of Data Pre-processing Comparison

	Workflow	*Number of Pre-processed genes*	*Number of Experiments (Tumour : Normal)*
RNA Microarray	Similar to RNA sequencing	27,933 (with duplicates)	125: 74
RNA Sequencing	Similar to RNA Microarray	13,956 (without duplicates)	497: 52
Protein Expression	Different from the other two	880 (with duplicates)	28: 8

3.5.3 Summary of Data

Based on the summaries in Chapter 3.5.1 and Chapter 3.5.2, we can summarize the statistical differences in data preparation of three types of data in Table 9, which answers the first two parts of our second research questions. The more detailed discussion will talk about in Chapter 5.2.

Table 9. Summary of Data Comparison

Summary		***Comparison***		
Data Input		*Measurement of Gene Expression Level*	*Number of Detected genes*	*Number of Experiments (Tumour : Normal)*
	RNA Microarray	Scaled intensity data	47,323 (with duplicates)	125: 74
	RNA Sequencing	Raw count data	19,947 (without duplicates)	497: 52
	Protein Expression	Scaled expression variation	1,224 (with duplicates)	28: 8
Data Pre-processing		*Workflow*	*Number of* Pre-processed genes	*Number of Experiments (Tumour : Normal)*
	RNA Microarray	Similar to RNA sequencing	27,933 (with duplicates)	125: 74
	RNA Sequencing	Similar to RNA Microarray	13,956 (without duplicates)	497: 52
	Protein Expression	Different from the other two	880 (with duplicates)	28: 8

4 Methods

4.1 Analysis of Differentially Expressed Genes

As mentioned in Chapter 2.2, the goal of DEG analysis is to extract genes with significantly different expression levels under tumour/normal conditions and identify whether the DEGs are up- or down-regulated based on the values of fold change on the log scale.

The tools used for DEG analysis included the R packages called limma[26], edgeR[27] and DESeq2[28].

For microarray data, we used limma tools by applying generalised linear models and empirical Bayes methods for analysis of DEGs[29].

For RNA-seq data, we used two of all tools (edgeR and limma) on raw counts and extracted the intersection of the results.

For protein expression data, we directly had the list of significant DEGs by performing t-tests as mentioned in Chapter 3.3.1, which were the row names of the matrix $ProExp\_gene\_GEM_{880\times36}$. Thus, we obtained 880 DEGs for protein expression data.

4.1.1 Brief Descriptions of the Tools

The data inputs in all three tools were the same: a GEM and a design matrix. The design matrix shows whether each sample is in tumour condition or normal condition. The differences between the three tools were determined using statistical methods.

There are two steps in the limma tool. First, a generalised linear model can be built using a GEM and a design matrix. Second, by using empirical Bayes methods, we obtained two gene-specific posterior estimates: coefficients and variance. The four main statistical methods in the limma package include the following: (1) using empirical Bayes methods to obtain posterior variance estimators; (2) adding weights to allow variation; (3) estimating gene-specific parameters; (4) using adjusted p-values to assess significant DEGs[26]. The output for the limma tool is a data frame containing a list of significant DEGs with adjusted p-values all less than 0.01 and their corresponding values of fold change (FC) on the log scale (logFC).

The edgeR tool is based on likelihood. First, it maximises the negative binomial conditional likelihood to estimate a dispersion coefficient across all genes. Second, it estimates tagwise dispersion coefficients using empirical Bayes methods. Last, it uses the false discovery rate (FDR) to assess significant DEGs[27]. The output for the edgeR tool is also a data frame containing a list of significant DEGs with FDR all less than 0.01and their corresponding values of logFC.

Since the DESeq2 tool is not useful for the three studied data types, we do not discuss it further.

The t-test is to compare the mean of two groups. The null hypothesis is that the true difference between the two group means is equal to zero, whilst the alternative hypothesis is that the true difference between the two group means is not equal to zero. Here, we define the tumour sample be group 1 and the normal sample be group 2. For gene i, the expression level of gene i across all tumour samples can be presented as $X_{i1} = (X_{i1}^1, X_{i1}^2, \cdots, X_{i1}^{28})$; the expression level of gene I acroos all normal samples can be presented as $X_{i2} = (X_{i2}^1, X_{i2}^2, \cdots, X_{i2}^8)$. Then, the t-

test statistic can be calculated as $d = \overline{X_{\iota 1}} - \overline{X_{\iota 2}}$. The null hypothesis can be presented as $H_0: d = 0$, whilst the alternative hypothesis can be presented as $H_1: d \neq 0$.

4.1.2 Results of Analysis of DEGs

We obtained a list of significant DEGs with adjusted p-values or FDRs less than 0.05. Also, we obtained values of logFC for the significant DEGs. A negative logFC indicates that the corresponding gene is down-regulated, while a positive logFC indicates that the corresponding gene is up-regulated. We performed functional enrichment based on the list of DEGs. As previously mentioned, we used DAVID with a list of DEGs to predict which TFs the DEGs are enriched in. We then obtained a list of TFs and their interacting DEGs. Then, our unweighted TF-DEG GRNs could be reconstructed.

4.1.3 Summary of Analysis of DEGs

Hereafter, we discuss statistical differences in the analysis of DEGs through the tools applied to the three data types. Table 10 provides a summary of the discussion presented in Chapter 4.1.1.

Table 10. Summary of DEGs Tools Comparison

Tools	*Data*	*Data Input*	*Statistical Method*	*Output*	*Result*
edgeR	RNA microarray	GEM + design matrix	Generalised linear model (posterior estimates)	DEG with logFC	$\text{FDR} < 0.01$
limma	RNA sequencing and RNA microarray	GEM + design matrix	Likelihood (dispersion estimates)	DEG with logFC	Adjusted $\text{p} - \text{values} < 0.01$
GEM itself	protein expression	GEM	T-test	DEG with t-test difference	T-test $\text{p} - \text{values} < 0.05$

For RNA microarray data, we used the results provided by the limma tool. We will get a list of significant DEGs with their corresponding logFC values (with adjusted p-values all less than 0.01).

For RNA-seq data, we used the intersection of the two results provided by two tools (the limma tool and the edgeR tool). We will get a list of significant DEGs with their corresponding logFC values (with both adjusted p-values and FDR values all less than 0.01).

For protein expression data, the number of DEGs was equal to the number of rows in ProExp_gene_GEM. We got a list of significant DEGs with their corresponding t-test difference values (with t-test p-values all less than 0.05).

4.2 Gene Regulatory Network

Generally speaking, in a GRN, nodes usually represent regulatory elements (genes, TFs or miRNAs), whilst edges represent regulatory relationships. Here, based on our three gene expression data, we can reconstruct the TF-gene GRN and gene-gene GRN for our subsequent reconstruction analysis.

To answer our first research question, we reconstructed the unweighted TF-DEG GRN with source nodes as AR-interacting TFs, target nodes as DEGs and edges as the presence of interactions between AR-interacting TFs and DEGs. Furthermore, we reconstructed different weighted TF-DEG GRN algorithms by using MI tools and comparing ‘known’ interactions and predicted interactions to measure the performance of each GRN algorithm. This process aimed to answer our third and fourth research questions.

Additionally, a GRN can be presented by an adjacency matrix $A = [a_{ij}]$ $(i = 1, \cdots, M; j = 1, \cdots N)$ that encodes whether/how a pair of nodes interacts. An unweighted GRN implies that its corresponding adjacency matrix only contains the values of 0 and ± 1, thereby showing whether pairs of nodes interact. In contrast, a weighted GRN implies that its corresponding adjacency matrix contains values with a range of $[-1,1]$, thereby showing how pairs of nodes interact with several relationship measurement indexes, such as correlation or MI.

4.2.1 Workflow for GRN Reconstruction

After DEG analyses and functional enrichment, we obtained a list of DEGs, a list of TFs and their interactions. Thus, the workflow for reconstructing GRN is clear (see Fig. 4).

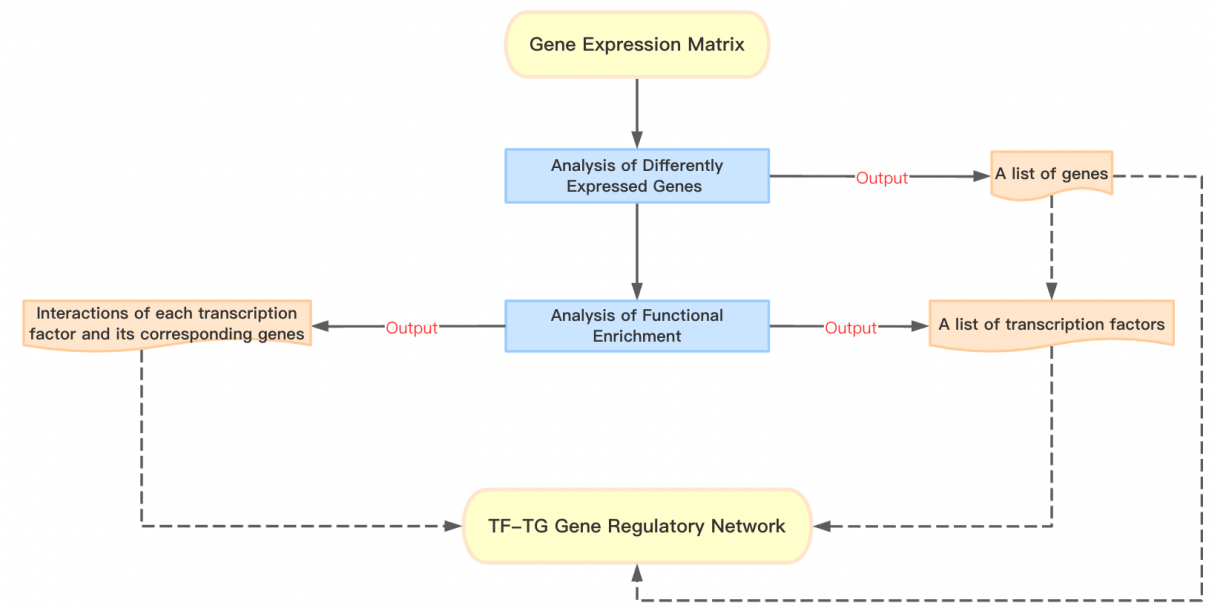

Figure 4. Workflow for GRN Reconstruction

4.2.2 GRN Methods

We used the GRN algorithms based on the mutual information tools for each data set. We chose four algorithms based on MI: ARACNE[30], CLR[31], MRNET[32], and C3NET[33]. The main reason for choosing these four algorithms was that they can all be achieved using R packages

and our process mainly utilise R. The second reason for these choices was that the outputs of the four algorithms are adjacency matrices, which enables us to measure the performance of algorithms.

Many coexpression-based GRN algorithms use different tools:

(1) Bayesian network structure learning tools, such as GeneNet algorithm[34], use correlation as relationship measurements and can predict directed GRNs with the outputs being data frames containing all edges listed in order of the magnitude of the partial correlation associated with each edge. This tool is achieved on JavaScript.

(2) Differential equation-based methods[35] are achieved on MATLAB.

(3) Gene co-expression networks[36] can also infer GRNs; however, co-expression networks require the clinical data set to assess the relationship between modules and traits. Here, for protein expression data, we could not assess the clinical data set of 36 samples.

Overall, these are the reasons why we do not use other tools for coexpression-based GRN algorithms and focus on MI tools instead.

MI captures the non-linear relationship between two random variables[37] and has also been used to measure the regulatory relationship between genes. Mathematically, MI measures the degree of dependence between two genes: X_i and X_j[38].

$$MI_{ij} = \sum_{X_i} \sum_{X_j} p(X_i, X_j) log_2 \frac{p(X_i, X_j)}{p(X_i)p(X_j)}$$

Where $p(X_i, X_j)$ is the joint probability distribution function of X_i and X_j, and $p(X_i)$ and $p(X_j)$, are the marginal probability distribution function of X_i and X_j, respectively[38].

The marginal probability distribution function of X_i can be written as

$$p(X_i) = \frac{1}{c} exp\left[-\sum_{i}^{N} \phi_i(X_i) - \sum_{i,j}^{N} \phi_{ij}(X_i, X_j) - \sum_{i,j,k}^{N} \phi_{ijk}(X_j, X_j, X_k) - \cdots\right]$$

where c is the normalization score, N is the number of genes, X_i is the GEP of gene i.

The ARACNE algorithm identifies interactions by estimating pairwise GEP MI. Since GRNs are reconstructed by DEGs under two conditions, the MI can be inferred from pairwise marginal. The second step of the ARACNE algorithm is to remove indirect interactions ($\phi_{ij} = 0$)[30].

CLR algorithm obtains a score related to the empirical distribution of MI values[31]. The score CLR_{ij} can be given between gene i and gene j as

$$CLR_{ij} = \sqrt{CLR_i^2 + CLR_j^2}$$

with

$$CLR_i = \max\left(0, \frac{MI_{ij} - \hat{\mu} MI_i}{\widehat{\sigma_{MI_i}}}\right)$$

and

$$CLR_j = \max\left(0, \frac{MI_{ji} - \hat{\mu} MI_j}{\widehat{\sigma_{MI_j}}}\right)$$

The parameters μ and σ are the mean and standard deviation of the empirical distribution of MI, respectively[31].

MRNET algorithm uses the feature selection technique to maximise the MI with the DEGs[32]. The feature selection technique uses a forward selection strategy to select features that are strongly conditioned by the first selected variables.

The first step of C3NET algorithm[33] is the same as ARACNE. The difference is in the second step. In this step, only one of the most prominent edges is selected for each gene. The C3NET algorithm aims to build the edge of every gene.

We propose qualitative features (described in Table 11) based on the different steps of the four algorithms and attempt to summarise the performance of these algorithms in Table 12.

Table 11. Description of Qualitative Features

Feature	*Description*
Linear Assumption	Does the algorithm assume the linear relationship between gene-gene pairs or TF-gene pairs?
Directed Network	Does the algorithm output the directed network?
Parameter Tuning	Does the algorithm require manual parameters tuning?
Threshold Setting	Does the algorithm require setting up thresholds?
Loop structure	Does the algorithm include the loop structure?
Prior Knowledge	Does the algorithm require prior knowledge?
Performance Measurement	Can the algorithm do performance measurement?

Table 12. Summary of Comparison of Features of GRN methods

	ARACNE	*MRNET*	*CLR*	*C3NET*
Linear Relationship	Yes	Yes	Yes	Yes
Directed Network	No	No	Yes	No
Parameter Tuning	No	Yes	Yes	No
Threshold Setting	No	No	No	No
Loop structure	Yes	No	No	No
Prior Knowledge	Yes	Yes	Yes	Yes
Performance Measurement	Yes	Yes	Yes	Yes

4.2.3 Performance Measurements of GRN Methods

By comparing the 'known' interactions and the interactions in our predicted GRNs, our performance assessments from machine learning were achieved. We often used a confusion matrix to present the results of these comparisons, resulting in the confusion matrix presented in Table 13.

Table 13. Confusion Matrix

True	*Predict*	
	Positive	*Negative*
Positive	TP	FN
Negative	FP	TN

As mentioned in Chapter 3.4, if there is a 'known' interaction between a TF-TG or gene-to-gene pair, the 'known' interaction between this pair will be represented as a positive. In contrast,

if there is no known interaction between a TF-gene or gene-to-gene pair, the relationship between this pair will be represented as a negative.

Additionally, TP indicates the number of TF-TG or gene-to-gene pairs that are predicted as positive and also true as positive. FP indicates the number of TF-TG or gene-to-gene pairs that are predicted as positive but true as negative. TN indicates the number of TF-TG or gene-to-gene pairs that are predicted as negative and also true as negative. FN indicates the number of TF-TG or gene-to-gene pairs that are predicted as negative but true as positive.

According to TP, FP, TN and FN, the performance of a GRN algorithm can then be measured using the ROC curve and AUC. Some rates related to performance measurements can be given by:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The ROC curve is a graph presenting the performance of the classification model at all classification thresholds, which is plotted by TPR versus FPR at different classification thresholds. The closer the ROC curve is to the upper left corner, the higher the FPR and lower the FPR of the test, that is, the higher the sensitivity rate, the better the performance of the algorithm.

AUC is the area under the ROC curve. That is, AUC measures the area under the entire ROC curve from point (0,0) to point (1,1), providing performance across all possible classification thresholds. However, AUC is not always desirable. Under the case that prioritizes minimizing the number of FP, AUC is not reliable. Instead, we can use the ROC curve.

5 Results

Based on the analyses of data and methods outlined in previous sections, we can now answer our four research questions:

- Based on our examination of GRNs from our three types of data, can some of the investigated AR-interacting TFs be enhancers or inhibitors of PCa??
- How different are the pre-processing methods for our data?
- How does network analysis differ when using different types of gene expression data?
- Which GRN algorithm and data type can provide the best performance model to make our inferences and analyses more reliable?

We answer these research questions in their own respective sections.

5.1 Answers to the First Research Question

5.1.1 Microarray Data

Firstly, after the analysis of DEGs using the limma package, we obtained a list of 105 DEGs with duplicate genes. We first examined the results of the DEG analysis for duplicate genes (see Table 14). In Table 14, it is significant that all the values of logFC for triplicate genes AMACR and duplicate genes HPN are positive, which indicates that genes AMACR and HPN are up-regulated genes. And all the values of logFC for triplicate genes MSMB, PGM5 and duplicate genes MYLK are negative, which indicates that genes MSMB, PGM5 and MYLK are down-regulated genes.

Table 14. Duplicate DEGs in Microarray Data

Probe ID	*Symbol*	*logFC*	*Adjusted P-value*
38009	*AMACR*	1.9801	<0.05
25106	*AMACR*	1.5406	<0.05
19913	*AMACR*	1.8421	<0.05
7639	*HPN*	1.7716	<0.05
37779	*HPN*	1.7213	<0.05
24804	*MSMB*	-1.9364	<0.05
38251	*MSMB*	-1.9125	<0.05
9925	*MSMB*	-1.6247	<0.05
8479	*MYLK*	-1.4137	<0.05
37939	*MYLK*	-1.0254	<0.05
23384	*PGM5*	-1.1569	<0.05
11782	*PGM5*	-1.1605	<0.05
36153	*PGM5*	-1.1802	<0.05

After summarising the results of the analysis of DEGs for duplicate genes, we got a list of $105 - 13 + 5 = 97$ DEGs in total. By exporting the heat map (Fig. 5) of these 97 DEGs based on $DEG\_GEM_{97\times199}$, we visualised that gene expression is significantly different under the two conditions.

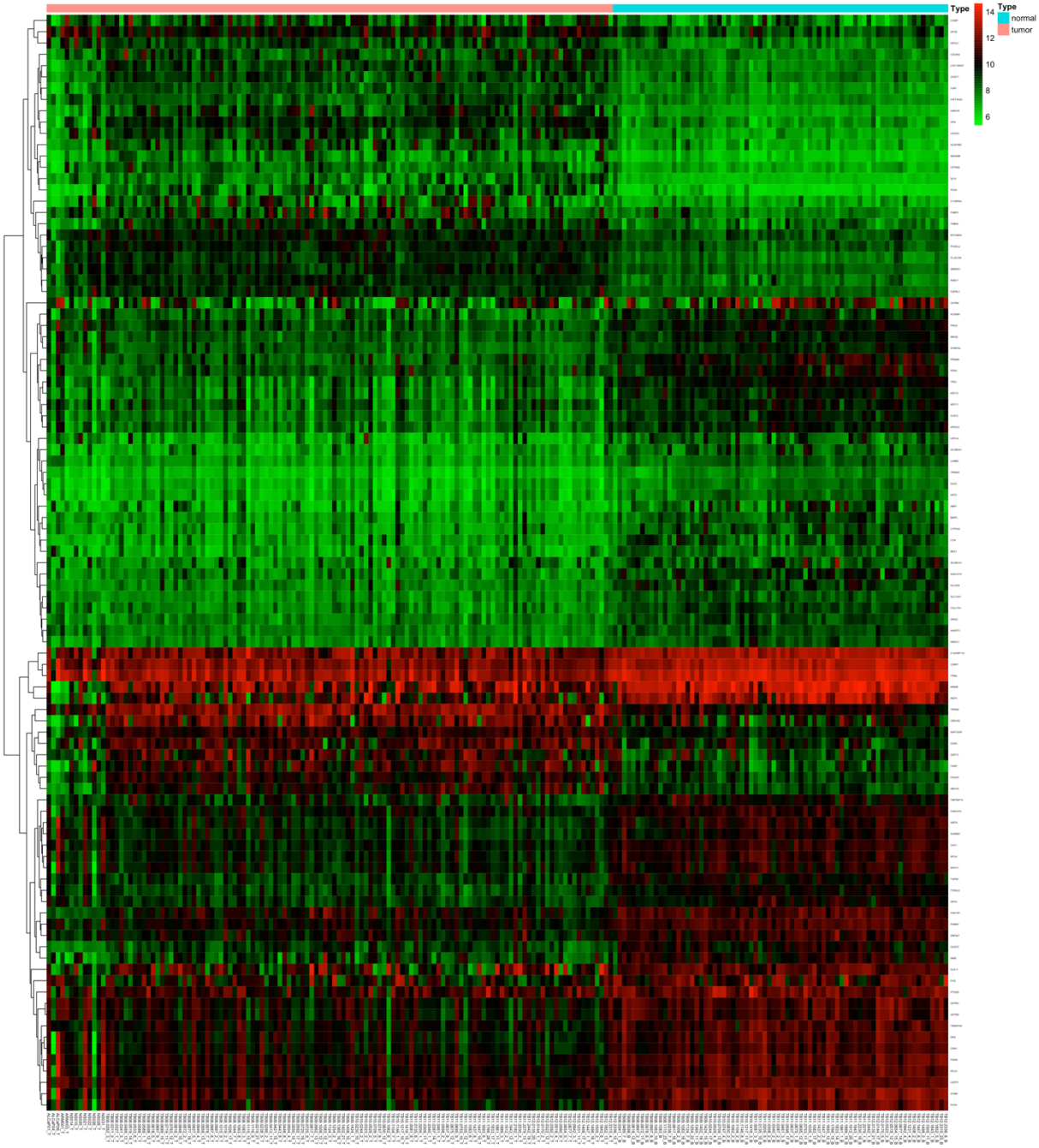

Figure 5. Heatmap of Microarray Data

Then, the results of the analysis of functional enrichment can be exported as a data frame containing a list of 39 TFs and the 1,559 interactions between 39 TFs and 97 DEGs.

The first twelve rows of the data frame can be shown in Table 15.

Table 15. Results of the Analysis of Functional Enrichment on RNA Microarray Data

TF	*Count*	*Percentage*	*Genes*	*FDR*
NFE2	42	47.19%	COL17A1, SYNM, BEX1, …	0.05560
ZIC3	42	47.19%	PRDM8, COL17A1, DLX1,…	0.21410
LMO2COM	57	64.04%	PRDM8, SYNM, PODXL2,…	0.25945
LYF1	40	44.94%	PRDM8, COL17A1, DLX1,…	0.33131
SOX9	44	49.44%	DLX1, PNCK, PARM1,…	0.33131
CEBPA	31	34.83%	DLX1, BEX1, PARM1,…	0.33131
NRSF	51	57.30%	PRDM8, SYNM, PODXL2,…	0.33131
SOX5	46	51.69%	PRDM8, COL17A1, DLX1, …	0.34915
SRY	41	46.07%	PRDM8, COL17A1, DLX1,…	0.34915
HMX1	44	49.44%	PRDM8, COL17A1, DLX1, …	0.34915
HTF	50	56.18%	PRDM8, PODXL2, PAK1IP1,…	0.37557
MYOD	51	57.30%	PRDM8, SYNM, PODXL2, …	0.37557

From Table 15, we can see that all FDRs are greater than 0.05, which implies that none of the differential expressed genes was enriched on these TFs (i.e. there were no interactions between the TFs and the differential genes).

Thus, since there were only insignificant TFs and insignificant interactions, an analysis of functional enrichment could not be completed. Furthermore, the reconstruction of the TF-TG GRN could not be completed using microarray data. This suggests that there is no great advantage to using microarray data to perform GRN reconstruction analysis in PCa studies.

5.1.2 RNA-seq Data

As noted in Chapter 4.1.3, we extracted the intersection of two results using two tools (edgeR and limma). A Venn diagram (Fig. 6) shows how the two results of the number of DEGs differ and intersect. The result using the edgeR tool includes 2,286 DEGs, the result using the limma tool includes 2,280 DEGs, and the intersection of two results includes 1,847 DEGs.

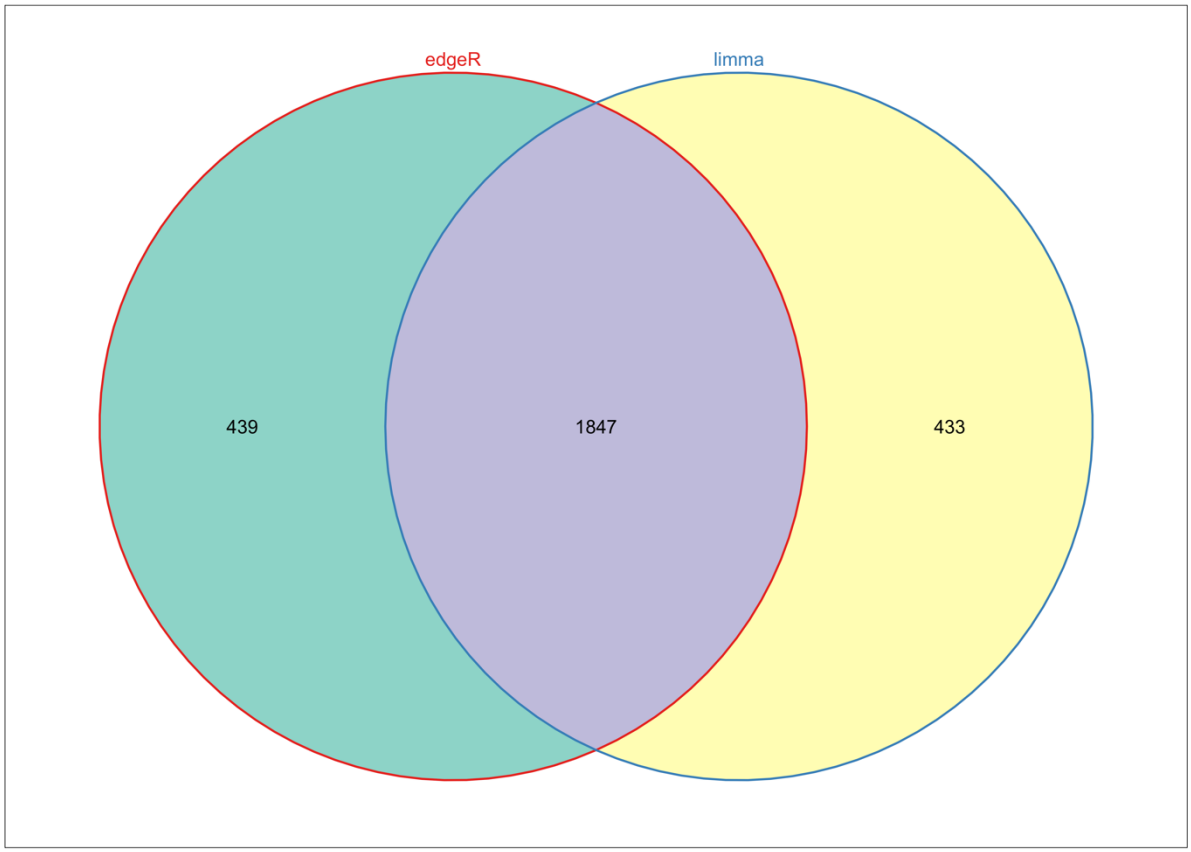

Figure 6. Venn Plot of Results of Analysis of DEGs

Overall, we obtained 1,847 DEGs. Similarly, the results of the DEG analysis were visualised by exporting the heat map (Fig. 7) of 1,847 DEGs based on $DEG\_GEM_{1,847\times549}$. In Fig. 7, it is evident that the gene expression levels significantly differ between normal and tumour samples. Also, we exported the first four rows of the data frame (Table 16) containing the results of the DEG analysis.

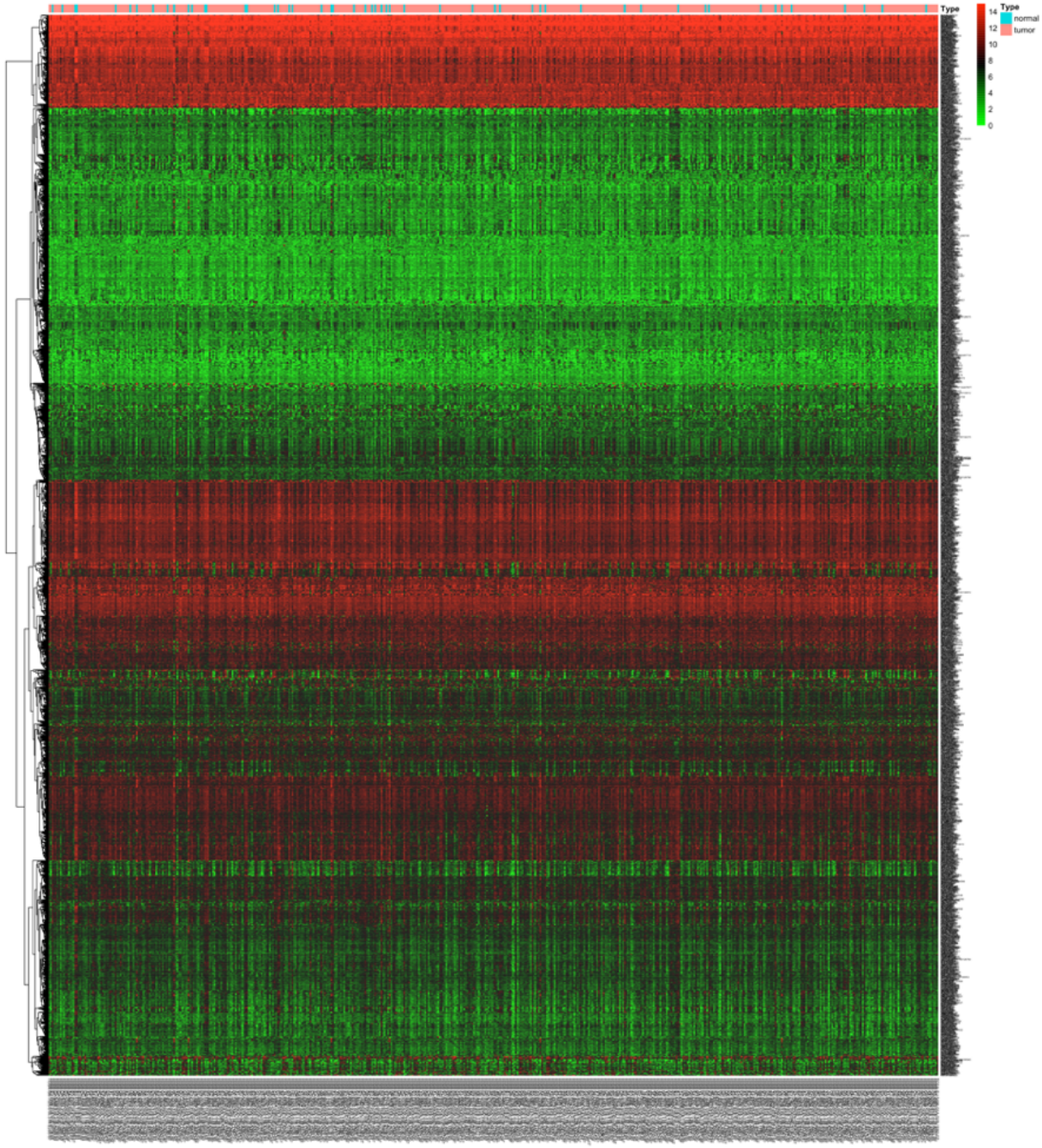

Figure 7. Heatmap of RNA-seq Data

Table 16. Results of Analysis of DEGs on RNA-seq Data

Gene	*logFC*	*Adjusted P-Value*
SERPINA5	-6.8922911	<0.01
MFSD2A	-6.2313122	<0.01
ACSL6	-5.6018902	<0.01
MCF2	-5.8536205	<0.01

In Table 16, the values of logFC for genes SERPINA5, MFSD2A, ACSL6 and MCF2 are all negative, which indicates that the four DEGs are down-regulated.

Then, the results of the analysis of functional enrichment were exported as a data frame containing a list of 171 TFs and the 123,016 interactions between 171 TFs and 1,847 DEGs. According to the investigated AR-interacting TFs (FOX-family, OCT1, GATA2, p53, NF1, AP1, IRF1 and STAT3), we obtained 12 AR-interacting TFs with statistical significance (FDR < 0.01) from the results of functional enrichment. The 12 AR-interacting TFs are AP1, NF1, FOXO3, FOXO1, FOXD3, STAT3, IFR1, FOXO4, p53, FOXJ2, GATA2 and OCT1 as source nodes in the unweighted TF-DEG GRN. The first twelve rows of the data frame are shown in Table 17.

Table 17. Results of Analysis of Functional Enrichment on RNA-seq Data

AR-interacting TF	*Count*	*Percentage*	*Genes*	*FDR*
AP1	972	59.63%	FLJ12825, MYLK, PREX2, …	<0.01
NF1	632	38.77%	CNTFR, MAML2, FLJ12825,…	<0.01
FOXO3	480	29.45%	VIT, MAML2, FLJ12825,…	<0.01
FOXO1	679	41.66%	CNTFR, MAML2, FLJ12825, …	<0.01
FOXD3	568	34.85%	VIT, MAML2, FLJ12825, …	<0.01
STAT3	754	46.26%	VIT, MAML2, FLJ12825, …	<0.01
IRF1	433	26.56%	VIT, MAML2, FLJ12825, …	<0.01
FOXO4	854	52.39%	FLJ12825, MYLK, PREX2,…	<0.01
P53	994	60.98%	FLJ12825, MYLK, PREX2,…	<0.01
FOXJ2	980	60.12%	FLJ12825, MYLK, PREX2,…	<0.01
GATA2	523	32.09%	ROPN1B, CNTFR, MAML2, …	<0.01
OCT1	1284	78.77%	FLJ12825, MYLK, PREX2, …	<0.01

From Table 17, it is evident that each AR-interacting TF interacts with a large number of genes. Most TFs interact with more than 50% of 1,847 DEGs. Since OCT1 has the most interactions with DEGs, we can conclude that OCT1 is the most active TF in PCa based on

our RNA-seq data. We also wanted to examine whether OCT1 can be an inhibitor or an enhancer in PCa. Thus, we exported the GRN.

Here, we selected 50 interactions from each AR-interacting TF to improve the clarity and verification of our GRN. After exporting the GRN plotted by Cytoscape (Fig. 8), we examined this GRN to answer our first research question on RNA-seq data. This TF-DEG GRN (Fig. 8) contains 162 nodes and 600 edges. The purple nodes are AR-interacting TFs, the orange nodes are up-regulated genes, and the green nodes are down-regulated genes. The edges represent the presence of interaction between TF-DEG pairs. All AR-interacting TFs—including OCT1—interact with both up- and down-regulated genes. Thus, our answer to the first research question is that one of the investigated AR-interacting TF OCT1 can be examined as the most active TF in PCa based on our RNA-seq data; however, we do not have sufficient confidence to state that OCT1 is an enhancer or inhibitor in PCa.

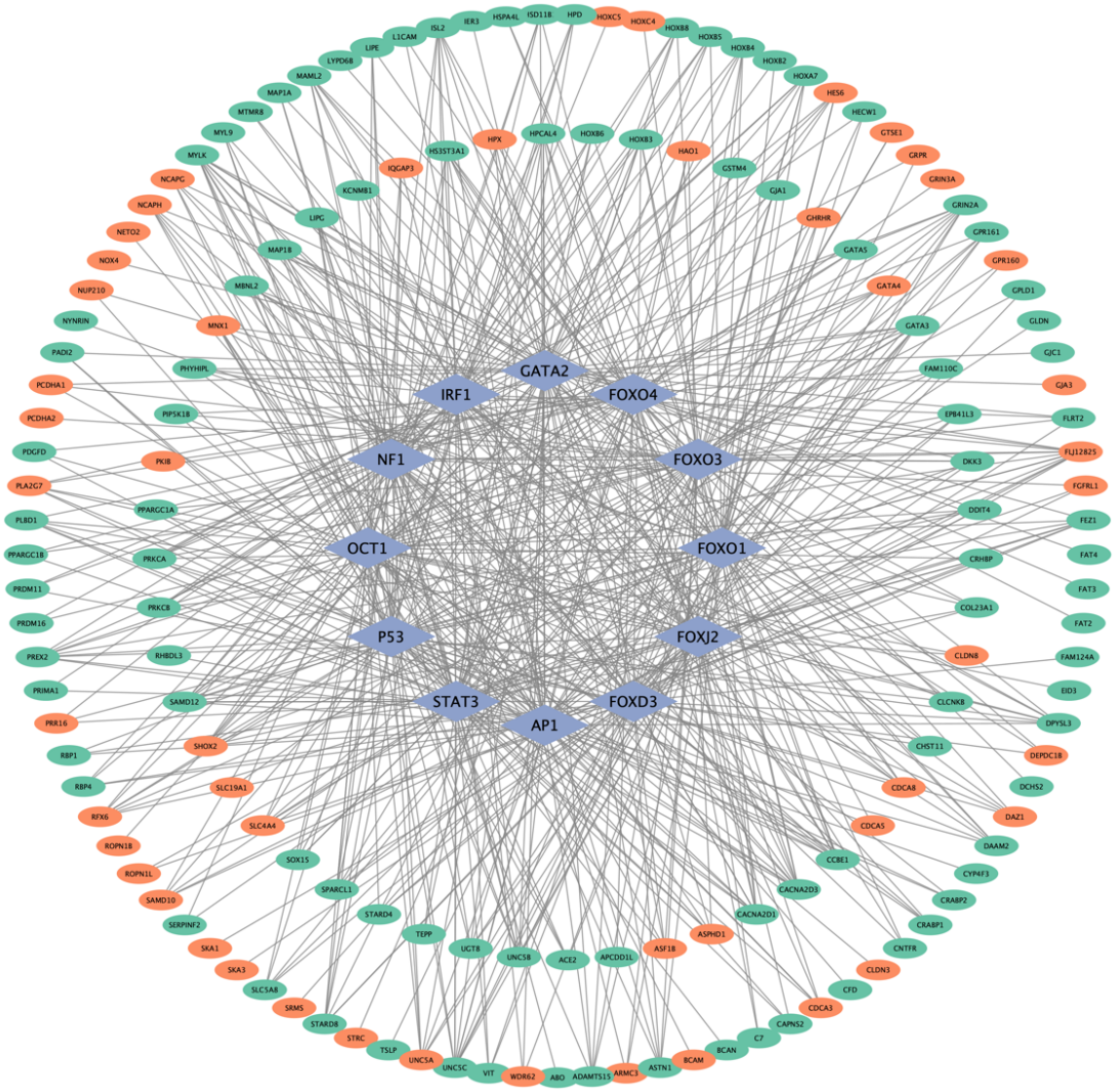

Figure 8. Unweighted TF-DEG GRN on RNA-seq Data

5.1.3 Protein Expression Data

We directly used the matrix $ProExp\_gene\_GEM_{880\times36}$ with duplicates genes. We first examined the results of the DEG analysis for duplicate genes (see Table 18). In Table 18, it is evident that all the values of t-test differences for duplicate genes NME1 are positive, which indicates that gene NME1 are up-regulated genes. And all the values of t-test differences for duplicate genes MXRA7, JUP and SEPT9 are negative, which indicates that genes MXRA7, JUP and SEPT9 are down-regulated genes. However, there are two duplicate genes NUDT4 and HNRNPD with both positive and negative t-test differences in Table 18. Currently, we

cannot define whether these two genes are up- or down-regulated. Thus, we kept all the values of t-test differences for these two duplicate genes NUDT4 and HNRNPD for analysis of functional enrichment and GRN reconstructions.

Table 18. Duplicate DEGs in Protein Expression Data

Protein ID	*Symbol*	*T-test Difference*	*T-test P-value*
Q9NZJ92	*NUDT4*	1.1145	<0.05
Q141031	*NUDT4*	-1.0500	<0.05
Q141032	*HNRNPD*	0.6857	<0.05
Q32Q12	*HNRNPD*	-0.3873	<0.05
P15531	*NME1*	0.3751	<0.05
P841571	*NME1*	0.3139	<0.05
P841572	*MXRA7*	-0.7499	<0.05
F5GWP8	*MXRA7*	-1.2592	<0.05
P14923	*JUP*	-1.2655	<0.05
Q9UHD81	*JUP*	-0.4199	<0.05
Q9UHD81	*SEPT9*	-2.7431	<0.05
Q9NZJ92	*SEPT9*	-0.4679	<0.05

After summarising the results of the analysis of DEGs for duplicate genes, we got a list of $880 - 12 + 8 = 876$ DEGs in total. By exporting the heat map (Fig. 9) of these 876 DEGs based on $DEG\_GEM_{876\times36}$, we visualised that gene expression is different under the two conditions.

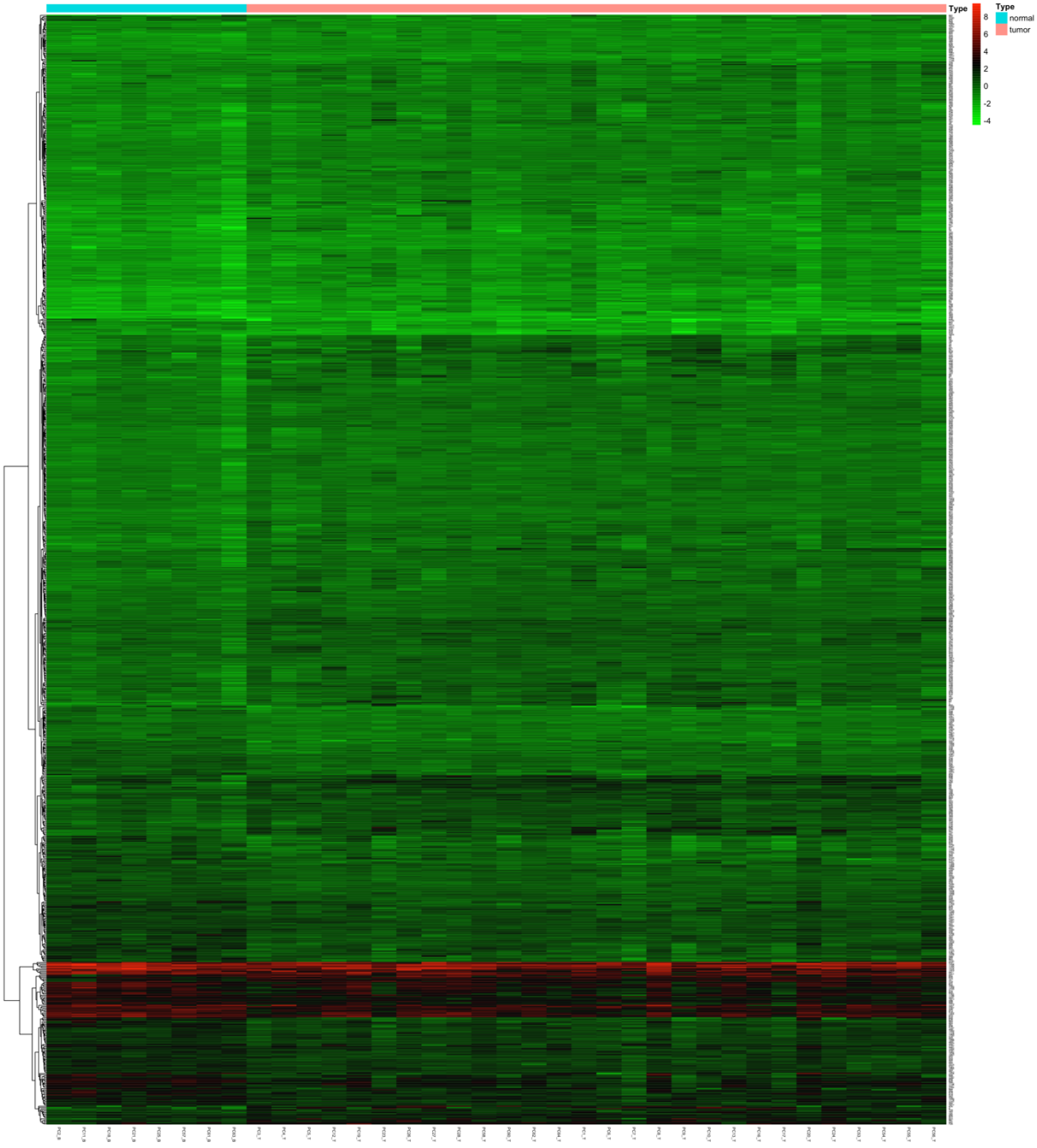

Figure 9. Heatmap of Protein Expression Data

After performing the analysis of functional enrichment, we also obtained a data frame containing a list of 156 TFs and the 58,343 interactions between 156 TFs and 876 DEGs. According to the investigated AR-interacting TFs (FOX-family, OCT1, GATA2, p53, NF1, AP1, IRF1 and STAT3), we got ten AR-interacting TFs with statistical significance (FDR < 0.01) from the results of functional enrichment. The ten AR-interacting TFs are AP1, NF1,

FOXJ2, FOXO1, FOXD4, STAT3, IFR1, p53, GATA2 and OCT1 as source nodes in the unweighted TF-DEG GRN. The first ten rows of the data frame can be shown in Table 19. The AR-interacting TF list in protein expression data has no FOXO3 and FOXD3 TFs which was investigated using RNA-seq data.

Table 19. Results of Analysis of Functional Enrichment on Protein Expression Data

AR-interacting TF	*Count*	*Percentage*	*Genes*	*FDR*
AP1	466	55.88	TDRKH, RPL4, TRIO, …	<0.01
NF1	272	32.61	TDRKH, RPL4,…	<0.01
FOXJ2	489	58.63	TDRKH, RPL5, RPL30, …	<0.01
FOXO1	308	36.93	RPL5, TRIO, CLPB,…	<0.01
FOXO4	412	49.40	TDRKH, TRIO, CLPB, …	<0.01
STAT3	378	45.32	TDRKH, RPL4, …	<0.01
IRF1	195	23.38	TCERG1, RPL30, …	<0.01
P53	542	64.99	TDRKH, RPL4, RPL5, …	<0.01
GATA2	293	35.13	EIF4A1, TRIO, RPL3,…	<0.01
OCT1	686	82.25	TDRKH, RPL4, RPL5, …	<0.01

From Table 19, it is evident that each AR-interacting TF interacts with a large number of genes. Four of TFs interact with more than 50% of the 876 DEGs. Since OCT1 has the most interactions with DEGs, we can draw the same conclusion as in Chapter 5.1.2, which is that OCT1 is the most active TF in PCa based on our RNA-seq data. We also exported the GRN to examine whether OCT1 can be an inhibitor or enhancer in PCa.

As the process of GRN reconstruction mentioned in Chapter 5.1.2, we selected 50 interactions of each AR-interacting TF. This TF-DEG GRN (Fig. 10) contains 165 nodes and 500 edges. The purple nodes are AR-interacting TFs, the orange nodes are up-regulated genes and the green nodes are down-regulated genes. The edges represent the presence of interaction between TF-TG pairs. Most DEGs are up-regulated genes. This may be because we used t-test statistics to test the expression variants under different conditions, which may not be very reliable. All AR-interacting TFs—including OCT1—mostly interact with up-regulated genes.

Thus, our answer to the first research question is that one of the investigated AR-interacting TFs, OCT1, can be examined as the most active TF and an enhancer in PCa based on our protein expression data.

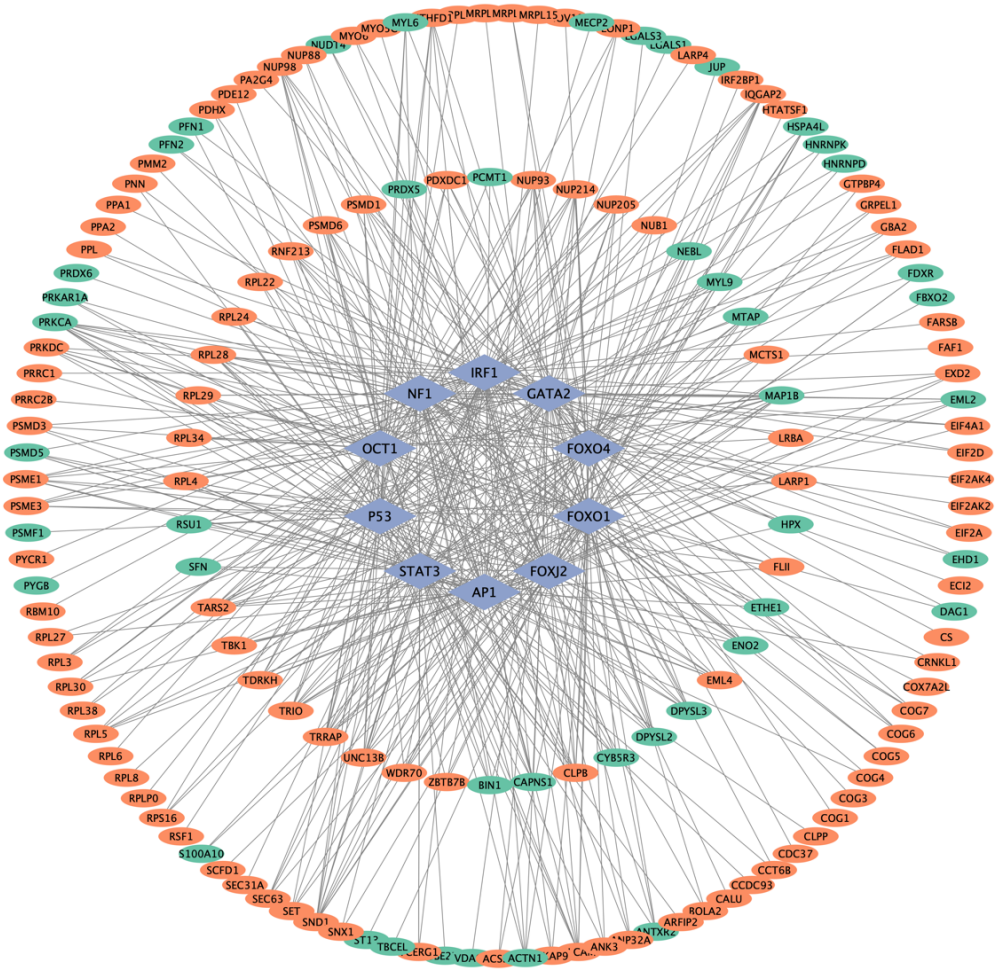

Figure 10. Unweighted TF-DEG GRN on Protein Expression Data

5.1.4 Summary of Answers to the First Research Question

Overall, one of the investigated AR-interacting TFs, known as OCT1, can be examined as the most active TF in PCa based on our RNA-seq and protein expression data. Moreover, OCT1 can be examined as an enhancer in PCa based on protein expression data.

We can also conclude that RNA-seq and protein expression data offer some advantages over RNA microarray data in PCa-related studies.

5.2 Answers to the Second Research Question

In this study, we answered three aspects of differences in data pre-processing: data input, data pre-processing and DEG analysis. Since DEG analyses represent preparations for GRN reconstruction, we treated these as pre-processing steps. Moreover, since we have discussed the differences in data (inputs and pre-processing) in Chapter 3.5, we mainly discuss the differences in analyses of DEGs here. Furthermore, we answer our second research question through a comprehensive summary.

5.2.1 Differences in analyses of DEGs

As noted in Chapter 2.3, the differences in DEG analyses can be divided into three parts: the differences in tools based on data types, the numbers of DEGs and the concordance of DEGs between three data studies. The differences in tools based on data types were discussed in Chapter 4.1.3, and the numbers of DEGs have been described in Chapter 5.1. We calculated the concordance of DEGs by calculating the ratio of the number of the same genes to the number of all genes. A discussion of these points is summarised in Table 20.

Table 20. Comparison of DEGs

Data	*Tools*	*Number of DEGs*	*Number of TFs*	*Number of Interactions*	*Overlapping DEGs with RNA-seq*	*Overlapping DEGs with Protein*
RNA Microarray	limma	97	39	1,559	68 ($\frac{68}{97} = 70.1\%$)	7 ($\frac{7}{97} = 7.22\%$)
RNA Sequencing	edgeR and limma	1,847	171	123,016	-	-
Protein Expression	GEM itself	876	156	58,343	65 ($\frac{65}{876} = 7.42\%$)	-

5.2.2 Summary of Answers to the Second Research Question

We integrated the comparisons of entire data pre-processing from Table 9 and Table 20 to Table 21. Thus, we can answer our second research question following three parts of the differences by our plan in Chapter 2.2: differences in data input, data pre-processing and the analysis of DEGs.

Table 21. Comparisons of Entire Data Pre-processing

Data Input	*Measurement of gene expression level*	*Number of detected genes*	*Number of experiments (Tumour : Normal)*		
RNA Microarray	Scaled intensity data	47,323 (with duplicates)	125: 74		
RNA Sequencing	Raw count data	19,947 (without duplicates)	497: 52		
Protein Expression	Scaled expression variation	1,224 (with duplicates)	28: 8		
Data Pre-processing	*Workflow*	*Number of detected genes*	*Number of experiments (Tumour : Normal)*		
RNA Microarray	Similar to RNA sequencing	27,933 (with duplicates)	125: 74		
RNA Sequencing	Similar to RNA Microarray	13,956 (without duplicates)	497: 52		
Protein Expression	Different from the other two	880 (with duplicates)	28: 8		
Analysis of DEGs	*Tools*	*Number of DEGs*	*Number of TFs*	*Number of Interactions*	*Overlapping DEGs with RNA-seq*	*Overlapping DEGs with Protein*
RNA Microarray	limma	97	39	1,559	68 ($\frac{68}{97}$ = 70.1%)	7 ($\frac{7}{97}$ = 7.22%)
RNA Sequencing	edgeR and limma	1,847	171	123,016	-	-
Protein Expression	GEM itself	876	156	58,343	65 ($\frac{65}{876}$ = 7.42%)	-

In Table 21, it is evident that RNA-seq data offer some advantages over microarray and protein expression data in data pre-processing since the number of DEGs, the number of TFs and the number of interactions were the largest.

Additionally, RNA microarray data clearly has the largest number of detected genes, but the number of DEGs, TFs and interactions are all lower than those in RNA-seq data. Because the RNA microarray data is scaled data, and we lack the information on how the data is scaled. And the limma tool outputs the values of logFC, which is also a scaled value. This may remain some problem. However, we used the limma tool and the edgeR tool on the raw counts for RNA-seq data. This might cause the differences between RNA microarray data and RNA-seq data.

For protein expression data, since all t-test p-values are less than 0.05 (not 0.01), it might cause the number of DEGs to get larger. And this also caused the number of FP to increase. Thus, the results of the analysis of DEGs for protein expression data are not reliable.

5.3 Answers to the Third and Fourth Research Question

5.3.1 Answers to the Third Research Question

Since the reconstruction of the TF-TG GRN could not be completed using microarray data (discussed in Chapter 5.1.1), we reconstructed gene-to-gene GRN based on MI tools.

Additionally, since we began with the gene expression matrix containing only differentially expressed data, we use DEG_GEM here. The DEG_GEM for RNA microarray data is the matrix $camcap\_DEG\_GEM_{97\times199}$, whilst the DEG_GEM for RNA-seq data is the matrix $TCGA\_DEG\_GEM_{1,847\times549}$ and the DEG_GEM for protein expression data is the matrix $ProExp\_DEG\_GEM_{876\times36}$.

Also, the validation data for the three data types were different. We obtained a set of known interactions from the BioGRID database. There are 32 ‘known’ interactions between the 97

DEGs in the RNA microarray data, 28,468 ‘known’ interactions between the 1,847 DEGs in the RNA-seq data and 6,494 ‘known’ interactions between the 876 differentially expressed genes in the protein expression data.

We then compared the inferred network with known interactions from the BioGRID database and performed performance assessments by comparing ROC curves and AUC values.

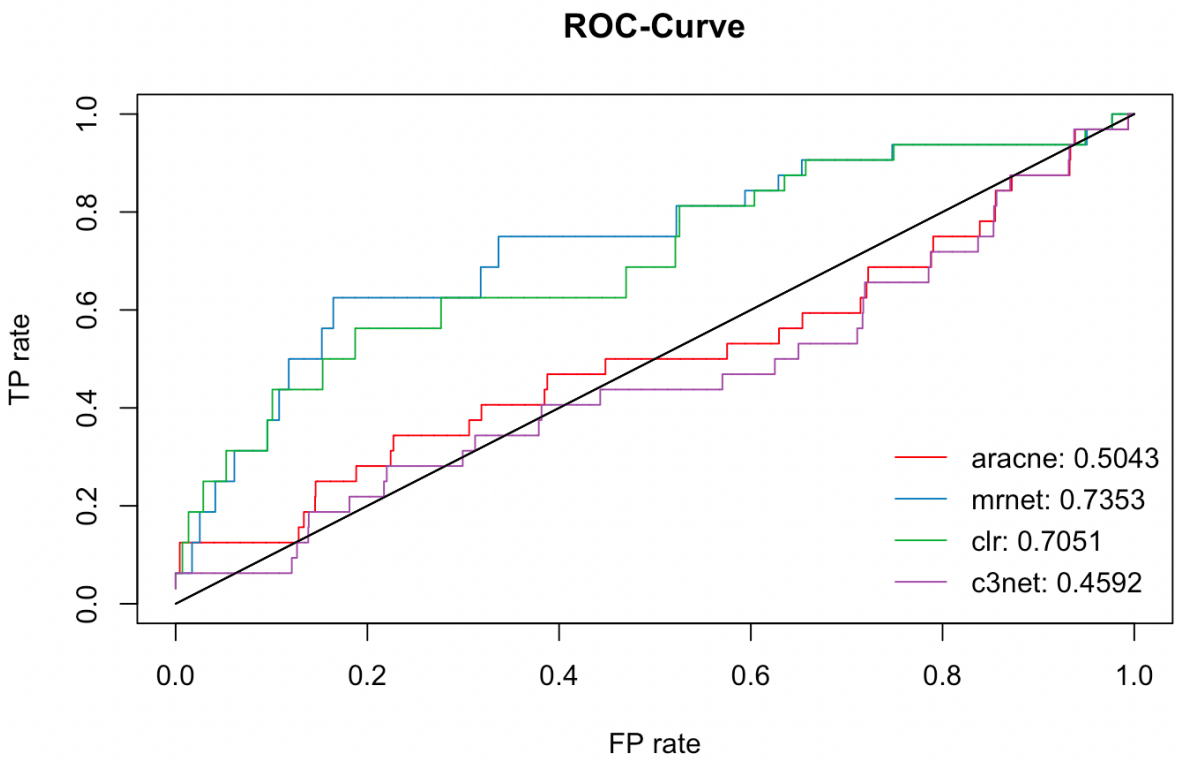

Figure 11. ROC with Corresponding AUC on RNA Microarray Data

In Fig. 11, the obtained ROC curve and corresponding AUC based on RNA microarray data are presented. As we mentioned in Chapter 4.2.3, under the case that prioritizes minimizing the number of FP, AUC is not helpful. We see the ROC curve instead. In Fig. 11, the ROC curve for the MRNET algorithm is the closest to the upper left corner. Thus, it is evident that the

MRNET algorithm performs best when comparing the inferred network on RNA microarray data with known interactions from the BioGRID database.

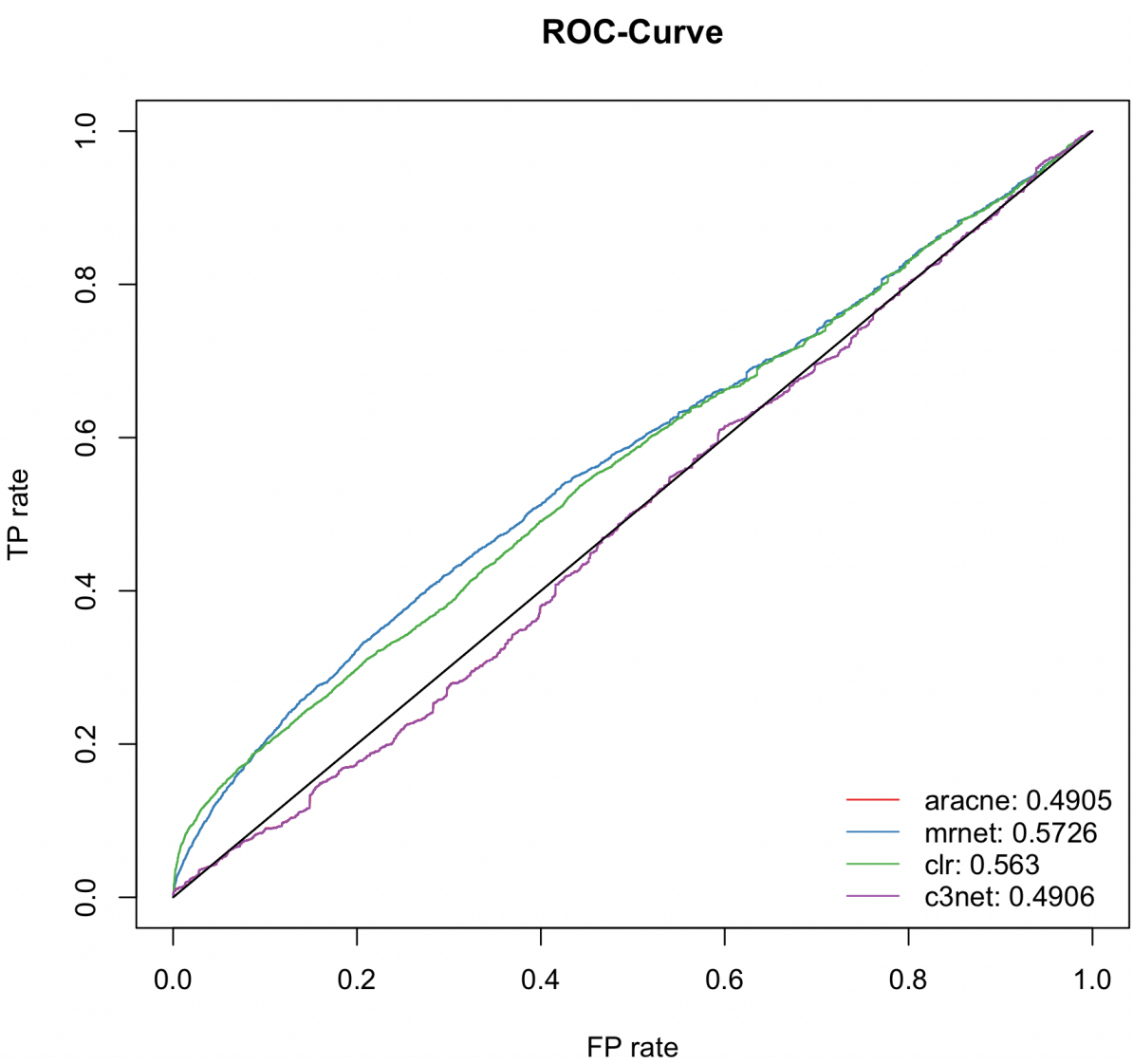

Figure 12. ROC with Corresponding AUC on RNA-seq Data

In Fig.12, the obtained ROC curve and the value of AUC on RNA-seq data are presented. We want to minimise the number of FP, so we look at the ROC curve. In Fig. 12, the ROC curve for the MRNET algorithm is the closest to the upper left corner. Thus, it can be observed that the MRNET algorithm performs best when comparing the inferred network on RNA-seq data with known interactions from the BioGRID database.

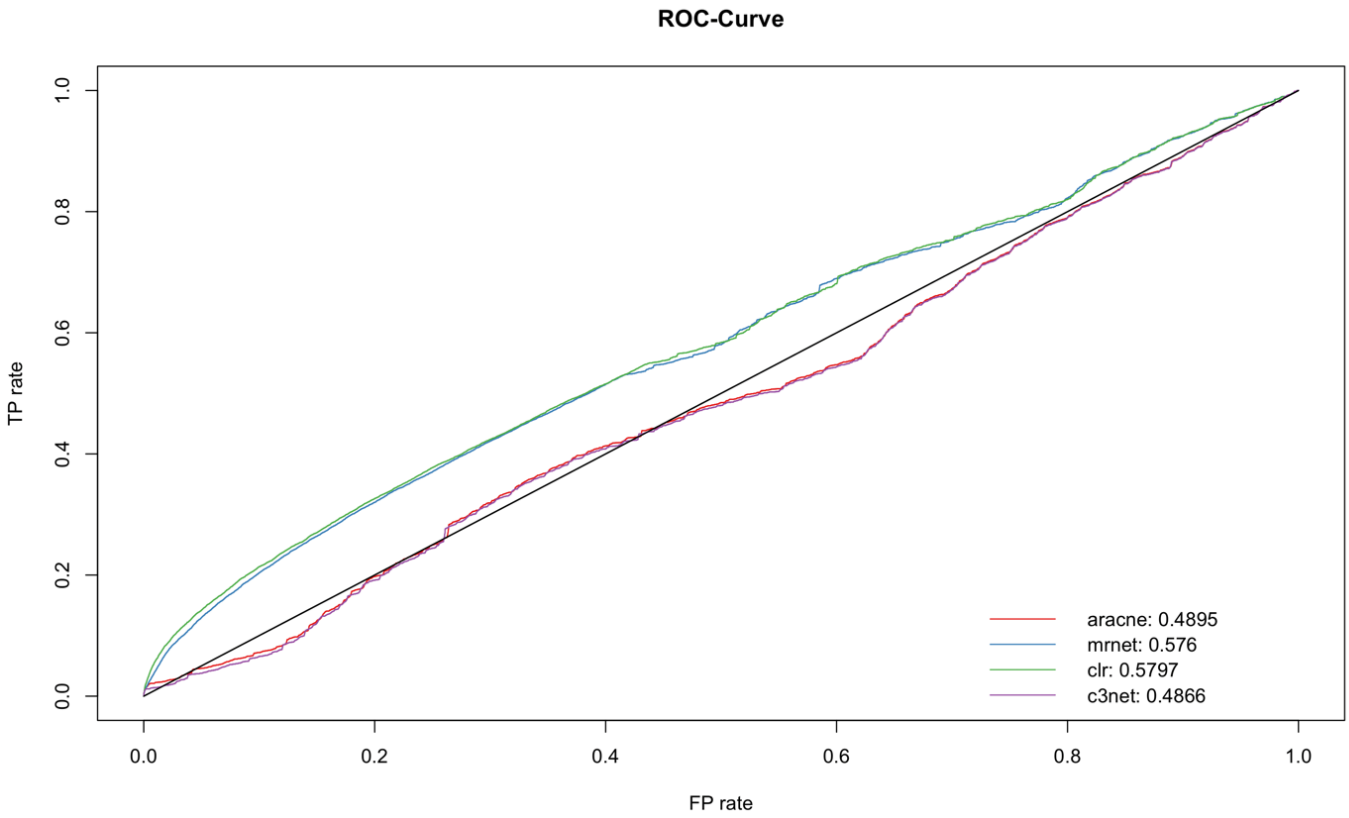

Figure 13. ROC with Corresponding AUC on Protein Expression Data

Fig. 13 presents the obtained ROC curve and the value of AUC based on protein expression data. Again, we want to minimise the number of FP, we look at the ROC curve which is the closest to the upper left corner. Thus, it can be observed that the CLR performed best when comparing the inferred network based on protein expression data with known interactions from the BioGRID database.

Now we can answer our third research question from the Table 22.

Table 22. Performance Measurement on Three Data

	Number of DEGs	*Number of 'known' interactions*	*Best Performing Algorithm*
RNA Microarray	97	32	MRNET
RNA Sequencing	1,847	28,468	MRNET
Protein Expression	876	6,494	CLR

For RNA microarray data, the MRNET algorithm performs best; for RNA-seq data, the MRNET algorithm performs best; for protein expression data, the CLR performed best when comparing the inferred network based on protein expression data with known interactions from the BioGRID database.

Since the number of DEGs and the number of the 'known' interactions were the largest for RNA-seq data among the three data, the number of TP and TN will be the least. This caused the smallest AUC for RNA-seq data. In contrast, the number of DEGs and the number of the 'known' interactions were the smallest for RNA microarray data among the three data. The number of FP and TP increased a lot. Thus, this is the most unreliable result of all. In terms of protein expression data, we have discussed that the results of analysis of DEGs for this data is not reliable in Chapter 5.2.2, so that the GRN reconstructions for this data is not reliable.

5.3.2 Answers to the Fourth Research Question

Then we can answer our fourth research questions in the Table 23.

Table 23. Comparisons of Performance Assessments

Data	*Best Performing Algorithm with corresponding AUC*
RNA Microarray	MRNET (0.7353)
RNA Sequencing	MRNET (0.5726)
Protein Expression	CLR (0.5797)

From Table 23, the AUC value is the lowest for the best-performing algorithm based on RNA-seq data. As we mentioned, under the case that prioritizes minimizing the number of FP, AUC is not reliable. Additionally, this may be because it has the largest number of 'known' interactions.

Moreover, as we stated before, the results of the analysis of DEGs for protein expression data are not reliable, and the results for RNA microarray data remain some problems. Thus, RNA-seq data offer some advantages over microarray and protein expression data in data preparation.

In summary, we still believe that RNA-seq data offer advantages over the other two types of studied data in PCa-related research. And the best-performing algorithm for RNA-seq data is the MRNET algorithm.

5.4 Answers to Summary of Comparisons on data

We created an integrated table (Table 24) to show the specific differences in each part of the comparisons on data we summarised in Chapter 2.3

Table 24. Answers to Summary of Comparisons on data

Data Input	*Measurement of gene expression level*	*Number of detected genes*	*Number of experiments (Tumour : Normal)*			
RNA Microarray	Scaled intensity data	47,323 (with duplicates)	125: 74			
RNA Sequencing	Raw count data	19,947 (without duplicates)	497: 52			
Protein Expression	Scaled expression variation	1,224 (with duplicates)	28: 8			
Data Pre-processing	*Workflow*	*Number of detected genes*	*Number of experiments (Tumour : Normal)*			
RNA Microarray	Similar to RNA sequencing	27,933 (with duplicates)	125: 74			
RNA Sequencing	Similar to RNA Microarray	13,956 (without duplicates)	497: 52			
Protein Expression	Different from the other two	880 (with duplicates)	28: 8			
Analysis of DEG	*Tools*	*Number of DEGs*	*Number of TFs*	*Number of Interactions*	*Overlapping DEGs with RNA-seq*	*Overlapping DEGs with Protein*
RNA Microarray	limma	97	39	1,559	68 ($\frac{68}{97}$ = 70.1%)	7 ($\frac{7}{97}$ = 7.22%)
RNA Sequencing	edgeR and limma	1,847	171	123,016	-	-
Protein Expression	GEM itself	876	156	58,343	65 ($\frac{65}{876}$ = 7.42%)	-
Performance Measurement	*Number of DEGs*	*Number of 'known' Interactions*	*Best Performing Algorithm*			
RNA Microarray	97	32	MRNET			
RNA Sequencing	1,847	28,468	MRNET			
Protein Expression	876	6,494	CLR			

6 Discussions

Reconstructing GRNs from expression data based on co-expression is a very difficult task. However, perhaps more important than details on the methods of analysis, the type and amount of data required to construct a reasonable model remains an open question. Solving this problem is difficult for several reasons. We discussed statistical comparisons of the entire GRN reconstruction process using different types of gene expression data to explain why the algorithms perform differently when given different data.

Despite the usefulness of this study, certain limitations to our analysis remain. For example, we only focused on one tool-based GRN algorithms.

For mutual information-based algorithms such as ARACNE and CLR, the network structure is determined by the degree of dependence between gene pairs. These MI network methods can infer directionality and potential causality whilst also being able to predict feedforward loops more accurately; however, the performance of linear cascades is limited.

The different MI-based algorithms were only predicted based on the partial knowledge of the network. This caused some FPs may be predicted as TPs. However, the performance of algorithms is limited.

Reference

[1] Sîrbu, A., Crane, M., & Ruskin, H. J. (2015). Data integration for microarrays: Enhanced inference for gene regulatory networks. Microarrays, 4(2), 255-269.
[2] Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J. Clin. 2010; 60: 277–300.
[3] Debes JD, Tindall DJ. Mechanism of androgen-refractory prostate cancer. N. Engl. J. Med. 2004; 351: 1488–90.
[4] Takayama, K., & Inoue, S. (2013). Transcriptional network of androgen receptor in prostate cancer progression. International journal of urology : official journal of the Japanese Urological Association, 20(8), 756–768. https://doi.org/10.1111/iju.12146
[5] Takayama, K., & Inoue, S. (2013). Transcriptional network of androgen receptor in prostate cancer progression. International journal of urology : official journal of the Japanese Urological Association, 20(8), 756–768. https://doi.org/10.1111/iju.12146
[6] Wang Q, Li W, Liu XS et al. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol. Cell 2007; 27: 380–92.
[7] Guseva NV, Rokhlin OW, Bair TB, Glover RB, Cohen MB. Inhibition of p53 expression modifies the specificity of chromatin binding by the androgen receptor. Oncotarget 2012; 3: 183–94.
[8] Jia L, Berman BP, Jariwala U et al. Genomic androgen receptor-occupied regions with different functions, defined by histone acetylation, coregulators and transcriptional capacity. PLoS One. 2008; 3: e3645.
[9] Agarwal SK, Guru SC, Heppner C, et al. Menin interacts with the AP1 transcription factor JunD and represses JunD-activated transcription. Cell, 1999, 96(1): 143–152.
[10] Cheng, Y., Wang, D., Jiang, J., Huang, W., Li, D., Luo, J., ... & Xu, Y. (2020). Integrative analysis of AR-mediated transcriptional regulatory network reveals IRF1 as an inhibitor of prostate cancer progression. The Prostate, 80(8), 640-652.
[11] Yeh, H. Y., Cheng, S. W., Lin, Y. C., Yeh, C. Y., Lin, S. F., & Soo, V. W. (2009). Identifying significant genetic regulatory networks in the prostate cancer from microarray data based on transcription factor analysis and conditional independency. BMC medical genomics, 2(1), 1-19.
[12] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Ste- paniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. Cell, 2000, 102(1): 109−126.
[13] Zhang, H., He, L., & Cai, L. (2018). Transcriptome sequencing: RNA-seq. In Computational systems biology (pp. 15-27). Humana Press, New York, NY.
[14] Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., & Giorgi, F. M. (2020). Gene regulatory network inference resources: A practical overview. Biochimica et biophysica acta. Gene regulatory mechanisms, 1863(6), 194430. https://doi.org/10.1016/j.bbagrm.2019.194430
[15] Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. https:// doi.org/10.1038/nprot.2008.211
[16] Margolin, A. A., & Califano, A. (2007). Theory and limitations of genetic network inference from microarray data. Annals of the New York Academy of Sciences, 1115(1), 51-72.
[17] Dunning M (2022). _prostateCancerCamcap: Prostate Cancer Data_. R package version 1.25.0.
[18] Smyth, G. K., Michaud, J., & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics, 21(9), 2067-2075.
[19] Chiu, C. C., Chan, S. Y., Wang, C. C., & Wu, W. S. (2013). Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC systems biology, 7(6), 1-13.
[20] Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences, 107(21), 9546-9551.
[21] Silva, T. C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., & Noushmehr, H. (2016). TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research, 5.
[22] Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., ... & Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research, 44(8), e71-e71.
[23] Baghfalaki, T., Ganjali, M., & Berridge, D. (2016). Missing Value Imputation for RNA-Sequencing Data Using Statistical Models: A Comparative Study. J. Stat. Theory Appl., 15(3), 221-236.
[24] Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences, 107(21), 9546-9551.

[25] Iglesias-Gato, D., Wikström, P., Tyanova, S., Lavallee, C., Thysell, E., Carlsson, J., ... & Flores-Morales, A. (2016). The proteome of primary prostate cancer. *European urology*, *69*(5), 942-952.
[26] Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.
[27] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformatics, 26(1), 139-140.
[28] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 1-21.
[29] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 3(1).
[30] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006, March). ARACNE: an algorithm for the rereconstruction of gene regulatory networks in a mammalian cellular context. In BMC bioinformatics (Vol. 7, No. 1, pp. 1-15). BioMed Central.
[31] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., ... & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology, 5(1), e8.
[32] Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. EURASIP journal on bioinformatics and systems biology, 2007, 1-9.
[33] Gokmen Altay, Frank emmert-Streib (2010). Structural Influence of gene networks on their inference: Analysis of C3NET. Submitted. URL http://cran.r-project.org/web/packages/c3net/index.html
[34] Schaefer J, Opgen-Rhein R, Strimmer. K (2021). _GeneNet: Modeling and Inferring Gene Networks_. R package version 1.2.16, https://CRAN.R-project.org/package=GeneNet.
[35] Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., & Giorgi, F. M. (2020). Gene regulatory network inference resources: A practical overview. Biochimica et biophysica acta. Gene regulatory mechanisms, 1863(6), 194430. https://doi.org/10.1016/j.bbagrm.2019.194430
[36] Peter Langfelder, Steve Horvath (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. Journal of Statistical Software, 46(11), 1-17.
[37] Laarne, P., Zaidan, M. A., & Nieminen, T. (2021). Ennemi: Non-linear correlation detection with mutual information. SoftwareX, 14, 100686.
[38] Bellot, P., Olsen, C., Salembier, P., Oliveras-Vergés, A., & Meyer, P. E. (2015). NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference. BMC bioinformatics, 16(1), 1-15.