TECHNISCHE UNIVERSITÄT DRESDEN

Center for Molecular and Cellular Bioengineering (CMCB)

BIOTEC

Master's program Molecular Bioengineering

Correlation-based detection of activation /repression transcription factor regulation from gene expression

A Thesis

By

Ming-Ju Kuo

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science

First Assessor: Dr. Carlo Vittorio Cannistraci

Second Assessor: Prof. Dr. Francis Stewart

Date of submission: 20.09.2019

Abstract

In the present study, we investigate methods that can indicate the regulation type of genes. For this purpose, many reverse-engineering algorithms inferring gene regulatory network (GRN) have been developed, but are not the only alternative for this task. Here, we applied correlation-based methods to different types of genome-wide high-throughput data to predict their gene regulation type, and validated their results by comparing against a gold standard regulation data constructed by a collection of information from published transcription factor articles. Additionally, we compare the correlation-based methods to published network-based methods by using both experimental data and simulated data. The results show that correlation-based methods perform well and can outperform regulatory network algorithms, being valid alternatives to gene regulation type predictions.

Acknowledgments

I would like to express my gratitude to people who made it possible for me to complete this master thesis. Firstly, I would like to thank Dr. Carlo Cannistraci for giving me a chance to work on my thesis and leading me to step in the bioinformatics field even if I did not have any informatics background. And I would like to thank also Claudio Durán who always solves the problem I have met and helps me a lot on the project. I want to thank Sara Ciucci for teaching me how to code from the beginning. Then I want to thanks my classmate and friends for assisting me with my study. Finally, I would like to thank my parents who always support me to pursue my degree. Without their support, I cannot reach so far.

Table of Contents

List of Figures

Abbreviations

CAGE	Cap Analysis of Gene Expression
ChIP	Chromatin Immunoprecipitation
DNA	Deoxyribonucleic Acid
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GRN	Gene Regulatory Network
GRNVBEM	Gene Regulatory Network Variational Bayesian Expectation-Maximization
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
RNA-seq	RNA-sequencing
mRNA	messenger RNA
siRNA	small interfering RNA
SINCERITIES	SINgle CEll Regularized Inference using Time-satmped Expression profileS
TN	True Negative
TP	True Positive
TSNI	Time Series Network Identification

1. Introduction

Although cancer research has been concentrated for decades, gene regulation in cancer cells is still unclear. To elucidate the properties of cancer and mechanism of carcinogenesis, we need to investigate the cancer gene regulatory which affects biological processes including cancer development, progression, and metastasis (Hollern *et al.*, 2014)(Zhang *et al.*, 2016). Compared to wild-type cells, cancer cells have different gene expression due to changed transcription factors which are regulatory proteins binding to DNA to activate or repress certain gene transcriptions. Therefore, the relationship and understanding between transcription factors and its regulated gene is a crucial factor to cancer therapeutic targets.

Nowadays, high-throughput technologies, such as microarray, can build the genome-scale gene expression profile by measuring the concentration of mRNA in a sample. There are two types of gene expression profile: (1) steady-state expression profile, which is collected from various conditions or samples at the same time. (2) dynamic expression profile, which is collected over multiple time-points (Wang *et al.*, 2008). These genome-scale gene expression profiles provide the information of each gene expression level and enable us to achieve the association between genes by reverse-engineering algorithms.

The application of reverse engineering to this problem involves to find expression level of genes which vary among steady-state or dynamic expression profile dataset and therefore can be used to infer gene regulatory networks (GRN) from these changes. Mining the interaction between two genes is the aim of reverse engineering. However, the interaction between two genes does not prove necessarily physical interaction. If an algorithm infers the positive correlation between two genes, G1 and G2, there would be three possible interpretations: (1) G1 and G2 both are regulated by a transcription factor, meaning that G1 and G2 are siblings. (2) G1 directly regulates G2. (3) G1 regulates a transcription factor which physically activates G2, meaning that G1 is a grandparent of G2 (Haury *et al.*, 2012).

The evaluation of reverse engineering algorithms needs a ground truth or gold standard to validate the performance of the inferred GRN. Elsewhere, reverse engineering algorithms are applied to simulated data generated from in silico networks which are the gold standards for the simulated data (Schaffter, Marbach and Floreano, 2011).

In this study, we aim to investigate the trend of transcription factor regulation from high-throughput technologies. Genome-scale microarray profiles (Barretina *et al.*, 2012), RNA sequencing (RNA-seq) profiles (Klijn *et al.*, 2015), and cap analysis gene expression (CAGE) profiles (Kawaji *et al.*, 2017) were obtained from published articles. We surveyed the transcription factor literature to manually build the used gold standard list; and moreover, we used simulated datasets and their respective network reference. We applied the correlation-based methods to both experimental and simulated expression profiles, and for comparison purposes, we used as well published network-based algorithms which can indicate the activation and repression of gene associations (Bansal, Gatta and di Bernardo, 2006)(Sanchez-Castillo *et al.*, 2018)(Papili Gao *et al.*, 2018). After algorithms implementation, the method's performance was evaluated by comparing the results to the gold standard.

1.1 Gene expression

In the biological system, the genetic information is obtained from the nucleotide sequence and plays a vital role in protein synthesis. In 1958, Francis Crick suggested the central dogma of biology which describes how the gene information becomes its product, protein, and RNA. There are biological processes involved in the central dogma: (1) The genetic information in organisms is encoded in DNA and represent as the certain nucleotide sequence which can be passed to offspring through the DNA replication. (2) During the development of organisms, the genetic information in DNA is served as a template that can be used to produce an mRNA molecule whose process is called transcription. (3) Protein is decoded and formed from the mRNA sequence and will be used in organisms to fulfill its biological purpose. The synthesis of protein is called translation. The phenotypes according to their genotypes by transcription and translation is the basic gene expression and therefore mRNA which involves in both transcription and translation can be the indicator to signify and quantify the gene expression.

1.1.1 DNA to protein

DNA is a double strand molecule containing the genetic information which is encoded to nucleotides or so-called bases (Adenine, Guanine, Cytosine ,and Thymine). Each of the nucleotides pairs specifically with another, Adenine with Cytosine and Guanine with Thymine. To pass the information to offspring, DNA replication is important and should be accurate. DNA replication is the synthesis of DNA which is the process using the original DNA as a template to replicate the identical DNA molecule by pairing the nucleotides on parent DNA according to the base pair rule. During the DNA replication, a double-strand DNA is rewound into two single strands and both strands can serve as templates to form two DNA molecules which are the same as the parent.

Nevertheless, the gene expression needs other processes to enable a certain DNA fragment to generate bioactivity products and this would need transcription and translation. In the transcription section, a double-strand DNA molecule is rewound and serves as a template for producing the transcript in the beginning. The RNA polymerase read and recognizes the specific bases on the rewound DNA strand and then take the unique nucleotide to pair the template DNA. After pairing, a single complementary strand molecule called mRNA is formed and separated from the single- strand DNA. The bases in mRNA are different from the bases in DNA, the base Thymine is changed by Uracil instead.

Translation is a process to produce protein which is the macromolecule composed of amino acid sequences. Once the mRNA is formed, it is transferred to the cytoplasm and associated with the ribosomes. A specific sequence of amino acid is produced according to the sequence of mRNA which can be translated from a three-nucleotide subsequence codon to a corresponding amino acid. Peptide chains are formed by ribosomes and tRNA using the translation machinery. At the end of the translation phase, the completed peptide would be folded into the specific structural protein whose function will be to regulate other biological processes.

$$DNA \leftrightarrow RNA \rightarrow Protein$$

1.1.2 Transcription factor

The process of the protein-producing is well controlled and regulated in the cell. Transcription factors are the protein that can associate with the specific DNA sequence and activate or inhibit the RNA polymerase recruitment. The DNA-binding domain of the transcription factor can be one or more and therefore the transcription factor can regulate the mRNA production by using the DNA-binding domain. There are four domains in the general transcription factor, DNA-binding domain, transcription regulation domain, nuclear localization signal domain and oligomerization site, and the transcription factor governs gene expression through these four domains: (1) The transcription factor can recognize the DNA nucleotide sequence by DNA-binding domain (2) Transcription regulation domain includes activation and repression function. For repression, the transcription repression domain can repress the mRNA transcription by binding on the gene operon or inhibiting other transcription factors. (3) Nuclear localization signal allows the transcription factors entering the nuclei to meet the DNA. (4) Oligomerization site is the region where different transcription factors interact with each other.

The main function of transcription factors is the activation or repression of certain gene expression by increasing or decreasing the amount of its mRNA. Recently, to validate the direct relationship between the transcription factor and its regulated gene, researchers use ChIP assay which can reveal the direct association of the transcription factor. Additionally, the knock-out methods, such as siRNA, are used to uncover the positive regulation and negative regulation of the transcription factor. Combining the knock-out methods and ChIP assay, researchers can learn the regulated gene of the transcription factor and its regulation type (activation or repression).

1.2 High through-put technologies

Because genes in the cell highly interact with each other, gene expression in the cell is much more complex. To understand the interaction between genes, the quantification of gene expression is essential. However, evaluating genes one-by-one is time-consuming and inefficient. Thus, many technologies have been developed in the last decades to quantify the expression of every gene simultaneously. These technologies either measure the mRNA concentration or protein concentration to portrait the gene expression in cells and tissues. In our study, we focused on the technologies that quantify in RNA level which can directly represent the genes transcribed in cells.

1.2.1 Microarray

Microarray is a basic type of high-throughput method and enables researchers to obtain genomic data accurately and efficiently (Jaksik *et al.*, 2015). By using nowadays high-tech, thousands of DNA molecules are well arranged in a centimeter silicon square chip as probes to detect target nucleic acid. According to the base pair rule, the probes can pair with the target transcriptome from samples.

In the microarray experiment, mRNAs from cells or tissues are reverse-transcript to the sample cDNAs which are labeled with the fluorescence molecule. Then the probes on the chip are hybridized with the fluorescence sample cDNAs. By using the base pair property, the probes can pair with the specific sample cDNAs, so-called hybridization. When the hybridization takes place on the chip, the fluorescence signal will show and the strength of fluorescence represents the abundance of mRNA. Therefore, the concentration of mRNAs can be quantified by the fluorescence signal.

1.2.2 RNA sequencing (RNA-seq)

Due to the development of next-generation sequencing, RNA-seq technology is widely used (Chu and Corey, 2012). In contrast to the microarray, the probe DNA is not required. RNA-seq is a transcriptomics research method based on next-generation sequencing and able to quantify mRNA in transcriptome samples. The first step in RNA-seq technology is to extract the transcriptome from cells or tissues and to transcribe from the sample transcriptome to cDNAs. To analyze the samples, the cDNAs are fragmented and sequenced by proper methods. Finally, the sequencing data of cDNA fragments are compared with genome data libraries to identify the transcriptome. The transcriptome can be quantified by the number of reads during the sequencing step.

1.2.3 Cap Analysis of Gene Expression (CAGE)

In the past decade, Hayashizaki developed Cap analysis of gene expression (CAGE) . As well as microarray and RNA-seq, CAGE can be used to determine the mRNA abundance. In the first step of CAGE, the capped mRNA is reversely transcribed to cDNA and this full-length cDNA is attached to the linker which can introduce the recognition site for the restriction enzyme. The restriction enzyme cleaves the cDNA with a linker at 20 and 18 nucleotides from the recognition site. Then the second linker is ligated with cleaved cDNA and the cleaved cDNA with the linkers is amplified by PCR. Finally, the cDNA is measured by high-throughput sequencing. The cleaved cDNA is a label of the mRNA and the concentration of cDNA from the high-throughput sequencing represents the abundance of the mRNA (Kodzius *et al.*, 2006).

1.3 Reverse engineering

The gene regulatory network (GRN) is important for many researches, such as cancer and developmental biology. Understanding the GRN can help biologists to decipher the mechanism of interactions between genes and to solve the target problem of certain genes. However, gene regulation is a complex process and an important issue. There are numerous high-throughput technologies nowadays and these technologies allow us to explore the gene expression pattern in the cells, tissues or organisms. Due to the expression profiles, reverse engineering can reconstruct GRN through algorithms. In the past decade, many computational methods of reverse engineering have been developed to unravel the GRN problem by using different expression profiles.

Figure. 1 illustrates the principal of reverse engineering. The basic way to infer the GRN of the organism by computational models is firstly to obtain the mRNA from cells. Since the many high-throughput technologies have been developed, the mRNA level is easy to quantify and the expression profiles are obtained by technologies. Subsequently, the gene expression profiles are the input data and the computational methods are applied to infer the GRN by computing the relationships in the gene expression profiles (Lee and Tzou, 2009).

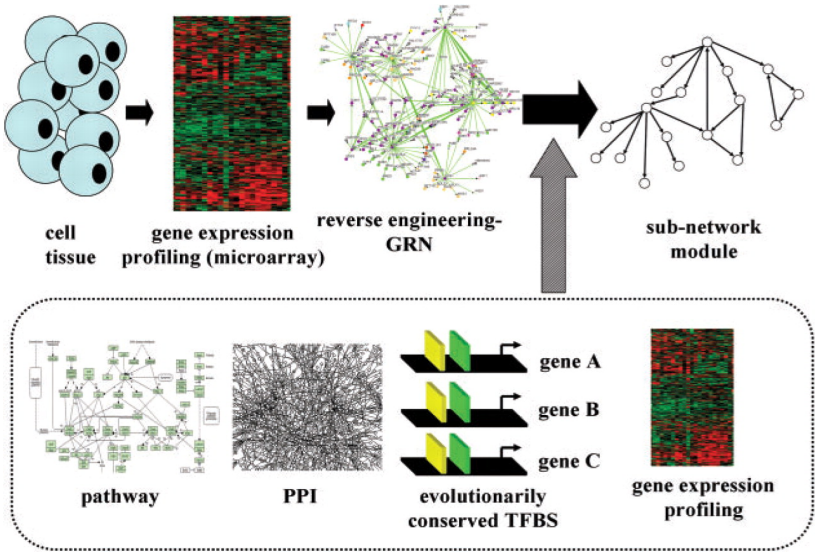

Figure 1. **The workflow of the GRN inference from gene expression profiles (Lee and Tzou, 2009)**

There are two types of gene expression profiles: dynamic and steady-state. The dynamic type is the time-series expression profiles which are measured in different time-points and show the change of mRNA level over time. Another is the steady-state expression profiles which are collected from different samples at a time. Both types of data suit different reverse-engineering methods. Besides the experimental gene expression profiles, the simulated expression profiles generated from *in silico* networks are often used for the evaluation of reverse-engineering methods as well. In this study, we used both experimental and simulated data.

The reverse-engineering methods can reconstruct the GRN from the gene expression profiles, but the inferred GRN is not always consistent with the real network. Therefore, it is important to evaluate the performance of reverse-engineering methods. The gold standard which is the real interactions between genes is the standard to evaluate the methods. The outputs of the methods are usually the putative regulatory interactions that are taken to compare to the gold standard interactions. The comparison assesses the number of true positive (TP, the putative interactions are in the gold standard), false positive (FP, the putative interactions are not in the gold standard), false negative (FN, the gold standard interactions which are not inferred) and true negative (TN, the interactions which are not inferred are not in the gold standard). Then indexes are used to measure the performance including sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)), precision (TP/(TP+FP)), FP rate (FPR, 1-specificity) and area under the curve (AUC, the curve is plotted with sensitivity against FPR) (Haury *et al.*, 2012).

1.3.1 Pearson correlation coefficient

Pearson correlation coefficient is used to calculate the linear relationship between two variables based on the covariance and the standard deviation of the two variables. The strength of the relationship of Pearson correlation coefficient is normalized between 1 and -1. The value on +1 or -1 indicates the perfect correlation, while the value on 0 means the two variables are independent. The positive sign manifests the positive correlation between variables and the negative sign is the negative correlation.

Definition:

$$\rho x, y = \frac{cov(x, y)}{\sigma x \sigma y}$$

Where cov is the covariance and σx, σy are the standard deviation of variables *x* and *y*. For the size n sample, we can estimate the covariance and the standard deviation and obtain the correlation coefficient r from the following equation:

$$r = \frac{\sum_{i=1}^{n}(xi - \mu x)(yi - \mu y)}{\sqrt{\sum_{i=1}^{n}(xi - \mu x)^2}\sqrt{\sum_{i=1}^{n}(yi - yx)^2}}$$

Where μx and μy are the mean of variable *x* and *y*.

1.3.2 Spearman's rank correlation coefficient

Spearman correlation coefficient can detect the non-linear relationship of two variables by considering the rank of values in the two variables. The value of Spearman correlation coefficient is normalized between 1 and -1.

Definition:

$$\rho(rx, ry) = \frac{cov(r(x), r(y))}{\sigma r(x)\ \sigma r(y)}$$

Where $cov(rx, ry)$ denote the covariance of the rank of variables $r(x)$, $r(y)$. σrx and σry are the standard deviations of the rank variables. For the size n sample, the correlation coefficient ρ can be used as the following equation:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2 - 1)}$$

Where the di is the difference between the two rank variables, $r(x) - r(y)$.

1.3.3 Signed distance correlation

Distance correlation or distance covariance serves to measure the dependence of two random vectors in arbitrary dimension and the value of distance correlation is always positive where zero in distance correlation implies independence. The method can be used in the bivariate or multivariate data to measure complicated dependence structures. For example, the distance covariance of n dimension vectors a and b need to calculate the Euclidean distance of all pairwise elements (Székely and Rizzo, 2009).

$$akl = \| Xk - Yl \|, \qquad k, l = 1,2, \dots, n$$

$$bkl = \| Xk - Yl \|, \qquad k, l = 1,2, \dots, n$$

ajk and bjk are the $n \times n$ matrices and then matrices akl and bkl are doubly centered.

$$Akl \approx akl - arow - acol + mean(akl)$$

$$Bkl \approx bkl - brow - bcol + mean(bkl)$$

Where $arow$ and $brow$ are the matrices of n column vectors of the mean values of each row, $acol$ and $bcol$ are the matrices of n row vectors of the mean values of each column, and $mean(akl)$ and $mean(bkl)$ are the n-by-n matrices that contain the identical mean values of akl and bkl. Then the distance covariance is derived from the square root of

$$Vn^2 = \frac{1}{n^2} \sum_{k,l=1}^{n} Akl\, Bkl = dCov(a, b)$$

In this study, we suggested a signed distance correlation. To determine the positive and negative correlation, the sign of distance correlation is determined by

$$Dkl = \| (ak, bk) - (al, bl) \|, \qquad k, l = 1,2, \dots, n$$

$$Dklsign = (ak - al) \times (bk - bl), \qquad k, l = 1,2, \dots, n$$

$$D = \sum_{k,l}^{n} sign(Dklsign) \times Dkl$$

Where ak and bk are the k-th element in vectors a and b, and al and bl are the l-th element in vectors a and b. The sign of distance covariance is

$$sign(dCov(a,b)\ ) = sign(D)$$

Then, the signed distance correlation is obtained by combining the sign and the distance covariance.

$$Signed\ distance\ correlation = sign(dCov(a,b)) \times dCov(a,b)$$

2. Materials and Methods

2.1 High-throughput gene expression profiles

Datasets of high-throughput steady-state gene expression profiles, including microarray, RNA-seq and CAGE dataset, were obtained from the published human cancer cell lines articles (Barretina *et al.*, 2012)(Klijn *et al.*, 2015) and the FANTOM5 project (Kawaji *et al.*, 2017). There are 266 cancer cell lines samples from various tissues and mRNA expression levels of 18926 human genes in the microarray dataset. The RNA-seq dataset contains 238 samples and expression data of 53478 transcripts. For the CAGE dataset, there are 31 cancer cell lines samples and 24649 annotations of transcripts.

To test in this study the performance of published GRN inference algorithms and the correlation-based methods, experimental and simulated time-series gene expression profiles were used. The experimented time-series gene expression dataset comes from the human THP-1 myeloid monocytic leukemia cell line which was done by a single-cell experiment (Kouno *et al.*, 2013). The two simulated data were downloaded from the DREAM4 challenge and were generated by the GeneNetWeaver software (Marbach *et al.*, 2009)(Schaffter, Marbach and Floreano, 2011)(Stolovitzky, Monroe and Califano, 2007)(Stolovitzky, Prill and Califano, 2009).

2.2 Gold standard construction

The gold standard were collected from published articles and were validated by experimental methods. The ChIP assay conducted in the experiments were chosen to validate the direct regulations. The gold standard list contains the transcription factors, their regulated gene and the type of the regulation, from were we find 503 positive regulations and 126 negative regulations. For the time-series data, the gold standard of the experimental THP-1 single-cell dataset was validated by Matrix RNAi analysis (Kouno *et al.*, 2013)(Tomaru *et al.*, 2009) and the gold standards of the simulated data are provided by the DREAM4 challenge in the GeneNetWeaver website.

2.3 Normalization and steady-state data analysis

The RNA-seq and the CAGE strategies measured multiple transcripts of a single gene. To procedure with the analysis, the expression level of a gene was represented by an average value of all the transcripts of that gene. Then the z-score and log normalization were applied to the datasets before the correlation calculation. Later, the correlation methods (Pearson, Spearman and signed distance correlation) were applied to the normalized datasets.

To measure the performance of the correlation-based methods by the constructed gold standard list, the transcription factor genes and the regulated genes in both constructed gold standard and the normalized datasets were extracted. Subsequently, the correlations between extracted transcription factor genes and regulated genes were computed and compared to the gold standard list (Figure 2).

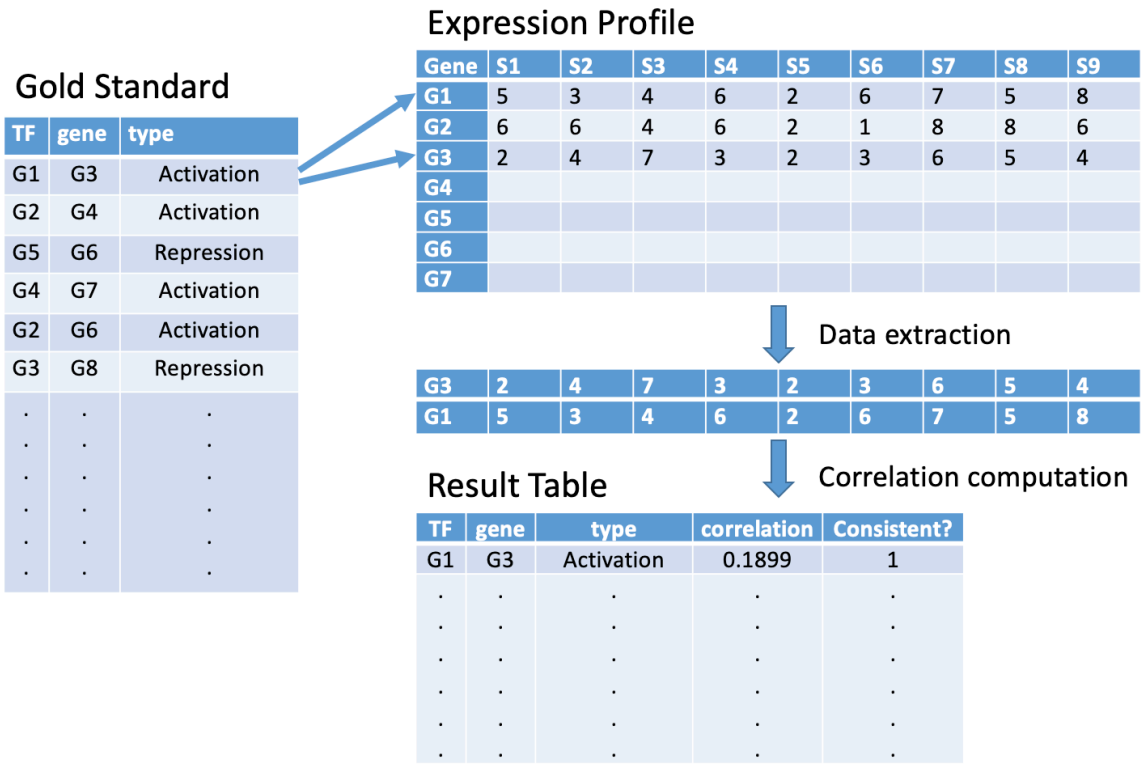

Figure 2. **The workflow of methods result and the gold standard comparison**

In this study, the positive regulation (activation) was defined as the positive class, while the negative regulation (repression) was defined as the negative class. Hence, different performance measures such as sensitivity, specificity, F1-score positive, F1-score negative, area under receiver operating characteristics curve (AUC), precision-positive and precision-negative can be calculated (see the following equations).

Equations:

$$TP = \ number\ of\ inferred\ positive\ correlations\ which\ consist\ with\ the\ truth$$

$$TN = \ number\ of\ inferred\ negative\ correlations\ which\ consist\ with\ the\ truth$$

$$FP = \ number\ of\ inferred\ positive\ correlations\ \ which\ do\ NOT\ consist\ with\ the\ truth$$

$$FN = \ number\ of\ inferred\ negative\ correlations\ which\ do\ NOT\ consist\ with\ the\ truth$$

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

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

$$precision\ positive = \frac{TP}{TP + FP}$$

$$precision\ negative = \frac{TN}{TN + FN}$$

$$F1\ positive = \ 2 \times \left(\frac{precision\ positive\ \times\ sensitivity}{precision\ positive + sensitivity}\right)$$

$$F1\ negative = \ 2 \times \left(\frac{precision\ negative\ \times\ specificity}{precision\ negative + specificity}\right)$$

2.4 Network-based reverse-engineering methods and dynamic expression data analysis

To compare the performance of correlation-based methods, reverse-engineering methods that can reveal the regulation type between genes were chosen, such as TSNI (Bansal, Gatta and di Bernardo, 2006), GRNVBEM (Sanchez-Castillo *et al.*, 2018) and SINCERITIES (Papili Gao *et al.*, 2018). The output of the reverse-engineering methods consists on a list containing regulators, regulated genes and the regulation type. Therefore, the performance of the methods can be evaluated by comparing the output list with the gold standard of input datasets. For these reverse-engineering methods, the input dataset was the dynamic expression profile. Subsequently, the THP1 single-cell expression profile and simulated expression profiles were also analyzed by these and the correlation-based methods.

3. Results

3.1 Correlation-based method

We used the correlation-based methods to analyze the genome-wide microarray, RNA-seq and CAGE datasets. Note that in this section the gold standard ground truth is constructed by published experimental data. To compare the datasets we first present the performance results of different technologies across different normalizations (Figure 3). Here, each bar in the plot was calculated by averaging the performances obtained by three different correlation measures (Pearson, Spearman and signed distance correlations) for each case. The result shows that, following correlation directions (positive or negative) it is easier to get the up- or down-regulations by means of microarray compared to the other two technologies, followed by RNA-seq in a second place in a normalization-free framework (Figure 3A). However, it is difficult to say assertively that RNA-seq worked better than CAGE since in the specificity and F1-score negative, the CAGE dataset showed better results. For the log normalization framework (Figure 3B), the results follow similar trends compared to the datasets without normalization, where the microarray dataset shows the best performance. Whereas the results in the Z-score normalization framework (Figure 3C) displays variety. The sensitivity which indicates the true positive rate of activation associations shows that the CAGE dataset with Z-score is the highest (also compared to other normalization scenarios) while the specificity which indicates the true negative rate (repression associations) shows that the microarray dataset has better outcome than for the case of the CAGE technology. Therefore, the microarray dataset has better detection than the RNA-seq and the CAGE dataset in all the normalizations when it comes to correlation methods analysis.

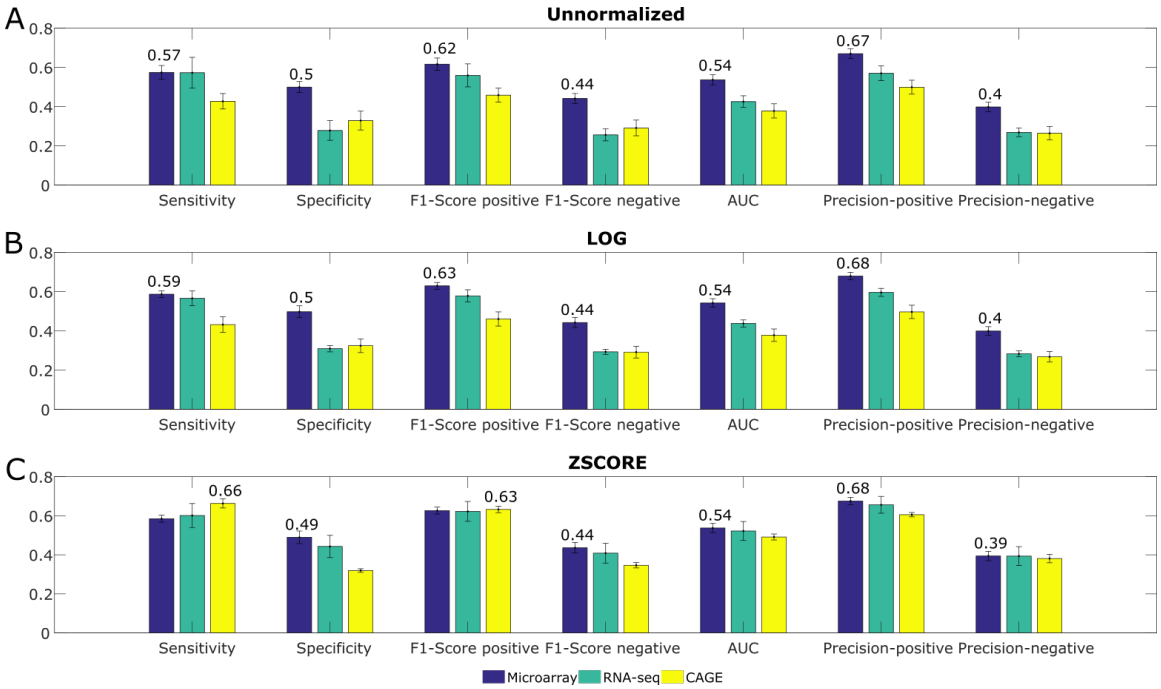

Figure 3. **Performance of technologies across different normalizations.**

The different bar colors represent the datasets obtained by the technologies in question: microarray (blue), RNA-seq (green) and CAGE (yellow), and each group of three bars represent an evaluation by a different measure. Each bar consists on the average performance obtained by three correlation measures. A) Raw datasets (unnormalized), B) datasets with log normalization and C) datasets with Z-score normalization.

Next, we considered that different correlation-based methods may fit for different dataset, so we perform the result of technologies across different correlation-based methods, where the average performance was taken according to the different normalization frameworks (Figure 4). For all the correlation-methods employed, the microarray always has the highest sensitivity and specificity that shows the best performance to detect gene activation and repression. Taking all together from Figure 3 and Figure 4, the microarray dataset seems to be the best option for regulation type detection in correlation-based methods and the different normalization frameworks.

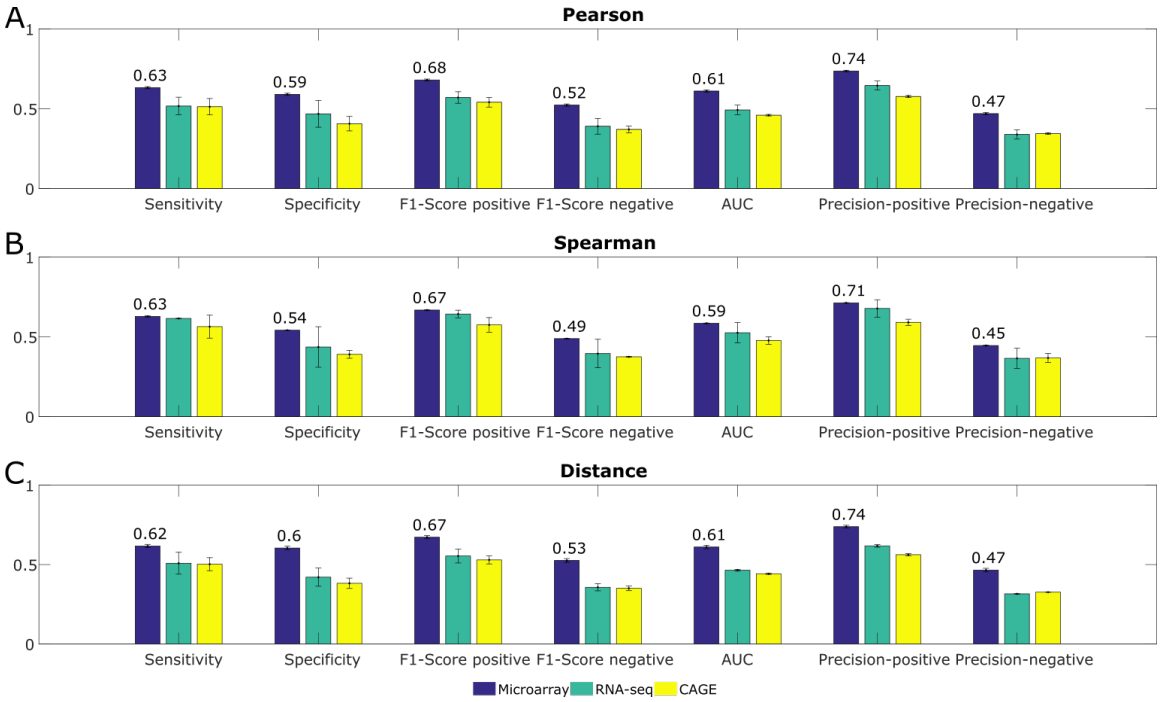

Figure 4. Performance of technologies across different correlations.

The three different bar colors represent the datasets obtained by the technologies: the microarray (blue), RNA-seq (green) and CAGE (yellow) dataset, and each group of three bars represents an evaluation by different measure. Each bar represents the average performance obtained by three normalization measures. A) Datasets measured by Pearson correlation, B) datasets measured by Spearman correlation and C) datasets measured by signed distance correlation.

To learn which normalization is suitable for each technology, we compared the different normalizations performance within each dataset. We calculated the average statistical measures of different correlation methods in the microarray, RNA-seq and CAGE datasets, and checked their performances (Figure 5). In the microarray dataset (Figure 5A), the performances of different normalizations are not much different and therefore it is infer that the type of normalization slightly affects regulation type detection here. Nevertheless, the Z-score normalization provides the strongest performance in the RNA-seq and CAGE datasets (Figure 5B and 5C). The difference between these two relies that Z-score is much better for repression associations detection according to the highest specificity in RNA-seq, whereas for the CAGE dataset, we observed that the Z-score normalization has the strongest performances in sensitivity, F1-score positive and precision-positive suggesting that the Z-score normalization has better activation associations detection for the CAGE dataset.

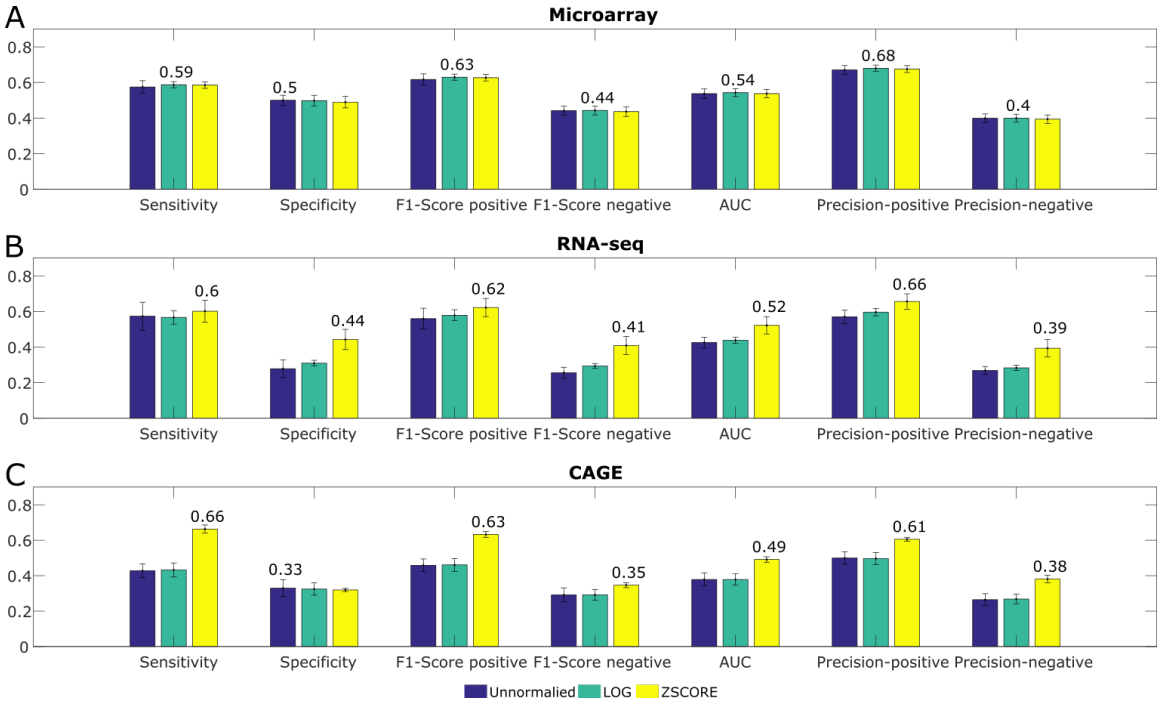

Figure 5. The performance of normalizations across different technologies.

The three different bar colors represent the datasets with different normalization: Unnormalized (blue), log normalization (green) and Z-score normalization (yellow), and each group of three bars represents an evaluation by different measure. Each bar represents the average performance obtained by three correlation measures. A) microarray dataset, B) RNA-seq dataset and C) CAGE dataset.

To determine the performance of the three suggested correlation methods, we took the average performances from each correlation method according to different normalization frameworks and presented them for each dataset technology (Figure 6). We observed that Pearson correlation and signed distance correlation in the microarray dataset (Figure 6A) have overall higher performance than Spearman correlation although spearman is still close to the two previous mentioned correlations. However, Spearman correlation is relatively better when it is applied to both RNA-seq and CAGE datasets (Figure 6B and 6C).

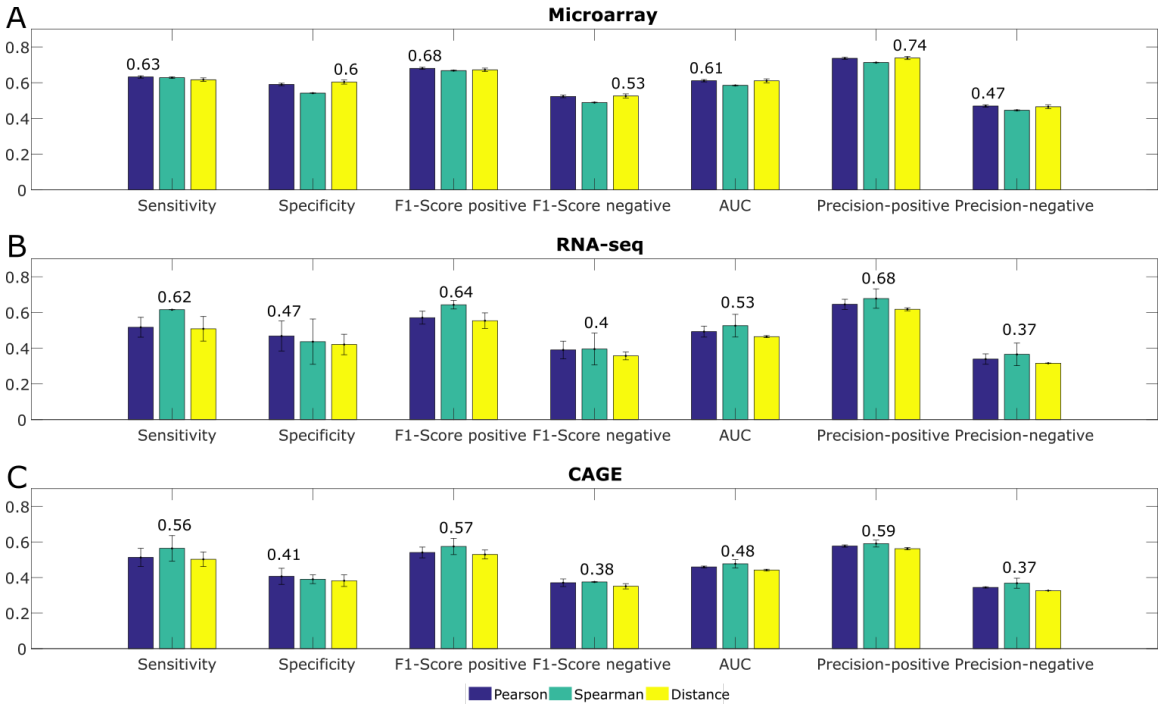

Figure 6. **Performance of correlations across different technologies.**

The three different bar colors represent the datasets measured by different correlations: Pearson correlation (blue), Spearman correlation (green) and signed distance correlation (yellow), and each group of bars represents an evaluation by different measure. Each bar represents the average performance obtained by three normalization measures. A) microarray dataset, B) RNA-seq dataset and C) CAGE dataset.

To know the best performance of the correlation with a normalization, we displays the best AUC of the correlation in each technologies (Figure 7). In the microarray technologies, the signed distance correlation shows the best performance when applied to the non-normalized microarray dataset, and Pearson correlation applied to the log normalized is the second place (Figure 7A). In both RNA-seq and CAGE, Spearman with Z-score normalization has the best performance, and Pearson correlation with Z-score normalization is following (Figure 7B and 7C). To sum up, the signed distance correlation applied to the raw microarray dataset and Spearman correlation applied to both Z-score-normalized RNA-seq and CAGE datasets perform best.

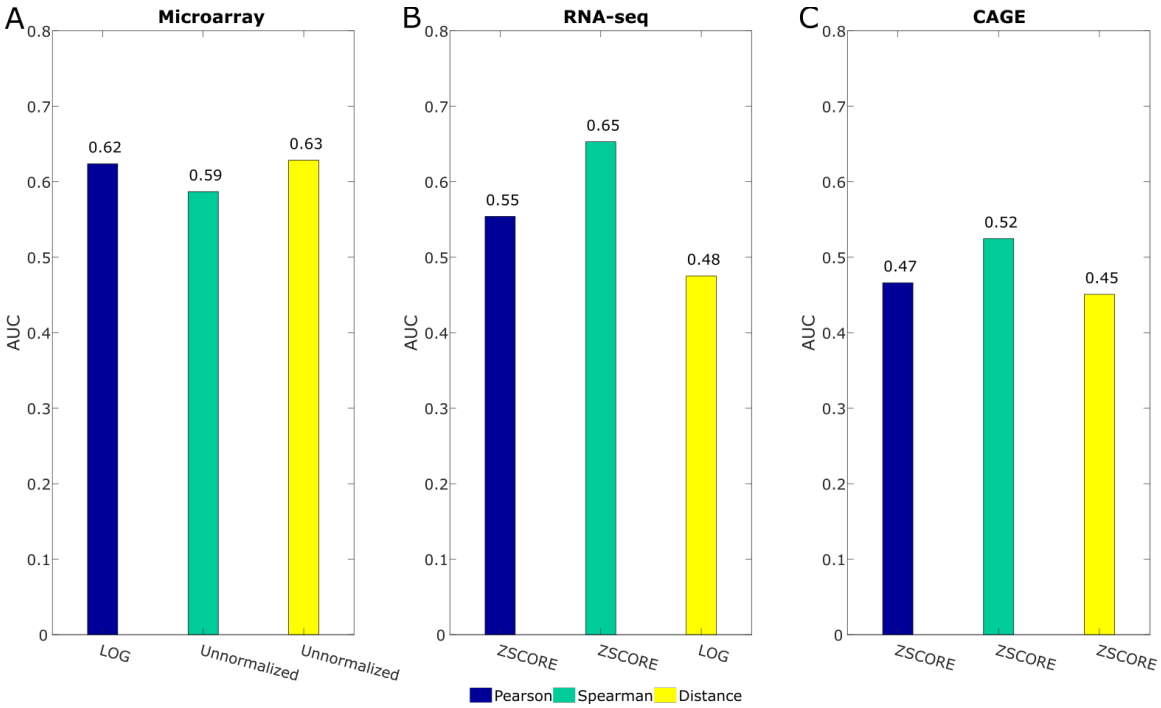

Figure 7. Best performance correlation across different technologies.

Each bar represents the best AUC of a correlation combined with a normalization. The x axis represents the normalization of the dataset. The order of the bar (from left to right): Pearson correlation (blue), Spearman correlation (green), signed distance correlation(yellow). A) Microarray, B) RNA-seq and C) CAGE.

3.1 Network-based method comparison

In this section, we compared the correlation-based methods against the GRN-based methods, such as TSNI, GRNVBEM and SINCERITIES, by using the THP1 single-cell expression profiles, simulated 1 and simulated 2 datasets. The THP1 single cell expression profiles contains 8 time-points and simulated data 1 and simulated data 2 have 21 time-points. The gold standard of THP1 data was constructed by the matrix RNA assay (Tomaru *et al.*, 2009) and the gold standards of the simulated data are synthetic networks from GeneNetWeaver. We first compared the datasets for each normalization framework and we obtained similar situations across them. Both simulated data have much stronger performance than the experimental THP1 single-cell microarray data, and here, the simulated data 1 shows better results than the simulated data 2 (Figure 8).

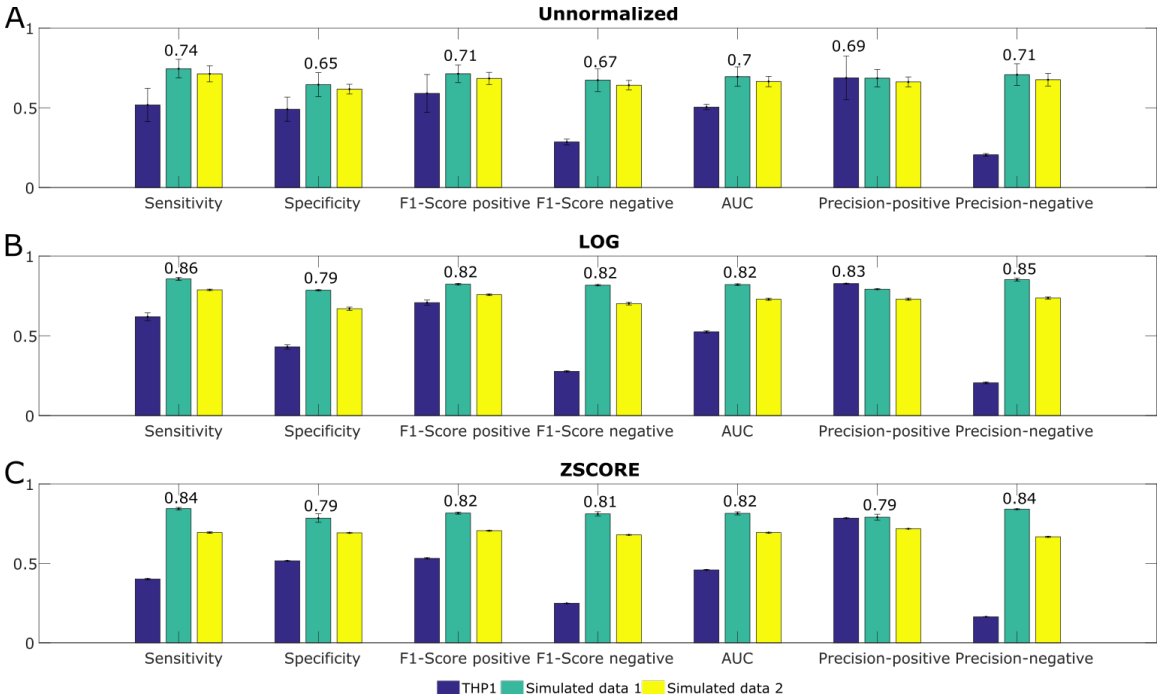

Figure 8. **Performance of datasets across different normalizations.**

The three different bar colors represent the different datasets: The THP1 single-cell dataset (blue), simulated data 1 (green) and simulated data 2 (yellow), and each group of three bars represents an evaluation by different measure. Each bar represents the average performance obtained by different methods measures. A) Raw datasets (unnormalized), B) datasets with the log normalization and C) datasets with Z-score normalization.

We then compared the datasets according to correlation-methods by averaging the different normalization results (Figure 9). In concordance to Figure 8, Figure 9 also suggests that the best dataset to obtain gene regulation types is the simulated data 1 and across all the correlation-based methods, the simulated data 1 and simulated data 2 have much higher performance values than THP1 with exception on the precision-positive. This could be explain because, from the few positives regulations detected in THP1, several of them are true.

This clear trends however seems to differ for the network-based methods, TSNI, GRNVBEM and SINCERITIES (Figure 10). In the TSNI, the THP1 dataset has the highest sensitivity value, but in both GRNVBEM and SINCERITIES methods the THP1 dataset has the lowest sensitivity. Notably, the THP1 dataset has the highest precision-positive value and the lowest precision-negative across all the methods. Since the sensitivity of the THP1 dataset is not the highest, the high precision-positive implies that the number of false positive detection is low in the THP1 dataset. In conclusion, correlation-methods seems to obtain higher performances than network-based methods for the detection of gene regulation type.

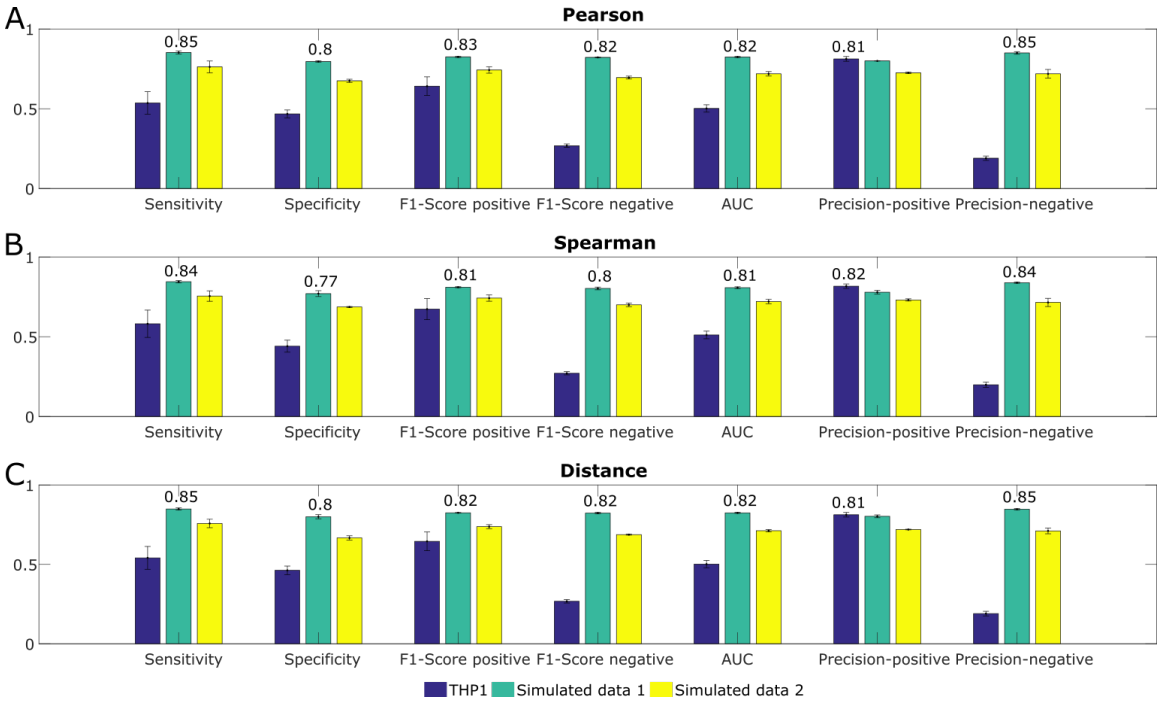

Figure 9. **Performance of datasets across different correlations.**

The three different bar colors represent the different datasets: The THP1 single-cell dataset (blue), simulated data 1 (green) and simulated data 2 (yellow), and each group of three bars represents an evaluation by different measure. Each bar represents the average performance obtained by different normalization measures. A) Datasets measured by Pearson correlation, B) datasets measured by Spearman correlation and C) datasets measured by signed distance correlation.

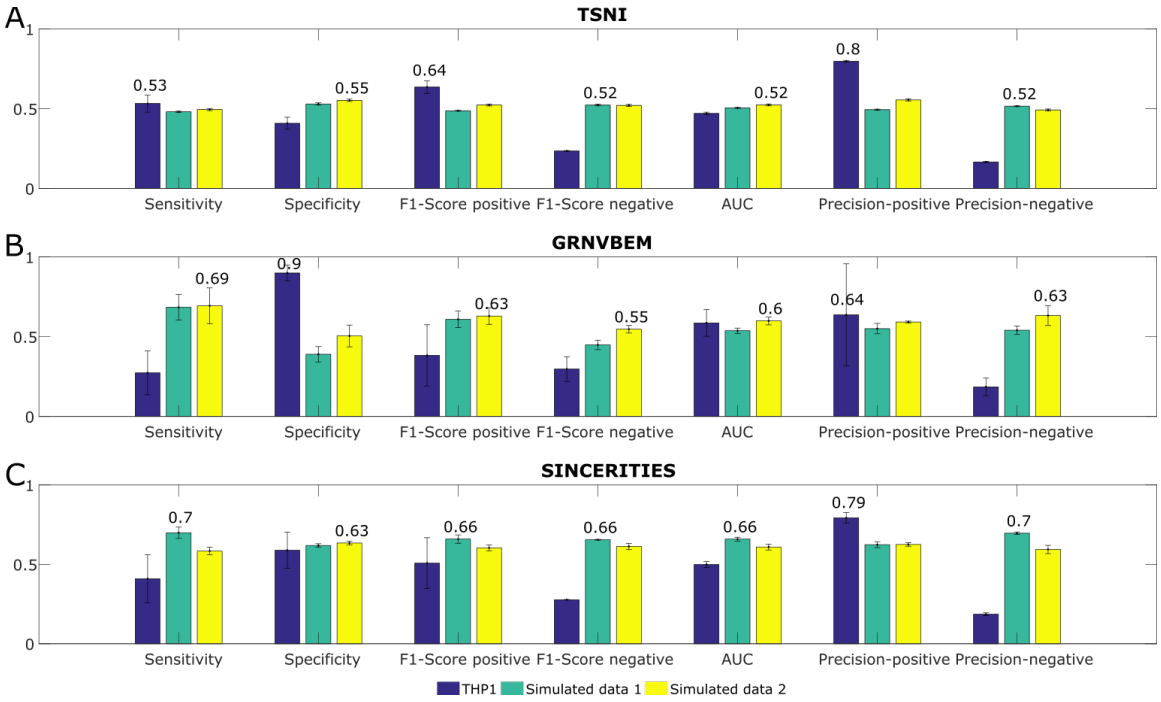

Figure 10. **Performance of datasets across different network-based methods.**

The three different bar colors represent the different datasets: The THP1 single-cell dataset (blue), simulated data 1 (green) and simulated data 2 (yellow), and each group of three datasets bars represents an evaluation by different measure. Each bar represents the average performance obtained by different normalization measures. A)Datasets are measured by TSN1, B) datasets measured by GRNVBEM and C) datasets measured by SINCERITIES.

To compare the normalizations in different dataset, we performed the normalizations comparison result across all the datasets, and the different results are obtained by averaging the different correlation and network-based method performances (Figure 11). The result shows that the log normalization seems to be the best option since it fits for all methods and datasets. However, the performances between normalizations do not differ tremendously. In general, it seems that normalized data works better than to not apply a normalization.

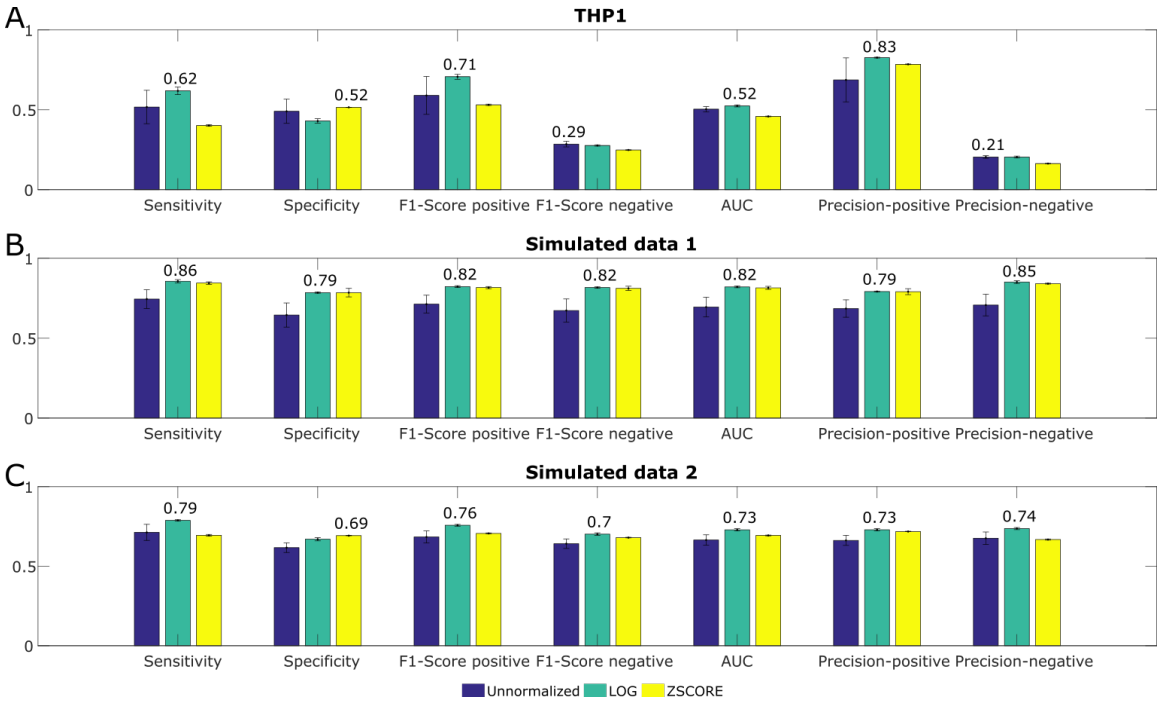

Figure 11. Performance of normalizations across different datasets.

The three different bar colors represent the datasets with different normalizations: Unnormalized (blue), log normalization (green) and Z-score normalization (yellow), and each group of three bars represents an evaluation by different measure. Each bar represents the average performance obtained by different method measures. A) THP1 single-cell dataset, B) simulated data 1 and C) simulated data 2.

It is important to identify which methods are suitable for regulation type detection. Therefore, we compared all the correlation-based and network-based methods across all the datasets (Figure 12). In the experimental THP1 dataset, the networked-based method GRNVBEM seems to be better than other methods for repression detection, contrary of what happen for activation detection (specificity is the highest and the sensitivity is the lowest, with all measures having high variance). In both simulated datasets, we observed that the correlation-based methods (with similar performances) are relatively better than network-based methods and the best network-based method seems to be SINCERITIES for the detection of gene type regulation.

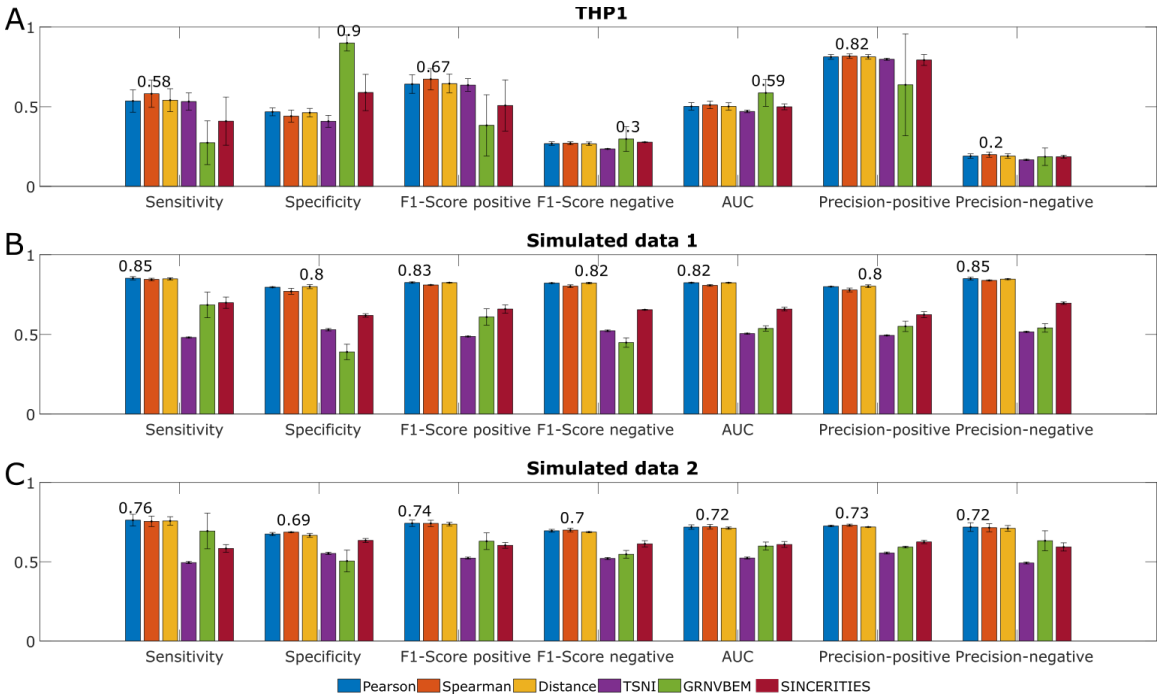

Figure 12. **Performance of methods across different datasets.**

The six different bar colors represent the datasets measured by different methods: Pearson correlation (blue), Spearman correlation (orange), signed distance correlation (yellow), TSNI (purple), GRNVBEM (green) and SINCERITIES (red), and each group of six bars represents an evaluation by different measure. Each bar represents the average performance obtained by different normalization measures. A) THP1 single-cell dataset, B) simulated data 1 and C) simulated data 2

Then we present the best performance of each method in the THP1 dataset, simulated data 1 and simulated data 2 (Figure 13). The GRNVBEM applied to the THP1 dataset with Z-score normalization shows the best performance (AUC = 0.71) and the second places are Spearman correlation and signed distance correlation applied to the raw THP1 data (AUC = 0.54). In Figure 12, it is shown that in both simulated data, the different methods are in general better than in THP1. Here in the simulated data 1, all the correlation-based method have good performance, and Pearson correlation applied to the log dataset and signed distance correlation applied to the Z-score dataset have the highest AUC (AUC = 0.83). For the simulated data 2, the correlation-based methods are also good, since the best performances are the Pearson with the log dataset and the Spearman with raw data (AUC = 0.74). Interestingly, Pearson correlation, Spearman correlation and SINCERITIES with same normalization in three datasets have the best performances.

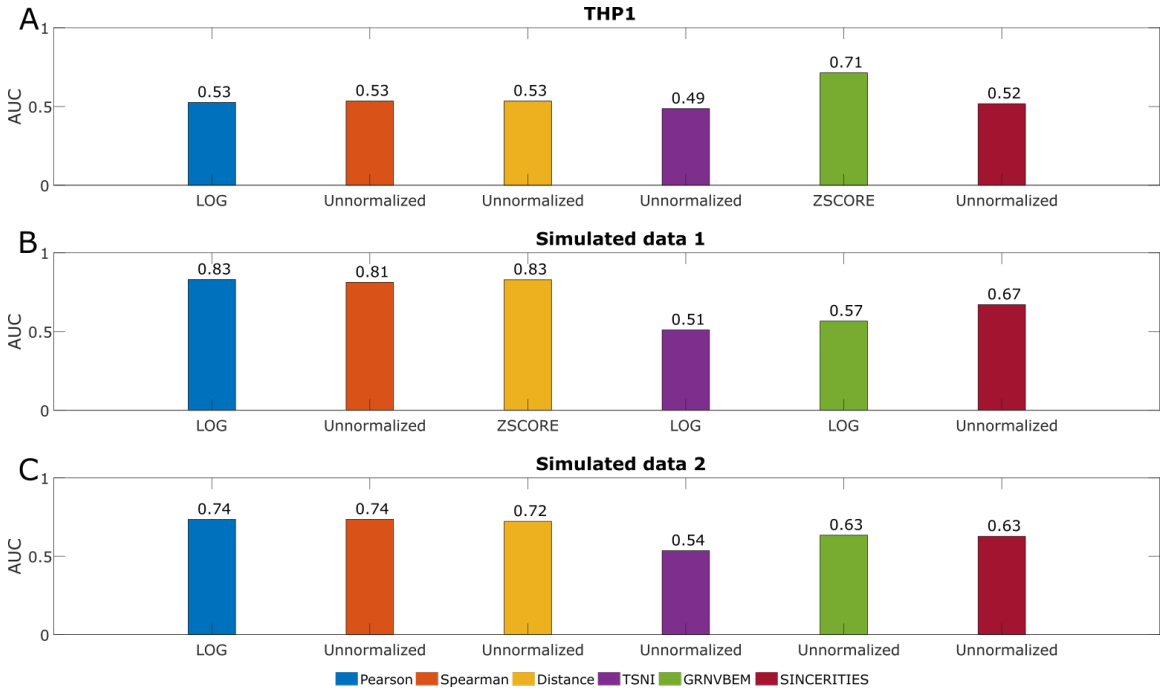

Figure 13. **Best performance of methods across different datasets.**

There are six bars in a subplot and each bar represents the best AUC of a method combined with a normalization. The order of the bar (from left to right): Pearson correlation, Spearman correlation, signed distance correlation, TSNI, GRNVBEM and SINCERITIES. The datasets: A) THP1 single-cell data, B) simulated data 1 and C) simulated data 2.

We generated different levels of noise data for the THP1 single-cell dataset, simulated data1 and simulated data 2 and we checked the performance of each method (Figure 14). The normalized datasets were chosen according to Figure 13. For example, we applied Pearson correlation to the different noise levels of log normalized THP1 dataset, the log normalized simulated data 1 and the log normalized simulated data 2.The signal noise ratio (SNR) value determine the noise that the data contains and the lower the value the higher is the noise presented. From the THP1 dataset, the AUC and precision value of most methods are not much different when the noise is getting higher. However, the GRNVBEM has huge variance in different noise level datasets. Additionally, the GRNVBEM cannot detect the negative correlation in SNR 10 data. The noisy THP1 data result implies that GRNVBEN is bias to some datasets. In the simulated data 1 and simulated data 2, there is a trend since the AUC value and precision of methods are weak in the strong noisy data. In noisy simulated data result, we still observed that the correlation-based methods have better detection than the networked-based methods for gene regulation types.

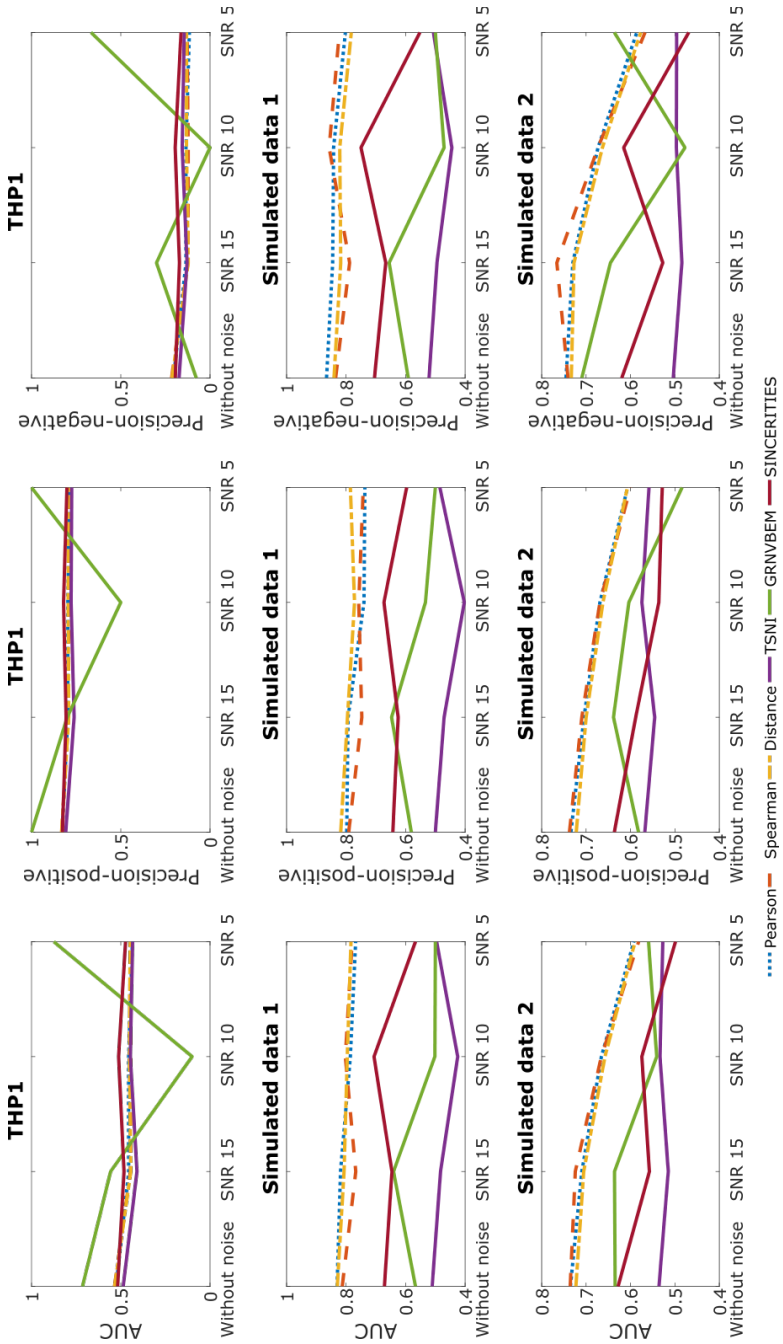

Figure 14. Noise data.

Six curves represent different methods. Dashed lines correspond to correlation methods whereas solid lines to GRN methods: Pearson correlation (blue), Spearman correlation (orange), signed distance correlation (yellow), TSNI (purple), GRNVBEM (green) and SINCERITIES (red). Each column of subplots display a statistical measure (AUC, precision positive and precision negative) and each row of subplots display different dataset (THP1, Simulated data 1 and 2). The different methods are applied to different normalized data according to the best performance result (Figure 13). The data on x-axis from left to right are normalized or raw data, SNR 15 data, SNR 10 data and SNR 5 data.

4. Discussion

This study focuses on the determination of regulation types for transcription factors from different expression profiles. We suggested that correlation-based methods can be used for this purpose and we compared them in different normalization and datasets situations. It is important to consider that gene regulation is a complicated process and after regulation type prediction, the causality is still a need for experimental validation. For example, the correlation-based methods can only identify the relationship between two genes, but this relationship can occur with more than one possibility of interaction between both genes. Thus, all interaction predictions of transcription factors should be verified by experiments (Penfold and Wild, 2011).

Additionally, one drawback of this study is data-driven since most of the transcription factors upregulate their regulated gene and therefore there is an unbalanced number of positive regulations and negative regulations in the gold standard data we constructed. For the microarray, RNA-seq and CAGE part, we observed that specificity values are always lower than sensitivity values which causality may be the same as for the constructed gold standard data.

Another issue in the expression profiles is encountered in the CAGE dataset, which was shown to have the weakest performance. This could be explained because the genome-wide CAGE dataset contains only 33 samples, a lower amount than in the case of the microarray and RNA-seq datasets. Additionally the mRNA level of many genes are not detected by CAGE which can cause the inaccuracy.

In the first section result, we can find that the performances of different technologies have same pattern. The microarray datasets has the better performance than other two high-throughput technologies, RNA-seq and CAGE regardless of normalizations and correlations. But we tested only one dataset of each technology. This should be confirmed by testing more datasets.

There are many published state-of-art methods to infer networks from expression profiles. However, many of them are not showing the regulation type in their inferred network (Chan, Stumpf and Babtie, 2017)(Deshpande *et al.*, 2019)(Huynh-Thu *et al.*, 2010)(Pratapa *et al.*, no date)(Specht and Li, 2017)(Qiu *et al.*, 2018). Therefore, we only chose the network-based methods which can indicate the activation and repression in the inferred network, such as TSNI, GRNVBEM and SINCERITIES.

Nevertheless, without considering the drawbacks, and taking into account that all methods were applied in the same frameworks, a clear improvement in performance for gene regulation type prediction is gained with correlation methods compared with GRN-methods.

In conclusion, here we manifested that the correlation-based methods can be useful to identify the activation and repression of transcription factors in gene regulation which provides a completely new point in the network inference research. The datasets generated by different high-throughput technologies might carry different properties and therefore the performance results might differ between them. Hence, it may improve the accuracy when the methods applied to normalized datasets, Pearson correlation in our case for example. Finally, comparing to the networked-based methods, although correlation-based methods provided better performances to detect the relationship between genes, more studies in this direction are needed for the correct application of these correlation-based methods.

References

Bansal, M., Gatta, G. Della and di Bernardo, D. (2006) 'Inference of gene regulatory networks and compound mode of action from time course gene expression profiles', *Bioinformatics*, 22(7), pp. 815–822. doi: 10.1093/bioinformatics/btl003.

Barretina, J. *et al.* (2012) 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*. doi: 10.1038/nature11003.

Chan, T. E., Stumpf, M. P. H. and Babtie, A. C. (2017) 'Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures', *Cell Systems*. doi: 10.1016/j.cels.2017.08.014.

Chu, Y. and Corey, D. R. (2012) 'RNA sequencing: Platform Selection, Experimental Design, and Data Interpretation', *Nucleic Acid Therapeutics*.

Deshpande, A. *et al.* (2019) 'Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data', *bioRxiv*. doi: 10.1101/534834.

Haury, A.-C. *et al.* (2012) 'TIGRESS: Trustful Inference of Gene REgulation using Stability Selection', *BMC Systems Biology*, 6(1), p. 145. doi: 10.1186/1752-0509-6-145.

Hollern, D. P. *et al.* (2014) 'The E2F Transcription Factors Regulate Tumor Development and Metastasis in a Mouse Model of Metastatic Breast Cancer', *Molecular and Cellular Biology*, 34(17), pp. 3229–3243. doi: 10.1128/MCB.00737-14.

Huynh-Thu, V. A. *et al.* (2010) 'Inferring Regulatory Networks from Expression Data Using Tree-Based Methods', *PLoS ONE*. Edited by M. Isalan, 5(9), p. e12776. doi: 10.1371/journal.pone.0012776.

Jaksik, R. *et al.* (2015) 'Microarray experiments and factors which affect their reliability', *Biology Direct*. doi: 10.1186/s13062-015-0077-2.

Kawaji, H. *et al.* (2017) 'The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types', *Scientific Data*. doi: 10.1038/sdata.2017.113.

Klijn, C. *et al.* (2015) 'A comprehensive transcriptional portrait of human cancer cell lines', *Nature Biotechnology*. doi: 10.1038/nbt.3080.

Kodzius, R. *et al.* (2006) 'Cage: Cap analysis of gene expression', *Nature Methods*. doi: 10.1038/nmeth0306-211.

Kouno, T. *et al.* (2013) 'Temporal dynamics and transcriptional control using single-cell gene expression analysis', *Genome Biology*. doi: 10.1186/gb-2013-14-10-r118.

Lee, W. P. and Tzou, W. S. (2009) 'Computational methods for discovering gene networks from expression data', *Briefings in Bioinformatics*. doi: 10.1093/bib/bbp028.

Marbach, D. *et al.* (2009) 'Generating realistic in silico gene networks for performance assessment of reverse engineering methods', *Journal of Computational Biology*. doi: 10.1089/cmb.2008.09TT.

Papili Gao, N. *et al.* (2018) 'SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles', *Bioinformatics*. doi: 10.1093/bioinformatics/btx575.

Penfold, C. A. and Wild, D. L. (2011) 'How to infer gene networks from expression profiles, revisited', *Interface Focus*, 1(6), pp. 857–870. doi: 10.1098/rsfs.2011.0053.

Pratapa, A. *et al.* (no date) 'Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data', *bioRxiv*. doi: http://dx.doi.org/10.1101/642926.

Qiu, X. *et al.* (2018) 'Towards inferring causal gene regulatory networks from single cell expression Measurements', *bioRxiv*, p. 426981. doi: 10.1101/426981.

Sanchez-Castillo, M. *et al.* (2018) 'A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data', *Bioinformatics*. doi: 10.1093/bioinformatics/btx605.

Schaffter, T., Marbach, D. and Floreano, D. (2011) 'GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods', *Bioinformatics*. doi: 10.1093/bioinformatics/btr373.

Specht, A. T. and Li, J. (2017) 'LEAP: Constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering', *Bioinformatics*. doi: 10.1093/bioinformatics/btw729.

Stolovitzky, G., Monroe, D. and Califano, A. (2007) 'Dialogue on reverse-engineering assessment and methods: The DREAM of high-throughput pathway inference', in *Annals of the New York Academy of Sciences*. doi: 10.1196/annals.1407.021.

Stolovitzky, G., Prill, R. J. and Califano, A. (2009) 'Lessons from the DREAM2 Challenges', *Annals of the New York Academy of Sciences*. doi: 10.1111/j.1749-6632.2009.04497.x.

Székely, G. J. and Rizzo, M. L. (2009) 'Brownian distance covariance', *Annals of Applied Statistics*. doi: 10.1214/09-AOAS312.

Tomaru, Y. *et al.* (2009) 'Regulatory interdependence of myeloid transcription factors revealed by Matrix RNAi analysis', *Genome Biology*, 10(11), p. R121. doi: 10.1186/gb-2009-10-11-r121.

Wang, X. *et al.* (2008) 'Short time-series microarray analysis: Methods and challenges', *BMC Systems Biology*. doi: 10.1186/1752-0509-2-58.

Zhang, X. *et al.* (2016) 'FRA1 promotes squamous cell carcinoma growth and metastasis through distinct AKT and c-Jun dependent mechanisms', *Oncotarget*. doi: 10.18632/oncotarget.9110.

Declaration of Research Integrity and Good Scientific Practice

I hereby certify that I have authored this Master's Thesis entitled "Correlation-based detection of activation /repression transcription factor regulation from gene expression" and without undue assistance from third parties. No other than the resources and references indicated in this thesis have been used. I have marked both literal and accordingly adopted quotations as such. There were no additional persons involved in the spiritual preparation of the present thesis. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.

Place, date

signature
Name