Regulation Expression Pathway Analysis (REPA): A novel method to facilitate biological interpretation of high throughput expression profiling data

Master of Science

Department of Computer Science

Memorial University of Newfoundland

Pranjal Patra

June 2015

Abstract

In the past decade there have been great advances and emergence of new techniques in the field of gene expression profiling. As the popularity of these techniques grew, the amount of data that gets generated has also grown. The task of analyzing this data to create a global picture to identify the biological pathways that are relevant to the study has been addressed by many. These approaches (collectively termed as enrichment analysis) have also grown in sophistication and accuracy making them the default step following a gene profiling experiment. However, enrichment analysis approaches do not provide pointers to likely regulators in their results.

In this project we built a system called Regulation Expression Pathway Analysis or REPA to facilitate the biological interpretation of results from high throughput gene expression profiling experiments. In particular, we provide researchers with gene sets that were most active in the biological phenomenon under study and their likely regulators. Users can input the gene expression profile data from their expression profiling experiments in REPA and get a list of disturbed gene sets and inferred transcription factors that possibly regulate these gene sets.

To build this system first we processed the transcription factor binding data from the ENCODE project to quantify the strength of regulation that each transcription factor has on each gene set. Then we build a gene expression enrichment analysis system that can analyze the gene expression profiling data and list the most active

gene sets. Finally we combine the results from the previous two steps to arrive at a more complete picture that gives users information about not only the most active gene sets, but also about the most likely regulators of these gene sets.

Acknowledgements

I wish to thank several people for their advice and support throughout this dissertation process. Special gratitude is owing to Dr. Lourdes Peña-Castillo, my supervisor, whose direction, consideration and encouragement were exemplary and greatly appreciated.

Guidance from many professors especially Dr. Oscar Meruvia-Pastor and support from my colleagues and friends was helpful in completing this thesis.

I also wish to dedicate this work to Maa and Baba for all their love and sacrifices for me and Rippy.

Contents

List of Figures

Chapter 1

Introduction

It is now common knowledge that the entire genetic information of an organism is coded in its DNA. Therefore the knowledge of the exact sequence of the DNA of an organism is a valuable resource in the quest to understand how that organism functions. In recent times great advances were made in the development of techniques that enable high speed DNA sequencing. Any method or technology that can determine the exact order of the four bases of DNA is called a DNA sequencing technique. Another method that have grown in popularity is the expression profiling techniques which allow researchers to check the activity levels of thousands of genes at once. These gene expression profiling techniques are extremely useful in determining the functions of genes. The data generated from such methods is huge, but without proper interpretation, is not of much use.

Our project is in the field of bioinformatics, which is an interdisciplinary field dealing with the development and use of computer software and databases to facilitate and enhance biological research. The main objective of this project is to facilitate the biological interpretation of the results from gene expression profiling experiments. In particular, we provide researchers with gene sets (or biological pathways) associated to the biological phenomenon under study and their likely regulators. Gene sets are set of genes which have some feature in common such as genes that are involved in a pathway. To achieve this goal we built a software system called Regulation Expression Pathway Analysis (REPA) where users can input the gene expression profile data from their expression profiling experiments and get a list of disturbed gene sets and inferred transcription factors that possibly regulate these gene sets.

REPA uses an enrichment analysis approach called Functional Class Scoring or FCS. Enrichment analysis are a set of software tools and techniques which attempt to interpret the data from gene expression profiling experiments by finding functions and pathways that summarize the observations. Over the past decade as expression profiling grew in popularity, the need for accurate enrichment analysis also grew. Many algorithms and software tools were developed to address this. An in depth review can be found in the paper (Khatri, Sirota, and Butte 2012).

In REPA, we have mainly two modules. The first module links the transcription factors to individual gene sets. The second module performs enrichment analysis on the gene expression data. Combining the results from these two modules allows REPA to predict three things:

- Transcription factors that may regulate a given pathway. This is the result from the first module.

- Pathways that are affected in the given experiment. This is the result from the second module.

- Transcription factors that are most likely regulating the pathways that are most affected in the study. This is the result of combining the output of the two modules.

Over a hundred systems were developed in the past decade for performing enrichment analysis in gene expression data. Some of the most widely used tools are Gene Set Enrichment Analysis or GSEA (Subramanian et al. 2005), and Parametric Analysis of Gene set Enrichment or PAGE (Kim and Volsky 2005). In 2009 a new method called GAGE, or Generally Applicable Gene-set Enrichmen (Luo et al. 2009), was published which could handle datasets of different sample sizes or experimental designs. GAGE showed significantly improved results compared to GSEA and PAGE (Luo et al. 2009). For validation, we compared the second module of REPA to GAGE. The novel aspect of REPA is that we are using hypothesis based statistical testing to find regulators that control entire gene sets. Then we combine the results from the enrichment analysis module to present more detailed analysis of the data obtained from gene profiling experiments. Previous tools only perform enrichment analysis on the gene expression data and provide gene sets that are perturbed in the experiment whereas REPA also provides information about the likely regulators of the perturbed pathways.

This thesis is organized in 5 chapters. After this initial introductory chapter, we discuss the necessary biology that is required to understand this project in chapter 2. We also look at the work that has been done so far in this area, describe the problem statement, and existing solutions. In chapter 3 we take a detailed look at REPA and all its components. System validation and comparison is presented in chapter 4 followed by a conclusion in chapter 5. The work described in this thesis has been accepted for publication in the IEEE/ACM Transactions on Computational Biology and Bioinformatics journal and presented at the Great Lakes Bioinformatics Conference 2015.

Chapter 2

Background Knowledge and Related Work

This chapter describes the biological concepts required to understand the work done.

2.1 Flow of information in biological systems

The process of transmission of the genetic information from the genome of an organism to its phenotype (i.e. the expression of observable characteristics as an individual) is a complex process. A simplified description is provided here, as it is necessary to the understanding of this project.

Figure 2.1 shows how genetic information flows within a biological system. This was first proposed by Frank Crick in 1958 and published later in 1970 (Crick et al. 1970). Generally this information flows from DNA to DNA (replication), DNA to RNA (transcription) and RNA to proteins (translation).

To fully understand the diagram we need to learn more about the macromolecules such as DNA, RNA, and proteins, along with processes such as replication, translation,

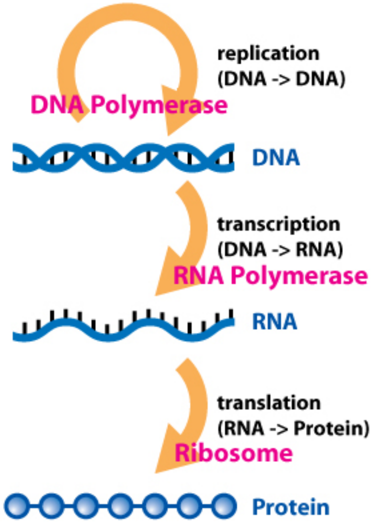


Figure 2.1: Flow of information in biological systems (Horspool 2008)

and transcription. This section describes each of them one by one.

2.1.1 Deoxyribonucleic acid (DNA)

The substance that is responsible for carrying the genetic information from parents to offspring in most living organisms, including all prokaryotic and eukaryotic cells and in many viruses, is an organic chemical of a complex molecular structure called Deoxyribonucleic acid, or DNA (Klug et al. 2012).

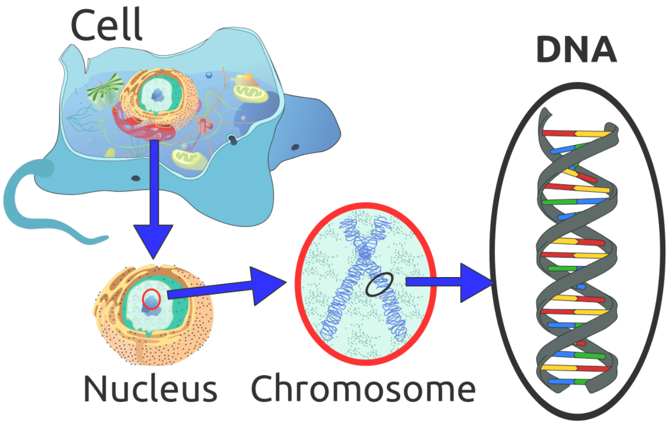


Figure 2.2: Location of DNA in a cell (Mariana Ruiz 2012)

As shown in figure 2.2, DNA resides in the nucleus of eukaryotic cells, where inside the chromosomes, the DNA is condensed in a DNA - protein complex called chromatin. When the chromatin is uncoiled, its main component is revealed: the DNA molecule.

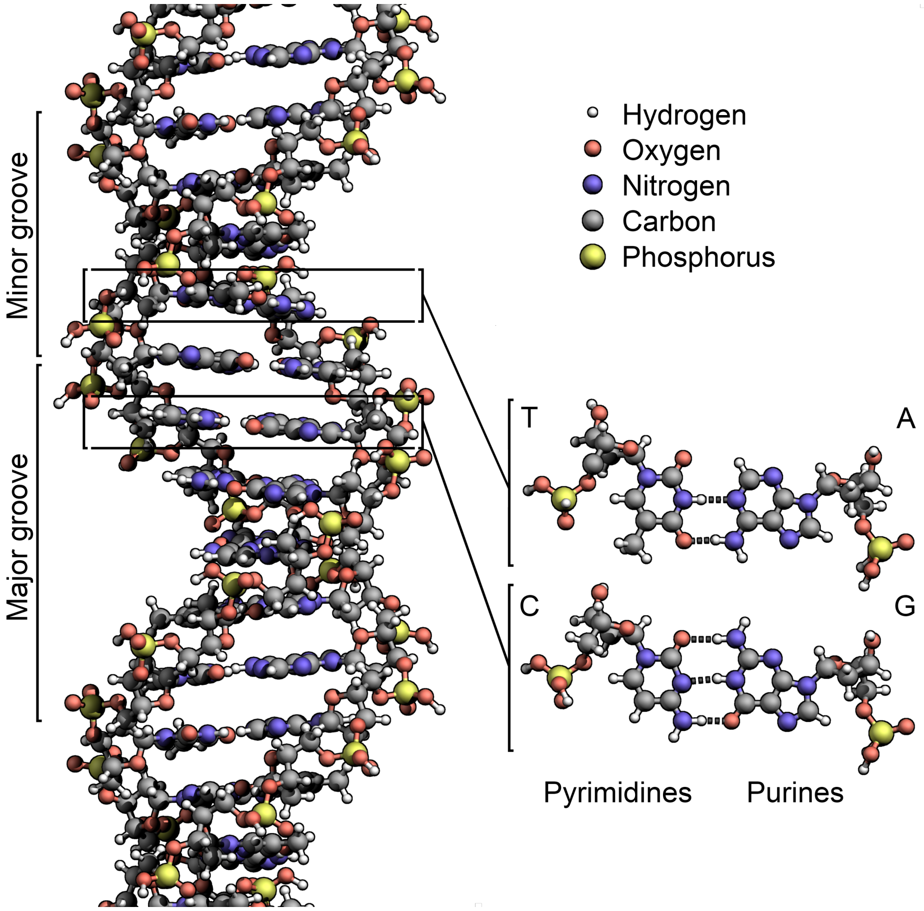


Figure 2.3: The structure of the DNA double helix (Zephyris 2011)

DNA has a double helix structure (Watson, Crick, et al. 1953), that looks like a long spiraling ladder (figure 2.3). It is formed of millions of elemental molecules, called nitrogenous bases, linked together in chains. The sequence in which the nitrogenous bases are linked amounts to a code that determines the characteristics of an individual, such as their eye color. These coded instructions are called genes. Genetic information is different for every individual making each of us unique.

The genetic material, or the DNA, of an organism contains instructions to control all everyday cellular activities (Hunter 2012). Bases, or nucleotides, are the building blocks of DNA, and there are four types: Adenine, Guanine, Cytosine and Thymine. The structure of these bases are given in the figure 2.3.

The configuration of the DNA molecule is highly stable, allowing it to act as a template for the replication of new DNA molecules, as well as for the production (transcription) of the related RNA (ribonucleic acid) molecule.

2.1.2 Ribonucleic acid (RNA)

RNA is also a nucleic acid like DNA but unlike DNA it is a single stranded molecule. Another difference between the two is that instead of thymine the fourth base pair in RNA is uracil. The structural differences between the two nucleic acids are shown in figure 2.4.

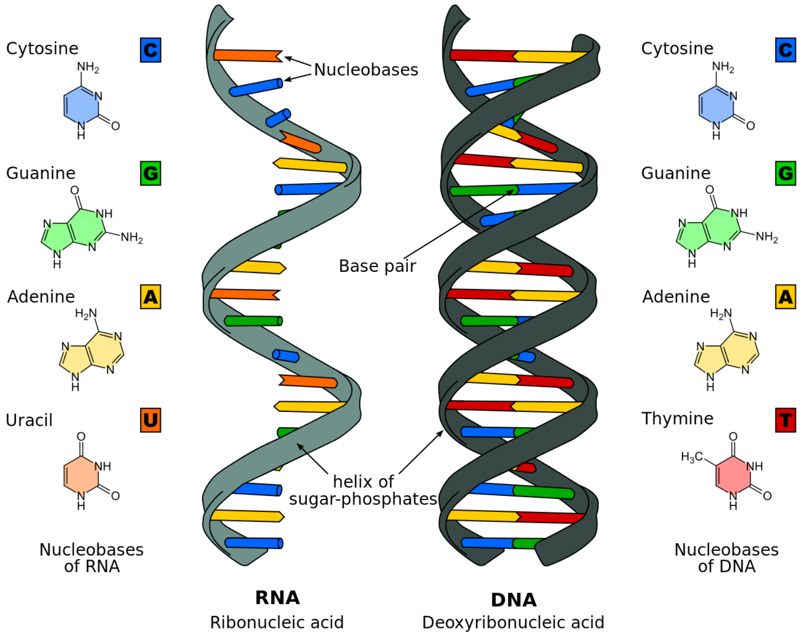


Figure 2.4: Comparison of a single-stranded RNA and a double-stranded DNA (Sponk 2011)

There are different types of RNA molecules, but in this thesis we are only interested in the RNA molecule whose main function is to carry the genetic information from the DNA to proteins via the steps of transcription and translation. This type of RNA molecule is called messenger RNA or mRNA (Hunter 2012).

mRNA carries the coding instructions for protein synthesis from DNA to the ribosome. During translation, the mRNA molecule specifies the sequence of the amino acids in a polypeptide chain and thereby provides a template for joining amino acids. (Pierce 2005). The folding of these amino acid chains gives birth to a protein molecule.

2.1.3 Proteins

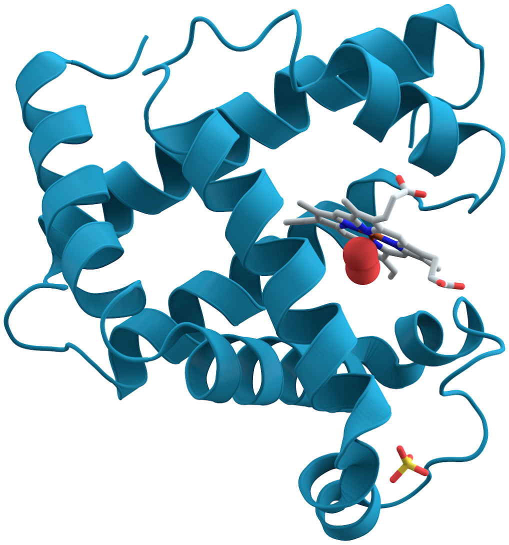


Figure 2.5: Myoglobin protein 3D structure (AzaToth 2008)

Proteins are large macromolecules that perform a wide array of functions. Each organism uses thousands of different proteins in their life span (Hunter 2012). Some functions that proteins perform are the catalyzation of metabolic reactions, the replication of DNA, responses to stimuli, and the transportation of molecules from one location to another, among many more.

For example, the protein myoglobin is an iron and oxygen binding protein. It is commonly found in the muscle tissues of vertebrates. It's 3D structure is shown in figure 2.5 (Kendrew et al. 1958).

2.1.4 Gene

A gene is a unit of heredity in a living organism. A gene is usually responsible for influencing certain characteristics of the organism. It is normally a stretch of DNA that codes for a type of protein, or for an RNA molecule that has a function. Genes only have an effect on the cell when they are expressed (transcribed).

2.1.5 Gene expression

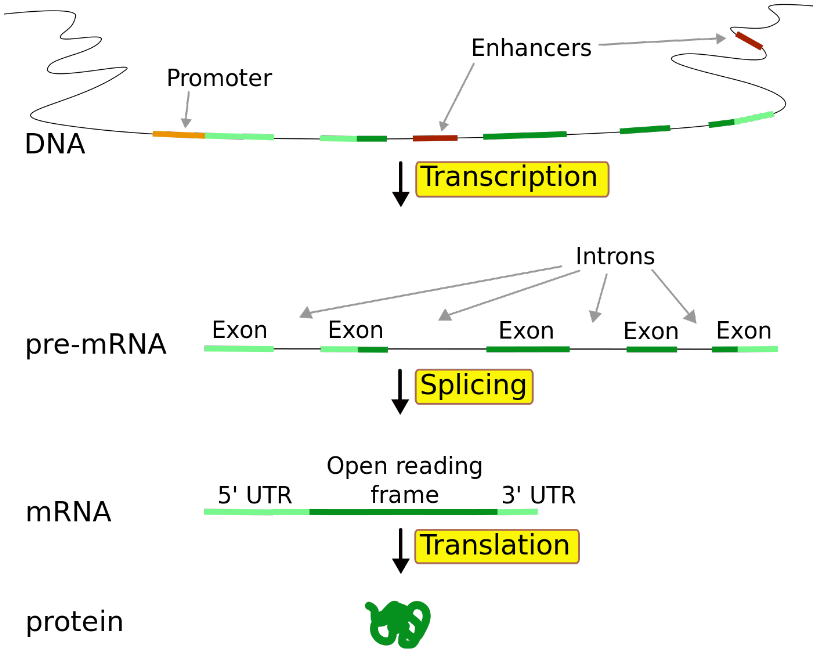


Figure 2.6: Steps in gene expression (Forluvoft 2007)

To be able to perform its functions, a gene needs to be expressed. The process of gene expression allows the information from a gene to be used in the synthesis of a functional gene product such as a protein molecule. Gene expression is of high importance because by controlling which genes are expressed and which are not expressed in a given scenario, a cell can decide its phenotype and which proteins to synthesis. The process of gene expression involves several steps as shown in figure 2.6.

A section of the DNA is first transcribed and then translated. This section is called the transcription unit. Just above the transcription unit there is a sequence of nucleotides which defines where the transcription unit begins. This is known as the promoter region.

2.1.6 Transcription

Transcription is the process by which the information contained in a section of DNA (a gene) is transferred to a newly assembled piece of RNA. It is facilitated by RNA polymerase and transcription factors (described in section 2.2). In eukaryotic cells protein encoding transcripts (pre-mRNA) must be processed further in order to ensure translation.

2.1.7 Translation

In translation, messenger RNA (mRNA) produced by transcription is decoded by ribosomes to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein molecule.

2.1.8 Putting it all together

So far in this section we have seen that, the genetic information that is passed on from parents to offspring in most living organism is stored in DNA. Genes are a unit of heredity, and the entire genome of an organism could contain thousands of genes. An individual gene is usually a small stretch of DNA, that when expressed, codes for a specific protein. The process of gene expression involves several steps, namely transcription, splicing, and translation. The final product of gene expression is commonly a functional protein molecule. This way, the information that was passed on from the organism's parents gets expressed and performs real functions in the organism.

2.1.9 Selective gene expression

Not every gene is expressed in all cells at all times. By controlling which genes are active, a cell can take on special characteristics and respond to its environment. Muscle cells and neurons have the same DNA, but perform different functions because they express different sets of genes. Transcription factors are one of the mechanisms to regulate which genes are expressed. In the nucleus, the DNA is condensed in the chromatin. In places close to where genes are being expressed, there are often zones of naked DNA. Transcription factors bind to these naked DNA sequences and regulate gene expression (Lyons 2012).

2.2 Transcription Factors

As mentioned above, transcription factors play a part in the regulation of gene expression. Transcription factors are protein molecules that bind to a specific DNA sequence. Once bound to the matching DNA sequence, the transcription factor molecule can promote or block the transcription of a nearby gene. The location where the transcription factor attaches itself to the DNA is called the transcription factor binding site. After binding itself the transcription factor regulates a gene that is spatially near the binding location, for example, by making it easier for the RNA polymerase to attach to the gene's promoter region. In most cases the gene lies downstream to the transcription factor binding site but in some cases, due to the complex nature of the three dimensional structure of the chromatin, a transcription factor can regulate a gene that is thousands of base pairs away but is close to the binding site in three dimensional space.

By promoting (as an activator), or blocking (as a repressor), the recruitment of RNA polymerase during transcription, transcription factors regulate the level of gene expression. RNA polymerase is the enzyme that performs the transcription of genetic information from DNA to RNA. Some transcription factors perform this function with other proteins in a protein complex while some do it alone.

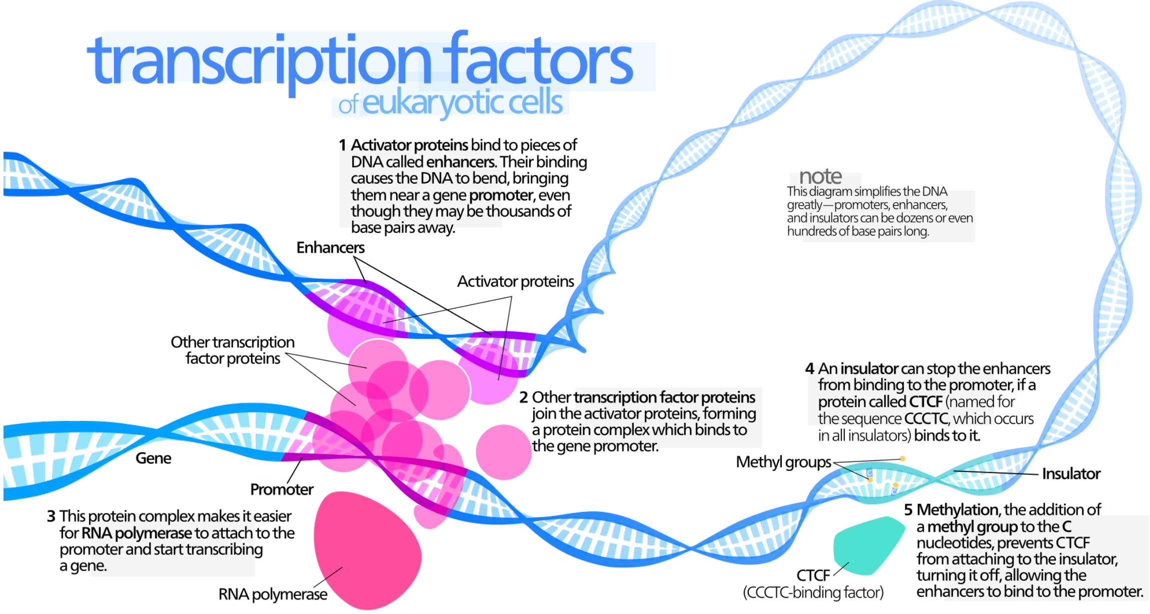

transcription factors
of eukaryotic cells
1 Activator proteins bind to pieces of DNA called enhancers. Their binding causes the DNA to bend, bringing them near a gene promoter, even though they may be thousands of base pairs away.
note
This diagram simplifies the DNA greatly—promoters, enhancers, and insulators can be dozens or even hundreds of base pairs long.
Enhancers
Activator proteins
Other transcription factor proteins
2 Other transcription factor proteins join the activator proteins, forming a protein complex which binds to the gene promoter.
4 An insulator can stop the enhancers from binding to the promoter, if a protein called CTCF (named for the sequence CCCTC, which occurs in all insulators) binds to it.
Gene
Promoter
Methyl groups
Insulator
3 This protein complex makes it easier for RNA polymerase to attach to the promoter and start transcribing a gene.
5 Methylation, the addition of a methyl group to the C nucleotides, prevents CTCF from attaching to the insulator, turning it off, allowing the enhancers to bind to the promoter.
RNA polymerase
CTCF
(CCCTC-binding factor)

Figure 2.7: Transcription factors working as activators (Kelvinsong 2012)

A demonstration of how these proteins affect the level of gene expression is given in figure 2.7. In this figure, several transcription factors are working together to create a protein complex that makes it easier for RNA polymerase to attach to the promoter region and start transcribing the gene. The gene that is being regulated here is located in a distant part of the DNA but due to the three dimensional folding of DNA in the chromatin, it is spatially close to the transcription factor binding site.

By regulating the gene expression, transcription factors enable different cells to perform different functions. For example, different genes are turned on in liver cells than those in skin cells and different genes are turned on in cancer cells than in healthy cells. Through the action of transcription factors, the various cells of the body, which all have the same genome, can function differently.

Roughly 8% of genes in the human genome encode transcription factors (Broad 2014). They play important roles in development, the sending of signals within the cell, and the events in a cell that lead to division and duplication, known as the cell cycle. Several human diseases are linked to mutations in transcription factors, such as hearing loss, congenital heart disease, and cancer (Villard 2004; Peters et al. 2002; Schott et al. 1998).

2.3 Biological Pathways

A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in that cell. A pathway can trigger the assembly of new molecules, such as a lipids or proteins. Pathways can also turn genes on and off, or spur a cell to move. Some of the most common biological pathways are involved in metabolism, the regulation of genes, and the transmission of signals.

2.4 ChIP-X Experiments

Several in vivo experimental technologies such as ChIP-chip (Iyer et al. 2001), ChIP-seq (Johnson et al. 2007), ChIP-PET (Wei et al. 2006) and DamID (Peric-Hupkes et al. 2010) provide details about possible binding sites for transcription factors at a genome-wide level. These four methods together are referred to as ChIP-X. The sites discovered using the ChIP-X methods are near genes and are found when the chromatin structure of a specific cellular state allows binding of a particular transcription factor. This means that unlike possible binding sites found using in vitro approaches, the possibility of these sites to have actual biological significance is much higher. Results from such experiments report the binding of specific transcription factors to DNA in proximity of a target gene's location. Such experiments commonly list hundreds to thousands of potential regulatory interactions (Lachmann et al. 2010).

2.5 Gene Expression Profiling

Gene expression profiling is the measurement of the abundance level (the expression) of thousands of transcripts at once, to create a global picture of cellular state. These profiles can, for example, distinguish between cells that are actively dividing, or show how the cells react to a particular treatment.

A DNA microarray (also commonly known as a DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously, or to genotype multiple regions of a genome. Each DNA spot contains segments of a specific DNA sequence, known as probes (or reporters or oligos). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA (also called anti-sense RNA) sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence- labelled targets to determine relative abundance of the targets in the sample. RNA-seq refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content. The technique has been rapidly adopted in studies of diseases like cancer.

2.6 Pathway Analysis of Gene Expression Data

Gene expression profiling experiments allow biologists to measure the activity levels of genes. The data that is generated from such experiments usually is several megabytes long. For example in (Emery et al. 2009) and (Lavery et al. 2008), the research teams have performed typical gene expression profiling experiments and the data after refinement is 88 MB and 21.7 MB in size respectively. To derive useful information from this large quantity of data is a challenge. In the last decade or so, as experiments of such nature have gained popularity, several approaches have been devised to help researchers understand the meaning of this data and to determine the biological activities taking place in the cells under observation.

This thesis is in the research area of facilitating the researchers who are performing such experiments to understand the biological processes that are active in their studied cellular state. It is done by performing statistical tests on the data obtained from high throughput gene expression profiling experiments. Several individual research groups have made important contributions to this field and this project built upon the work that has been done so far. In the following sections, we discuss all those techniques and databases that are related to this project.

2.7 Gene sets formation

Individual genes are annotated based on their functions, position and other characteristics. For example, functional annotation for a gene can be its association with a particular function in a metabolic pathway. Any information about the function of this gene is a functional annotation of the gene. These annotations are stored in publicly available databases. Using these publicly available gene annotations it is possible to create gene sets by taking all the genes that have a common annotation and clubbing them together. Usually every biological pathway, such as metabolic or signaling pathways, are associated with certain genes. Thus, by clubbing together genes based on their functions, we can connect biological processes or pathways to sets of genes. An example of such a gene set could be the KEGG pathway Glycerolipid metabolism (hsa00561) (Kanehisa and Goto 2000). Based on published studies, there are 49 genes associated with the pathway (Norbeck et al. 1996; Karlsson et al. 1997; Berg et al. 2001). These genes form the glycerolipid metabolism gene set.

2.8 Enrichment Analysis (Pathway Analysis)

High-throughput gene expression profiling techniques, such as DNA microarray and RNA-Seq, allow researchers to simultaneously measure genome-wide levels of gene expression under specific biological conditions. Statistical approches such as limma (Smyth 2005) and edgeR (Robinson, McCarthy, and Smyth 2010) are then used to identify differences in gene expression between two or more conditions. Enrichment analysis or pathway analysis is an analytical approach to interpret the results of a gene expression profiling experiments with respect to gene sets. It is the process with which we associate observed changes in gene expression with cellular functions and/or metabolic pathways. Without such an analytical process, it will be very difficult to comprehend which biological pathways are most active in the particular case under study.

Gene expression profiling experiments usually generate a list of differentially expressed genes. Here is an example of how the data that is obtained from such an experiment looks (Lavery et al. 2008).

EntrezID	8hCont1	8hTrt_8hCult1	8hCont2	8hTrt_8hCult2
10000	6.666482	6.727039	7.859644	7.888743
10001	9.874859	9.873068	9.838792	9.757909
10002	5.524512	5.651697	5.299609	5.146715
10003	4.604491	4.661876	4.790255	4.705559
10004	7.904135	7.883218	8.00505	7.962769
...				
...				

Table 2.1: Expression profiling data sample

Such a list can be very useful in identifying genes that may have roles in a particular phenomena or phenotype. However, for many researchers this list would not be sufficient in providing insight into the underlying biology of the condition being studied. To individually study each gene and interpret the meaning would be a very complex and time consuming process. Therefore, categorizing the genes based on their common functional annotation helps in two ways.

Reduced complexity By grouping the genes into sets of genes, with each gene set targeting a specific pathway or function, the complexity is reduced to just a few hundred pathways for the experiment.

Higher explanatory power Pathways that have different activity levels between two conditions would generally have a higher explanatory power than just a list of genes would (Glazko and Emmert-Streib 2009).

Hence the lower complexity and higher explanatory power of enrichment analysis has made it a de-facto in post gene expression analysis. Broadly, there are three different approaches for performing enrichment analysis (Khatri, Sirota, and Butte 2012). They are :

- Over-Representation Analysis (ORA) approach

- Functional Class Scoring (FCS) approach.
- Pathway Topology (PT) approach.

We look at them one by one in the following sections.

2.8.1 Over-Representation Analysis (ORA) Approach

The growth in the popularity of High-throughput sequencing, and also the development of public gene set repositories such as Gene Ontology (GO) or Kyoto Encyclopedia of Genes and Genomes (KEGG), fueled the immediate need for functional analysis of microarray gene expression data. To tackle these problems the Over-Representation Analysis (ORA) approach was devised.

From the expression levels observed in the sequencing experiment, a list of significant genes that were over-expressed or under-expressed is created. To create this list, an arbitrary cut-off p-value is set. For example, a researcher may say that all genes that have a p-value less than, or equal to, 0.05 qualify as significant genes. Next, for each pathway (or gene set) the numbers of genes that are present in this list are counted. Then, by using statistical analysis techniques, such as tests based on the hyper geometric, chi-square, or binomial distribution, it is determined whether more genes belonging to the gene set are present in the list than expected by chance.

This is a very simple technique, but it sheds some light on the gene sets that are under or over expressed. ORA has a few shortcomings too, as discussed next.

Limitations of Over-Representation Analysis (ORA) Approach

Even though the ORA approach is the most popular approach, it has several shortcomings.

Firstly, the statistical tests (e.g., hyper geometric distribution, binomial distribution, chi-square distribution, etc.) ignore the measurements found for the genes in the gene expression experiments. As this data is ignored, all the genes that make the list are treated equally despite their varying levels of expression. The list of significant genes is generated based on an arbitrary threshold and the individual genes that do not make the threshold are discarded. Genes whose expression levels fall in the border of the threshold also have some significance but are totally ignored. This is a disadvantage of having a hard cutoff threshold.

Secondly, one of the goals of gene expression analysis is to understand how interactions between various gene products occur as the levels of gene expression changes. By considering that all the genes are independent, ORA significantly reduces its ability to analyze complex biological interactions that include several gene products. Because the ORA techniques consider all genes as equal and independent, it fails to provide any insight in this regard.

Finally, this approach works with the assumption that all the pathways are independent to each other, which is not true. For example, in signaling pathways in KEGG, there is a presence of growth factors that activate the MAPK signaling pathway. This signaling pathway in turn activates the cell cycle pathway. ORA methods do not account for such inter-pathway interactions and dependences.

2.8.2 Functional Class Scoring (FCS) Approach

In most biological systems, significant effects on pathways can be caused by large changes in individual genes, but they can also be caused by weaker coordinated changes in the expression levels of several functionally related genes. By clubbing such related genes into a gene set such that a gene set represents a biological pathway, we can detect such effects. Almost all the FCS based methods have mainly three steps:

Step 1: Calculate Gene Level Statistic

First a gene level statistic is computed from the molecular measurement data obtained from high-throughput expression analysis experiments such as DNA microarray or RNA-Seq. This is done by calculating differential expression for each of the genes. Several statistical methods, such as correlation of molecular measurements with phenotype (Pavlidis et al. 2004), Q-statistic (Goeman et al. 2004), signal-to-noise ratio (Subramanian et al. 2005), t-statistic (Tian et al. 2005) or Z-score (Kim and Volsky 2005) can be used to represent the expression levels.

Step 2: Calculate pathway level statistics

Next, gene level statistics for all genes in a given gene set are aggregated into a single pathway level statistic. There are several statistical methods to do this but some of the more common ones are Kolmogorov - Smirnov (Smirnov 1944), the Wilcoxon rank sum test (Mann, Whitney, et al. 1947), or to take the sum, mean or median of the gene level statistics. Whatever method is chosen to implement this, it's power can depend on factors such as the proportion of the genes present in the pathway that were differentially expressed, the actual size of the pathway (i.e. the number of genes present in the pathway) and the amount of correlation that exists between the various genes in the pathway. Even though multivariate statistics should show better results as they also account for inter-dependencies among genes, it has been observed that for higher cut-offs ($pValue \leq 0.001$), the uni-variate statistics show more power, and for less stringent cut-offs ($pValue \leq 0.05$) the uni-variate statistics show equal power (Khatri, Sirota, and Butte 2012).

Step 3: Assessing statistical significance of the pathway level statistic

In this step the statistical significance is computed by using a null hypothesis. There are mainly two ways to do the testing: competitive null hypothesis testing and self-contained null hypothesis testing. In the former method, class labels (i.e. phenotypes) for each sample are permuted, and comparison is made between the set of genes in a given pathway with itself. In the latter method, gene labels are permuted for each pathway, and comparisons are made between the set of genes that are in the pathway, with the set of genes that are not in the pathway. The size of the gene sets remains the same.

Advantage of using FCS

Some of the limitations described above related to using the ORA approach have been addressed in the FCS approach. This helps FCS provide better results and deeper insight into the underlying biology of any given condition than those provided by ORA. For example, FCS does not require any arbitrary cut-off threshold for dividing the genes into significant and non-significant groups. It uses all the available molecular measurements for its analysis. FCS uses the molecular measurement information to detect coordinated changes in expression of genes in some pathways. By detecting such coordinated changes, FCS can give us information about dependence between genes.

Limitations to FCS

FCS analyzes each pathway independently, hence a problem arises when a single gene is part of multiple pathways. In such a case, a given gene might be over-expressed because it is playing an important role in a particular pathway, but this expression level will be considered while evaluating the pathway level statistic of other pathways that the gene is a part of. Another limitation arises when the statistical method used to implement FCS is a rank based method. In such a case, the value obtained in the experiment is not considered in the analysis, but only the rank assigned is considered.

2.8.3 Pathway Topology (PT)-Based Approach

Pathway topology is the newest technique available for performing enrichment analysis. It is similar to FCS, except for how pathway topology based-approaches compute the gene level statistics. Several publicly available pathway knowledge bases hold information about gene products that interact with each other, how those products interact and where they interact in a given pathway. ORA and FCS do not utilize this knowledge. An example of a PT-based approach is ScorePAGE proposed by (Rahnenfuhrer et al. 2004). ScorePAGE computes similarity between each pair of genes in a pathway. The similarity is measured by calculating the correlation or covariance between the two genes. The similarity score is comparable to the gene level score in FCS based approaches. Then, these scores are averaged to arrive at the pathway level statistics. However, ScorePAGE divides the similarity score with the number of reactions needed to connect the two genes in the pathway. This strategy assigns varying weights to the pairwise similarity scores.

Limitations

Some of the common limitations with this strategy are:

- Pathway topology depends upon the cell type and the condition being studied. Hence this information is not readily available and is usually fragmented in various knowledge bases. As the annotation becomes more comprehensive and complete, these approaches are expected to perform better.

- No existing PT-based approach can collectively model and analyze high-throughput data as a single dynamic system.

Chapter 3

Methodology and Implementation

3.1 Motivation

In the modern day world a plethora of data is produced everyday in laboratories performing high throughput experiments such as DNA Sequencing and Gene Expression Profiling. Trying to analyze and understand the mechanisms of the biological processes under study is a non-trivial task. To extract knowledge from this data there is a need for new computer programs to perform analytics and help interpret this growing amount of data.

Some of the common questions that arise after any RNA sequencing or DNA microarray experiment are: what biological pathways were affected in the sample, and which transcription factors were regulating these biological pathways. Providing answers to these questions can help researchers to make new discoveries or provide direction to their future research.

Currently there are programs to identify the biological pathways that are getting affected. As discussed in Khatri et al (Khatri, Sirota, and Butte 2012) programs such as Gage (Luo et al. 2009) can identify the gene sets that are over or under expressed in the sample by statistical testing. There are also programs such as ChEA (Kou et al. 2013) as described in the previous chapter that provide information about which transcription factors have their targets over-represented in a list of genes. Presently though, there is no such program that provides the full picture by predicting which transcription factor is actually regulating the pathways that are most affected in the study sample.

In this project, we combine the results of gene set enrichment analysis (FCS) and Chip-X data driven analysis to make predictions that tell the researcher which transcription factors might regulate pathways affected in the sample being studied.

3.2 Overview of the system

The software system that we built to materialize our idea had to perform several tasks. First, it should quantify to what degree a given transcription factor regulates any particular gene set. To achieve this, we performed a functional class scoring on the data obtained from the ENCODE project. Second, the software should identify which gene set was most affected in a given experiment. After conducting an experiment, researchers can use this software to perform functional class scoring on the data obtained from their experiment to reveal the biological pathways that were most affected. Finally, the system should combine the results of the above two parts to arrive at the final results. The final results will be a list of transcription factors and gene sets that are playing an important role in the phenomena under study.

We named our system REPA for Regulation Expression Pathway Analysis.

3.3 REPA description

We will now look at the inputs to the system. Then, we will see each module in a greater detail. The modules of the system are as follows:

- Linking transcription factor binding sites with promoter regions of known human genes
- Creating the Regulation Database
- Expression Analysis
- Combining Regulation and Expression Results

3.3.1 Inputs to the system

There are mainly three inputs to the system.

Gene sets from gene annotation databases

As described in the previous chapter, a gene set is a group of genes that has some common functional annotation. A gene set may represent a biological pathway and include all the genes that play a role in that pathway.

For this project we gather our gene sets from two sources:

- **Molecular Signatures Database or MSigDB**

  MSigDB (Subramanian et al. 2005; Liberzon et al. 2011) is a collection of annotated gene sets that were made publically available by various research groups such as Reactome (Vastrik et al. 2007), Biocarta (Nishimura 2001), Gene Ontology (Ashburner et al. 2000), Gene Arrays, KEGG (Kanehisa and Goto 2000; Kanehisa et al. 2014) and more.

MSigDB contains gene sets that are kept classified in six collections.

H Hallmark gene sets

Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression (Total 50 gene sets).

C1 Positional gene sets

These are gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene. These gene sets are helpful in identifying effects related to chromosomal deletions or amplifications, dosage compensation, epigenetic silencing, and other regional effects (Total 326 gene sets).

C2 Curated gene sets

These are gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. The gene sets in this group can be further classified into the following groups (Total 4725 gene sets).

CGP: chemical and genetic perturbations These gene sets represent expression signatures of chemical and genetic perturbations. For each perturbation there is usually two sets: one set consisting of genes that show increase in expression levels (XXX_UP) and another set consisting of genes that show lower expression levels denoted by (XXX_DOWN) (Total 3395 gene sets).

CP: Canonical pathways These gene sets are from pathway databases.

These are usually compiled by domain experts and are canonical representation of any given biological process (Total 1330 gene sets).

CP:BIOCARTA: BioCarta gene sets These are genes derived from BioCarta pathway database (Nishimura 2001) (Total 217 gene sets).

CP:KEGG: KEGG gene sets Genes derived from KEGG pathway database (Kanehisa and Goto 2000; Kanehisa et al. 2014) (Total 186 gene sets).

CP:REACTOME: Reactome gene sets Genes derived from Reactome pathway database (Vastrik et al. 2007) (Total 674 gene sets).

C3 Motif gene sets

Gene sets that contain genes that share a cis-regulatory motif that is conserved across the human, mouse, rat, and dog genomes (Xie et al. 2005) (Total 836 gene sets).

MIR: microRNA targets Gene sets that contain genes that share a 3'-UTR microRNA binding motif (Total 221 gene sets).

TFT: transcription factor targets Gene sets that contain genes that share a transcription factor binding site defined in the TRANSFAC database (version 7.4) (Wingender 2008). Each of these gene sets is annotated by a TRANSFAC record (Total 615 gene sets).

C4 Computational gene sets

Computational gene sets defined by mining large collections of cancer-related microarray data (Total 858 gene sets).

C5 GO gene sets

These are the gene sets that are named after GO terms and contain genes annotated by that term (Total 1454 gene sets). The gene sets in this group can be further classified into the following groups.

BP: GO biological process Gene sets derived from the Biological Process Ontology (Total 825 gene sets).

CC: GO cellular component Gene sets derived from the Cellular Component Ontology (Total 233 gene sets).

MF: GO molecular function Gene sets derived from the Molecular Function Ontology (Total 396 gene sets).

C6 Oncogenic signatures

These gene sets represent signatures of cellular pathways which are often dis-regulated in cancer. The majority of signatures were generated directly from microarray data from NCBI GEO (Edgar, Domrachev, and Lash 2002) or from internal unpublished profiling experiments which involved perturbation of known cancer genes. In addition, a small number of oncogenic signatures were curated from scientific publications (Total 1454 gene sets).

C7 Immunologic signatures

Gene sets that represent cell states and perturbations within the immune system (Total 1454 gene sets).

- **KEGG: Kyoto Encyclopedia of Genes and Genomes**

Kyoto Encyclopedia of Genes and Genomes (KEGG, `http://www.genome.jp/kegg/` or `http://www.kegg.jp/`) is a database that provides manually curated gene sets (Kanehisa and Goto 2000; Kanehisa et al. 2014). KEGG collects information on functional annotations of various DNA elements and integrates this information. Genes from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism and the ecosystem. Then this information is used to create a knowledge base for organizing experimental knowledge in computable forms; namely, in the forms of KEGG pathway maps.

Any KEGG pathway includes a list of all the genes which play a functional role in the pathway and also other details such as the pathway map and any diseases linked with this pathway. The genes in a given pathway form a gene set. In REPA we used the gene sets from KEGG along with the gene sets from MSigDB. We included directly KEGG pathways as gene sets, as we noted that MSigDB does not include the most recent version of the KEGG.

We represent gene sets in the format that is used by MSigDB. Namely first column gives the name of the gene set. The second column is not used by the program but specifies a URL that gives more information about the specific gene set. After that we have a list of entrez ids that specify the genes that are present in the gene set. The file is a tab delimited text file without headers. For example, table 3.1 shows the format of a sample gene set.

Gene Set Name	URL	Entrez IDs of members
KEGG_STEROID_BIOSYNTHESIS	http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_STEROID_BIOSYNTHESIS.html	6646,4047,6713,10682,1595,1717,1594,1718,51478,6307,2222,6309,3988,1056,7108,50814,8435
⋮	⋮	⋮

Table 3.1: Sample Gene Set

Transcription Factor Binding Data (TFBD)

As described in the previous chapter in section 3.4.1 one of the goals of the public research project Encyclopedia of DNA Elements or ENCODE (Consortium et al. 2012) was to identify the transcription factor binding sites. The data produced under this project from numerous Chip-Seq (Johnson et al. 2007) and Chip-Chip (Iyer et al. 2001) experiments provide information about the binding sites of 160 transcription factors in 44 cell lines. This data was then processed and used by ChEA2 (Kou et al. 2013). ChEA2 also made the processed data freely available for download on their website (Kou 2014).

The ENCODE data compiled by ChEA2 includes 920 experiments done in 44 cell-lines profiling 160 transcription factors for a total of approximately 1.4 million transcription-factor / target-gene interactions (Kou et al. 2013). This data is given in the bed format which includes the transcription factor, cell line, start and end of peaks, score and signal. The score and signal values are directly proportional to the strength of the binding between DNA and the given transcription factor at any given site. The score value is derived from the signal and lies between 0 and 1000 and it is proportional to the maximum signal strength. We use the signal strength as input to our algorithm.

Genomic positions of human genes

To link the transcription factor binding sites to the genes that the transcription factors likely regulate, we needed the locations of both the transcription factor binding sites and the genes. As mentioned above, we got the ENCODE transcription factor binding sites data from ChEA2. For a list of all human genes and their position we used Ensembl's Biomart (Kinsella et al. 2011) website `http://www.ensembl.org/biomart/martview/`. Using this website we gathered the genomic positions of all human genes. The version that was used was `Ensembl Release version 71` and the data set was `Homo sapiens genes (GRCh37.p10) (may 21, 2013)`. We included all genes coding for lincRNA, miRNA and proteins.

A sample of the data downloaded from Ensembl's biomart website is given in table 3.2.

Ensembl Gene ID	Associated Gene Name	Chromosome Name	Strand	Gene Start (bp)
ENSG00000248425	AC006296.2	4	1	14392063
ENSG00000236437	AP001891.1	11	-1	116367616
ENSG00000240567	RP11-3P17.4	3	1	161144215
ENSG00000235357	RP1-159G19.1	6	-1	80513300
⋮	⋮	⋮	⋮	⋮

Table 3.2: Ensembl Biomart Human Gene List

Gene expression profiling experiment data

After processing the results obtained from a high throughput gene expression profiling experiments such as DNA micro-array or RNA-seq a list with all the genes in the experiment with their corresponding level of activity is generated. To represent the gene level activity various types of statistics can be used. Some of the popular statistics user are log fold change, Z-statistic, t-statistic, etc. Table 3.3 gives an example of how the data looks.

Entrez ID	logFC	AveExpr	t-statistic	P.Value	adj.P.Val
5228	4.066771195	8.984909943	30.01850012	1.75E-06	0.030867069
8200	-2.937752131	9.096127188	-22.30635557	6.85E-06	0.060528957
3400	3.107500295	8.990624466	18.99550122	1.43E-05	0.066519298
133	-3.11990311	9.026165624	-17.69540658	1.98E-05	0.066519298

Table 3.3: Sample gene expression profiling experiment data

3.3.2 Step 1: Linking transcription factor binding sites with promoter regions of human genes

As discussed in (Kristiansson et al. 2009) “A gene’s promoter region is traditionally (if loosely) defined with respect to its transcription start site (TSS): 1000-3000 base pairs upstream, and 100-300 basepairs downstream.” In our project we focus on cis-regulated regions which are within 3000 base pairs upstream of the genes as the majority of the binding sites are located there. We ignored the enhancer regions which are binding sites but typically at a much larger distance from the TSS.

We found all such instances where in the ENCODE data, a transcription factor binding site existed within 3000 base pairs upstream of a gene's TSS.

Inputs to this part were the bed file with the location of the DNA - transcription factor binding peaks and the bed file with the genomic coordinates from the gene's transcription start site to 3000 base pairs upstream for all human genes as described in section 3.3.1. We used Bedtools (Quinlan and Hall 2010) to accomplish this. Intersect command in Bedtools was used to find those promoter regions where a binding event occured. The command used was as follows:

```
intersectBed −a encode.bed −b geneList.bed −wo
```

The $-wo$ flag writes side by side the original A and B entries plus the number of base pairs of overlap between the two features.

3.3.3 Step 2: Creating the Gene Set - Transcription Factor Binding Database

The goal of this step is to associate a transcription factor with given gene sets based on the binding strength of that transcription factor with the promoter region of the genes in each gene set. In other words we want to find out whether a transcription factor may regulate a given gene set. To achieve this we ran a functional class scoring with the transcription factor binding signals associated to human genes and the gene sets obtained from MSigDB and KEGG.

As described in the previous chapter, a functional class scoring approach has three steps. First is to calculate the gene level statistic. Next step is to find a gene set level statistic using the gene level statistic and finally using null hypothesis testing we find out the statistical significance of the gene sets.

A step by step description of the process follows.

Inputs

1. Gene sets in the format described in the section 3.3.1.
2. Transcription factor binding data from the output obtained in step 1 described in section 3.3.2.

Step 2.1: Parsing and loading input files The two input files are parsed and loaded in memory. A hash table is used to store the data to quickly access the values based on transcription factor name and the gene's Entrez Id.

Steps 2.2 through 2.6 are performed for each transcription factor - gene set pair.

Step 2.2: Map TFBSs to strongest signal Get the signal values from the transcription factor binding sites database for each of the gene in the gene set. If multiple values exist from different cell lines, the maximum value is considered.

If there are values for at least 5 genes in the gene set we proceed to the next step else we skip to the next gene set - transcription factor pair.

Steps 2.3 and 2.4 are executed 1000 times.

Step 2.3: Generate vector with random signal values To run the permutation test, we first generate a gene set of randomly selected genes of the same size as that of the real gene set. We collect the signal values for this randomly chosen gene set from the transcription factor binding database as done in step 2.2 for the real gene sets.

Step 2.4: Statistical hypothesis testing We use the following three statistical tests in our program with the real gene set values and the random gene set values as input.

Two-sample pooled test (*ALGLIB project*; Broad 2014)

This test checks hypotheses about the fact that the means of two random variables X and Y which are represented by samples x_S and y_S are equal. In our case x_S represents the signals from real gene sets we found in step 2.2 and y_S represents the randomly selected signal in the step 2.3. The test works correctly under the following conditions:

- Both random variables have a normal distribution
- Dispersions are equal (or slightly different)
- Samples are independent.

During its work, the test calculates t-statistic:

$$t = \frac{\overline{x_S} - \overline{y_S}}{\sqrt{\frac{\sum(x_i - \overline{x_S})^2 + \sum(y_i - \overline{y_S})^2}{N_x + N_y - 2}\left(\frac{1}{N_x} + \frac{1}{N_y}\right)}}$$

N_X and N_Y are the sizes of X and Y respectively.

Note 1 If X and Y have a normal distribution, the t-statistic will have Student's distribution with $N_X + N_Y - 2$ degrees of freedom. This allows the use of the Student's distribution to define a significance level corresponding to the value of the t-statistic.

Note 2 If X or Y are not normal, t will have an unknown distribution and, strictly speaking, the t-test is inapplicable. However, according to the central limit theorem, as the sample sizes increase, the distribution of t tends to be normal. Therefore, if sample sizes are big enough, we can use the t-test even if X or Y is not normal. But there is no way to find what values for N_X and N_Y are big enough. These values depend on how X and Y deviate from the normal distribution.

After running this test we store the right-tailed p-value. The null hypothesis for the right-tailed test is that the mean of y_S is less than or equal to the mean of x_S.

Two-sample unpooled test (*ALGLIB project*) Similar to the two-sample pooled test, this test checks hypotheses about the fact that the means of two random variables X and Y which are represented by samples x_S and y_S are equal. The test works correctly under the following conditions:

- Both random variables have a normal distribution
- Samples are independent.

Unlike the previous test, for this test, dispersion equality is not required.

During its work, the test calculates the t-statistic:

$$t = \frac{\overline{x_S} - \overline{y_S}}{\sqrt{\frac{Var(x_S)}{N_X} + \frac{Var(y_S)}{N_Y}}}$$

If X and Y have a normal distribution, the t-statistic will have Student's distribution with DF degrees of freedom:

$$DF = \frac{(N_X - 1)(N_Y - 1)}{(N_Y - 1)c^2 + (N_X - 1)(1 - c^2)}$$

$$c = \frac{\frac{Var(X_S)}{N_X}}{\frac{Var(X_S)}{N_X} + \frac{Var(X_S)}{N_Y}}$$

This allows the use of the Student's distribution to define the significance level corresponding to the value of the t-statistic. Note 2 from the two sampled pooled test section is also applicable to this test.

Mann-Whitney U test The Mann-Whitney U-test (Mann, Whitney, et al. 1947; McKnight and Najab 2010; Bochkanov and Bystritsky 1947) is a non-parametric method which is an alternative to the two-sample Student's t-test. This test is used to compare medians of non-normal distributions X and Y. The test works correctly under the following conditions:

- X and Y are continuous distributions (or discrete distributions well-approximating continuous distributions)
- X and Y have the same shape. The only possible difference is their position (i.e. the value of the median)
- The number of elements in each sample is not less than 5
- The samples are independent
- Scale of measurement should be ordinal, interval or ratio (i.e. test could not be applied to nominal variables)

Here is a simple step by step description of how the Mann-Whitney U test works:

- Both samples (having sizes N and M) are combined into one array which is sorted in ascending order keeping track of which sample the element had come from.
- After sorting, each element is replaced by its rank (its index in array, from 1 to $N + M$).
- Then the ranks of the first sample elements are summarized and the U-value is calculated using the following formula:

$$U = N \times M + \frac{N(N+1)}{2} - \sum_{x_i} Rank(x_i)$$

The mean of U equals $0.5 \times N \times M$. If U is close to this value, the medians of X and Y are close to each other. If we know distribution quantiles, we can get the significance level corresponding to the value of U.

- For a big enough N and M, U could be approximated by the normal distribution with a mean of $0.5 \times N \times M$ and a standard deviation of

$$\sigma = \sqrt{\frac{N \times M(N+M+1)}{12}}$$

The p-value is calculated from the mean and standard deviation.

All the three tests mentioned above namely two-sample pooled test, two-sample unpooled test and Mann-Whitney U-test returns three p-values:

p-value for two-tailed test The null hypothesis here is that the medians for the two samples are equal.

p-value for left-tailed test The null hypothesis here is that the median of y_S is greater than or equal to the median of x_S.

p-value for right-tailed test The null hypothesis is that the median of y_S is less than or equal to the median of x_S.

Step 2.5 For three pre-determined thresholds of significance namely 0.05, 0.01 and 0.001 we count number of times the mean / median of the values of the actual gene set is to the right of the mean / median of the values of the random gene set with a p-value of:

- less than 0.001
- between 0.01 and 0.001

- between 0.05 and 0.01
- greater than 0.05

Transcription Factor	E2F6
Gene set name	DOANE_BREAST_CANCER_CLASSES_UP
Total genes in the gene set	72
Genes for which values were found	14
Pooled tTest [< 0.001]	11
UnPooled tTest [< 0.001]	7
MWU test [< 0.001]	5
Pooled tTest [< 0.01]	74
UnPooled tTest [< 0.01]	75
MWU test [< 0.01]	70
Pooled tTest [< 0.05]	238
UnPooled tTest [< 0.05]	237
MWU test [< 0.05]	249
Pooled tTest [Remaining]	677
UnPooled tTest [Remaining]	681
MWU test [Remaining]	676
Pooled tTest Average p-value	0.1937
UnPooled tTest Average p-value	0.194217
MWU test Average p-value	0.18896
Genes in gene set for which values were found	347, 253190, 51181 ....

Table 3.4: Sample of the gene set - transcription factor binding database

Results After running the program for all the transcription factors and gene sets available, the generated results are as shown in table 3.4. Due to space constrains, the columns are presented as rows and rows as columns.

3.3.4 Step 3: Expression Analysis

The purpose of this step is to identify which gene set is over-expressed or under-expressed in the given biological sample. To achieve this goal, we used a functional class scoring approach.

The inputs to the system are as follows:

1. Gene sets in the format described in the section 3.3.1

2. High throughput Gene expression profiling experiment data in the format described in section 3.3.1.

Step 3.1: Parse and load input data

First the two data files i.e. the expression levels and gene set database are read and loaded in the memory.

The following steps are executed for each available gene set in the database.

Step 3.2: Get expression levels for member genes of the gene set

For each of the genes present in the selected gene set, their corresponding values are retrieved and stored in a vector.

If the number of values found is greater than 5, proceed with the following steps.

Step 3.3: Create a random value vector of equal size

Here we select values from the expression levels files and insert them into a new vector. The number of values randomly selected and inserted are equal to the number of values found in the previous step.

Step 3.4: Running the statistical hypothesis tests

To understand the statistical significance of the values found in step 2, we compare them with the randomly generated values from step 3. We use the following three statistical hypothesis tests for this:

1. Two-sample pooled test

2. Two-sample unpooled test

3. Mann-Whitney U test

This step is similar to the step 4 in section 3.3.3 and the detailed description of the tests are also given in the same section.

Results After running the tests for all the available gene sets, we arrive at p-values for each gene set for the particular case. We can conclude that the gene sets with the lowest p-values are either over-expressed or under-expressed. In either case their behavior is different from what is expected hence it can be concluded that there is a strong likelihood that the gene set is playing an important role in the cellular condition under study.

3.3.5 Step 4: Combining two p-values

From the two previous steps we have the following information:

From Step 2 A p-value quantifying the level of binding a transcription factor has on a gene set.

From Step 3 A p-value representing the level of perturbation shown by any given gene set in the experimental condition under study.

The next step is to combine the two p-values to obtain another p-value that represents the likelihood of a transcription factor regulating the gene set in the biological condition under study.

To achieve this we follow the method suggested in the article titled 'Combining p-values via averaging' (Vovk 2012). In this article, the authors explains an old result by Rüschendorf which shows that the p-values can be combined by scaling up their average by a factor of two. In our case since the there are only two p-values, we calculate the combined p-value as below:

$$p_{combined} := \frac{2}{K}(p_1 + p_2 + ... + p_K)$$

with $K = 2$

3.4 Brief survey of existing approaches for associating transcription factors to gene lists

The task of associating transcription factors with genes based on specific biological conditions has been tackled by some researchers. The work done in this field that we came across has been described and compared with in this section.

3.4.1 ChIP-X Enrichment Analysis (ChEA)

ChEA is a software tool that utilizes ChIP-X experiments data for linking transcription factors to gene expression changes by computing over-representation of transcription factor targets in an input list of genes. ChEA essentially counts the number of targets in a list and compares them with the number of targets that were identified in the database, i.e. an ORA approach of transcription factor targets on an input list of genes (Lachmann et al. 2010).

ChEA is based on a manually curated database from the literature reporting ChIP-X experiment results. In this database, each record contains a list of genes potentially regulated by a specific transcription factor under a specific condition. This database was then used as the prior knowledge base to analyze mRNA expression data where enrichment analysis was performed. The current database as of September 2014 has the following statistics:

- Transcription Factors: 209
- Publications: 237
- Genes: 47197

- Total Entries: 483786

ChEA is commonly used after a genome-wide gene expression profiling study is performed. The steps that follow are: First, a list of genes that significantly changed their expression levels is prepared and given as an input to the ChEA software. Next, the software computes over-representation for targets of transcription factors per study in the ChIP-X database. To compute statistical enrichment, ChEA implements the Fisher exact test with Bonferroni's correction, where the proportions for the test are the number of genes in the input list, the number of genes identified in the ChIP-X experiment, the genes that are shared among the two lists, and the number of overall targets in the ChIP-X database. Finally, ChEA reports a ranked list of ChIP-X experiments that show statistically significant overlap with the input list. Identified genes from the input list, potentially regulated by a specific transcription factor, are also connected and visualized as a network, using known protein - protein interactions.

The ENCODE (Consortium et al. 2012; Raney et al. 2011) project was started with the main purpose of finding the functional elements of the human genome. Along with assigning function to DNA elements, a big part of the ENCODE project is to identify the transcription factor binding sites on the entire human genome.

The results from the experiments performed under ENCODE provide us with details about the location in the genome where a transcription factor binds and also the intensity of its binding. ChEA2 (Kou et al. 2013) also includes in its database all ChIP-X experiments from the ENCODE project.

Similarities between ChEA and REPA

The two programs are similar in the following ways:

- Both ChEA and REPA attempt to identify transcription factors regulating a collection of genes thereby predicting likely regulators of biological systems under study.

- The data used by both ChEA2 and REPA comes from the ENCODE project.

Differences between ChEA and REPA

- REPA uses a functional class scoring algorithm instead of the over representation analyses approach used by ChEA. This allows REPA to identify gene sets instead of transcription factors that may regulate a list of genes that may or may not be co-functional. By using over representation analyses approach ChEA suffers from disadvantages arising from using hard cutoffs.

3.4.2 Inferring condition-specific transcription factor function from DNA binding and gene expression data

'CRACR' (McCord et al. 2007) (Combination Rank-order Analysis of Condition-specific Regulation; pronounced 'cracker'), derives information about condition-specific gene regulation and transcription factor activity by combining comprehensive, condition-independent protein binding microarray (PBM) (Berger and Bulyk 2009) data for a given TF with gene expression microarray data under a variety of biological conditions. Specifically, CRACR searches for conditions in which differentially expressed genes are enriched or genes whose upstream intergenic regions (IGRs) contain a pattern to which a transcription factor has significant preference in PBM data. In contrast to earlier studies, CRACR integrates PBM-derived transcription factor sequence preference data with gene expression data without imposing arbitrary cut-offs that define which IGRs are 'bound' or which genes are 'differentially expressed'. In addition, CRACR uses rank order statistics, which facilitates comparison of gene expression data from different microarray platforms.

To predict the condition specific functions of yeast Saccharomyces cerevisiae's transcription factors, the team first collected 1327 publicly available gene expression microarray data sets for Saccharomyces cerevisiae. Each of these datasets refers to a specific cellular condition. Next, for each of these conditions, the genes were ordered based on their expression fold change levels. At the top were the genes that were highly induced, and at the bottom the ones that were repressed. Parallel to the previous step, ranks were assigned to the genes according to the PBM P-values of transcription factors binding to their upstream IGRs. Then using a rank based statistical test, a comparison was made between the PBM defined ranks of similarly expressed genes within a sliding foreground window to the ranks of a length-matched background set of genes outside this window. The result of this statistical test yields a value which is referred to as the enrichment score and represents the degree to which PBM-derived target genes of a given TF are significantly enriched within each window of similarly expressed genes. The statistical significance of the maximum enrichment in a condition is derived by permutation testing. Using the method described above, CRACR can list expression conditions in which predicted transcription factor target genes show statistical significance in expression levels. From such a list of cellular conditions, one can hypothesize about the functions of the transcription factor.

Difference between CRACR and our approach

- CRACR focuses on similarly expressed genes whereas REPA looks at gene sets that represent biological pathways. Similarly expressed genes may or may not belong to the same gene set.
- CRACR attempts to derive information about condition specific gene regulation and transcription factor binding whereas the goal of this project is to help provide researchers information about which biological pathways are active in

the biological condition they are studying along with the transcription factors that are most likely to regulate these pathways.

- CRACR ranks the genes based on their expression levels and then for each transcription factor, it compares the p-values (derived from in-vitro PBM experiments) of similarly expressed genes with the p-values of the other genes in the list. PBM data does not account for the cell's chromatin state. On the other hand our approach is dependent on in-vivo ChIP-X data which takes into account the cell's chromatin state and hence the results obtained can be expected to have a higher biological significance although chromatin state depends on cell's state and cell line.

- CRACR used data from studies on Saccharomyces cerevisiae (yeast) whereas this project focuses on human data from the ENCODE project.

Chapter 4

Results and Comparison

The goal of this chapter is to assess the accuracy and strength of REPA. For the first module, i.e. regulation pathway analysis, we identified TF - gene set associations with a strong signal and predicted these as true TF - gene set associations. Then we analyzed the accuracy of these predictions. Since we expect many of the predictions to be novel, we might not find any evidence in the existing literature confirming them even when the predictions are correct. This is probably the most significant contribution of REPA as it points to new research avenues.

For the second module, i.e. enrichment analysis, we look at the results obtained after running REPA with sample data from studies involving gene expression analysis and compare the results with preexisting applications.

Finally we combine the results of the two above mentioned parts and present an alternative analysis of the data.

4.1 Module 1: Regulation Pathway Analysis

The data obtained from the ENCODE project provided transcription factor binding sites information for 131 distinct human transcription factors. The MSigDB and

KEGG pathway libraries provided a total of 10192 gene sets. The results obtained after running the first module of REPA with this data are discussed in the following sections.

4.1.1 Inferred associations

REPA calculated a permutation-based p-value for a total of 803711 transcription factor - gene set associations and predicted that 68008(8.5%) of these pairs are true associations. To explore the effect of the significance threshold used in the Mann-Whitney U test in the distribution of permutation-based p-values, we set the significance threshold at 0.001, 0.01 and 0.05.

Figure 4.1 shows the distribution of permutation based p-values. The plot has following features that indicate a successful differentiation between real and non-existing Transcription factor - gene set associations.

1. A peak on the left side, containing Transcription factor - gene set associations with strong signal for which the null hypothesis was rejected.
2. A uniform distribution between the interval [0.015,1].
3. A peak at the upper end at 1.

The last two features correspond to transcription factor - gene set bindings for which the null hypothesis was accepted.

We consider the associations that form the peak on the left side as the predicted true associations. By performing a visual inspection of figure 4.1 we can see that the peak ends at about 0.015 (shown by the vertical red line). We set this as the cut-off for significant associations.

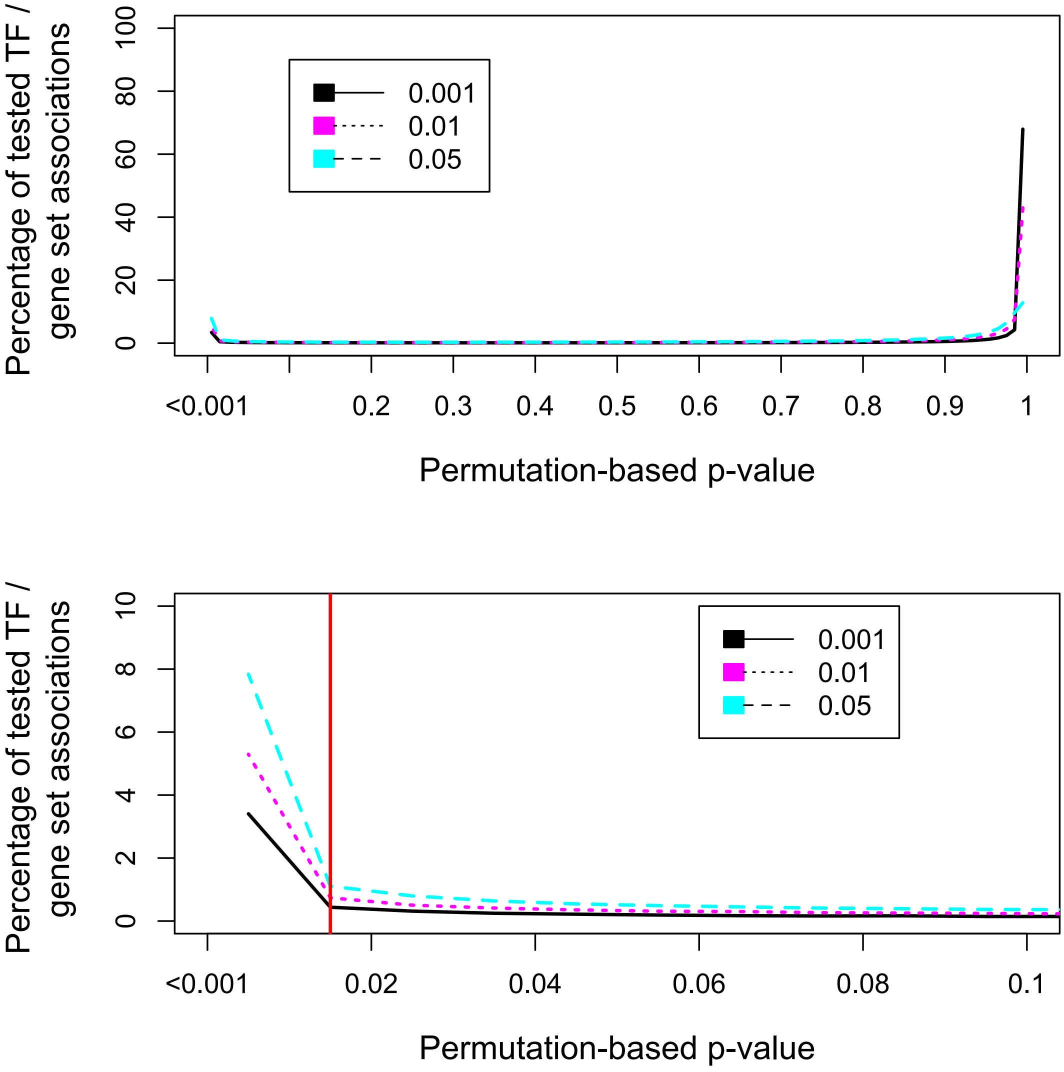


Figure 4.1: Transcription factor - Gene set association distribution

The number of REPA's predictions varied from 68008 to 29317 depending upon the significant threshold used for Mann-Whitney U test. Since the permutation based p-value distribution remain stable at the different significance thresholds for the Mann-Whitney U test, we set the significant threshold of Mann-Whitney U test to 0.05. There are total 68,008 (8.5%) associations that form this peak.

4.1.2 Number of REPA's predictions per transcription factor

As discussed earlier, we executed module one with over 10,000 gene sets as input. Several of these gene sets were created based on the location of their member genes or were computationally generated as described in section 3.3.1. For performing further analysis including taking a closer look at specific predictions, we only focus on manually curated gene sets that represent functions or pathways. These gene sets are more widely studied and usually include findings from more than a single study. Hence, from now on, we considered predictions involving a total of 2,677 gene sets from the following sources.

- Canonical pathways and functions representing biological pathways curated by domain experts. This includes pathway databases such as KEGG (282 gene sets), GO (1454 gene sets), REACTOME (674 gene sets) and BIOCARTA (217 gene sets).

- Hallmark gene sets representing specific well-defined biological states or processes and display coherent expression (Total 50 gene sets).

Out of the 68,008 predictions, 8,948 (13.2%) pass this criterion. Figure 4.2 shows the number of REPA's predictions per transcription factor. There are 88 unique transcription factors that could be found in REPA's predictions. 74 (84%) are associated to at least one gene set. 28 (37.8%) are associated to more than 50 gene sets and three (3.4%), namely GR, POL2 and YY1, are associated with more than 500 gene sets.

Next we discuss why it makes sense that these three transcription factors are predicted to regulate a large amount of gene sets.

Glucocorticoid Receptor (GR) It is known that GR regulates diverse cellular functions (such as mitosis and apoptosis) and essential biological processes (such

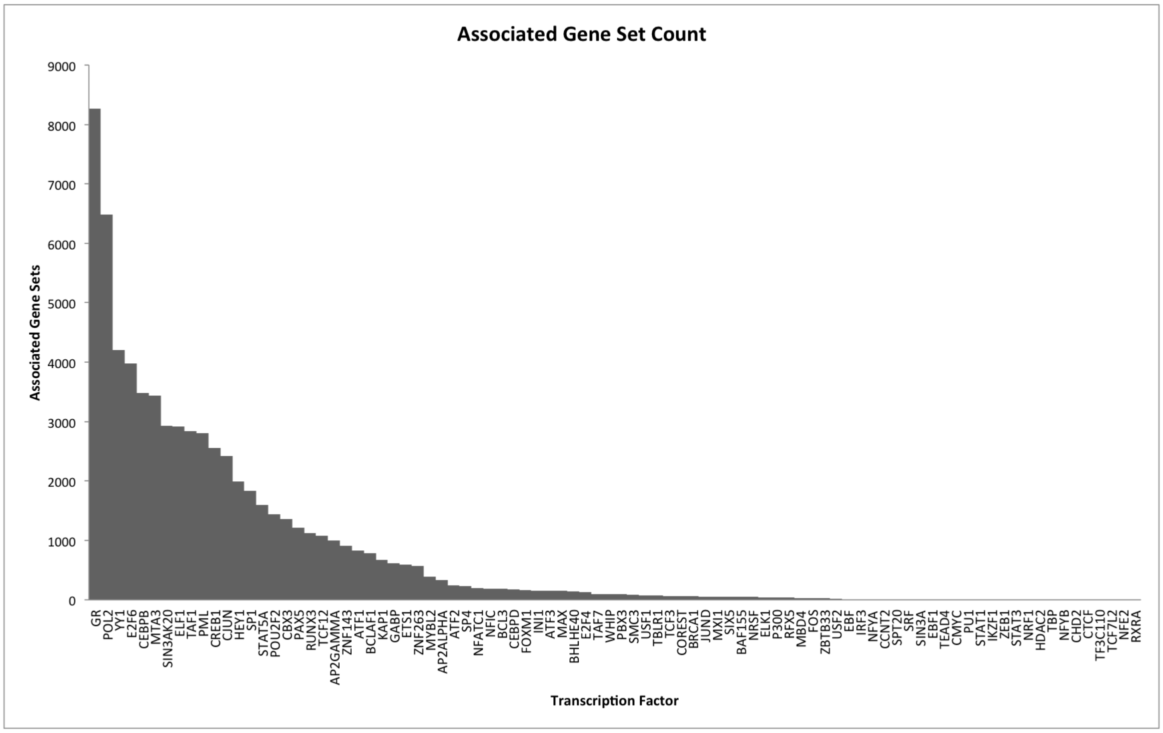


Figure 4.2: Predictions per transcription factor

as growth, development, metabolism, and behaviour), and is expressed in most cell types (Zhou and Cidlowski 2005; Lu and Cidlowski 2005). To explain how a single transcription factor can regulate such a variety of processes, it has been proposed that different GR isoforms allow for regulation of genes in a cell type specific manner, and that each GR isoform regulates both a common and a unique group of genes in each cell type (Zhou and Cidlowski 2005; Lu and Cidlowski 2005).

Polymerase (RNA) II (DNA directed) polypeptide A (POL2) The *POL2* gene encodes for the largest subunit of RNA polymerase II, the polymerase that synthesize mRNA in eukaryotes. POL2 is a general transcription factor that initiates transcription and is responsible for transcriptional regulation (Orphanides, Lagrange, and Reinberg 1996).

Yin Yang 1 (YY1) YY1 is an ubiquitously expressed transcription factor that regulates cell proliferation and differentiation, and is a multifunctional mediator of different signaling pathways that modulates an impressive and increasing list of genes (Deng et al. 2010)

Thus, one would expect that these three transcription factors regulate many different gene sets and REPA's results reflect this.

4.1.3 Number of REPA's predictions per gene set

We also looked at the number of REPA's predictions per gene set. There were 1,980 canonical pathways, GO gene sets and hallmark gene sets predicted to be regulated by at least one transcription factor. Out of these 1,980, 30 (or 1.5%) were associated to more than 30 transcription factors. Figure 4.3 shows gene sets associated by REPA with at least 25 transcription factors. Many of the gene sets listed in Figure 4.3 are tightly regulated cellular processes. We examined the literature related to the regulation of the Reactome pathway and the KEGG pathway associated to the largest number of transcription factors.

The Reactome pathways metabolism of RNA and cell cycle mitotic were the Reactome pathways associated to the largest number of transcription factors. The Reactome pathway metabolism of RNA has been deleted from Reactome since version 50 (current version is 52); thus, we investigated the cell cycle mitotic pathway which contains 325 genes. Cell cycle regulation is critical for growth and development, and its misregulation plays an important role in diseases such as cancer. There is a large variety of cell cycle programs within a single species that corresponds to specific cell types, developmental stages or physiological conditions (Harashima, Dissmeyer, and Schnittger 2013). In the budding yeast, cell cycle is controlled by a large

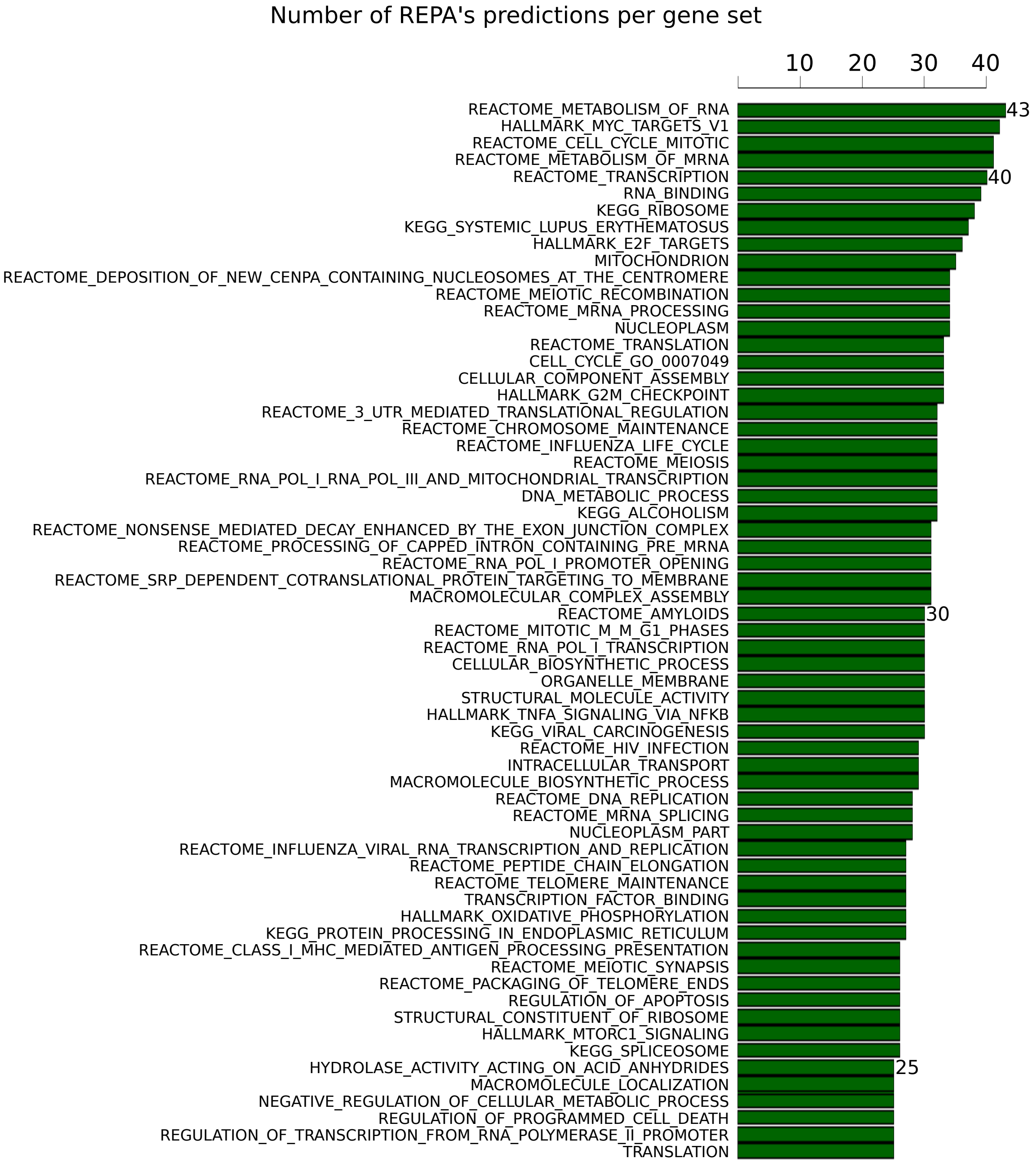


Figure 4.3: Predictions per gene set

and complex interacting network of regulatory proteins, and the general organization of this control system is conserved across the Eukaryota (Haase and Wittenberg 2014). REPA associated the following 41 transcription factors with the Reactome cell cycle mitotic pathway (literature supporting the involvement of these transcription factors during the cell cycle is referred to after the corresponding transcription factor): AP2ALPHA (Prasov and Glaser 2012), AP2GAMMA, ATF1 (Bandyopadhyay et al. 2014), BCLAF1, CBX3, CEBPB, CJUN, CREB1, E2F4, E2F6 (Haase and Wittenberg 2014), ELF1, ELK1 (Demir and Kurnaz 2013), ETS1 (Oikawa and Yamada 2003), FOXM1 (Wierstra and Alves 2007), GABP (Imaki et al. 2003), GR (Zhou and Cidlowski 2005; Lu and Cidlowski 2005), HEY1, INI1 (Versteege et al. 2002), KAP1 (White et al. 2012), MAX (Amati and Land 1994), MTA3, MXI1 (Lee and Ziff 1999), MYBL2 (Joaquin and Watson 2003), PAX5, PBX3, PML, POL2, POU2F2 (Prasov and Glaser 2012), RUNX3, SIN3AK20, SMC3, SP1, STAT3, STAT5A, TAF1, TBLR1, TCF12, WHIP, YY1, ZNF143, and ZNF263. Thus, we found literature support for 14 (or 34%) of these transcription factors. The central components of the cell-cycle control system, cyclin-dependent protein kinases (CDKs), are missing from this list because they were missing from the linked-transcription factor binding data used as REPA's input.

The KEGG ribosome pathway consists of 135 genes including the ribosomal proteins and ribosomal RNAs. The mechanisms regulating ribosome biogenesis are only partially understood, and they are the focus of current research. Recently, ribosome biogenesis has been linked to various diseases and aging, and studies have revealed an elaborate control of ribosome biogenesis that requires coordinate regulation of all three RNA polymerases and that includes feedback and feed-forward loops (Lempiainen and Shore 2009; Thomson, Ferreira-Cerca, and Hurt 2013). A large number of transcription factors have been implicated in ribosome biogenesis; for instance, roughly 80 factors have been associated in the maturation of the 60S subunit (Thomson, Ferreira-Cerca, and Hurt 2013). REPA predicted the following 38 transcription factors to be associated with the KEGG ribosome pathway (supporting literature is referred to after the corresponding transcription factor): ATF1, ATF2, ATF3, BCL3, BCLAF1, CBX3, CEBPB, CJUN, CREB1 (Nosrati, Kapoor, and Kumar 2014), ELF1 (Algire et al. 2002), ETS1, GABP (Perry 2005), GR (Shah et al. 2002), HEY1, IRF3, MTA3, MYBL2, NFATC1, NFIC, NRSF, PML (Vilotti et al. 2012), POL2 (Lempiainen and Shore 2009), POU2F2, RFX5, RUNX3, SIN3AK20, SIX5, SP1 (Perry 2005), SP4, STAT5A, TAF1 (Lin et al. 2002), TAF7, TBLR1, TCF12, WHIP, YY1 (Perry 2005), ZNF143, and ZNF263. Thus, we found literature support for 9 (or 24%) of these transcription factors.

We found an intriguing gene set (i.e., systemic lupus erythematosus (SLE)) among the top 10 genes sets shown in Figure 4.3, and thus decided to look at the transcription factors inferred by REPA to be associated with this disease. SLE is an autoimmune disease with more than 40 genes and loci identified as associated with this disease. However, these genes and loci only account for 10 to 20% of disease heritability. This indicates that there are many factors still to be identified (Frangou, Bertsias, and Boumpas 2013). REPA inferred 37 transcription factors associated with the KEGG systemic lupus erythematosus pathway which consists of 138 genes. Out of those 37 transcription factors predicted to regulate genes in the SLE pathway, we found literature support for 13 (or 35%). The transcription factors associated by REPA with SLE are the following (supporting literature is referred to after the corresponding transcription factor): ATF2, ATF3 (Cai et al. 2014), BCLAF1, CBX3, CEBPB, CEBPD, ETS1 (Lu et al. 2015), FOS (Frangou, Bertsias, and Boumpas 2013), FOXM1, GR (Chen et al. 2015), HEY1, INI1, JUND (Tenbrock et al. 2007), MBD4 (Balada et al. 2007), MTA3, MYBL2, NFATC1 (Tenbrock et al. 2007), NFIC, P300 (Leung et al. 2015), PAX5 (Dozmorov, Wren, and Alarcón-Riquelme 2014), PML, POL2, POU2F2, RUNX3 (Jeffries et al. 2011), RXRA, SIN3AK20, SP1 (Hikami et al. 2011), SP4, STAT5A, TAF1, TAF7, TBP (Chauhan et al. 2004), TCF12, TCF3, TEAD4, YY1 (Zhao et al. 2012), and ZBTB33.

4.1.4 Literature based evaluation of REPA's predictions

Since REPA is the first approach to systematically infer relationships between transcription factors and gene sets, there is no benchmark available that can be used to do a direct comparison. Therefore, to gain intuition into the quality of REPA's predictions and assess the level of precision, we performed a literature analysis on 50 randomly selected predictions. The number of predictions we could examine was limited by available resources as literature curation is a time consuming effort. For this analysis, we focused only on predictions that associate transcription factors to well established manually curated gene sets from sources such as KEGG, Biocarta, Gene Ontology, MSigDB's hallmark collection and Reactome. These manually curated gene sets are well studied unlike several other gene sets which could be created on the basis of a single study. To avoid over-weighting particular transcription factors, we only allowed two predictions per transcription factor. Related literature was searched using PubMed, Disgenet (Bauer-Mehren et al. 2011), and ChEA (Lachmann et al. 2010). Table 4.1 contains the list of 50 randomly selected REPA's predictions that we investigated for literature support. The number of genes in the gene sets examined varied from 57 to 320, and the percentage of genes in the gene set with a binding signal value varied from 18% to 94% (see third column of Table 4.1).

We classified the evidence found in the literature into four types:

1. Direct evidence

2. Binding evidence

3. Indirect evidence

4. Refuting evidence

Transcription Factor	Gene Set (GS)	# of genes in GS with a binding signal value / # of genes in GS	p-Value	Type of evidence	Reference
ATF2	Meiosis	38/116	0.001	NE	
ATF2	Ribosome	68/135	0.001	I	Johnson et al. 2003
BCLAF1	Alcoholism	52/180	0.001	NE	
BCLAF1	Chromosome maintenance	49/122	0.001	I	Lee et al. 2012b
CEBPB	Spliceosome	69/131	0.001	B	Lefterova et al. 2010
CEBPB	Viral carcinogenesis	128/206	0.001	D	Watanabe-Okochi et al. 2013
CJUN	FOXO signalling	36/133	0.001	D	Xu et al. 2011
CREB1	E2F targets	105/200	0.001	B	Zhang et al. 2005b; Martianov et al. 2010
CREB1	MYC targets v1 (Hallmark)	104/200	0.001	B	Zhang et al. 2005b
E2F6	Pathways in cancer	145/327	0.001	D	Oberley, Inman, and Farnham 2003
E2F6	Signaling by the B Cell Receptor	59/126	0.001	B	Lam et al. 1999
ELF1	Huntington's disease	100/183	0.001	B	Hollenhorst et al. 2007
ELF1	Spliceosome	61/131	0.001	B	Hollenhorst et al. 2007
FOXM1	Alcoholism	49/180	0.001	NE	
GABP	Metabolism of RNA	118/330	0.001	B	Hollenhorst et al. 2007; Wallerman et al. 2009
GABP	Ribosome	64/135	0.001	D	Donadini et al. 2006
GR	Acute myeloid leukemia	54/57	0.001	D	Haarman et al. 2002
GR	Neurotrophin signalling	110/120	0.001	I	Adachi et al. 2014
HEY1	HIV infection	101/207	0.001	I	Pinzone et al. 2015; Wang et al. 2014b
HEY1	RNA transport	75/164	0.001	NE	
JUND	Alcoholism	48/180	0.001	I	Taqi et al. 2011
KAP1	miRNAs in cancer	69/296	0.012	D	Min et al. 2013
MBD4	Amyloids	22/83	0.001	NE	
MBD4	Systemic lupus erythematosus	25/138	0.002	D	Balada et al. 2007
MTA3	Herpes simplex infection	106/188	0.001	NE	
MTA3	Ribosome biogenesis in eukaryotes	40/85	0.004	NE	
MYBL2	Alcoholism	52/180	0.001	NE	
NFATC1	Systemic lupus erythematosus	45/138	0.001	D	Lu et al. 2008
NFIC	Viral carcinogenesis	107/206	0.005	I	Schuur et al. 1995
PAX5	Epstein Barr virus infection	101/203	0.001	D	Tierney et al. 2007
PML	HTLV I infection	129/263	0.001	D	Ariumi et al. 2003
PML	Viral carcinogenesis	103/206	0.001	D	Singh et al. 2013
POL2	Bacterial invasion of epithelial cells	65/76	0.001	D	Lutay et al. 2013
POL2	Cytokine Cytokine receptor interaction	156/271	0.001	NE	
POU2F2	mTORC1 signaling	89/200	0.001	NE	
POU2F2	Transcriptional misregulation in cancer	47/179	0.001	NE	
RUNX3	Metabolism of RNA	140/330	0.001	NE	
RUNX3	Protein processing in endoplasmic reticulum	86/167	0.001	I	Evans et al. 2011
SIN3AK20	Metabolism of mRNA	101/284	0.001	D	Dong et al. 2007
SP1	Alcoholism	63/180	0.001	D	Harada et al. 1998
SP1	Carbon metabolism	51/105	0.002	D	Lin, Lai, and Chau 2011
STAT5A	Herpes simplex infection	89/188	0.001	R	Kriesel et al. 2004
TAF1	Huntington's disease	87/183	0.001	I	Kaji et al. 2005
TAF1	Ribosome	65/135	0.001	D	Lin et al. 2002
TCF3	Meiotic synapsis	26/73	0.004	B	Cole et al. 2008
TCF12	Class I MHC mediated antigen processing & presentation	112/251	0.001	NE	
TCF12	E2F targets	110/200	0.001	NE	
YY1	Oxidative phosphorylation	61/133	0.001	D	Lescuyer, Martinez, and Lunardi 2002
YY1	Spliceosome	62/131	0.001	B	Mendenhall et al. 2010
ZNF143	RNA Transport	70/164	0.009	I	Yuan et al. 2007

Table 4.1: REPA's predictions evaluated based on the literature. In type of evidence; B indicates "binding", D "direct", I "indirect", NE "No Evidence" and R "Refuting Evidence"

Direct evidence indicates that current literature directly links a transcription factors with a given gene set; for example, the association of the transcription factor YY1 with oxidative phosphorylation is supported by promoter analysis for a complex I gene (NDUFS8) (Lescuyer, Martinez, and Lunardi 2002).

Binding evidence indicates that targets of a transcription factor identified by a published Chip-Seq or Chip-ChIP study are over-represented in a given gene set. This type of evidence was found using ChEA. In the case of binding evidence, current literature does not directly discuss the link of that transcription factor with the given gene set.

Indirect evidence indicates that current literature suggests the involvement of a transcription factor in the regulation of a given gene set; for example, in (Taqi et al. 2011), is suggested that a single nucleotide polymorphism (SNP) in the promoter region of PDYN, which is associated with alcohol-dependence, may impact PDYN transcription in the human brain and that this SNP is located within a regulatory region that may be targeted by the transcription factor JUND.

Finally, refuting evidence indicates that current literature contains experimental data against the prediction; for example, REPA associated STAT5A with herpes simplex infection; however, current literature (Kriesel et al. 2004) shows that STAT1, but not STAT5A, binds to the herpes simplex virus latency-associated transcript promoter.

Out of the 50 predictions investigated, 17 (34%) were supported by direct evidence and 9 (18%) by binding evidence (see Table 4.1). These 26 (52%) associations supported by direct or binding evidence were correct or likely to be correct associations. Out of the 50 associations examined, 23 had indirect evidence or could neither be confirmed nor refuted by current literature. One (2%) of the 50 associations investigated was considered to be incorrect or likely to be incorrect based on current literature. These results suggest that REPA's precision lies above 52%. This level of precision is quite promising as several of REPA's predictions are expected to be novel and therefore lack literature support. Moreover, gene function predictions with similar precision levels have been successfully used in yeast (Peng et al. 2003) and mouse (Peña-Castillo et al. 2008). Based on this, we expect REPA's predictions to be an useful resource to guide further biological research.

4.1.5 Estimation of REPA's recall

The Collection 3: Transcription Factor Targets (C3-TFT) of MSigDB consists of 615 gene sets that contain genes that share a transcription factor binding site defined in the TRANSFAC (version 7.4, `http://www.gene-regulation.com/`) database. Each of these gene sets is annotated by a TRANSFAC record. Additionally, the transcription factor known to bind the given motif is provided for 500 of these 615 gene sets. In total, 282 transcription factors are matched to a given DNA-binding motif in the C3-TFT collection. Out of these 282 transcription factors, 33 are also present in the linked-TFBD used as REPA's input, and REPA generated a prediction for 26 of them (see Figure 4.4). The 7 transcription factors for which REPA did not make a prediction had binding signal values for very few genes (four of them had binding signal values for less than 505 genes). To estimate REPA's recall, we counted the number of transcription factors associated by REPA with the corresponding C3-TFT gene set; for instance, since REPA associated SP1 to the V$SP1_01 gene set (which consists of genes whose promoter regions contain the motif GGGGCGGGGT which matches annotation for SP1) we counted SP1 as successfully retrieved by REPA.

REPA associated 14 TFs to their corresponding C3-TFT gene set. Based on this, REPA's recall is at 53.8% (14/26).

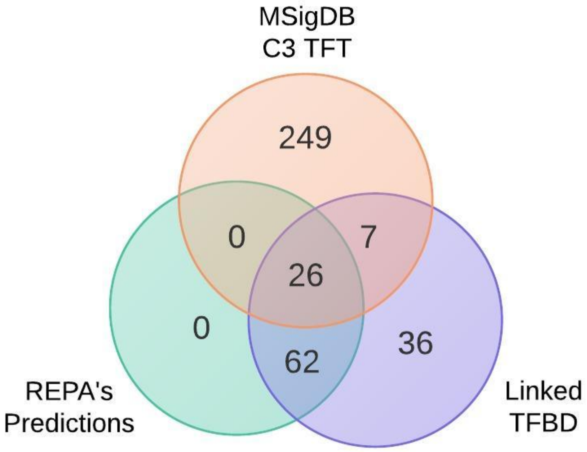


Figure 4.4: Venn diagram indicating the number of TFs in common between MSigDB C3-TFT collection, ENCODE's TFBD and REPA's predictions.

4.2 Module 2: Enrichment analysis

Over the past decade several methods and software tools were developed to address the task of enrichment analysis. Many approaches adapted functional class scoring (FCS) as their preferred technique. As the methods grew in sophistication and became more matured the accuracy of the results increased (Khatri, Sirota, and Butte 2012; Khatri and Drăghici 2005).

To test the performance and accuracy of the enrichment analysis using FCS module of REPA (Regulation Expression Pathway Analysis) we compare our results against those of Generally Applicable Gene-set Enrichment (GAGE). GAGE is arguably the most widely used gene set enrichment analysis system. GAGE was published in the year 2009 (Luo et al. 2009). Before GAGE, tools such as Gene Set Enrichment Analysis or GSEA (Subramanian et al. 2005) and Parametric Analysis of Gene set Enrichment or PAGE (Kim and Volsky 2005) were still widely used tools but those methods had limited usage because they couldn't handle datasets of different sample sizes or experimental designs. GAGE overcame those limitations and when compared with GSEA and PAGE showed significantly improved results as discussed in (Luo et al. 2009). The datasets that were used for the comparison between GAGE, GSEA and PAGE in (Luo et al. 2009) were of varying sample sizes, experimental designs and microarray profiling techniques. We compare REPA with GAGE using the same datasets along with one more dataset that was derived from a study done to understand infection of influenza A viruses.

The purpose of this comparison is to confirm that REPA's results are mostly in agreement with GAGE's results. Module 2 was added to REPA for the convenience of running the complete analysis within the same software. However, results from REPA's module 1 (regulatory enrichment) can be combined with results from other tools for gene set analysis of expression data such as GAGE. REPA's module 1 is the novel contribution of this thesis.

4.2.1 Description of the datasets

To perform the comparisons we used the datasets provided by the R package "gageData" (Luo 2013). Gagedata is a supporting data package for the software package, GAGE (Luo et al. 2009). It contains microarray datasets that GAGE uses in its paper to compare itself against GSEA and PAGE. Therefore the data supplied here is also useful for our purposes since we can use it to run REPA and compare it against GAGE.

The R package "gageData" contains the microarray datasets from the following two experiments.

Case 1: BMP6 treated vs untreated hMSC

In this study, microarrays were used to profile the global gene expression in human mesenchymal stem cells (hMSC). These hMSCs were treated with an endogenous regulator Bone morphogenetic protein 6 or BMP6. The dataset contains a total of 4 gene chip measurements from duplicate experiments each with paired measurements of human MSC with or without 8 hours of BMP6 treatment (Luo et al. 2009). This is a typical small dataset with as few as two samples per condition. BMP6 treated samples and controls are one-on-one matched. This dataset is also registered as GSE13604 in Gene Expression Omnibus or GEO (Edgar, Domrachev, and Lash 2002).

Case 2: Dataset derived from breast cancer study on 12 patients

This study covers 12 breast cancer patients each with histologically normal (HN), atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS) RMA samples (Emery et al. 2009). Due to space constraints, the dataset was split into two halves. The gage package only includes the first half dataset for 6 patients as the example dataset of this study as gse16873. Most of the comparison analyses that was done by GAGE were done on these samples. For our comparison purposes we use the gse16873. This dataset is also registered as GSE16873 in GEO.

Case 3: Infection of influenza A viruses

The goal of this study was to investigate the early host responses of influenza A viruses in human lung epithelial cells (Gerlach et al. 2013a). Gene expression profiling was performed on host transcriptional responses of well-differentiated, primary human bronchial epithelial cells during infection of influenza A viruses. We used gene expression data of uninfected cells and cells infected with H1N1pdm isolates from a nonfatal case (A/KY/136/09). The data set contains 3 biological replicates for each group. This dataset is also registered as GSE48466 in GEO.

4.2.2 Running GAGE with the datasets

To run GAGE with the first dataset we took the following steps:

1. Load the gage and gageData libraries.

   ```
   library(gageData)
   library(gage)
   ```

2. Load the gene sets.

   ```
   c2.gs=readList("msig5_kegg.txt")
   ```

3. Run GAGE with the following arguments.

   ```
   bmp6.c2.p <- gage(bmp6, gsets = c2.gs, ref =c(1,3),
   samp = c(2,4), same.dir = FALSE, compare = "as.group")
   ```

   Description of the parameters:

   - bmp6 contains the gene expression data.
   - gsets is a named list. Each element contains a gene set.
   - ref = c(1,3) is a numeric vector and indicates that in this data, columns 1 and 3 store the expression levels of the controls.
   - samp = c(2,4) indicates that columns 2 and 4 store the expression levels of the treated samples.

- same.dir is a boolean. It indicates whether the input gage result test for changes in a gene set toward a single direction (all genes up or down regulated, same.dir = TRUE) or changes towards both directions simultaneously (same.dir = FALSE). Since in REPA we consider absolute scores (fold change, log ratio etc.) we set same.dir as FALSE.
- compare = “as.group”. This argument indicates which comparison scheme to be used. “as.group” indicates that group-on-group comparison between the controls and treated samples.

4. Finally we select only the cases that gave complete results and are significant with their adjusted p-values less than 0.01.

```
results <- bmp6.c2.p$greater[complete.cases(bmp6.c2.p$greater),]
gage.bmp6.sigResults <- results[results[,"q.val"] < 0.01,]
```

The same procedure was followed for the datasets from case 2.

For case 3, we executed gage with the following parameters:

```
gage_results_influenza <- gage(gse48466_perGene,
gsets = msigv5_kegg.gs, ref = 10:12, samp =4:6,
same.dir = TRUE, compare = "unpaired")
```

4.2.3 Running REPA with the datasets

In this section we go through the steps that were taken to run our system with the datasets provided in the the R package “gageData” (Luo 2013) to perform the comparison between the two systems.

Preprocessing with Limma

Limma or Linear Models for Microarray Data is a R package for differential expression analysis of data arising from microarray experiments (Smyth 2005). We use Limma with the raw data from the experiments to fit a linear model to the expression data for each gene.

Inputs to Limma:

1. **Design matrix** The design matrix for the smaller dataset BMP6 is stored in a text file called "13604DesignMatrix.txt". It stores the information shown in the following table.

	Control	BMP6_Treated
8hCont1	1	0
8hTrt_8hCult1	0	1
8hCont2	1	0
8hTrt_8hCult2	0	1

Table 4.2: 13604 Design Matrix

The design matrix for the larger dataset follows the same pattern.

2. **Microarray data**

The second input to Limma is the raw expression level data from the microarray experiments. Table 4.3 shows the first few rows of the bmp6 dataset.

	8hCont1	8hTrt_8hCult1	8hCont2	8hTrt_8hCult2
10000	6.666482	6.727039	7.859644	7.888743
10001	9.874859	9.873068	9.838792	9.757909
10002	5.524512	5.651697	5.299609	5.146715
10003	4.604491	4.661876	4.790255	4.705559
10004	7.904135	7.883218	8.00505	7.962769
⋮	⋮	⋮	⋮	⋮

Table 4.3: bmp6 expression data sample

The expression data table for the larger breast cancer patient dataset also follows the same pattern.

Steps in running Limma

1. Load the library Limma using the following command.

   ```
   library(limma)
   ```

2. Next load the experiment design matrix.

   ```
   designMatrix <- read.table("13604DesignMatrix.txt")
   ```

3. Fit linear model for each gene from the given series of arrays. The variable bmp6 holds the data shown in Table 4.3

   ```
   fit <- lmFit(bmp6, designMatrix)
   ```

4. The next step is the contrast step, which uses the contrasts.fit() function. This allows the fitted coefficients to be compared in as many ways as there are questions to be answered, regardless of how many or how few these might be.

   ```
   contrastFit <- makeContrasts(BMP6_Treated-Control, levels=fit)
   ```

5. In this step we calculate the empirical Bayes statistics. The eBayes() function is used to rank genes in order of evidence for differential expression.

   ```
   eBayesTable <- eBayes(contrastFit)
   ```

6. Finally we extract a table of the top-ranked genes from the linear model. bmp6 contains the data from the gagedata library and is shown in the Table 4.3.

```
results <- topTable(eBayesTable, adjust = ‘‘fdr’’,
 coef = 1, number = nrow(bmp6))
```

The “results” variable contains the output from Limma. Table 4.4 shows few lines from the Limma results.

Gene	LogFC	Avg Expr	t	P.Val	Adj P.Val	B
5228	4.066771195	8.984909943	30.01850012	1.75E-06	0.030867069	2.719129297
8200	-2.937752131	9.096127188	-22.30635557	6.85E-06	0.060528957	2.461454621
3400	3.107500295	8.990624466	18.99550122	1.43E-05	0.066519298	2.261904761
133	-3.11990311	9.026165624	-17.69540658	1.98E-05	0.066519298	2.15765175
3399	3.819738229	11.38550593	17.22309034	2.24E-05	0.066519298	2.115050187

Table 4.4: Results after running Limma with bmp6 data

We input the Limma results along with the file containing all the gene sets to REPA. Few rows of REPA’s output are shown in the following table.

Test Name	Gene Set Name	Total Genes	Values Found	<0.001	<0.01	<0.05	>0.05	Average pValue
Two-sample student’s pooled t-test	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	62	60	0	1	12	987	0.461956
Two-sample student’s unpooled t-test	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	62	60	0	1	12	987	0.461977
Mann-Whitney U-test	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	62	60	0	1	3	996	0.521978
Two-sample student’s pooled t-test	KEGG_CITRATE_CYCLE_TCA_CYCLE	32	30	0	0	0	1000	0.695003
Two-sample student’s unpooled t-test	KEGG_CITRATE_CYCLE_TCA_CYCLE	32	30	0	0	0	1000	0.694662
Mann-Whitney U-test	KEGG_CITRATE_CYCLE_TCA_CYCLE	32	30	0	0	0	1000	0.717507
Two-sample student’s pooled t-test	KEGG_PENTOSE_PHOSPHATE_PATHWAY	27	26	0	0	3	997	0.414412
Two-sample student’s unpooled t-test	KEGG_PENTOSE_PHOSPHATE_PATHWAY	27	26	0	0	2	998	0.414801
Mann-Whitney U-test	KEGG_PENTOSE_PHOSPHATE_PATHWAY	27	26	0	0	9	991	0.511711

Table 4.5: Enrichment analysis module results

4.2.4 Comparison between results from GAGE and REPA

In this section we compare the results that were obtained from running GAGE and REPA with the same gene expression level data and the same gene sets.

Case 1: BMP6 treated vs untreated hMSC

The results obtained after running GAGE and REPA were stored in two lists. Both the systems were run with the same gene sets and datasets as inputs.

- The list from GAGE contained 133 gene sets that had an adjusted P.value of less than 0.01.
- The list from REPA contains 142 gene sets that had a P.value of less than 0.01.

We expected to see some significant overlap of the two as both the systems were trying to solve the same problem. To analyze the results we look at the results in the following three ways.

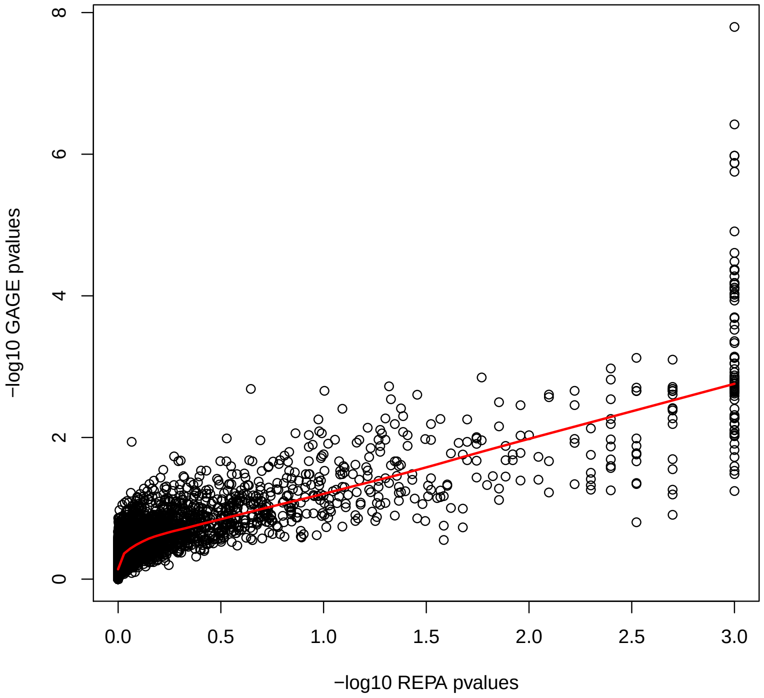


Figure 4.5: REPA vs GAGE plot for BMP6 dataset

The Spearman's correlation coefficient between the significant results obtained from the two systems is 0.491. Correlation coefficient for all the gene sets together is 0.542. In Figure 4.5 each gene set is represented by a circle. On the X-axis we have the log of REPA p-values and on Y-axis we have the log of GAGE p-values. This indicates that there is an agreement in how both systems rank the gene sets.

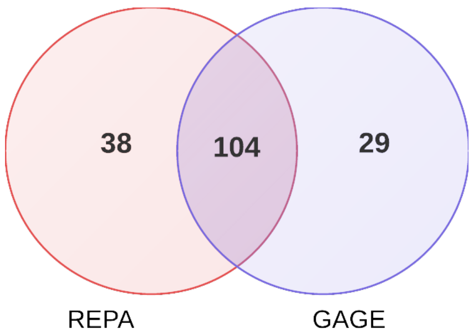


Figure 4.6: BMP6 dataset Venn diagram

Out of all the total 146 gene sets that were predicted to be significant by both systems combined, 22.2% were unique to REPA, 16.9% were unique to GAGE and 60.8% were common to both. The Venn diagram of this comparison is shown in figure 4.6.

Case 2: Dataset derived from breast cancer study on 12 patients

Similar to case 1, we get two lists of gene sets, one each from running GAGE and REPA. For comparing the results we again look at the correlation coefficients and the plots. GAGE predicted 463 significant gene sets where as REPA predicted 543 significant gene sets.

The Spearman's correlation coefficient for the two lists is 0.316 for significant gene sets and 0.547 for all gene sets combined. Figure 4.7 shows the plot between the two results.

In this case there were total 658 gene sets that were predicted to be significant by both systems combined. 17.5% were unique to GAGE whereas 29.6% were unique to REPA. 52.9% were common to both. The Venn diagram of this comparison is shown in figure 4.8.

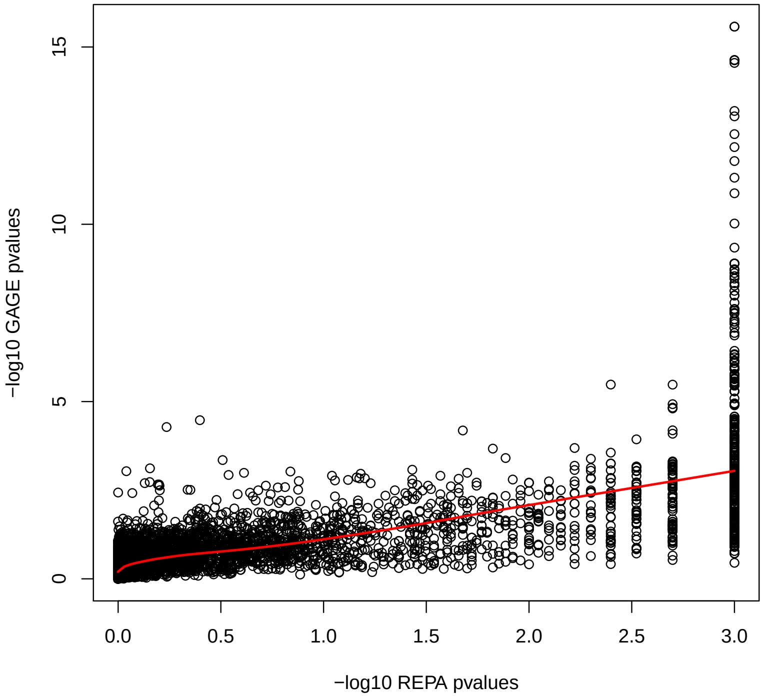


Figure 4.7: REPA vs GAGE plot for breast cancer dataset

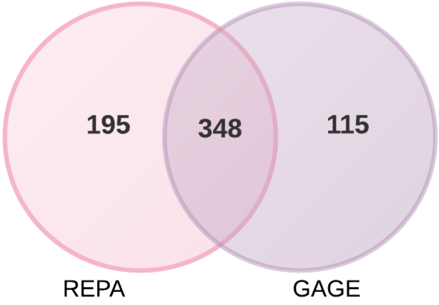


Figure 4.8: Breast cancer dataset Venn diagram

Case 3: Infection of Influenza A viruses

Here, we looked at gene expression profiling data obtained from a study performed to investigate the early host responses of seasonal and pandemic influenza A viruses in primary well-differentiated human lung epithelial cells (Gerlach et al. 2013b). We used gene expression data of uninfected cells and cells infected with H1N1pdm isolates from a nonfatal case (A/KY/136/09). We performed enrichment analysis using both GAGE and REPA on this dataset. Then we compared the list of significant gene sets produced by both the systems and made the following conclusions:

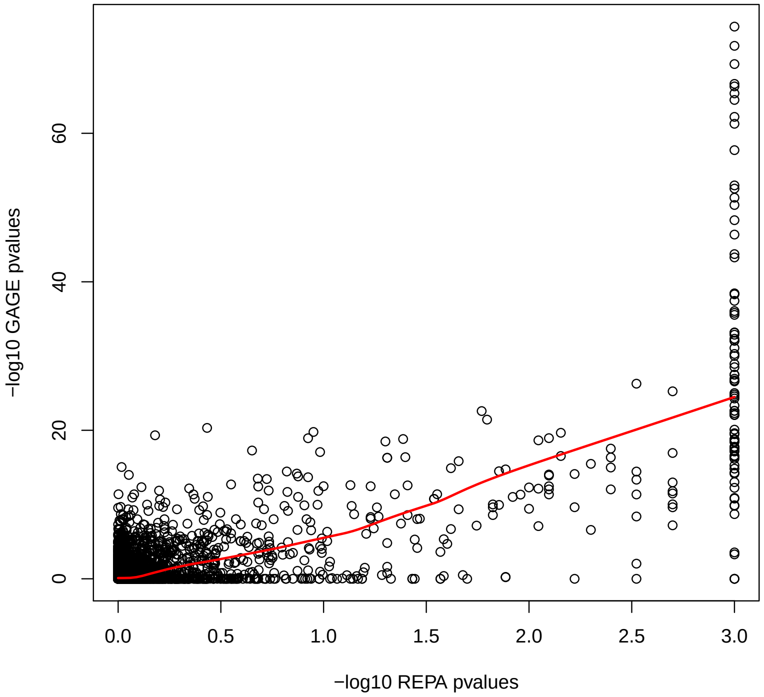


Figure 4.9: REPA vs GAGE plot for Infection of Influenza A viruses

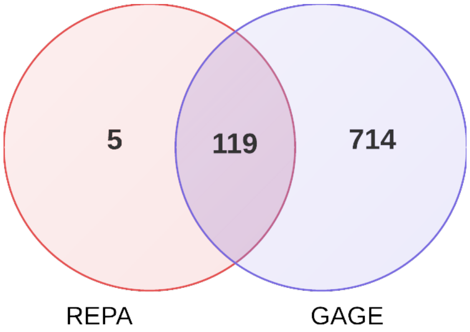


Figure 4.10: Infection of Influenza A viruses dataset Venn diagram

1. Out of the provided 10192 gene sets, REPA predicted 124 (1.22%) gene sets as significant where GAGE produced a much larger list of 833 (8.17%) significant gene sets.

2. The spearman's correlation coefficient (0.6726835 overall; 0.6480139 for significant gene sets) and the plot in figure 4.9 shows a higher degree of agreement in the rankings of the two systems when compared to the previous two cases.

3. Looking at the venn diagrams in figure 4.10, we can see that 119 out of 124 or 95.97% of the gene sets predicted by REPA are also present in the GAGE's list of significant gene sets.

In a specific case like this where there is a more comprehensive list generated by GAGE, we can substitute the results obtained by REPA for module 2 with the results obtained by GAGE or any other competing software and use them to process with the results obtained in REPA's module 1.

4.3 Module 3: Combining the results from the first two modules and deriving conclusions

The next and final step was to combine the results obtained from the first two modules and make final conclusions on the data. We looked at each of the three cases described in the previous section. Although we used all the available 10,192 gene sets for performing the enrichment analysis in module 2, in this section we only focus on the canonical pathways and manually curated gene sets. Therefore gene sets from the following sources:

- KEGG (283 gene sets)

- Reactome (674 gene sets)
- Biocarta (217 gene sets)
- Gene Ontology (1,454 gene sets)
- MSigDB's Hallmark gene sets (50)

Total 2,678 gene sets.

4.3.1 Case 1: BMP6 treated vs untreated hMSC

The final results were obtained by combining the outputs of the first two modules. The list of the statistically significant gene sets (p-value < 0.01 for MWU test with significant threshold = 0.05) and their likely regulators are given below in table 4.6.

Differentially expressed gene sets	**REPA's predicted putative regulators**
(H) Genes regulated by NF-kB in response to TNF	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, AP2GAMMA, CEBPD, ATF2, CREB1, SP1, JUND, NFIC, SIN3AK20, E2F6, ELF1, PML, BCLAF1, P300, CJUN, MTA3, RUNX3, STAT5A, TCF3, BCL3, POU2F2, NFATC1, BAF155
(H) Genes down-regulated in response to ultraviolet (UV) radiation	POL2, GR, YY1, TAF1, HEY1, CREB1, SP1, E2F6
(H) Genes up-regulated in response to Interferon Gamma	POL2, CEBPB, PAX5, GR, YY1, HEY1, IKZF1, EBF, E2F6, ELF1, PML, CJUN, MTA3, RUNX3, STAT5A, SMC3, POU2F2
(H) Inflammatory response	POL2, CEBPB, GR, E2F6, MTA3, POU2F2
(H) Genes defining early response to estrogen	POL2, GR, E2F6
(K) TGF-beta signaling pathway	POL2, GR, YY1, TAF1, E2F6
(H) APOPTOSIS	POL2, CEBPB, GR, TCF12, YY1, TAF1, SIN3AK20, E2F6, PML, MTA3, POU2F2
(K) HTLV-I infection	POL2, CEBPB, GR, TCF12, YY1, TAF1, HEY1, CREB1, SP1, SIN3AK20, EBF, E2F6, ELF1, PML, BCLAF1, ZNF143, CJUN, MTA3, RUNX3, STAT5A, POU2F2, NFATC1, CBX3, KAP1
(K) TNF signaling pathway	POL2, CEBPB, PAX5, GR, TCF12, YY1, E2F6, ELF1, PML, BCLAF1, MTA3, RUNX3, POU2F2
(K) Pathways in cancer	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, CREB1, SP1, SIN3AK20, E2F6, ELF1, PML, ZNF143, ATF1, CJUN, MTA3, RUNX3, STAT5A, POU2F2
(G) Anatomical structures morphogenesis	
(H) Genes up-regulated in response to alpha interferon proteins	POL2, CEBPB, GR, MTA3
(K) Cytokine-cytokine receptor interaction	POL2, CEBPB, GR, E2F6

Table 4.6: Case 1: Final results with the statistically significant gene sets and their likely regulators

The analysis links the Bmp6 (Bone Morphogenetic Protein 6) to 37 transcription factors. We could find evidence linking 6 of these 37 transcription factors to the Bmp6 protein. Many of the remaining 31 predictions could be novel and could be discovered to be true in the future.

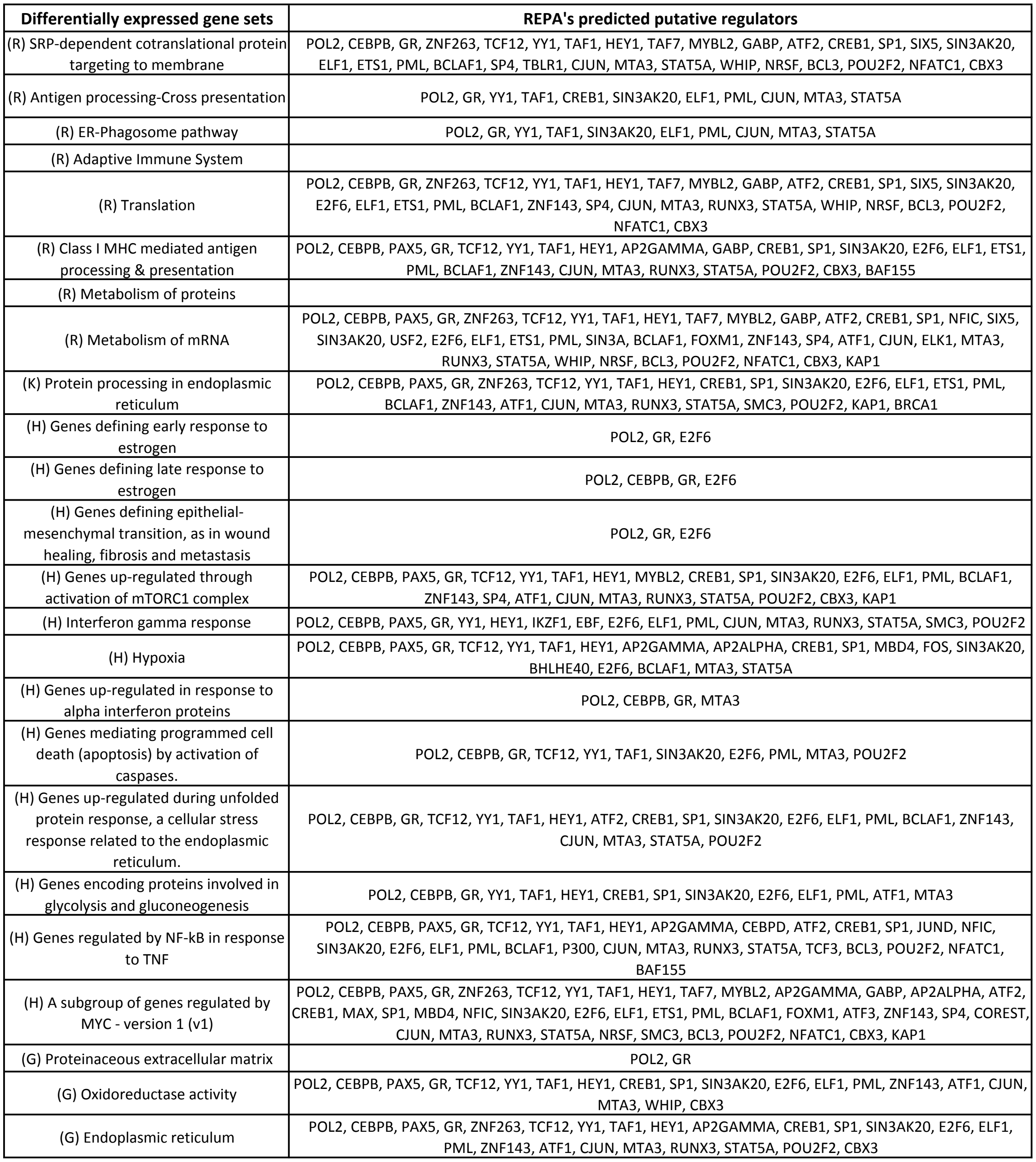
AP2GAMMA, ATF1, ATF2, BAF155, BCL3, BCLAF1, CBX3, CEBPB (Lin et al. 2013), CEBPD, CJUN, CREB1, E2F6, EBF, ELF1, GR (Kano et al. 2005), HEY1 (Sivertsen et al. 2007), IKZF1, JUND, KAP1, MTA3, NFATC1, NFIC, P300, PAX5, PML (Topić et al. 2013), POL2, POU2F2, RUNX3, SIN3AK20, SMC3, SP1 (Zhang et al. 2005a), STAT5A, TAF1, TCF12, TCF3, YY1 (Lee et al. 2004) and ZNF143.

4.3.2 Case 2: Dataset derived from breast cancer study on 12 patients

Out of the total 10,192 gene sets, REPA listed 543 as perturbed. 24 out of these 543 gene sets come from sources such as KEGG, Biocarta, GO, Reactome or MSigDB's Hallmark.

The results from module 1 shows that REPA also predicted likely regulators for 22 out of these 24 gene sets. Figure 4.11 provides a full list of these 24 gene sets and the associated transcription factors.

This list provides clues about likely regulatory mechanisms underlying the observed gene expression changes. REPA's predictions associated 59 transcription factors with these 24 gene sets identified as differentially expressed in the breast cancer expression profiling study. Out of these 59 TFs, 45 (or 76.3%) have previously been directly linked to breast cancer; namely, AP2ALPHA (McPherson, Woodfield, and Weigel 2007) AP2GAMMA (McPherson, Woodfield, and Weigel 2007) ATF1 (Jones et al. 2012) ATF2 (Lau and Ronai 2012) ATF3 (Yin et al. 2010) BAF155 (Wang et al. 2014a) BCL3 (Choi et al. 2010) BCLAF1 (Savage et al. 2014) BHLHE40 (Wu et al. 2014; Cadenas et al. 2014) BRCA1 (Wang and Di 2014) CBX3 (Choi, Park, and Lee 2012) CEBPB (Abreu and Sealy 2010) CJUN (Xu et al. 2013) COREST (Vicent et al. 2013) CREB1 (Phuong et al. 2014) E2F6 (Oberley, Inman, and Farnham 2003) EBF (Geng et al. 2007) ELF1 (Gerloff et al. 2011) ELK1 (Laliotis et al. 2013)



Differentially expressed gene sets	REPA's predicted putative regulators
(R) SRP-dependent cotranslational protein targeting to membrane	POL2, CEBPB, GR, ZNF263, TCF12, YY1, TAF1, HEY1, TAF7, MYBL2, GABP, ATF2, CREB1, SP1, SIX5, SIN3AK20, ELF1, ETS1, PML, BCLAF1, SP4, TBLR1, CJUN, MTA3, STAT5A, WHIP, NRSF, BCL3, POU2F2, NFATC1, CBX3
(R) Antigen processing-Cross presentation	POL2, GR, YY1, TAF1, CREB1, SIN3AK20, ELF1, PML, CJUN, MTA3, STAT5A
(R) ER-Phagosome pathway	POL2, GR, YY1, TAF1, SIN3AK20, ELF1, PML, CJUN, MTA3, STAT5A
(R) Adaptive Immune System	
(R) Translation	POL2, CEBPB, GR, ZNF263, TCF12, YY1, TAF1, HEY1, TAF7, MYBL2, GABP, ATF2, CREB1, SP1, SIX5, SIN3AK20, E2F6, ELF1, ETS1, PML, BCLAF1, ZNF143, SP4, CJUN, MTA3, RUNX3, STAT5A, WHIP, NRSF, BCL3, POU2F2, NFATC1, CBX3
(R) Class I MHC mediated antigen processing & presentation	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, AP2GAMMA, GABP, CREB1, SP1, SIN3AK20, E2F6, ELF1, ETS1, PML, BCLAF1, ZNF143, CJUN, MTA3, RUNX3, STAT5A, POU2F2, CBX3, BAF155
(R) Metabolism of proteins	
(R) Metabolism of mRNA	POL2, CEBPB, PAX5, GR, ZNF263, TCF12, YY1, TAF1, HEY1, TAF7, MYBL2, GABP, ATF2, CREB1, SP1, NFIC, SIX5, SIN3AK20, USF2, E2F6, ELF1, ETS1, PML, SIN3A, BCLAF1, FOXM1, ZNF143, SP4, ATF1, CJUN, ELK1, MTA3, RUNX3, STAT5A, WHIP, NRSF, BCL3, POU2F2, NFATC1, CBX3, KAP1
(K) Protein processing in endoplasmic reticulum	POL2, CEBPB, PAX5, GR, ZNF263, TCF12, YY1, TAF1, HEY1, CREB1, SP1, SIN3AK20, E2F6, ELF1, ETS1, PML, BCLAF1, ZNF143, ATF1, CJUN, MTA3, RUNX3, STAT5A, SMC3, POU2F2, KAP1, BRCA1
(H) Genes defining early response to estrogen	POL2, GR, E2F6
(H) Genes defining late response to estrogen	POL2, CEBPB, GR, E2F6
(H) Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis	POL2, GR, E2F6
(H) Genes up-regulated through activation of mTORC1 complex	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, MYBL2, CREB1, SP1, SIN3AK20, E2F6, ELF1, PML, BCLAF1, ZNF143, SP4, ATF1, CJUN, MTA3, RUNX3, STAT5A, POU2F2, CBX3, KAP1
(H) Interferon gamma response	POL2, CEBPB, PAX5, GR, YY1, HEY1, IKZF1, EBF, E2F6, ELF1, PML, CJUN, MTA3, RUNX3, STAT5A, SMC3, POU2F2
(H) Hypoxia	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, AP2GAMMA, AP2ALPHA, CREB1, SP1, MBD4, FOS, SIN3AK20, BHLHE40, E2F6, BCLAF1, MTA3, STAT5A
(H) Genes up-regulated in response to alpha interferon proteins	POL2, CEBPB, GR, MTA3
(H) Genes mediating programmed cell death (apoptosis) by activation of caspases.	POL2, CEBPB, GR, TCF12, YY1, TAF1, SIN3AK20, E2F6, PML, MTA3, POU2F2
(H) Genes up-regulated during unfolded protein response, a cellular stress response related to the endoplasmic reticulum.	POL2, CEBPB, GR, TCF12, YY1, TAF1, HEY1, ATF2, CREB1, SP1, SIN3AK20, E2F6, ELF1, PML, BCLAF1, ZNF143, CJUN, MTA3, STAT5A, POU2F2
(H) Genes encoding proteins involved in glycolysis and gluconeogenesis	POL2, CEBPB, GR, YY1, TAF1, HEY1, CREB1, SP1, SIN3AK20, E2F6, ELF1, PML, ATF1, MTA3
(H) Genes regulated by NF-kB in response to TNF	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, AP2GAMMA, CEBPD, ATF2, CREB1, SP1, JUND, NFIC, SIN3AK20, E2F6, ELF1, PML, BCLAF1, P300, CJUN, MTA3, RUNX3, STAT5A, TCF3, BCL3, POU2F2, NFATC1, BAF155
(H) A subgroup of genes regulated by MYC - version 1 (v1)	POL2, CEBPB, PAX5, GR, ZNF263, TCF12, YY1, TAF1, HEY1, TAF7, MYBL2, AP2GAMMA, GABP, AP2ALPHA, ATF2, CREB1, MAX, SP1, MBD4, NFIC, SIN3AK20, E2F6, ELF1, ETS1, PML, BCLAF1, FOXM1, ATF3, ZNF143, SP4, COREST, CJUN, MTA3, RUNX3, STAT5A, NRSF, SMC3, BCL3, POU2F2, NFATC1, CBX3, KAP1
(G) Proteinaceous extracellular matrix	POL2, GR
(G) Oxidoreductase activity	POL2, CEBPB, PAX5, GR, TCF12, YY1, TAF1, HEY1, CREB1, SP1, SIN3AK20, E2F6, ELF1, PML, ZNF143, ATF1, CJUN, MTA3, WHIP, CBX3
(G) Endoplasmic reticulum	POL2, CEBPB, PAX5, GR, ZNF263, TCF12, YY1, TAF1, HEY1, AP2GAMMA, CREB1, SP1, SIN3AK20, E2F6, ELF1, PML, ZNF143, ATF1, CJUN, MTA3, RUNX3, STAT5A, POU2F2, CBX3

Figure 4.11: Breast cancer case study results

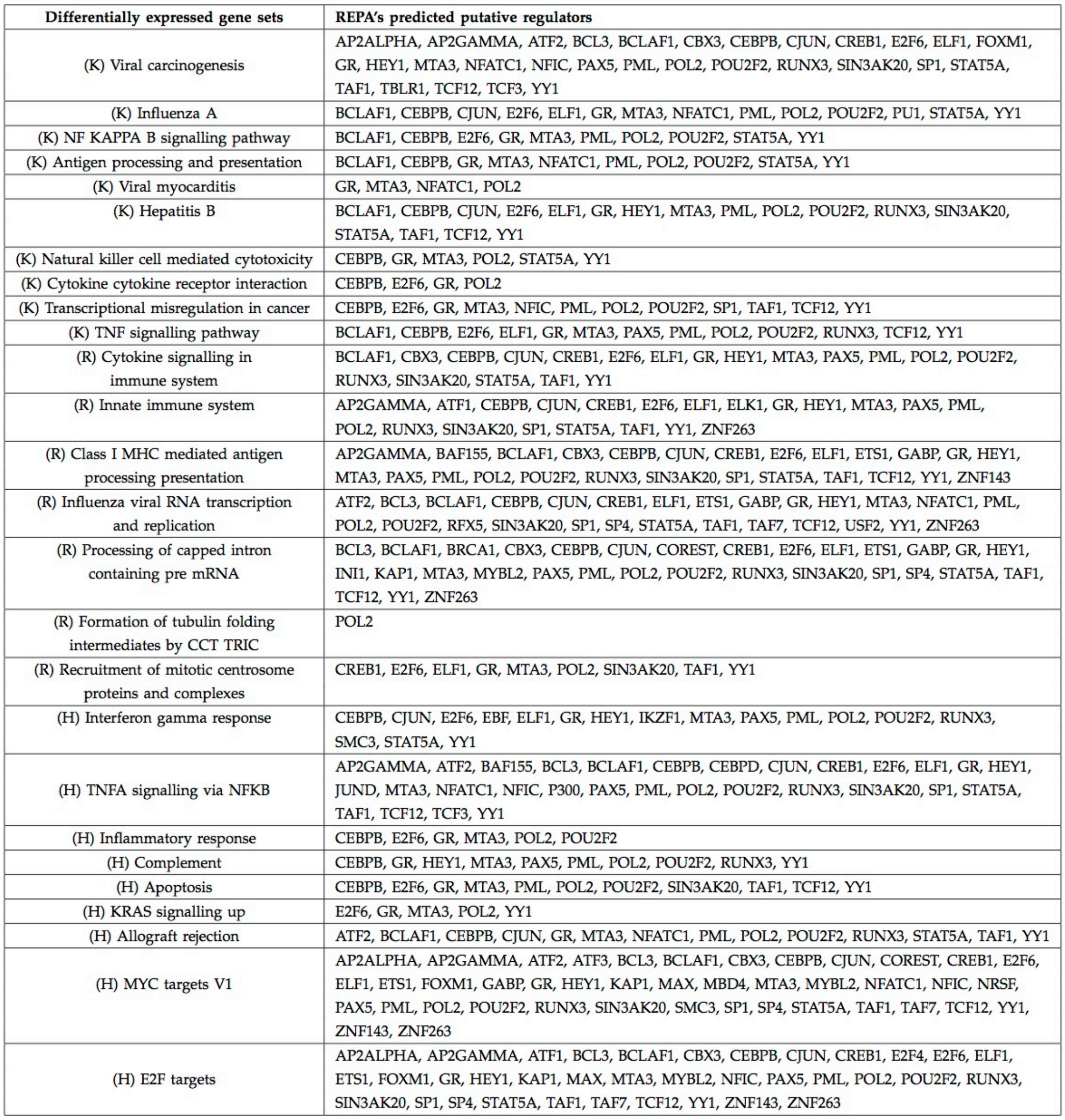
ETS1 (Furlan et al. 2014) FOXM1 (Koo, Muir, and Lam 2012) GABP (Thompson, MacDonald, and Mueller 2011) GR (Vilasco et al. 2011) HEY1 (Bolós et al. 2013) IKZF1 (Yang, Luo, and Wei 2010) KAP1 (Addison et al. 2015) MAX (Iliopoulos, Rotem, and Struhl 2011) MBD4 (Cunha et al. 2014) MTA3 (Fujita et al. 2003) MYBL2 (Shi et al. 2012) NFATC1 (Yiu et al. 2011) NFIC (Eeckhoute et al. 2006) NRSF (Bronson et al. 2010) PAX5 (Moelans, Verschuur-Maes, and Diest 2011) PML (Carracedo et al. 2012) POL2 (Han et al. 2014) RUNX3 (Bai et al. 2013) SMC3 (Wernicke et al. 2011) SP1 (Kong et al. 2014) SP4 (Wu et al. 2009) STAT5A (Zeng et al. 2014) TAF1 (McDonnell et al. 1995) TBLR1 (Li et al. 2014) TCF12 (Lee et al. 2012a) and YY1 (Lieberthal et al. 2009; Wan et al. 2012).

These results indicate that REPA's precision may be higher than the one suggested by our literature-based evaluation done in section 4.1.4. Moreover, REPA's predictions suggested 14 additional transcription factors that may play a role in breast cancer. These are CEBPD, FOS, JUND, P300, POU2F2, SIN3A, SIN3AK20, SIX5, TAF7, TCF3, USF2, WHIP, ZNF143 and ZNF263. Some of these 14 additional regulators have already been implicated in other types of cancer.

4.3.3 Case 3: Infection of influenza A viruses

In the previous section, we discussed that when we used REPA's enrichment analysis module to analyze this data set, we got a list of 124 perturbed gene sets. GAGE, on the other hand, predicted 833 gene sets to be active. Since GAGE's list was more comprehensive and includes 119 out of 124 of REPA's predictions, we substituted the results of module 2 with GAGE's results.

To reduce the amount of redundancy in GAGE results, we compared every pair of differentially expressed gene sets. We obtained the number of genes in common between each pair and removed those gene sets with a significant overlap with another gene set (p-value < 0.0001 using the hypergeometric distribution) and at least 50% of their genes lying on the intersection between both gene sets. This filtering reduced the number of gene sets by 70%; however, the number of associated TFs decreased by only 7%. This indicates that REPA's predictions are replicated in different annotation scheme.



Differentially expressed gene sets	REPA's predicted putative regulators
(K) Viral carcinogenesis	AP2ALPHA, AP2GAMMA, ATF2, BCL3, BCLAF1, CBX3, CEBPB, CJUN, CREB1, E2F6, ELF1, FOXM1, GR, HEY1, MTA3, NFATC1, NFIC, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, SP1, STAT5A, TAF1, TBLR1, TCF12, TCF3, YY1
(K) Influenza A	BCLAF1, CEBPB, CJUN, E2F6, ELF1, GR, MTA3, NFATC1, PML, POL2, POU2F2, PU1, STAT5A, YY1
(K) NF KAPPA B signalling pathway	BCLAF1, CEBPB, E2F6, GR, MTA3, PML, POL2, POU2F2, STAT5A, YY1
(K) Antigen processing and presentation	BCLAF1, CEBPB, GR, MTA3, NFATC1, PML, POL2, POU2F2, STAT5A, YY1
(K) Viral myocarditis	GR, MTA3, NFATC1, POL2
(K) Hepatitis B	BCLAF1, CEBPB, CJUN, E2F6, ELF1, GR, HEY1, MTA3, PML, POL2, POU2F2, RUNX3, SIN3AK20, STAT5A, TAF1, TCF12, YY1
(K) Natural killer cell mediated cytotoxicity	CEBPB, GR, MTA3, POL2, STAT5A, YY1
(K) Cytokine cytokine receptor interaction	CEBPB, E2F6, GR, POL2
(K) Transcriptional misregulation in cancer	CEBPB, E2F6, GR, MTA3, NFIC, PML, POL2, POU2F2, SP1, TAF1, TCF12, YY1
(K) TNF signalling pathway	BCLAF1, CEBPB, E2F6, ELF1, GR, MTA3, PAX5, PML, POL2, POU2F2, RUNX3, TCF12, YY1
(R) Cytokine signalling in immune system	BCLAF1, CBX3, CEBPB, CJUN, CREB1, E2F6, ELF1, GR, HEY1, MTA3, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, STAT5A, TAF1, YY1
(R) Innate immune system	AP2GAMMA, ATF1, CEBPB, CJUN, CREB1, E2F6, ELF1, ELK1, GR, HEY1, MTA3, PAX5, PML, POL2, RUNX3, SIN3AK20, SP1, STAT5A, TAF1, YY1, ZNF263
(R) Class I MHC mediated antigen processing presentation	AP2GAMMA, BAF155, BCLAF1, CBX3, CEBPB, CJUN, CREB1, E2F6, ELF1, ETS1, GABP, GR, HEY1, MTA3, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, SP1, STAT5A, TAF1, TCF12, YY1, ZNF143
(R) Influenza viral RNA transcription and replication	ATF2, BCL3, BCLAF1, CEBPB, CJUN, CREB1, ELF1, ETS1, GABP, GR, HEY1, MTA3, NFATC1, PML, POL2, POU2F2, RFX5, SIN3AK20, SP1, SP4, STAT5A, TAF1, TAF7, TCF12, USF2, YY1, ZNF263
(R) Processing of capped intron containing pre mRNA	BCL3, BCLAF1, BRCA1, CBX3, CEBPB, CJUN, COREST, CREB1, E2F6, ELF1, ETS1, GABP, GR, HEY1, INI1, KAP1, MTA3, MYBL2, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, SP1, SP4, STAT5A, TAF1, TCF12, YY1, ZNF263
(R) Formation of tubulin folding intermediates by CCT TRIC	POL2
(R) Recruitment of mitotic centrosome proteins and complexes	CREB1, E2F6, ELF1, GR, MTA3, POL2, SIN3AK20, TAF1, YY1
(H) Interferon gamma response	CEBPB, CJUN, E2F6, EBF, ELF1, GR, HEY1, IKZF1, MTA3, PAX5, PML, POL2, POU2F2, RUNX3, SMC3, STAT5A, YY1
(H) TNFA signalling via NFKB	AP2GAMMA, ATF2, BAF155, BCL3, BCLAF1, CEBPB, CEBPD, CJUN, CREB1, E2F6, ELF1, GR, HEY1, JUND, MTA3, NFATC1, NFIC, P300, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, SP1, STAT5A, TAF1, TCF12, TCF3, YY1
(H) Inflammatory response	CEBPB, E2F6, GR, MTA3, POL2, POU2F2
(H) Complement	CEBPB, GR, HEY1, MTA3, PAX5, PML, POL2, POU2F2, RUNX3, YY1
(H) Apoptosis	CEBPB, E2F6, GR, MTA3, PML, POL2, POU2F2, SIN3AK20, TAF1, TCF12, YY1
(H) KRAS signalling up	E2F6, GR, MTA3, POL2, YY1
(H) Allograft rejection	ATF2, BCLAF1, CEBPB, CJUN, GR, MTA3, NFATC1, PML, POL2, POU2F2, RUNX3, STAT5A, TAF1, YY1
(H) MYC targets V1	AP2ALPHA, AP2GAMMA, ATF2, ATF3, BCL3, BCLAF1, CBX3, CEBPB, CJUN, COREST, CREB1, E2F6, ELF1, ETS1, FOXM1, GABP, GR, HEY1, KAP1, MAX, MBD4, MTA3, MYBL2, NFATC1, NFIC, NRSF, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, SMC3, SP1, SP4, STAT5A, TAF1, TAF7, TCF12, YY1, ZNF143, ZNF263
(H) E2F targets	AP2ALPHA, AP2GAMMA, ATF1, BCL3, BCLAF1, CBX3, CEBPB, CJUN, CREB1, E2F4, E2F6, ELF1, ETS1, FOXM1, GR, HEY1, KAP1, MAX, MTA3, MYBL2, NFIC, PAX5, PML, POL2, POU2F2, RUNX3, SIN3AK20, SP1, SP4, STAT5A, TAF1, TAF7, TCF12, YY1, ZNF143, ZNF263

Figure 4.12: Infection of influenza A viruses case study results

In the original study, pathway analyses were performed using Ingenuity Pathway Analysis (IPA, Ingenuity Systems) software. As in the original study, we identified gene sets related to cytokine signalling, interferon signalling, apoptosis, complement system, and antigen presentation. In addition, we identified several influenza-related gene sets (see figure 4.12).

Upon identification of the differentially expressed gene sets, we used REPA's predictions to list putative regulators of those differentially expressed gene sets (see figure 4.12). REPA's predictions associated 58 TFs with the gene sets identified as differentially expressed in the influenza infection transcriptional profiling study. These 58 TFs are the following (supporting literature is referred to after the corresponding TF): AP2ALPHA, AP2GAMMA, ATF1, ATF2 (Hrincius et al. 2010), ATF3 (Whitmore et al. 2007), BAF155, BCL3, BCLAF1, BRCA1, CBX3, CEBPB (Zhu et al. 2010), CEBPD, CJUN (Cannon et al. 2014), COREST, CREB1 (Liu et al. 2012), E2F4 (Zhu et al. 2010), E2F6 (Zhu et al. 2010), EBF, ELF1, ELK1 (Harii et al. 2005), ETS1, FOXM1, GABP, GR (Ge et al. 2011), HEY1, IKZF1, INI1, JUND, KAP1, MAX, MBD4, MTA3, MYBL2, NFATC1 (Zhang et al. 2009), NFIC, NRSF, P300, PAX5 (Savitsky and Calame 2006), PML (Li et al. 2009), POL2, POU2F2 (Bussfeld et al. 1997), PU1, RFX5, RUNX3, SIN3AK20, SMC3, SP1 (Barbier et al. 2012), SP4, STAT5A, TAF1, TAF7, TBLR1, TCF12, TCF3, USF2, YY1, ZNF143, and ZNF263. We found literature linking with influenza infection 14 (or 24%) of these TFs. Additionally, ATF3 and SP1 were found to be expressed in cells infected with pandemic influenza A virus (H1N1pdm) but not in cells infected with seasonal influenza virus (Gerlach et al. 2013a).

4.4 Summary

In this chapter, we present the results of evaluating REPA. Our results suggested that between 24% and 76% of the regulatory associations predicted by REPA are likely correct, and that REPA's recall is around 54%. In addition, REPA's module 2 results are mostly in agreement with the gene sets obtained by GAGE, a widely used tool for gene set analysis of expression data. This suggests that REPA's novel predicted regulatory associations may indeed be useful to guide further biological investigations.

Chapter 5

Conclusion

To measure the activity levels of various genes, biologists often use gene expression profiling experiments. After obtaining the activity levels of all the genes and analyzing the data from such experiments, researchers then resort to enrichment analysis techniques. Enrichment analysis is usually the next step after performing a gene expression profiling experiment because it helps researchers gain a better understanding of the cellular activities and processes relevant to the study performed. Transcription factors are regulator proteins that control the level of activity of various genes. A transcription factor usually regulates a gene by binding to its promoter region. Transcription factor binding (TFB) data for the entire human genome was made publicly available under the ENCODE project. Even though there is an important relationship between transcription factors and biological pathways, so far enrichment analysis techniques have not looked at those connections.

In this thesis we developed a novel method of analyzing TFB data and combining it with gene set enrichment analysis. An article describing REPA has been accepted for publication in the IEEE/ACM Transactions on Computational Biology and Bioinformatics journal. We also built a software application, REPA (Regulation

Enrichment Pathway Analysis), as an implementation of this approach. We explored the appropriate statistical testing methods for our purpose and evaluated the results of this new approach.

The greatest advantage of using REPA is that it allows the user to see a more complete picture of the cellular activity by providing information about which transcription factors may regulate the genes being affected in an expression profiling study. Our hope is that by informing researchers about likely regulators, they might be able to identify future research paths that could have been overlooked.

We can further improve the accuracy and scope of REPA by looking at certain aspects of the system. By adding TFB data for other species we can use the same system to analyze other species. We can also use orthology relationships to transfer information between species. We also want to test the system with biological data from ongoing wet lab experiments. Finally, having the system as a web service will make the tool accessible to a larger research community.

We found literature backing several of the predictions relating transcription factors to gene sets, and we found that there was a significant overlap between the results of GAGE and the enrichment analysis part of REPA. Even though the results from these tests appear to be promising, the final test would be in the hands of the researchers in terms of whether or not they find REPA useful and continue using it as their preferred enrichment analysis technique.

One of the challenges we faced while doing this project is the lack of transcription factor binding data for all the known human transcription factors. There are over 1,391 known sequence-specific DNA-binding human transcription factors (Vaquerizas et al. 2009) however, in this project, we have information of only 120 transcription factors. The ENCODE project was the first effort to undertake this massive task of functionally annotating human DNA, but with the advent of new DNA sequencing techniques more and more such data can be generated for a much lower cost (Resnick 2011). As these technologies become less expensive and start producing more data, our approach will have the information it needs to become more accurate. Therefore it is fair to say that the biggest significance of this project is that it takes enrichment analysis in a new direction where it will be possible to provide information on not only which gene sets are interesting, but also how they are regulated.

Appendix A

R script used for comparing REPA with GAGE.

A.1 Dataset 1: 8 hours BMP6 treated vs untreated human mesenchymal stem cells

```
# 13604 Comparison Script

# To install the gage package, start R and enter:
source(”http://bioconductor.org/biocLite.R”)
biocLite(”gage”)

#To install the auxillary data for gage package:
biocLite(”gageData”)

# Load Libraries
```

```
library(gageData)
library(gage)


# Set working dir
setwd("GAGE_Comparison")


# Run gage and get results of analysis
c2.gs=readList("c2_c5_c6_hallmark_kegg_12April2015.txt")
data(bmp6)
lapply(c2.gs[1:3],head)
head(rownames(bmp6))
bmp6.c2.p <- gage(bmp6, gsets = c2.gs, ref =c(1,3), samp = c
    (2,4), same.dir = FALSE, compare = "as.group")
results <- bmp6.c2.p$greater[complete.cases(bmp6.c2.p$greater
    ),]
gage.bmp6.sigResults <- results[results[,"q.val"] < 0.01,]

resultsAllSystems <- list()
resultsAllSystems$gageBMP6 <- row.names(gage.bmp6.sigResults)

write.table(file = "gage13604_results.txt", results, sep = "\
    t")

system <- read.table("13604_system_results.txt", sep = "\t",
    header = FALSE, stringsAsFactors = FALSE)
system$pvalue <- 1.001 - (system[,"V4"] + system[,"V5"])/1000
```

```
row.names(system) <- system[,"V1"]

system <- system[row.names(results),]

length(intersect(system[ system[,"pvalue"] < 0.01, "V1"], row
    .names(results[results[,"q.val"] < 0.01,])))

significantBoth <- union(system[ system[,"pvalue"] < 0.01, "
    V1"], row.names(results[results[,"q.val"] < 0.01,]))

# For all gene sets
cor(system[ row.names(results[results[,"q.val"] < 0.01,]), "
    pvalue"], results[results[,"q.val"] < 0.01, "q.val"],
    method = "spearman")
# For significant gene sets
cor(system[ significantBoth, "pvalue"], results[
    significantBoth, "q.val"], method = "spearman")

# Generate REPA vs GAGE plot.
pdf("Plot_13604_Comparison.pdf")
plot(-log10(system[ row.names(results), "pvalue"]), -log10(
    results[, "q.val"]),  main = "", ylab = "-log10␣GAGE␣
    pvalues", xlab = "-log10␣REPA␣pvalues")
lines(lowess(x= -log10(system[ row.names(results),"pvalue"])
    , y = -log10(results[, "q.val"])), col = "red", lwd = 2)
```

```
dev.off()


# Generate venn diagram
library(gplots)
pdf("Venn_13604_Comparison.pdf")
venn(list(GAGE= row.names(results[results[,"q.val"] < 0.01,
    ]), REPA = system[system[, "pvalue"] < 0.01, 1] ))
dev.off()
```

A.2 Dataset 2: Derived from breast cancer study on 12 patients

```
# To install the gage package, start R and enter:
source("http://bioconductor.org/biocLite.R")
biocLite("gage")


#To install the auxillary data for gage package:
biocLite("gageData")


# load the gage and gageData libraries:
library(gageData)
library(gage)


# load the gene sets
c2.gs <- readList("gene_sets.txt")
```

```
#item Run GAGE with the following arguments.
data(gse16873)
lapply(c2.gs[1:3],head)
head(rownames(gse16873))
gse16873.c2.p <- gage(gse16873, gsets = c2.gs, ref =c(1,3,5),
    samp = c(2,4,6), same.dir = FALSE, compare = ”as.group”)
results <- gse16873.c2.p$greater[complete.cases(gse16873.c2.p
   $greater),]
gage.gse16873.sigResults <- results[results[,”q.val”] <
   0.01,]
write.table(file = ”gage_results_16873.txt”, results, sep = ”
   \t”)

resultsAllSystems <- list()
resultsAllSystems$gageGSE16873 <- row.names(gage.gse16873.
   sigResults)
write.table(file = ”gage_results_16873.txt”, results, sep = ”
   \t”)

system <- read.table(”GSE16873_Repa_Results.txt”, sep = ”\t”,
    header = FALSE, stringsAsFactors = FALSE)
system <- system[, c(1:5)]
system$pvalue <- 1.001 - (system[,”V4”] + system[,”V5”])/1000
row.names(system) <- system[,”V1”]
system <- system[row.names(results),]
```

```
length(intersect(system[ system[,”pvalue”] < 0.01, ”V1”], row
    .names(results[results[,”q.val”] < 0.01,])))

significantBoth <− union(system[ system[,”pvalue”] < 0.01, ”
    V1”], row.names(results[results[,”q.val”] < 0.01,]))
# For all gene sets
cor(system[ row.names(results[results[,”q.val”] < 0.01,]), ”
    pvalue”], results[results[,”q.val”] < 0.01, ”q.val”],
    method = ”spearman”)
# For significant gene sets
cor(system[ significantBoth, ”pvalue”], results[
    significantBoth, ”q.val”], method = ”spearman”)

# Generate REPA vs GAGE plot.
pdf(”Plot_16873_Comparison.pdf”)
plot(−log10(system[ row.names(results), ”pvalue”]), −log10(
    results[, ”q.val”]),  main = ””, ylab = ”−log10␣GAGE␣
    pvalues”, xlab = ”−log10␣REPA␣pvalues”)
lines(lowess(x= −log10(system[ row.names(results),”pvalue”])
    , y = −log10(results[, ”q.val”])), col = ”red”, lwd = 2)
dev.off()

# Generate venn diagram
library(gplots)
pdf(”Venn_16873_Comparison.pdf”)
```

```
venn(list(GAGE= row.names(results[results[,"q.val"] < 0.01,
    ]), REPA = system[system[, "pvalue"] < 0.01, 1] ))
dev.off()

write.table(file = "gage_sig_res.txt", results[results[,"q.
    val"] < 0.01,], sep = "\t")
write.table(file = "repa_sig_res.txt", system[ system[,"
    pvalue"] < 0.01, "V1"], sep = "\t")
```

A.3 Dataset 3: Infection of influenza A viruses

```
# To install the gage package, start R and enter:
source("http://bioconductor.org/biocLite.R")
biocLite("gage")

#To install the auxillary data for gage package:
biocLite("gageData")

# Load Libraries
library(gageData)
library(gage)

# Set working dir
setwd("GAGE_Comparison")

### Code for Influenza A
```

```
msigv5_kegg.gs <- readList("msig5_kegg.txt")
length(msigv5_kegg.gs)

#Early host responses of seasonal and pandemic influenza A
   viruses in primary well-differentiated human lung
   epithelial cells.
#PMID: 24244384 Gerlach2013

# load series and platform data from GEO
library(Biobase)
library(GEOquery)

gse48466_raw <- getGEO("GSE48466", GSEMatrix =TRUE)
if (length(gse48466_raw) > 1) idx <- grep("GPL570", attr(
   gse48466_raw, "names")) else idx <- 1
gse48466_raw <- gse48466_raw[[idx]]

# load NCBI platform annotation
gpl <- annotation(gse48466_raw)
platf <- getGEO(gpl, AnnotGPL=TRUE)
ncbifd <- data.frame(attr(dataTable(platf), "table"))

probe_entrez <- ncbifd[, c("ID", "Gene.ID")]
probe_entrez <- probe_entrez[probe_entrez$Gene.ID != "", ]
probe_entrez <- probe_entrez[!grepl("///", probe_entrez$Gene.
   ID, fixed = TRUE) ,] # filter promiscuous probes
```

```
row.names(probe_entrez) <- probe_entrez[,1]
head(probe_entrez)

gse48466 <- exprs(gse48466_raw)
head(gse48466)

length(intersect(row.names(gse48466), probe_entrez[,”ID”]))

#Get average intensity per gene Entrez ID
gse48466_perGene <-  apply(gse48466, 2, function(c, f) {
  tapply(c,f,mean)
},  probe_entrez[row.names(gse48466),2]   )

gse48466_perGene <- log2(gse48466_perGene)

gage_results_influenza <- gage(gse48466_perGene, gsets =
    msigv5_kegg.gs, ref = 10:12, samp =4:6, same.dir = TRUE,
    compare = ”unpaired”)



results <- gage_results_influenza$greater[complete.cases(gage
    _results_influenza$greater),]
gage.influenza.sigResults <- results[results[,”q.val”] <
    0.01,]

resultsAllSystems <- list()
```

```
resultsAllSystems$gageinfluenza <- row.names(gage.influenza.
    sigResults)
write.table(file = "gage_results_influenza.txt", results, sep
    = "\t")

system <- read.table("repa_results_influenza.txt", sep = "\t"
    , header = FALSE, stringsAsFactors = FALSE)
#system <- read.table("repa_results_13604.txt", sep = "\t",
    header = FALSE, stringsAsFactors = FALSE)
system <- system[, c(1:5)]
system$pvalue <- 1.001 - (system[,"V4"] + system[,"V5"])/1000
row.names(system) <- system[,"V1"]
system <- system[row.names(results),]

length(intersect(system[ system[,"pvalue"] < 0.01, "V1"], row
    .names(results[results[,"q.val"] < 0.01,])))

significantBoth <- union(system[ system[,"pvalue"] < 0.01, "
    V1"], row.names(results[results[,"q.val"] < 0.01,]))
# For all gene sets
cor(system[ row.names(results[results[,"q.val"] < 0.01,]), "
    pvalue"], results[results[,"q.val"] < 0.01, "q.val"],
    method = "spearman")
# For significant gene sets
cor(system[ significantBoth, "pvalue"], results[
    significantBoth, "q.val"], method = "spearman")
```

```
# Generate REPA vs GAGE plot.
pdf("Plot_influenza_Comparison.pdf")
plot(-log10(system[ row.names(results), "pvalue"]), -log10(
    results[, "q.val"]),  main = "", ylab = "-log10 GAGE 
    pvalues", xlab = "-log10 REPA pvalues")
lines(lowess(x= -log10(system[ row.names(results),"pvalue"])
    , y = -log10(results[, "q.val"])), col = "red", lwd = 2)
dev.off()

# Generate venn diagram
library(gplots)
pdf("Venn_influenza_Comparison.pdf")
venn(list(GAGE= row.names(results[results[,"q.val"] < 0.01,
    ]), REPA = system[system[, "pvalue"] < 0.01, 1] ))
dev.off()


# 13606 Comparison Script

# Load Libraries
library(gageData)
library(gage)
```

```
# Set working dir
setwd(”GAGE_Comparison”)

# Run gage and get results of analysis
c2.gs=readList(”c2.all.v4.0.entrez.gmt”)
data(bmp6)
lapply(c2.gs[1:3],head)
head(rownames(bmp6))
bmp6.c2.p <- gage(bmp6, gsets = c2.gs, ref =c(1,3), samp = c
    (2,4), same.dir = FALSE, compare = ”as.group”)
results <- bmp6.c2.p$greater[complete.cases(bmp6.c2.p$greater
    ),]
gage.bmp6.sigResults <- results[results[,”q.val”] < 0.01,]

resultsAllSystems <- list()
resultsAllSystems$gageBMP6 <- row.names(gage.bmp6.sigResults)

write.table(file = ”gage13606_results.txt”, results, sep = ”\
    t”)

system <- read.table(”13606_system_results.txt”, sep = ”\t”,
    header = FALSE, stringsAsFactors = FALSE)
system$pvalue <- 1.001 - (system[,”V4”] + system[,”V5”])/1000

row.names(system) <- system[,”V1”]
```

```
system <- system[row.names(results),]

length(intersect(system[ system[,"pvalue"] < 0.01, "V1"], row
    .names(results[results[,"q.val"] < 0.01,])))

significantBoth <- union(system[ system[,"pvalue"] < 0.01, "
    V1"], row.names(results[results[,"q.val"] < 0.01,]))

# For all gene sets
cor(system[ row.names(results[results[,"q.val"] < 0.01,]), "
    pvalue"], results[results[,"q.val"] < 0.01, "q.val"],
    method = "spearman")
# For significant gene sets
cor(system[ significantBoth, "pvalue"], results[
    significantBoth, "q.val"], method = "spearman")

# Generate REPA vs GAGE plot.
pdf("Plot_13606_Comparison.pdf")
plot(-log10(system[ row.names(results), "pvalue"]), -log10(
    results[, "q.val"]),  main = "", ylab = "-log10␣GAGE␣
    pvalues", xlab = "-log10␣REPA␣pvalues")
lines(lowess(x= -log10(system[ row.names(results),"pvalue"])
    , y = -log10(results[, "q.val"])), col = "red", lwd = 2)
dev.off()
```

```
# Generate venn diagram
library(gplots)
pdf(”Venn_16873_Comparison.pdf”)
venn(list(GAGE= row.names(results[results[,”q.val”] < 0.01,
    ]), REPA = system[system[, ”pvalue”] < 0.01, 1] ))
dev.off()
```

Bibliography

Abreu, Maria M and Linda Sealy (2010). "The C/EBPbeta isoform, liver-inhibitory protein (LIP), induces autophagy in breast cancer cell lines". In: *Exp Cell Res* 316.19, pp. 3227–38. DOI: 10.1016/j.yexcr.2010.07.021.

Adachi, N. et al. (2014). "New insight in expression, transport, and secretion of brain-derived neurotrophic factor: Implications in brain-related diseases". eng. In: *World journal of biological chemistry* 5.4, pp. 409–428.

Addison, Joseph B et al. (2015). "KAP1 promotes proliferation and metastatic progression of breast cancer cells". In: *Cancer Res* 75.2, pp. 344–55. DOI: 10.1158/0008-5472.CAN-14-1561.

Algire, Mikkel A et al. (2002). "Development and characterization of a reconstituted yeast translation initiation system". In: *RNA* 8.3, pp. 382–97.

Amati, B and H Land (1994). "Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death". In: *Curr Opin Genet Dev* 4.1, pp. 102–8.

Ariumi, Y. et al. (2003). "Distinct nuclear body components, PML and SMRT, regulate the trans-acting function of HTLV-1 Tax oncoprotein". eng. In: *Oncogene* 22.11, pp. 1611–1619.

Ashburner, Michael et al. (2000). "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1, pp. 25–29.

AzaToth (2008). *Myoglobin.* URL: `http://commons.wikimedia.org/wiki/File%3AMyoglobin.png`.

Bai, Jin et al. (2013). "RUNX3 is a prognostic marker and potential therapeutic target in human breast cancer". In: *J Cancer Res Clin Oncol* 139.11, pp. 1813–23. DOI: `10.1007/s00432-013-1498-x`.

Balada, E. et al. (2007). "Transcript overexpression of the MBD2 and MBD4 genes in CD4+ T cells from systemic lupus erythematosus patients". eng. In: *Journal of leukocyte biology* 81.6, pp. 1609–1616.

Bandyopadhyay, Sushobhana et al. (2014). "The basic leucine zipper domain transcription factor Atf1 directly controls Cdc13 expression and regulates mitotic entry independently of Wee1 and Cdc25 in Schizosaccharomyces pombe". In: *Eukaryot Cell* 13.6, pp. 813–21. DOI: `10.1128/EC.00059-14`.

Barbier, Diane et al. (2012). "Influenza A induces the major secreted airway mucin MUC5AC in a protease-EGFR-extracellular regulated kinase-Sp1-dependent pathway". In: *Am J Respir Cell Mol Biol* 47.2, pp. 149–57. DOI: `10.1165/rcmb.2011-0405OC`.

Bauer-Mehren, A. et al. (2011). "Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases". eng. In: *PloS one* 6.6, e20284.

Berg, Stefan et al. (2001). "Sequence Properties of the 1, 2-Diacylglycerol 3- Glucosyltransferase from Acholeplasma laidlawiiMembranes". In: *Journal of Biological Chemistry* 276.25, pp. 22056–22063.

Berger, Michael F and Martha L Bulyk (2009). "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors". In: *Nature protocols* 4.3, pp. 393–411.

Bochkanov, S and V Bystritsky. *ALGLIB project.*

Bolós, Victoria et al. (2013). “Notch activation stimulates migration of breast cancer cells and promotes tumor growth”. In: *Breast Cancer Res* 15.4, R54. DOI: 10.1186/bcr3447.

Broad, Institute (2014). *Transcription Factor*. https://www.broadinstitute.org/education/glossary/transcription-factor.

Bronson, Michael W et al. (2010). “Estrogen coordinates translation and transcription, revealing a role for NRSF in human breast cancer cells”. In: *Mol Endocrinol* 24.6, pp. 1120–35. DOI: 10.1210/me.2009-0436.

Bussfeld, D et al. (1997). “Expression of transcription factor genes after influenza A virus infection”. In: *Immunobiology* 198.1-3, pp. 291–8. DOI: 10.1016/S0171-2985(97)80049-6.

Cadenas, Cristina et al. (2014). “Loss of circadian clock gene expression is associated with tumor progression in breast cancer”. In: *Cell Cycle* 13.20, pp. 3282–91. DOI: 10.4161/15384101.2014.954454.

Cai, Xinming et al. (2014). “PIKfyve, a class III lipid kinase, is required for TLR-induced type I IFN production via modulation of ATF3”. In: *J Immunol* 192.7, pp. 3383–9. DOI: 10.4049/jimmunol.1302411.

Cannon, Georgetta et al. (2014). “Early activation of MAP kinases by influenza A virus X-31 in murine macrophage cell lines”. In: *PLoS One* 9.8, e105385. DOI: 10.1371/journal.pone.0105385.

Carracedo, Arkaitz et al. (2012). “A metabolic prosurvival role for PML in breast cancer”. In: *J Clin Invest* 122.9, pp. 3088–100. DOI: 10.1172/JCI62129.

Chauhan, R et al. (2004). “Over-expression of TATA binding protein (TBP) and p53 and autoantibodies to these antigens are features of systemic sclerosis, systemic lupus erythematosus and overlap syndromes”. In: *Clin Exp Immunol* 136.3, pp. 574–84. DOI: 10.1111/j.1365-2249.2004.02463.x.

Chen, Hongbo et al. (2015). "Hypermethylation of glucocorticoid receptor gene promoter results in glucocorticoid receptor gene low expression in peripheral blood mononuclear cells of patients with systemic lupus erythematosus". In: *Rheumatol Int.* DOI: 10.1007/s00296-015-3266-5.

Choi, Hee June et al. (2010). "Bcl3-dependent stabilization of CtBP1 is crucial for the inhibition of apoptosis and tumor progression in breast cancer". In: *Biochem Biophys Res Commun* 400.3, pp. 396–402. DOI: 10.1016/j.bbrc.2010.08.084.

Choi, Jae Duk, Mi Ae Park, and Jong-Soo Lee (2012). "Suppression and recovery of BRCA1-mediated transcription by HP1 via modulation of promoter occupancy". In: *Nucleic Acids Res* 40.22, pp. 11321–38. DOI: 10.1093/nar/gks947.

Cole, M. F. et al. (2008). "Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells". In: *Genes Dev.* 22.6, pp. 746–755.

Consortium, ENCODE Project et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Crick, Francis et al. (1970). "Central dogma of molecular biology". In: *Nature* 227.5258, pp. 561–563.

Cunha, Stéphanie et al. (2014). "The RON receptor tyrosine kinase promotes metastasis by triggering MBD4-dependent DNA methylation reprogramming". In: *Cell Rep* 6.1, pp. 141–54. DOI: 10.1016/j.celrep.2013.12.010.

Demir, Ozlem and Isil Aksan Kurnaz (2013). "Phospho-Ser383-Elk-1 is localized to the mitotic spindles during cell cycle and interacts with mitotic kinase Aurora-A". In: *Cell Biochem Funct* 31.7, pp. 591–8. DOI: 10.1002/cbf.2944.

Deng, Zhiyong et al. (2010). "Yin Yang 1: a multifaceted protein beyond a transcription factor". In: *Transcription* 1.2, pp. 81–4. DOI: 10.4161/trns.1.2.12375.

Donadini, A. et al. (2006). “GABP complex regulates transcription of eIF6 (p27BBP), an essential trans-acting factor in ribosome biogenesis”. eng. In: *FEBS letters* 580.8, pp. 1983–1987.

Dong, X. et al. (2007). “Transcriptional activity of androgen receptor is modulated by two RNA splicing factors, PSF and p54nrb”. In: *Mol. Cell. Biol.* 27.13, pp. 4863–4875.

Dozmorov, Mikhail G, Jonathan D Wren, and Marta E Alarcón-Riquelme (2014). “Epigenomic elements enriched in the promoters of autoimmunity susceptibility genes”. In: *Epigenetics* 9.2, pp. 276–85. DOI: 10.4161/epi.27021.

Edgar, Ron, Michael Domrachev, and Alex E Lash (2002). “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository”. In: *Nucleic acids research* 30.1, pp. 207–210.

Eeckhoute, Jérôme et al. (2006). “A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer”. In: *Genes Dev* 20.18, pp. 2513–26. DOI: 10.1101/gad.1446006.

Emery, Lyndsey A et al. (2009). “Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression”. In: *The American journal of pathology* 175.3, pp. 1292–1302.

Evans, David M et al. (2011). “Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility”. In: *Nat Genet* 43.8, pp. 761–7. DOI: 10.1038/ng.873.

Forluvoft (2007). *Gene2-plain.* URL: http://commons.wikimedia.org/wiki/File:Gene2-plain.svg#mediaviewer/File:Gene2-plain.svg.

Frangou, E. A., G. K. Bertsias, and D. T. Boumpas (2013). “Gene expression and regulation in systemic lupus erythematosus”. eng. In: *European journal of clinical investigation* 43.10, pp. 1084–1096.

Fujita, Naoyuki et al. (2003). "MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer". In: *Cell* 113.2, pp. 207–19.

Furlan, Alessandro et al. (2014). "Ets-1 controls breast cancer cell balance between invasion and growth". In: *Int J Cancer* 135.10, pp. 2317–28. DOI: 10.1002/ijc.28881.

Ge, Xingyi et al. (2011). "Influenza virus infection induces the nuclear relocalization of the Hsp90 co-chaperone p23 and inhibits the glucocorticoid receptor response". In: *PLoS One* 6.8, e23368. DOI: 10.1371/journal.pone.0023368.

Geng, Li-Yi et al. (2007). "Expression of SNC73, a transcript of the immunoglobulin alpha-1 gene, in human epithelial carcinomas". In: *World J Gastroenterol* 13.16, pp. 2305–11.

Gerlach, Rachael L et al. (2013a). "Early host responses of seasonal and pandemic influenza A viruses in primary well-differentiated human lung epithelial cells". In: *PLoS One* 8.11, e78912. DOI: 10.1371/journal.pone.0078912.

— (2013b). "Early host responses of seasonal and pandemic influenza a viruses in primary well-differentiated human lung epithelial cells". In: *PloS one* 8.11, e78912.

Gerloff, Alice et al. (2011). "Protein expression of the Ets transcription factor Elf-1 in breast cancer cells is negatively correlated with histological grading, but not with clinical outcome". In: *Oncol Rep* 26.5, pp. 1121–5. DOI: 10.3892/or.2011.1409.

Glazko, Galina V and Frank Emmert-Streib (2009). "Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets". In: *Bioinformatics* 25.18, pp. 2348–2354.

Goeman, Jelle J et al. (2004). "A global test for groups of genes: testing association with a clinical outcome". In: *Bioinformatics* 20.1, pp. 93–99.

Haarman, E. G. et al. (2002). “In vitro glucocorticoid resistance in childhood leukemia correlates with receptor affinity determined at 37 degrees C, but not with affinity determined at room temperature”. eng. In: *Leukemia* 16.9, pp. 1882–1884.

Haase, Steven B and Curt Wittenberg (2014). “Topology and control of the cell-cycle-regulated transcriptional circuitry”. In: *Genetics* 196.1, pp. 65–90. DOI: 10.1534/genetics.113.152595.

Han, Z. et al. (2014). “Estrogen Induced RNA Polymerase II Stalling in Breast Cancer Cell Line MCF7”. In: *Bioinformatics Research and Applications: 10th International Symposium, ISBRA 2014, Zhangjiajie, China, June 28-30, 2014, Proceedings*. Ed. by M. Basu, Y. Pan, and J. Wang. Lecture Notes in Computer Science. Springer International Publishing. ISBN: 9783319081717.

Harada, S. et al. (1998). “A new genetic variant in the Sp1 binding cis-element of cholecystokinin gene promoter region and relationship to alcoholism”. eng. In: *Alcoholism, Clinical and Experimental Research* 22.3 Suppl, 93S–96S.

Harashima, Hirofumi, Nico Dissmeyer, and Arp Schnittger (2013). “Cell cycle control across the eukaryotic kingdom”. In: *Trends Cell Biol* 23.7, pp. 345–56. DOI: 10.1016/j.tcb.2013.03.002.

Harii, Norikazu et al. (2005). “Thyrocytes express a functional toll-like receptor 3: overexpression can be induced by viral infection and reversed by phenylmethimazole and is associated with Hashimoto’s autoimmune thyroiditis”. In: *Mol Endocrinol* 19.5, pp. 1231–50. DOI: 10.1210/me.2004-0100.

Hikami, K. et al. (2011). “Association of a functional polymorphism in the 3’-untranslated region of SPI1 with systemic lupus erythematosus”. eng. In: *Arthritis and Rheumatism* 63.3, pp. 755–763.

Hollenhorst, P. C. et al. (2007). "Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family". eng. In: *Genes and development* 21.15, pp. 1882–1894.

Horspool, Daniel (2008). *Central Dogma of Molecular Biochemistry with Enzymes.* URL: http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology#mediaviewer/File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg.

Hrincius, Eike R et al. (2010). "CRK adaptor protein expression is required for efficient replication of avian influenza A viruses and controls JNK-mediated apoptotic responses". In: *Cell Microbiol* 12.6, pp. 831–43. DOI: 10.1111/j.1462-5822.2010.01436.x.

Hunter, Lawrence E (2012). *The Processes of Life.* The MIT press.

Iliopoulos, D., A. Rotem, and K. Struhl (2011). "Inhibition of miR-193a expression by Max and RXRalpha activates K-Ras and PLAU to mediate distinct aspects of cellular transformation". eng. In: *Cancer research* 71.15, pp. 5144–5153.

Imaki, Hiroyuki et al. (2003). "Cell cycle-dependent regulation of the Skp2 promoter by GA-binding protein". In: *Cancer Res* 63.15, pp. 4607–13.

Iyer, Vishwanath R et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF". In: *Nature* 409.6819, pp. 533–538.

Jeffries, Matlock A et al. (2011). "Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus". In: *Epigenetics* 6.5, pp. 593–601.

Joaquin, M and R J Watson (2003). "Cell cycle regulation by the B-Myb transcription factor". In: *Cell Mol Life Sci* 60.11, pp. 2389–401. DOI: 10.1007/s00018-003-3037-4.

Johnson, C. R. et al. (2003). “Requirement for SAPK-JNK signaling in the induction of apoptosis by ribosomal stress in REH lymphoid leukemia cells”. eng. In: *Leukemia* 17.11, pp. 2140–2148.

Johnson, David S et al. (2007). “Genome-wide mapping of in vivo protein-DNA interactions”. In: *Science* 316.5830, pp. 1497–1502.

Jones, Dylan T et al. (2012). “Gene expression analysis in human breast cancer associated blood vessels”. In: *PLoS One* 7.10, e44294. DOI: 10.1371/journal.pone.0044294.

Kaji, R. et al. (2005). “Molecular dissection and anatomical basis of dystonia: X-linked recessive dystonia-parkinsonism (DYT3)”. In: *J. Med. Invest.* 52 Suppl, pp. 280–283.

Kanehisa, Minoru and Susumu Goto (2000). “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1, pp. 27–30.

Kanehisa, Minoru et al. (2014). “Data, information, knowledge and principle: back to metabolism in KEGG”. In: *Nucleic acids research* 42.D1, pp. D199–D205.

Kano, Y. et al. (2005). “Regulatory roles of bone morphogenetic proteins and glucocorticoids in catecholamine production by rat pheochromocytoma cells”. In: *Endocrinology* 146.12, pp. 5332–5340.

Karlsson, Olof P et al. (1997). “Lipid dependence and basic kinetics of the purified 1, 2-diacylglycerol 3-glucosyltransferase from membranes of Acholeplasma laidlawii”. In: *Journal of Biological Chemistry* 272.2, pp. 929–936.

Kelvinsong (2012). *Transcription Factors.* URL: http://commons.wikimedia.org/wiki/File:Transcription_Factors.svg#mediaviewer/File:Transcription_Factors.svg.

Kendrew, John C et al. (1958). “A three-dimensional model of the myoglobin molecule obtained by x-ray analysis”. In: *Nature* 181.4610, pp. 662–666.

Khatri, Purvesh and Sorin Drăghici (2005). "Ontological analysis of gene expression data: current tools, limitations, and open problems". In: *Bioinformatics* 21.18, pp. 3587–3595.

Khatri, Purvesh, Marina Sirota, and Atul J Butte (2012). "Ten years of pathway analysis: current approaches and outstanding challenges". In: *PLoS computational biology* 8.2, e1002375.

Kim, Seon-Young and David J Volsky (2005). "PAGE: parametric analysis of gene set enrichment". In: *BMC bioinformatics* 6.1, p. 144.

Kinsella, Rhoda J et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space". In: *Database* 2011, bar030.

Klug, William S et al. (2012). *Concepts of genetics.* Ed. 10. Pearson Education, Inc. Chap. 1.

Kong, Ling-Min et al. (2014). "A regulatory loop involving miR-22, Sp1, and c-Myc modulates CD147 expression in breast cancer invasion and metastasis". In: *Cancer Res* 74.14, pp. 3764–78. DOI: 10.1158/0008-5472.CAN-13-3555.

Koo, Chuay-Yeng, Kyle W Muir, and Eric W-F Lam (2012). "FOXM1: From cancer initiation to progression and treatment". In: *Biochim Biophys Acta* 1819.1, pp. 28–37. DOI: 10.1016/j.bbagrm.2011.09.004.

Kou, Yan (2014). *ChEA2 official website data download page.* URL: http://amp.pharm.mssm.edu/ChEA2/index.html#data.

Kou, Yan et al. (2013). "ChEA2: Gene-Set Libraries from ChIP-X Experiments to Decode the Transcription Regulome". In: *Availability, Reliability, and Security in Information Systems and HCI.* Springer, pp. 416–430.

Kriesel, J. D. et al. (2004). "STAT1 binds to the herpes simplex virus type 1 latency-associated transcript promoter". eng. In: *Journal of neurovirology* 10.1, pp. 12–20.

Kristiansson, Erik et al. (2009). “Evolutionary forces act on promoter length: identification of enriched cis-regulatory elements”. In: *Molecular biology and evolution* 26.6, pp. 1299–1307.

Lachmann, Alexander et al. (2010). “ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments”. In: *Bioinformatics* 26.19, pp. 2438–2444.

Laliotis, Aggelos et al. (2013). “Immunohistochemical study of pElk-1 expression in human breast cancer: association with breast cancer biologic profile and clinicopathologic features”. In: *Breast* 22.1, pp. 89–95. DOI: 10.1016/j.breast.2012.09.013.

Lam, E. W. et al. (1999). “Modulation of E2F activity in primary mouse B cells following stimulation via surface IgM and CD40 receptors”. In: *Eur. J. Immunol.* 29.10, pp. 3380–3389.

Lau, Eric and Ze’ev A Ronai (2012). “ATF2 - at the crossroad of nuclear and cytosolic functions”. In: *J Cell Sci* 125.Pt 12, pp. 2815–24. DOI: 10.1242/jcs.095000.

Lavery, Karen et al. (2008). “BMP-2/4 and BMP-6/7 differentially utilize cell surface receptors to induce osteoblastic differentiation of human bone marrow-derived mesenchymal stem cells”. In: *Journal of biological chemistry* 283.30, pp. 20948–20958.

Lee, Chun-Chung et al. (2012a). “TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer”. In: *J Biol Chem* 287.4, pp. 2798–809. DOI: 10.1074/jbc.M111.258947.

Lee, K. H. et al. (2004). “SMAD-mediated modulation of YY1 activity regulates the BMP response and cardiac-specific expression of a GATA4/5/6-dependent chick Nkx2.5 enhancer”. In: *Development* 131.19, pp. 4709–4723.

Lee, T C and E B Ziff (1999). “Mxi1 is a repressor of the c-Myc promoter and reverses activation by USF”. In: *J Biol Chem* 274.2, pp. 595–606.

Lee, Y. Y. et al. (2012b). “BCLAF1 is a radiation-induced H2AX-interacting partner involved in ÎH2AX-mediated regulation of apoptosis and DNA repair”. In: *Cell Death Dis* 3, e359.

Lefterova, M. I. et al. (2010). “Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages”. eng. In: *Molecular and cellular biology* 30.9, pp. 2078–2089.

Lempiainen, H. and D. Shore (2009). “Growth control and ribosome biogenesis”. eng. In: *Current opinion in cell biology* 21.6, pp. 855–863.

Lescuyer, P., P. Martinez, and J. Lunardi (2002). “YY1 and Sp1 activate transcription of the human NDUFS8 gene encoding the mitochondrial complex I TYKY subunit”. eng. In: *Biochimica et biophysica acta* 1574.2, pp. 164–174.

Leung, Yiu Tak et al. (2015). “Interferon regulatory factor 1 and histone H4 acetylation in systemic lupus erythematosus”. In: *Epigenetics* 10.3, pp. 191–9. DOI: 10.1080/15592294.2015.1009764.

Li, Weizhong et al. (2009). “Differential suppressive effect of promyelocytic leukemia protein on the replication of different subtypes/strains of influenza A virus”. In: *Biochem Biophys Res Commun* 389.1, pp. 84–9. DOI: 10.1016/j.bbrc.2009.08.091.

Li, Xinghua et al. (2014). “Transducin ()-like 1 X-linked receptor 1 promotes proliferation and tumorigenicity in human breast cancer via activation of beta-catenin signaling”. In: *Breast Cancer Res* 16.5, p. 465. DOI: 10.1186/s13058-014-0465-z.

Liberzon, Arthur et al. (2011). “Molecular signatures database (MSigDB) 3.0”. In: *Bioinformatics* 27.12, pp. 1739–1740.

Lieberthal, Jason G et al. (2009). "The role of YY1 in reduced HP1alpha gene expression in invasive human breast cancer cells". In: *Breast Cancer Res* 11.3, R42. DOI: 10.1186/bcr2329.

Lin, C. Y. et al. (2002). "The cell cycle regulatory factor TAF1 stimulates ribosomal DNA transcription by binding to the activator UBF". eng. In: *Current biology : CB* 12.24, pp. 2142–2146.

Lin, C. Y. et al. (2013). "Microarray analysis of gene expression of bone marrow stem cells cocultured with salivary acinar cells". In: *J. Formos. Med. Assoc.* 112.11, pp. 713–720.

Lin, Heng-Huei, Shao-Chuan Lai, and Lee-Young Chau (2011). "Heme oxygenase-1/carbon monoxide induces vascular endothelial growth factor expression via p38 kinase-dependent activation of Sp1". In: *J Biol Chem* 286.5, pp. 3829–38. DOI: 10.1074/jbc.M110.168831.

Liu, Li et al. (2012). "Influenza A virus induces interleukin-27 through cyclooxygenase-2 and protein kinase A signaling". In: *J Biol Chem* 287.15, pp. 11899–910. DOI: 10.1074/jbc.M111.308064.

Lu, M. C. et al. (2008). "Nifedipine suppresses Th1/Th2 cytokine production and increased apoptosis of anti-CD3 + anti-CD28-activated mononuclear cells from patients with systemic lupus erythematosus via calcineurin pathway". eng. In: *Clinical immunology (Orlando, Fla.)* 129.3, pp. 462–470.

Lu, N. Z. and J. A. Cidlowski (2005). "Translational regulatory mechanisms generate N-terminal glucocorticoid receptor isoforms with unique transcriptional target genes". eng. In: *Molecular cell* 18.3, pp. 331–342.

Lu, Xiaoming et al. (2015). "Lupus Risk Variant Increases pSTAT1 Binding and Decreases ETS1 Expression". In: *Am J Hum Genet.* DOI: 10.1016/j.ajhg.2015.03.002.

Luo, Weijun (2013). *gageData: Auxillary data for gage package*. R package version 2.0.3.

Luo, Weijun et al. (2009). "GAGE: generally applicable gene set enrichment for pathway analysis". In: *BMC bioinformatics* 10.1, p. 161.

Lutay, N. et al. (2013). "Bacterial control of host gene expression through RNA polymerase II". eng. In: *The Journal of clinical investigation* 123.6, pp. 2366–2379.

Lyons, Blair (2012). *How Genes are Expressed: Transcription Factors*. http://vimeo.com/30034882.

Mann, Henry B, Donald R Whitney, et al. (1947). "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* 18.1, pp. 50–60.

Mariana Ruiz sponk, Tryphon Magnus Manske Radio89 (2012). *Eukaryote DNA*. URL: http://commons.wikimedia.org/wiki/File:Eukaryote_DNA.svg.

Martianov, I. et al. (2010). "Cell-specific occupancy of an extended repertoire of CREM and CREB binding loci in male germ cells". In: *BMC Genomics* 11, p. 530.

McCord, Rachel Patton et al. (2007). "Inferring condition-specific transcription factor function from DNA binding and gene expression data". In: *Molecular systems biology* 3.1.

McDonnell, D P et al. (1995). "Cellular mechanisms which distinguish between hormone- and antihormone-activated estrogen receptor". In: *Ann N Y Acad Sci* 761, pp. 121–37.

McKnight, Patrick E and Julius Najab (2010). "Mann-Whitney U Test". In: *Corsini Encyclopedia of Psychology*.

McPherson, Lisa A, George W Woodfield, and Ronald J Weigel (2007). "AP2 transcription factors regulate expression of CRABPII in hormone responsive breast carcinoma". In: *J Surg Res* 138.1, pp. 71–8. DOI: 10.1016/j.jss.2006.07.002.

Mendenhall, E. M. et al. (2010). “GC-rich sequence elements recruit PRC2 in mammalian ES cells”. eng. In: *PLoS genetics* 6.12, e1001244.

Min, D-J et al. (2013). “MMSET stimulates myeloma cell growth through microRNA-mediated modulation of c-MYC”. In: *Leukemia* 27.3, pp. 686–94. DOI: 10.1038/leu.2012.269.

Moelans, Cathy B, Anoek H J Verschuur-Maes, and Paul J van Diest (2011). “Frequent promoter hypermethylation of BRCA2, CDH13, MSH6, PAX5, PAX6 and WT1 in ductal carcinoma in situ and invasive breast cancer”. In: *J Pathol* 225.2, pp. 222–31. DOI: 10.1002/path.2930.

Nishimura, Darryl (2001). “BioCarta”. In: *Biotech Software & Internet Report: The Computer Software Journal for Scient* 2.3, pp. 117–120.

Norbeck, Joakim et al. (1996). “Purification and Characterization of Two Isoenzymes of DL-Glycerol-3-phosphatase from Saccharomyces cerevisiae”. In: *Journal of Biological Chemistry* 271.23, pp. 13875–13881.

Nosrati, Nagisa, Neetu R Kapoor, and Vijay Kumar (2014). “Combinatorial action of transcription factors orchestrates cell cycle-dependent expression of the ribosomal protein genes and ribosome biogenesis”. In: *FEBS J* 281.10, pp. 2339–52. DOI: 10.1111/febs.12786.

Oberley, M. J., D. R. Inman, and P. J. Farnham (2003). “E2F6 negatively regulates BRCA1 in human cancer cells without methylation of histone H3 on lysine 9”. eng. In: *The Journal of biological chemistry* 278.43, pp. 42466–42476.

Oikawa, Tsuneyuki and Toshiyuki Yamada (2003). “Molecular biology of the Ets family of transcription factors”. In: *Gene* 303, pp. 11–34.

Orphanides, G., T. Lagrange, and D. Reinberg (1996). “The general transcription factors of RNA polymerase II”. eng. In: *Genes and development* 10.21, pp. 2657–2683.

Pavlidis, Paul et al. (2004). “Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex”. In: *Neurochemical research* 29.6, pp. 1213–1222.

Peña-Castillo, Lourdes et al. (2008). “A critical assessment of Mus musculus gene function prediction using integrated genomic evidence”. In: *Genome Biol* 9 Suppl 1, S2. DOI: 10.1186/gb-2008-9-s1-s2.

Peng, Wen Tao et al. (2003). “A panoramic view of yeast noncoding RNA processing”. In: *Cell* 113.7, pp. 919–33.

Peric-Hupkes, Daan et al. (2010). “Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation”. In: *Molecular cell* 38.4, pp. 603–613.

Perry, Robert P (2005). “The architecture of mammalian ribosomal protein promoters”. In: *BMC Evol Biol* 5, p. 15. DOI: 10.1186/1471-2148-5-15.

Peters, Linda M et al. (2002). “Mutation of a transcription factor, TFCP2L3, causes progressive autosomal dominant hearing loss, DFNA28”. In: *Human molecular genetics* 11.23, pp. 2877–2885.

Phuong, Nguyen Thi Thuy et al. (2014). “Aromatase induction in tamoxifen-resistant breast cancer: Role of phosphoinositide 3-kinase-dependent CREB activation”. In: *Cancer Lett* 351.1, pp. 91–9. DOI: 10.1016/j.canlet.2014.05.003.

Pierce, Benjamin A (2005). *Genetics: A conceptual approach*. Macmillan.

Pinzone, M. R. et al. (2015). “Epstein-barr virus- and Kaposi sarcoma-associated herpesvirus-related malignancies in the setting of human immunodeficiency virus infection”. In: *Semin. Oncol.* 42.2, pp. 258–271.

Prasov, Lev and Tom Glaser (2012). “Dynamic expression of ganglion cell markers in retinal progenitors during the terminal cell cycle”. In: *Mol Cell Neurosci* 50.2, pp. 160–8. DOI: 10.1016/j.mcn.2012.05.002.

Quinlan, Aaron R and Ira M Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842.

Rahnenfuhrer, Jorg et al. (2004). "Calculating the statistical significance of changes in pathway activity from gene expression data". In: *Statistical Applications in Genetics and Molecular Biology* 3.1, p. 1055.

Raney, Brian J et al. (2011). "ENCODE whole-genome data in the UCSC genome browser (2011 update)". In: *Nucleic acids research* 39.suppl 1, pp. D871–D875.

Resnick, Richard (2011). *Welcome to the genomic revolution.* URL: `https://www.ted.com/talks/richard_resnick_welcome_to_the_genomic_revolution#t-110999` (visited on 09/19/2014).

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140.

Savage, Kienan I et al. (2014). "Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability". In: *Mol Cell* 54.3, pp. 445–59. DOI: `10.1016/j.molcel.2014.03.021`.

Savitsky, David and Kathryn Calame (2006). "B-1 B lymphocytes require Blimp-1 for immunoglobulin secretion". In: *J Exp Med* 203.10, pp. 2305–14. DOI: `10.1084/jem.20060411`.

Schott, Jean-Jacques et al. (1998). "Congenital heart disease caused by mutations in the transcription factor NKX2-5". In: *Science* 281.5373, pp. 108–111.

Schuur, E. R. et al. (1995). "Nuclear factor I interferes with transformation induced by nuclear oncogenes". eng. In: *Cell growth and differentiation : the molecular biology journal of the American Association for Cancer Research* 6.3, pp. 219–227.

Shah, O Jameel et al. (2002). “The activated glucocorticoid receptor modulates presumptive autoregulation of ribosomal protein S6 protein kinase, p70 S6K”. In: *J Biol Chem* 277.4, pp. 2525–33. DOI: 10.1074/jbc.M105935200.

Shi, Hong et al. (2012). “Prognostic impact of polymorphisms in the MYBL2 interacting genes in breast cancer”. In: *Breast Cancer Res Treat* 131.3, pp. 1039–47. DOI: 10.1007/s10549-011-1826-2.

Singh, N. et al. (2013). “Downregulation of tumor suppressor gene PML in uterine cervical carcinogenesis: impact of human papillomavirus infection (HPV)”. In: *Gynecol. Oncol.* 128.3, pp. 420–426.

Sivertsen, E. A. et al. (2007). “Inhibitory effects and target genes of bone morphogenetic protein 6 in Jurkat TAg cells”. In: *Eur. J. Immunol.* 37.10, pp. 2937–2948.

Smirnov, N.V. (1944). “Approximate distribution laws for random variables, constructed from empirical data”. In: *Uspekhi Mat. Nauk* 10, 179—206.

Smyth, Gordon K (2005). “Limma: linear models for microarray data”. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by R. Gentleman et al. New York: Springer, pp. 397–420.

Sponk (2011). *Difference DNA RNA*. URL: http://commons.wikimedia.org/wiki/File%3ADifference_DNA_RNA-uk.svg.

Subramanian, Aravind et al. (2005). “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–15550.

Taqi, M. M. et al. (2011). “Prodynorphin promoter SNP associated with alcohol dependence forms noncanonical AP-1 binding site that may influence gene expression in human brain”. eng. In: *Brain research* 1385, pp. 18–25.

Tenbrock, K. et al. (2007). "Altered signal transduction in SLE T cells". eng. In: *Rheumatology (Oxford, England)* 46.10, pp. 1525–1530.

Thompson, Crista, Gwen MacDonald, and Christopher R Mueller (2011). "Decreased expression of BRCA1 in SK-BR-3 cells is the result of aberrant activation of the GABP Beta promoter by an NRF-1-containing complex". In: *Mol Cancer* 10, p. 62. DOI: 10.1186/1476-4598-10-62.

Thomson, Emma, Sébastien Ferreira-Cerca, and Ed Hurt (2013). "Eukaryotic ribosome biogenesis at a glance". In: *J Cell Sci* 126.Pt 21, pp. 4815–21. DOI: 10.1242/jcs.111948.

Tian, Lu et al. (2005). "Discovering statistically significant pathways in expression profiling studies". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.38, pp. 13544–13549.

Tierney, R. et al. (2007). "Epstein-Barr virus exploits BSAP/Pax5 to achieve the B-cell specificity of its growth-transforming program". eng. In: *Journal of virology* 81.18, pp. 10092–10100.

Topić, Iva et al. (2013). "Bone morphogenetic proteins regulate differentiation of human promyelocytic leukemia cells". In: *Leukemia research* 37.6, pp. 705–712.

Vaquerizas, Juan M et al. (2009). "A census of human transcription factors: function, expression and evolution". In: *Nature Reviews Genetics* 10.4, pp. 252–263.

Vastrik, Imre et al. (2007). "Reactome: a knowledge base of biologic pathways and processes". In: *Genome biology* 8.3, R39.

Versteege, Isabella et al. (2002). "A key role of the hSNF5/INI1 tumour suppressor in the control of the G1-S transition of the cell cycle". In: *Oncogene* 21.42, pp. 6403–12. DOI: 10.1038/sj.onc.1205841.

Vicent, Guillermo Pablo et al. (2013). "Unliganded progesterone receptor-mediated targeting of an RNA-containing repressive complex silences a subset of hormoneinducible genes". In: *Genes Dev* 27.10, pp. 1179–97. DOI: 10.1101/gad.215293.113.

Vilasco, Myriam et al. (2011). "Glucocorticoid receptor and breast cancer". In: *Breast Cancer Res Treat* 130.1, pp. 1–10. DOI: 10.1007/s10549-011-1689-6.

Villard, Jean (2004). "Transcription regulation and human diseases". In: *Swiss medical weekly* 134, pp. 571–579.

Vilotti, S et al. (2012). "The PML nuclear bodies-associated protein TTRAP regulates ribosome biogenesis in nucleolar cavities upon proteasome inhibition". In: *Cell Death Differ* 19.3, pp. 488–500. DOI: 10.1038/cdd.2011.118.

Vovk, Vladimir (2012). "Combining p-values via averaging". In: *arXiv preprint arXiv: 1212.4966*.

Wallerman, O. et al. (2009). "Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing". In: *Nucleic Acids Res.* 37.22, pp. 7498–7508.

Wan, Meimei et al. (2012). "Yin Yang 1 plays an essential role in breast cancer and negatively regulates p27". In: *Am J Pathol* 180.5, pp. 2120–33. DOI: 10.1016/j.ajpath.2012.01.037.

Wang, Li and Li-Jun Di (2014). "BRCA1 and estrogen/estrogen receptor in breast cancer: where they interact?" In: *Int J Biol Sci* 10.5, pp. 566–75. DOI: 10.7150/ijbs.8579.

Wang, Lu et al. (2014a). "CARM1 methylates chromatin remodeling factor BAF155 to enhance tumor progression and metastasis". In: *Cancer Cell* 25.1, pp. 21–36. DOI: 10.1016/j.ccr.2013.12.007.

Wang, X. et al. (2014b). "Latency-associated nuclear antigen of Kaposi sarcoma-associated herpesvirus promotes angiogenesis through targeting notch signaling effector Hey1". In: *Cancer Res.* 74.7, pp. 2026–2037.

Watanabe-Okochi, N. et al. (2013). “The shortest isoform of C/EBPbeta, liver inhibitory protein (LIP), collaborates with Evi1 to induce AML in a mouse BMT model”. eng. In: *Blood* 121.20, pp. 4142–4155.

Watson, James D, Francis HC Crick, et al. (1953). “Molecular structure of nucleic acids”. In: *Nature* 171.4356, pp. 737–738.

Wei, Chia-Lin et al. (2006). “A global map of p53 transcription-factor binding sites in the human genome”. In: *Cell* 124.1, pp. 207–219.

Wernicke, Mario et al. (2011). “Breast cancer and the stromal factor. The ”prometastatic healing process” hypothesis”. In: *Medicina (B Aires)* 71.1, pp. 15–21.

White, David et al. (2012). “The ATM substrate KAP1 controls DNA repair in heterochromatin: regulation by HP1 proteins and serine 473/824 phosphorylation”. In: *Mol Cancer Res* 10.3, pp. 401–14. DOI: 10.1158/1541-7786.MCR-11-0134.

Whitmore, Mark M et al. (2007). “Negative regulation of TLR-signaling pathways by activating transcription factor-3”. In: *J Immunol* 179.6, pp. 3622–30.

Wierstra, Inken and Jürgen Alves (2007). “FOXM1, a typical proliferation-associated transcription factor”. In: *Biol Chem* 388.12, pp. 1257–74. DOI: 10.1515/BC.2007.159.

Wingender, Edgar (2008). “The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation”. In: *Brief Bioinform* 9.4, pp. 326–32. DOI: 10.1093/bib/bbn016.

Wu, Fei et al. (2009). “Role of SP transcription factors in hormone-dependent modulation of genes in MCF-7 breast cancer cells: microarray and RNA interference studies”. In: *J Mol Endocrinol* 42.1, pp. 19–33. DOI: 10.1677/JME-08-0088.

Wu, Y. et al. (2014). “Involvement of c-Myc in the proliferation of MCF-7 human breast cancer cells induced by bHLH transcription factor DEC2”. ENG. In: *International journal of molecular medicine.*

Xie, Xiaohui et al. (2005). “Systematic discovery of regulatory motifs in human promoters and 3 UTRs by comparison of several mammals”. In: *Nature* 434.7031, pp. 338–345.

Xu, Ping et al. (2011). “JNK regulates FoxO-dependent autophagy in neurons”. In: *Genes Dev* 25.4, pp. 310–22. DOI: 10.1101/gad.1984311.

Xu, Yan et al. (2013). “Inhibition of proliferation of estrogen receptorpositive MCF7 human breast cancer cells by tamoxifen through cJun transcription factors”. In: *Mol Med Rep* 7.4, pp. 1283–7. DOI: 10.3892/mmr.2013.1306.

Yang, Liming, Yuming Luo, and Jifu Wei (2010). “Integrative genomic analyses on Ikaros and its expression related to solid cancer prognosis”. In: *Oncol Rep* 24.2, pp. 571–7.

Yin, Xin et al. (2010). “ATF3, an adaptive-response gene, enhances TGFsignaling and cancer-initiating cell features in breast cancer cells”. In: *J Cell Sci* 123.Pt 20, pp. 3558–65. DOI: 10.1242/jcs.064915.

Yiu, Gary K et al. (2011). “NFAT promotes carcinoma invasive migration through glypican-6”. In: *Biochem J* 440.1, pp. 157–66. DOI: 10.1042/BJ20110530.

Yuan, Chih-Chi et al. (2007). “CHD8 associates with human Staf and contributes to efficient U6 RNA polymerase III transcription”. In: *Mol Cell Biol* 27.24, pp. 8729–38. DOI: 10.1128/MCB.00846-07.

Zeng, Yan et al. (2014). “Inhibition of STAT5a by Naa10p contributes to decreased breast cancer metastasis”. In: *Carcinogenesis* 35.10, pp. 2244–53. DOI: 10.1093/carcin/bgu132.

Zephyris (2011). *DNA Structure+Key+Labelled*. URL: http://commons.wikimedia.org/wiki/File:DNA_Structure%2BKey%2BLabelled.png#mediaviewer/File:DNA_Structure%2BKey%2BLabelled.png.

Zhang, M. et al. (2005a). “Activation of bone morphogenetic protein-6 gene transcription in MCF-7 cells by estrogen”. In: *Chin. Med. J.* 118.19, pp. 1629–1636.

Zhang, X. et al. (2005b). “Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues”. In: *Proc. Natl. Acad. Sci. U.S.A.* 102.12, pp. 4459–4464.

Zhang, Yongliang et al. (2009). “MKP-1 is necessary for T cell activation and function”. In: *J Biol Chem* 284.45, pp. 30815–24. DOI: 10.1074/jbc.M109.052472.

Zhao, H et al. (2012). “An intronic variant associated with systemic lupus erythematosus changes the binding affinity of Yinyang1 to downregulate WDFY4”. In: *Genes Immun* 13.7, pp. 536–42. DOI: 10.1038/gene.2012.33.

Zhou, J. and J. A. Cidlowski (2005). “The human glucocorticoid receptor: one gene, multiple proteins and diverse responses”. eng. In: *Steroids* 70.5-7, pp. 407–417.

Zhu, Wei et al. (2010). “A whole genome transcriptional analysis of the early immune response induced by live attenuated and inactivated influenza vaccines in young children”. In: *Vaccine* 28.16, pp. 2865–76. DOI: 10.1016/j.vaccine.2010.01.060.