Pathway Analysis for ChIP-seq Data

By Atsalaki Xanthoula

BSc, Technological Educational Institute of Crete, 2014

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF INFORMATICS ENGINEERING

SCHOOL OF ENGINEERING

TECHNOLOGICAL EDUCATIONAL INSTITUTE OF CRETE

2016

Approved by:

Professor Manolis Tsiknakis

Statement of Originality

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution or any other purpose. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except as specified in references, acknowledgments or in footnotes. I certify that the intellectual content of this thesis is the product of my own work and all the assistance received in preparing this thesis and sources have been acknowledged.

Atsalaki Xanthoula

Acknowledgements

I would like to express the deepest appreciation and my utmost gratitude to my supervisor Prof. Manolis Tsiknakis. His exceptional scientific knowledge, experience and creative thinking have been a source of inspiration and motivation for me while his patience and encouragement were determinant for the accomplishment of this thesis. In addition, I would like to express my sincere and warm acknowledgement in the continuous support, constructive suggestions and guidance of my thesis advisor Lefteri Koumaki.

Lastly, I would like to thank Charles Joly the author and maintainer of ENCODExplorer package from the Bioconductor packages that was really willing to help in fixing all the bugs the R packages had from time to time immediately during this thesis.

Abstract

With the completion of the human genome sequence, attention turned to identifying and annotating its functional DNA elements and the interactions among them through different phenotypes. Epigenetics is a field of biology that studies these interactions.
On one hand, an important issue in deciphering the epigenetic code is whether two given histone modifications, transcription factors or chromatin modifiers are co-enriched on the same locus [1]. To resolve these issues, ChIP-seq has been developed, where one protein is immunoprecipitated from a chromatin sample, and a second protein is subsequently immunoprecipitated from chromatin eluted from the first ChIP [1]. On the other hand, analysis of GRN can help in identifying important or core regulatory genes (TFs and miRNAs) that play significant role in controlling the specificity of gene expression during a biological process [2]. These two keys, can unravel the mystery behind diseases and their treatment, and can be a powerful tool when combined in software tools in the hands of biologists. The objective of the thesis is to explore the effect of ChIP-seq data, coming from specific proteins under specific conditions, in functional sub-pathways for specific phenotype. A Shiny application combined with R programming language was developed for the download and analysis of ChIP-seq data from the ENCODE Experiment ChIP-seq Matrix and the extended version of the open source pathway analysis tool MinePath was used for the identification and visualization of functional sub-pathways.

Περίληψη

Με την ραγδαία εξέλιξη της τεχνολογίας και την ολοκλήρωση του προγράμματος χαρτογράφησης του Ανθρώπινου Γονιδιώματος δημιουργήθηκε η ανάγκη και οι επιστήμονες στράφηκαν στην αναγνώριση των σχέσεων που εκφράζονται ανάμεσα στα γενετικά στοιχεία του DNA του ανθρώπινου κυττάρου σε διαφορετικούς φαινότυπους. Η Επιγενετική είναι ο κλάδος της Βιολογίας που μελετά αυτές τις συσχετίσεις στα στοιχεία του ανθρώπινου DNA . Η Επιγενετική στις μέρες μας χρησιμοποιεί νέας γενιάς τεχνολογίες αλληλούχισης, όπως η ChIP-seq τεχνολογία και τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό. Η ανάπτυξη εφαρμογών web based λογισμικού που χρησιμοποιεί την τεχνολογία ChIP-seq για την ανάλυση και αποκρυπτογράφηση του ανθρώπινου γονιδιώματος είναι στο

επίκεντρο του ενδιαφέροντος των βιοπληροφορικών στην προσπάθεια τους να ανακαλύψουν νέες μεθόδους αντιμετώπισης των ασθενειών.

Από την άλλη μεριά, η ανάλυση των Ρυθμιστικών Δικτύων Γονιδίων μπορεί να βοηθήσει στον εντοπισμό σημαντικών ρυθμιστικών γονιδίων (TFs/μεταγραφικοί παράγοντες και miRNAs) που παίζουν σημαντικό ρόλο στον έλεγχο της έκφρασης του γονιδίου κατά τη διάρκεια μιας βιολογικής διαδικασίας. Αυτά τα δύο κλειδιά, μπορούν να ξετυλίξουν το μυστήριο πίσω από τις ασθένειες και τη θεραπεία τους, και μπορεί να αποτελέσουν ισχυρό εργαλείο όταν συνδυαστούν με εργαλεία λογισμικού στα χέρια των βιολόγων. Ο στόχος της διατριβής είναι η διερεύνηση της επίδρασης των δεδομένων ChIP-seq , που προέρχονται από ειδικές πρωτεΐνες κάτω από συγκεκριμένες συνθήκες, πάνω σε λειτουργικά υπομονοπάτια για συγκεκριμένους φαινότυπους. Για τα πειράματα η βάση δεδομένων που θα χρησιμοποιηθεί είναι ο Experiment ChIP-seq Matrix από το ENCODE Project. Για την ανάκτηση και την ανάλυση των δεδομένων δημιουργήσαμε μια web-based εφαρμογή υλοποιημένη με Shiny και R που τραβάει αυτόματα τα δεδομένα από την βάση βιοδεδομένων του ENCODE. Τέλος το εργαλείο ανάλυσης MinePath χρησιμοποιήθηκε για την ταυτοποίηση και την απεικόνιση των λειτουργικών υπομονοπατιών αυτών.

Table of Contents

List of Figures

List of Abbreviations

CTCF	CCCTC-binding factor(11 zing finger protein)
ChIP	Chromatin ImmunoPrecipitation
ChIP-seq	Chromatin ImmunoPrecipitation Sequencing
DHS	DnaseI Hypersensitive Sites
DNA	DeoxyriboNucleic Acid
ENCODE	Encyclopedia of DNA Elements
FDR	False Discovery Rate
FTP	File Transfer Protocol
GC-CONTENT	Guanine-Cytosine Content
GO	Gene Ontology
GRN	Gene Regulatory Network
HF	Histone Modification
HGP	Human Genome Project
HTTP	HyperText Transfer Protocol
IDR	Irreproducible Discovery Rate
IP	ImmunoPrecipitation
KEGG	Kyoto Encyclopedia of Genes and Genomes
mRNA	Message RNA
miRNA	Micro RNA
NGS	Next Generation Sequencing
NHGRI	National Human Genome Research Institute
NSC	Normalized strand cross-correlation
P-VALUE	Probability Value
Q-VALUE	Quality Value
RNA	RiboNucleic Acid
RSC	Relative strand cross-correlation
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
tRNA	Transfer RNA
TSS	Transcription Site

List of Biology Terms

Budding Yeast	Any of various unicellular, nucleated, usually rounded fungi that reproduce by budding; some are fermenters of carbohydrates, and a few are pathogenic for humans.
Cell	The structural and functional unit of all organisms; an autonomous self-replicating unit that may exist as functional independent unit of life (as in the case of unicellular organism), or as sub-unit in a multicellular organism that is specialized into carrying out particular functions towards the cause of the organism as a whole.
Cis-regulatory modules	A stretch of DNA, usually 100-1000 DNA base pairs in length, where a number of transcription factors can bind and regulate expression of nearby genes and regulate their transcription rates.
Cistrome	The sum of DNA binding sites of a specific transcription factor
Deoxyribonucleic acid	The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These four bases are attached to the sugar/phosphate to form the complete nucleotide, as shown for adenosine monophosphate
Exon	Any part of a gene that will become a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing. The term exon refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts. In RNA splicing, introns are removed and exons are covalently joined to one another as part of generating the mature messenger RNA
Gene Expression	The process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA
Gene Regulation	The informal term used to describe any mechanism used by a cell to increase or decrease the production of specific gene

	products (protein or RNA). Cells can modify their gene expression patterns to trigger developmental pathways, respond to environmental stimuli, or adapt to new food sources. All points of gene expression can be regulated. This includes transcription, RNA processing and transport, translation and post-translational modification of a protein, and mRNA degradation.
Genome	An organism's complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism. In humans, a copy of the entire genome, more than 3 billion DNA base pairs, is contained in all cells that have a nucleus
Genomics	The branch of molecular biology concerned with the structure, function, evolution and mapping of genomes
Intron	Any nucleotide sequence within a gene that is removed by RNA splicing during maturation of the final RNA product. The term intron refers to both the DNA sequence within a gene and the corresponding sequence in RNA transcripts
miRNA	A small non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals and some viruses that functions in RNA silencing and post-transcriptional regulation of gene expression.
Promoters	A region of DNA that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes, on the same strand and upstream on the DNA (towards the 5' region of the sense strand). Promoters can be about 100–1000 base pairs long
Protein	Any of a group of complex organic macromolecules that contain carbon, hydrogen, oxygen, nitrogen and usually sulfur and are composed of one or more chains of amino acids. Proteins are fundamental components of all living cells and include many substances, such as enzymes, hormones and antibodies, which are necessary for the proper functioning of an organism.

	They are essential in the diet of animals for the growth and rep air of tissue and can be obtained from foods such as meat, fish, eggs, milk and legumes
RNA	A nucleic acid that is generally single stranded (double stranded in some viruses) and plays a role in transferring information from DNA to protein-forming system of the cell
Transcription	The process of making RNA from a DNA template by RNA polymerase
Transcription Factor	A protein that binds to specific DNA sequences, thereby controlling the rate of transcription of genetic information from DNA to messenger RNA. Transcription factors perform this function alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes
Transcription Factor Binding Sites	The region of the gene to which a TF binds is called a transcription factor binding site. These sites are a subset of DNA binding sites. Overall, these sites can be defined as short segments of DNA that are specifically bound by one or more proteins with various functions
Transcriptome	The set of all messenger RNA molecules in one cell or a population of cells
Translation	The process in which cellular ribosomes create proteins. In translation, messenger RNA (mRNA), produced by transcription from DNA, is decoded by a ribosome to produce a specific amino acid chain or polypeptide

1 Introduction

Why do some siblings look alike but have different eye color or blood types? How do people get old and suffer from different diseases? These are major questions that scientists more than half a century try to answer by studying human or other organisms through thorough analysis of their Deoxyribonucleic acid, more commonly known as DNA, which is a complex molecule that contains all of the information necessary to build and maintain an organism. All living organisms have DNA within their cells. In fact, nearly every cell in a multicellular organism possesses the full set of DNA required for that organism. However, DNA does more than specify the structure and function of living things, it also serves as the primary unit of heredity in organisms of all types. In other words, whenever organisms reproduce, a portion of their DNA is passed along to their offspring.

Since the first illustration of the double helical model of DNA by Watson and Crick in the 1950s, there have been great development in genome research and biotechnology that have drastically influenced the disease diagnosis and treatment. Many scientists worked together and large collaborative biological projects such as Human Genome Project (HGP)[1] and ENCODE Project[2] that are still in progress, revealed crucial biological information of Human Genome. The Human Genome holds an extraordinary trove of information about human development, physiology, medicine and evolution [3]. A genome is an organism's complete set of genetic instructions. Each genome contains all of the information needed to build that organism and allow it to grow and develop.

A challenge facing researchers today is that of piecing together and analyzing the plethora of data currently being generated through these genomics. Many web tools, platforms and software have been developed in order to analyze biological interactions that takes place in the human cells. At the molecular level in living cells biological pathways represent complex biological interactions. Gene regulatory networks (GRNs) are one of the major categories for such biological pathways.

GRNs are logic maps that state in detail the inputs into each cis-regulatory module, so that one can see how a given gene is fired off or on at a given time and place [4]. They

[1] https://www.encodeproject.org/
[2] https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

also provide specifically testable sets of predictions of just what target sites are hardwired into the cis-regulatory DNA sequence. The specific linkages constituting these networks provide a causal structure function answer to the question of how any given aspect of development is ultimately controlled by heritable genomic sequence information. The architecture reveals features that can never be appreciated at any other level of analysis but that turn out to embody distinguishing and deeply significant properties of each control system [4]. Figure 1 shows such a GRN between ADRB2 (Adrenoceptor Beta 2, a protein coding gen) and cancer-specific genes [5].

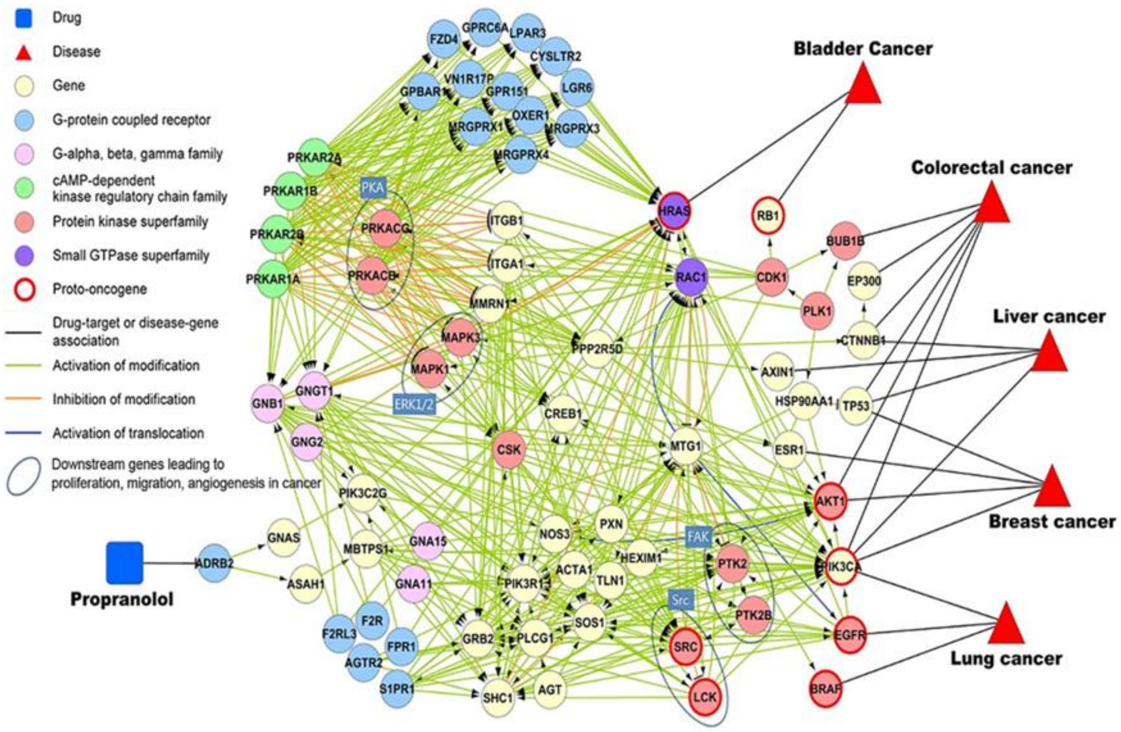


Figure 1: Gene Regulatory Network (GRN) between ADRB2 and cancer-specific genes *[5]*

While the development of new technologies is revolutionizing genome-wide analysis and scientists' abilities to have a better understanding of the biological meaning, inferring gene regulatory networks from such data is still a major challenge in systems biology [6]. Demand for analyzing very large datasets is increasing, especially with the introduction of ChIP-sequencing which is a recent method of Next Generation Sequencing (NGS) used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the cistrome of DNA-associated proteins; i.e. the DNA binding sites of a transcription factor. The sites are usually represented in the form of peaks. Sites or binding sites in biochemistry are regions on a protein or a piece of DNA or RNA to which molecules form a chemical bond. Large quantities of data generated from ChIP-seq experiments require effective computational analysis for uncovering biological mechanisms.

Chromatin immunoprecipitation combined with the next-generation DNA sequencing technologies (ChIP-seq) becomes a key approach for detecting genome-wide sets of genomic sites bound by proteins, such as Transcription Factors (TFs). Several methods and open-source tools have been developed to analyze ChIP-seq data.

1.1 Scope and Objective

In a biological process specific genes are switched on (activated) or off (repressed). Analysis of GRN can help in identifying important or core regulatory genes (TFs and miRNAs) that play significant role in controlling the specificity of gene expression during a biological process [7].These core regulatory genes are candidates for further experimental investigation and potential targets for therapeutic intervention [8] [9]. Analysis of GRNs also enables quantitative modeling of gene expression which can be used for rational design of molecular approaches to target specific biological processes [10] and infer new biology [11].

While the analysis of GRNs is well described in bacteria and yeast [2], similar analysis in higher organisms such as humans is challenging for a variety of reasons. Firstly, our knowledge of regulatory interactions between genes is incomplete, which is further complicated by the fact that the interactions may vary across different tissues [2]. Secondly, GRNs in higher organisms are highly complex as each regulatory molecule has dozens to thousands of targets and correspondingly a gene is usually targeted by multiple regulators. There is also cross-regulation and auto-regulation among genes. Such multiplicity of interconnections and loops makes the human GRN resemble a tangled hairball which is more challenging to analyze than a yeast gene network [12] [13]. Lastly, gene expression is regulated at multiple levels in higher organisms and thereby transcriptional and post-transcriptional regulations represent only a fraction of total regulatory apparatus [14].

In addition, the advent of NGS has enabled researchers to study biological systems at a level never before possible [15]. ChIP-seq has displaced earlier methods to investigate Protein-DNA interactions almost entirely. Being able to analyze these interactions genome-wide has increased our understanding of transcription factor biology, chromatin modification and transcription. ChIP-seq technology allows high-fidelity mapping of different regulators, such as transcription factors (TFs) or epigenetic modifications to genomic locations, thus providing a basis for profiling transcriptional or epigenetic regulatory relationships. Accurate binding features of these factors on the genomic sequences can also be used to identify regulatory modules and reconstruct gene regulatory networks (GRNs) [16].

So, combining NGS (ChIP-seq) and GRNs can help the biologists to discover gene pathways associated with the different expression of genes in different phenotypes and help them understand complex biological processes such as cell cycle, cell differentiation, cell apoptosis, diseases and other. As a result, through the analysis of a disease pathway analysis and by finding TFs that may disrupt pathways that play a key role in the specific disease they can discover treatments of those diseases. ChIP-seq can provide important insights towards gene regulatory process particularly in combination with transcriptomic profiles from expression microarrays or RNA-seq, since ChIP-seq can help identify genes directly regulated by the factor [17]. So, how can we combine ChIP-seq and GRNs for a powerful approach to further understanding the molecular bases of complex diseases? What information of great biological value can we gain by analyzing ChIP-seq data and visualizing their Gene Regulatory Networks along with equivalent expression data among different phenotypes?

In this Master Thesis, we try to shed light to these questions by retrieving programmatically ChIP-seq peak files from a big genome project, The Encode Project, find the genes that are most likely to be expressed when a specific antibody target interacts with a specific human cell type, annotate the genes and visualize them in a GRN analysis web based tool MinePath [18]. We conducted several tests with specific TFs on human cells for specific phenotypes and reported the results. Our work can be of great assistance to many scientists in their analysis and research of the interactions of the genes in disease pathways or treatments. Information obtained using Next Generation Sequencing along with the integration of GRNs allows researchers to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.

1.2 Thesis Overview

The thesis is organized in seven chapters, as follows:

- **Chapter 1:** the current chapter, includes a brief introduction of the topic, highlights the scope and the main objectives of this dissertation.
- **Chapter 2:** provides the theoretical background for understanding basic genetic terms about ChIP-seq technology, GRNs and the statistical terms p-value and FDR.
- **Chapter 3:** presents an elaborate review of related work in an effort to identify, review and analyze the findings of all related studies published during the last six years.
- **Chapter 4:** explains in detail the technical implementation. In this chapter, the application that was developed during this master thesis is described in detail along with the libraries and functions from the Bioconductor. Moreover, the ENCODE guidelines are presented along with the Data Library of the ENCODE Experiment ChIP-seq matrix. Lastly, the MinePath, a web-based GRN visualization tool is presented.
- **Chapter 5:** presents the validation results for the analysis of specific ChIP-seq data and the Gene Regulatory Network that was inferred from them.
- **Chapter 6:** includes the discussion over the results.
- **Chapter 7:** is the last chapter of this thesis and includes the conclusion and the possible directions to follow in future work.

2 Background

Genomics in the last decade have tremendously developed and a new era has arisen in the decryption of the human genome .ChIP-chip was the star of chromatin analysis until ChIP-seq came along and stole the limelight. ChIP-seq uses the same chromatin IP (ImmunoPrecipitation) procedures as ChIP-chip with the difference that it couples it with quantitative next-generation sequencing technology to detect enrichment peaks. After the IP sample is generated, it is prepared for sequencing using any of the Next Generation Sequencing technologies. Using ChIP-seq technology we can identify DNA-interactions with specific proteins or other chromatin polymorphisms and specify the TFBSs (Transcription Factor Binding Sites) of them. Identification of transcription regulatory elements in a genome is an actively evolving topic in modern molecular biology. The major class of these elements is represented by TFBSs. Modern high-throughput techniques, such as ChIP-chip ChIP-Seq, allow genome-scale mapping of TF occupancy in a given cell type and state [19].

In addition, Gene Regulatory Networks play crucial role to decipher the human gene expression. Drawing Gene Regulatory Networks that connect TFs to their predicted target genes can uncover gene modules that implement a particular function. Both ChIP-seq and Gene Regulatory Networks can fully reveal biological procedures of great significance in medicine, diseases and in general genetics.

2.1 ChIP-seq

ChIP-seq is a next generation sequencing technology that combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

ChIP

ChIP is a powerful method to selectively enrich for DNA sequences bound by a particular protein in living cells. The ChIP process enriches specific crosslinked DNA-protein complexes using an antibody against the protein of interest. Oligonucleotide adaptors are then added to the small stretches of DNA that were bound to the protein of interest to enable massively parallel sequencing.

Sequencing

After size selection, all the resulting ChIP-DNA fragments are sequenced simultaneously using a genome sequencer. A single sequencing run can scan for genome-wide associations with high resolution, meaning that features can be located precisely on the chromosomes.

A ChIP-seq analysis is shown in Figure 2.

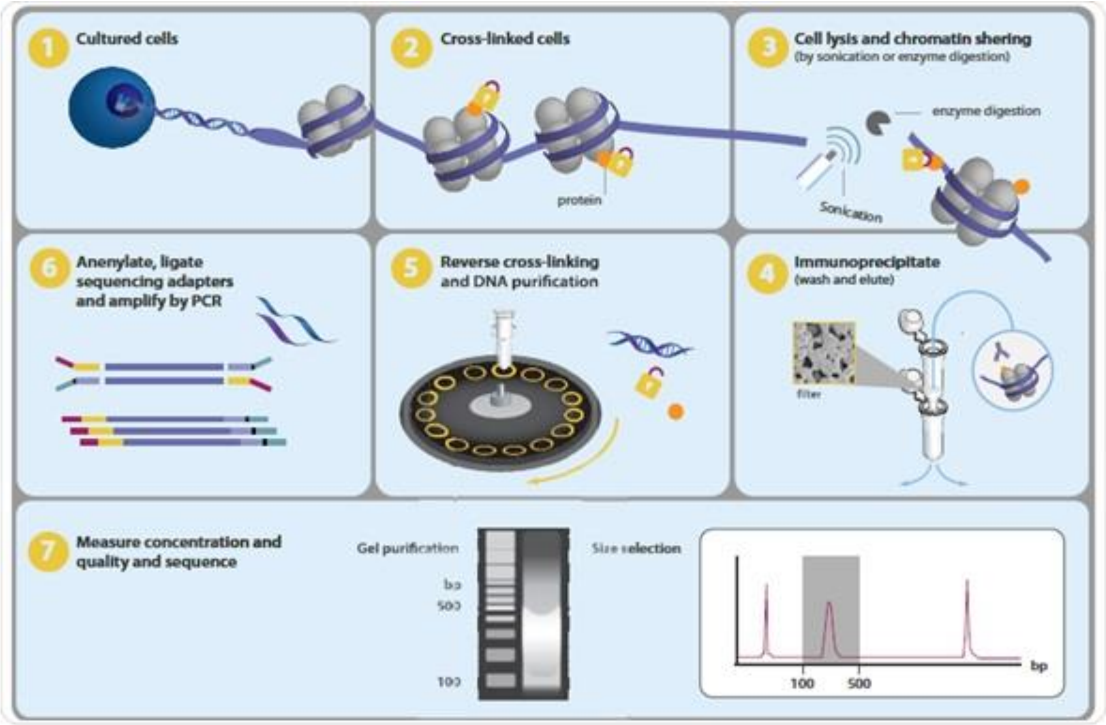


Figure 2: Overview of ChIP-seq process[3]

2.2 Gene Regulatory Networks

Gene regulatory networks (GRNs) are logic maps that state in detail the inputs into each cis-regulatory module, so that one can see how a given gene is fired off at a given time and place [4]. In transcriptional regulation, proteins called transcription factors (TFs) regulate the transcription of their target genes to produce messenger RNA (mRNA), whereas in post-transcriptional regulation microRNAs (miRNAs) cause degradation and repression of target mRNAs. These interactions are represented in a GRN by adding edges linking TF or miRNA genes to their target mRNAs. Since these physical interactions are fixed, we can represent a GRN as a static network even though regulatory interactions occur dynamically in space and time [2].

The interaction in a GRN could be of many types such as activation, inhibition, catalysis, binds to, co-cited [18]. An indicative example of the pathways in cell Apoptosis GRN from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database is shown in Figure 3.

[3] As retrieved from http://www.news-medical.net/ in 30 of August 2016

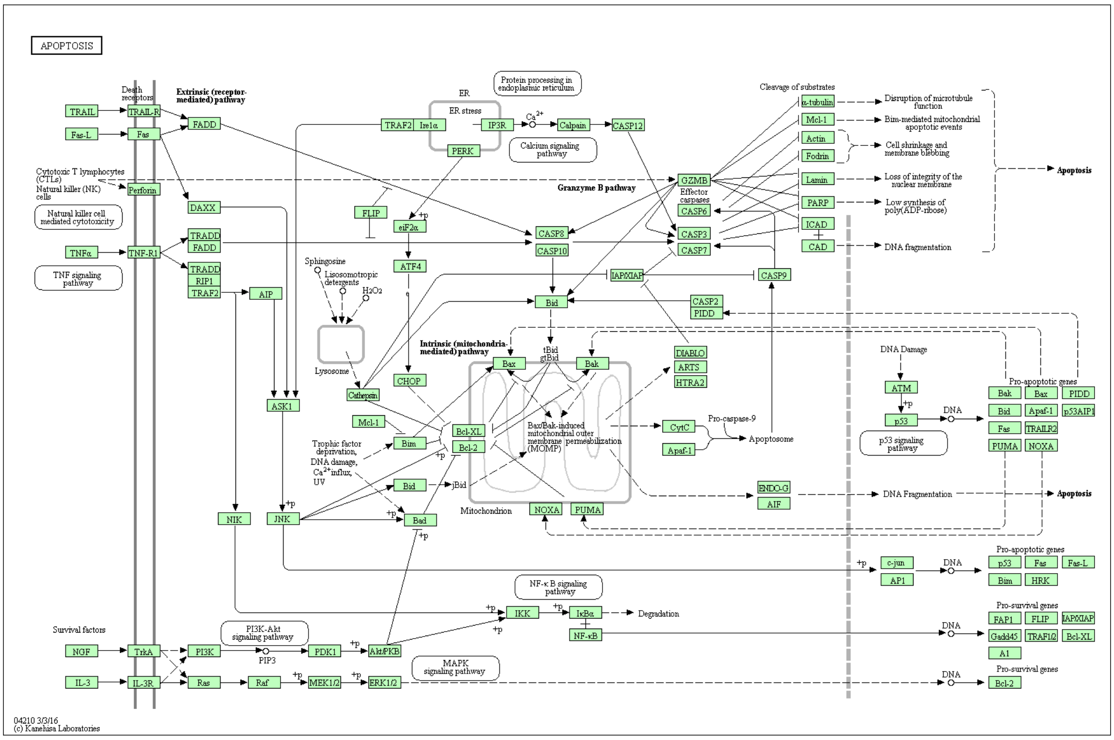


Figure 3: Apoptosis GRN from KEGG

2.3 P-Value

In the analysis of genetic association studies, a parameter of statistical significance, a p-value, is used to determine the certainty of an association. A p-value provides the probability that a given result from a test is due to chance. Andrea S. Foulkes in his book “Applied Statistical Genetics with R: For Population-based Association Studies” describes the statistical significance (p-value) in genetic studies thoroughly [20].

In statistics, if you want to draw conclusions about a null hypothesis H_0 (reject or fail to reject) based on a p-value, you need to set a predetermined cutoff point where only those p-values less than or equal to the cutoff will result in rejecting H_0.

While 0.05 is a very popular cutoff value for rejecting H_0, cutoff points and resulting decisions can vary, some people use stricter cutoffs, such as 0.01, requiring more evidence before rejecting H_0, and others may have less strict cutoffs, such as 0.10, requiring less evidence.

If H_0 is rejected (that is, the p-value is less than or equal to the predetermined significance level), the researcher can say she’s found a statistically significant result. A result is statistically significant if it’s too unlikely to have occurred by chance assuming H_0 is true. If you get a statistically significant result, you have enough evidence to reject the claim, H_0, and conclude that something different or new is in effect (that is, H_a - Alternative Hypothesis). The significance level can be thought of as the highest possible p-value that would reject H_0 and declare the results statistically significant. Following are the general rules for making a decision about H_0 based on a p-value:

- If the p-value is less than or equal to your significance level, then it meets your requirements for having enough evidence against H_0; you reject H_0.
- If the p-value is greater than your significance level, your data failed to show evidence beyond a reasonable doubt; you fail to reject H_0.

However, if you plan to make decisions about H_0 by comparing the p-value to your significance level, you must decide on your significance level ahead of time.

So, either you have enough evidence to say it's false (in which case you reject H_0) or you don't have enough evidence to say it's false (in which case you fail to reject H_0).

As a result of all the above a few guidelines we follow to make a decision (reject or fail to reject H_0) based on a p-value when our significance level is 0.05 are:

- If the p-value is less than 0.01 (very small), the results are considered highly statistically significant — reject H_0.
- If the p-value is between 0.05 and 0.01 (but not super-close to 0.05), the results are considered statistically significant — reject H_0.
- If the p-value is really close to 0.05 (like 0.051 or 0.049), the results should be considered marginally significant — the decision could go either way.
- If the p-value is greater than (but not super-close to) 0.05, the results are considered non-significant — you fail to reject H_0.

For example a p-value=0.05 means that 1 in 20 or there is a 5% chance that a result is really no different from the null hypothesis. This is valid for one test, in case of 20.000 tests there will be 1000 potential false positives.

A false positive error, or in short false positive, commonly called a "false alarm", is a result that indicates a given condition has been fulfilled, when it actually has not been fulfilled, erroneously a positive effect has been assumed. In the case of "crying wolf" – the condition tested for was "is there a wolf near the herd?". The actual result was that there had not been a wolf near the herd. The shepherd wrongly indicated there was one, by calling "Wolf, wolf!".

2.4 FDR and Q-Value

The FDR (False Discovery Rate) or q-value is similar to the well-known p-value, except it is a measure of significance in terms of the false discovery rate rather than the false positive rate [21]. A nice review of FDR in epigenetics can be found here[4].

False positives

A positive is a significant result, i.e. the p-value is less than your cut off value, normally 0.05. A false positive is when you get a significant difference where, in reality, none exists. As mentioned above the p-value is the chance that this data could occur given no difference actually exists. So, choosing a cut off of 0.05 means there is a 5% chance that we make the wrong decision.

The multiple testing problem

When we set a p-value threshold of, for example, 0.05, we are saying that there is a 5% chance that the result is a false positive. In other words, although we have found a statistically significant result, there is, in reality, no difference in the group means. While 5% is acceptable for one test, if we do lots of tests on the data, then this 5% can result in a large number of false positives. For example, if there are 2000 compounds in an experiment and we apply a t-test to each, then we would expect to get 100 (i.e. 5%) false positives by chance alone. This is known as the multiple testing problem.

Multiple testing and the False Discovery Rate

While there are a number of approaches to overcoming the problems due to multiple testing, they all attempt to assign an adjusted p-value to each test or reduce the p-value threshold from 5% to a more reasonable value. Many traditional techniques such as the Bonferroni correction are too conservative in the sense that while they reduce the number of false positives, they also reduce the number of true discoveries [22]. The False Discovery Rate approach is a more recent development. This approach also determines adjusted p-values for each test. However, it controls the number of false discoveries in those tests that result in a discovery (i.e. a significant result). Because of this, it is less conservative that the Bonferroni approach and has greater ability (i.e. power) to find truly significant results.

Another way to look at the difference is that a p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies

[4] http://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf

that 5% of significant tests will result in false positives. The latter will result in fewer false positives.

So, q-values are the name given to the adjusted p-values found using an optimized FDR approach. The FDR approach is optimized by using characteristics of the p-value distribution to produce a list of q-values[5].

In this master thesis, in the peak files that there were no q-values an FDR adjusted p-value was estimated in the experiments that only p-value was given. We used p.adjust R function that takes as a parameter the adjustment method , which in our case is BH(Benjamini & Hochberg or its alias "fdr") [23]. All the binding sites were estimated using an FDR (or q-value) threshold <0.01.

All the above can also be reviewed and confirmed in John D. Storey and Robert Tibshirani paper [22].

3 Literature Review

Chromatin immunoprecipitation (ChIP) followed by genomic tiling microarray hybridization (ChIP-chip) or massively parallel sequencing (ChIP-seq) are two of the most widely used approaches for genome-wide identification and characterization of in vivo protein-DNA interactions [24]. These two methods are been used by biology scientists the last 15 years and each of them has its own advantages and disadvantages. On the other hand, GRNs can be used to visualize such interactions. Gene regulation is a general name for a number of sequential processes, the most well-known and understood being transcription and translation, which control the level of a gene's expression, and ultimately result with specific quantity of a target protein. A gene regulation system consists of genes, cis-elements, and regulators. The regulators are most often proteins, called transcription factors, but small molecules, like RNAs and metabolites, sometimes also participate in the overall regulation. The interactions and binding of regulators to cis-elements in the cis-region of genes controls the level of gene expression during transcription. The cis-regions serve to aggregate the input signals, mediated by the regulators, and thereby effect a very specific gene expression signal. The genes, regulators, and the regulatory connections between them, together with an interpretation scheme form gene networks [25].

[5] As retrieved from http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/pq-values.aspx in 15th of July 2016

In the literature, there are many software tools developed for the analysis of ChIP-seq data and the analysis of GRNs. There also tools that combine those two but not so efficiently from the view of a most coherent, holistic and simple analysis tool to identify DNA interactions.

3.1 ChIP – Chip

The first studies for ChIP – on- Chip were published in 1999 when a team of scientists tried to analyze the distribution of cohesion along budding yeast chromosome III [26]. The Chip-on-Chip technique was then applied successfully in the three papers that were published in 2000 and 2001 from a group of biology scientists [27] [28] [29]. Many more studies were published using this technique of ChIP – Chip to investigate interactions between proteins and DNA. Michael J. Buck and Jason D. Lieb, biology scientists, describe the ChIP-Chip experimental procedure in their paper in 2004 [30].

Chip-on-chip analysis combines chromatin immunoprecipitation and DNA microarray analysis to identify protein-DNA interactions that occur in living cells. Protein-DNA interactions are captured in vivo by chemical crosslinking [31]. There are 8 main steps in ChIP-chip analysis. In the first step the protein of interest is cross-linked with its DNA binding site. Then the cells are disintegrated using sonication and we get double-stranded chunks of DNA fragments. In the next step, only the cross-linked DNA with the protein of interest is filtered out of the set of DNA fragments, using an antibody specific to the protein of interest. The antibodies may be attached to a solid surface or some other physical property that allows separation of cross-linked complexes and unbound fragments. This procedure is called immunoprecipitation (IP) of the protein. This can be done with specific antibody to the native protein. The cross-linking of protein of interest-DNA complexes is reversed (usually by heating) and the DNA strands are purified. After an amplification and denaturation step, the single-stranded DNA fragments are tagged, e.g. Cy5 or Alexa 647.

Finally, the fragments are saturated on the surface of the DNA microarray. Whenever a labeled fragment "finds" a complementary fragment on the array, they will hybridize and form again a double-stranded DNA fragment. (Figure 4) shows the ChIP-Chip workflow described.

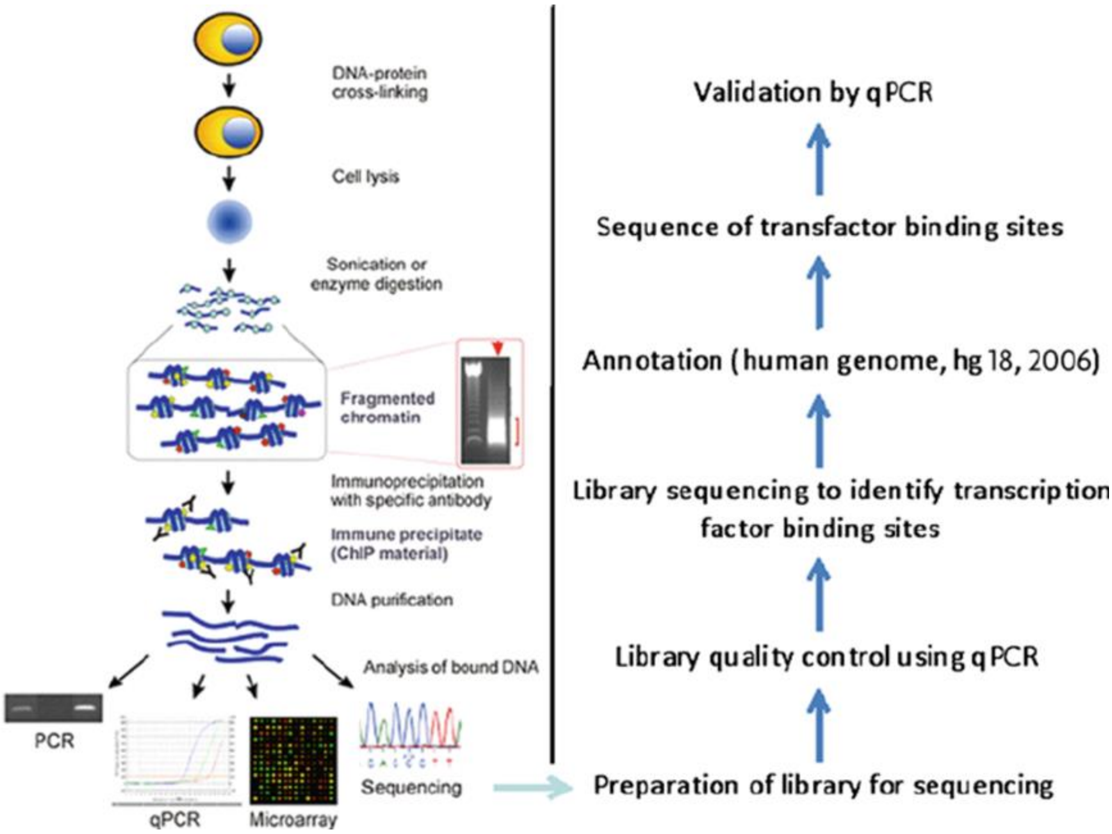


Figure 4: ChIP on Chip overview

Many software tools have been developed for chip-chip analysis such as ChIP-on-Chip online (CoCo) [32], Amadeus [33], CATCHprofiles [34], ChIP-on-chip Analysis Suite (CoCAS) [35],TileMap [36], a web-based analysis tool ChIPSeek (for ChIP-Chip and ChIP-Seq data analysis) [37], Cis-regulatory Element Annotation System(CEAS, for ChIP-Chip and ChIP-Seq data analysis [38]) and many other R packages in Bioconductor and algorithms. A ChIP-Chip overview can be found in Lee's et.al paper «Chromatin immunoprecipitation and microarray-based analysis of protein location» [39] .

3.2 ChIP-seq, a next generation sequencing method

ChIP-seq has become a widely adopted genomic assay in recent years to determine binding sites for transcription factors or enrichments for specific histone modifications [40]. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) belongs to the NGS (Next Generation Sequencing) and is a technique for genome wide profiling of DNA-binding proteins, histone modifications, or nucleosomes in living cells [41]. Enabled by the tremendous progress in NGS, ChIP-Seq offers higher resolution, less noise, and greater coverage than its array-based predecessor ChIP-chip. With the decreasing cost of sequencing, ChIP-Seq has become an indispensable tool for studying gene regulation and epigenetic mechanisms [42].

Chromatin immunoprecipitation sequencing, or ChIP-seq, combines ChIP with next-generation sequencing. ChIP-seq protocols have been adapted from ChIP-chip methods: proteins are cross-linked to their bound DNA by formaldehyde treatment, cells are homogenized, and chromatin is sheared and immunoprecipitated with antibody-bound magnetic beads. The immunoprecipitated DNA is then used as the input for a next-generation sequencing library prep protocol, where it is sequenced and analyzed for DNA binding sites [43]. The outline of the above procedure of ChIP-Seq is shown in (Figure 5).

The above steps come from the biochemistry view, from the view of the bioinformatics the main steps in ChIP-seq data analysis(Figure 6 ***:*** *Work Flow for the computational Analysis*) are Quality Filtering, Read Mapping, Peak Calling (Detecting Enriched regions and Peak Annotation) and Motif Analysis(Finding de novo Motifs, Motif Site Prediction and Motif scoring and Annotation) [44] [45]. Many ChIP-Seq Data analysis tools have been developed and published since the adoption of the ChIP-seq technology for revealing the human genome. Searching the literature we found many tools and software that scientists or biology researchers use to analyze and visualize ChIP-seq data. There are various ChIP-seq applications that can be categorized according to the step or the type of analysis a biologist wants to implement. There are many software tools for two of the steps we described above in the ChIP-seq pipeline, the peak calling and the motif discovery – e.g. ChIP-seq Peak Finder, BayesPeak, BroadPeak, MACS a widely used software tool, Peakzilla, PeakSeq, Bamm, CisFinder, CTF and many others, you can find great reviews of them and many others of other categories of ChIP-seq analysis along with their publications in http://omictools.com/chip-seq-category.

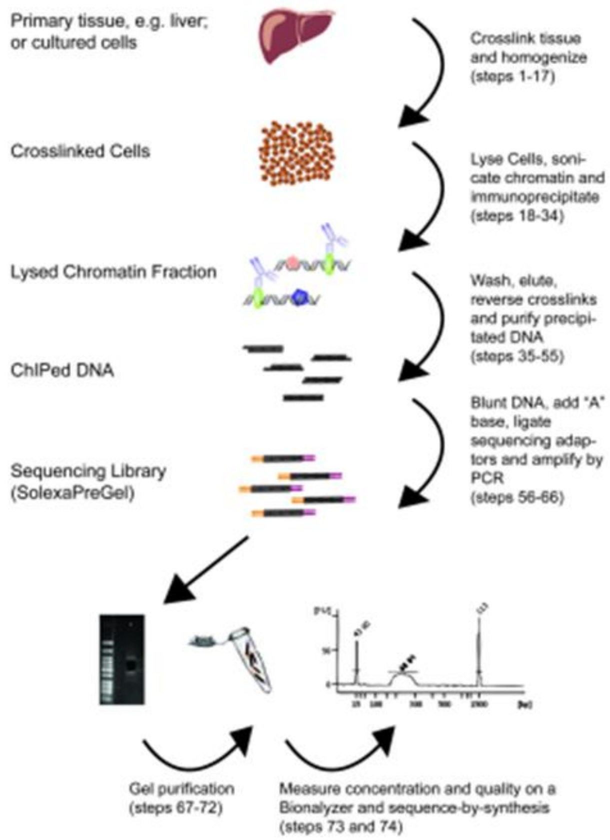


Figure 5: Outline of ChIP-seq procedure [43]

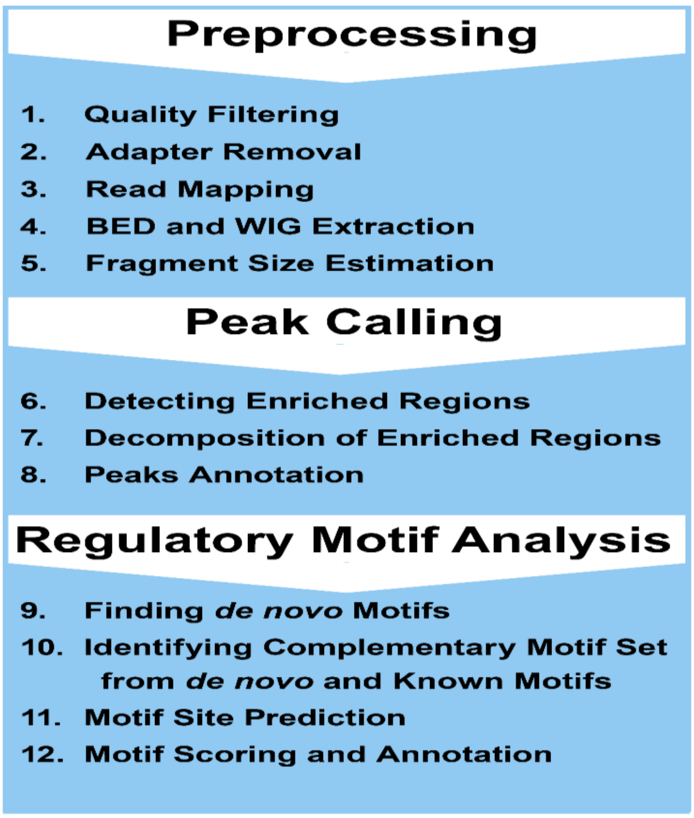


Figure 6 : Work Flow for the computational Analysis of ChIP-seq data

In his work Sebastian Steinhauser [40], reviewed 14 tools for ChIP-seq data analysis that are most often used by the biology researchers and categorized them as shown in (Figure 7).

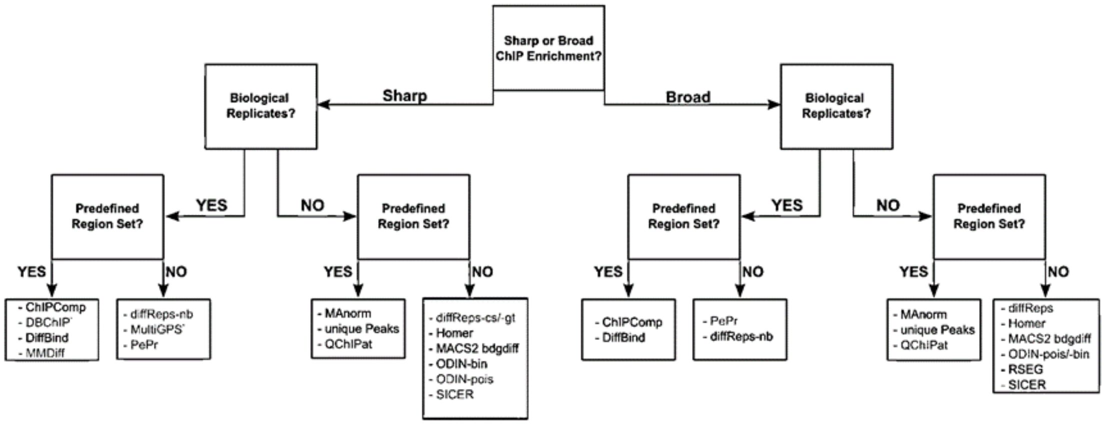


Figure 7: Decision tree indicating the proper choice of tool depending on the data set: shape of the signal (sharp peaks or broad enrichments), presence of replicates and presence of an external set of regions of interest. They have indicated in dark the name of the tools that give good results using default settings, and in gray the tools that would require parameter tuning to achieve optimal results: some tools suffer from an excessive number of DR (PePr, ODIN-pois), an insufficient number of DR (QChIPat, MMDiff, DBChIP) or from an imprecise definition of the DR for sharp signal (SICER, diffReps-nb). MultiGPS has been explicitly developed for transcription factor ChIP-seq *[40]*

Moreover, an excellent and thorough review for the software tools that are available for ChIP-seq analysis published in 2012 by Nuno A. Fonseca [46].He summarized many of them that are widely used by biology researchers as shown in (Figure 8). In this paper Nuno A. Fonseca classified the numerous software tools that map the generated reads to a reference sequence. The answer to the question "Which is the best of them" is difficult and depends on the scientists goals, the different type of data (e.g. miRNA, RNA, ChIP, and bisulfite) to analyze, the speed and accuracy they want to have in their experiments. Despite some recent evaluation studies (Bock et al., 2010; Li and Homer, 2010; Chatterjee et al., 2012) determining the most accurate and fastest mappers for a particular application is still difficult [46]. A regularly updated compendium of mappers can be found here[6] .

[6] http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Mapper	Min. RL	Max. RL	Mismatches	Indels	Gaps	Align. reported	Alignment	Parallel	QA	PE	Splicing	Data
BFAST		*	Y	Y	Y	B,R,U	G	SM	N	Y	N	DNA
Bismark	16	10 K	Score	Score	N	U	—	SM	Y	Y	N	Bisulphite
BLAT	11	5000K	Score	Score	Y	B	L	N	N	N	*de novo*	DNA
Bowtie	4	1 K	Score	Score	N	A,B,R,S	G L	SM	Y	Y	N	DNA
Bowtie2	4	5000K	Score	Score	Y	A,B,R,S	G L	SM	Y	Y	N	DNA
BS Seeker	—	—	3	0	N	U	—	SM	Y	N	N	Bisulphite
BSMAP	8	144	15	0	N	B,S,U		SM	N	Y	N	Bisulphite
BWA	4	200	Y	8	Y	R,S	G	SM	Y	Y	N	DNA
BWA-SW	4	1000K	0.1	0.1	Y	R,S	L	SM	Y	N	N	DNA
BWT-SW		1 K	Score	Score	Y	A		N	N	N	N	DNA
CloudBurst		1 K	Y	Y	Y	A,B	G	Cloud	N	N	N	DNA
DynMap	18	8 K	5	0	N	B	L	N	N	N	N	DNA
ELAND		32	2	0	N	B		N	N	N	N	DNA
Exonerate	20	*	Score	Score	Y	B,S	G L	N	N	N	*de novo*	DNA
GEM	0	4294M	1.0	1.0	Y	A, S	G	SM	Y	Y	Lib and *de novo*	DNA
GenomeMapper	12	2 K	10	10	Y	A,B,R	G	SM	N	N	N	DNA
GMAP	8	*	Y	Y	Y	B	G L	SM	N	N	*de novo*	DNA
GNUMAP	16	1 K	Score	Score	Y	B	G	SM/DM	Y	N	N	DNA
GSNAP	8	250	Y	Y	Y	A,B,U,S	G L	SM	N	Y	Lib and *de novo*	DNA
MapReads	10	120	Score	0	N	S		N	Y	N	N	DNA
MapSplice	—	—	3		Y	B	—	SM	N	Y	*de novo*	RNA
MAQ	8	63	Y	Y	N			N	Y	Y	N	DNA
MicroRazerS	10	*	Score	0	N	S	G	N	N	N	N	miRNA
MOM			Y	0	N	A	L	SM	N	Y	N	DNA
MOSAIK	15	1000	Y	Y	Y	A,B	G	SM	Y	Y	N	DNA
mrFAST	25	300	Score	6	N	A,B	G	N	N	Y	N	miRNA
mrsFAST	25	200	Y	0	N	A	G	N	N	Y	N	miRNA
Mummer 3	10	*	Y	Y	Y	A,B	G	N	N	N	N	DNA
Novoalign	30	300	8	2	N	A, B, R, U, S	G	SM/DM/Cloud	Y	Y	Lib	DNA
PASS	23	1 K	Y	Y	Y	A,B	G	SM	Y	Y	*de novo*	DNA
Passion	—	—	Y	Y	Y	U	—	SM	Y	Y	*de novo*	RNA
PatMaN	1	*	Y	Y	N	A	G	N	N	N	N	miRNA
PerM	20	128	9	0	Y	A,U	G	DM	Y	Y	N	DNA
ProbeMatch	36	50	3	Y	N	A,B		N	N	N	N	DNA
QPALMA	—	—	Y	Y	Y	B	L	N	Y	N	Lib and *de novo*	RNA
RazerS	11	*	Score	Score	Y	A,B,S	G	N	N	Y	N	DNA
REAL	4	*	Score	N	N	B, U	G	SM	Y	N	N	DNA
RMAP	11	10 K	Y	0	N	B,S		N	Y	Y	N	DNA
RNA-Mate	—	—	Y	0	N	S	—	DM	Y	N	Lib	RNA
RUM	—	—	Y	Y	Y	B	—	SM	N	Y	*de novo*	RNA
SeqMap	15	500	5	3	N	A		SM	N	N	N	DNA
SHRiMP	14	1 K	Score	Score	Y	B,S	G	SM	N	Y	N	DNA
SHRiMP 2	30	1 K	Y	Score	N	B,U,S	G	SM	Y	Y	N	DNA
Slider		62	3	0	N	B,S		N	Y	Y	N	DNA
Slider II		93	Y		N	B,S		N	N	Y	N	DNA
Smalt	4	2048M	Score	Score	N	A,B,R,U,S	L	SM	Y	Y	N	DNA
SOAP	7	60	5	3	N	B,R S		SM	N	Y	N	DNA
SOAP2	27	1 K	2	0	Y	A,B,R	L	SM	N	Y	N	DNA
SOAPSplice	13	3 K	5	2	Y	U	—	SM	Y	Y	*de novo*	RNA
SOCS		64	Y	0	N	A,B		SM	Y	N	N	DNA
SpliceMap	—	—	0.1		Y	A	—	SM	N	Y	Lib and/or *de novo*	RNA
SSAHA	15	*	Y	Y	Y	B,S	G L	N	N	N	N	DNA
SSAHA2	15	48 K	Score	Score	N	B,S	L	N	N	Y	N	DNA
Stampy	4	4 K	0.15	30	N	B,R,S	G	N	Y	Y	N	DNA
Supersplat			0	0	Y	A,U	G	N	N	N	*de novo*	RNA
TopHat	—	—	2	0	N	B,S	—	SM	Y	Y	*de novo*	RNA
VMATCH			Score	Score	Y	A,B,S	G L	N	N	N	N	DNA
WHAM	5	128	5	3	N	A,B,R,U,S	G	N	Y	Y	*de novo*	DNA
X-Mate	—	—	Y	0	N	S	—	DM	Y	N	Lib	DNA
ZOOM	12	240	Y	Y	N	B,S,U	G	SM/DM	Y	Y	N	DNA

Figure 8: Read length limits are shown in the first two data columns: minimum read length (Min. RL) and maximum read length (Max. RL). Unless otherwise stated the unit is base pairs, K denotes kilobases (1000 bases), M denotes megabases (1000K bases), and * denotes a (unknown) large number. The support for mismatches and short indels is presented in the fourth and fifth columns, respectively, including when possible the maximum number of allowed mismatches and indels: by default the value is in bases; in some cases, the value is presented as a proportion of the read size; or as score, meaning that mapper uses a score function. The Gaps column indicates whether consecutive insertions or deletions are allowed during alignment. The Alignments reported column indicates the alignments reported when a read maps to multiple locations: A-all, B-best, R-random, U-unique alignments only (no multi-maps), and S-user defined number of matches. The Alignment column indicates whether the reads are aligned end-to-end (Globally) or not (Locally). The Parallel column indicates whether the mapper can be run in parallel and, if yes, how: using an SM or/and a DM computer. The QA (Quality awareness) column indicates whether the mapper uses read quality information during the mapping. The support for paired reads is indicated in the PE column. The Splicing column indicates, for the RNA mappers, whether the

detection of splice junctions is made de novo or through user provided libraries (Lib). Yes is abbreviated as Y, and No is abbreviated as N. A cell in the table is filled with '—' when a third-party mapper is used to perform the alignment *[46]*

In their work Verfaillie, Imrichova, Janky and Aerts introduced two interesting tools iRegulon and iCisTarget. These tools perform regulatory sequence analysis (motif discovery) and integrate and mine large collections of existing regulatory data, such as ChIP-Seq, DHS-seq, and FAIRE-seq. While iRegulon focuses on sets of co-expressed genes, iCisTarget also analyses genomic regions as input [47].

Lastly, an evaluation of many of the tools that were reviewed above were selected and evaluated by Elizabeth G. Wilbanks and Marc T. Facciotti in their publication «Evaluation of Algorithm Performance in ChIP-Seq Peak Detection» depicted in (Figure 9) [48]. They measured their sensitivity, accuracy and usability and compared their performance. They concluded that eleven ChIP-seq analysis programs of varying algorithmic complexity identify protein binding sites from common empirical datasets with remarkably similar performance with regards to sensitivity and specificity. They also observed a few significant differences between the performances of these programs on their simulated datasets at increasing noise thresholds.

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

Figure 9: ChIP-seq peak calling programs selected for evaluation

3.3 Advantages and Disadvantages of ChIP-Chip and ChIP-Seq

Fueled by rapid development of the second generation high-throughput sequencing technologies in the past few years, ChIP-seq has emerged as an attractive alternative to ChIP-chip [42]. Both techniques has its own advantages and disadvantages relating to cost, bias, data analysis, specificity and significance of the biological results they produce.

ChIP-seq generally produces profiles with a better signal-to-noise ratio, and allows detection of more peaks and narrower peaks. The set of peaks identified by the two technologies can be significantly different, but the extent to which they differ varies depending on the factor and the analysis algorithm [24]. Importantly, in "ChIP-chip versus ChIP-seq : lessons for experimental design and data analysis" [24] the authors found that there is a significant variation among multiple sequencing profiles of input DNA libraries and that this variation most likely arises from both differences in experimental condition and sequencing depth. Moreover, the advances in NGS technologies enable ChIP-seq to be conducted at greater genome coverage at lower price, and recover weaker binding events.

As described in Ho's et.al review [24] there are major differences between these two technologies when compared.

From the Data Management and genome alignment view, ChIP-chip is a) a relatively easy and flexible technology, because one microarray corresponds to less than 1 Gigabit data and the coordinates of each probes are known from the beginning, b) more agile when it comes to submitting the data in a database, c) easy to store in your computer and easy to run different algorithms for analysis. On the other hand, ChIP-seq is a relatively hard method because a) raw data and analysis generate approximately more than one terrabyte data, b) FTP or HTTP protocols must be developed to submit the data to a database, c) large capacity storage servers must be established and the analysis must be done with very powerful computers for computing and memory. Moreover, there are challenges in data analysis when these two technologies used to identify peaks in the same study. A comparison of features for the peaks identification are shown in (Figure 10***:*** *Comparison of peak identification features with ChIP-chip and ChIP-seq analysis*). A biological experiment comparison is shown in (Figure 11***:*** *Biological Experiment Comparison ChIP-chip VS ChIP-seq*) from previous studies. Further and more thorough analysis of this biological experiment comparison is presented with figures and statistical data in Ho's et.al study [24].

	ChIP–chip	ChIP–seq
Maximum resolution	Array-specific, generally 30–100 bp	Single nucleotide
Coverage	Limited by sequences on the array; repetitive regions are usually masked out	Limited only by alignability of reads to the genome; increases with read length; many repetitive regions can be covered
Cost	US$400–800 per array (1–6 million probes); multiple arrays may be needed for large genomes	Currently US$1,000–2,000 per lane (using the Illumina Genome Analyzer); 6–15 million reads before alignment
Source of platform noise	Cross-hybridization between probes and nonspecific targets	Some GC bias can be present
Experimental design	Single- or double-channel, depending on the platform	Single channel
Cost-effective cases	Profiling of selected regions; when a large fraction of the genome is enriched for the modification or protein of interest (broad binding)	Large genomes; when a small fraction of the genome is enriched for the modification or protein of interest (sharp binding)
Required amount of ChIP DNA	High (a few micrograms)	Low (10–50 ng)
Dynamic range	Lower detection limit; saturation at high signal	Not limited
Amplification	More required	Less required; single-molecule sequencing without amplification is available
Multiplexing	Not possible	Possible

Figure 10: Comparison of peak identification features with ChIP-chip and ChIP-seq analysis

Studies	Findings
IP: NRSF **Evaluation:** motif occurrence **Peak caller:** CisGenome **Ref:** (Ji et al. 2008)	1. Clear global correlation 2. Peaks in ChIP-chip are wider 3. ChIP-seq peaks are more likely to contains conserved NRSF motifs, and therefore more likely to be true positives
IP: PolII and STAT1 **Evaluation:** Comparison with ChIP-PCR of selected genomic regions. **Peak caller:** PeakSeq **Ref:** (Rozowsky et al. 2009)	1. ChIP-seq generates fewer false positives 2. ChIP-seq generates more peaks 3. ChIP-seq peaks are generally closer to the binding site motif 4. Findings consistent with (Robertson et al. 2007)
***IP:** FOXA1* ***Evaluation:** Motif occurrence* ***Peak caller:** MACS* ***Ref:** (Zhang et al. 2008)*	1. ChIP-chip has more binding sites at 1% FDR 2. Many false negative in ChIP-chip is due to lack of microarray probes. 3. ChIP-chip peaks with higher enrichment are more likely to be discovered by ChIP-seq. 4. Average peak width in ChIP-chip is twice of that of ChIP-seq peaks. 5. ChIP-seq peaks are localized to the sequence motif.

Figure 11: Biological Experiment Comparison ChIP-chip VS ChIP-seq

One more advantage of ChIP-seq is that this method is independent from hybridization artefacts such as dye effects, tiling resolution and the influence of GC-content of the oligoprobe. Another advantage is the fact that ChIP-seq is not dependent on a defined array-design, for example heterochromatic regions that are generally not represented on microarrays. Although ChIP-seq does not also allow the analysis of repetitive elements, unique regions within heterochromatin can be analyzed, e.g. to give some information of replication initiation of this part of the chromatin. A major disadvantage is that 100–150bp fragments are preferentially sequenced, which enhances the bias towards small DNA fragments and is the reason for large complexes and structures not being efficiently sequenced [49].

To summarize, ChIP-seq brought a new era in genome analysis because it offers high resolution, a more distinctive signal especially to noise ratio and a detection of more peaks compared to ChIP-chip, along with the ongoing cost reduction ChIP-seq technique is becoming more and more dominant in the study of transcriptional regulatory pathways and networks. However, because ChIP-seq datasets are massive and complex, their analysis requires advanced statistical methods, efficient computational algorithms and user-friendly software for processing and visualization. Nowadays, ChIP-Seq is increasingly being used and preferred over its predecessor for mapping protein–DNA interactions in-vivo on a genome scale.

3.4 ChIP-Seq Pipeline Steps

Several studies [50] [45] [44] [51] [52] in the literature describe the pipeline of ChIP-seq data analysis which can be summarized in the following five key steps: 1) Map the reads to a reference genome – the goal in this step is to identify for each short read in the dataset, all the locations in a reference genome that show perfect or near perfect matches to the read, 2) Background Estimation – in this step all the reads that are unrelated to the binding events of interest can be regarded, as «background reads», 3)Peak Calling – the peak regions and their significance are identified 4) Gene Assignment and Peak Annotation – after obtaining a list of peak coordinates, it is important to study the biological implications of the protein-DNA bindings and associate each peak to its nearest gene and 5) De Novo Motif Analysis – in this last step the binding motifs are recovered from the peak sequences as well as from their orthologous sequences.

Lastly, the basic steps of the ChIP–seq assay have been also reviewed elsewhere [51] and were summarized for transcription factors and for histone modifications.

3.5 Methodologies coupling ChIP-seq & Gene regulatory networks

ChIP-seq is the most direct way to identify the binding sites of a single DNA-binding protein or the locations of modified histones [51]. Also, biological networks provide a comprehensive overview of biological systems [53]. They enable better understanding of the system and can shed light on the function of genes and other molecular compounds. Among other applications, they have been used for discovery and prediction of gene interactions, gene functions and disease–gene associations [53]. Such biological networks are the GRNs. So, coupling these two together, ChIP-seq technology and GRNs can create a very powerful tool in the hands of biologists for the decryption of the human genome and the prevention of diseases.

To mine gene expression data sets effectively, analysis frameworks need to incorporate methods that identify intergenic relationships within enriched biologically relevant sub pathways [54]. Elucidating gene regulatory network (GRNs) from large scale experimental data remains a central challenge in systems biology [55]. The advent of high-throughput data generation technologies has allowed researchers to fit theoretical models to experimental data on gene-expression profiles. A numerous of GRNs applications along with their publications can be found in http://omictools.com/gene-regulatory-networks-category.

By searching the literature we found some great projects that provide the bioinformaticians with ChIP-seq data from lab experiments and pathway analysis of these biological data.

Firstly, The Human Genome Project, which is an international scientific research project with the goal of determining the sequence of chemical base pairs which make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and functional standpoint. It remains the world's largest collaborative biological project.

Secondly, KEGG (Kyoto Encyclopedia of Genes and Genomes), which is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances[7]. KEGG is utilized for bioinformatics research and education, including data analysis in genomics, metagenomics, metabolomics and other omics studies, modeling and simulation in systems biology, and translational research in drug development. The KEGG biological pathways are represented with GRNs. However, KEGG doesn’t analyze recent or any ChIP-seq data for biological experiments.

Thirdly, the ENCODE (Encyclopedia of DNA Elements) Consortium, which is an international collaboration of research groups funded by the National Human Genome

---

[7] http://www.kegg.jp/kegg

Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

In addition to these, several programs that have been published for the analysis of ChIP-seq data, they often focus on the peak detection step and are usually not well suited for thorough, integrative analysis of the detected peaks and furthermore they do not infer a Gene Regulatory Network which its biological significance have already been discussed. Currently, there are many software tools implementing different approaches to identify TFBSs within ChIP-Seq peaks.

TEAC, a Topology Enrichment Analysis Framework for detecting activated biological sub-pathways requires as input a file with gene expression data and standard pathways from KEGG and returns a set of activated ranked sub-pathways. However, it has some limitations - it takes only the KEGG pathways and does not extend beyond transcriptional analyses of the events underlying yeast cellular responses to nitrogen stress [54].

Another web-based tool for pathway analysis on ChIP-seq data is rTRM [56], which is a web tool for predicting transcriptional regulatory modules for ChIP-seqed transcription factors. It requires to choose organism (human or mouse), a transcription factor, two files in a specific form with the gene data and some network and plot parameters and it returns a gene path as shown in Figure 12***: rTRM a web tool for transcriptional regulatory modules***.

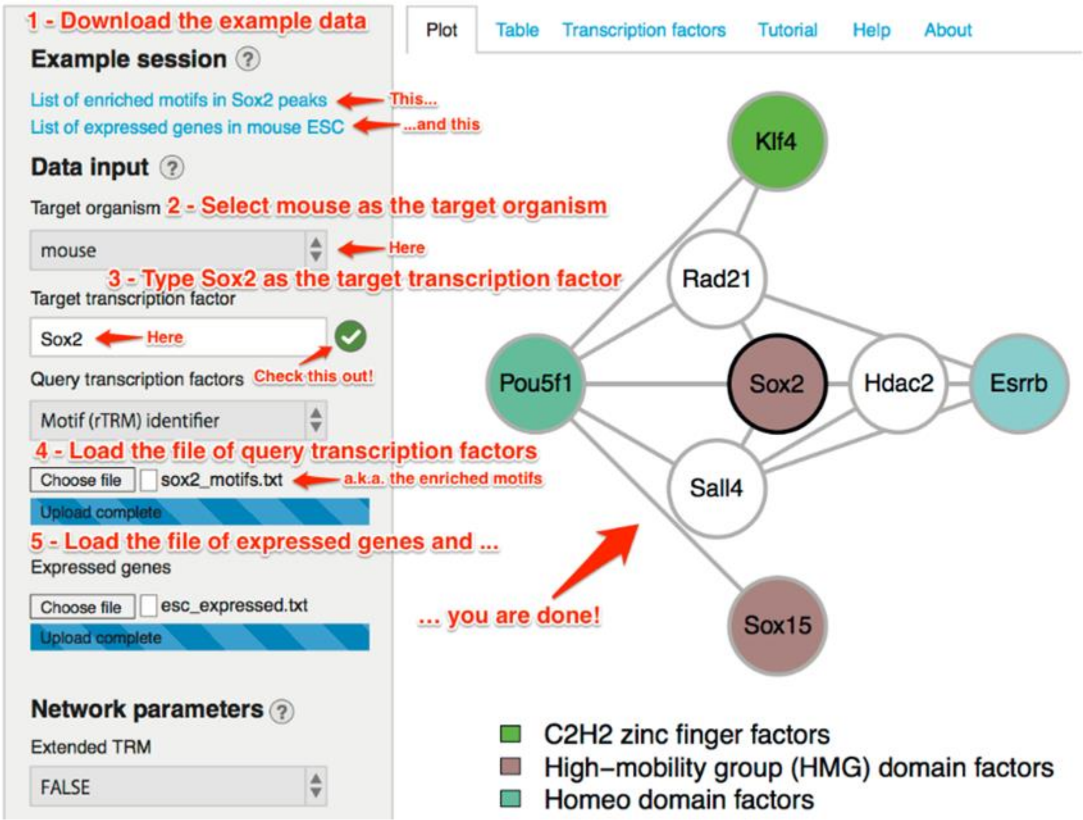


Figure 12: rTRM a web tool for transcriptional regulatory modules

However it doesn't provide a clear and thorough view of all the biological pathways and the interactions between the genes that are expressed in a regulation procedure. Moreover, an enriched motifs file from ChIP-seqed regions and already analyzed ChIP-seq data with the expressed genes, which participate in the regulation given a specific transcription factor, is needed.

CMGRN (Constructing Multilevel Gene Regulatory Networks) [16], an integrative web server to unravel hierarchical interactive networks at different regulatory levels. The developed method used the Bayesian network modeling to infer causal interrelationships among transcription factors or epigenetic modifications by using ChIP-seq data. Moreover, CMGRN used Bayesian hierarchical model with Gibbs sampling to incorporate binding signals of these regulators and gene expression profile together for reconstructing gene regulatory networks.

Although, CMGRN is an easy-to-use bioinformatics tool to interpret ChIP-seq high-throughput data, the user has to provide two files with a specific format, regulatory signal of TFs, epigenetic modifications or microRNAs file and a gene expression data file. A final network exploring interactions of both regulator-regulator and regulator–gene will be presented to the user (Figure 13).

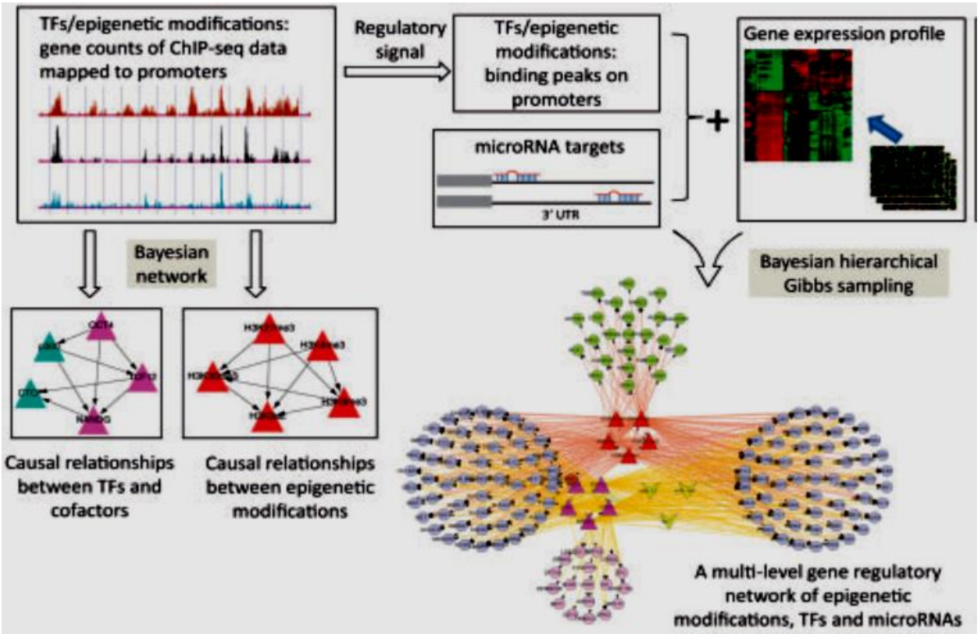


Figure 13: An overview of data-integrated analysis and regulatory network construction. The triangle nodes represent TFs (pink), epigenetic regulators (red) and cofactors (blue), whereas the V-type nodes represent microRNAs (yellow). Target genes were indicated by using the circle nodes *[16]*

Wang and Qin in their work [57], enhanced ChIP-Array [58] and created ChIP-Array 2 which is able to incorporate ChIP-X and transcriptome data, long-range chromatin interaction, open chromatin region and histone modification data to construct GRNs (Figure 14***:*** *Workflow (A) and results (B) of ChIP-Array 2. (A) Direct targets are identified by combining ChIP-X and transcriptome data. Interplays between the TF of interest and other regulatory factors/target genes are supported by other omics data. Then indirect targets are detected by curated ChIP-X data or predicted TFBSs with the assistance of other omics data. (B) The results are composed of four parts: the resulting GRN shown in CytoscapeWeb, motif enrichment analysis by MEME Suite, functional enrichment analysis, and visualization in JBrowse*).

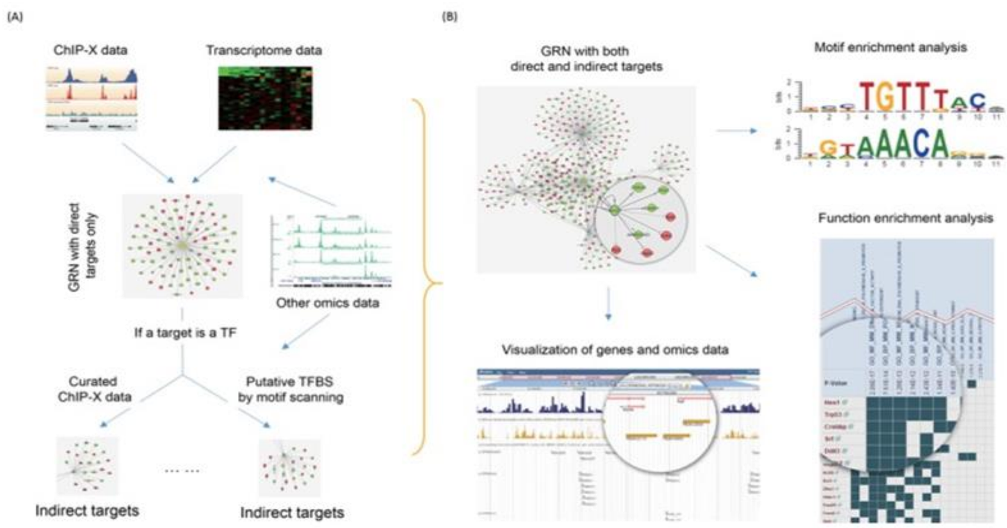


Figure 14: Workflow (A) and results (B) of ChIP-Array 2. (A) Direct targets are identified by combining ChIP-X and transcriptome data. Interplays between the TF of interest and other regulatory factors/target genes are supported by other omics data. Then indirect targets are detected by curated ChIP-X data or predicted TFBSs with the assistance of other omics data. (B) The results are composed of four parts: the resulting GRN shown in CytoscapeWeb, motif enrichment analysis by MEME Suite, functional enrichment analysis, and visualization in JBrowse

Although the existing web-based software can predict potential target genes of the multiple regulators, they are not able to discover hierarchical organizations formed by cross-interactions between the regulators and genes simultaneously. Most programs are run from the command line and require variable degrees of data formatting and computation expertise to implement and they are very complex [48]. Currently, there is a lack of effective web resources to generate the topology of networks controlled by the interacting factors at transcriptional, post transcriptional and epigenetic layers [16].

4 Technical Implementation

The labor in this Master Thesis is divided into three phases. In the first phase, we download programmatically ChIP-seq peak files from the ENCODE ChIP-seq Experiment Matrix according to a user's query and analyze them using Bioconductor's packages for ChIP-seq Analysis. So, we created a user-friendly interface (shinny app) in the statistically programming language R using Shiny Studio that gets users selection, analyzes the equivalent ChIP-seq peak files obtained from ENCODE Project and creates a file which has five (5) columns: the gene id's (EntrezID), the p-value or the q-value equivalent of what is given for each file, the score, the signal value and the gene name. EntrezID is the coding system the KEGG database uses for the genes identification. Score denotes indicates how dark the peak will be displayed in a browser (0-1000). Signal value is the measurement of overall (usually, average) enrichment for the region, which is the average (between replicates) read count over the region. If the signal value is high, it means that a lot of chromatin from that region is pulled down by the IP and sequenced.

In detail, the application chooses the optimal IDR (Irreproducible Discovery Rate) thresholded peak files (IDR<0.01). There are many file types to choose from the ENCODE Experiment ChIP-seq Matrix. For the needs of this work all the peak files with format type bed, bed narrowPeak,bed broadPeak, broadPeak and narrowPeak were chosen. We choose all the binding sites near TSS and then we obtain peaks within 5kb upstream of TSS within the gene because transcription factors are proteins that bind to DNA, typically upstream from and close to the transcription start site of a gene, and regulate the expression of that gene by activating or inhibiting the transcription machinery [59]. We threshold the qvalue of the peaks and we choose only the peaks that have FDR or qvalue<0.01 because 1% FDR is the most commonly accepted value for peaks of good quality [17]. We map the genes with their corresponded Entrez Id's using EnsDb.Hsapiens.v75 a Bioconductor Mapping Library and we get the selected TF's binding sites in that cell line. We save the mapped genes in a txt file.

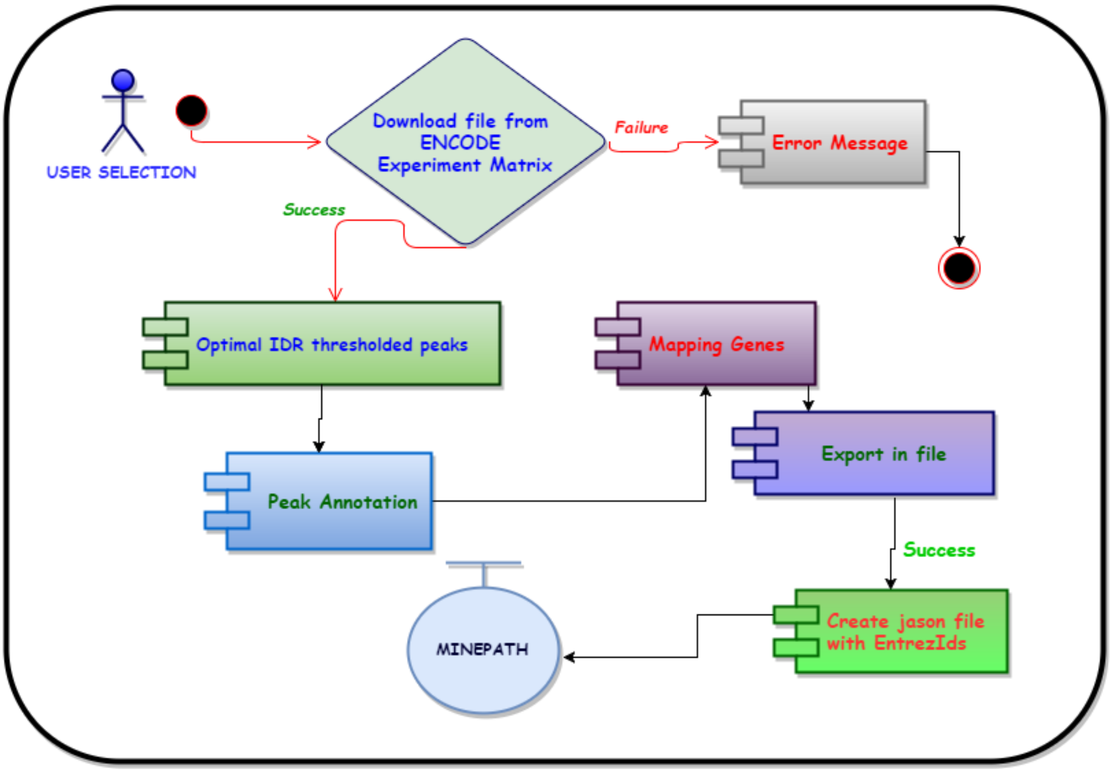
In the second phase of this work, GRN visualization tool called MinePath, its extended version was used to accept and analyze ChIP-seq peak data files. The app, also, gives the ability to user to find the overlapping peaks of two different peak files. Lastly, in the third phase, we conducted many experiments with specific antibody target in a human cell type of various cancer phenotype breast cancer, glioma, e.t.c.

4.1 Bioconductor and R – Shiny Studio

Bioconductor [60] is a free open source software for bioinformatics, meaning that all developers from the scientific community are able to contribute software. It provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the statistical programming language R, which facilitates data visualization and downstream analysis for the statistically - inclined user, and it has two releases each year, 1211 software packages and an active user community. The latest version is 3.3.1, which is the version that was used in this master thesis.

We developed an application using Bioconductor packages through Shiny[8], a web application framework for R powered by RStudio. Shiny is an R package which uses a reactive programming model to simplify the development of R-powered web applications. We used RStudio platform to develop the application and FORTH Institute's server to host the application and run it. The application is written in R language and it downloads ChIP-seq data from the ENCODE project according to the users selection using queries and filters on ENCODE data.

The data retrieved are described below in Data Library section of this Master Thesis. The main functionality of this Shiny application is to analyze a ChIP-seq peak file and map the biological data retrieved on the Human Genome reference sequence version GRCh38 by using a mapping function. The output of the Shiny application is a file with all the genes that form the TFBSs of the specific experiment, which will be visualized along with their gene expression data in the extended version of MinePath [18]. An overview of the pipeline implemented in this Master Thesis is shown in (Figure 15: *Diagram of the* . We have also added one more functionality in the Shiny application that allows the user to intersect two different ChIP-seq file peaks and receive a file with all the overlapping peaks between these files for further differential analysis study.



[8] http://shiny.rstudio.com/

Figure 15: Diagram of the pipeline

4.2 Libraries and functions

The most important Libraries and Functions used for the shiny application development are described below (Figure 16***:*** *Flowchart of Shiny RStudio app R packages*):

1) The ENCODExplorer Library. This package allows user to quickly access ENCODE project files metadata and give access to helper functions to query the ENCODE rest api, download ENCODE datasets and save the database in SQLite format. A reference manual can be downloaded from here[9].

2) The function searchEncode was used from the ENCODExplorer Library and users' choices were passed as input in a query form. Some errors and Null restrictions overcame with the use of tryCatch function which provides a mechanism for handling unusual conditions, including errors and warnings in R, more can be found here[10].

3) The downloadEncode function from the same Library was used to download the proper file with the peaks in BED format.

4) The function annotatePeakInBatch from ChIPpeakAnno Bioconductor package was used to annotate peaks by annoGR object in the given range for the list of peaks from the BED file already retrieved from ENCODE Project database. We used it to obtain the peaks within 5kb distance of TSS, upstream of the gene body. We have chosen these genes because the promoters of a gene can be found upstream of the gene [61] [62].

5) The library (TSS.human.NCBI36): This library has TSS Annotation for Human Sapiens (NCBI36) obtained from BiomaRt. The TSS.human.NCBI36 exposes an annotation database generated from NCBI. It's the latest and biggest from the view of genes updated database[11].

6) The library (org.Hs.eg.db) is an R object that contains mappings between Entrez Gene identifiers and GenBank accession numbers. This object is a simple mapping of Entrez Gene identifiers[12] to all possible GenBank accession numbers. Mappings

---

[9] https://www.bioconductor.org/packages /3.3/bioc/manuals/ENCODExplorer/
[10] http://www.inside-r.org/r- doc/base/signalCondition
[11] Different annotation knowledge databases in http://ccb.jhu.edu/software/tophat/igenomes.shtml
[12] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

were based on data provided by: Entrez Gene Data[13] with a date stamp from the source of 14 March 2016. A reference manual can be found here[14].

7) The addGeneIDs function was used to add common IDs to annotated peaks such as gene symbol, entrez ID, ensemble gene id and refseq id leveraging organism annotation dataset.

8) The biomaRt library is an R package that enables retrieval of large amounts of data in a uniform way without the need to know the underlying database schemas or write complex SQL queries. Examples of BioMart databases are Ensembl,COSMIC, Uniprot, HGNC, Gramene, Wormbase and dbSNP mapped to Ensembl. These major databases give biomaRt users direct access to a diverse set of data and enable a wide range of powerful online queries from gene annotation to database mining. A reference manual can be downloaded from here[15].

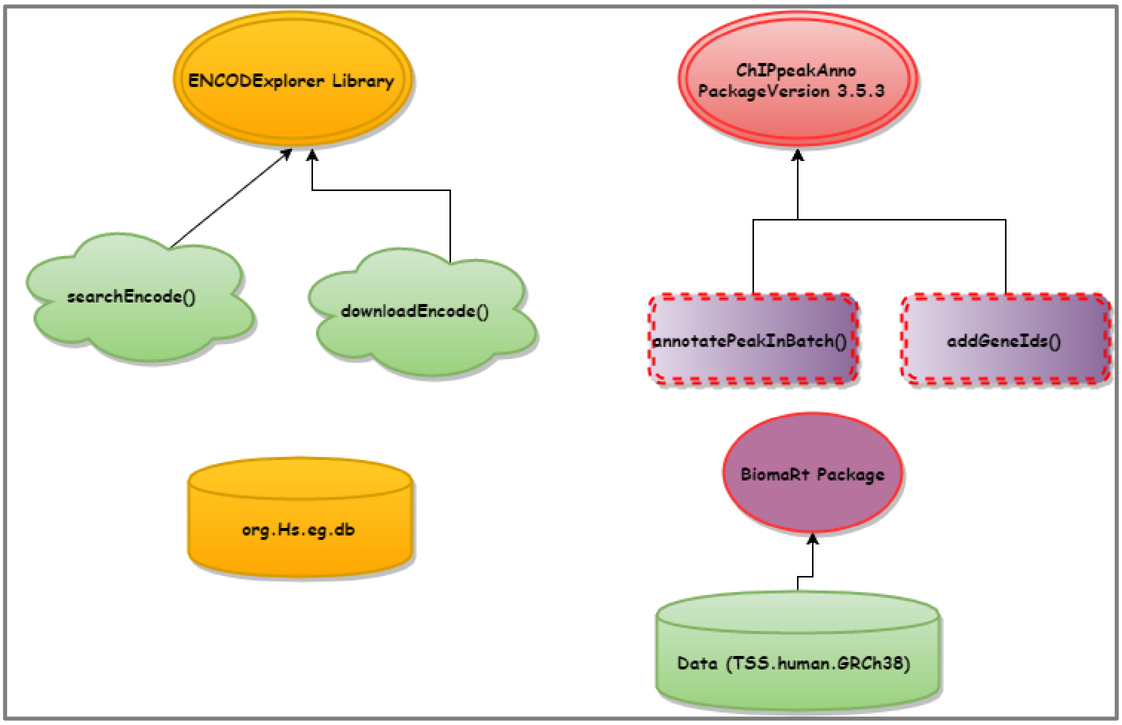


Figure 16: Flowchart of Shiny RStudio app R packages

9) The ChIPpeakAnno package Version 3.5.3 was used from the Shiny application to annotate the peak files retrieved from the ENCODE project official site. This package provides Batch annotation of the peaks identified from either ChIP-seq, ChIP-chip experiments or any experiments resulted in large number of chromosome ranges. The

[13] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
[14] https://bioconductor.org /packages/release/data/annotation/manuals/org.Hs.eg.db/
[15] https://bioconductor.org/packages/release/bioc/manuals/biomaRt /man/biomaRt.pdf

package includes functions to retrieve the sequences around the peak, obtain enriched Gene Ontology (GO) terms, and find the nearest gene, exon, miRNA or custom features such as most conserved elements and other transcription factor binding sites supplied by users. Starting 2.0.5, new functions have been added for finding the peaks with bi-directional promoters with summary statistics (peaksNearBDP), for summarizing the R topics documented: occurrence of motifs in peaks (summarizePatternInPeaks) and for adding other IDs to annotated peaks or enrichedGO (addGeneIDs).This package leverages the biomaRt, IRanges, Biostrings,BSgenome, GO.db, multtest and stat packages [61]. A reference manual can be downloaded from here[16].

4.3 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

The ENCODE Consortium has adopted uniform guidelines for the most common ENCODE experiments. The guidelines have evolved over time, as technologies have changed. The current guidelines are informed by results gathered during the ENCODE project[17].

All ChIP-seq experiments were performed at least in duplicate, and were scored against an appropriate control designated by the production groups (either input DNA or DNA obtained from a control immunoprecipitation). All ENCODE ChIP experiments follow guidelines and specific standards. ENCODE guidelines for antibody and immunoprecipitation characterization are described in detail in the paper the ENCODE Consortium published in 2012 [63]. A quick updated guide can be found here[18].

4.3.1 Replication, sequencing depth, library complexity, and reproducibility

Biological replicate experiments from independent cell cultures, embryo pools, or tissue samples are used to assess reproducibility. Initial RNA polymerase II ChIP-seq experiments showed that more than two replicates did not significantly improve site discovery [64]. Thus, the ENCODE Consortium set as standards that all ChIP

[16] https://www.bioconductor.org/packages/3.3/bioc/manuals/ChIPpeakAnno/man/ChIPpeakAnno.pdf
[17] Here http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001046#s5 there is a detailed review for the ENCODE Data Analysis
[18] https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ChIP_DNase_FAIRE_DNAme_v2_2011.pdf

measurements would be performed on two independent biological replicates. The IDR analysis methodology [65] is used to assess replicate agreement and set thresholds. For experiments with poor values for quality metrics, additional replicate(s) have been generated. Briefly, ENCODE produces replicate data for most experiments to quantify reliability. Biological replicates involve different biological samples, e.g., different tissue preparations for cell growth and expansion when cell lines are used. Biological replicates are contrasted with technical replicates, for which different sequencing libraries are prepared from the same sample, or different sequencing lanes for the same library. Reads from different replicates are stored in separate files and should include flow cell and lane ID. If multiple lanes are used for the same biological or technical replicate, they are stored in the same file (after a QC check to eliminate failed lanes), with information on flow cell and lane ID included. For experiments that produce paired-end reads, the two reads in each pair are stored in two separate files, with the reads in the same order in the two files.

A few of the ENCODE ChIP experimental design guidelines for replication, sequencing depth, library complexity, and reproducibility are [63]:

Sequencing and library complexity

For each ChIP-seq point-source library, ENCODE's goal is to obtain ≥10 million uniquely mapping reads per replicate experiment for mammalian genomes, with a target NRF (nonredundancy fraction) ≥0.8 for 10 million reads. The corresponding objective for modENCODE point-source factors is to obtain ≥2 M uniquely mapped reads per replicate, ≥0.8 NRF. The modENCODE target for broad-source ChIP-seq in Drosophila is ≥5 million reads, and the ENCODE provisional target for mammalian broad-source histone marks is ≥20 million uniquely mapping reads at NRF ≥0.8. *The distribution of NRF values for all ENCODE data sets is shown in Figure 17.*

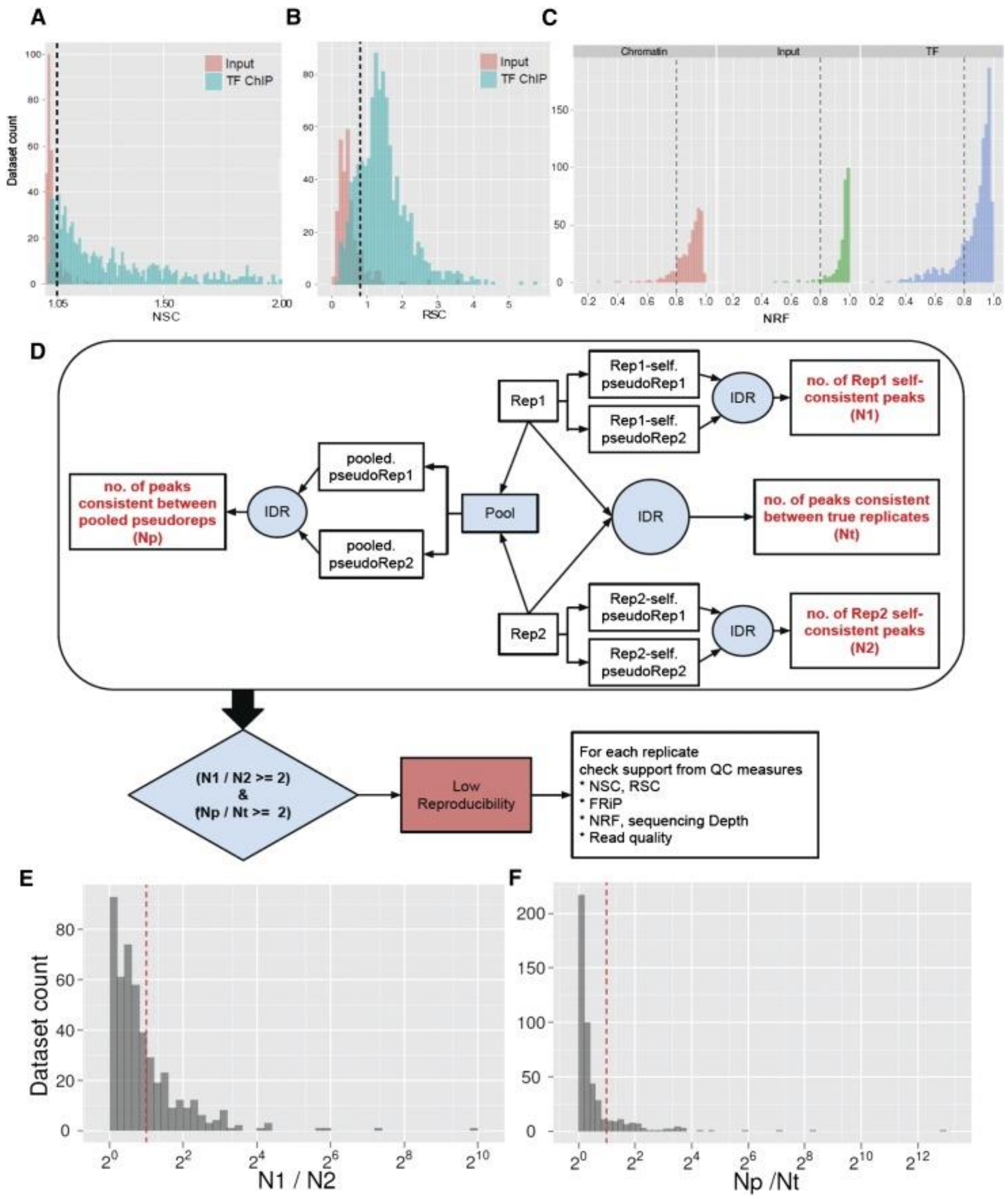


Figure 17: Analysis of ENCODE data sets using the quality control guidelines *[63]*

Control libraries

<u>ENCODE generates and sequences a control ChIP library for each cell type, tissue, or embryo collection and sequences the library to the appropriate depth (i.e., at least equal to, and preferably greater than, the most deeply sequenced experimental library).</u> If cost constraints allow, a control library should be prepared from every chromatin preparation and sonication batch, although some circumstances can justify fewer control libraries. Importantly, a new control is always performed if the culture

conditions, treatments, chromatin shearing protocol, or instrumentation is significantly modified.

Reproducibility

Experiments are performed at least twice to ensure reproducibility. For ENCODE data to pass criteria for submission, concordance is determined from analysis using the IDR methodology [63]*, and a third replicate is performed if the standard is not reached* (Figure 17, D). Cut-offs for identifying highly reproducible peaks for use in subsequent analyses *can be determined by IDR (typically using a 1% threshold).*

A practical goal is to maximize site discovery by optimizing immunoprecipitation and sequencing deeply, within reasonable expense constraints. For point-source factors in mammalian cells, a minimum of 10 million uniquely mapped reads are used by ENCODE for each biological replicate (providing a minimum of 20 million uniquely mapped reads per factor. For broad areas of enrichment, the appropriate number of uniquely mapped reads is currently under investigation, but at least 20 million uniquely mapped reads per replicate for mammalian cells is currently being produced for most experiments.

Within ENCODE, a set of data quality thresholds has been established for submission of ChIP-seq data sets. These have been constructed based on the historical experiences of ENCODE ChIP-seq data production groups with the purpose of balancing data quality with practical attainability and are routinely revised. The current standards are below and the performance of ENCODE data sets against these thresholds is shown in Figure 17. So, A few of ChIP-seq quality assessment guidelines are:

Cross-correlation analysis

The current ENCODE practice is to calculate and report NSC and RSC for each experiment. For experiments with NSC values below 1.05 and RSC values below 0.8, they currently recommend that an additional replicate be attempted or the experiment explained in the data submission as adequate based on additional considerations.

Irreproducible discovery rate (IDR)

The following guidelines have been established for mammalian cells (optimal parameter may differ for other organisms). *Biological replicates are performed for each ChIP-seq data set and subjected to peak calling. IDR analysis is then performed with a 1% threshold. For submission to ENCODE, they currently require that the number of bound regions identified in an IDR comparison between replicates to be at least 50% of the number of regions identified in an IDR comparison between two* *"pseudoreplicates" generated by pooling and then randomly partitioning all available reads from all replicates (Np/Nt < 2). To ensure similar weighting of individual replicates for identifying binding regions, they further recommend that the number of significant peaks identified using IDR on each individual replicate (obtained by partitioning reads into two equal groups for the IDR analysis) be within a factor of 2 of one another (N1/N2 < 2)(Figure 17). Data sets which fail to meet these criteria may still be deposited by ENCODE experimenters, provided that at least three experimental replicates have been attempted and a note accompanies these data sets explaining which parameters fail to meet the standards and providing any technical information that may explain this failure. This guideline is for point source features; metrics are still being determined for broad peak analyses.*

Updated information about the performance of ENCODE data sets against these quality metrics and tools for determining these metrics will be forthcoming through the ENCODE portal[19].

A simpler heuristic for establishing reproducibility was previously used as a standard for depositing ENCODE data and was in effect when much of the currently available data was submitted. According to this standard, either 80% of the top 40% of the targets identified from one replicate using an acceptable scoring method should overlap the list of targets from the other replicate, or target lists scored using all available reads from each replicate should share more than 75% of targets in common. As with the current standards, this was developed based on experience with accumulated ENCODE ChIP-seq data, even though with a much smaller sample size.

4.3.2 Peak Calling

Since every ENCODE dataset is represented by at least two biological replicate experiments, a novel measure of consistency and reproducibility of peak calling results between replicates, known as the Irreproducible Discovery Rate (IDR), was used to determine an optimal number of reproducible peaks. Code and detailed step-by-step instructions to call peaks using the IDR method are available[20]. In general, two peak callers were used for the analysis of the enriched regions that were identified from the ChIP-seq analysis, SPP caller and MACS [66] peak caller. They used the MACS peak

[19] http://encodeproject.org/ENCODE/
[20] https://sites.google.com/site/anshulkundaje/projects/idr

caller to identify regions of enrichment over a wide range of signal strength[21]. Enriched regions were scored on individual replicates, pooled data (reads pooled across replicates) and on subsampled pseudoreplicates (obtained by pooling reads from all replicates and randomly subsampling, without replacement, two pseudoreplicates with half the total number of pooled reads).

They also, used MACS2 to identify three types of regions of enrichment: (i) narrow peaks of contiguous enrichment (narrowPeaks) that pass a Poisson *p*-value threshold of 0.01; (ii) broader regions of enrichment (**broadPeaks**) that pass a Poisson *p*-value threshold of 0.1 (using MACS2's broad peak mode); (iii) gapped/chained regions of enrichment (**gappedPeaks**) defined as broadPeaks that contain at least one strong narrowPeak.

In order to obtain reliable regions of enrichment, they restricted to enriched regions identified using pooled data that were also independently identified in both pseudoreplicates. The coverage and conservation analysis only used histone modification datasets from the Broad Institute Production group. They used the gappedPeak representation for the histone marks with relatively compact enrichment patterns. These include H3K4me3, H3K4me2, H3K4me1, H3K9ac, H3K27ac and H2AFZ.

SPP caller was also used by the labs the ENCODE takes the ChIP-seq data. In brief, the SPP peak caller [67] was used with a relaxed peak calling threshold (FDR = 0.9) to obtain a large number of peaks (maximum of 300K) that span true signal as well as noise (false identifications). The IDR method analyzes a pair of replicates, and considers peaks that are present in both replicates to belong to one of two populations: a reproducible signal group or an irreproducible noise group. Peaks from the reproducible group are expected to show relatively higher ranks (ranked based on signal scores) and stronger rank-consistency across the replicates, relative to peaks in the irreproducible groups. Based on these assumptions, a two-component probabilistic copula-mixture model is used to fit the bivariate peak rank distributions from the pairs of replicates. The method adaptively learns the degree of peak-rank consistency in the signal component and the proportion of peaks belonging to each component. The model can then be used to infer an IDR score for every peak that is found in both replicates. The IDR score of a peak represents the expected probability that the peak belongs to

[21] https://sites.google.com/site/anshulkundaje/projects/encodehistonemods

the noise component, and is based on its ranks in the two replicates. Hence, low IDR scores represent high-confidence peaks. An IDR score threshold of 0.02 (2%) was used to obtain an optimal peak rank threshold on the replicate peak sets (cross-replicate threshold). If a dataset had more than two replicates, all pairs of replicates were analyzed using the IDR method. The maximum peak rank threshold across all pairwise analyses was used as the final cross-replicate peak rank threshold. Reads from replicate datasets were then pooled and SPP was once again used to call peaks on the pooled data with a relaxed FDR of 0.9. Pooled-data peaks were once again ranked by signal-score. The cross-replicate rank threshold learned from the replicates was used to threshold the ranked set of pooled-data peaks.

Any thresholds based on reproducibility of peak calling between biological replicates are bounded by the quality and enrichment of the worst replicate. Valuable signal is lost in cases for which a dataset has one replicate that is significantly worse in data quality than another replicate. A rescue pipeline was used for such cases in order to balance data quality between a set of replicates. Mapped reads were pooled across all replicates of a dataset, and then randomly sampled (without replacement) to generate two pseudo-replicates with equal numbers of reads. This sampling strategy tends to transfer signal from stronger replicates to the weaker replicates, thereby balancing cross-replicate data quality and sequencing depth. These pseudo-replicates were then processed using the IDR method in order to learn a rescue threshold. For datasets with comparable replicates (based on independent measures of data quality), the rescue threshold and cross-replicate thresholds were found to be very similar. However, for datasets with replicates of differing data quality, the rescue thresholds were often higher than the cross-replicate thresholds, and were able to capture true peaks that showed statistically significant and visually compelling ChIP-seq signal in one replicate but not in the other. Ultimately, for each dataset, the best of the cross-replicate and rescue thresholds were used to obtain a final consolidated optimal set of peaks.

All peak sets were then screened against a specially curated empirical blacklist of regions in the human genome and peaks overlapping the blacklisted regions were discarded[22]. Briefly, these artifact regions typically show the following characteristics:

---

[22] A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep

- Unstructured and extreme artefactual high signal in sequenced input-DNA and control datasets, as well as open chromatin datasets irrespective of cell type identity.
- An extreme ratio of multi-mapping to unique mapping reads from sequencing experiments.
- Overlap with pathological repeat regions such as centromeric, telomeric and satellite repeats that often have few unique mappable locations interspersed in repeats.

4.4 MinePath - Pathway Analysis tool

MinePath[23], a web-based platform aiming to facilitate and ease the identification and visualization of differentially active paths or subpaths within a GRN, using gene-expression data. The methodology takes advantage of the topology and the underlying regulatory mechanisms of GRNs, including the direction and the type of the engaged interactions (e.g. activation/expression, inhibition). Each GRN sub-path is interpreted according to Kauffman's principles and semantics: (i) the network is a directed graph with genes (inputs and outputs) being the graph nodes and the edges between them representing the causal links between them, i.e., the regulatory reactions(ii) each node can be in one of the two states, 'ON', the gene is expressed or up-regulated (i.e., the respective substance being present) or, 'OFF', the gene is not-expressed or targeted from a specific gene and (iii) time is viewed as proceeding in discrete steps - at each step the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it [18].The extended version [68] of MinePath was used as a visualization tool for the ChIP-seq data we infer from the Shiny application.

4.5 ENCODE Data Library

After a thorough search of ChIP-seq data repositories the ENCODE Project database was chosen [69] for many reasons including integrity of data, the ENCODE's sql database that someone can query for mass download of files matching specific criteria and the strict guidelines and criteria it has for the data uploaded and published by scientists. The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the NHGRI. The goal of

---

[23] www.minepath.org

ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. ENCODE investigators employ a variety of assays and methods to identify functional elements. The discovery and annotation of gene elements is accomplished primarily by sequencing a diverse range of RNA sources, comparative genomics, integrative bioinformatics methods, and human curation. Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing[24].

The files selected for this Master Thesis from ENCODE ChIP-seq Experiment Matrix through queries and filters in ENCODE's database are bed peak files of type broadpeak, narrowPeak, bed and bed files produced with optimal IDR values. All these files are ChIP-seq data from different cell lines targeted with different antibody target (TF or HM).

4.6 ENCODE ChIP-Seq Experiment Matrix

The ENCODE Experiment Matrix is a set of web pages that visually summarize the types of data produced by the ENCODE project during the first production phase (September 2007 until today). The data summarized here is all hosted at UCSC as browser tracks and downloadable files. The grid on the main Experiment Matrix page shows the number of experiments for each cell type/assay pairing. The ChIP-seq Experiment Matrix page provides a more detailed view of the chromatin immunoprecipitation experiment subset, showing experiments by cell type and antibody target. The companion Experiment Summary page lists the number of experiments by assay type alone and may include annotations that are cell-type independent (annotations on the reference genome).

An ENCODE experiment is defined as a biochemical assay and follow-on data analyses performed on a single cell type by a single lab. Data from an experiment is typically displayed in multiple browser tracks that offer different views of the data (e.g.

[24] As retrieved from https://www.encodeproject.org/ on 22nd of May 2016

enrichment signal graph, peak calls) and is available for download in multiple file formats (e.g. sequence alignments in BAM format, signal graph in bigWig format) that support different analysis methods. Data for multiple replicates are included in a single experiment. The information above retrieved from the new official site of ENCODE Project[25] which contains ChIP-seq data from experiments from different research labs. For the purpose of this thesis, all the data from the ChIP-seq Experiment matrix, a sub-matrix of the ENCODE experiment matrix, can be downloaded programmatically with the Shiny application we created. The ENCODE ChIP-seq Experiment matrix has 118 different Human cell lines, 187 transcription factors and 12 Histone modifications. Every cell in the matrix is a biological experiment that is conducted in one human cell line and an antibody target (Histone Modification or Transcription Factor).

5 Experimental Validation

Following the exploratory analysis presented in the previous chapter and in order to verify the findings obtained as well as in evaluating the overall accuracy of our computational pipeline and its methods, we engaged in a subsequent study, employing the required ChIP-seq datasets from Literature, conducting an analysis of the data from ENCODE ChIP-seq Experiment Matrix, cross-correlating them and experimenting with RNA-seq combined with our ChIP-seq data to infer new knowledge.

5.1 Validation based on Literature

Following the technical implementation presented in the previous chapter and in order to validate the overall accuracy of our computational pipeline and its methods we engaged in several experiments. So, we show how we can use ChIP-seq data from specific phenotype and along with their RNA-seq data visualize them and gain insight of specific biological interactions that happen during transcription and translation.

5.1.1 CTCF binding sites in lung cancer cells

At first, we conducted an experiment to show that the binding sites derived from the Shiny application we developed are reliable and can be confirmed. We used the ChIP-seq peak files from Zheng et al paper [70] of transcription factor CTCF in lung cancer

[25] https://www.encodeproject.org/

cell line A549(adenocarcinomic human alveolar basal epithelial cells). CTCF is an 11 Zinc Finger Protein that binds different DNA target sequences and proteins. CTCF is a chromatin binding factor that plays an essential role in oocyte and preimplantation embryo development by activating or repressing transcription. It also, plays a central role in multiple complex genomic processes, including transcription, imprinting and long-range chromatin interactions and sub-nuclear localization. It seems to act as tumor suppressor and plays a critical role in the epigenetic regulation. Among its related pathways are Chromatin Regulation / Acetylation and Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3. Lastly, GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and chromatin binding[26].

The authors of the paper in their experiments, identified 54.642 ChIP-seq enriched regions of CTCF in A549 cell line. The peaks were called using SPP. The set of peaks reproducible were identified based on an irreproducible discovery rate (IDR) of 0.25%. We annotated the peaks using a web based tool from Cistrome, called BETA-minus, that predict the transcription factors direct target genes. We used David annotation tool and we identified 1835 discrete target genes.

We then implemented our pipeline, for the ChIP-seq data our Shiny app downloaded from ENCODE, for CTCF in A549 cell line and found 1943 discrete target genes (Figure 18).

[26] http://www.genecards.org/cgi-bin/carddisp.pl?gene=CTCF

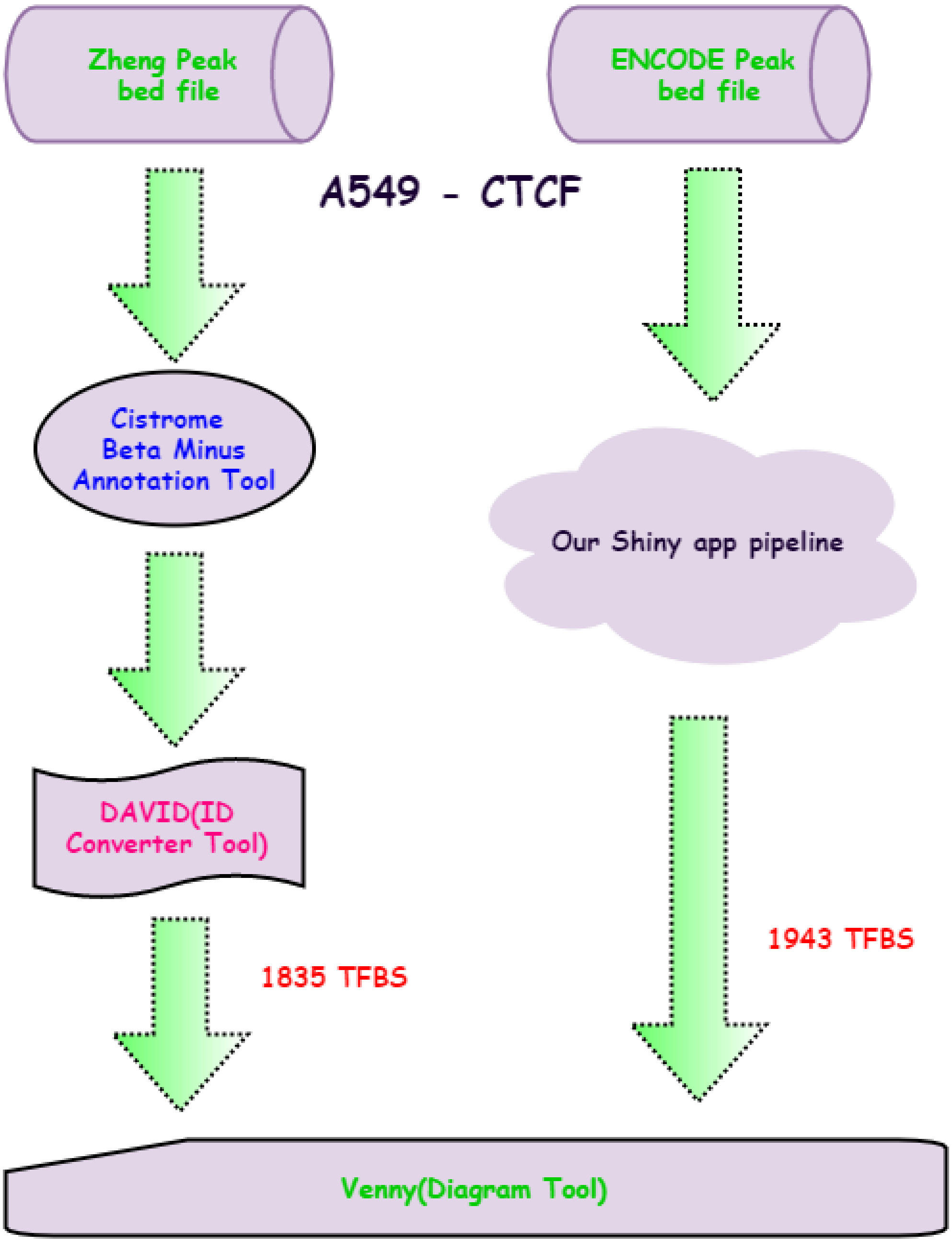


Figure 18 : Validation Experiment for CTCF in A549

Lastly, we cross-correlated the binding sites identified from DAVID and the binding sites we identified using Venny 2.1.0 (a web based tool for drawing venn diagrams) (Figure 19).

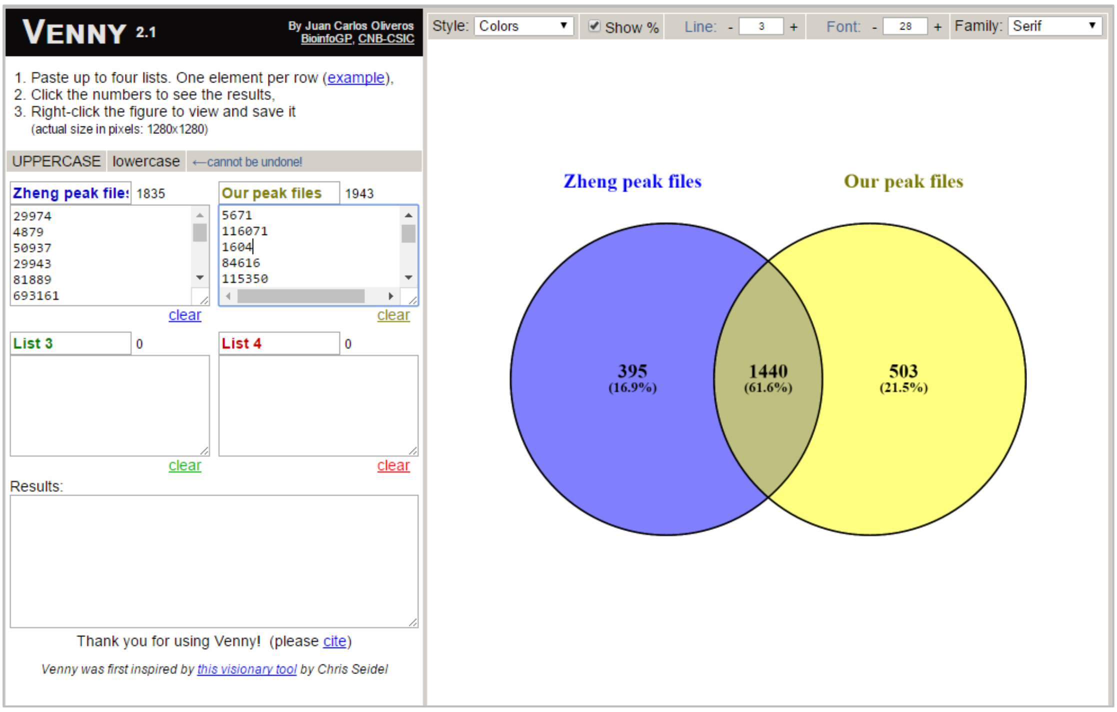


Figure 19 : Venn diagram for overlapping peaks

We found that 1.441 potential target genes were in common, a quiet respective number of overlapping peaks. Our peak calling was done by ENCODE with IDR<0.01 , for that reason and many others including difference in knowledge base at the time of mapping, different conditions in the lab, different thresholds there is small difference in the peaks identified by us and the peaks identified by Gertz. Moreover, different peak callers have their own rationality, and are different in positional accuracy of predicted binding sites [17].

The above results prove the reliability of our Shiny application pipeline.

5.1.2 STAT3 in Glioma

STAT3[27] (Signal Transducer And Activator Of Transcription 3 or Acute-Phase Response Factor) is a protein coding gene. Diseases associated with STAT3 include is autoimmune disease, multisystem, infantile-onset(admio): A disorder characterized by early childhood onset of a spectrum of autoimmune manifestations affecting multiple organs, including insulin-dependent diabetes mellitus and autoimmune enteropathy or celiac disease[28]. Other features include short stature, non-specific dermatitis, hypothyroidism, autoimmune arthritis, and delayed puberty. STAT3 has emerged as a

[27] http://www.genecards.org/cgi-bin/carddisp.pl?gene=STAT3
[28] http://www.malacards.org/card/autoimmune_disease_multisystem_infantile_onset

key initiator and master regulator of mesenchymal transformation in malignant gliomas [71]. Among its related pathways are Endometrial cancer and Adipocytokine signaling pathway. GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and sequence-specific DNA binding. Lastly, this protein mediates the expression of a variety of genes in response to cell stimuli, and thus plays a key role in many cellular processes such as cell growth and apoptosis. It is known that STAT3 is involved in the development of multiple tumors [72].

Glioma is a type of tumor that starts in the brain or spine. It is called a glioma because it arises from glial cells. The most common site of gliomas is the brain [73]. Gliomas make up about 30% of all brain and central nervous system tumors and 80% of all malignant brain tumors[29] [74].

To prove the biological significance of the pathway analysis of the ChIP-seq data, we propose in this master thesis, and that pathway analysis using ChIP-seq data could aid researchers to determine the biological relevance of the binding sites over functional sub-paths and provide insights for new disease treatments, we studied Koumakis et al paper [68] and confirmed the results of their approach by using our pipeline in ChIP-seq data from CHEA (ChIP Enrichment Analysis) database.

They visualized the binding sites from ChIP-seq data of CTCF on patients with Glioma from Zhang's et al paper [75] in MinePath, the same visualization tool for pathway analysis we used for the purposes of this master thesis, and identified that a specific path of glioma patients only can be disrupted. They found a sub-pathway of WNT signaling KEGG pathway for glioma samples which starts from gene WNT16, activates FZD10 and in turn activates DVL1 which is associated with AXIN1 and with CTNNB1 activates the hub gene lymphocyte enhancer factor-1 (LEF1) which continue to the activation of various proteins and the alternation of cell cycle (Figure 20). Their findings were validated from the literature. Xingchun Gao et al [76] and Yanwei Liu et al [77] papers validate that the specific sub-path holds for glioma is considered as one of the key elements for glioblastoma cell proliferation, migration, invasion, and cancer stem-like cell self-renewal.

[29] As retrieved from https://en.wikipedia.org/wiki/Glioma#cite_note-1 on September 17 2016

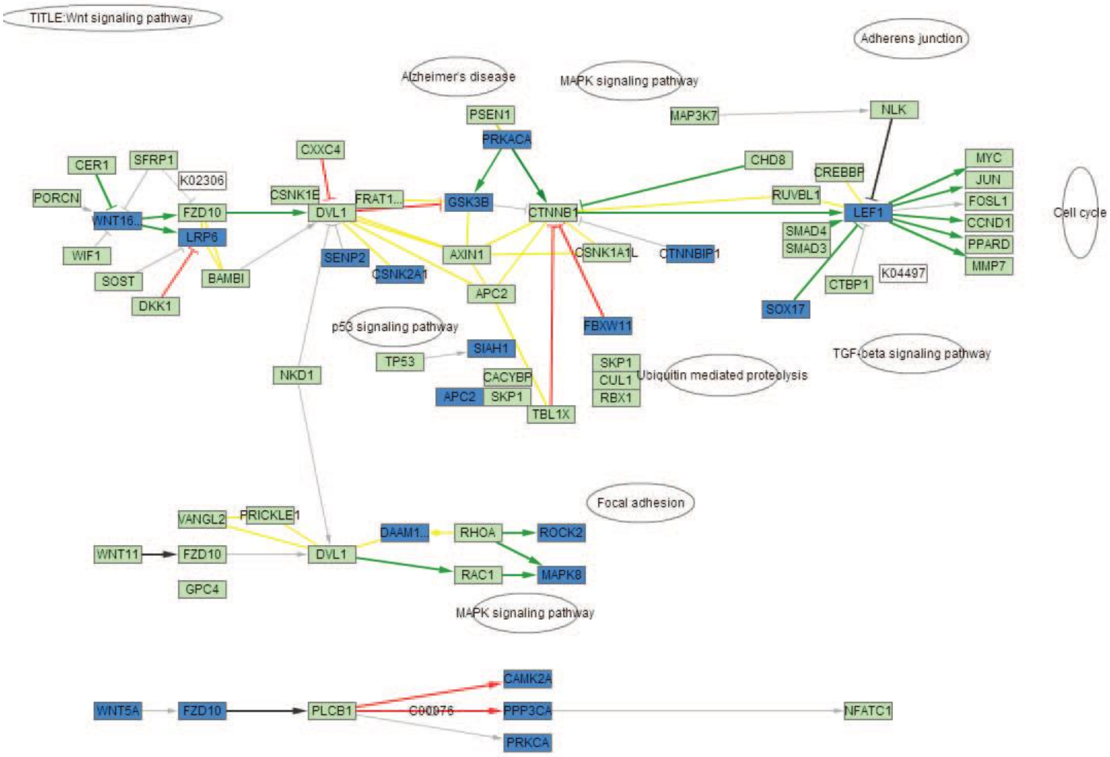


Figure 20 : WNT functional sub-paths and binding sites of STAT3 *[68]*

We searched the harmonizome [78] (a collection of processed datasets gathered to serve and mine knowledge about genes and proteins) and downloaded the bed peak files with the enriched binding regions, of STAT3 in U87 human cells, that Alexander Lachmann et al found and delivered in a web based interactive application called ChIP Enrichment Analysis (ChEA) [79]. We implemented our pipeline in those peaks and we predicted the potential target genes of STAT3.We used the web based Gene Conversion tool AbsIDconvert [80] (Absolute Gene ID Conversion Tools) to convert the gene symbols of the target genes into Entrez Ids so as to load them in MinePath for analysis.

Lastly, we ran a pathway analysis in glioma dataset in MinePath. For the feasibility study we used a microarray dataset, as in Koumakis et al paper, proposed by [81] for glioma and healthy samples. The reference dataset is a merging of two different studies using as classes the glioma cases from the GSE4271 [82] (100 samples) versus the control cases from the GSE1133 [83] (158 samples). We then chose Wnt signaling pathway, the same signaling pathway in Koumakis et al paper, loaded the U87 STAT3 target genes we found with our pipeline in this pathway and we observed the same findings for the gene LEF1 and the sub-pathway it is involved in (Figure 21).

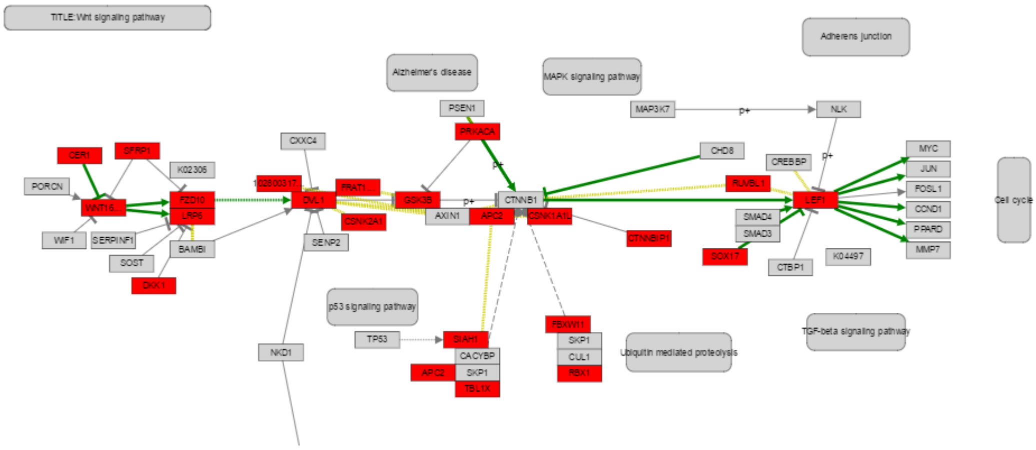


Figure 21 : LEF1 in Wnt signaling pathway for ChIP-seq data of U87 cell line (glioblastoma) with STAT3 (binding sites from ChEA)

To conclude, we validated our pipeline analysis and confirmed it with the one in Koumakis et al paper by using public microarray expression datasets for glioma and the KEGG human GRNs as proof of concept and we identified and confirmed disrupted sub-paths due to STAT3 on functional glioma pathways.

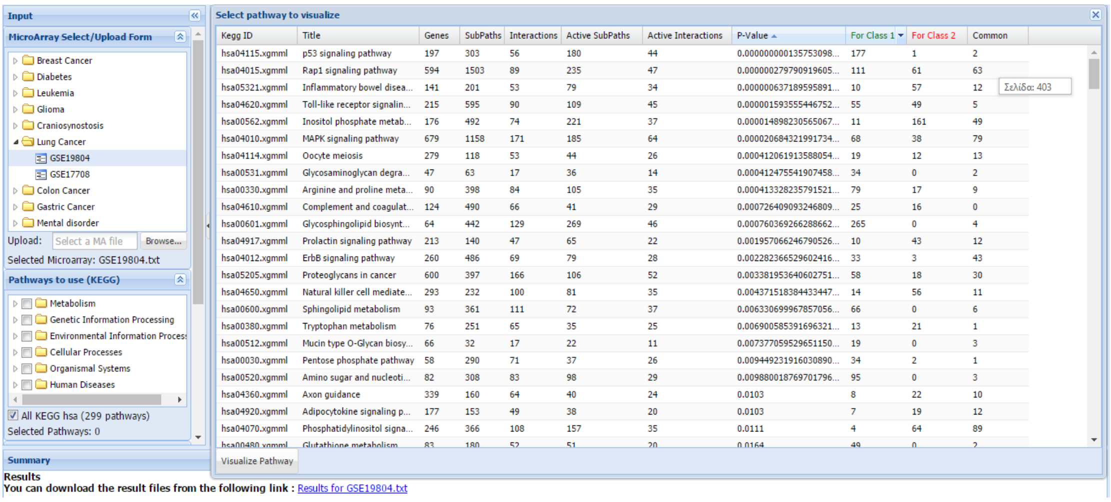
5.1.3 Lung cancer, CTCF and p53 signaling pathway

Lung cancer, also known as lung carcinoma, is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. If left untreated, this growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as primary lung cancers, are carcinomas. The two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC) [84].

Resistance to radio- and chemotherapy is a major problem in treatment responses of lung cancer. In this disease, biological markers, that can be predictive of response to treatment for guiding clinical practice, still need to be validated. Radiotherapy and most chemotherapeutic agents directly target DNA and in response to such therapies, p53 functions as a coordinator of the DNA repair process, cell cycle arrest, and apoptosis. Currently there are two approaches undertaken to target p53 and its regulators with an overall goal either to activate p53 in cancer cells for killing or to inactivate p53 temporarily in normal cells for chemoradiation protection [85].

Moreover, CTCF is a remarkably versatile, ubiquitous, and highly conserved zinc finger (ZF) protein and has been implicated in diverse cellular processes, including transcriptional regulation, alternative splicing, insulation, imprinting, X-chromosome inactivation, and higher-order chromatin organization [86]. CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. Among the various genomic CTCF target sites is the gene encoding the tumor suppressor protein p53. p53 is a sequence-specific transcription factor essential in the cellular response to DNA damage and other types of cellular stress [87]. Contingent on the level of DNA damage, p53 can initiate signaling pathways toward cell cycle arrest, senescence, or apoptosis to avoid oncogenic transformation [87].

So, we ran our Shiny application and chose CTCF transcription factor in lung cancer cell line. We inserted the binding sites they were derived in MinePath. For the feasibility study we chose a dataset for lung cancer (60 samples) and healthy samples (60 samples) (GSE19804). We used MinePath over the produced microarray dataset (60 lung cancer samples versus 60 healthy samples) and all the human KEGG pathways (299 in total). Figure 22 shows the significant pathways for lung cancer versus healthy according to MinePath.



Kegg ID	Title	Genes	SubPaths	Interactions	Active SubPaths	Active Interactions	P-Value	For Class 1	For Class 2	Common
hsa04115.xgmml	p53 signaling pathway	197	303	56	180	44	0.000000000135753098...	177	1	2
hsa04015.xgmml	Rap1 signaling pathway	594	1503	89	235	47	0.000000279790919605...	111	61	63
hsa05321.xgmml	Inflammatory bowel disea...	141	201	53	79	34	0.000000637189595891...	10	57	12
hsa04620.xgmml	Toll-like receptor signalin...	215	595	90	109	45	0.000001593555446752...	55	49	5
hsa00562.xgmml	Inositol phosphate metab...	176	492	74	221	37	0.000014898230565067...	11	161	49
hsa04010.xgmml	MAPK signaling pathway	679	1158	171	185	64	0.000020684321991734...	68	38	79
hsa04114.xgmml	Oocyte meiosis	279	118	53	44	26	0.000412061913588054...	19	12	13
hsa00531.xgmml	Glycosaminoglycan degra...	47	63	17	36	14	0.000412475541907458...	34	0	2
hsa00330.xgmml	Arginine and proline meta...	90	398	84	105	35	0.000413328235791521...	79	17	9
hsa04610.xgmml	Complement and coagulat...	124	490	66	41	29	0.000726409093246809...	25	16	0
hsa00601.xgmml	Glycosphingolipid biosynt...	64	442	129	269	46	0.000760369266288662...	265	0	4
hsa04917.xgmml	Prolactin signaling pathway	213	140	47	65	22	0.001957066246790526...	10	43	12
hsa04012.xgmml	ErbB signaling pathway	260	486	69	79	28	0.002282366529602416...	33	3	43
hsa05205.xgmml	Proteoglycans in cancer	600	397	166	106	52	0.003381953640602751...	58	18	30
hsa04650.xgmml	Natural killer cell mediate...	293	232	100	81	35	0.004371518384433447...	14	56	11
hsa00600.xgmml	Sphingolipid metabolism	93	361	111	72	37	0.006330699967857056...	66	0	6
hsa00380.xgmml	Tryptophan metabolism	76	251	65	35	25	0.006900585391696321...	13	21	1
hsa00512.xgmml	Mucin type O-Glycan biosy...	66	32	17	22	11	0.007377059529651150...	19	0	3
hsa00030.xgmml	Pentose phosphate pathway	58	290	71	37	26	0.009449231916030890...	34	2	1
hsa00520.xgmml	Amino sugar and nucleoti...	82	308	83	98	29	0.009880018769701796...	95	0	3
hsa04360.xgmml	Axon guidance	339	160	64	40	24	0.0103	8	22	10
hsa04920.xgmml	Adipocytokine signaling p...	177	153	49	38	20	0.0103	7	19	12
hsa04070.xgmml	Phosphatidylinositol signa...	246	366	108	157	35	0.0111	4	64	89
hsa00480.xgmml	Glutathione metabolism	83	180	52	51	20	0.0164	49	0	2

Figure 22: Significant Pathways according to MinePath for lung cancer versus healthy samples

Apart from the significant pathways, MinePath provides information for functional sub-paths for each phenotype. With such a functionality we can identify which are the functional sub-paths for lung cancer and what effect could have the binding sites of specific ChIP-seq data. For our experiment we used the binding sites of CTCF in A549 lung cancer cell line we derived from our Shiny application.

According to MinePath (Figure 22) the most significant pathway for the differentially expressed sub-paths based on the gene expression data is the p53 with a p-value less than 10^{-8}. Looking at the p53 pathway (Figure 23) in the MinePath viewer we can conclude that p53 pathway is mainly functional for lung cancer samples (green relation between genes) as we expected.

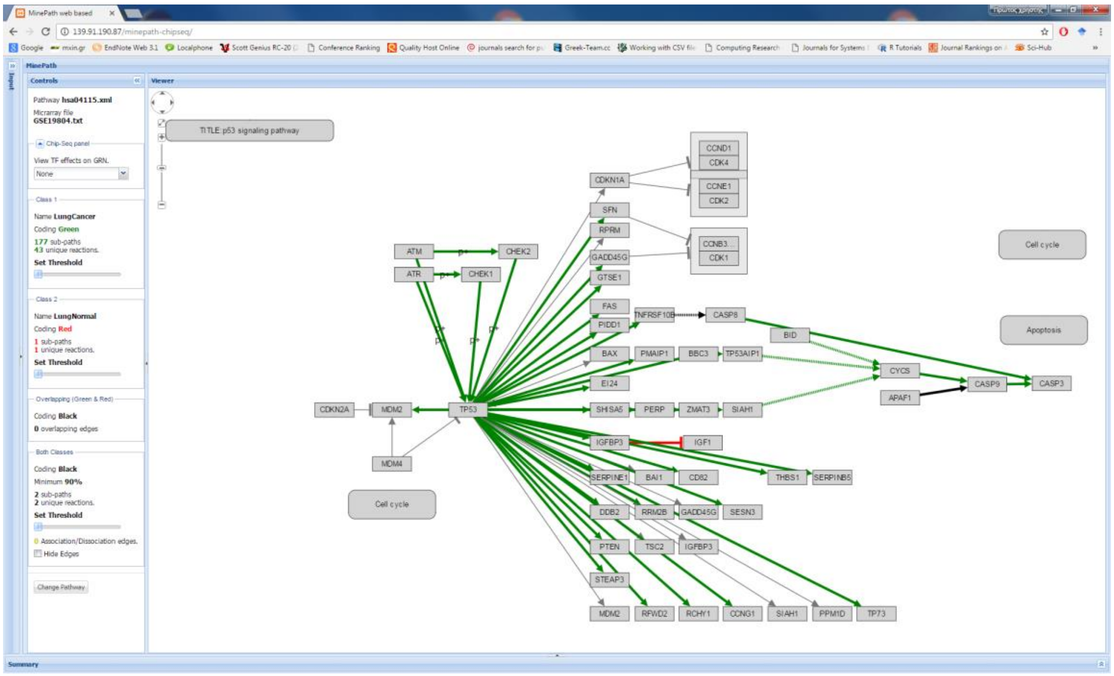


Figure 23 : p53 signaling pathway

A research question would be "What alerted p53 pathway only for lung cancer samples?" One of the known cancer related roots, which can alter the p53 pathway, can be found in the MAPK signaling pathway, TP53. TP53 gene that plays a key role in p53 pathway is expressed in MAPK pathway. The MAPK/ERK pathway (also known as the Ras-Raf-MEK-ERK pathway) is a chain of proteins in the cell that communicates a signal from a receptor on the surface of the cell to the DNA in the nucleus of the cell. When one of the proteins in the pathway is mutated, it can become stuck in the "on" or "off" position, which is a necessary step in the development of many cancers. Components of the MAPK/ERK pathway were discovered when they were found in cancer cells. Drugs that reverse the "on" or "off" switch are being investigated as cancer treatments [88].

As a result, we have chosen MAPK signaling pathway, one of the significant pathways according to MinePath and GGEA. In Figure 24 we visualize the functional sub-paths for lung cancer and healthy samples along with the binding sites of the CTCF on the MAPK signaling pathway.

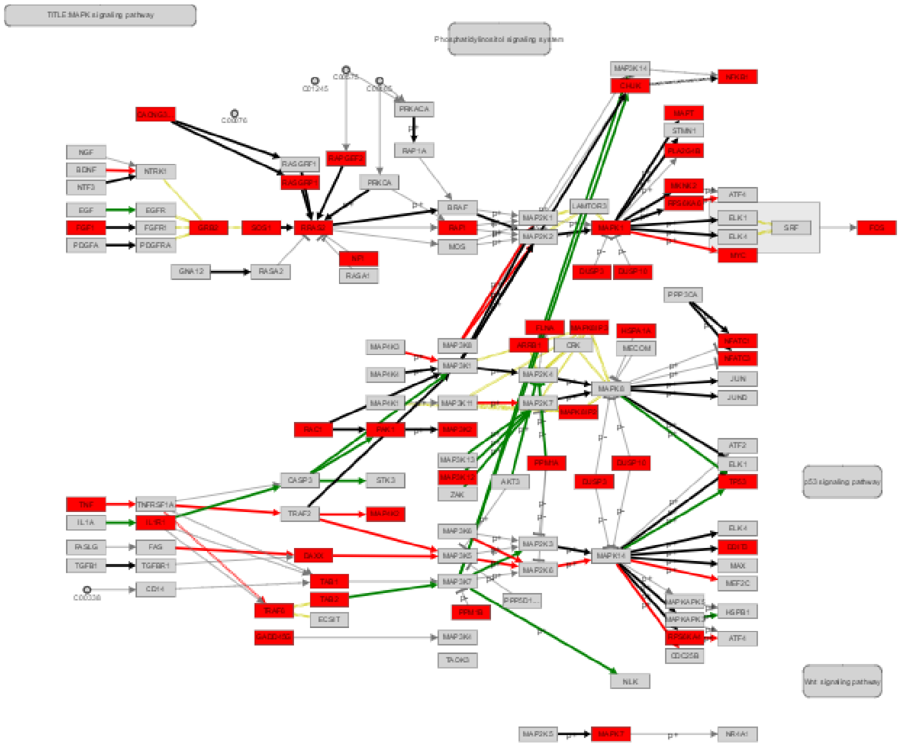


Figure 24 : MAPK functional sub-paths and binding sites of CTCF

Each node (box) is one protein with grey color being the default color from KEGG. The red color in the nodes indicate that the specific protein, gene or gene group is affected by the ChIP-seq which in our case it the CTCF ChIP-seq. For the edges (gene to gene reactions) we have four different colors:

- Green indicates functional reaction for the lung cancer samples
- Red indicates functional reaction for the healthy samples
- Black indicates relations that are almost always functional for both phenotypes (functional in over 90% of the samples for lung cancer and for healthy samples)
- Grey indicates inactive relations
- Yellow indicates the association/disassociation relations, which are considered physical associations and always hold.

As we can see from Figure 25 there is a clear path active only for glioma samples (green reactions) which starts from TAB2, activates MAP3K7 and in turn activates MAP2K3 which activates MAPK14, and in turn activates TP53, which continue to the alteration of p53 signaling pathway, apoptosis and alteration of the cell cycle.

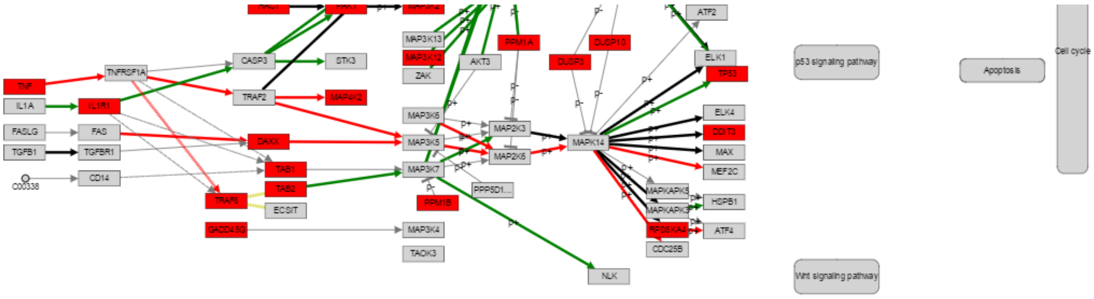


Figure 25 : TAB2-TP53 Sub-path for lung cancer samples

Given the high p53 mutation frequency in lung cancer which likely impairs some of the p53-mediated functions, a role of p53 as a predictive marker for treatment responses has been suggested [89]. Thus, p53 becomes the most appealing target for mechanism-driven anticancer drug discovery [85]. So, having in our disposal the ChIP-seq CTCF in lung cancer cell line data we can identify that a specific path can be disrupted since tp53 is one of the binding sites of CTCF along with the TAB2 which is the starting point of the specific sub-path. As a result, disruption of tp53 could lead to disruption of p53 signaling pathway that in turn affects Apoptosis and the cell cycle [87] in pathway MAPK pathway that is a pathway that is common among cancers and plays a key role in cancer treatment [88].

6 Discussion

High-throughput technologies, such as ChIP-seq, have made the collection of genome-wide data in cells, tissues and model organisms easier and cheaper. These data allow one to investigate biological aspects of cell functionality and to better understand previously unexplored disease etiologies. Pathway analysis using ChIP-seq data could aid researchers to determine the biological relevance of the binding sites over functional sub-paths and provide insights for new disease treatments. We created an application that produces ChIP-seq binding sites which when combined with expression data in a Gene Regulatory Network visualization tool (MinePath) can give us insight of significant biological procedures. So combining the extended version of MinePath, which provides a simple mechanism to visualize functional sub-paths in GRNs and the binding sites from ChIP-seq data for a specific protein we can visualize the binding sites over the functional and non-functional sub-paths. With such a merging a researcher can immediately understand the effect of a ChIP-seq for specific phenotype and identify at once which functional sub-paths will be affected after this effect [68].

So, the objective of this Master Thesis, to explore the effect of ChIP-seq data, coming from specific proteins under specific conditions in functional subpathways for specific phenotype (cancer vs. non-cancer) ,visualize these sub-pathways and create new significant knowledge for specific biological processes that occur and help scientists to gain insight for new disease treatments was met.

7 Conclusions

The sole nucleotide sequence of a gene does not explain its functions nor its regulation. Gene transcription is specified by DNA structure and by its accessibility to the basal transcription machinery [90]. The mechanism of transcriptional regulation of coding genes is one of the basic phase in biology dogma in systems biology. Transcriptional factors (TFs) are proteins that regulate several target genes by binding DNA motifs at the transcriptional level [91] [92]. Some investigators have reported that TFs take part in many important biological functions and human diseases, such as cell differentiation, proliferation, immune response, apoptosis, cardiac diseases, and tumor development [93] [94]. Interpreting the regulation of TFs is helpful for understanding their regulatory function in complex biological systems. A physical interaction of TFs, chromatin-modifying enzymes (histone acetyl/methyl transferases and deacetylases/ demethylases) and other accessory proteins with DNA is needed to modulate transcription dynamics, determining cell fate [95] . The binding of transcription factor proteins (TFs) to DNA promoter regions upstream of gene TSSs is one of the most important mechanisms by which gene expression, and thus many cellular processes, are controlled. Though in recent years many new kinds of data have become available for identifying transcription factor binding sites (TFBSs)- ChIP-seq among them .ChIP-seq technology is used primarily for the analysis of the interactions of DNA with proteins(TFs, histone or other chromatin-modifying enzymes) and it finds all the binding sites of these proteins . Nowadays ChIP-Seq is the gold standard for studying TF-chromatin interactions in vivo [19] and in silico.

On the other hand, biological pathways that represent complex interactions between proteins in a molecule in living cells (GRNs) can help biologists to infer new knowledge for the biological interactions in a molecular level. The availability of several gene expression datasets generated from knock-out cells for one or few TFs has made possible to infer GRNs. Reconstructing GRNs using gene expression data has been one of the most widely studied problems in the last decade [96].

In addition, there are not general tools that allow comparing the developed methods for gene expression prediction and GRN on the same benchmarks. As a result, it is now very difficult for biologists to carry on data integration. So as to facilitate biologists in such a task we strongly emphasize the need to develop new and intuitive explorative tools for the integration of ChIP-seq and RNA-seq data. Moreover, we believe such tools should be designed in the spirit of reproducible research [97] to allow reproducibility and transparent verification of published results and to improve transfer of knowledge [96].

REFERENCES

[1] P. Collas, "The Current State of Chromatin Immunoprecipitation," *Mol Biotechnol,* vol. 45, pp. 87-100, 14 January 2010.

[2] V. Narang, M. A. Ramli, A. Singhal, P. Kumar, G. d. Libero, M. Poidinger and C. Monterola, "Automated Identification of Core Regulatory Genes in Human Gene Regulatory Networks," *PLOS Computational Biology,* pp. 1-28, 22 September 2015.

[3] Nature, "Initial sequencing and analysis of the human genome," *Macmillan Magazines Ltd,* no. 409, pp. 861-921, 15 February 2001.

[4] Davidson, M. Levine and H. Eric, "Gene regulatory networks for development," *Proceedings of the National Academy of Sciences,* vol. 102, no. 14, p. 4936–4942, 5 April 2005.

[5] Min, Jaegyoon and Youngmi, "A Network-Based Classification Model for Deriving Novel Drug-Disease Associations and Assessing Their Molecular Actions," *PLOS ONE,* vol. 9, no. 10, pp. 1-12, 30 October 2014.

[6] J. Qin, Y. Hub, F. Xu, H. K. Yalamanchili and J. Wang, "Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods," *Methods,* vol. 67, no. 3, pp. 294-303, 5 March 2014.

[7] V. Narang, M. A. Ramli, A. Singhal, P. Kumar, G. d. Libero, M. Poidinger and C. Monterola, "Automated Identification of Core Regulatory," *PLOS-Computational Biology,* pp. 1-28, 22 September 2015.

[8] R. M. Buckingham and W. Peter, "Gene Regulatory Networks and Transcriptional Mechanisms that Control Myogenesis," *Developmental Cell,* pp. 225-238, 10 February 2014.

[9] J. C. Chen, M. J. Alvarez, F. Talos, H. Dhruv, G. E. Rieckhof, A. Lyer, K. L. Diefes, K. Aldape, M. Berens, M. M. Shen and A. Califano, "Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks," *CELL,* vol. 159, no. 2, pp. 402-414, 9 October 2014.

[10] Y. Nishio, Y. Usuda, K. Matsui and H. Kurata, "Computer-aided rational design of the phosphotransferase system for enhanced glucose uptake in Escherichia coli," *Molecular Systems Biology,* vol. 4, no. 160, pp. 1-12, 15 January 2008.

[11] R. C. McLeay, T. Lesluyes, G. C. Partida and T. L. Bailey, "Genome-wide in silico prediction of gene expression," *BIOINFORMATICS,* vol. 28, no. 21, p. 2789–2796, 6 September 2012.

[12] C. Cheng, K.-K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, M. Snyder and M. Gerstein, "Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data," *PLoS Computational Biology,* vol. 7, no. 11, pp. 1-15, 17 November 2011.

[13] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj and A. P. Boyle, "Architecture of the human regulatory network derived from ENCODE data," *Nature,* vol. 489, pp. 91-100, 5 September 2012.

[14] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen and M. Selbach, "Global quantification of mammalian gene expression control," *Nature,* vol. 473, p. 337–342, 18 May 2011.

[15] E. L. V. Dijk, H. Auger, Y. Jaszczyszyn and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics,* vol. 30, no. 9, pp. 418-426, 6 August 2014.

[16] D. Guan, J. Shao, Y. Deng, P. Wang, Z. Zhao, Y. Liang, J. Wang and B. Yan, "CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data," *Bioinformatics,* vol. 30, no. 8, pp. 1190-1192, 2 January 2014.

[17] H. Shin, X. D. Tao Liu, Y. Zhang and X. S. Liu, "Computational methodology for ChIP-seq analysis," *Quantitative Biology,* vol. 1, no. 1, pp. 54-70, 2013.

[18] L. Koumakis, "Computational Methods for Knowledge Discovery from Heterogeneous Data Sources: Methodology and Implementation on Biological and Molecular Sources," September 2014.

[19] V. G. Levitsky, I. V. Kulakovskiy, N. I. Ershov, D. Y. Oshchepkov, V. J. Makeev, T. C. Hodgman and T. I. Merkulova, "Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data," *BMC Genomics*, vol. 15, no. 80, pp. 1-12, 25 January 2014.

[20] A. S. Foulkes, Applied Statistical Genetics with R: For Population-based Association Studies, Springer, 2009.

[21] J. D. S. a. R. Tibshirani, "Statistical significance for genomewide studies," *PNAS,* vol. 100, no. 16, pp. 9440-9445, 5 August 2003.

[22] R. Tibshirani and J. D. Storey, "Statistical significance for genomewide studies," *PNAS,* p. 9440–9445, 5 August 2003.

[23] Hochberg and Benjamini, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B,* vol. 57, pp. 289-300, 1995.

[24] J. W. Ho, E. Bishop, P. V. Karchenko, N. Negre, K. P. White and P. J. Park, "ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis," *BMC Genomics,* vol. 12, no. 134, 2011.

[25] V. Filkov, Identifying Gene Regulatory Networks from Gene Expression Data.

[26] Y. Blat and N. Kleckner, "Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region," *Cell,* vol. 98, no. 2, pp. 249-259, 23 July 1999.

[27] Lieb, Liu, Botstein and Brown, "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association," *Nature Genetics,* vol. 28, no. August, pp. 327-334, August 2001.

[28] L. Vishwanath, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder and P. O. Brown, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature,* vol. 409, no. 6819, pp. 533-538, 25 January 2001.

[29] B. R. R. J. J. Wyrick, O. Aparicio, E. G. Jennings, l. Simon, J. Zeitlinger, J. chreiber, N. a. nett, E. Kanin and T. Volkert, "Genome-wide location and function of DNA binding proteins," *Science,* vol. 290, no. 5500, pp. 2306-2309, 22 December 2000.

[30] M. J. Buck, "ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments," *Genomics,* vol. 83, no. 3, pp. 349-360, April 2004.

[31] T. I. Lee, S. E. Johnstone and R. A. Young, "Chromatin immunoprecipitation and microarray-based analysis of protein location," *Nature,* pp. 729-748, 13 July 2006.

[32] Girardot, Sklyar, Grosz, Huber and Furlong, "CoCo: a web application to display, store and curate ChIP-on-chip data integrated with diverse types of gene expression data.," *Bioinformatics,* vol. 23, no. 6, pp. 771-773, 15 March 2007.

[33] Linhart, Halperin and Shamir, "Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets," *Genome Res,* vol. 18, no. 7, pp. 1180-1189, 18 July 2008.

[34] F. Nielsen, K. Markus, R. Friborg, L. Favrholdt, H. Stunnenberg and H. Huynen, "CATCHprofiles: clustering and alignment tool for ChIP profiles," *PLoS One,* vol. 7, no. 1, pp. 1-8, 4 January 2012.

[35] T. Benoukraf, P. Cauchy, R. Fenouil, A. Jeanniard, F. Koch, S. Jaeger, D. Thieffry, J. Imbert, J. Andrau, S. Spicuglia and P. Ferrier, "CoCAS: a ChIP-on-chip analysis suite," *Bioinformatics,* vol. 25, no. 7, pp. 954-955, 4 February 2009.

[36] H. Ji and W. Wong, "TileMap: create chromosomal map of tiling array hybridizations," vol. 21, no. 18, pp. 3629-3636, 26 July 2005.

[37] T. Chen, H. Li, C. Lee, R. Gan, P. Huang, T. Wu, C. Lee, Y. Chang and P. Tang, "ChIPseek, a web-based analysis tool for ChIP data," *BMC Genomics,* vol. 15, no. 1, pp. 1-13, 30 June 2014.

[38] H. Shin, H. Liu, A. Manrai and X. Liu, "CEAS: cis-regulatory element annotation system," *Bioinformatics,* vol. 25, no. 19, pp. 2605-2606, 18 August 2009.

[39] T. I. Lee, S. E. Johnstone and R. A. Young, "Chromatin immunoprecipitation and microarray-based analysis of protein location," *NATURE PROTOCOLS,* pp. 729-748, 13 July 2006.

[40] S. Steinhauser, N. Kurzawa, R. Eils and C. Herrmann, "A comprehensive comparison of tools for differential ChIP-seq analysis," *Briefings in Bioinformatics,* no. October 2015, pp. 1-14, 13 January 2016.

[41] D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *SCIENCE,* pp. 1497-1502, 8 June 2007.

[42] P. J. Park, "ChIP-Seq: advantages and challenges of a maturing technology," *PMC,* vol. 10, no. 10, pp. 669-680, 12 October 2012.

[43] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield and D. T. Odom, "ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions," *PMC,* pp. 240-248, 11 June 2014.

[44] S. BERGER, S. OMIDI, M. PACHKOV, P. ARNOLD, N. KELLEY, S. SALATINO and E. V. NIMWEGEN, "Crunch: Completely Automated Analysis of ChIP-seq Data," *bioRxiv,* pp. 1-31, 8 March 2016.

[45] T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim and J. Zhang, "Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data," *PLOS Computational Biology,* vol. 9, no. 11, pp. 1-8, 9 November 2013.

[46] N. A. Fonseca, J. Rung, A. Brazma and J. C. Marioni, "Tools for mapping high-throughput sequencing data," *BIOINFORMATICS,* vol. 28, no. 24, pp. 3169-3177, 11 October 2012.

[47] A. Verfaillie, H. Imrichova, R. Janky and S. Aerts, "iRegulon and i-cisTarget: Reconstructing Regulatory Networks Using Motif and Track Enrichment," *Current Protocols in Bioinformatics,* 17 December 2015.

[48] E. G. Wilbanks and M. T. Facciotti, "Evaluation of Algorithm Performance in ChIP-Seq Peak Detection," *PLoS ONE,* vol. 5, no. 7, pp. 1-12, 8 July 2010.

[49] P. Papior and A. Schepers, "Why are we where we are? Understanding replication origins and initiation sites in eukaryotes using ChIP-approaches," *Chromosome Research,* vol. 2010, no. 18, pp. 63-77, 2010.

[50] W. M. Wong and W. Hung, "THE ANALYSIS OF CHIP-SEQ DATA," in *Methods in Enzymology*, vol. 497, C. Voigt, Ed., Elsevier, 2011, pp. 58-67.

[51] T. S. Furey, "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions," *Nature,* pp. 840-852, 23 October 2012.

[52] G. Ambrosini, R. Dreos and P. Bucher, "Principles of ChIP-seq Data Analysis Illustrated with Examples," *GENOMICS AND COMPUTATIONAL BIOLOGY,* vol. 1, no. 1, pp. 1-12, 18 September 2015.

[53] D. A. a. R. Shamir, "Constructing module maps for integrated analysis of heterogeneous biological networks," *Nucleic Acids Research,* p. 4208–4219, 4 February 2014.

[54] C. J. A. K. a. D. Z. Thair Judeh, "TEAK: Topology Enrichment Analysis frameworK for detecting activated biological subpathways," *Nucleic Acids Research,* pp. 1-13, 24 December 2012.

[55] T. Hase, S. Ghosh, R. Yamanaka and H. Kitano, "Harnessing Diversity towards the Reconstructing of Large Scale Gene Regulatory Networks," *PLOS Computational Biology,* vol. 9, no. 11, pp. 1-16, 21 November 2013.

[56] L. Fengkai and M.-S. Diego, "rTRM-web: a web tool for predicting transcriptional regulatory modules for ChIP-seq-ed transcription factors," *Gene,* vol. 546, no. 2, pp. 417-420, 10 August 2014.

[57] P. Wang, Q. Jing, Q. Yiming, Z. Yun, L. Y. Wang, M. J. Li, M. Q. Zhang and J. Wang, "ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks," *Nucleic Acids Research,* vol. 43, no. Web Server Issue, p. W264–W26, 27 April 2015.

[58] P. Wang, J. Qin, Q. Y. Zhu, L. Wang, M. Li, M. Zhang and J. Wang, "ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor," *Nucleic Acids Research,* vol. 39, no. Web Server Issue, pp. W430-W436, 17 May 2011.

[59] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favoro, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi and Grazi, "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology,* vol. 23, no. 1, pp. 137 - 144, 6 January 2005.

[60] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus and R. Irizarry, "Bioconductor: open software development for computational biology

and bioinformatics," *Genome Biology,* vol. 5, no. 10, pp. 1-16, 15 September 2004.

[61] L. J. Zhu, C. Gazin, N. D. Lawson, H. Pages, S. M. Lin, D. S. Lapointe and M. R. Green, "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data," *BMC Bioinformatics,* vol. 11, no. 237, pp. 1-10, 11 May 2010.

[62] Z. L, "Integrative analysis of ChIP-chip and ChIP-seq dataset," in *Tilling Arrays*, vol. 1067, Humana Press, 2013, pp. 105-124.

[63] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein and K. I, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia," *Genome Research,* vol. 22, no. 9, pp. 1813-1831, September 2012.

[64] J. Rozowsky, G. Euskirchen, R. Auerbach, Z. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder and M. Gerstei, "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Natural Biotechnology,* vol. 27, no. 1, pp. 66-75, January 2009.

[65] Q. Li, J. Brown, H. Huang and P. Bickel, "Measuring reproducibility of high-throughput experiments," *Annals of Applied Statistics,* pp. 1752-1779, 21 October 2011.

[66] T. L. Q. Z. a. X. S. L. Jianxing Feng, "Identifying ChIP-seq enrichment using MACS," *Nat Protoc,* vol. 7, no. 9, pp. 1-24, 7 September 2012.

[67] P. V. Kharchenko, M. Y. Tolstorukov and P. J. Park, "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *National Biotechnology,* vol. 26, no. 12, p. 1351–1359, December 2008.

[68] L. Koumakis, G. Potamias, K. Marias and M. Tsiknakis, "An algorithmic approach for the effect of transcription factor binding sites over functional gene regulatory networks," in *Bioinformatics and Bioengineering (BIBE)-2015 IEEE 15th International Conference*, 2016.

[69] ENCODE Project Consortium, "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature*, vol. 489, no. 7414, pp. 57-74, 6 September 2012.

[70] H. Wang, M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers and

J. A. Stamatoyannopoulos, "Widespread plasticity in CTCF occupancy linked to DNA methylation," *Genome research,* vol. 22, no. 9, pp. 1680-1688, September 2012.

[71] M. S. Carro, W. K. Lim, M. J. Alvarez, R. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape and A. Califano, "The transcriptional network for mesenchymal transformation of brain tumors," *Nature,* vol. 463, no. 7279, pp. 318-325, 23 December 2009.

[72] H. Armanious, P. Gelebart, J. Mackey and M. Yupo, "STAT3 upregulates the protein expression and transcriptional activity of β-catenin in breast cancer," *International journal of clinical and experimental pathology 3,* vol. 3, no. 7, pp. 654-664, 25 July 2010.

[73] Jacoby, A. N. Mamelak and B. Douglas, "Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601)," *Expert Opinion in Drug Delivery,* vol. 4, no. 2, pp. 175-186, 5 March 2007.

[74] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer genetics,* vol. 205, no. 12, pp. 613-621, December 2012.

[75] J.-X. Zhang, J. Zhang, W. Yan, Y.-Y. Wang, L. Han, X. Yue, N. Liu, Y.-P. You, T. Jiang, P.-Y. Pu and C.-S. Kang, "Unique genome-wide map of TCF4 and STAT3 targets using ChIP-seq reveals their association with new molecular subtypes of glioblastoma," *Neuro-Oncology,* vol. 15, no. 3, pp. 279-289, 07 January 2013.

[76] G. Xingchun, M. Yajing, M. Yue and J. Weilin, "LEF1 regulates glioblastoma cell proliferation, migration, invasion, and cancer stem-like cell self-renewal," *Tumor Biology,* vol. 35, no. 11, pp. 11505-11511, November 2014.

[77] Y. L. W. Y. W. Z. L. C. G. Y. Z. B. Y. W. H. W. C. K. T. Jiang, "MiR-218 reverses high invasiveness of glioblastoma cells by targeting the oncogenic transcription factor LEF1," *Oncology Reports,* pp. 1013-1021, 5 July 2012.

[78] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott and A. Maayan, "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins," *Database,* pp. 1-17, 31 May 2016.

[79] H. X. K. I. B. R. M. a. A. M. Alexander Lachmann, "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments," *Bioinformatics,* vol. 26, no. 19, pp. 2438-2444, 1 October 2010.

[80] F. Mohammad, R. M. Flight, B. J. Harrison, J. C. Petruska and E. Rouchka, "AbsIDconvert: An absolute approach for converting genetic identifiers at different granularities," *BMC Bioinformatics,* vol. 13, no. 229, pp. 1-22, 12 September 2012.

[81] L. Koumakis, V. Moustakis, M. Zervakis, D. Kafetzopoulos and G. Potamias, "Coupling Regulatory Networks and Microarays:Revealing Molecular Regulations of Breast Cancer Treatment Responses. Artificial Intelligence: Theories and Applications," *Lecture Notes in Computer Science,* no. 7297, pp. 239-246, 2012.

[82] C. BM, J. Smith, Y. Chen and J. Chen, "Reversing HOXA9 oncogene activation by PI3K inhibition: epigenetic mechanism and prognostic significance in human glioblastoma," *Cancer Research,* vol. 70, no. 2, pp. 453-462, 15 January 2010.

[83] S. Al, Wiltshire, Batalov and H. Lapp, "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc Natl Acad Sci U S,* vol. 101, no. 16, pp. 6062-6067, 20 April 2004.

[84] "WIkipedia," 2015. [Online]. Available: https://en.wikipedia.org/wiki/Lung_cancer. [Accessed 18 August 2016].

[85] Sun, Z. Wang and Yi, "Targeting p53 for Novel Anticancer Therapy," *Transl Oncol,* vol. 3, no. 1, pp. 1-12, February 2010.

[86] d. L. W. Holwerda SJB, "CTCF: the protein, the binding partners, the binding sites and their chromatin loops," *Philos Trans R Soc Lond B Biol Sci,* vol. 368, no. 1620, 19 June 2013.

[87] Vousden and Prives, "Blinded by the light: the growing," *Cell,* vol. 137, no. 3, pp. 413-431, 1 May 2009.

[88] R. J. Orton, O. E. Sturm, V. Vyshemirsky, M. Calder, D. R. Gilbert and W. Kolch, "Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway," *Biochemical Journal,* vol. 392, no. 2, pp. 249-261, 1 December 2015.

[89] Viktorsson, D. Petris and Lewensohn, "The role of p53 in treatment responses of lung cancer," *Biochem Biophys Res Commun.*, vol. 331, no. 3, pp. 868-880, 10 June 2005.

[90] C. Angelini and V. Costa, "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems," *Cell and Developmental Biology*, vol. 2, pp. 1-8, 17 September 2014.

[91] B. Xu, D. E. Schones, Y. Wang, H. Liang and G. Li, "A Structural-Based Strategy for Recognition of Transcription Factor Binding Sites," *PLoS ONE*, vol. 8, no. 1, pp. 1-10, 8 January 2013.

[92] A. S. Bais, N. Kaminski and P. V. Benos, "Finding subtypes of transcription factor motif pairs with distinct regulatory roles," *Nucleic Acids Research*, vol. 39, no. 11, pp. 1-13, 22 March 2011.

[93] L. Peiyao, X. Gang, L. Guiyuan and W. Minghua, "Function and mechanism of tumor suppressor gene LRRC4/NGL-2," *Molecular Cancer*, vol. 13, no. 266, pp. 1-7, 19 September 2014.

[94] W. Czaja, K. Y. Miller, M. K. Skinner, Miller and L. Bruce, "Structural and functional conservation of fungal MatA and human SRY sex-determining proteins," *Nature Communications*, vol. 5, no. 5434, pp. 1-6, 17 November 2014.

[95] Halfon, T. J. Atkinson and S. Marc, "Regulation of gene expression in the genomic context," *Computational and Structural Biotechnology Journal*, vol. 9, no. 13, 29 January 2014.

[96] V. Costa and C. Angelini, "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data : statistical solutions to biological problems," *Cell and Developmental Biology*, pp. 1-7, 17 September 2014.

[97] F. Russo and C. Angelini, "RNASeqGUI: a GUI for analysing RNA-Seq data," *Bioinformatics*, vol. 30, no. 17, pp. 2514-2516, 1 September 2014.

[98] C. Ross-Innes, G. Brown and J. Carroll, "A co-ordinated interaction between CTCF and ER in Breast Cancer Cells," *BMC Genomic*, vol. 12, no. 593, pp. 1-10, 2011.

[99] J. Rainer, "EnsDb.Hsapiens.v75: Ensembl based annotation package," 2016.