Statistical methods for predicting genetic regulation

by

Nisar Ahmed Shar

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

in the

Faculty of Biological Sciences

School of Molecular and Cellular Biology

November 2016

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter 5 and 6 of the thesis has appeared in publication as follows

Shar, Nisar A., M. S. Vijayabaskar, and David R. Westhead. **"Cancer somatic mutations cluster in a subset of regulatory sites predicted from the ENCODE data."** Molecular Cancer 15.1 (2016): 76.

I was responsible for carrying out this study and writing the paper along with the David R. Westhead. M.S. Vijayabaskar helped with data analysis, paper writing and supervision of the work. David R. Westhead conceived and designed the study.

Acknowledgements

First and foremost, I would like to thank my supervisor, David Westhead, for huge support, guidance and unending patience. He has been very kind and approachable throughout my PhD and I was lucky to have learned supervisor like him. Dave has been encouraging and kind to me whenever I get confused and frustrated.

I would like to thank everybody in the laboratory specially Francis, Fatin and Chulin for providing moral support and a good company whenever I needed. A special thanks to Vijaya Baskar and Mathew Care for helping me during PhD.

A big thanks to my family for their unending support and understanding my limitations being thousands of miles away from me.

Lastly, I would like to thank NED university of Engineering & Technology, Karachi and University of Leeds for sponsoring my PhD, and anyone who has helped me in anyway during my PhD duration.

Abstract

Transcriptional regulation of gene expression is essential for cellular differentiation and function, and defects in the process are associated with cancer. Transcription is regulated by the cis-acting regulatory regions and trans-acting regulatory elements. Transcription factors bind on enhancers and repressors and form complexes by interacting with each other to control the expression of the genes. Understanding the regulation of genes would help us to understand the biological system and can be helpful in identifying therapeutic targets for diseases such as cancer. The ENCODE project has mapped binding sites of many TFs in some important cell types and this project also has mapped DNase I hypersensitivity sites across the cell types.

Predicting transcription factors mutual interactions would help us in finding the potential transcription regulatory networks. Here, we have developed two methods for prediction of transcription factors mutual interactions from ENCODE ChIP-seq data, and both methods generated similar results which tell us about the accuracy of the methods. It is known that functional regions of genome are conserved and here we identified that shared/overlapping transcription factor binding sites in multiple cell types and in transcription factors pairs are more conserved than their respective non-shared/non-overlapping binding sites. It has been also studied that co-binding sites influence the expression level of genes. Most of the genes mapped to the transcription factor co-binding sites have significantly higher level of expression than those genes which were mapped to the single transcription factor bound sites.

The ENCODE data suggests a very large number of potential regulatory sites across the complete genome in many cell types and methods are needed to identify those that are most relevant and to connect them to the genes that they control. A penalized regression method, LASSO was used to build correlative models, and choose two regulatory regions that are predictive of gene expression, and link them to their respective gene.

Here, we show that our identified regulatory regions accumulate significant number of somatic mutations that occur in cancer cells, suggesting that their effects may drive cancer initiation and development. Harboring of somatic

mutations in these identified regulatory regions is an indication of positive selection, which has been also observed in cancer related genes.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ENCODE	**E**ncyclopedia **O**f **D**NA **E**lements
ChIP-seq	**C**hromatin **I**mmuno**P**recipitation followed by sequencing
DHSs	**D**Nase I **H**ypersensitivity **S**ites
TFs	**T**ranscription **F**actors
CREs	**C**is **R**egulatory **E**lements
CRRs	**C**andidate cis **R**egulatory **R**egions
snRNA	**s**mall **n**uclear **RNA**
TFIID	**T**ranscription **F**actor **IID**
TFIIH	**T**ranscription **F**actor **IIH**
TBP	**T**ATA **Bi**nding **P**rotein
TSS	**T**ranscription **S**tart **S**ite
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
COSMIC	**C**atalogue **O**f **S**omatic **M**utations **I**n **C**ancer
FPKM	**F**ragments **P**er **K**ilobase of exon per **M**illion reads
UCSC	**U**niversity of **C**alifornia, **S**anta **C**ruz
IDR	**I**rreproducible **D**iscovery **R**ate
TCGA	**T**he **C**ancer **G**enome **A**tlas
CML	**C**hronic **M**yelogenous **L**eukaemia
glm	**G**eneralized **L**inear **M**odel

Chapter 1

1 Introduction

This thesis is about transcriptional regulation of human genes, which is more complex than prokaryotes and some eukaryotic organisms for example yeast, eukaryotic genome contains introns (non-coding region) and large intergenic regions, these regions add more complexity to the regulation of genes. However, size of genome is not correlated with the genetic complexity for some organisms i.e., salamander and lilies. Both these organisms contain ten times more amount of DNA than in the human genome but they are not even more complex than human [1]. Transcriptional regulation in eukaryotes involves the regulation of genes by trans-acting and cis-acting regulatory regions. Trans-acting elements (transcription factors) bind on the cis regulatory regions (enhancers or repressors) to control the expression level of genes. However, distant cis regulatory regions bend themselves to bind with the promoters of corresponding genes [2]. These regulatory regions are short (50-1500bp) regions of DNA and can be located up to 1Mbp away from the transcription start site (TSS) [3]. Transcription factors (TFs) can mutually interact with each other, though some TFs bind directly on cis regulatory regions and some through other TFs (indirect binding). For example, transcription factors AP-2 and AP-3 interact with each other mutually and bind on the SV40 enhancer [4].

1.1 Overview of Molecular Biology

1.1.1 Gene regulation

Gene expression is the process by which information in our DNA is converted into a functional product such as protein and non-coding products such as transfer RNA (tRNA), small nuclear RNA (snRNA) and others. Diagrammatic representation of central dogma is shown in Figure 1.1. This figure contains two

important biological processes: transcription and translation. Transcription involves synthesis of mRNA from DNA and translation involves synthesis of protein using information coded in the mRNA.

Transcription is the first step of gene expression and can be controlled by enhancers, repressors and other factors. Transcription starts with the binding of TATA-binding protein (TBP) on the promoter region such as TATA box, this TBP is a subunit of Transcription Factor IID (TFIID). After binding of TBP, RNA polymerase along with five more TFs bind around the TATA box to form a pre-initiation complex. Transcription Factor II H (TFIIH) have role in separating opposing strand of double stranded DNA to provide the RNA polymerase access to the single stranded DNA template. Cis-regulatory regions such as enhancers can increase the rate of transcription and repressors can decrease the rate of transcription and these regulatory regions can be bound by TFs and also can interact with each other through looping as illustrated in Figure 1.2.

The mRNA is the product of transcription containing introns and exons: introns are non-coding regions and exons are coding regions. Transcription is followed by a mechanism called alternative splicing that involves removal of introns and joining of exons to form a mature mRNA that is ready for translation.

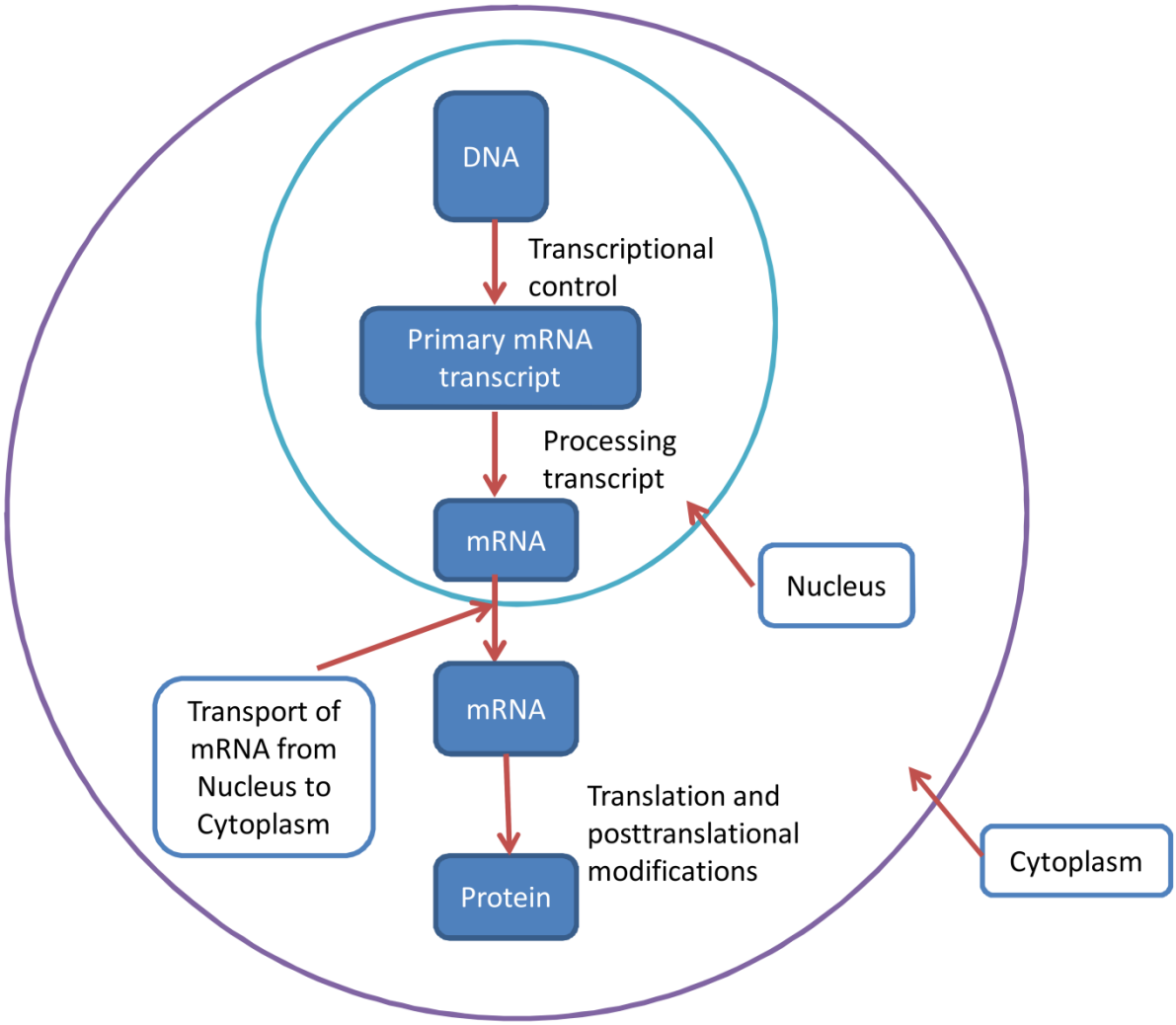


Figure 1. 1: Central dogma: This figure shows the different processes from transcription to translation. Transcription plays an important role in gene regulation and there are control elements that control the expression of genes. Transcription occurs in the nucleus producing an mRNA which is then transported to the cytoplasm for translation. This whole process is known as the central dogma.

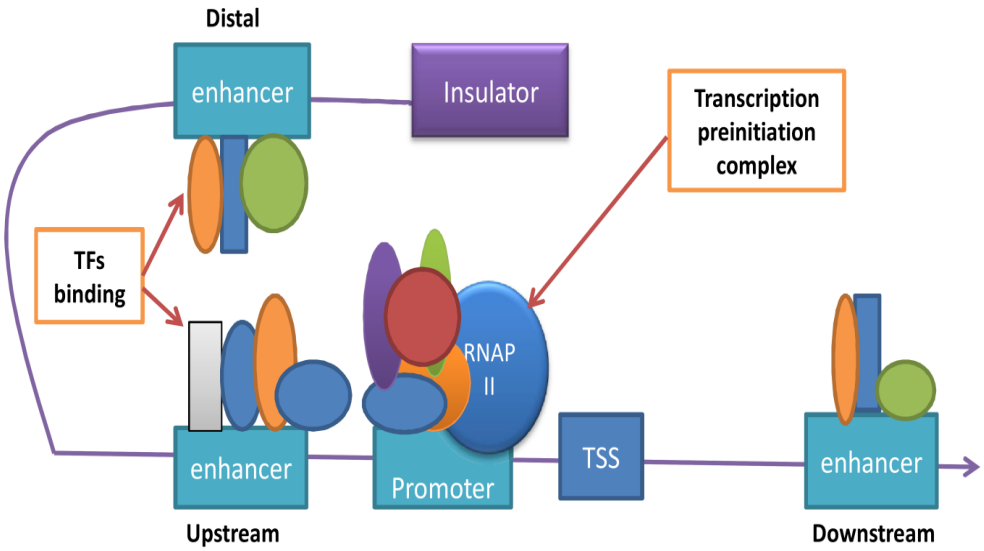


Figure 1. 2: Transcription: This figure shows the regulators of transcription with transcription factors bound to genomic regions. RNA polymerase II (RNAP II) initiates transcription by binding on the promoter along with other factors. Distal enhancers can interact with other regulatory regions by looping.

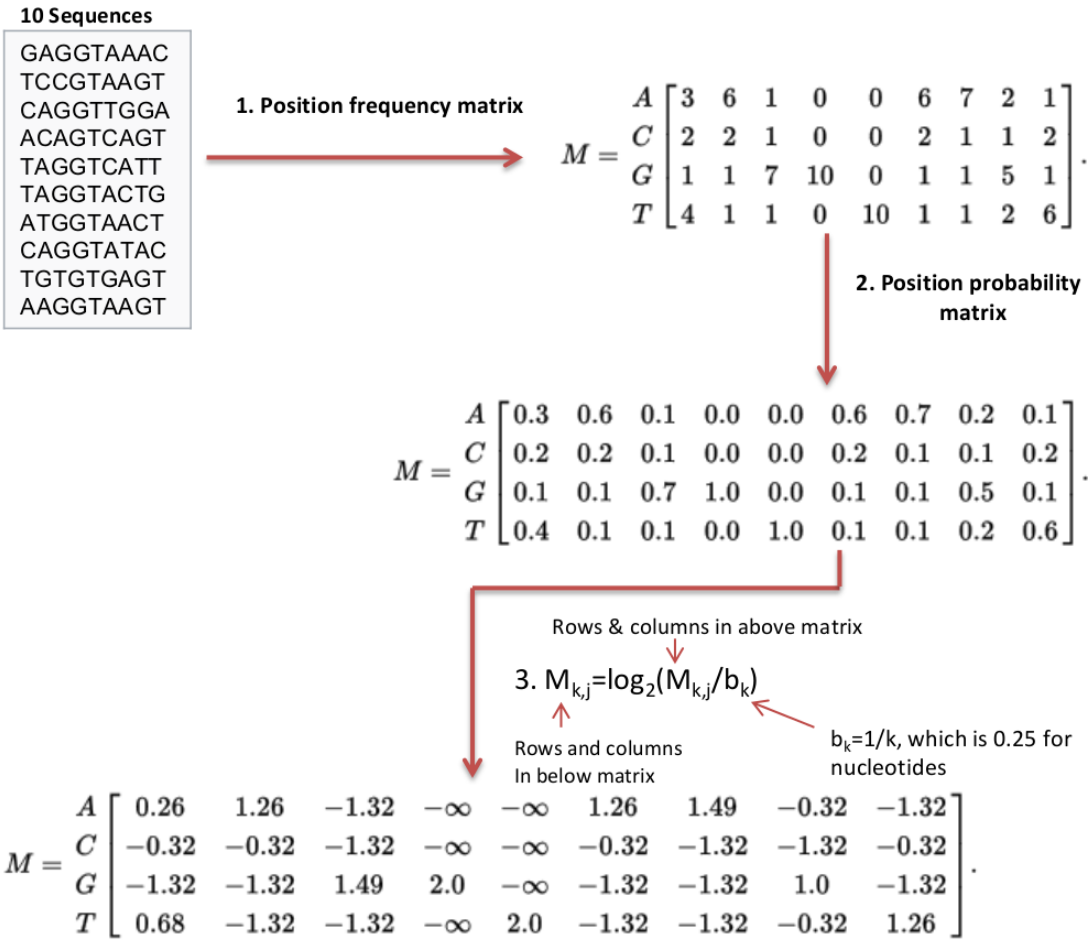
1.1.2 Transcription factors

A transcription factor (TF) is a protein that binds to the cis-regulatory regions to control the rate of transcription of genes. TFs bind on the DNA in sequence specific manner and these sequences are called DNA motifs. Transcription factors contain several structural elements which can mediate their binding on specific DNA sequences and these structural elements have been used to classify factors into families [5]. Transcription factors can bind directly on the DNA sequences followed by indirect binding of other transcription factors on the same DNA region to form the transcription factor complexes to influence the gene expression as shown in Figure 1.2. Combinatorial regulation of transcription can partly explain the complexity of gene regulation in higher eukaryotes [6]. Each transcription factor can bind to multiple binding sites and also multiple transcription factors can bind on the same genomic region. Transcription factor binding can help to understand different biological functions, for example several regulatory interactions can be inferred from TF binding patterns [7]. This pattern of transcription factor binding can help us to identify cis regulatory regions, as

Benjamin P. Berman and co-workers identified these regions by TF binding in Drosophila [8].TF binding sites can be functional sites, so they are usually conserved among different mammals but some studies have shown that TF binding sites are divergent in some organisms as they evolved [9].

There are different ways of representing transcription factor binding motifs, and one of them is Position weight matrices (PWM). These type of matrices are also known as position specific scoring matrices (PSSM). Position weight matrices are obtained from the set of aligned sequences that are functionally related and this method has been important for computational motif discovery. As discussed above TFs bind on the specific sequences that are motifs.

PWM construction starts with creation of position frequency matrix, which is the number of occurrences of each nucleotide at each position. 2nd step is the creation of position probability matrix (PPM) by dividing the nucleotide count at each position by the number of sequences. Now the elements in the PWMs can be calculated by $\log_2 (M_{k,j}/b_k)$, and $M_{k,j}$ are given in the position probability matrix (2nd step) as shown in Figure 1.3. While, b_k =1/k (1/4), which is 0.25 for nucleotides [10]. For example, 1st position of “A” in PWM can be calculated by $\log_2$ (0.3/0.25), and the result would be 0.26 as shown in Figure 1.3 (3rd step).



$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}.$$

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32 \\ -1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32 \\ 0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26 \end{bmatrix}.$$

Figure 1. 3: This figure shows the steps in construction of Position weight matrices, starting from the calculating position frequency matrix, followed by position probability matrix and then the creation of PWMs [11].

Polymorphisms usually don't contribute in causing any abnormality, however, several polymorphisms collectively can cause abnormality. Mutations are different from the polymorphisms, they can drive any abnormality i.e., cancer, especially if they are located in binding site of transcription factors that can lead to the dis-regulation of genes. If the highly consensus sequence is mutated in the TF binding motif then, it would have more effect on the function than if mutation is located in lesser consensus sequence.

1.1.3 Cis-acting regulatory regions

Cis-acting regulatory regions are regions of non-coding DNA that control the transcription of near-by genes. These regions regulate genes by acting as binding sites for TFs. Regulatory regions can be located upstream or downstream from

the genes that they control, a single region can control more than one gene and a single gene can be regulated by more than one regulatory regions. These regulatory regions can interact with the promoter of the concerned gene by looping and these looping interactions can be identified by the Hi-C technique [12].These controlling elements can be enhancers, repressors and silencers of the transcription. Enhancers binding on TFs can be seen in Figure 1.2. The enhancer trap method is commonly used to experimentally identify cis regulatory regions but here, we have identified enhancers from ChIP-seq, DNase-seq and RNA-seq data using statistical and machine learning methods.

1.1.4 Chromatin region

Chromatin consists of DNA and protein. DNA is wrapped around histone proteins to form nucleosomes. The formation of chromatin 1) allows the DNA to be packed into a smaller volume to fit in the cell; 2) strengthens the DNA macromolecule to allow mitosis; 3) stops DNA damage; and 4) controls gene expression and DNA replication.

1.1.4.1 Histone modifications

Histone modifications have an important role in transcriptional regulation, so here we also discuss histones and their modifications for understanding regulation of genes. There are 8 histone proteins wrapped by DNA in nucleosome that is a basic unit of DNA packaging. H2A, H2B, H3 and H4 are 4 core histones and each has 2 copies that are linked by linker DNA [13]. Histone modification is the covalent post-translation modification (PTM) to histone proteins, which can influence the gene expression by altering the chromatin structure or recruiting histone modifiers. There are at least eight different types of histone modifications which include methylation, acetylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, Deimination, and Proline isomerization, all these modifications have role in transcription [14]. Some examples of histone modifications involved in transcriptional regulation is detailed in the Table 1.1.

Table 1. 1: Histone modifications involved in transcriptional regulation

Type of modification	**H3K4**	**H3K9**	**H3K14**	**H3K27**	**H3K79**
mono-methylation	Activation [15]	Activation [16]		Activation [16]	Activation [16]
di-methylation		Repression [17]		Repression [17]	Activation [18]
tri-methylation	Activation [19]	Repression [16]		Repression [16]	Activation [18], Repression[16]
acetylation		Activation [19]	Activation [19]	Activation [20]	

H3k4me3 (trimethylation of histone H3 at lysine 4) modification is mentioned in the Table 1.1, this modification loses the chromatin condense structure. This modification recruits chromatin remodelling factors such as CHD1 and BPTF and these factors open the chromatin that become accessible for transcription factor binding and this TF binding regulates the transcription of gene [21]. H3K27ac (acetylation of histone H3 at lysine 27) modification represents the active regulatory region [20]. H3K79me3 (tri-methylation of histone H3 at lysine 79) modification activates the transcription in yeast cells but have role of repressor in human T cells [16].

1.1.4.2 Chromatin looping

Genomic regions i.e., promoters and enhancers interact with each other through looping. DNA bends itself because of acetylation so that regions of genome can interact with each other for regulation of genes. There are factors such as CTCF, which harbours insulator activity when they are present between enhancer and gene promoter [22].

1.1.4.3 Epigenetic regulation/ DNA methylation

DNA methylation is one of the mechanisms that cells use to regulate genes. DNA methylation is a process by which methyl groups are added to the DNA molecule. This addition can change the activity of a DNA segment without changing the DNA sequence. Usually methylation act as a transcriptional repressor if it is located in gene promoter. Methylation occurs at the CpG sites where Cytosine is methylated to form 5-methylcytosine [23], dense methylation is known to involve in silencing of CpG rich promoters [24]. It has been observed that methylation occurs at CpG islands remain consistent between tissues, normal and cancer samples. However, differences in methylation events have been observed at a short distances from CpG islands [25]. It has been observed that CpG islands are less susceptible to change than other regions, therefore, they remain conserved. Some of the researchers have identified conserved promoters from the CpG islands and they also observed that these CpG islands co-localize with the H3K4me3, which suggest their role in gene regulation [26].

1.2 Available data

1.2.1 ENCODE

Transcriptional regulation involves regulatory elements for example TFs and regulatory regions where TFs bind to control the expression of genes as discussed above. In order to understand the transcriptional regulation of genes, we need TF binding, open chromatin and gene expression data. The ENCODE (Encyclopedia of DNA Elements) consortium has generated this data, and that is explained below.

The ENCODE consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI), this project was planned after the completion of Human Genome Project to understand the functional elements in the genome. This project involves more than 30 research groups and more than 400 scientists. The major aim of this project is to annotate

protein coding genes, non-coding regions and pseudogenes. Only 1 percent of the human genome encodes for approximately 20,000 genes; scientists expected from this project to help us in understanding the functional role of the non-coding part of the genome. The purpose of ENCODE project is to link variable expression levels with the development of disease. This project has provided data for understanding the functioning and regulation of the genes. Data has helped us to understand that abrupt level of gene expression is linked with the cancer.

There are also some controversies over the ENCODE project, specifically regarding claim of this project that 80% of genome have function which contradict with the perception that 98% of human genome is junk DNA. Several papers have been published regarding this controversy [27]. Despite controversies, this projected has created impact in molecular biology, helped in understanding regulatory mechanism of genes. Still, there is a lot to be mined from this huge dataset.

ENCODE has generated ChIP-seq data, DNase-seq, FAIRE-seq, 5C and RNA-seq data. DNase-seq, FAIRE-seq, and ChIP-seq have helped us in understanding the proximal and distal regulatory regions [28]. A schematic diagram showing methods used by ENCODE is shown in Figure 1.4.

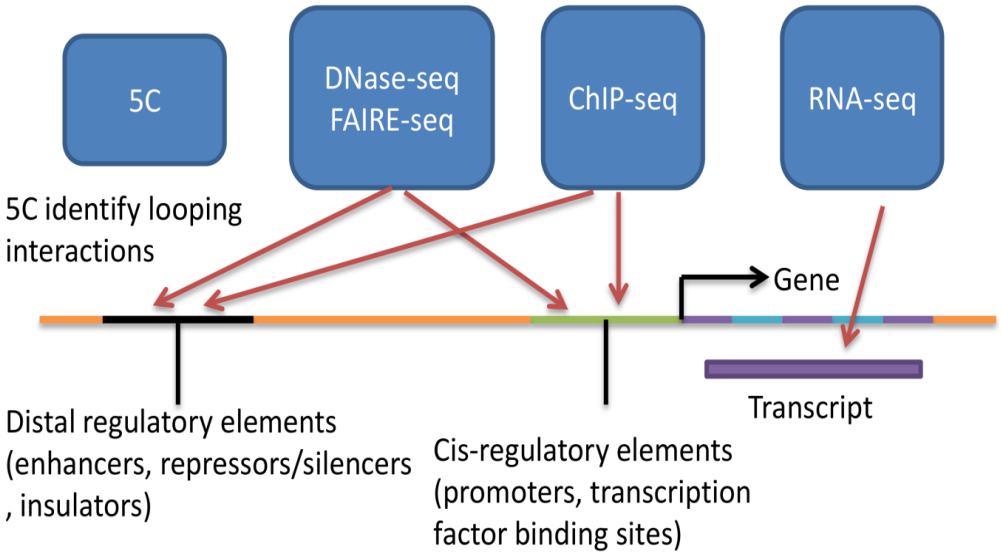


Figure 1. 4: ENCODE methods: A schematic diagram of major methods that are used by ENCODE and their relevance to detect functional elements. Reproduced from [28].

ENCODE has created a positive impact in molecular biology research. Computational biologists have developed several methods [29] and techniques from this data, and analysis of this data has led to the discovery of novel regulatory elements that can help to identify therapeutic targets for the cancer. Development of new methods and analysis of this data have been done by bioinformatics would help biologist to target specifically for identification of novel elements. Several studies have been performed in wet laboratory and their results have been analysed along with the analysis of the ENCODE data to produce novel findings. Scientist have also predicted cell type specific expression from the ENCODE DHS data [30]. This project has helped us in determining the role of biological factors such as replication timing, and gene expression on the rate of mutation [31].

Below, I have discussed some important methods such as ChIP-seq, DNase-seq and RNA-seq used by ENCODE and these techniques also have relevance in our study.

1.2.1.1 ChIP-seq

The ChIP-seq technique is chromatin immunoprecipitation followed by sequencing [32] and it is used to determine the interaction between proteins and DNA in the cell. There is also another technique called ChIP-chip to identify the interactions between the DNA and protein, this technology combines ChIP (Chromatin Immunoprecipitation) with DNA microarray (chip). In ChIP-seq, an antibody is used to bind to a specific epitope and high throughput sequencing is performed on an enriched sample to determine the binding sites in the genome most often bound by the protein to which antibody is directed. Antibodies can be used to any chromatin-associated protein including transcription factors, specific chemical modifications on histone proteins and chromatin binding proteins. After the mapping of reads to the genome, ENCODE has used different peak caller algorithms such as SPP, PeakSeq and MACs. The output of these callers generally ranks regions by absolute signal (read numbers) or by computed significance of enrichment (e.g., p values and false discovery rates). Each peak calling algorithm rely on different statistical methods to calculate the p values and

false discovery rates, hence p values from different algorithms can't be compared [33].

In the ENCODE project, 119 transcription factor binding sites have been mapped using ChIP-seq data on 72 tissue types, of which 87% were sequence specific transcription factors [34]. Recently the ENCODE Analysis working group (AWG) has re-processed all the ChIP-seq data uniformly using a high quality peak caller with irreproducible discovery rate (IDR), and the consortium has re-produced 690 ChIP-seq datasets representing 161 unique regulatory factors in 91 human cell types [35], but it is an ongoing process and probably now they have produced more data.

The ChIP-seq technique adapted in ENCODE is outlined in Figure 1.5. This procedure initially involves treatment of cells with the chemical agent, usually formaldehyde, for cross linking proteins to DNA. The next step is the disruption of cells, sonication or digestion through enzymes to split the chromatin into fragments of 100-300bp. This is followed by immunoprecipitation, which is the enrichment of the protein bound by DNA with a specific antibody. Cross linking of protein to DNA is then reversed after immunoprecipitation and the enriched DNA is purified. The DNA is analysed through high-throughput sequencing. ENCODE has generated replicates by using different antibodies for certain transcription factors [33].

ENCODE has generated uniform peaks which contain TF binding sites and they are in the BED format. There are tools which can be used for the analysis of BED files such as BEDOPS, that is an open source command line toolkit for comparing BED datasets with highly efficient and scalable Boolean and other set operations [36]. Another efficient tool for the analysis of BED files is the Bed tools [37].

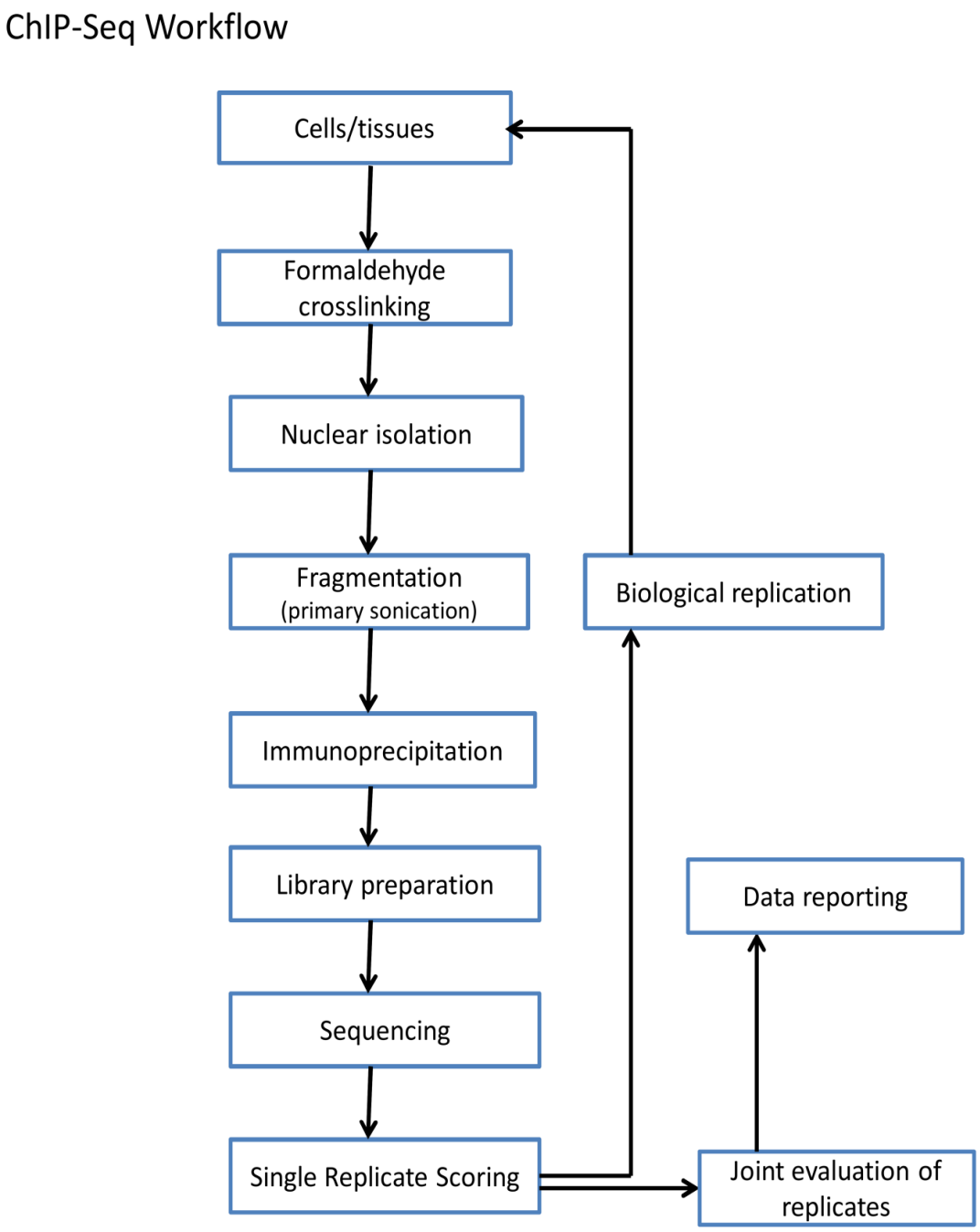


Figure 1. 5: ChIP-seq experiment workflow: All the steps involved in the ChIP-seq procedure are described in this figure. ENCODE has also generated replicates for certain transcription factors by using different type of antibodies. Reproduced from Landt, S.G., et al [33].

1.2.1.2 DNase-seq

DNase-seq technique is used to determine DNase I hypersensitive sites. The DNase I enzyme preferentially cuts chromatin preparations at exposed regions followed by high throughput sequencing to determine those sites ‘hypersensitive’ to DNase I, corresponding to open chromatin [38]. DNA in hypersensitive site is less compact and it allows DNA binding proteins such as TFs to bind there. ENCODE has mapped 2.89 million DNase I hypersensitive sites by DNase-seq technique in 125 cell lines [34].

A DNase-seq protocol is shown in Figure 1.6. The procedure for production of peaks and signal intensities are discussed in the ChIP-seq technique.

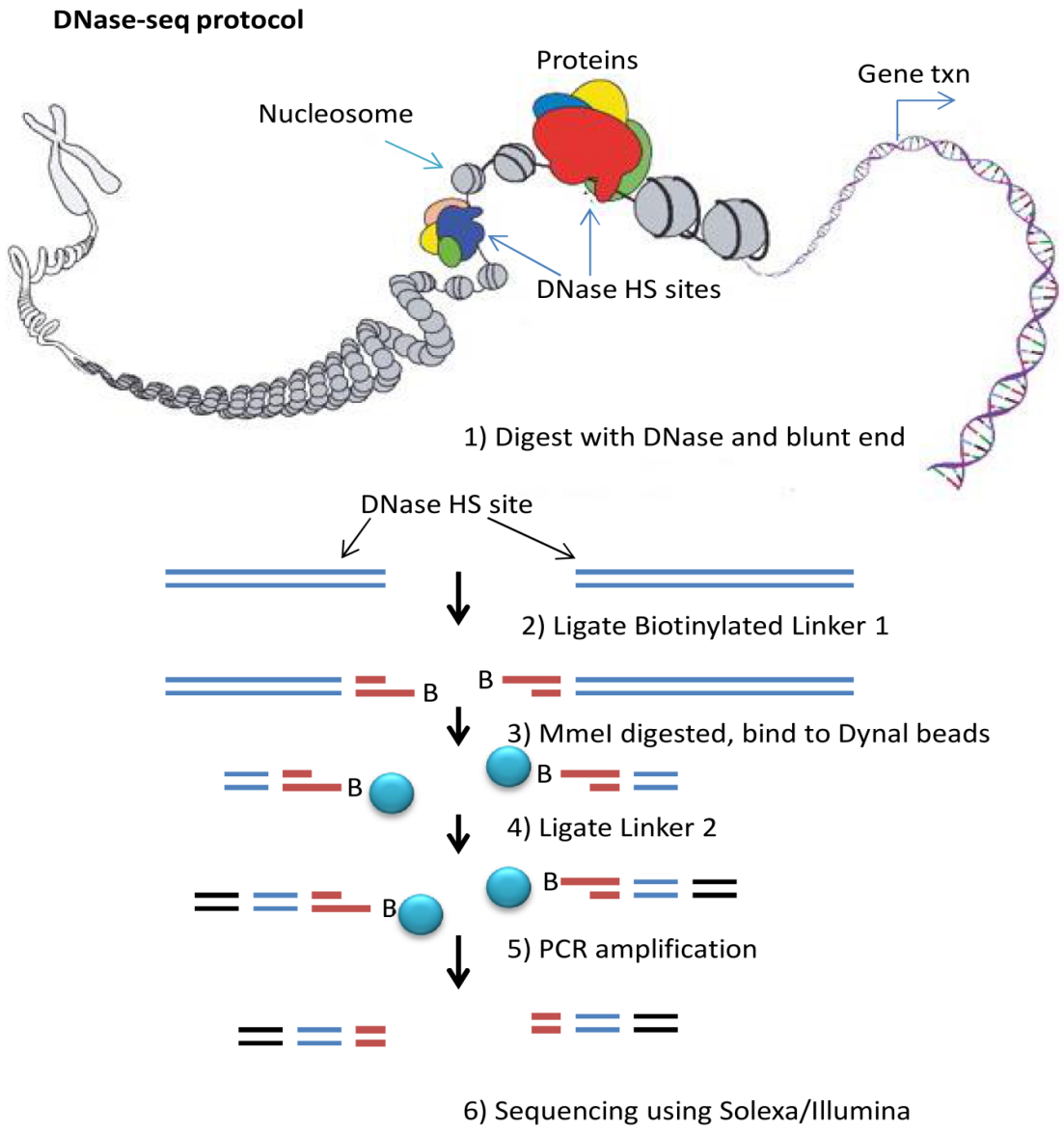


Figure 1. 6: DNase-seq protocol: This procedure starts with the cell lysis with detergent to release the nuclei. Then DNase I concentrations are used to digest the nuclei and the fragments are embedded in the low melt agarose plugs to decrease additional random shearing. Blunt-ended DNA strands are generated here, which are then ligated to biotinylated linker 1 (represented by red bar). This is followed by the removal of excess linker and biotinylated fragments by digestion with MmeI, and captured by streptavidin coated Dynal beads. Linker 2 (represented by the blue bars) is ligated to the overhanging DNA strands, which were generated by MmeI. PCR amplified the ditagged 20bp DNAs followed by the sequencing by Illumina/Solexa [39].

1.2.1.3 RNA-seq

RNA-seq is used to measure transcription and it involves the isolation of RNA sequences through different purification techniques followed by high throughput sequencing [40]. RNA-seq technique is a cost effective and powerful method for transcriptome analysis and this method is helpful in finding novel exons and junctions as it does not need probe selection [41]. An overview of the RNA-seq method is shown in Figure 1.7. This technique starts with the conversion of a population of RNA (fractionated either as poly (A) + or total) into a library of cDNA fragments with adapters attached to one or both ends. Each molecule is sequenced in a high-throughput manner with or without amplification to achieve the short sequences from one or both ends. The read size ranges from 30-400bp; these reads are then aligned to the reference transcript or reference genome after the sequencing to generate genome scale transcription map; which consists of both the level of expression and transcriptional structure for each gene. FPKM (Fragments Per Kilobase of exon per Million reads) values are measure of level of gene expression and can be calculated from RNA-seq reads. In FPKM, “Fragment” is the fragment of DNA formed by two paired end reads, “Per Kilobase of exon” is fragments counts are normalised by dividing with the total length of al exons in the gene, “per Million reads” means this value is again normalised against the library size [42].

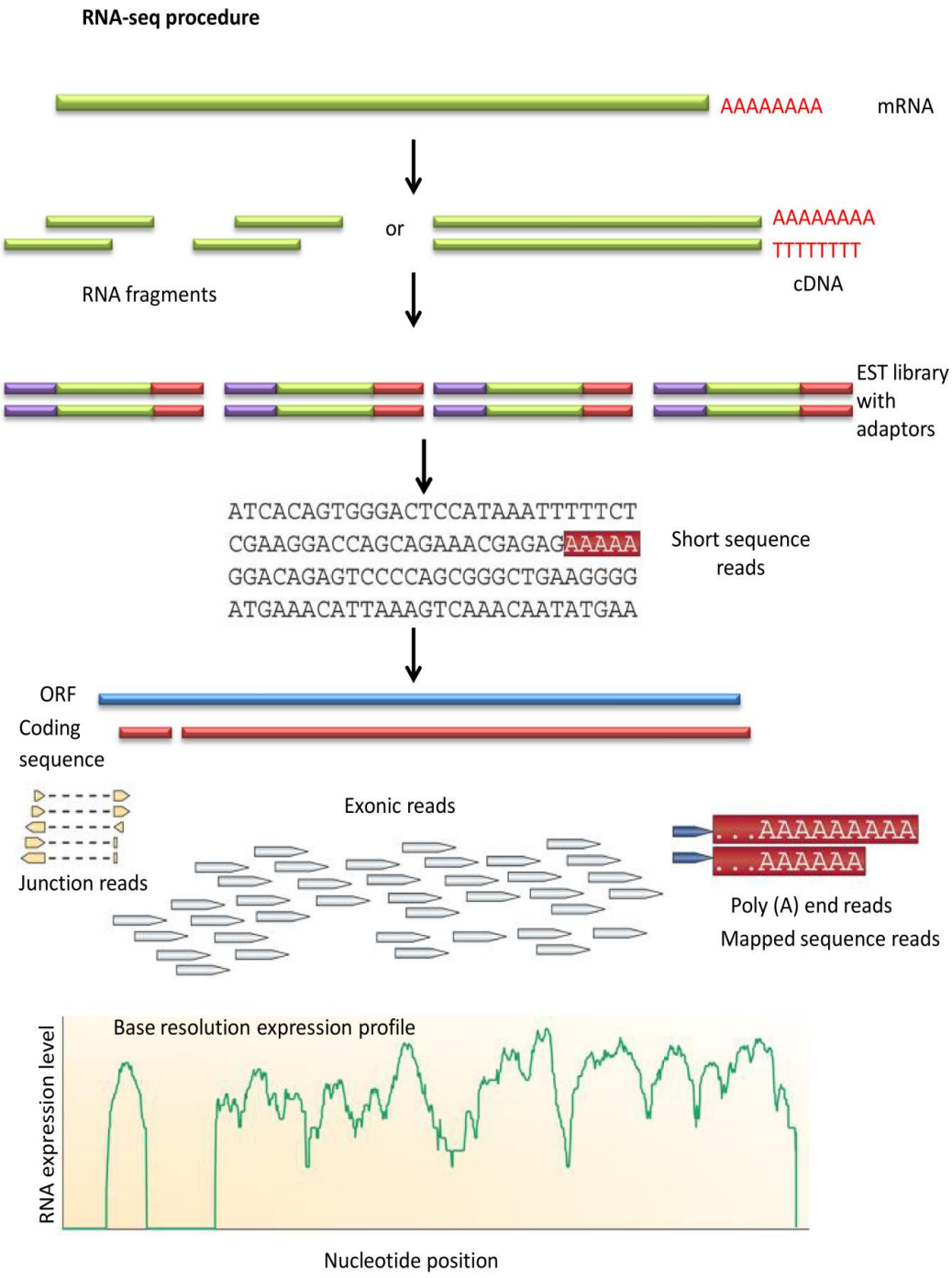


Figure 1. 7: RNA-seq: This figure shows the overview of RNA-seq. Reproduced from Wang, Z. et al. [42]. It starts with the conversion of mRNA into cDNA that followed by EST library and then sequence reads.

1.2.2 Conservation data

Functional regulatory regions can be evolutionary conserved and conservation data might be helpful in identifying novel regulatory elements that are important for controlling expression of genes. Genomic regions shared between cells and shared binding sites among transcription factors could be conserved because they can be functionally significant. Functionally important regions remain conserved even though living organisms have been evolving for millions of years but the rate of evolution is slow and spontaneous. Conservation information can be predictive of functional regions from non-coding DNA [43] and highly conserved elements are linked with the function [44]. Organisms share their genomic regions as they evolve from the same ancestors, and most of these shared regions are significant and necessary for the survival of these organisms. These shared regions in different organisms are called conserved regions and they can be located in cis regulatory regions. Organisms were grouped together according to their biological similarities such as vertebrates, insects, *Caenorhabditis* and *Saccharomyces*. Human is a vertebrate and conserved regions were identified with other vertebrates and other groups of eukaryotes.

Conserved elements of 46 organisms including vertebrates, insects, and *Caenorhabditis* and *Saccharomyces* were identified by a computer program called phastCons [45] . These conservation scores can help us in identifying potential cis-acting regulatory regions but conservation of these regions vary depending on the tissue types, for example, enhancers in myocardial cells are weakly conserved [46]. Adam et al. identified regulatory regions from conserved non-coding elements [47]. Pennacchio et al. have used conservation data for the identification of non-coding sequences such as cis-acting regulatory regions [48]. There are some known examples of enhancers where genes are controlled by conserved enhancers such as homeobox gene Hoxb-1, and this gene express earlier because of conserved retinoic acid response element [49]. Therefore, we have also used the conservation analysis for filtering the potential enhancers.

1.3 Cancer

Alteration in transcriptional regulation can lead to high and low levels of gene expression that result in abnormal cell growth, and this unwanted growth in cells can cause cancer. Transcriptional regulation works in mechanistic way, deregulation of cell proliferation and supressed apoptosis (programmed cell death), together can cause cancer [50].

Studying transcriptional regulation can be helpful in identifying therapeutic targets for cancer. This is common disease and it can occur in all age groups. The most commonly occurring cancers in men and women are prostrate and breast cancer respectively. In United States, this disease is responsible for 25% deaths [51]. In United Kingdom, cancer causes 29% of total deaths in 2011. There are more than 200 types of cancer and there are five different groups of cancer such as carcinomas, Lymphomas, Leukaemia's, Brain tumours, and Sarcomas, these groups are based on the type of cell from cancer starts. In United Kingdom, almost of half of all cancer deaths are because of lung, bowel, breast, or prostate cancer in 2014. Cancer can occur because of molecular changes in the cells, and usually it starts from a single cell. Mutations disturb the mechanism of cell regulation that results in altered gene expression which ultimately leads to the cancer, mostly these mutations are somatic. In several studies, it has been observed that high level of gene expression is linked with the cancer [52]. Certain studies show that cancer somatic mutations are influenced by the certain factors such as base pair composition, and replication timing [53].

There are several databases of cancer somatic mutations such as TCGA (The Cancer Genome Atlas) [54], ICGC (International Cancer Genome Consortium) [55], and COSMIC (Catalogue of Somatic Mutations in Cancer) [56]. Only COSMIC database is manually curated cancer database and this also contains genes which are known to involve in cancer. Most of the ENCODE cell types are cancer cells such as Helas3 and Hepg2 that suggest that their data can be used for understanding the regulation of genes, specifically cancer genes.

1.4 Statistical methods of machine learning

Transcriptional regulation genome wide involves large number of regulatory regions and it is challenging to link these potential regulatory regions to their target genes. We need to filter those regulatory regions that are predictive of gene expression for their respective target genes. Statistical methods of machine learning can be helpful in predicting potential regulatory regions.

Machine learning is a technique to learn for predicting the better results in the future based on what they have learned in the past. Machines can be trained with the existing information and they can predict on the basis of their training. Machine learning generally used in optical character recognition, face detection, spam filtering, topic spotting, spoken language understanding, medical diagnosis, customer segmentation, fraud detection and weather detection. In addition, this technique can be used to predict gene expression. Brown and co-workers used support vector machines, a type of machine learning to classify the microarray gene expression data [57]. Different cancer outcome such as diffuse large B-cell lymphoma can be predicted by machine learning [58]. Machine learning is helpful in analysis of genomic datasets including genomic, metabolomic or proteomic data. Machine learning can be supervised, semi-supervised and unsupervised and each has different applications for different problems. Supervised learning is analysing data wherein classes are already assigned. However, unsupervised learning involves analysing data wherein classes are not known.

These machine learning methods can be used to identify splice sites, promoters, enhancers and transcription start sites (TSS) [59]. Cross-validation has been used to test the performance of machine learning algorithms; a single cross validation is equal to one observation to test how our algorithm performs. We have used different statistical methods in this Thesis and they are explained in the introduction of their respective chapters.

1.5 Thesis objectives and structure

Transcriptional regulation involves binding of TFs on cis regulatory regions and these TFs form complex to switch on and off the transcription of genes. Understanding transcriptional regulation requires study of TFs mutual interactions, binding pattern of TFs in different cell types, conservation analysis of TF binding sites, influence of TF binding on gene expression level, identifying cis regulatory regions (enhancers and repressors) and linking them to their respective genes. We have studied all these elements of transcriptional regulation in following chapters. (We have used hg19 version of genome across this Thesis).

Chapter 2

Transcription factors interact with each other and bind on the DNA. In this Chapter, we have developed methods from ENCODE ChIP-seq data for prediction of TF-TF mutual interactions using two statistical methods i.e., Poisson distribution and randomisation.

Chapter 3

We have done conservation analysis of shared (overlapped sites among different cell types) TFBS vs. cell type specific TFBS (Transcription Factor Binding Sites); and overlapped TFBS vs. non overlapped TFBS within a particular cell type; here we also looked for influence of co-bound and single bound TFBS on gene expression.

Chapter 4

Prediction of cis regulatory regions is significant because this prediction would help experimentalists to validate cis regulatory regions and to understand gene regulation. Here, we have built linear models by integrating candidate cis regulatory regions (CRRs) obtained from ChIP-seq and DNase-seq with RNA-seq data, and potential regulatory regions were chosen by four methods mostly based on genomic features for example TF binding and conservation score.

Chapter 5

In this Chapter, we have built correlative models using LASSO (Least Absolute Shrinkage and Selection Operator) in two different approaches. In 1st approach, LASSO built models with all possible CRRs (on average 42) per transcript, and in 2nd approach we gave CRRs filtered by higher TF binding and higher conservation score.

Chapter 6

Regulatory regions predicted in previous chapters can be helpful in understanding cancer somatic mutations, knowing that these regions obtained from ENCODE cell types and most of these cell types are cancerous. Therefore, there is relevance in mapping somatic mutations on these regulatory regions. Here, we tested the significant difference in accumulating cancer somatic mutations between regulatory regions chosen and rejected by LASSO.

Chapter 7

Discussion and future work.

Chapter 2

2 Preliminary studies of potential methods for predicting transcription factor interactions by binding site overlap analysis

2.1 Introduction

Transcription factors have an important role in the regulation of genes and their role is explained in the 1st chapter. These factors are proteins which bind directly or indirectly on cis-acting regulatory regions. Direct binding involves binding of a transcription factor directly on the cis regulatory regions, while indirect binding is the binding on the cis regulatory regions via other transcription factors. Transcription factors mutually interact with each other and form complexes for the combinatorial regulation of genes. Therefore, predicting mutual interactions is important for understanding the regulation of genes.

2.1.1 TF Co-association

Transcription factors co-associate (mutually interact) in a combinatorial and context-specific fashion. Different combinations of factors bind different targets and their binding affects each other. Moreover, transcription factors often show different co-association patterns in gene-proximal and distal regions [60].

Gerstein et al., developed a method, where they focus on specific region bound by a transcription factors and examined the binding of all other transcription factors in that region. They generated a co-binding map by obtaining normalised binding signals of overlapping peaks of all TFs [60].

Similarly, Kazemian and co-workers have predicted transcription factor interactions by overlapping the transcription factor binding sites. They assessed the significance of co-binding of a TF pair by testing the over representation of

particular orientation(one binding site for each TF in a TF pair) using a binomial test [61].

The ENCODE project has systematically mapped regions of transcription, transcription factor co-association, chromatin structure and histone modification. The project reported co-association involving 114 out of a possible 117 transcription factors in proximal and distal regions. These include known associations, such as Jun and Fos, and some novel associations, such as TCF7L2 with HNF4-α and FOXA2 (Figure 2.1). They have also identified regions bound by multiple transcription factors representing high occupancy transcription factor regions [29].

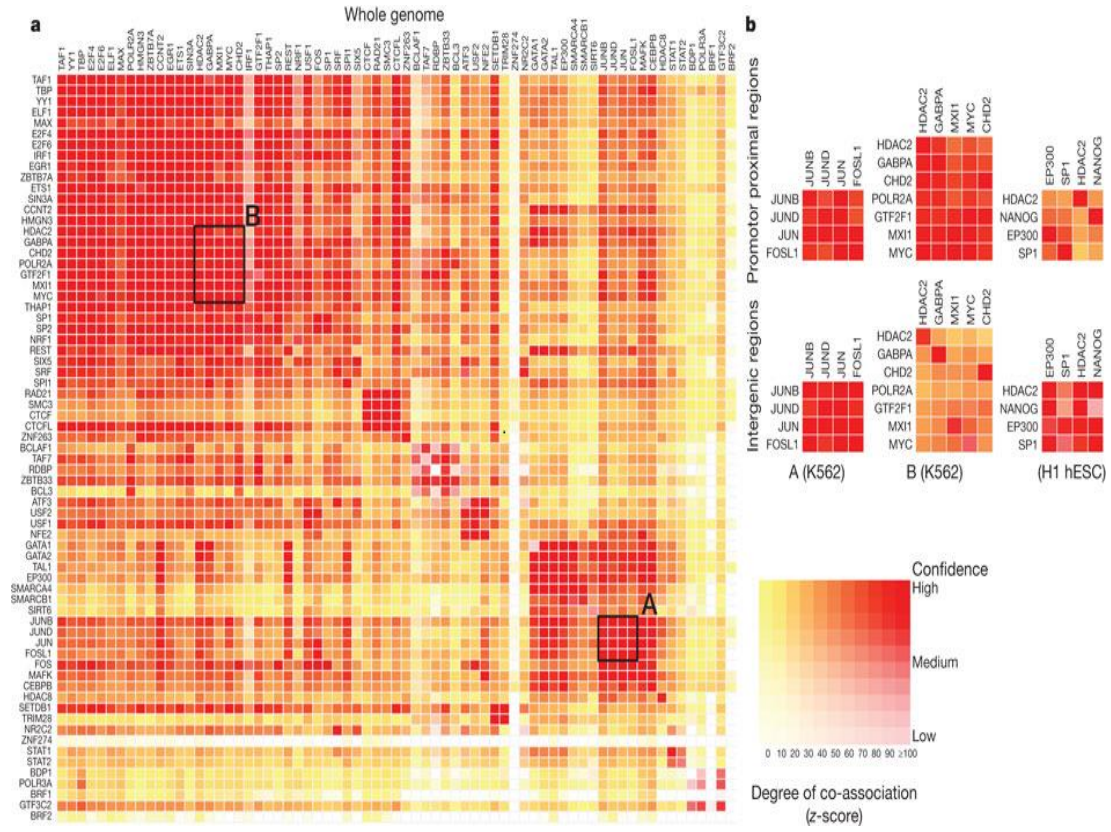


Figure 2. 1: Co-association between transcription factors. **a**. Co-associations of transcription factor pairs across the entire genome in K562 cells. The colour intensity shows the extent of association (from red (strongest), orange, to yellow (weakest)). **b**. Three classes of behaviour are shown. The first column shows a set of associations between TFs independent of location in promoter and distal regions; whereas the second column shows a set of associations in promoter-proximal regions. Both of these columns are highlighted on the genome-wide co-association matrix (**a**) by the labelled boxes A and B, respectively. The third column represents a set of transcription factors that show stronger association in distal regions. Reproduced from [29]. (Copyright permission is not required for this figure as it is here only for education purposes).

2.1.2 Distinguishing indirect transcription factor occupancy

If two transcription factors have the same binding site then there can be several possibilities but main two of them are: 1. they can bind to same site at different times, and 2. Co-binding by two TFs when only one has recognised motif in that region suggests that other might bind indirectly. Neph and co-workers discovered

many known protein-protein interactions such as CTCF-YY1 and TAL1-GATA1 as well as many novel associations by integrating ChIP-seq and DNase I footprint data. ChIP-seq peak containing DNase I footprint motif represents direct binding and ChIP-seq without compatible motif shows indirect binding. They also stated that transcription factors can be mutually interacting when the frequency with which indirectly bound sites of one transcription factor coincide with directly bound sites of a second factor is higher than expected by chance [62].

Gordan and co-workers developed a method to identify direct TF binding and indirect TF binding by using TF binding motifs. They concluded that a TF can interact directly with the DNA if it has compatible motif in ChIP-chip data otherwise it would interact indirectly. When they applied this to yeast ChIP-chip data, they found only 48% TF binding directly and 16% binding indirectly. In the remaining 36%, none of the motif was able to explain the ChIP-chip data because either the motifs set were incomplete or the data were too noisy [63].

In this chapter, we have developed two statistical methods based on the Poisson distribution and randomisation for prediction of TF-TF mutual interactions. Several other researchers have also predicted the co-association of transcription factors as discussed above but our methodology is different. Here, we have discussed results from two cell types i.e., Gm12878 and K562; however this method can be applied to other cell types and newly generated data.

2.2 Methods and data

2.2.1 Dataset

Publicly available ChIP-seq data from ENCODE (2012 release) which contains information on TF binding sites, was used in this study. This ChIP-seq data (uniform peaks) was retrieved and analysed. The data contains biological replicates for some transcription factors because it was produced by several laboratories sometimes using different antibodies. The ENCODE ‘Analysis working group’ has already generated uniform peaks (peak is a TF binding site) from these datasets by using the Irreproducibility Discovery Rate (IDR: measures

consistency between replicates) [64], and this analysis was used in the work reported here.

All 439 datasets with uniform peaks (ENCODE generated uniform peaks by using IDR) were retrieved for five cell types (K562, Gm12878, Hepg2, Helas3, and H1hesc). These cell types were selected because significant number of transcription factors binding sites were mapped. Table 2.1 shows the number of transcription factors mapped by ENCODE in these five cell types along with their corresponding tissue types and karyotype.

Table 2. 1: This table shows the number of transcription factors mapped in five cell lines along with the tissue types and karyotype.

Tissue type	Cells	Karyotype	No. of TFs mapped in each cell type
K562	Erythroleukemia/ bone marrow	cancer	100
Gm12878	EBV transformed B-cell lymphoblastoid	normal	73
Hepg2	Liver hepatocellular cells	cancer	57
Helas3	Cervix/adenocarcinoma	cancer	54
H1hesc	Human embryonic stem cells	normal	47

2.2.2 Methods

These datasets also contain binding sites of CTCF, CTCFL, POL2 and POL3, but these proteins were not considered for prediction of TF-TF mutual interactions. The ENCODE has generated biological replicates using different antibodies for a single transcription factor; replicate with highest number of binding sites were considered for particular transcription factor.

In the next step raw data was processed because chromosomes were not in the order so Bedtools sort [37] was used to sort them. Within each dataset binding sites that are overlapped by at least one base pair were merged into a single site by using Bedtools merge [37], because they belong from a single binding site. Overlapping of TF binding sites was assessed as significantly overlapping TF pairs interact with each other. Here sites will be considered overlapping if they overlap by at least one base pair.

A Python program was written to identify overlapping binding between transcription factor pairs.

The fraction of overlaps for transcription factor pairs (i.e., X and Y) was calculated using the following formula:

1. Fraction of X overlaps with Y = Number of overlaps between X and Y / Number of peaks in X

2. Fraction of Y overlaps with X = Number of overlaps between Y and X / Number of peaks in Y

A high percentage of overlaps at the actual peak sizes given by ENCODE were found because peaks are much bigger than the actual TF binding site of the transcription factor. This is controlled by the size of sequence fragments typically ~ 200 bases. The binding site is usually located in the centre of the peak and its size varies from TF to TF [65]. Therefore, it was decided to fix the size of the peaks by extending a set distance on each side from the centre of the actual peaks.

To check the overlaps at fixed sizes, the summit (centre of the peak) was calculated by adding chromosomes start of the peak (2nd column) with coordinate (10th column) in narrow peak format as shown in **Appendix I.**

Subsequently peak sizes were extended using Python program and Bedtools slop (i.e., a tool used for extending peak on both sides). We decided to check overlaps at peak sizes 20, 50, 74, 150, 350, 600, 800, 1000, 1500 and 2000; and also required to finalize the peak size where significant overlaps can be found for each TF pair.

P values were required to check whether the overlaps are real or by chance. Therefore, two statistical methods, randomisation and Poisson distribution, were optimised to test the significance of the overlaps. Here, we set a null hypothesis that the transcription factors bind independently to the genome. In this case, we would still expect some level of overlapping binding sites where they happen to bind on the same place by chance. If there are more overlaps than expected, then that is the evidence that TFs tend to bind in the same place. The null hypothesis would be rejected and that is the evidence that they might interact in some way. So statistical test tells us about significant overlaps and are evidence of interaction. There can be some problems with these statistical tests that binding might not be equally likely everywhere in the genome and in the extreme case if in reality parts of the genome were inaccessible to binding by any TF. To tackle such issues, we have considered accessible genome that is derived from the TF binding sites and we have optimised it by just considering % of accessible genome size.

2.2.2.1 Randomisation

Randomisation is a process of producing a sequence of random variables illustrating a process whose result cannot follow a deterministic pattern. Randomisation has been used in the clinical trials and in designing experiments. Mao et al. used randomisation for significant gene selection using gene expression data where they used the partial least squares discriminant analysis (PLSDA) models to test the significance of the genes for classification of cancer [66]. Randomisation was used here as a significance testing approach. We have optimised a method on the basis of randomisation for the prediction of TF-TF mutual interactions and overlaps of transcription factors were evaluated either they are random by chance or real.

To assess the probability of chance overlaps between the binding sites of TF1 and TF2 the peak size S (for example S=20) was first fixed as described above. The peaks for TF1 were then treated as reference set. Peaks were considered to overlap if the centre of 1 peak lies anywhere within the other peak, and this gives a number of actual overlapping peaks for TF1 and TF2. To assess the probability

of random overlaps the union set of binding sites for all TFs was used. Sets of peaks containing the same number of peaks as the actual TF2 set were then randomly selected from this union and the actual overlap with TF1 peaks determined. Repeating this for 5000 such random sets allow the calculation (also calculation of Mu which is the average number of overlaps in the random sets then p values were calculated from Mu and real overlaps) of p value as the proportion of random sets having the same or greater overlap as seen in the actual set.

2.2.2.2 Poisson distribution

The Poisson distribution is a discrete probability distribution which applies to any process that produces count data from independent events with a fixed average count, and it has been used in testing the transcript expression profiles [67] [68]. Here, we have used the Poisson distribution to evaluate the overlaps of transcription factor pairs.

The probability of binding events per region can be calculated by following equation

$$\mathbf{P(k)} = \frac{\lambda^k e^{-\lambda}}{k!}$$

K=number of overlapping sites

λ = It is the expected number of overlapping binding sites based on the null hypothesis of independent binding sites to the genome

In the Poisson distribution, expected λ was calculated by following formula:

$\lambda = N_1 N_2 / N_{ag}$

Where:

N_1: number of bases covered by peaks from TF2 ((number of peaks) x (peak size))

N_2: number of peaks for TF1.

N_{ag}: Accessible genome size. Nag was estimated from union set of all binding sites from all TFs

Estimating accessible genome size was challenging and we have discussed it in the below subsection.

2.2.2.3 Optimisation of method

Significance of our method will increase if the lambda of Poisson distribution is similar to the Mu of the randomisation (lambda and Mu were discussed above). Peak sizes are required to fix for each transcription factor pair because ENCODE ChIP-seq peak sizes are bigger than the actual TF binding site. Therefore, transcription factors were overlapped at 20, 50, 74, 150, 350, 600, 800, 1000, 1500 and 2000 base pairs. Peak sizes for each transcription factor pair were chosen where the overlap significance was highest after multiple testing. Higher significance level indicates that transcription factors are overlapping more than the expected by chance at chosen peak size.

Results show that most of the TF pairs have highly significant overlaps. We might have over-estimated the size of accessible genome. To reduce over-estimation, we considered 75%, 50%, 40%, 30%, 20% and 10% size of accessible genome and compared the resulting significant TF-TF interactions. Size of accessible genome was optimised at 20% because at this size most of the known TF interacting pairs are overlapping significantly and overlaps for novel TF-TF interacting pairs can be ranked in order of their significance.

2.2.2.4 Multiple testing correction

Multiple testing correction is an important step to identify the false positives or to identify false significant interactions in this case. As we have discussed above, different peak sizes were used such as 20, 50, 74, 150, 350, 600, 800, and1000 base pairs, and p values were calculated from Poisson distribution for each peak size separately. These p values were extremely small, so it was not possible for even python float to handle them because the programming language considers every extremely small value as a zero but we need to rank actual values for multiple testing correction. Therefore, it was decided to take negative log of the p values and divide this resultant negative logarithm with 10. These logarithm

values were converted into the p values for multiple testing correction. We corrected these p values globally and concatenated all the p values from 20, 50, 74, 150, 350, 600, 800, 1000 base pairs and applied the Benjamini Hochberg test [69], a type of multiple testing correction. Again, we took the negative log of corrected p values. Peak sizes were then chosen if a particular TF pair has highly significant interaction (lowest corrected p value or highest negative log of it).

2.2.2.5 Validation of method

Known protein-protein interaction of TFs were retrieved from Biogrid (V.3.2.114) [70] and IntAct (downloaded data on 28th February 2014) [71] databases. In addition to that orthologous information on TF-TF mutual interaction in mouse was retrieved and all these interactions were unified into a single matrix to correlate with the p values calculated from randomisation and the Poisson distribution (with significant TF pair overlaps after global multiple testing corrections).

In the last step, p values were correlated with the known protein-protein interactions. All these known interactions were checked whether they have significant p values or not. We also checked those TF pairs which are not known interacting transcription factors but they have significant p values, so that these pairs might be novel interacting TF pairs.

2.3 Results

Methods were developed using the Poisson distribution and randomisation to predict mutual TF-TF interactions. Here, we have presented result of two major ENCODE cell types i.e., Gm12878 and K562.

2.3.1 Gm12878 cell type

ENCODE categorised Gm12878 cell line as a Tier-1 along with the K562 and H1 human embryonic stem cells (H1hesc). This lymphoblastoid cell line that

represents mesoderm cell lineage was produced from the blood of normal female donor. ENCODE has generated 90 datasets of this cell line representing 73 transcription factors and other regulatory elements.

2.3.1.1 Intersection (overlapping) of transcription factors

Initially transcription factors were considered with their ENCODE ChIP-seq peak sizes, which are usually large (referred to as actual peak size below), and subsequently all these transcription factors were overlapped with all possible combinations. Many transcription factors have high percentage of overlaps because they are overlapping at actual peak sizes or these overlapping transcription factors pairs are likely to interact [62]. These ENCODE actual peak sizes are larger than the size of the transcription factor binding site, as these transcription factor binding sites size ranges from 5-31 nucleotides, on average 10 nucleotides in eukaryotes [72].

We have discussed in the introduction of this chapter that binding sites are assumed to be located in vicinity of centre of the peak and it has larger size than actual binding site because of resolution of ChIP-seq. Therefore, it was decided to fix the size of the peaks to reduce false positives.

It is hard to decide whether the TF binding sites are overlapping because each transcription factor has different binding site size and has a different protein size. For example, if the protein size of the transcription factor is big and even if it has a small binding site, it will occupy a fragment on the DNA (in case of direct binding) that is proportional to the protein size. Gene specific transcription factors are smaller in size compared with those transcription factors which recruit mediators, RNA polymerases, histone modifiers and nucleosome remodellers. The size of gene specific transcription factors is ~50 kDa and size of the recruiters (large protein complexes) range from 1 to 3 MDa [73]. Therefore, it is likely that each TF pair will overlap significantly at different peak size.

We overlapped TF1 peak centre with the TF2 with different peak sizes (20, 50, 74, 150, 350, 600, 800, 1000, 1500 and 2000 base pairs) separately. It determines TF1 is lying anywhere in the binding site of TF2. Ultimately statistical

significance of the overlap will decide that either the transcription factors would interact or not.

There are two things which need to be considered: 1) To test the significance of overlaps by using statistical tests such as randomisation and the Poisson distribution; and 2) To optimize peak size for each transcription factor pair to account for the difference in protein size and binding site size.

2.3.1.2 Statistical significance of the overlaps.

Two statistical tests, randomisation and the Poisson distribution, were performed to evaluate the significance of overlaps between transcription factor pairs.

We started with the randomisation, which is a time consuming process. A large number of random sets (5000) were generated from the union of all transcription factors binding sites with actual ChIP-seq peak sizes as discussed in the methods section.

However, Poisson distribution is simpler than randomisation and we have developed both methods successfully, as lambda of Poisson distribution is similar to the Mu of randomisation (Mu is the average number of overlaps in the random sets). The Poisson approximation is confirmed by the randomisation process, therefore, p values are likely also to be similar. Actually, similar results increases the significant of the results as Poisson distribution may have some false discovery rate (may have poor fit) [74] or randomisation may result in false positives.

Figure 2.2 shows scatter plot, where Mu of randomisation is plotted against lambda of Poisson distribution. Both values are well correlated as shown in the Figure.

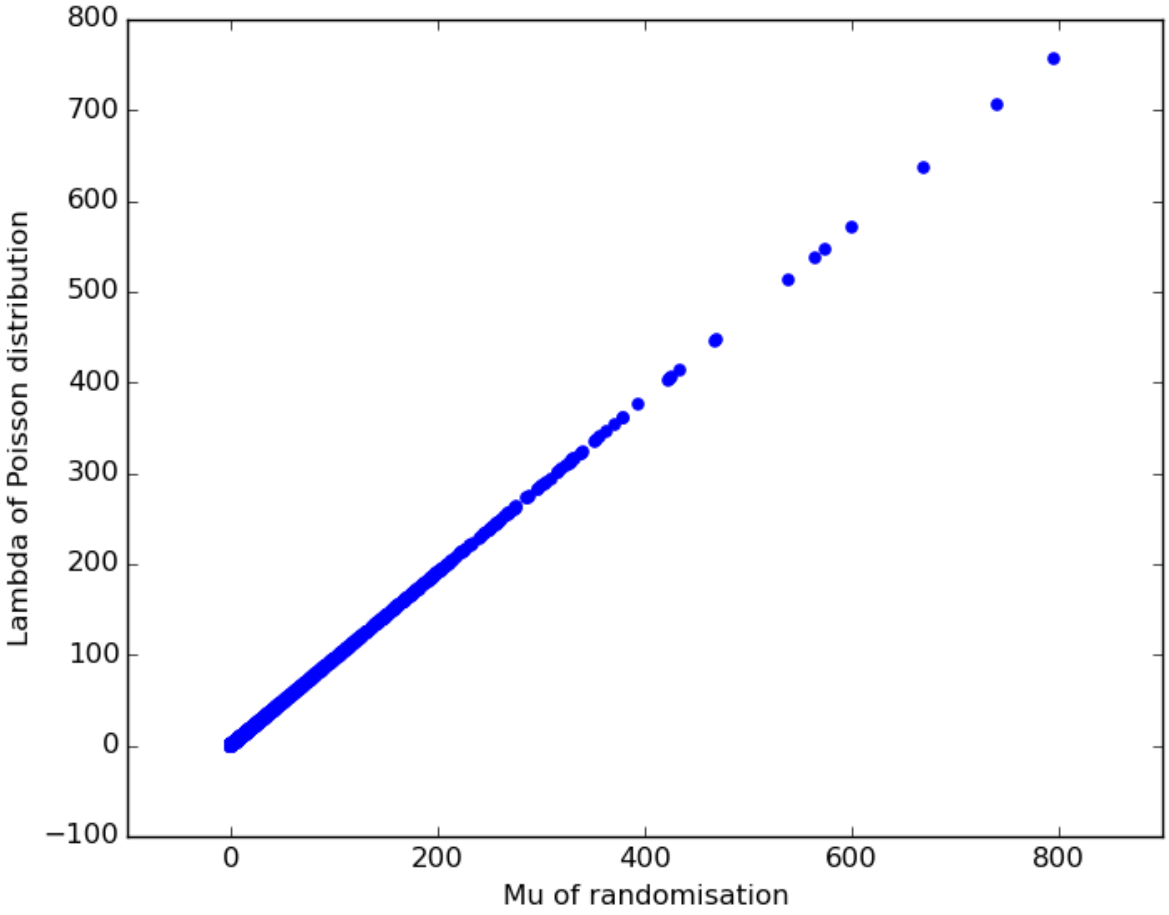


Figure 2. 2: Here Mu of randomisation is plotted against the lambda of Poisson distribution and both values are similar as they are well correlated.

2.3.1.3 Optimised results

After correcting p values by multiple testing correction and choosing the particular size of the peak for each TF pair as discussed in methods section, then we ranked TF pairs by the negative log of corrected p values and by ratio of real and expected overlap. A total of 1006 TF pairs with highly significant corrected p values were selected (1006 TF pairs have p values<0.01, this p value is a lot lesser but was normalised to separate highly significant predicted interacting TF pairs). These 1006 TF pairs were ranked by the negative log of corrected p values and ratio of real and expected overlap. Ratio of real and expected overlaps and logarithm of corrected p values are shown in Figure 2.3 and 2.4, respectively. In Figure 2.3, ratio of real and expected overlap was calculated by the real overlaps/expected overlaps and these expected overlaps (lambda) were

calculated by the Poisson distribution. There are fewer outliers that have extremely high ratio of overlaps. For example, Brca1-Zbtb33 TF pair has 124.45 ratio of overlap as its expected overlap is low (3.83) and real overlap is high (477), Brca1-Corestsc has 67.9 ratio, which is the 2nd highest ratio. Only 7 TF pairs have ratios of overlap higher than 20. However, a high ratio of real and expected overlaps in TF pairs does not mean that these pairs interact with high level of significance.

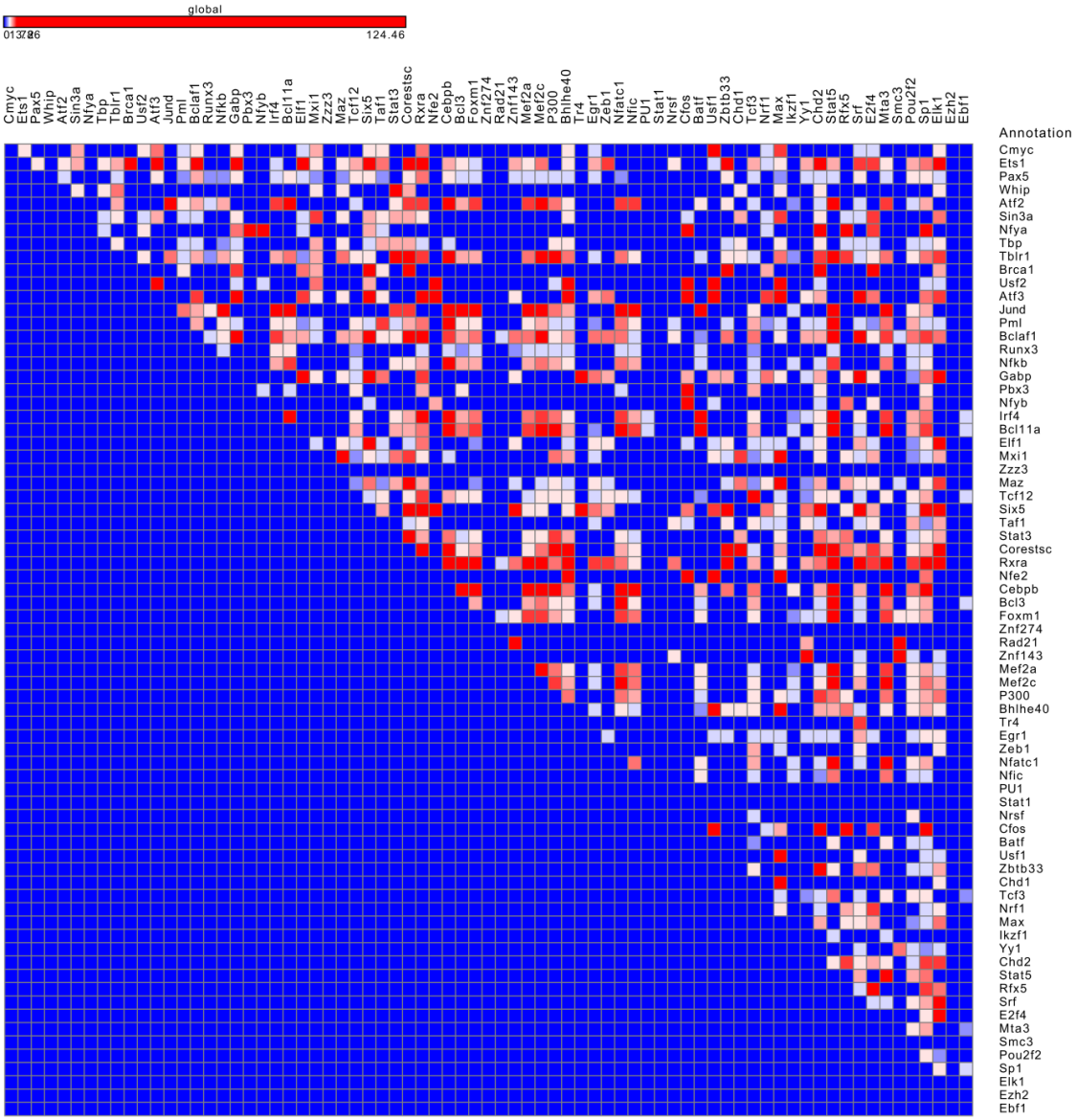


Figure 2. 3: (Only upper triangle). This heat map shows the ratio of real and expected overlap for 1006 TF pairs considered for analysis are represented by all the colours except blue. TF pairs were ranked by the ratio of real/expected overlap and top 300 pairs are represented with the red colour.

There are several TF pairs with p values that are highly significant, with high ratio of real and expected overlap such as Ets1-Elk1, Cfos-Sp1 and several others shown as red squares in the Figure 2.3 and 2.4.

The top 1006 TF pairs with high negative log of p values are shown in the Figure 2.4. There are a few cases where the TF pair has significant corrected p values but low ratio of real and expected overlap as compared to the p value. Examples of such TF pairs are Pax5-Pou2f2 and Atf2-Sp1 transcription factor pairs.

There are cases where TF pairs have high ratio of real and expected overlap but their corrected p values are not low as compared to the ratio of overlap between real and expected. Examples of such cases are Rxra-Elk1, Cmyc-Usf1 transcription factor pairs. Transcription factors such as Jund, Batf, Nfkb, Irf4, Bcl11a, Atf2, Foxm1, Nfic, Bcl3, Cebpb, Mta3, Nfatc1, Stat5, Mef2a, Mef2c, and Tblr1 form cluster as shown in Figure 2.4. Most of these transcription factors have important role in Gm12878 cells. We can find several TF pairs overlapping significantly in Figure 2.1 (K562 cell type, predicted by ENCODE) and Figure 2.4 (Gm12878 cell type, predicted by our method) e.g., TAF1 and YY1. There are cases where TF pair is not overlapping significantly in both figures e.g., JUND-SIX5. There are also cases, where a TF pair is overlapping significantly in Figure 2.4, but not in the Figure 2.4 for example SIX5-SP1. This is possibly because, some transcription factors have cell type specific functions.

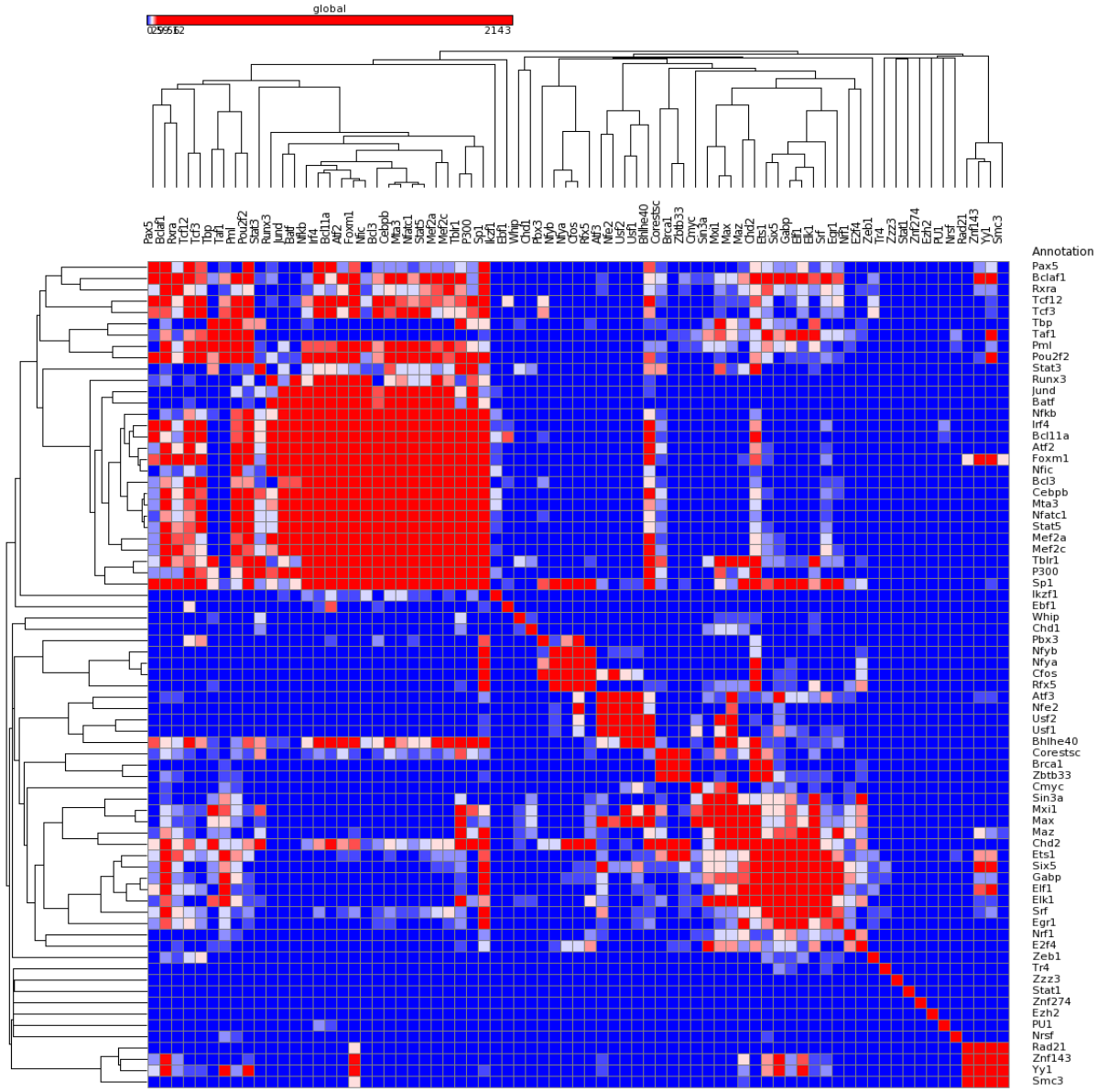


Figure 2. 4: This heat map shows the negative log of p values for 1006 TF pairs considered for further analysis are represented by all colours except blue. The red squares show TF pairs with highest negative log of p values. TF pairs with high significance of overlaps are clustered together by hierarchical clustering.

Peak sizes (binding sites of transcription factors) were optimised for each TF pair, with the maximum peak size of 350 base pairs. The actual binding sites are short ranging from 5-31 nucleotides, but here peak sizes are big because of the physical size of the transcription factor protein. As discussed above that some transcription factors bind directly on DNA and some bind indirectly [63].

Optimised peak sizes of 350, 150, 74, 50 and 20 base pairs are represented by white, light blue and red squares as shown in Figure 2.5.

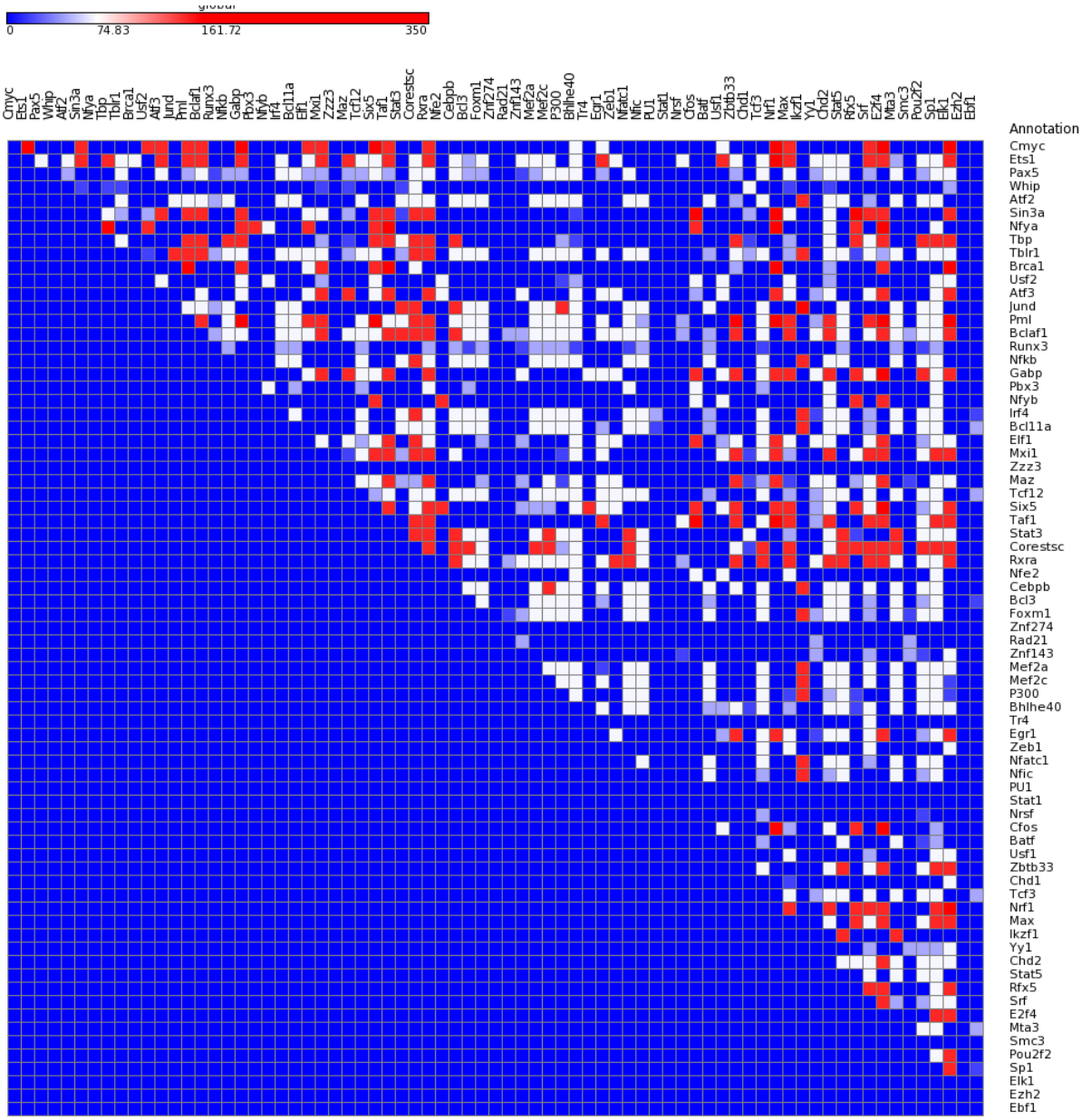


Figure 2. 5: This heat map (only upper triangle) shows peak sizes for 1006 TF pairs represented by all colours except blue. TF pairs with peak size 350 and 150 base pairs are represented by red colour. TF pairs with peak size 74 are represented by white squares, and pairs with peak size 50 are represented by half blue squares. While peak size 20 is represented by light blue squares. Scale is also shown on the top of this figure.

We plotted the optimal peak size identified by our method against the average size of the proteins involved but they did not correlate well possibly because only small portion of the proteins bind on the DNA. This plot is shown in Figure 2.6.

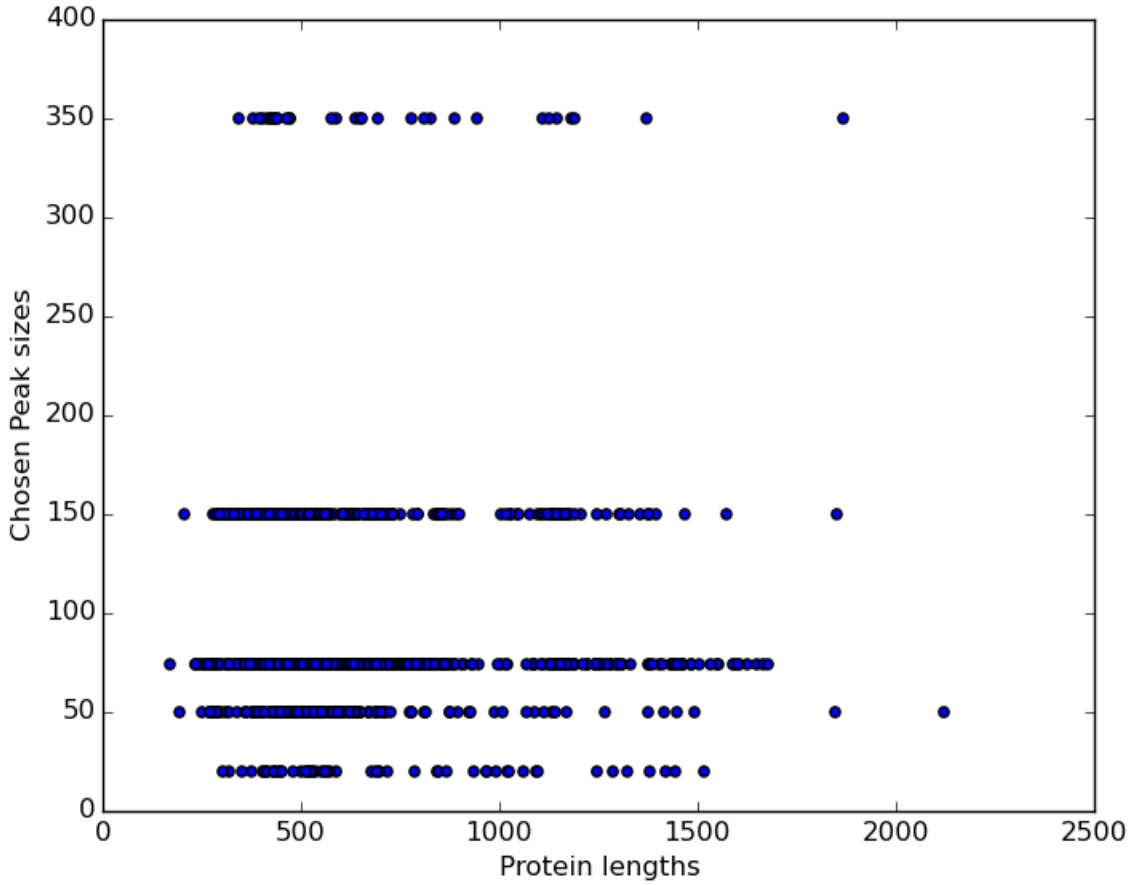


Figure 2. 6: Here, chosen (optimal) peak sizes are plotted against the average size of protein lengths and there is no correlation between them. This analysis was performed on 1006 TF pairs considered for analysis.

2.3.1.4 Validation (Comparison of significant overlaps and known protein-protein interactions)

Known interactions of transcription factors were retrieved from Biogrid [70] and IntAct [71] databases along with orthologous interactions in mouse which were also retrieved from IntAct database. Orthologous [75] interactions are also important because their proteins share common ancestry, suggesting that protein-protein interactions can also be conserved in different organisms [76].

Therefore, we have considered orthologous interactions as known interactions. Known and protein–protein interactions by our method are shown in Figure 2.7.

There are several known interactions and these have also been identified by our methods. Examples of such interactions are Cmyc-Pml, Nfya-Sp1, Srf-Elk1, Srf-Sp1, Rxra-Sp1, Rxra-Pou2f2, Srf-Elk1, Srf-Sp1 transcription factor pairs. There are several novel interacting transcription factor pairs that were predicted by our method, examples of such interactions are Cmyc-Elk1, Ebf1-Bcl3, Ebf1-Sp1, Ebf1-Mta3, Ets1-Elk1, Ets1-Sp1, Stat3-Elk1, Cmyc-E2f4, Runx3-Sp1 and Pax5-Elk1 transcription factor pairs. Statistics of all predicted and known interactions are detailed in Table 2.2. Those interactions which are predicted by our method and also known to interact are limited. Therefore, we got insignificant results when we applied hypergeometric distribution on known, not known, predicted and not predicted interactions. Result is not significant because several known interactions are might be cell type specific.

Table 2. 2: This table shows the summary of predicted, not predicted, known to interact and not known to interacting TF pairs- These known TF pairs are not enriched.

	Predicted	**Not predicted**
Known to interact	95	168
Not known to interact	911	1454

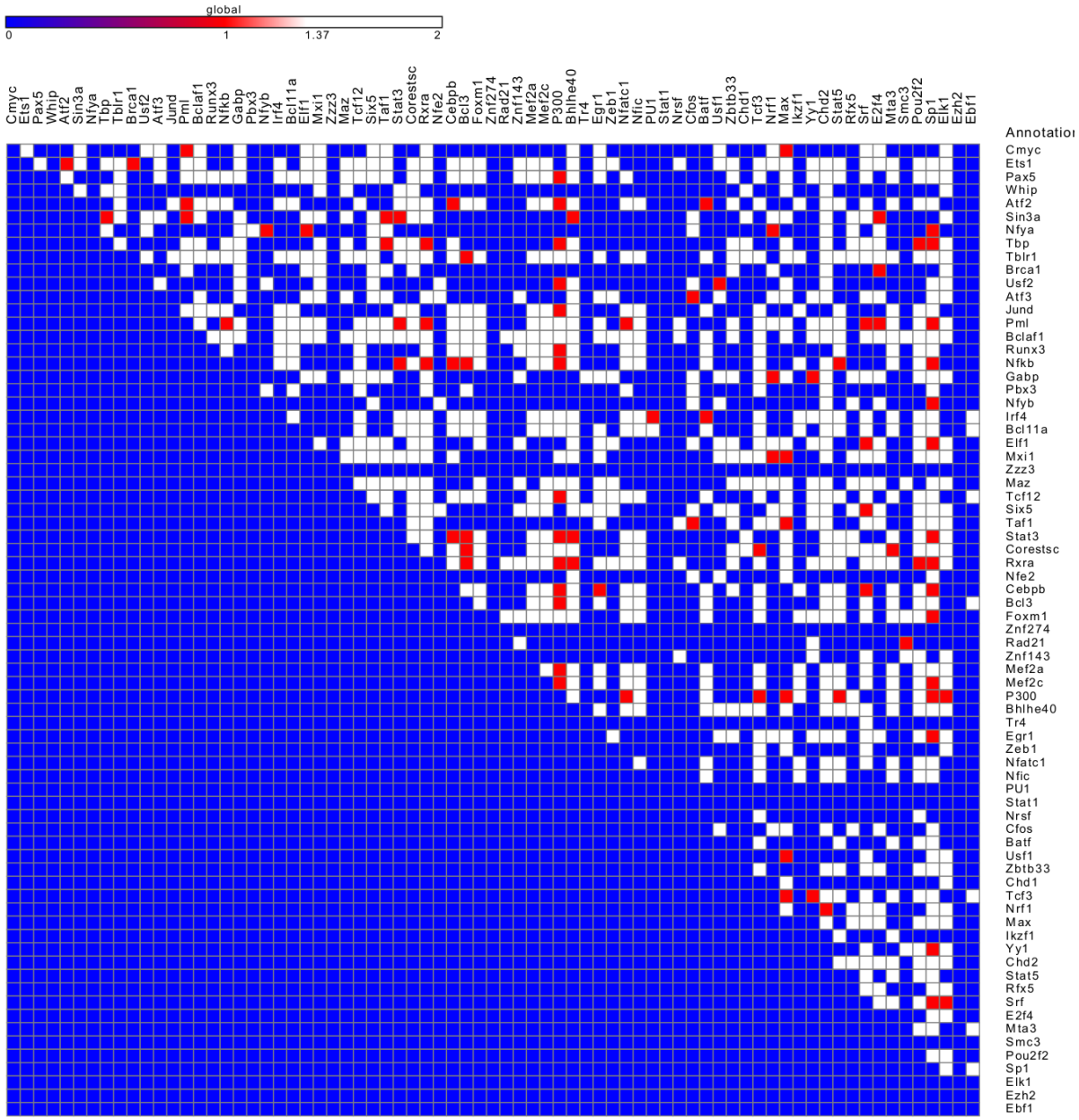


Figure 2. 7: This heat map (only upper triangle) shows known TF-TF interactions along with novel interactions predicted by our method. Red squares represent the predicted known interactions. White squares represent novel interactions predicted by our method. There are 1006 known and novel TF pairs shown in this figure. Blue squares (in upper triangle) represent the non-interacting transcription factor pairs.

2.3.2 K562 cell type

ENCODE has mapped 100 transcription factors in this cell type [34]. This is one of the cell types considered for identifying transcription factor mutual interactions. K562 cells are erythroleukemia type cells that were derived from a 53 year old female CML (chronic myelogenous leukemia) patient [77]. These cells have the ability to develop characteristics similar to early stage granulocytes, erythrocytes and monocytes [78].

2.3.2.1 Optimised results for K562 cell type

The same procedure used to evaluate Gm12878 data was adopted to identify the possible interactions between the transcription factors in the K562. There were a total of 4950 TF pairs from which very few TF pairs are known to interact. Therefore, computational methods need to be developed for prediction of transcription factor mutual interactions. Here we present the results after multiple testing correction (Benjamini and Hochberg) as all other procedures are similar to method used in the Gm12878 cell line. After global multiple testing correction (where we mixed all pairs overlapped at different peak sizes such as 20, 50, 74, 150, 350, 600, 800, 1000 base pairs), peak sizes with overlap that is highly significant than other peak sizes were chosen (significant interactions are those where real overlaps are higher than the average number of overlaps in random set). All the transcription factor pairs at their optimal peak sizes were ranked by negative log of their corresponding corrected p values. Only 1804 TF pairs have less than 0.01 corrected p value (Actual value is extremely low as the initial log was divided by 10) and they were considered for further analysis (We increased these values so that we can rank them. Otherwise large number of highly significant TF pairs have zero p value as python can’t handle extremely low values).The negative log of corrected p values, the ratio of real and expected overlap and the optimal peak sizes for these 1804 transcription factors are shown in Figure 2.8, 2.9 and 2.10, respectively.

In Figure 2.8, top 500 TF pairs with highly significant interactions are represented by red squares, and the TF pairs with similar pattern of overlap significance are clustered together by hierarchical clustering. Examples of predicted known TF

interactions are Cfos-Chd2, Cfos-Sp1 and Gata2-Tead4. There are transcription factors whose interaction are not significant in this set of 500 TF pairs, for example Egr1, Cebpb etc. Transcription factors such as Sirt6, Hdac2, Pml, Nrsf, Trim28, Egata2, Gata1, Gata2, Stat5, Nr2f2, Tead4, Tal1, P300, and Tblr1 form cluster in the Figure 2.8 (K562 cells). There is interesting observation in Gm12878 and K562 cells (Figure 2.4 & 2.8), that Smc3, Rad21, and Znf143 cluster together in right bottom of both figures. There are several TF pairs whose significant overlap is predicted by ENCODE method (Figure 2.1) and by our method (Figure 2.8) in K562 cells, examples of such pairs are MYC-ATF3, and HDAC2-ETS1. There are cases where a TF pair is overlapping significantly in our method (Figure 2.8), but not in ENCODE method e.g., SMC3-JUND and BCLAF1-TAF1. There are also TF pairs which are not overlapping significantly in both methods e.g., BCLAF1-RAD21.

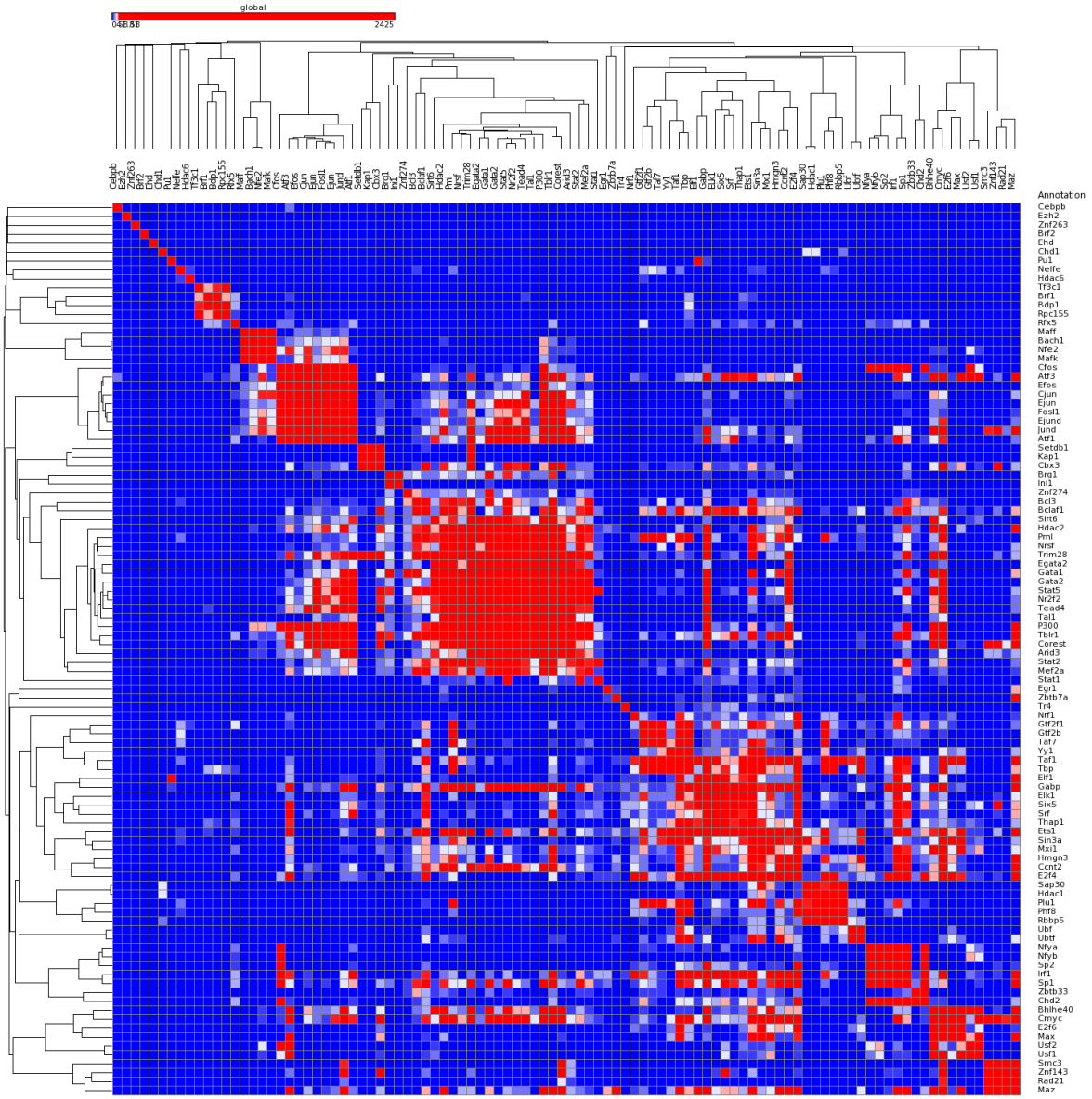


Figure 2. 8: This figure shows the negative logarithm of p values for K562 transcription factor pairs. Red squares show the highly significant transcription factor pairs, and remaining TF pairs from 1804 set (1304=1804-500) are represented with white, and light blue squares. Blue squares represent TF pairs with low level of significance and they are not part of 1804 set. TF pairs with similar pattern of overlap significance are clustered together by hierarchical clustering.

In Figure 2.9, top 500 TF pairs by highest ratio of real and expected overlap are represented with the red squares. Examples of such cases are Cfos-Chd2, Tal1-Tead4, Atf3-Ejun. Examples of TF pairs which are not included in the top 500 set are Cfos-Tead4, Elf1-Chd2, Zbtb33-Tead4.

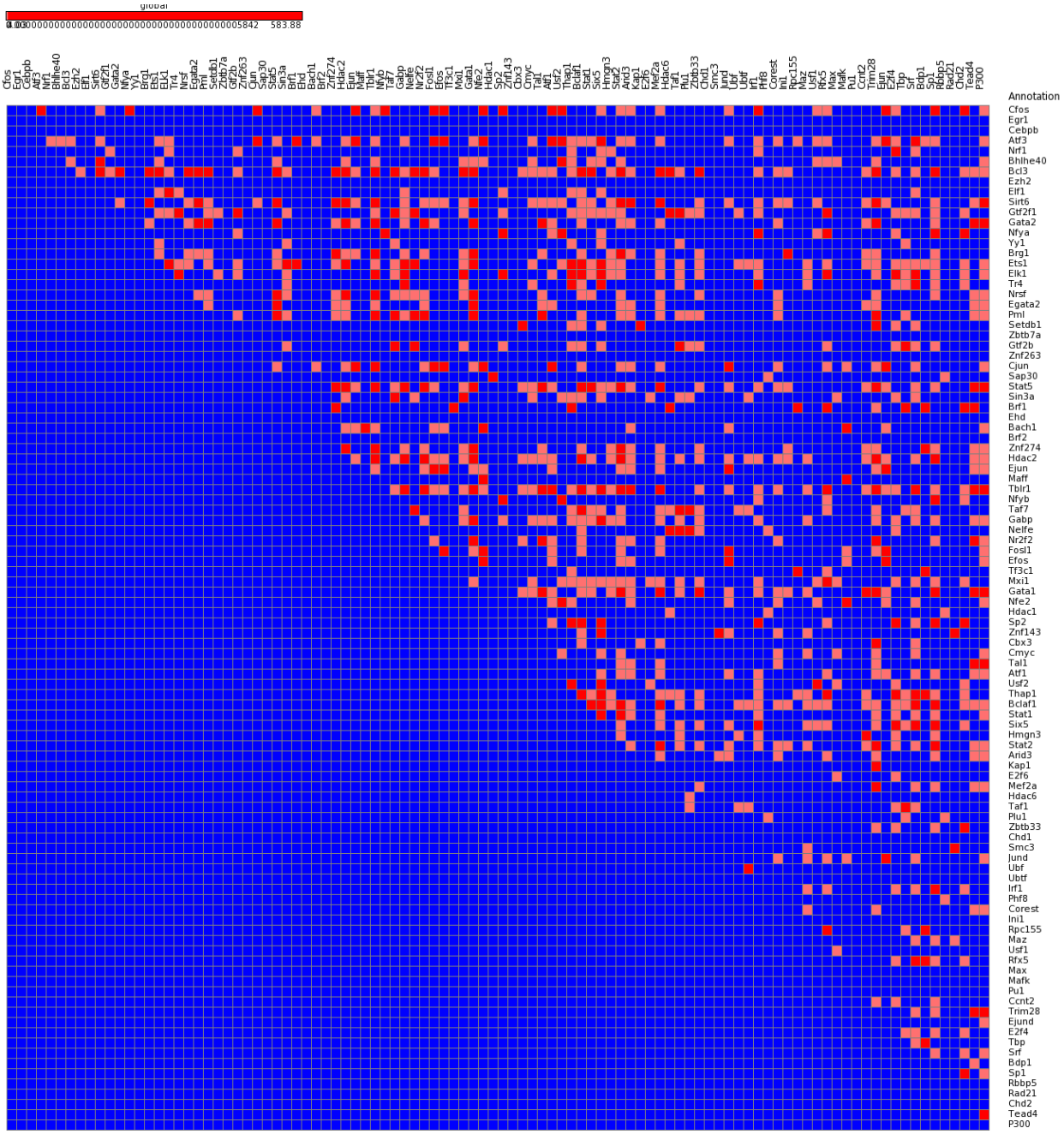


Figure 2. 9: This figure (only upper triangle) shows the ratio of real and expected overlap for 1804 TF pairs represented by all colours except blue. The red squares represent 500 TF pairs with highest ratio of real and expected overlap, and light red squares represent remaining1304 TF pairs (1804-500) from 1804 TF pairs set.

There are several TF pairs that are included in top 500 pairs in both sets (negative log of corrected p values and ratio of real and expected overlap). Examples of such cases are Cfos-Chd2, Gata2-Rad21 and Sp2-Ch2 TF pairs.

In below Figure (Figure 2.10) most of transcription factors have 74 base pairs peak size. There are also several transcription factors that interact mutually at peak sizes of 600, 350 and 150 base pairs.

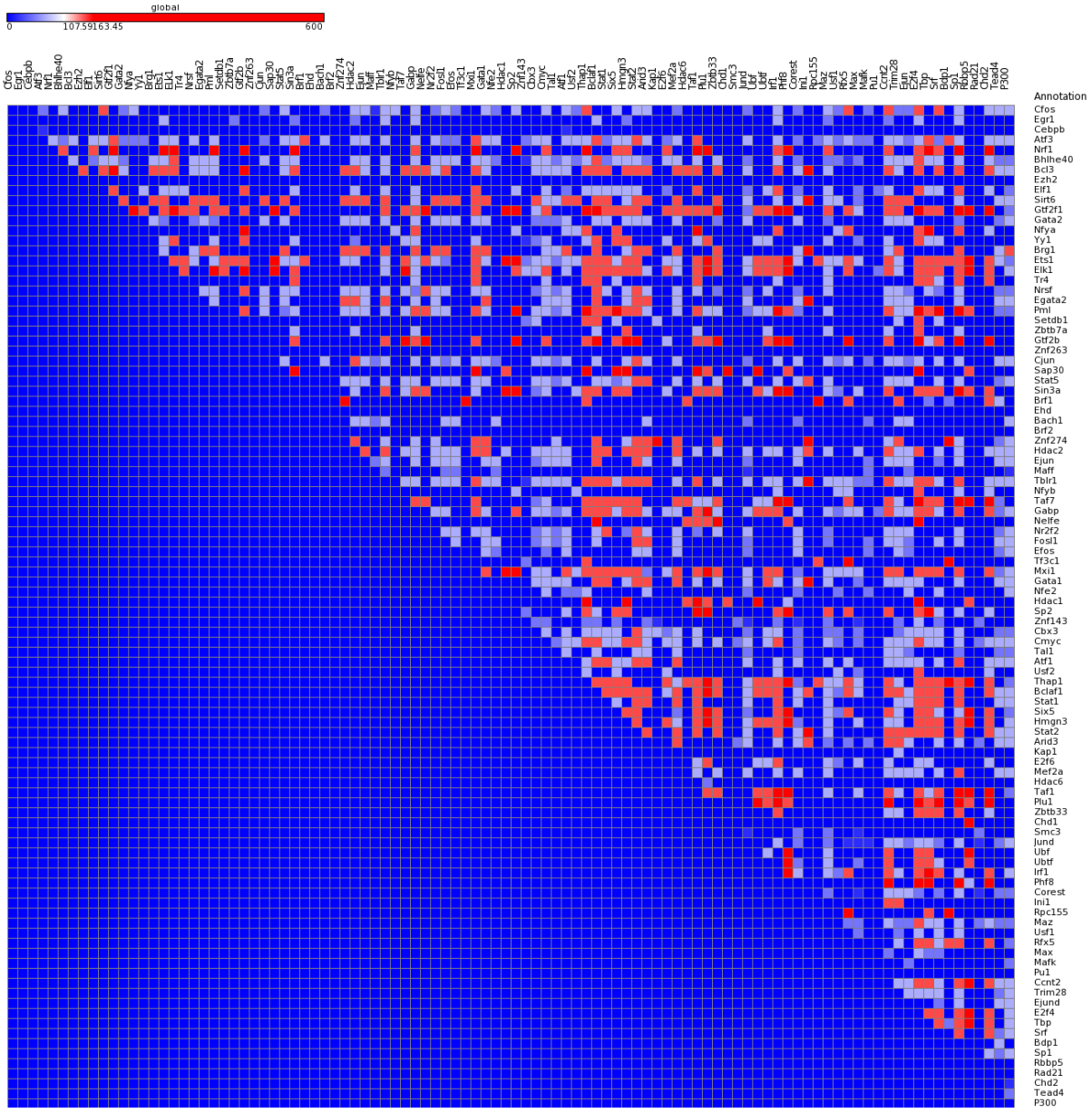


Figure 2. 10: This heat map (only upper triangle) shows peak sizes for 1804 transcription factor pairs with corrected p values lower than 0.01 (actually, this p value is extremely low in reality as initial log was divided by the 10). Red squares represent peak sizes of 600, 350, and 150 bases. Light red squares represent 74 peak size, white squares represent 50 and light blue squares represent 20 bases peak size. Blue squares represent TF pairs other than 1804 (3146=4950-1804).

Here, we also plotted the chosen peak sizes against the average length of the proteins but there is no correlation between them as shown in Figure 2.11. This

is because only small portion of protein binds on the DNA and there is no relationship between the size of protein and size of TF binding site.

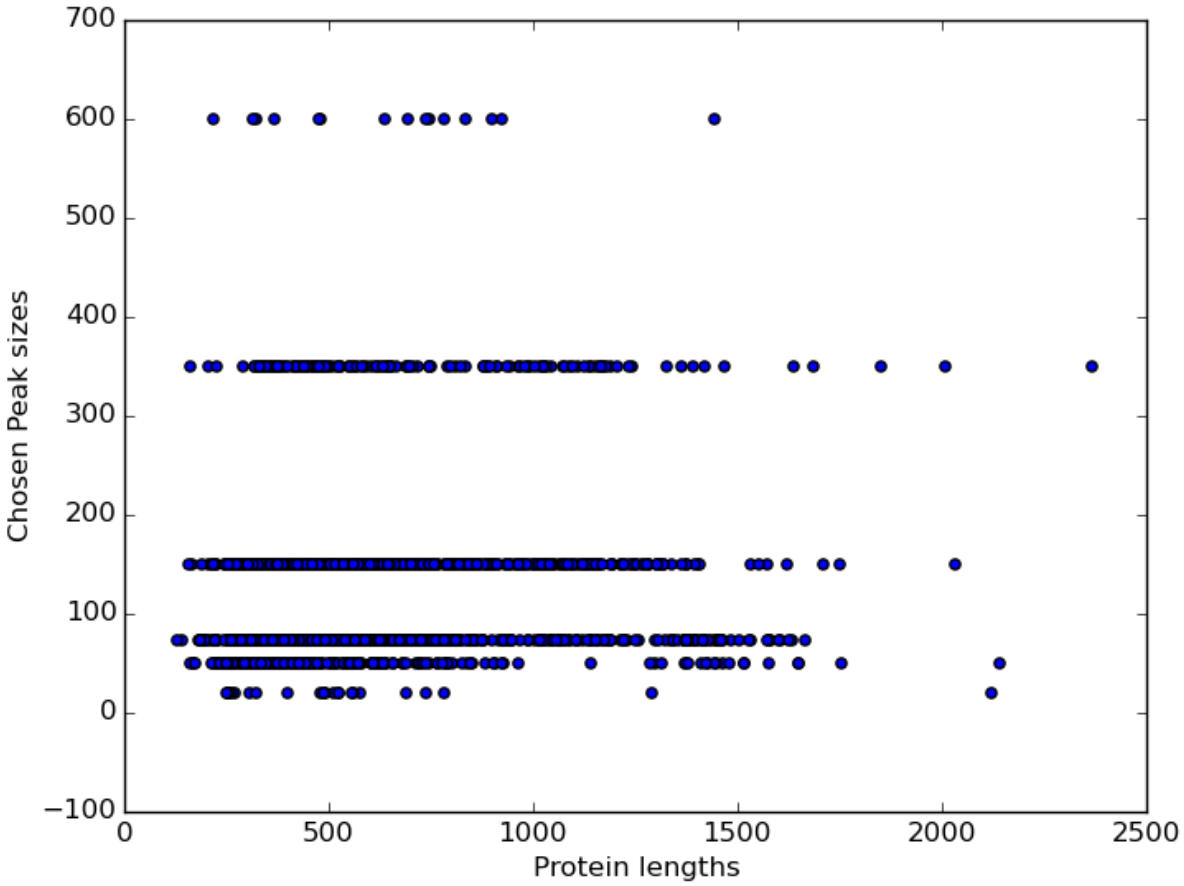


Figure 2. 11: This figure shows the plot where chosen peak sizes are plotted against the average protein lengths but there is no correlation between them.

2.3.2.2 Validation (Comparison of significant overlaps and known protein-protein interactions)

As mentioned in the method section, known protein-protein interactions were retrieved from Biogrid [70] and IntAct [71] databases. Orthologous interactions in mouse were also considered as evolutionary conserved transcription factors can interact with similar transcription factors in different organisms [76].

Known and novel transcription factor interactions predicted by our method are shown in Figure 2.12.

Several of our identified transcription factor interactions are known and they are represented with the red squares in Figure 2.12. Examples of such interactions are Cfos-Atf3, Egr1-Sp1, Nrf1-Chd2, Nfya-Sp1, Tal1-Sp1, Cmyc-Sp1 and Cfos-Jund TF pairs.

Most of our identified interactions are not known to interact and they are represented with white spots in Figure 2.12, these identified interactions can be considered as novel interactions. Examples of such cases are Atf3-Tead4, Tblr1-Chd2, Cmyc-Rad21, Mxi1-Chd2, Bcl3-Srf, Znf274-Tead4 transcription factor pairs.

Statistics of predicted and known interactions are shown in Table 2.3. Those interactions which are predicted by our method and also known to interact are limited, therefore, we got insignificant results when we applied hypergeometric distribution on known, not known, predicted and not predicted interactions. Result is not significant because several known interactions are might be cell type specific.

Table 2. 3: This table shows the summary of predicted, not predicted, known to interact and not known to interacting TF pairs. These known TF pairs are not enriched.

	Predicted	**Not predicted**
Known to interact	149	232
Not known to interact	1655	2914

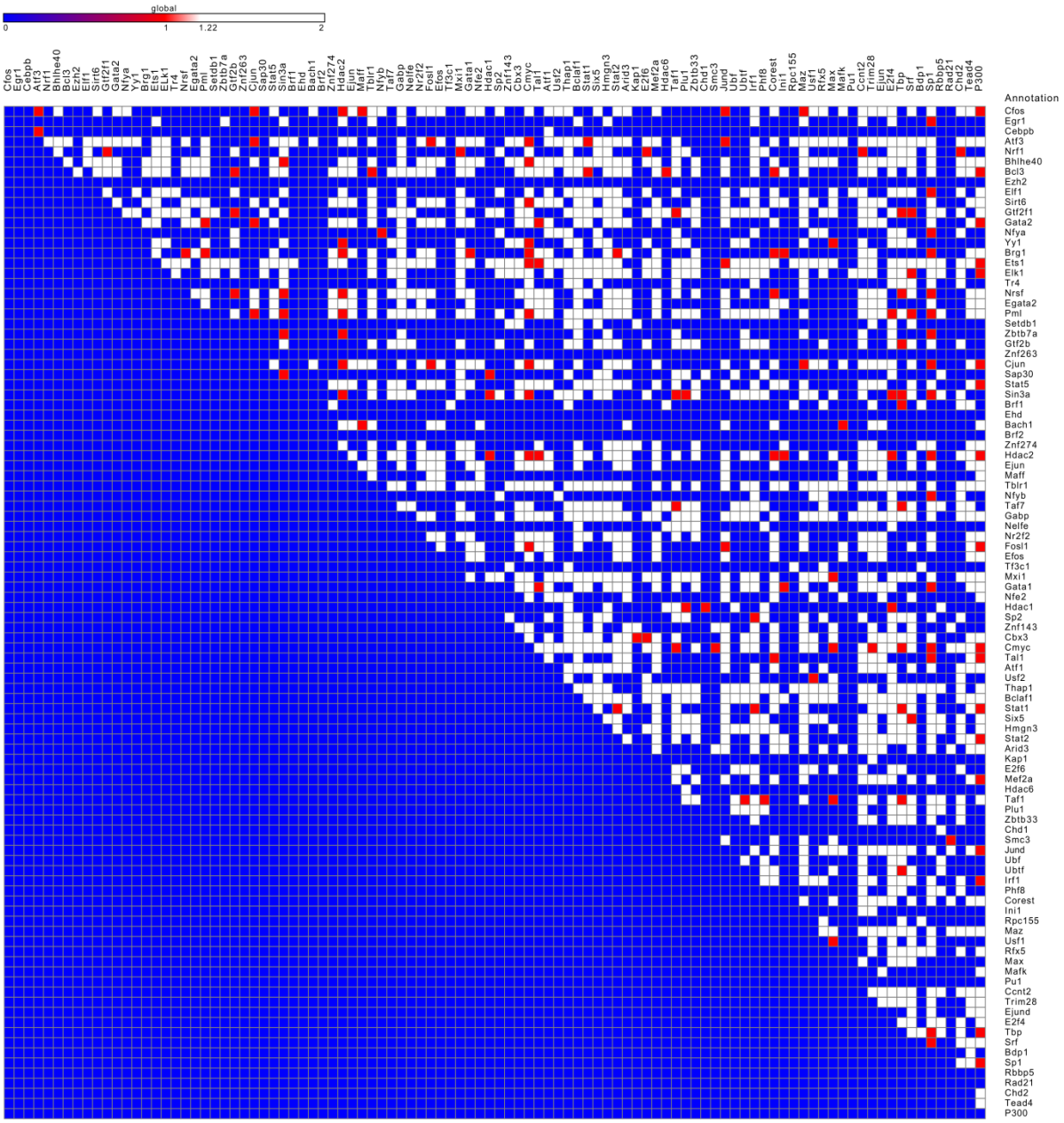


Figure 2. 12: This figure (only upper triangle) shows known TF-TF interactions along with the interactions identified by our method. Red squares represent the predicted known interactions. White squares represent novel interactions predicted by our method. Blue squares (in upper triangle) show the non-interacting transcription factor pairs (not predicted by our methods).

2.4 Discussion

Methods for prediction of transcription factor interactions were optimised using two statistical methods: randomisation and the Poisson distribution. Here we have presented results using data from two cell types (Gm12878 and K562), however our method can also be applied to data from other cell types. Binding sites for each transcription factor were also optimised, with binding sites (peak sizes) ranging from 20 to 350 base pairs. However, actual binding sites are small range from 5 to 31 nucleotide, but here transcription factor binds directly and indirectly on DNA i.e., transcription factor A sits on transcription factor B or vice versa and transcription factor protein size is bigger.

There are several transcription factors which are predicted by our method and they are also known to interact, examples of such pairs are Cmyc-Max, Cfos-Atf3 TF pairs. Table 2.2 and 2.3 show the statistics of predicted, not predicted interactions and known interacting TF pairs. In Gm12878 cell type, we have predicted 1006 interactions, 95 of them are known to interact. However, 1006 interactions number came because of setting a threshold, and that was difficult to adjust. It was difficult to set a threshold for size of the accessible genome. Therefore, we tried different percentages of accessible genome sizes, and then we chose size of accessible genome, where large number of predicted interactions are known.

There are approximately 168 known interacting TF pairs that were not predicted by our method, because they might interact in other cell types. Interactions of c-Myc (Cmyc) with 10 other transcription factors are known but our method didn't predicted them because c-Myc has a different pattern of binding in different cell types as this TF has different regulation mechanism in each cell type [79].

ENCODE has also identified co-associations between transcription factor pairs in K562 cell type. We compared few well associated TF pairs in Figure 2.1 (ENCODE method:K562 cells) with our results from Gm12878 cell type in Figure 2.4 and found many TF pairs overlapping significantly in both methods, even though cell types were different. However, all TF pairs in both methods and in

both cell types didn't behave similarly possibly because of their cell type specific function.

In K562 cell type (Table 2.3), our methods have predicted 1804 TFs interactions, and 149 of them are known to interact. However, there are approximately 232 known interacting TF pairs that were not predicted by our method possibly because their interactions are cell type specific. If we look at the same case of c-Myc as discussed for Gm12878 cell type, there are 21 interactions are known for c-Myc in K562 cell type, 18 of them are predicted by our method. We compared our method results in the Figure 2.8 with the ENCODE results in Figure 2.8, both results are from the same cell type (K562). We found several TF pairs overlapping significantly in both methods.

There are several transcription factor pairs, which were predicted to interact in one cell type but not in other cell type. Example of such cases is the Cfos-Jund pair which was predicted to interact in K562 cells and this is also a known interaction. However, Cfos-Jund pair interaction was not predicted in the Gm12878 cells. Examples of TF pairs, which were not predicted to interact in both cell types are Cebpb-Cmyc, Cebpb-Gabp.

There are existing methods for predicting association of TF pairs and one of them by ENCODE [34] is discussed above. Researchers have also predicted TF interactions from the primary sequence of transcription factor protein [80], though it is different from our method. There are existing methods for predicting protein-protein interactions which combine other biological features with the primary sequence. Asa et al., developed a kernel method, which combines protein sequences, gene ontology annotations, local properties of the network, and homologous interactions in other species for predicting protein-protein interactions [81].

Further, overlapping binding sites between TF pairs might have some functional role, and it is also possible that these overlapping sites are more conserved than the unique sites. We have done conservation analysis on overlapping binding sites and non-overlapping binding sites between TF pairs and cell types in the next chapter.

Chapter 3

3 Conservation analyses of transcription factor binding sites and effect of co-binding sites on gene expression

There are many binding sites for each TF, so many in fact that we suspect many of them do not have biological significance. Therefore, how might we determine which sites are biological interesting? Conservation possibly is one of indicators of functional TF binding site, as functional TF binding sites are less susceptible to change; hence, they might be evolutionary conserved [82]. Researchers have identified the binding sites for several TFs on the basis of conservation, and they suggest role of TFs in regulation of genes which are involved in prostate cancer [83]. In another study, it was observed that functional TF binding sites have higher conservation score than those TF binding sites whose function is not yet identified [84]. In addition, correlation of gene expression with the co binding of TFs can be the indication of biological importance.

This chapter has two parts, the first part contains the results of conservation analysis of transcription factor binding sites and the second part contains the analysis of effect of transcription factor binding sites and co-binding sites on gene expression.

3.1 Conservation analysis of transcription factor binding sites

Conservation data (phastCons) for 46 organisms were downloaded from the human genome browser at UCSC [85]. In phastCons files, conservation score was calculated using phylogenetic hidden Markov models [86]. The wiggle-formatted data were processed and conservation score for each base pair of transcription factor binding sites were retrieved. Shared binding sites for each TF in multiple cell types and overlapped binding sites for a TF pair may have some biological significance and conservation analysis of such sites with unique and non-overlapped sites would help us to understand their biological significance. Therefore, we aimed to answer the following hypothetical questions: 1) Are shared binding sites for a transcription factor in multiple cell types more

conserved than unique binding sites?; and 2) Are overlapping binding sites more conserved than non-overlapping binding sites in a TF pair within a particular cell type?

3.1.1 Are shared binding sites for a transcription factor in multiple cell types more conserved than cell type specific binding sites?

3.1.1.1 Methods

Shared binding sites among cell types for single transcription factors have biological importance and they might be conserved [87].

Common transcription factors in five cell types (K562, Gm12878, Hepg2, Helas3, and H1hesc) were identified but not enough transcription factors are common in these five cell types; therefore, we looked into a set of three cell types. There are 29 common transcription factors between K562, Gm12878 and Hepg2 cell types (i.e., Atf3). Binding sites for Atf3 in K562, Gm12878, and Hepg2 were divided into those shared (overlapping) between all cell types, and those unique (specific) to one cell type. This exercise was also repeated for the other 28 transcription factors.

As a result, we now have two data sets (i.e., shared and unique binding sites) for each transcription factor. The conservation median values (median of conservation score) of these two data sets were then compared, and plotted as box plots. The significance of difference between the two data sets were tested by the Kruskal-Wallis test [88]. This test assumes that these data sets came from population with the same distribution, so the null hypothesis is the median values for both data sets are same. If the p values from Kruskal-Wallis test turned out to be significant (p value <0.01) then the null hypothesis will be rejected. Data sets with higher median and significant p values would be considered more conserved than the other data sets.

3.1.1.2 Results

There are 29 common transcription factors from K562, Gm12878 and Hepg2 cell types were analysed and their shared and unique binding sites were identified. Statistics of shared, unique/cell type specific binding sites in three cell types and known motifs for 29 transcription factors are shown in the Table 3.1. These motifs were retrieved from the JASPAR database [89]. This database contains a curated non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes. The number of binding sites for a particular TF between cell types vary a lot (i.e., Mxi1-highlited in yellow in Table 3.1), which might indicates the functional difference of TF in different cell types. Other important observation is that, shared binding sites are smaller in number than cell type specific binding sites. This is evident even for those transcription factors with similar number of binding sites in all three cell types.

Two transcription factors (Ezh2 and Znf274-highlighted in dark yellow in Table 3.1) do not have any shared binding sites, therefore, they cannot be compared with unique binding sites. With the exception of Tr4, Ezh2 and Znf274 have small number of binding sites across all three cell types as compared to other transcription factors, hence, they have less chance of overlap. In addition, with no motif has been identified so far for these TFs, binding sites found in specific cell types are of less confidence binding sites. Ezh2 is the Enhancer of Zeste Homolog2/ Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit. This TF protein functions as an oncogene and it is involved in cell cycle, cell invasion and glioma stem cell maintenance [90]. Znf274 (Zinc Finger Protein 274) is a global transcriptional repressor [91]. All other transcription factors (27) have higher conservation score median for the shared binding sites and the significance of difference between shared and cell type specific binding sites was tested by Kruskal-Wallis test (see Table 3.2). Shared binding sites are significantly more conserved in all 27 transcription factors than cell type specific binding sites.

Table 3. 1: This table provides numbers of shared, unique/cell type specific binding sites in three cell types and motifs for 29 transcription factors are mentioned here

TFs	Number of TF binding sites					Motifs
	K562	Gm12878	Hepg2	Shared	Unique	
Tr4	587	1263	2953	61	526	
Srf	4717	8544	5314	416	4301	TGACCATATATGGTCA
Smc3	23598	30517	30797	4488	19110	
Atf3	16011	1677	3291	275	15736	
Nrsf	15849	6906	12828	1364	14485	
Jund	40052	2472	32275	35	40017	1.GGTGACTCATCC 2. tatGATGATGTCATC
Usf1	18521	9778	21890	1482	17039	gtCACGTGACC
Mxi1	6711	17735	20371	335	6376	
Taf1	15246	14278	16659	473	14773	
Yy1	24059	30994	17876	1655	22404	CAAgATGGCgGC
Elf1	27780	23008	18001	1735	26045	AACCCGGAAGTg
Nrf1	4211	5683	1902	483	3728	GCGCtTGCGCA
P300	25881	17461	27913	34	25847	
Rfx5	2201	4341	6017	193	2008	GTTgCCATGGcAAC
Cebpb	38715	5786	56629	60	38655	aTTGCGCAAT
Chd2	7797	15597	5169	540	7257	
Gabp	14393	6566	10109	1211	13182	
Sp1	7206	18248	25477	459	6747	gCCCCgCCCCc
Usf2	3083	9022	6291	519	2564	gtCAtGTGACc
Sin3a	12700	10392	16459	202	12498	
Cmyc	31092	3690	4413	85	31007	gAgCACGTGGT
Max	46171	12542	11854	535	45636	atCACGTGt
Tbp	17558	14893	13806	599	16959	gTATAAAAggtgg
Bhlhe40	22497	13986	14628	912	21585	atCACGTGAc
Zbtb33	3285	2144	2879	290	2995	TCTCGCGagactg
Rad21	34725	40019	54315	12977	21748	
Maz	33323	18952	12090	607	32716	
Ezh2	1685	2472	3286	0	1685	
Znf274	1997	233	245	0	1983	

Table 3.2 contains p values and median for shared and unique binding sites. There are transcription factors with extremely low p values for example Smc3, Nrsf, Yy1, Rad21, Maz and Usf1. This analysis shows that shared binding sites across different cell types are more conserved than unique binding sites for most transcription factors. Here Jund transcription factor have higher p value than others but it is still significant. This TF is the functional component of the AP1 transcription factor complex and encoded by the intron less gene and it has two known binding motifs as shown in Table 3.1. This transcription factor has been known to protect cells from p53-dependent senescence and apoptosis [92]. AP1 (Activator protein 1) regulates gene expression in response to several stimuli for example bacterial and viral infections and it is a heterodimer protein complex which composed of Fos, Jun, Atf and JDP protein families [93].

Table 3. 2: This table shows the p values calculated by the Kruskal-Wallis test for shared and unique binding (cell type specific) sites of common transcription factors in K562, Gm12878 and Hepg2 cell line. All shared sites are significantly evolutionary conserved than the unique binding sites. Shared and unique median are also mentioned here.

TFs	P values	Shared median	Unique median	TFs	P values	Shared median	Unique median
Tr4	1.725e-37	0.335	0.015	Cebpb	1.815e-77	0.123	0.004
Srf	6.505e-80	0.272	0.06	Chd2	0.0	0.544	0.009
Smc3	0.0	0.786	0.013	Gabp	4.249e-33	0.096	0.007
Atf3	2.11e-213	0.403	0.01	Sp1	0.0	0.267	0.008
Nrsf	0.0	0.618	0.004	Usf2	1.088e-75	0.466	0.004
Jund	7.095e-16	0.051	0.006	Sin3a	1.676e-203	0.815	0.008
Usf1	0.0	0.082	0.003	Cmyc	4.657e-120	0.806	0.004
Mxi1	3.661e-264	0.592	0.007	Max	0.0	0.558	0.003
Taf1	0.0	0.942	0.01	Tbp	0.0	0.436	0.006
Yy1	0.0	0.955	0.01	Bhlhe40	0.0	0.231	0.003
Elf1	0.0	0.436	0.005	Zbtb33	5.415e-213	0.336	0.013
Nrf1	0.0	0.314	0.003	Rad21	0.0	0.602	0.008
P300	8.459e-59	0.962	0.006	Maz	0.0	0.4525	0.007
Rfx5	5.490e-202	0.515	0.007				

As shown in the Table 3.2, conservation scores median for unique binding sites in three cell types are extremely small for all 27 TFs. On the contrary, conservation scores median for shared binding sites in three cell types are large except for few TFs (i.e., Jund, Usf1, and Gabp).

3.1.1.3 Discussion

Conserved regions of genome are important in understanding the shared functional role of transcription factors in multiple cell types. Therefore, we carried out an analysis on the shared and unique binding sites in three cell types for 29 common transcription factors. A total of 27 TFs shared binding sites are significantly more conserved than their respective unique binding sites which indicates that the functional regions (shared binding sites) are subject to purifying selection [94]. Ezh2 and Znf274 transcription factors have lesser binding sites in three cell types as compared to other transcription factors, therefore, they have less chance to share binding sites. DNA binding motifs are also not known for both these transcription factors, hence probably they would not bind on specific sequences.

Jund shared sites are less conserved than other TFs shared sites but still more significantly conserved than unique binding sites. This is may be because Jund has two different binding motifs. For all 27 TFs, medians of conservation scores for unique binding sites are extremely small which shows that unique sites evolved to specify the specific function of the cell types. Some TFs bind on the specific segment of DNA (motif) and these factors are called sequence specific TFs. Sequence specific TFs and their motifs are mentioned in the Table 3.1. Sequence specific TFs mostly bind directly on their DNA motifs and non-sequence specific TFs might bind anywhere on the DNA or indirectly on DNA through tethering [95].

3.1.2 Are shared sites more conserved than the non-shared sites in a TF pair within a particular cell type?

Conservation analysis for overlapping binding sites between transcription factors within a particular cell type may help us in understanding the biology behind the conservation, which will give indication about the transcription patterns in different organisms. Overlapping binding sites represent possible combinatorial regulation of genes and they are more likely to be functional.

3.1.2.1 Methods

We used the intersect function from bedtools [37] to find the binding sites that are occupied by TF pairs in the same cell type. Then, we divided the binding sites of TF pairs in the same cell type into those sites that overlap (are occupied by both TFs) and those that do not overlap (are occupied by only one TF). Conservation score for each base pair for overlapped and non-overlapped sites were calculated and subjected to the Kruskal-Wallis test [88] to determine the significance in median difference between the two data sets. This difference can be tested by Kruskal-Wallis test because our both samples are independent and have different sizes and this test is suitable for such samples. Null hypothesis is satisfied when both median values are the same. The p values from Kruskal-Wallis test were corrected by Benjamini & Hochberg test. If the corrected p value for a TF pair is significant then the null hypothesis will be rejected and both samples will be considered different although one of them is more dominant than other one. We applied this method to the K562, Gm12878, Hepg2, Helas3 and H1hesc cell types separately.

3.1.2.2 Results

Here, we have considered five cell types as discussed in the methods section. All the number of overlapping (shared) and non-overlapping (non-shared) binding sites for a pair of TF were separated and their conservation scores were compared and separated into following two sets for five cell types separately.

1. TF pairs where shared (overlapping) binding sites are more conserved than

the non-shared (non-overlapping) binding sites.

2. TF pairs where non-shared (non-overlapping) binding sites are more conserved than the shared (overlapping) binding sites.

In the K562 cell type, the ENCODE project mapped 100 transcription factors, so the total possible TF pairs are binding sites of 4950. A total of 3877 TF pairs have higher conservation score in shared binding sites than non-shared binding sites, while 1073 TF pairs have higher conservation score in non-shared binding sites than shared binding sites. Shared binding sites are significantly more conserved than non-shared sites in 2926 TF pairs (corrected p value<0.01), while non-shared binding sites are significantly more conserved than shared binding sites in 264 TF pairs. Distribution of corrected p values for shared and non-shared binding sites are shown in Figure 3.1 A and B respectively. Statistics of this analysis are mentioned in the Table 3.3. It is worth looking for distributions of conservation scores in shared binding sites and unique binding sites in TF pairs. Two examples of conservation distribution from K562 cell type: CEBPB-ATF3 and CFOS-EGR1 are shown in Figure 3.2 A and B respectively. These density plots show bimodal pattern of distribution with large number of zero conservation scores for unique binding sites.

In the Gm12878 cell type, ENCODE has mapped 73 transcription factors, the possible TF pairs are binding sites of 2628. A total of 2205 TF pairs have higher conservation score in shared binding sites than non-shared binding sites, while 423 TF pairs have higher conservation in non-shared binding sites than shared binding sites. Shared binding sites are significantly more conserved in 1834 TF pairs than non-shared sites, while non-shared binding sites are significantly more conserved than shared binding sites in 79 TF pairs. Distribution of corrected p values for shared and non-shared binding sites are shown in Figure 3.1 C and D respectively. Statistics of this analysis are mentioned in the Table 3.3. It is important to understand that how conservation is distributed in shared and unique binding sites. Here, we have given two examples of conservation distribution from Gm12878 cell type: cMYC-ELF1 and ETS1-PAX5 are show in Figure 3.2 C and D. They also show bimodal pattern of distribution.

In the Hepg2 cell type, 57 transcription factors have been mapped by ENCODE, possible number of TF pairs are binding sites of 1596. A total of 1294 TF pairs have higher conservation score in shared binding sites than non-shared binding sites, while 302 TF pairs have higher conservation in non-shared binding sites than shared binding sites. Shared binding sites are significantly more conserved in 1070 TF pairs than non-shared sites, while non-shared binding sites are significantly more conserved than shared binding sites in 40 TF pairs. Distribution of corrected p values for shared and non-shared binding sites are shown in Figure 3.1 E and F respectively. Statistics of this analysis are mentioned in the Table 3.3. Conservation distribution examples from this cell type are MAFK-BHLHE40 and NRSF-MAX that are shown in Figure 3.2 E and F. These two sub figures also show similar pattern to four other sub figures. It can be seen from Figure 3.2 that shared binding sites are highly conserved than unique binding sites.

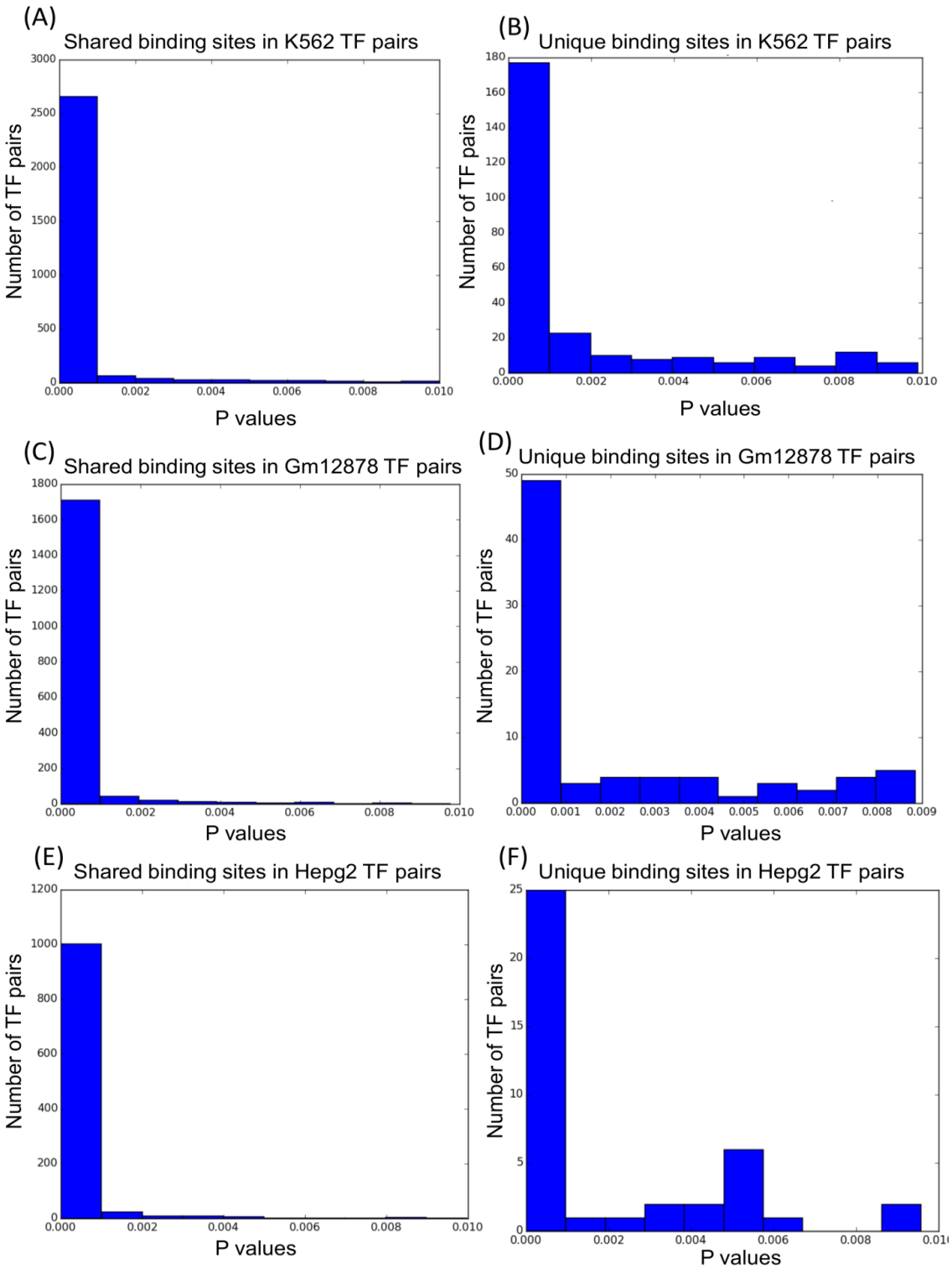


Figure 3. 1: (A) shows p values (p value <0.01) distribution for TF pairs where shared binding sites are more significantly conserved than non-shared binding sites in K562 cell type. (B) Shows p values distribution for TF pairs where non-shared binding sites are more significantly conserved than shared binding sites in K562 cells. Similar comparison is given in (C and D) for Gm12878 cell type and in (E and F) for Hepg2 cell type. In most of the TF pairs shared binding sites are more significantly conserved than unique binding sites.

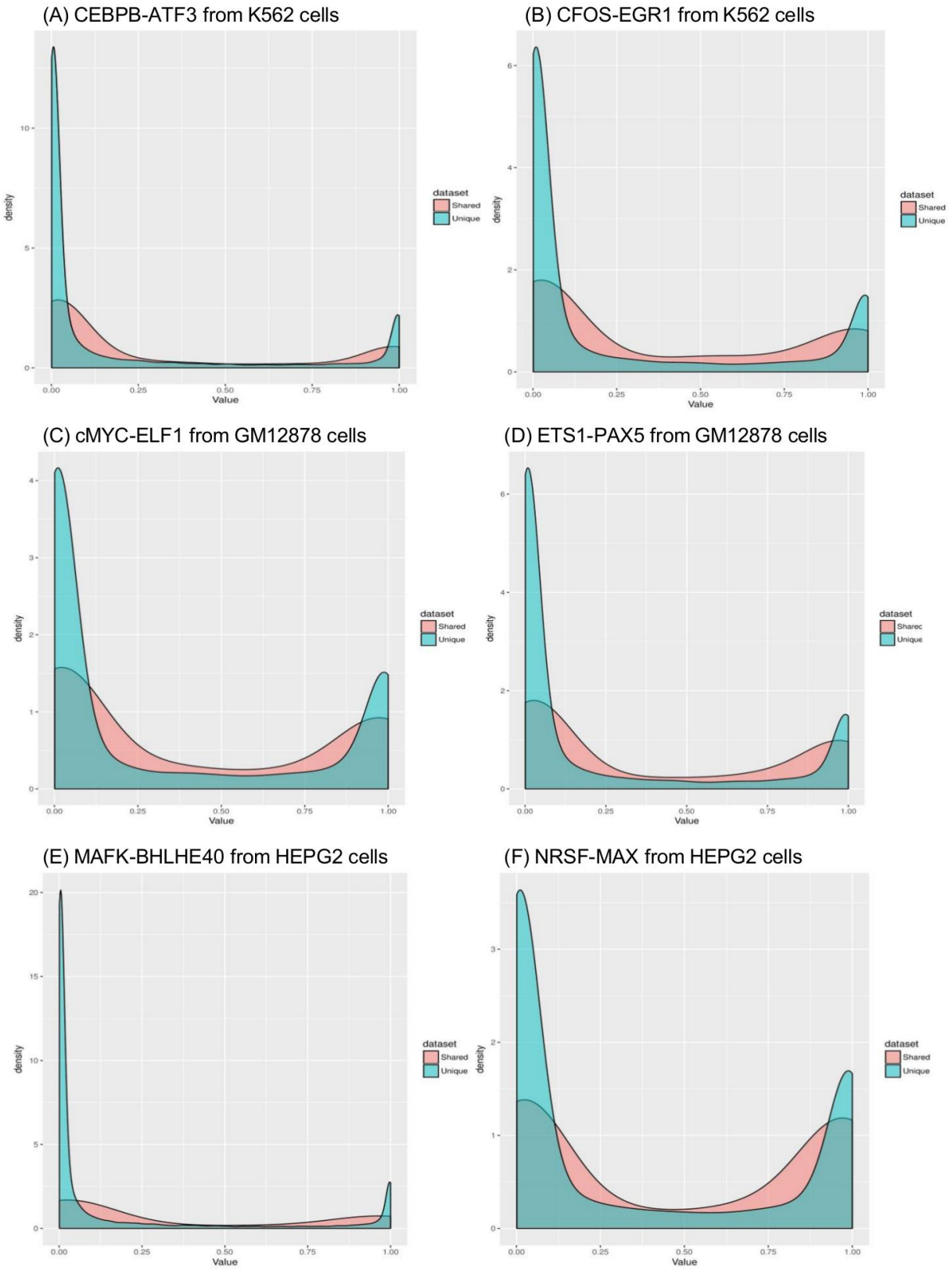


Figure 3. 2: This figure shows the density plots for conservation distribution in shared and unique binding sites between TF pairs. Here six examples are given from K562, Gm12878 and Hepg2 cell types, each example is represented by the single figure panel. Shared and unique distributions are shown in each density plot, suggesting bimodal pattern of distribution.

In the Helas3 cell type, ENCODE has mapped 54 transcription factors, so possible TF pairs are binding sites of 1431. A total of 960 TF pairs have higher conservation score in shared binding sites than non-shared binding sites but shared binding sites are significantly more conserved in 692 TF pairs than non-shared sites, while 471 TF pairs have higher conservation score in non-shared sites than shared sites but non-shared sites are significantly more conserved in 81 TF pairs than shared sites.
Distribution of corrected p values for shared and non-shared binding sites are shown in Figure 3.3 A and B respectively. Statistics of this analysis are detailed in Table 3.3.

In the H1hesc cell type, ENCODE has mapped 47 transcription factors, so possible TF pairs are binding sites of 1081. A total of 896 TF pairs have higher conservation score in shared binding sites than non-shared binding sites but shared sites are significantly more conserved in 698 TF pairs than non-shared binding sites, while 185 TF pairs have higher conservation score in non-shared sites than shared sites but non-shared sites are significantly more conserved in 25 TF pairs than shared sites. Distribution of p values for shared and non-shared binding sites are shown in Figure 3.3 C & D respectively. Statistics of this analysis are detailed in Table 3.3.

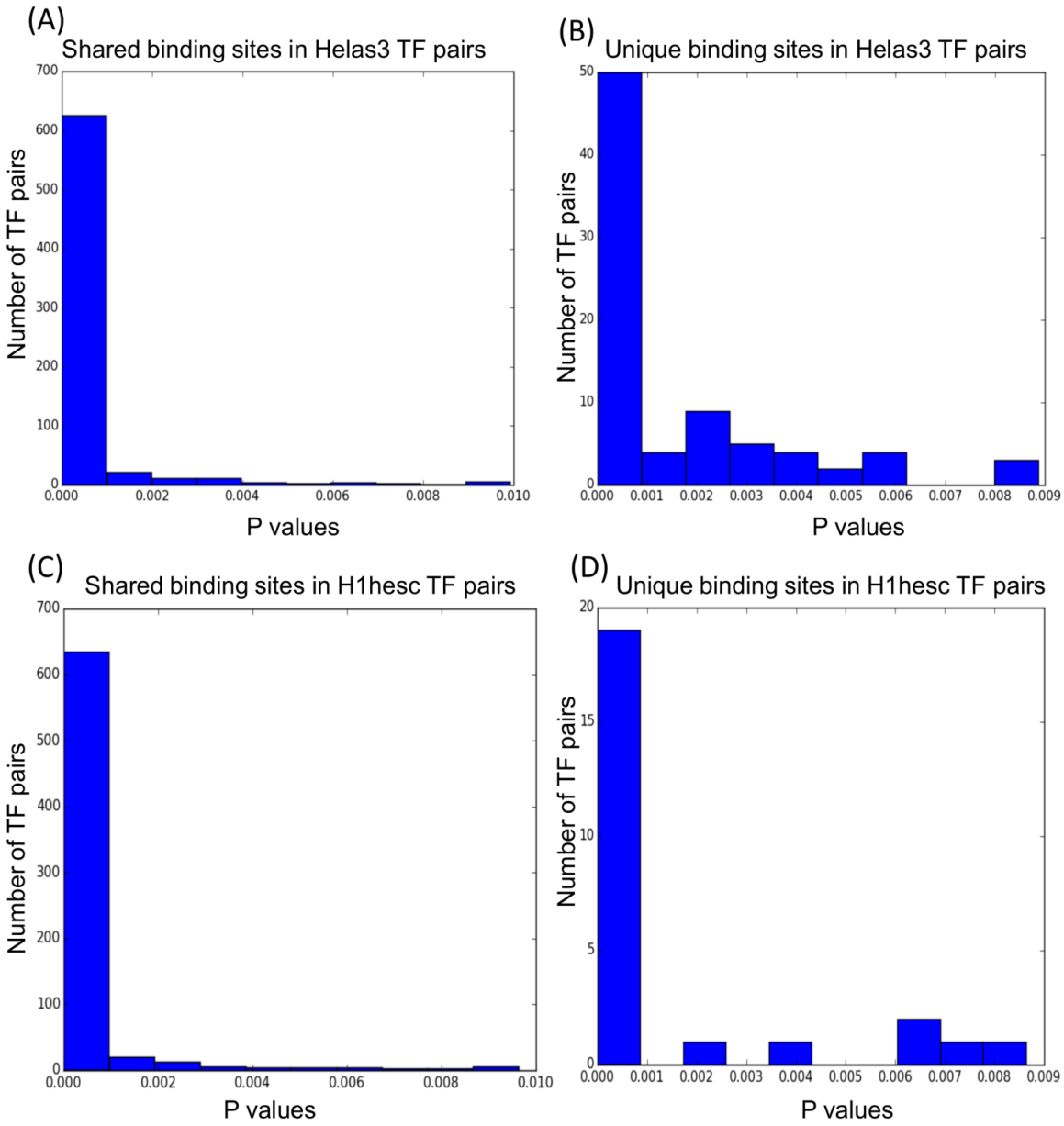


Figure 3. 3: (A) shows p values (p value <0.01) distribution for TF pairs where shared binding sites are significantly more conserved than non-shared binding sites in Helas3 cell type. (B) Shows p values distribution for TF pairs where non-shared binding sites are significantly more conserved than shared binding sites in Helas3 cells. Similar comparison is given in (C and D) for Gm12878 H1hesc. In most of the TF pairs shared binding sites are more significantly conserved than unique binding sites.

Table 3. 3: This table shows statistics about the significantly conserved shared and non-shared TF binding sites. In most of TF pairs, shared binding sites are more significantly conserved than the non-shared sites.

Cell lines	Total number of TF pairs	Number of TF pairs, where shared binding sites are more conserved than unique binding sites	Number of TF pairs, where shared binding sites are significantly more conserved than unique binding sites (corrected p values <0.01)	Number of TF pairs, where unique binding sites are more conserved than shared binding sites	Number of TF pairs, where unique binding sites are significantly more conserved than shared binding sites (corrected p values <0.01)
K562	4950	3877	2926	1073	264
Gm12878	2628	2205	1834	423	79
Hepg2	1596	1294	1070	302	40
Helas3	1431	960	692	471	81
H1hesc	1081	896	698	185	25

3.1.2.3 Discussion

Two questions have been answered in this section of the chapter. Firstly, are the binding sites for a particular transcription factor shared in multiple cell types are significantly more conserved than the unique binding sites? We have found that shared binding sites in K562, Gm12878 and Hepg2 are more conserved than unique binding sites in 27 transcription factors out of 29 transcription factors. Only Ezh2 and Znf274 transcription factors don't have any shared binding in these cell types. Conserved shared binding sites in multiple cell types indicate that they may share biological function in these cell types and may also be in different organisms [96]. Some transcription factors have ubiquitous roles so this might be a reason that their shared binding sites are conserved. Transcription factors form Transcriptional Regulatory Module (TRMs) by co-binding on DNA directly and indirectly through tethering. Transcription factor binding sites could be conserved among different organisms and there is an evidence for combinatorial regulation [97]. This evidence supports the idea that functional elements remain conserved across the cell types. Conservation of shared binding sites between cell types indicate about the direct and indirect binding of TFs on DNA. Several TFs are sequence specific factors as they bind specifically on specific sequence motifs and they may bind directly on the DNA and these sites were least evolved and they remain same in multiple cell types. However, TFs lack DNA motifs may be binding indirectly on an open chromatin region through tethering or bind on non-specifically on the DNA.

Secondly, are shared (overlapped binding) sites for a particular TF pair more conserved than the non-shared (non-overlapped) binding sites in particular cell type? This question was also answered with our finding that shared (overlapped binding) sites for a particular TF pair are more conserved than the non-shared (non-overlapped) binding sites in a particular cell type, and this pattern was observed in K562, Gm12878, Hepg23, Helas3 and H1hesc cells. Detailed result is mentioned in the Table 3.3. These shared sites are involved in the mutual interactions of transcription factors and their interactions have important role in the transcription of genes. There are also several TF pairs whose non-shared binding sites are significantly more conserved than shared binding sites (i.e.,

Cfos-Atf3, Cfos-Ehd, and Bcl3-Ets1 in K562 cells). Though shared as well as conserved sites have biological importance as they were least evolved because of their functional importance.

Shared binding sites between cell types for a particular TF and shared binding sites between TF pairs have functional importance and it is likely that these shared regions are GC regions possibly located in the CpG islands. These islands can be located in the promoter regions [98]. CpG islands are less susceptible to the mutations than other nucleotides [99]. Therefore, regions in the regulatory elements have higher chance to be conserved. Further, we have discussed CpG islands and DNA methylation in Chapter 1.

3.2 Mapping of transcription factors co-binding sites and single binding sites and their correlation with the gene expression

We asked the question whether the genes mapped near the genomic regions bound by a transcription factor pair have higher expression level than the genes mapped near the regions bound by single transcription factor.

Transcription factors binding events have an impact on regulation of genes. Large number of transcription factors would have chance to bind on open chromatin if it is large. Transcription factors interact with each other and bind directly or indirectly on DNA to form Transcription Regulatory Modules (TRMs). Binding of TFs on the cis regulatory regions leads to regulation of genes. The number of TFs binding events occurred at the particular region indicates the influence of that region on the regulation of nearby genes and that region is considered as a promoter or proximal enhancer [100]. There are also cis regulatory regions where large number of TF binding events are occurring and specific TFs bind there, and such regions are known as super enhancers as they enhance gene expression to a higher level [101]. Here, we have analysed how regions bound by two TFs and regions bound by single TF influence the expression of genes located within the distance of 2kb, 4kb, 6kb, 8kb, 10kb and 20kb.

3.2.1 Methods

Genomic regions bound by TF pair (overlapping sites) and regions bound by single TF (non-overlapping sites) were separated by using bedtools intersect [37]. Transcription factors were overlapped in such a way that TF1 with optimised peak size (Peak sizes were optimised by randomisation and the Poisson distribution in chapter 2) was overlapped with TF2 with peak size of 1bp (centre of the peak). If centre of TF2 binding site lies anywhere in the TF1 binding site then that site was considered as a co-bound site, otherwise sites from both TFs were considered as a single bound sites.

Both set of sites were mapped to the genes within 2kb, 4kb, 6kb, 8kb, 10kb and 20kb distances. Now we have to compare the expression levels of transcripts which were mapped to overlapping (bound by TF pair) binding sites with the

expression levels of transcripts which were mapped to the non-overlapping (bound by single TF) sites. It was assumed that the expression levels in both sets are the same and this assumption was tested by the Kruskal-Wallis test. P values from this test were converted into logarithm. If the expression median of non-overlapped sites mapped genes was higher than the expression median of overlapped sites mapped genes then the logarithm was multiplied with the "–" (minus sign) to convert that logarithm into positive value just to differentiate from the logarithm of overlapped sites mapped genes.

Now these values can be differentiated by the signs. Negative values are for the significantly higher expression of genes mapped to the overlapping sites than expression levels of genes mapped to the non-overlapped sites, while positive values are for the significantly higher expression of genes mapped to the non-overlapped sites than expression levels of genes mapped to the overlapped sites.

3.2.2 Results

This method was applied to the transcription factors from the K562 and Gm12878 cell lines. Although, this method can be applied to transcription factors from any cell line.

3.2.2.1 K562 cell type

We set a hypothesis that both data sets (genes from co-bound and single bound sites) are similar, but this hypothesis was rejected. It was found that genes mapped to the overlapping sites (co-bound) have higher expression levels than the genes mapped to the non-overlapping (unique) sites. It shows that combinatorial binding can enhance the expression level but some time these pairs act as a repressor when genes mapped to the sites bound by single transcription factor have higher expression than the genes mapped by sites bound by TF pair. Figure 3.4 shows the logarithm of p values calculated by Kruskal-Wallis test in 4950 TF pairs, and these sites were mapped to the genes within 2kb distance.

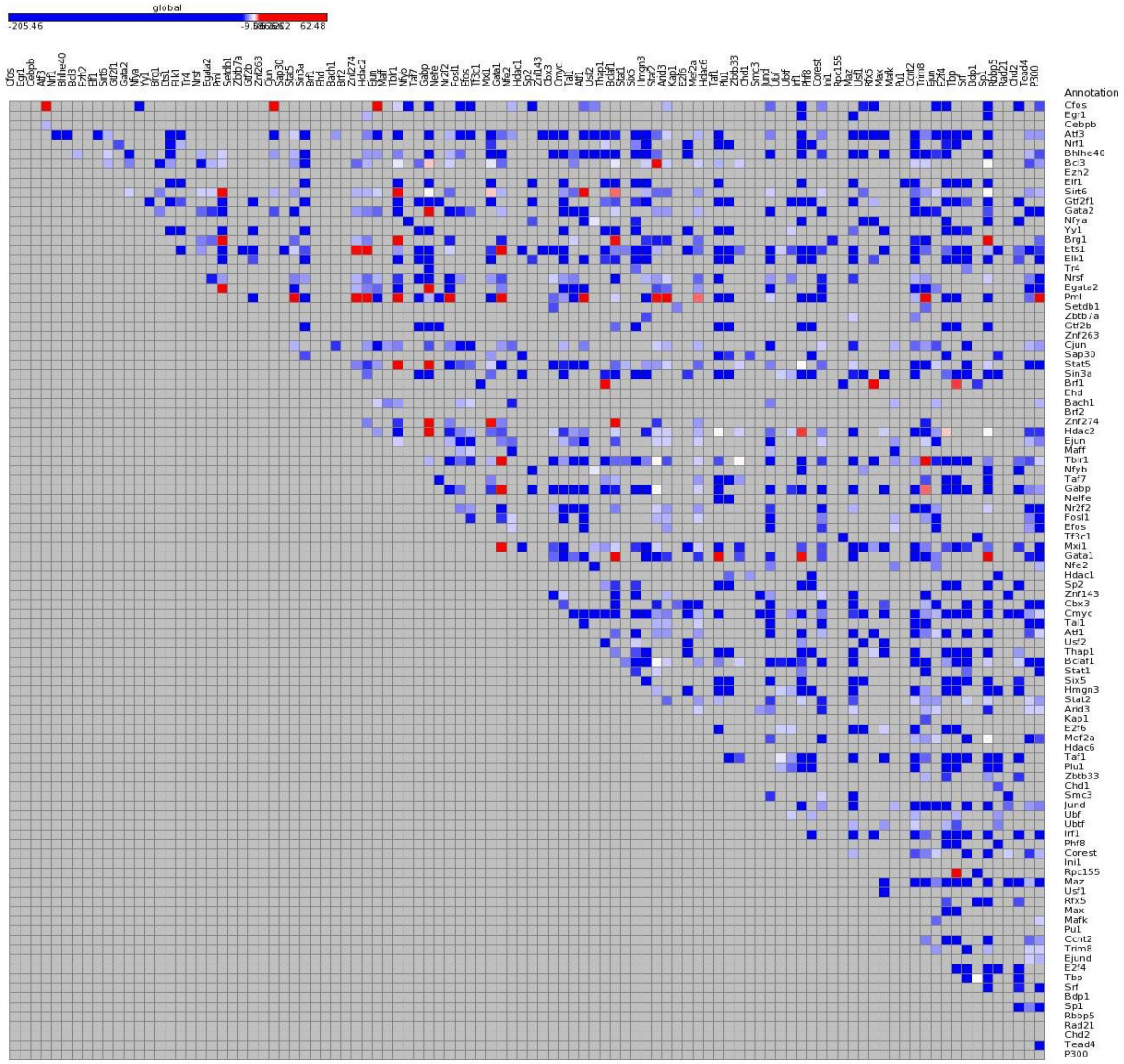


Figure 3. 4: In upper triangle, blue spots show the significantly higher expression level of genes mapped to the co-bound (overlapping) sites, while light blue squares represent same category but with less significance. The red squares show significantly higher expression level of genes mapped to the single bound sites (non-overlapping). Blue squares are large in number, which shows that co-bound sites enhance gene expression. These TF binding sites were mapped within the 2kb. While grey colour represents those pairs where there was no significant difference in expression levels of genes mapped to the co-bound sites and single bound sites.

There are some transcription factors which act as repressors in some cases when they co-bind with other transcription factors such as Pml (Promyelocytic Leukemia). TF pairs were considered as repressors of transcription when genes mapped to the sites bound by TF pair have lower expression level than the genes

mapped to the sites bound by single transcription factor. Gene codes for Pml transcription factor involves in a wide range of cellular processes such as apoptosis, transcriptional regulation, DNA damage response and tumor suppression. This transcription factor co-bind with Trim8, Arid3, Stat2, Atf1, Gata1, Nr2f2, Tblr1, Hdac2, Znf274, Stat5 and with P300 (P300 is not a transcription factor but regulates transcription through chromatin remodeling). These TF pairs act as repressors of transcription as shown in the Figure 3.4 (red spots in Pml row) [102]. We have found that when Pml co-bound with Sp1, Tbp, E2f4, Corest, Phf8, Irf1, Plu1, Taf1, Hmgn3, Tal1, Mxi1, Nelfe, Taf7, Ejun, Sin3a and Gtf2b then these TF pairs act as enhancers of transcription [103].
Other transcription factors such as Gata1, Ets1 and Brg1 also pairs with other transcription factors and act as enhancer and repressor of transcription. Majority of transcription factor pairs act as enhancer as shown in Figure 3.4 (blue spots).

As mentioned above these co-bound and single bound TF sites were mapped to the genes within the 2kb, 4kb, 6kb, 8kb, 10kb and 20kb distances.
Here, we have found that number of repressive pairs are dependent on the mapping distance. Consider the case of Pml, where10 other transcription factors forming a pair with Pml and act as a repressor when binding sites were mapped to the genes within 2kb. This number of TFs was decreased from 10 to 5 when binding sites were mapped with genes within 4kb distance.
When mapping distance was increased from 4kb to 6kb, number of TF pairs acting as repressors were decreased, again Pml pairs acting as repressors were decreased from 5 to 3 but Gata1 pairs remain same (three TF pairs were acting as repressor in 4kb). However, four Gata1 pairs were acting as repressor when sites were mapped within 2kb.
When mapping distance was increased from 6kb to 8kb, number of repressors were again decreased. Only one TF pairing with Pml (Pml-Stat2) was acting as repressor, while two TFs pairing with Gata1 (Gata1-Irf1 and Gata1-Taf1) acting as repressor.
When mapping distance was increased from 8kb to 10kb, number of repressors are again decreased to only few TF pairs such as Bcl3-Sp1, Brg1-Sp1, Brf1-Thap1 and Gata1 pairs remain same (Gata1-Irf1 and Gata1-Taf1) as they were in 8kb set. However, none of Pml pairs acted as repressor. Gata1-Irf1 and Gata1-

Taf1 along with the other TF pairs mentioned in 10kb set, they are constantly acting as repressors and may have some biological importance.

Finally when co-bound and single bound TF sites were mapped to the genes within 20kb, only Bcl3-Sp1 and Brf1-Thap1 pairs were acting as repressors, while most of the TF pairs were acting as enhancers represented in blue colour squares in Figure 3.5 heat map.

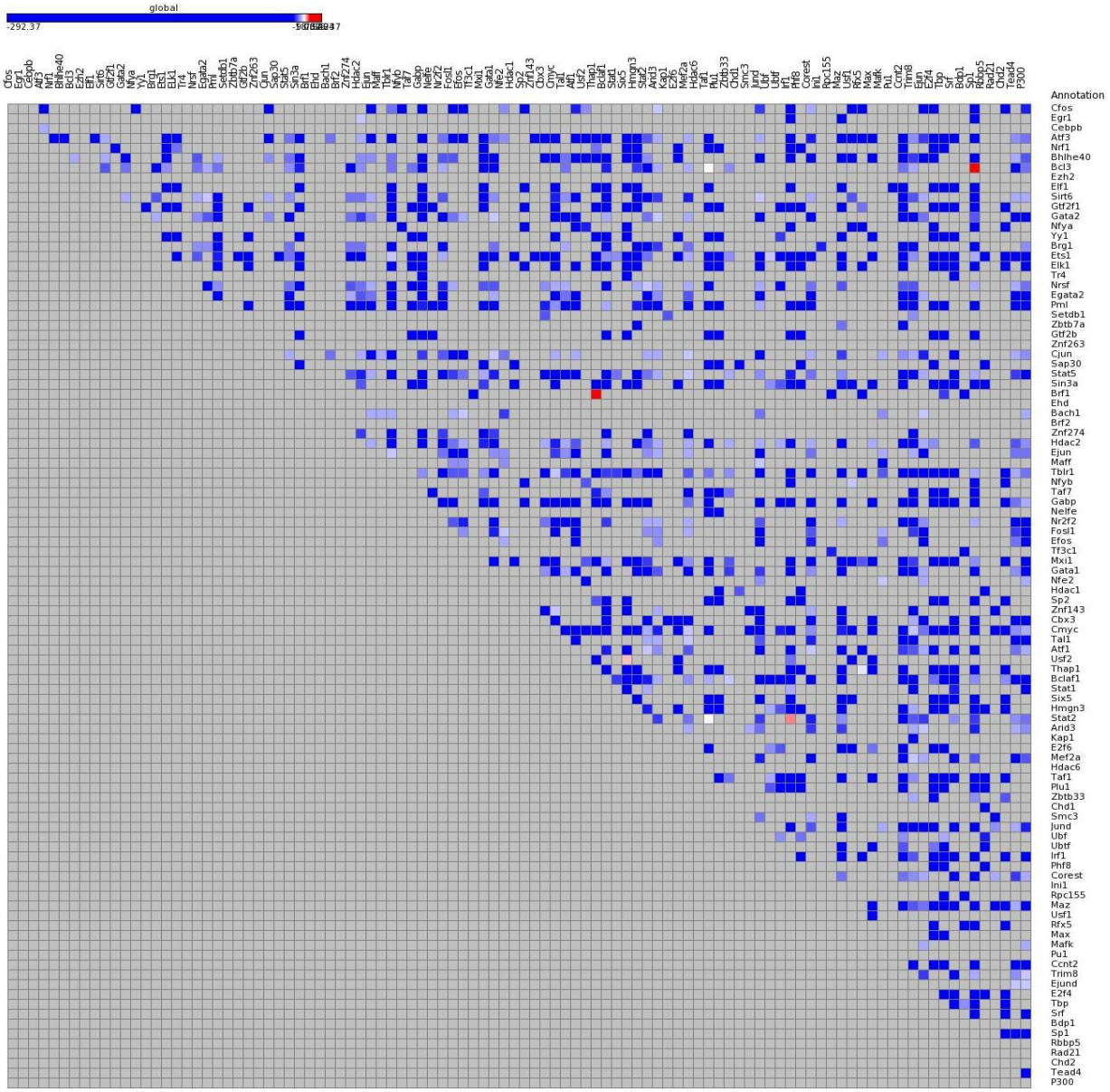


Figure 3. 5: (Only upper triangle). When the co-bound and single bound TF sites were mapped with the genes within 20kb distance, most of the TF pairs was found acting as enhancers spotted in blue colour, while light blue colour represents the enhancers but with less significance. Only two TF pairs was found acting as repressors. While grey colour means that for those pairs where there was no significant difference in expression levels of genes mapped to the co-bound sites and single bound sites.

3.2.2.2 Gm12878 cell line

ENCODE has mapped 73 transcription factors in this cell type, so the total

number of possible TF pairs are 2628. Here we followed the same procedures as we did for K562 cell type. All these sites were mapped to the genes within 2kb, 4kb, 6kb, 8kb, 10kb and 20kb distances. It was assumed that genes mapped with the genomic regions bound by TF pair have higher expression level than the genes mapped with the genomic region bound by single transcription factor. This assumption was tested by the Kruskal-Wallis test. Figure 3.6 shows the heat map representing gene expression significance levels, significantly higher expression of genes mapped to the co-bound sites are represented with blue squares and significantly higher expression of genes mapped with the single bound sites are represented with red squares, while other colours do not show significant dominance, these sites were mapped to the genes within 2kb.

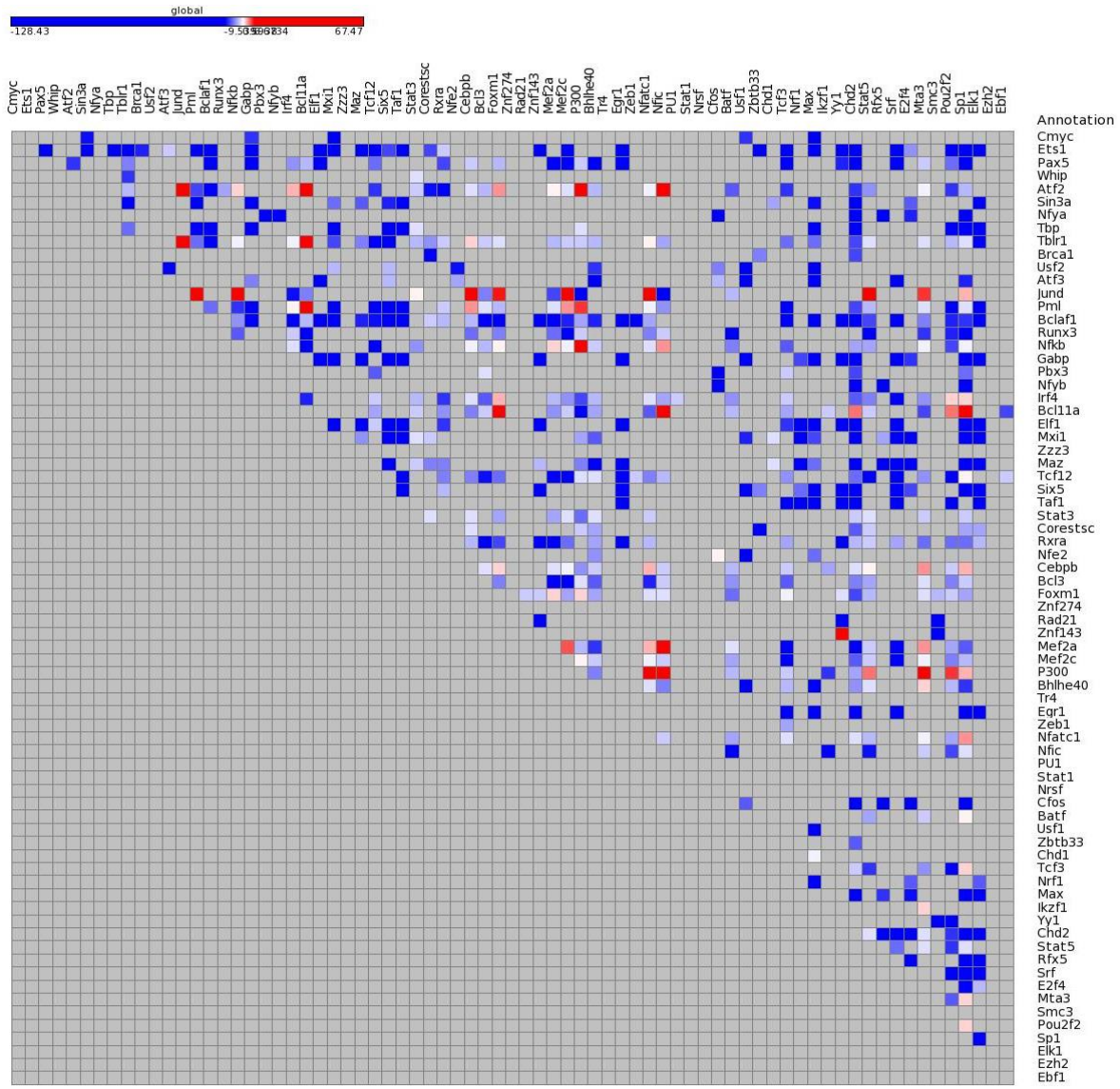


Figure 3. 6: (Only upper triangle). This heat map (here, sites were mapped to the genes within 2kb) shows the blue spots representing significantly higher expression levels of genes mapped to the genomic regions bound by TF pair, while light blue colour represent the same category but with less significance. Red squares show the significantly higher expression levels of genes mapped to the regions bound by the single transcription factor. While grey colour means that for those pairs where there was no significant difference in expression levels of genes mapped to the co-bound sites and single bound sites.

Genes mapped to the sites bound by TF pairs have higher expression level than the genes mapped to the sites bound by single transcription factor as shown in Figure 3.6 but there are some TF pairs where genes mapped to the single

transcription factor have higher expression level. Jund binding sites overlap with other transcription factors such as Mta3, Stat5, Nfatc1, Mef2c, Foxm1, Cebpb, Nfkb, Pml and these TF pairs act as repressors of transcription because genes mapped to the sites bound by single transcription factor have higher expression level than the genes mapped to the sites bound by TF pair [104]. Other transcription factors such as Atf2 and Bl11a binding sites pairs with other transcription factors and act as repressors of transcription as shown in Figure 3.6.

Number of repressors are dependent on the mapping distance as we have also observed similar pattern in the K562 cells. Now only one pair of Jund (Jund -Sp1) act as repressors. In this set, repressors of Pml and co-factors were increased from 2 to 4. Pml and co-factors also involved in the repression of gene expression as shown in Figure 3.4.

When the mapping distance was increased from 4kb to 6kb, overall repressors were decreased and enhancer were increased but we have observed that Pml overlap with 6 other transcription factors and act as repressor [105], although Jund overlap with only two transcription factors and act as repressor of the transcription.

When mapping distance was increased from the 6kb to 8kb, similar trend with previous increment was observed where overall repressors were decreased as usual by increasing the mapping distance. Now Pml overlapped with only 5 transcription factors to repress the transcription as genes. Sp1 also co-bound with other transcription factors and act as repressor in this set, as well as in 4kb and 6kb sets.

When the mapping distance was increased from the 8kb to 10kb, number of repressors was decreased again as expected. Now Pml overlapped with the 4 other transcription factors to repress the transcription and Sp1 overlapped with other 4 transcription factors to repress the transcription.

Finally when the mapping distance was increased from 10kb to 20kb, number of repressors were reduced to only five TF pairs and they are involved in repression of transcription. Pml overlapped with other 4 transcription factors to repress the transcription and Chd2 overlapped with Stat3 to repress the transcription (5=4+1). Heat map of these TF pairs shown in Figure 3.7.

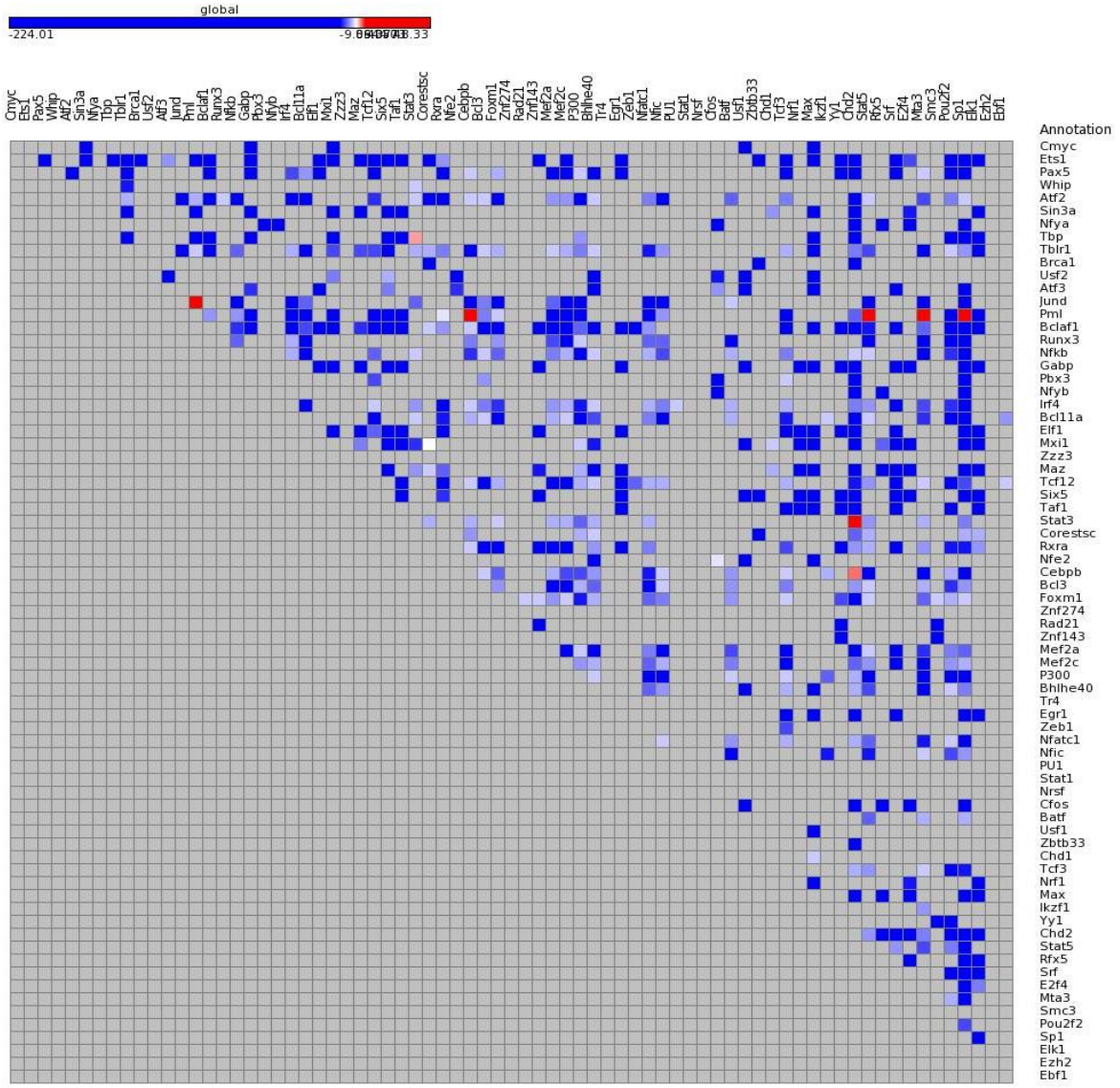


Figure 3. 7: Only upper triangle. (Binding sites were mapped to the genes within 20kb). Blue squares represent TF pair co-bound sites (bound by two transcription factors) mapped to the genes have significantly higher expression levels, light blue colour represent the same category but with less significance. Red squares show single bound (binding sites of single transcription factor) mapped to the genes have higher expression levels, while grey colour means that for those pairs there was no significant difference in expression levels of genes mapped to the co-bound sites and single bound sites.

3.2.3 Discussion

Transcription factor plays an important role in transcription of genes. They interact with each other and bind on the DNA to form a complex to regulate genes in combinatorial manner. We know that there are number of transcription factor binding events that occurred at certain region and has some influence on rate of gene expression. We expected that if TF pairs bind on the cis regulatory region then that region would have more influence on the gene expression level than the regions bound by single TF possibly because TFs complexes influence the expression level than the individual TF binding [106].

It was found that genes mapped to the sites bound by transcription factor pair have significantly higher expression than the genes mapped to the sites bound by single transcription factor, for majority of the cases; which supports the idea that quantity of transcription factors influence the rate of gene expression [107]. There are some transcription factors such as Pml, when they co-bound with other transcription factor such as Sp1 played a role of repressor because genes mapped to these overlapping sites have lower expression levels compared to the non-overlapping sites which have higher expression levels. This pattern was observed in both cell types (K562 and Gm12878) and other cell types might follow the similar pattern. Number of TF pairs acting as repressor is dependent on the mapping distance as number of repressors were decreased when mapping distance was increased from the 2kb to 20kb as mentioned in the results.

Chapter 4

4 Predicting cis-regulatory regions by using linear regression

4.1 Introduction

Genetic regulation is important for development and function of eukaryotic cells, and some principles of genetic regulation are already studied, however, the regulatory process of most of the genes is not known yet. On the genome scale our understanding of genetic regulation is poor, and we need to study the mechanisms of genetic regulation in multiple cell types.

Cis-regulatory regions are an important component of genetic regulation and these regions can be enhancers or repressors. Enhancers are cis-acting short regions (5-1500bp) of DNA that can be bound by transcription factors to activate the transcription of genes. They can be located up to 1Mbp (1,000,000 bp) away from transcription start site (TSS) of the gene and can be found upstream or downstream from the TSS. It is estimated that hundreds of thousands of enhancers are present in the human genome [108]. Some cis-regulatory regions repress the transcription of genes, and they are known as repressors of transcription. Enhancers and repressors are discussed in the 1st chapter.

There are several research studies which suggest that TF binding and conservation scores can be helpful in identifying potential CRRs (Candidate cis-Regulatory Regions). Mendelson [109] studied the expression of SP-A gene and it was observed that SP-A gene expression was increased as TTF-1 transcription factor binding was increased. Some studies [110] suggest that there is relationship between conservation of transcription factor binding events and the conservation of target gene expression. Previous studies [110] also suggested that genomic regions with extreme conservation can act as regulators of transcription but this is not always the case; regions with less conservation can also act as regulators. A few studies also suggested histone modifications and Pol II occupancy can identify the role of regulatory regions [111].

Several computational studies have been done to understand the gene regulation in different organisms, Wilczynski and co-workers built a probabilistic model that predicts certain aspects of gene expression by integrating TF occupancy and the

chromatin state [112].

Some methods for the identification of enhancers have already been published, and they have identified enhancers using different methods, which are different from our method.

Shen et al. [113] have developed a method to create CRMs (Cis Regulatory Modules) that uses the presence of H3k4me1 and absence of H3k4me3 to predict enhancers in mouse embryonic stem cells. They used p300 binding sites to train the model as proxies for the enhancers. Some other studies also showed that H3K4me1 and P300 binding site are the signatures of active enhancers [114].

Yip et al. [115] also created a cis-regulatory map linking their predicted CRMs with the genes whose expression they might control. To achieve that they 1st predicted CRMs in five cell types and merged all the CRMs that overlapped across the cell types. They computed correlation between signals of histone modification and transcript expression levels through Pearson correlation within 1Mbp (million base pairs).

Anderson et al. [116] used CAGE (cap analysis of gene expression) data to identify the CRMs and their target genes; CRMs were identified from short genomic regions with balanced, bi-directional and divergent transcription of short RNA molecules, as bi-directional capped RNA is signature of active enhancer. They have identified target genes for enhancers by correlating transcriptional activity at the CRMs and transcriptional activity at putative gene transcription start site (TSS) across a diverse set of human cells.

O'Connor et al. [117] identified the tissue specific CRMs; their model predicts the RNA expression level of gene from the TF binding events occurred at the CRMs associated with the gene.

There are also experimental methods for assessing enhancer activity, for example reporter genes. In this method, a reporter construct comprised of reporter gene and regulatory region can randomly be integrated into the genome; expression of gene would increase if it integrates near enhancer. Reporter gene usually encode for fluorescent protein such as gene encodes green fluorescent protein (GFP) and used as the marker for successful uptake of the gene of interest [118].

Computational and statistical methods can be developed to identify the

enhancers from sequencing data. Statistical methods such as linear regression and LASSO linear regression can be used to identify features relevant to enhancers [119].

In previous chapter we showed that sites bound by multiple TFs (co-bound sites) in the ENCODE data are associated with high expression genes. This suggests that the ENCODE data might provide a way to identify cis-control elements (cis-regulatory elements) for specific genes, by examining correlations between activity at those elements and the expression of nearby genes. We set out to investigate the hypothesis in this chapter. We might identify potential cis-regulatory regions from ENCODE data-genomic regions that are DHSs (DNase I hypersensitivity sites), that bind TFs, that bind multiple TFs, and that are conserved. We have discussed in the last chapter that how functional TF binding sites are evolutionary conserved. Therefore, we assume that evolutionary conserved regions would help us to predict the potential cis regulatory regions. There can be several challenges and problems in predicting cis regulatory regions. The fact that the enhancers/repressors may be located at long distance from their genes, the fact that there are many candidate regulatory regions and it is difficult to identify the important ones and link them to the genes, and the fact that regulation might be cell type dependent.

Possible Candidate cis Regulatory Regions (CRRs) can be obtained from the ChIP-seq and DNase-seq data. ENCODE has mapped approximately 119 transcription factors in multiple cell types as discussed in chapter 2. We choose 5 major ENCODE cell types as most of the 119 Transcription factors were mapped in these cells. The number of mapped TFs in each cell type is shown in Table 4.1. Some transcription factors binding sites were mapped in multiple cell types.

Table 4. 1: This table shows number of TFs mapped in five cell lines

S.No	Cell types	Number of TFs mapped
1	K562	100
2	Gm12878	73
3	Hepg2	57
4	Helas3	54
5	H1hesc	47

4.2 Methods

The following three points are the basic idea of method.

1. CRRs were identified from the ENCODE data (2012 release) on TF binding and DNase I hypersensitivity with some fixed distances of a gene TSS.
2. The idea is to use correlations between measures of activity at CRRs and the expression of nearby genes to identify relevant cis regulatory regions.
3. DHSs signal intensities at CRRs were adopted as an appropriate measure of activity.

Candidate cis regulatory regions (CRRs) were obtained from ChIP-seq data by taking the union of all the Transcription factor binding sites from 5 cell types (Table 4.1) followed by merging of overlapping sites. TF binding data for these regions were retrieved for each cell type separately (hg19 genome version was used here). In the 2nd step, DHSs signal intensities for each region were extracted from bigwig file by using bw tools [120] for 10 cell types (5 more cell lines were added). In last step, all these CRRs were allowed to map to all transcripts within a certain distance (distance from CRRs to TSS of transcript: 20Kb, 40kb and 100kb distances were considered).

If these CRRs mapped close to the particular transcript and DHSs signal intensities of these regions are highly correlated with the gene expression then these regions could be predicted to be cis regulatory regions.

4.2.1 Multiple Linear Regression

Linear regression is a statistical method to model the relationship between the dependent variable and several independent or explanatory variables. Linear regression can either be simple with one independent variable, or multiple with more than one independent variables. Here, we have used multiple linear regression, because, we have several independent variables for single dependent variable.

We based our model on DHS data (signal intensities) and RNA-seq data (FPKM), and assumed a simple linear relationship between transcript expression and DHSs signal intensities. Log (FPKM) values for dependent variable, and log (DHSs signal intensities) for independent variables for each transcript were given as input to the multiple linear regression R function which is mentioned below.

Stats=lm(Y ~ X1 + X2…Xn, data)

Here lm is the linear model, Y is the dependent variable and X1, X2 up to Xn are dependent variables and data is the input matrix containing all these variables.

Most of the transcripts were mapped to many CRRs; and we have just 10 expression data points, one for each cell type. The statistics of CRRs mapping to the nearby transcripts (genes) are detailed in the Table 4.2. There are several predictors (CRRs) for each transcript, with only 10 data points it is unreasonable to build models on more than a few predictors and therefore methods are needed to select predictors e.g., two well correlated chosen/selected CRRs per transcript.

Table 4. 2: Statistics of CRRs (Candidate cis Regulatory Regions) mapping with the transcripts within 20kb and 40kb distance

	20kb mapping distance	**40kb mapping distance**
Average CRRs per transcript	11.8	22.4
Minimum CRRs per transcript	1	1
Maximum CRRs per transcript	37	62

To achieve large number of significant (highly correlated) models; we chose to limit ourselves to genes expressed in reasonable number of cell lines (4 or more) since such data is more suited to regression modelling.

4.2.2 Fold changes

Fold change is a measure of changes from initial to a final value. If the initial value is A and final value is B then the fold change will be B/A, a change from 40 to 20 would be 0.5. The disadvantage of this method is that it is biased and may miss differentially expressed genes with large differences (B-A) [121]. This method can be used for analyzing gene expression data in RNA-seq data [122]. Here, we used regression to model the relationship of expression fold changes to DHS signal fold changes. These 10 data points were converted into 45 data points by applying fold change method.

4.2.3 Methods for choosing the Candidate cis Regulatory Regions (CRRs)

4.2.3.1 Method 1: Choosing CRRs according to high TF binding

Transcription factor binding on cis regulatory elements have important role in regulation of the genes, such as binding of c-Myc and its heterodimeric partner Max on cis regulatory region promotes the malignant transformation in Burkitt's lymphoma cells [123].

CRRs were mapped to all transcripts within the window of 20kb and 40kb distances. Only two CRRs were chosen according to highest ratio of TF binding events occurring in 1 of 5 cell types, ratios were considered because each cell type has different number of mapped transcription factors.

4.2.3.2 Method 2: Choosing CRRs by position in promoter region

CRRs were mapped to all transcripts within the 20kb distance. 1st CRR was chosen from within the 1kb of TSS (Transcription start site) of target gene (transcript) and 2nd CRR was chosen on the basis of ratio of TF binding events

occurred (excluding the surrounding 1kb regions of TSS). Here we are considering both proximal and distal regulatory regions. Region surrounding TSS is the promoter such as TATA box which is located 25-35 base pairs upstream from the transcription start site [124]; and this promoter along with the distal cis regulatory regions control the transcription of genes.

4.2.3.3 Method 3: Choosing CRRs by closest distance

In this method, CRRs were mapped to all transcripts within the 20kb distance. Two closest CRRs to the TSS were chosen if they are at equal distance from the TSS otherwise only single closest CRR was chosen.

4.2.3.4 Method 4: Choosing CRRs according to high conservation score

Cis regulatory regions can be conserved in evolution as evidenced by genes such as Hox4 [125]. Here, the conservation score for each tag (here tag is basically a sequencing read, each CRR contain several tags; certain regions of CRR are highly conserved and they may have some role in controlling regulation of genes) in candidate regulatory elements was calculated. One way is to consider the maximum score of any tag in CRR or take the mean of all tags score in a CRR. CRRs were mapped to the transcripts within the 20kb and only two CRRs were chosen on the basis of their higher mean conservation score and higher maximum conservation score.

4.3 Results

4.3.1 Choosing CRRs by method 1: High TF binding

Here, CRRs were mapped to the transcripts within 20kb and then we increased the distance just to analyze that how increase in mapping distance would affect the model building? CRRs for each transcript in 40kb are more than CRRs per transcript in 20kb as mentioned in the Table 4.2.

Here, 558 models have adjusted R-square >0.65 (adjusted R-square is modified form of R-square and can be calculated by comparing explanatory power of repression models) and 1680 models have adjusted R-square > 0.5 at the 20kb

mapping distance, and statistics of this model building are detailed in the Table 4.3. Two CRRs were chosen for each model (transcript) because we have only 10 predictors, therefore it is better to limit chosen control elements (independent variables) as discussed in the method section.

RHOB transcript (ENST00000272233.4) is the example of 20kb set and one of best correlated model shown in Figure 4.1 (A). Two chosen CRRs are highly and positively correlated with the transcription fold changes, this indicates that both these chosen regions might enhance the transcription in RHOB gene. Expression fold changes are increasing with the increase in regulatory inputs fold changes (from left bottom to right top) as shown in Figure 4.1(A).

When CRRs were mapped to transcripts within the 40kb, only 481 models have adjusted R-square >0.65, and statistics of model building are given in Table 4.3. Numbers of correlated models were decreased from when CRRs were mapped to transcripts within the 20kb because in 20kb, proximal regulatory regions have higher chance to be selected on the basis of TF binding and have higher probability to be correlated with the actual expression fold change of transcript.
One of the best correlation model example is SF3B1 (ENST00000470268.1) shown in Figure 4.1 (B). In this model (transcript), both chosen CRRs have 0.932 and 0.754 correlations between DHSs signal intensities fold change and gene expression fold change as shown in Figure 4.1 (C&D) respectively.

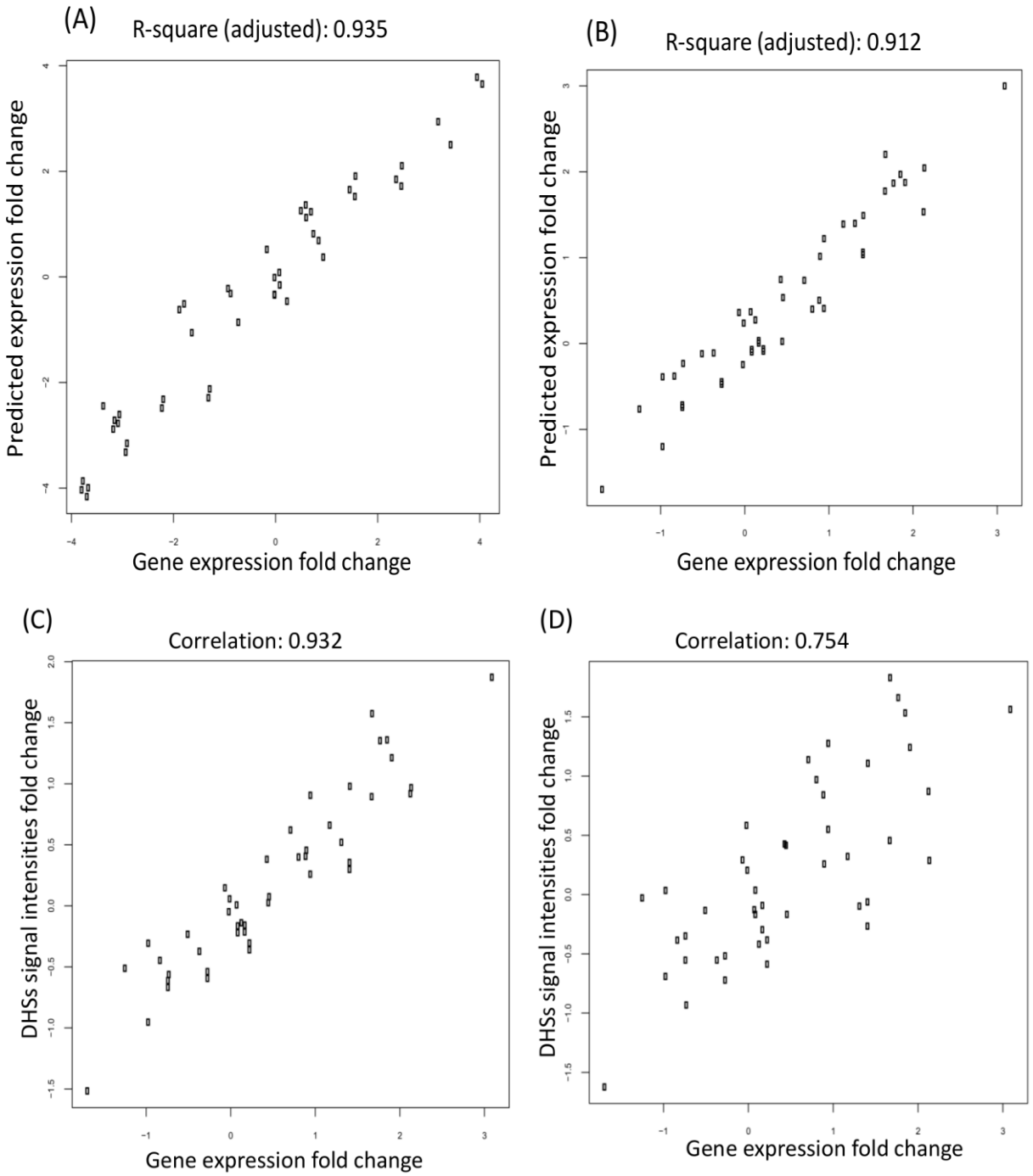


Figure 4. 1: (A) Well correlated model of RHOB showing correlation between predicted expression fold change and gene expression fold change. (B) SF3B1 example for 40kb showing correlation between predicted expression fold change and gene expression fold change. (C&D) Panels show chosen CRRs (CRR1 and CRR2) correlation between DHSs signal intensities fold change and gene expression fold change for SF3B1 gene.

A large number of models are not well correlated possibly because all chosen

CRRs according to high TF binding are not predictive of gene expression, ETS1 and LRP8 are example of not well correlated model are shown in Figure 4.2 (A) and (B) respectively.

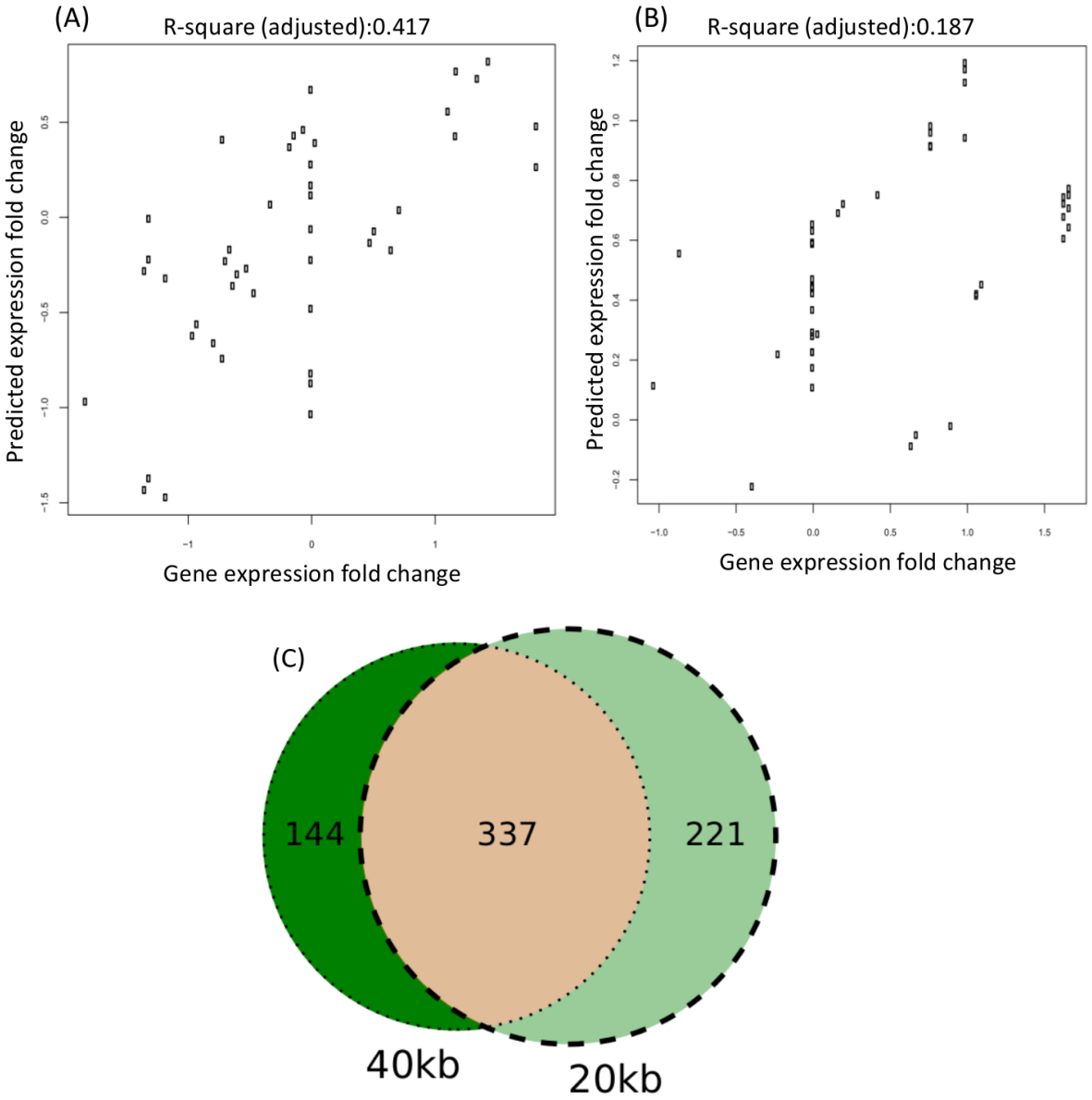


Figure 4. 2: (A) shows the correlation between gene expression fold change and predicted expression fold change in ETS1 gene, this is not well correlated model. (B) Shows the correlation between gene expression fold change and predicted expression fold change in LRP8 gene model where adjusted R-square is less than 0.2 (0.187). (C) 337 transcripts (models) have adjusted R-square>0.65 in both sets (When CRRs were mapped to transcripts within the 20kb and 40kb distance, hence they are two different sets).

A total of 337 models (These are same transcripts in both sets and have adjusted

R-square >0.65) have adjusted R-square >0.65 when CRRs were mapped to the transcripts within the 20kb and 40kb distance. There are unique models (models which don't have adjusted R-square>0.65 in both sets) and have adjusted square >0.65 as shown in Figure 4.2 (C).

4.3.2 Choosing CRRs by method 2: position in promoter region

Statistics of model building for this method are detailed in the Table 4.3. Here, lesser number of models have adjusted R-square>0.65 than method 1 (20kb mapping distance) (558-537=21). CCT7 (ENST00000464397.1) is one of the example of well correlated models, adjusted R-square for CCT7 gene is 0.914 and its correlation between gene expression fold change and predicted expression fold change is shown in Figure 4.3 (A), its high adjusted R-square is indication that CRRs chosen by this method are predictive of gene expression.

Table 4. 3: Statistics of model building by multiple linear regression for all 4 methods

	Method 1		**Method 2 (20kb)**	**Method 3 (20kb)**	**Method 4 (20kb)**	
	20kb	**40kb**			**Max**	**Mean**
Number of models attempted	20380	20512	20380	20380	20380	20380
Number of models built	19110	18828	18995	7041	17150	15811
Models with adjusted R-square>0.65	558	481	537	150	352	299
Models with adjusted R-square>0.5	1680	1513	1622	448	1096	901

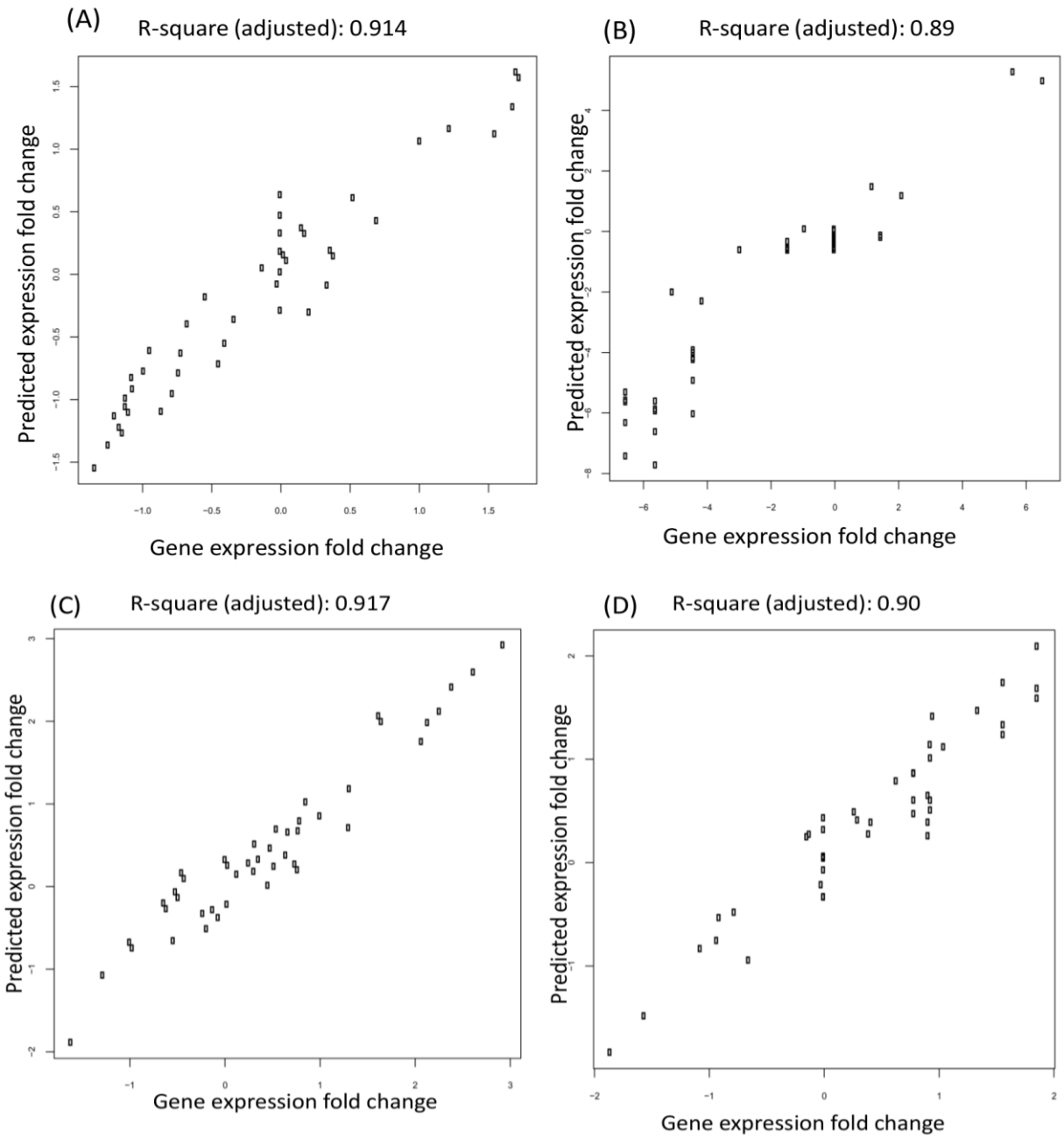


Figure 4. 3: (A) shows the correlation between gene expression fold change and predicted expression fold change in CCT7 gene (p value: 1.45e-23) from method 2. (B) Shows the correlation between gene expression fold change and predicted expression fold change in COL1A2 (p value: 1.25e-21) from method 3. (C) Shows the correlation between gene expression fold change and predicted expression fold change in ACIN1 (p value: 7.35e-24) from method 4 (maximum conservation score). (D) Shows the correlation between gene expression fold change and predicted expression fold change in PPM1H (p value: 3.139e-22) from method 4 (mean conservation score).

4.3.3 Choosing CRRs by method 3: Closest CRRs

In this method, 150 models have adjusted R-square>0.65, which is a smaller number than other methods according to adjusted R-square parameter, and statistics of model building are detailed in Table 4.3. This shows that just by choosing one or two closest CRRs won't help us to build well correlated models. COL1A2 (ENST00000297268.6) is one of the example of well correlated models, its correlation between gene expression fold change and predicted expression fold change is shown in Figure 4.3 (C).

4.3.4 Choosing CRRs by method 4: Conservation score

A total of 352 models have adjusted R-square>0.65 when CRRs were chosen on the basis of higher maximum conservation score. However, only 299 models have adjusted R-square>0.65 when CRRs were chosen according to higher mean conservation score, and statistics of this model building are detailed in the Table 4.3. A total of 250 models from 352 (maximum conservation score) and 299 (mean conservation score) sets are same, which is indication that both ways of choosing CRRs don't differ much; however, we have more well correlated models when we choose models according to the higher maximum conservation score than by choosing higher mean conservation score. ACIN1 (ENST00000473758.1) is one of best well correlated example of choosing CRRs according to higher maximum conservation score and PPM1H (ENST00000228705.4) is one of best well correlated example of choosing CRRs according to higher mean conservation score are shown in Figure 4.3 (C) and (D) respectively.

202 models have adjusted R-square> 0.65 in this method (high maximum conservation score) and in method 2, 213 models have adjusted R-square>0.65 in this method (high maximum conservation score) and in method 1 as shown in Figure 4.4.

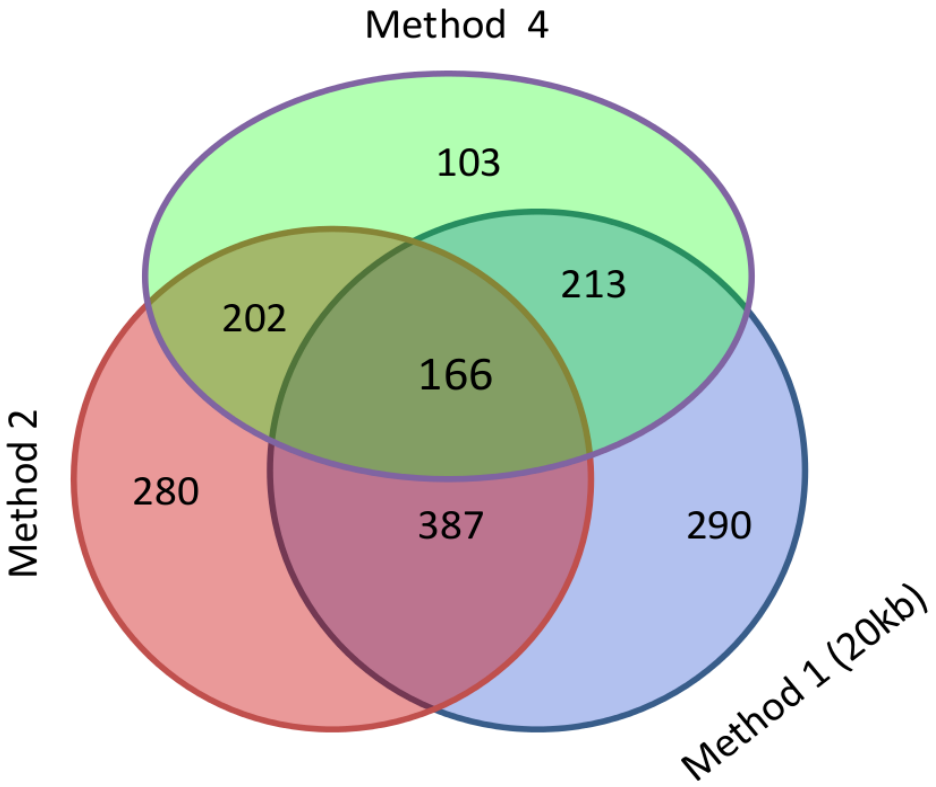


Figure 4. 4: This Venn diagram represents shared models which have adjusted R-square>0.65 from method 1, method 2 and method 4 (where maximum conservation score considered). A total of166 transcripts (models) have adjusted R-square in all three methods, and 387 transcripts have adjusted R-square> 0.65 in method 1 and method 2. A total of 213 models in method 1 and method 4, and 202 models in method 2 and method 4 have adjusted R-square >0.65.

4.3.5 Gene ontology enrichment analysis

We performed gene ontology enrichment analysis using DAVID [126] for the shared and unique gene sets linked to good models of expression in Figure 4.2 (C) and Figure 4.4. Figure 4.2 (C) contains comparison between models built in 20kb and 40kb mapping in 1st method. Similarly Figure 4.4 shows the comparison between method 1, 2 and 4. In gene ontology enrichment analysis, we have considered two ontology terms i.e., Biological process and Molecular function. Table 4.4 contains all enrichment analysis results for shared and unique models between methods. None of the gene set lies significantly in any biological process possibly because our genes are diverse and they are involved in different biological processes. However, almost all of the gene sets are involved in protein binding and poly (A) RNA binding molecular function.

Table 4. 4: This table shows the gene ontology enrichment analysis for the shared and unique gene sets linked to good models of expression built in method 1, 2 and 4.

	Biological process	**Molecular function**
Method 1: 20kb-40kb shared	None is significant	Protein binding (2.7E-5) and poly (A) RNA binding (2.6E-3)
Method 1: 20kb-40kb unique	None is significant	Protein binding (2.4E-5) and poly (A) RNA binding (3.3E-6)
Shared between method 1 & 2	None is significant	poly (A) RNA binding (9.3E-8) and protein binding (2.3E-6)
Unique in method 1 & 2	None is significant	Protein binding (2.3E-4) and identical protein binding (2.7E-2)
Shared between method 1 &4	None is significant	Protein binding (8.1E-6) and poly (A) RNA binding (4.8E-3)
Unique in method 1 &4	None is significant	Protein binding (6.0E-4) , poly (A) RNA binding (3.6E-3) and electron carrier activity (4.7E-2)
Shared between method 2 & 4	None is significant	Protein binding (3.2E-4) and poly (A) RNA binding (5.8E-4)
Unique in Method 2 & 4	None is significant	Protein binding (1.5E-5) and poly (A) RNA binding (1.9E-5)
Shared between method 1,2 and 4	None is significant	Protein binding (9.6E-5) and poly (A) RNA binding (3.0E-4)
Unique in method 1, 2 and 4	None is significant	Protein binding (2.2E-7) and poly (A) RNA binding (3.2E-4)

4.4 Discussion

In this chapter, we have investigated the question what is the best way to choose CRRs that predict gene expression and may regulate genes? Method 1 has the highest number of models with an adjusted R-square >0.65 as detailed in the Table 4.3, which indicates that TFs binding information is more predictive of gene expression than other parameters. In other studies, scientist had used TF binding sites for prediction of cis regulatory regions e.g., Benjamin et al., identified cis regulatory modules through TF binding sites clusters in Drosophila [8]. Researchers have also studied TF binding in different way, Theresa et al., estimate the activity of transcription factors at regulatory region by their collective effects on the gene expression levels [127]. There are other existing studies which suggest the association between transcription factors and gene expression levels. Chao et al., developed a method which can determine the TF activity changes from microarray expression profiles, and they evaluated the significance of TF activity changes by permutation tests [128]. Here, we have chosen regulatory regions on the basis of highest TF binding believing that highest TF binding is linked with the expression level of genes.

A total of 537 models have adjusted R-square>0.65 in method 2, which indicates that CRRs can be located in the proximal regions and also in distal regions, here the 2nd CRR was chosen according to highest amounts of TF binding, this again supports the previous argument that TF binding is helpful in choosing CRRs. Our analysis also supports the existing knowledge that large number of cis regulatory regions/promoters are located in the proximal regions [129].

Method 3 does not have large number of models with an adjusted R-square >0.65, which indicates that it is not necessary for genes to be regulated by their closest cis regulatory regions, as cis-regulatory regions can be located up to 1mb away from the TSS [130]. This method also suggest that choosing only one CRR won't lead to building better models (adjusted R-square>0.65). Here, our purpose is to choose regulatory regions which can predict gene expression. Therefore, regions which have significant role in regulation of genes can be present anywhere within the bracket of 1mb, and by limiting our method to only

closest CRR, we are losing important distance cis regulatory regions that lead to less number of well correlated models. Therefore, this method would not helpful in identifying cis regulatory regions.

Cis-regulatory regions can be conserved, hence we investigated that the conservation scores might be helpful in predicting cis-regulatory regions. In method 4 (maximum conservation score), 352 models have adjusted R-square>0.65, which shows that conservation can also be used to choose the CRRs. Both the maximum conservation score and the mean conservation score are considered here in choosing CRRs but according to the adjusted R-square, it is better to consider the maximum conservation score.

We performed gene ontology enrichment analysis using DAVID tool [126] on the well correlated shared and unique models between method 1, 2 and 4. All these set of genes were investigated for their significant involvement in biological process but none of genes set was involved in any biological process. This indicates that these genes are diverse and involved in several biological processes However, most of the genes set are involved in protein binding and poly (A) RNA binding molecular function. Though, we assume that these all the genes have diverse functions but significantly they fall into these two molecular functions.

In this chapter, we concluded that TF binding and conservation features can be helpful in identifying cis regulatory regions. We can choose best CRRs without splitting mapping window, as we have done in the method 1 and method 2. Based on these results, we have used TF binding and conservation score features together for filtering potential regulatory regions in the 5th chapter. Further, we also need to increase the mapping distance of transcripts to all CRRs, here we considered up to 40kb for only 1st method and it can be increased 100 kb in the next chapter.

Chapter 5

5 Predicting cis regulatory regions by using LASSO (Least Absolute Shrinkage and Selection Operator)

In Chapter 4, we used TF binding, conservation score and position of potential cis regulatory regions to choose the CRRs (Candidate Cis Regulatory Regions). There should be a way to penalize the CRRs which are not predictive of gene expression because choosing CRRs on the basis of functional features for example TF binding and conservation score may miss some important regulators. Therefore, we can use a regression method which can penalize CRRs that are not predictive of gene expression. In the previous Chapter, we have shown that TF binding and conservation score are helpful in choosing the best potential regulators. Therefore, we can still consider these features for filtering the number of CRRs per transcript.

There is an approach called LASSO, which is a type of regression that involves in penalizing the absolute size of the regression coefficients [131]. Qabaja et al., identified associations between disease and miRNA by LASSO regression using a disease signature [132]. Some researchers have used LASSO to select and classify biomarkers in genomic data [133]. Jie et al. [134] identified long non-coding RNA in mouse by integrative modelling of chromatin and genomic features. Ploeg and Steyerberg predicted biologically relevant features for classification of type of infections using LASSO [135]. Wang et al. [136] predicted the interactions between the microRNA and mRNA using LASSO. Several existing computational methods for predicting regulatory regions are discussed in the introduction of 4th chapter.

LASSO is important when the number of potential coefficients are large and penalizing them is required as in our case each transcript is mapped to a large number of CRRs. There is a tuning parameter called lambda for LASSO to penalize the irrelevant CRRs. The strength of LASSO can be increased and overfitting can be decreased by filtering the input if we consider lambda.min tuning parameter [137]. This lambda.min is the value of lambda that gives

minimum mean cross validated error. Tuning parameter lambda influences the prediction results and it is difficult to determine this tuning parameter [138] [139]. Further, LASSO is explained in the methods section.

Predicting enhancers would help researchers to understand the regulation of genes and experimental techniques are expensive and time consuming. There are experimental techniques for enhancer identification, for example enhancer trap method and also new technique called site-specific integration fluorescence-activated cell sorting followed by sequencing (SIF-seq) [140].

Our method of using LASSO to choose CRRs and their target gene is different from existing methods as several researchers had developed computational methods for identification of enhancers [141] [142].

In the beginning of this chapter we have predicted gene expression fold change by training LASSO models on CRR DHSs signal intensities fold change and transcript expression fold change. Later on we have predicted just gene expression by training LASSO models on CRR DHSs signal intensities and transcript expression and reason is explained in that section. In the start of this chapter CRRs were obtained from the ChIP-seq data only and then we obtained CRRs from both ChIP-seq and DHSs data. In the end of this chapter, we have predicted CRRs from LASSO method and LASSO with filtered input (CRRs were filtered by using TF binding and conservation scores) method.

5.1 LASSO with Fold change method

Here we used the LASSO to model the relationship between expression fold change and DHSs signal intensities fold change. Ten (10) data points (cell types) were converted into 45 fold changes, as in 4th chapter.

5.1.1 Method

All the CRRs were obtained from ChIP-sq data, as explained in the Chapter 4, and were allowed to map to transcripts within 40kb and also within 100kb separately (hg19 genome version was used). Those transcripts which expressed in at least in four cell types, as discussed in the previous chapter were chosen.

Following is the equation of LASSO where y is the dependent variable (log (FPKM) values) fold changes of transcript), x^i are the independent variables (log (DHSs signal intensities)) fold changes for the i^{th} candidate regulatory region, and n is the number of CRRs located within the chosen distance of TSS of particular transcript.

$$y = k_0 + \sum_{i=1}^{n} k_i x_i$$

Since n is typically greater than the number of cell types for which data were available, model fitting demanded a penalized approach to limit the number of non-zero k_i coefficients. We chose LASSO regression implemented in the R glmnet package [143]. Regularization and tuning parameter for the LASSO is lambda and LASSO penalizes the irrelevant predictors based on lambda, and penalty can be calculated as below.

$$\textbf{Penalty} = \lambda \sum_i |k_i|$$

As shown in above equation, penalty depends on lambda and number of non-zero coefficients. These coefficients/CRRs can be penalized by lambda.min (λ_{min}) and lambda.1se as shown in Figure 5.1 (C). Lambda.min is the value of lambda that gives minimum mean cross validated error and lambda.1se shows that error is within one standard error of the minimum.

5.1.2 Results

As mentioned in the method section that CRRs were mapped to all transcripts within the distance of 40kb and 100kb separately, therefore, results are split in two below parts.

5.1.2.1 40kb

We found that by using penalty parameter lambda.min which is described in methods, models typically have 4 or more non-zero coefficients which leads to the correlation>0.75 in more than 4000 models, and statistics of model building

are mentioned in the Table 5.1. Knowing the fact that this choice of lambda is based on minimum mean square error from cross validation. FTSJ2 (ENST00000242257.8) is the example where, LASSO chooses 4 CRRs for this transcript, and its rank is 2963 when models were ranked according to the descending order of correlation, but still predicted expression and observed expression fold change are highly correlated as shown in Figure 5.1 (A) that is probably overfitting. Therefore, we decided to reduce the risk of overfitting by applying stronger penalty and restricting LASSO models to at most 3 CRRs per transcript.

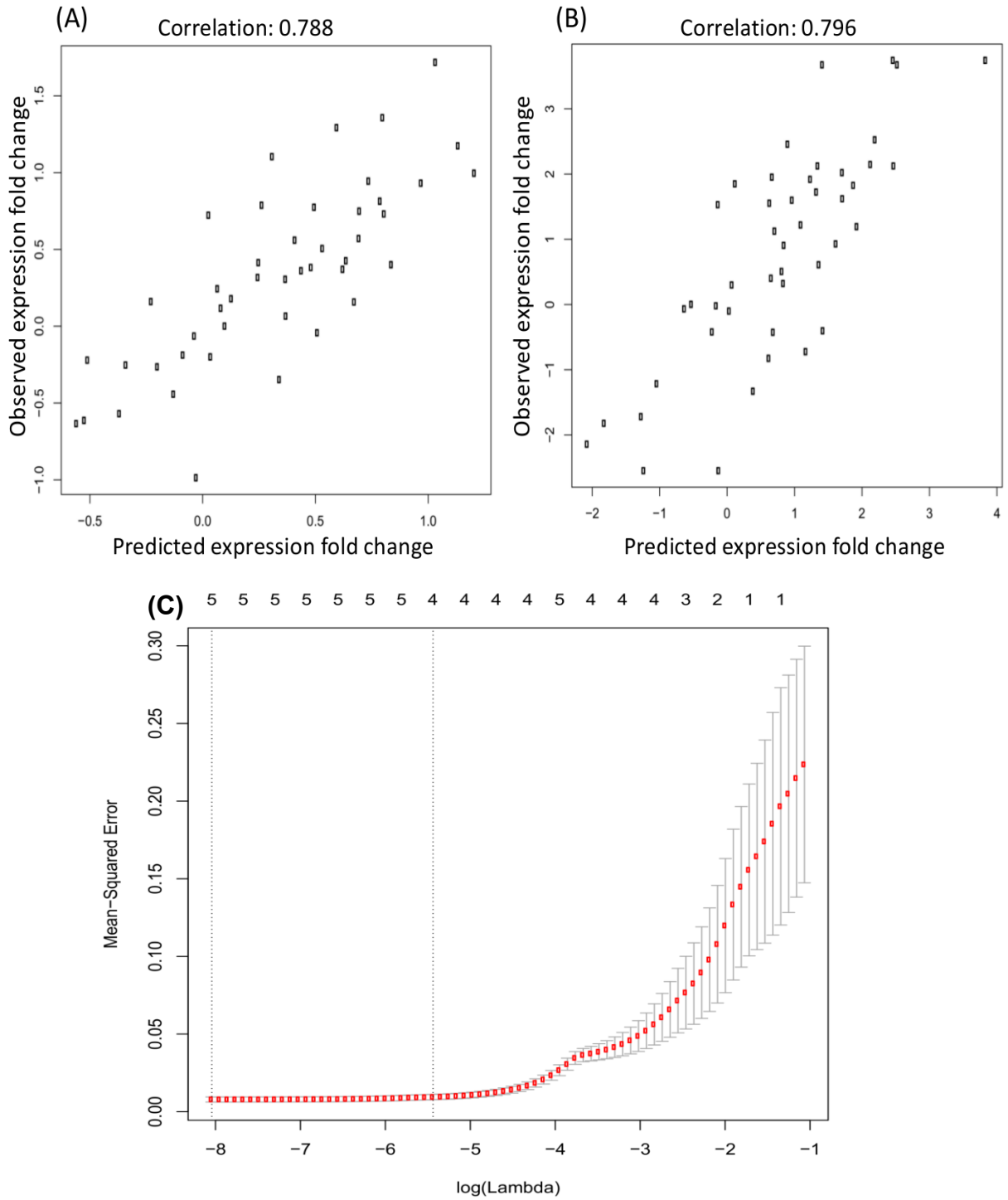


Figure 5. 1: (A) FTSJ2 is the example of 40kb mapping, where correlation (r=0.788) between observed expression fold change and predicted expression fold change is shown (B) IDI1 is the example of 100kb mapping, and here correlation (r=0.796) between observed expression fold change and predicted expression fold change is shown. (C) This is the example for lambda as a tuning parameter. Dotted lines show possible choices of λ at minimum mean-squared error (λ_{min}) and more conservatively at that value plus 1 standard error, and numbers above graph indicate number of predictive variables.

5.1.2.2 100kb

Here we have increased the mapping distance from 40kb to100kb, therefore, each transcript on average have 32 more CRRs than above method and this increase in number of candidates has caused over fitting. Here, 7437 models have correlation>0.75, and statistics of model building are detailed in the Table 5.1. IDI1 (ENST00000381344.3) is the example where LASSO chooses 4 CRRs for this transcript; and its rank based on descending order of correlation is 5481 and still it's highly correlated as shown in Figure 5.1 (B).

Table 5. 1: This table shows the statistics of model building for LASSO (40 & 100kb mapping)

	40kb	**100kb**
Number of models attempted	20512	20611
Number of models built	17670	19308
Models with correlation>0.75	4008	7437
Models with correlation>0.65	6940	11288
On average CRRs per transcript	22	54

5.1.3 Discussion

In this method, CRRs were mapped to all transcripts within 40kb and 100kb distances. Here, we have observed that well correlated models increase by increasing the mapping distance. Therefore it is better to consider the 100kb mapping distance. Well correlated models might increase because of increase in over fitting as CRRs for each transcript was increased by increasing the mapping distance. There are large number of well correlated models at lambda.min that leads to the overfitting as explained in the results section. Therefore, it was decided to restrict the chosen CRRs up to 3 per transcript.

5.2 LASSO without Fold change method

5.2.1 Introduction

In previous parts of thesis we aimed to predict expression fold changes from fold changes in activity at CRRs. However, in optimizing our methods we moved to the simpler approach of predicting absolute expression levels from absolute levels of activity at CRRs. This is better since there is no obvious cell type to use as a reference and inclusion of all possible fold changes in a model results in a set of variables that are not all independent.

To optimise our method, we added 5 more cell types increasing the number of data points on which each model is based, and therefore improving reliability. Huvec, Mcf7, Nhek, Nhlf and Sknshra cell types were added to the analysis (15=10+5). Later we realize that most of our mapped transcripts are not expressing in Hmec cells, therefore, this cell type was removed from the analysis and list of 14 cell types is given in Table 5.3 (page 115).

5.2.2 Method

We considered two ways of obtaining CRRs, based on ChIP-seq data and DNase-seq data (ENCODE 2012 release). Obtaining CRRs from ChIP-seq is explained in the Chapter 4, and the top 25% (based on peak intensity) DNase-seq data were considered for obtaining CRRs and procedure of derivation is similar to the obtaining CRRs from ChIP-seq. We analysed CRRs from ChIP-seq and DNase-seq separately to see which set of CRRs is the most predictive of gene expression.

Once were CRRs obtained then the DHSs signal intensities were extracted from bigwig files for each cell type and for all CRRs obtained from ChIP-seq and DNase-seq data. All CRRs were mapped to all transcripts within the 100kb distance. Here, we allow LASSO to choose only 3 best CRRs based on tuning parameter lambda to avoid over fitting. Only those transcripts, which express at least in 7 cell types were considered for further analysis (This 7 threshold was optimised to produce better results as we also tried other thresholds but this works better). Model was trained based on DHSs signal intensities of CRRs and

transcript expression values (FPKM) as explained in the previous section.

5.2.3 Results

Here, we built models from CRRs obtained from ChIP-seq and DNase-seq separately, results from both sets are given below.

5.2.3.1 CRRs obtained from ChIP-seq data

Here, LASSO built 16164 models successfully of which 1582 models have correlation>0.9. The statistics of model building are detailed in the Table 5.2. PPP1R11 (ENST00000376769.2) is the example of one of the best correlations and its correlation between observed expression and predicted expression is shown in Figure 5.2 (A).

Large number of models are well correlated as mentioned in the Table 5.2 but we need to analyze certain number of top correlated models and we thought to consider top 2000 models for further analysis. Chosen CRRs with positive coefficients (DHSs signal intensities are positively correlated with the expression) were considered as potential enhancers and otherwise potential repressors of gene expression. We considered the three coordinate model as discussed in the method section, so for 2000 models and we have 6000 chosen CRRs; 1618 of them are repressors while remaining 4382 chosen CRRs are enhancers.

A total of 403 chosen CRRs (enhancers and repressors) are regulating (here, we are using regulation term for those CRRs who are chosen by LASSO for particular transcript but we are not yet sure that these CRRs are regulating their respective transcript) more than one transcript, and 280 from these 403 (311 are enhancers, while 92 are repressors) are regulating different transcripts but the same gene. We have 4071 enhancers (not repeated), some of them potentially be regulating more than one transcript.

Other studies have suggested that different genes may be regulated by the same enhancers, and can be gene clusters [144]. These 4071 enhancers regulate 1633 genes and 265 of these enhancers are involved in controlling gene regulation of more than one gene.

These enhancers were overlapped with the transcription factors from K562,

Gm12878, Hepg2, Helas3 and H1hesc cell types separately. Number of transcription factors overlap with each enhancer from each cell type varies.

Number of overlapped TFs from these five cell types were added for each enhancer and plotted in histogram as shown in Figure 5.2 (B).

Figure 5.2 (C) shows the distance between the enhancers and transcription start site (TSS) of correlated transcript. Large number of enhancers lies in the proximal region and hundreds of enhancers are also located in the distal regions. In addition to that, enhancers are also located in upstream and downstream of TSS of the transcript.

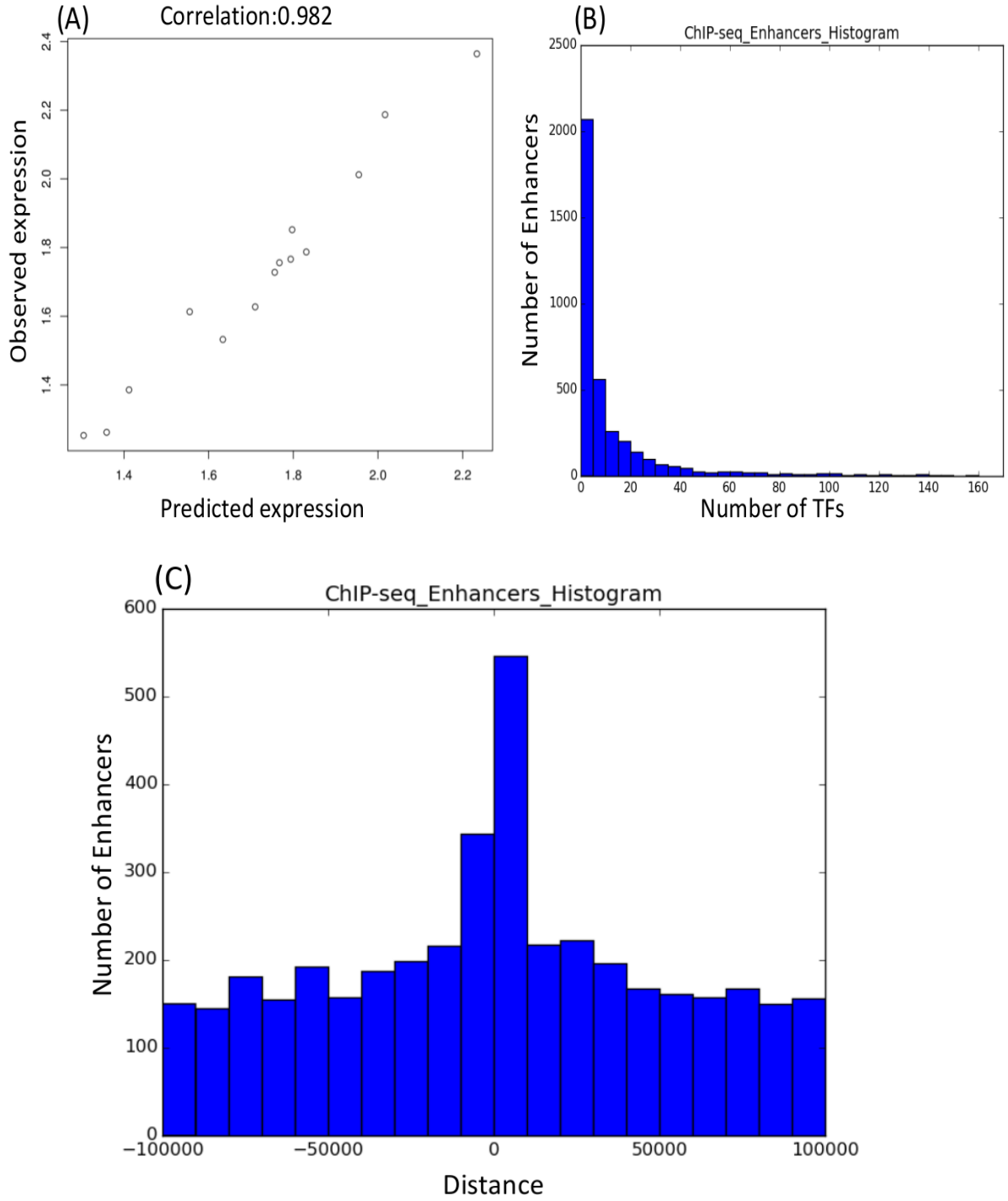


Figure 5. 2: (A) shows the correlation between observed expression and predicted expression in PPP1R11. (B) This histogram shows the number of TFs binding on the enhancers (where correlation is positive between DHSs signal intensities and transcript expression), i.e., up to five TFs bind on more than 2000 enhancers. (C) This histogram shows the distance of enhancers from their respective TSS, both upstream and downstream enhancers distances are shown here.

5.2.3.1.1 Validation

Our knowledge of experimentally validated enhancers in the human genome is limited, with the VISTA Enhancer database [145] contains only 1790, so validation is difficult. We only found 16 enhancers because we have considered limited number of cell types and only globally expressed transcripts. One enhancer of these 16 is known to regulate WNT5A gene and our method have also predicted that enhancer for WNT5A gene.

Table 5. 2: Statistics of model building by LASSO for CRRs obtained from ChIP-seq and DNase-seq data separately.

	CRRs obtained from ChIP-seq	CRRs obtained from DNase-seq
Number of models attempted	18098	17957
Number of models built	16164	16148
Models with correlation>0.9	1582	1258
Models with correlation>0.8	7920	6372
Enhancers in top 2000 models	4382	4566
Repressors in top 2000 models	1618	1434

5.2.3.2 CRRs obtained from DNase-seq data

LASSO built 16148 models successfully, here 1258 models have correlation> 0.9, and statistics of model building are detailed in the Table 5.2.

EPN1 (ENST00000270460.5) is one of the example of well correlated models shown in Figure 5.3 (A).

Again here, large number of models are well correlated but number of well correlated models are lesser than the CRRs obtained from ChIP-seq data as detailed in the Table 5.2. We considered top well correlated 2000 models for further analysis as we did for previous method, three CRRs were chosen by LASSO; therefore we have analyzed 6000 chosen CRRs. A total of 1434 of these CRRs are repressors (where CRRs DHSs signal intensities are negatively correlated with the transcript expression), while remaining 4566 are enhancers

(where CRRs DHSs signal intensities are positively correlated with the transcript expression). A total of 543 chosen CRRs (430 are enhancers and 113 are repressors) are found to regulate more than on transcript and 340 out of these 543 (430 are enhancers while 113 are repressors) are regulating multiple transcripts but their gene is same.

We have 4136 enhancers, and some of them are regulating more than one transcript from a single gene but very few also regulate more than one gene. These 4136 enhancers are regulating 1600 genes and 346 of these enhancers regulate more than one gene. Chosen enhancers were overlapped with all the TFs from K562, Gm12878, Hepg2, Helas3 and H1hesc cell types separately. Number of overlapped TFs from these five cell types were added for each enhancer and plotted in histogram as shown in Figure 5.3 (B).

Distances from enhancers and TSS of transcripts are plotted in histogram and shown in Figure 5.3 (C). Pattern of distances show that large number of enhancers lies in the proximal region, and some of the enhancers are also located in the upstream and downstream of their respective transcripts.

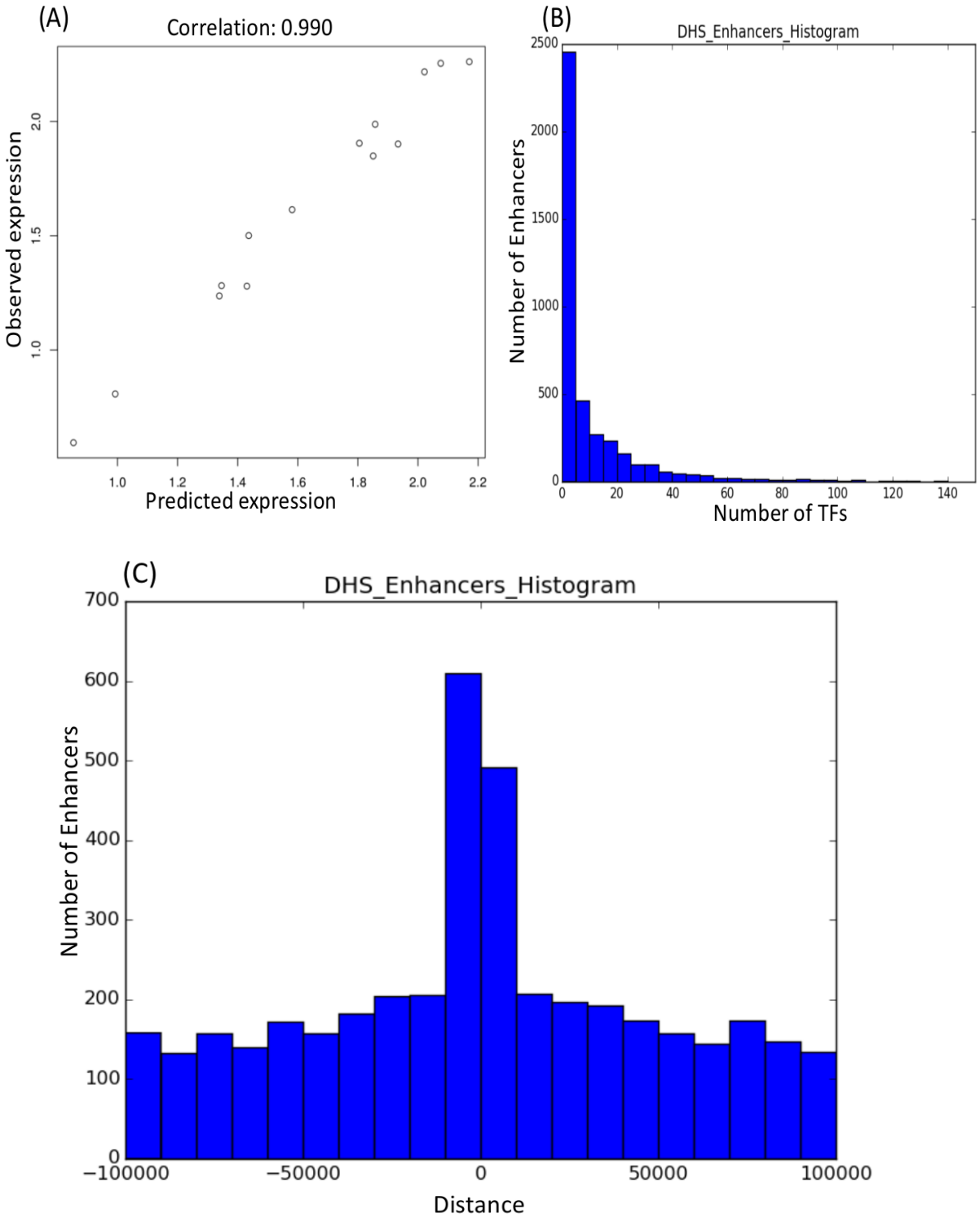


Figure 5. 3: (A) shows the correlation (r=0.99) between observed expression and predicted expression in EPN1. (B) This histogram shows the number of TFs bind on the enhancers (where correlation is positive between DHSs signal intensities and transcript expression), i.e., up to five TFs bind on approximately 2400 enhancers. (C) This histogram shows the distance of enhancers from their respective of TSS, both upstream and downstream enhancers distances are shown here.

5.2.3.2.1 Validation

Our knowledge of experimentally validated enhancers in the human genome is limited, with the VISTA Enhancer database [145] containing only 1790, so validation is difficult. We only found 10 predicted enhancers in VISTA, because we have considered limited number of cell types and only globally expressed transcripts. One enhancer of these 10 is known to regulate the HHEX gene and we have also predicted that enhancer for HHEX gene.

5.2.4 Discussion

Significant number of models are well correlated as mentioned in the Table 5.2, which indicates that CRRs obtained from ChIP-seq and DNase-seq data can be predictive of gene expression. We analysed top 2000 models, with 4382 enhancers and 4566 enhancers from ChIP-seq and DNase-seq data respectively, 1363 enhancers are common in both sets. Therefore, it would be better to obtain CRRs from merged set of ChIP-seq and DNase-seq data. In this section, we restricted the number of CRRs per transcript but still we need to optimise the quantity of chosen CRRs per transcript.

5.3 Optimised method for predicting regulatory region using LASSO

5.3.1 Introduction

Here we have predicted cis regulatory regions from integrative analysis of ChIP-seq, DNase-seq and RNA-seq data and here CRRs were obtained from merged set of ChIP-seq and DNase-seq data. We used LASSO to predict the models of gene expression in following two different ways.

1. All the CRRs (Candidate cis Regulatory Regions) were given as input to the LASSO.
2. CRRs were filtered on the basis of TF binding and conservation score.

We also used randomisation for estimating false discovery rate and these methods are explained in below methods section.

5.3.2 Methods

5.3.2.1 Dataset

A total of 14 cell types from ENCODE were considered for generating our model, the details of which are given in Table 5.3. Transcription factor binding sites (TFBS) and H3K27ac data, were obtained as ChIP-seq peaks for the five cell types. DNase-seq peaks and RNA-seq data for 14 cell types were obtained. In addition to that DHSs signal intensities for all identified CRRs were obtained from 14 cell types considered here. A schematic representation of the optimised methodology is shown in Figure 5.4 (flow chart).

Table 5. 3: This table shows the complete data set; tick mark in the column represent that data in the column was used for cell types in the row.

S.No	Cell types	ChIP-seq	DNase-seq	RNase-seq (FPKM)	H3K27ac
1	K562	✓	✓	✓	✓
2	Gm12878	✓	✓	✓	✓
3	Hepg2	✓	✓	✓	✓
4	Helas3	✓	✓	✓	✓
5	H1hesc	✓	✓	✓	✓
6	A549		✓	✓	
7	Ag04450		✓	✓	
8	Bj		✓	✓	
9	Hsmm		✓	✓	
10	Huvec		✓	✓	
11	Mcf7		✓	✓	
12	Nhek		✓	✓	
13	Nhlf		✓	✓	
14	Sknshra		✓	✓	

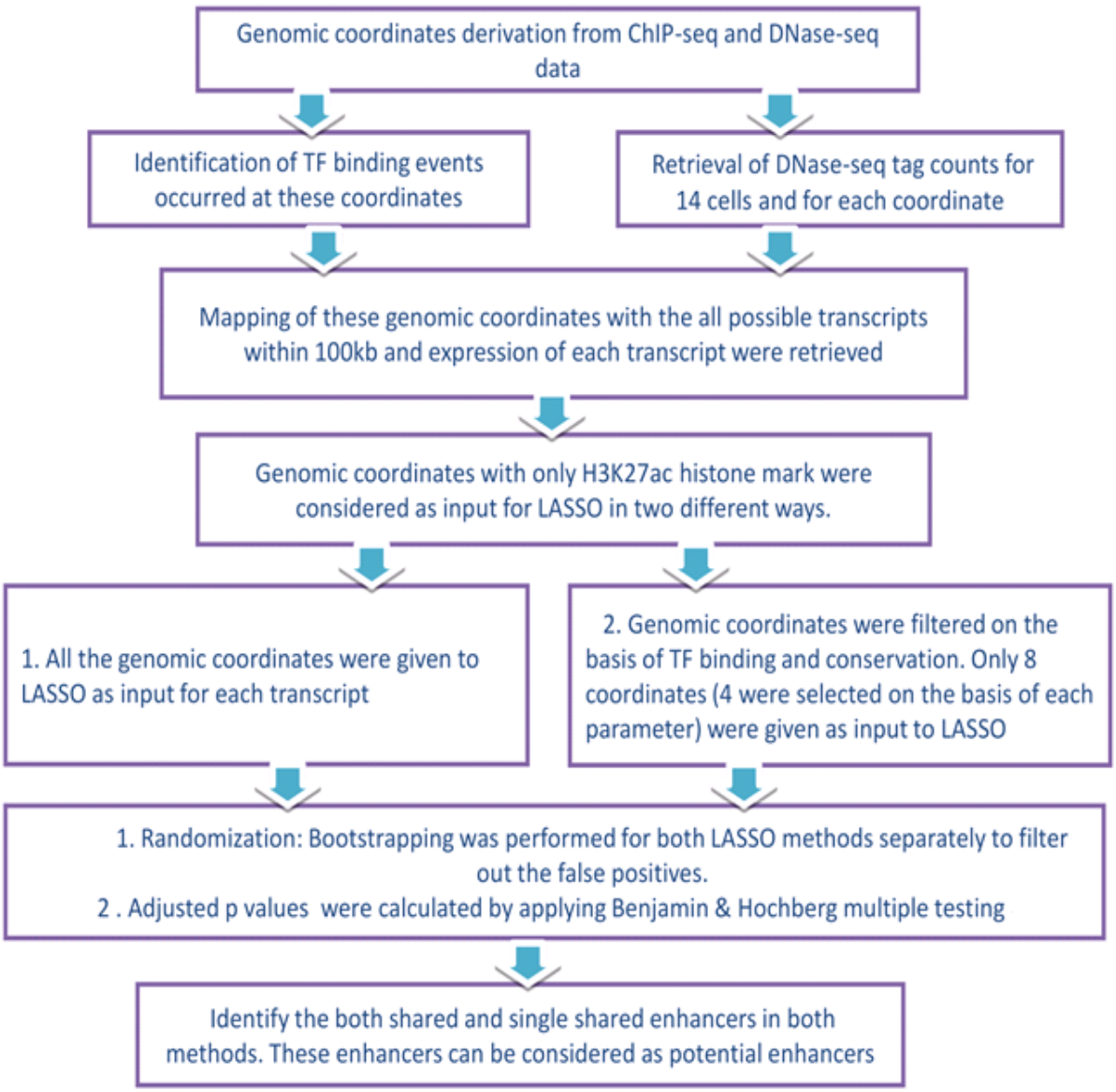


Figure 5. 4: This figure shows the methodology of predicting cis regulatory regions. Here, genomic coordinates are the CRRs and DNase-seq tag counts are DHSs signal intensities.

5.3.2.2 Identifying potential CRRs (Candidate cis Regulatory Regions)

CRRs (Candidate cis Regulatory Regions) were obtained from a merged set of ChIP-seq data for five cell types and DNase-seq data for 14 cell types, however, only top 25% of DNase-seq peaks (on the basis of peak intensity) were considered, these cell types are mentioned in the Table 5.3.

Transcription factor binding events occurring at each CRRs were retrieved by

overlapping these CRRs with the transcription factor binding sites for five cell types. Further, DHSs signal intensities were retrieved for all the obtained CRRs from 14 cell types. These CRRs were mapped to all the transcripts within the 100kb distance and transcript expression (FPKM) values were also retrieved for each transcript in 14 cell types.

Only those candidate cis regulatory regions (CRRs) which have H3K27ac histone mark were selected for further analysis, as this histone mark represents the active enhancers. This mark was considered because it shows that the enhancer is active [20] and we decided to filter those CRRs which don't contain this histone mark. Some researchers [146] have also considered the p300 binding sites for determining the potential enhancers but P300 mostly binds on the promoter and we considered H3k27ac mark instead of P300 to include most of the potential enhancers irrespective of their position to TSS. Transcripts which were expressed at least in 7 cell types (this threshold works better in building models than other thresholds-see in previous section) were considered further analysis. Each transcript was mapped to the several (on average ~42) candidate regulatory elements. Therefore, LASSO has to choose only best potential cis regulatory regions.

5.3.2.3 Predicting models of gene expression

LASSO is explained in the beginning of this Chapter and here, we have used LASSO regression in following two ways.

1. In 1st method, LASSO was allowed to choose the two best CRRs (those who are predictive of expression) and penalizes others, and we have referred this method in results section as a LASSO method. We have included R code for building LASSO models in **Appendix II**.

2. In 2nd method; CRRs were filtered by TF binding and conservation score, 4 CRRs were selected according to highest number of TF binding events and 4 CRRs were selected according to highest conservation score. We have referred this method in results as a filtered LASSO method. Here CRRs were filtered to reduce the false discovery rate and we will get more significant models by giving just biologically important CRRs as an input to the LASSO. We gave DHSs signal intensities of these 8 CRRs as an input along with the particular transcript

expression (FPKM) values of 14 cell types to the LASSO and LASSO was allowed to choose only 2 best CRRs.

5.3.2.4 Statistical testing by Randomisation

We performed randomisation to estimate the false discovery rate. Transcripts were ranked according to the correlation between predicted and observed transcript expression and we picked 4 transcripts with the high correlations, 4 with medium correlations (0.5) and 4 transcripts where LASSO did not find any CRRs (LASSO didn't build models).Models were chosen from each category to avoid any bias towards any group because these randomisations will be used for all the groups. However, distributions of correlations predicted from randomised CRRs were similar in all the groups. DHSs signal intensities in each transcript were shuffled and LASSO was allowed to pick the best two CRRs, and this exercise was repeated 50,000 times. So, we have 600,000 = (12× 50,000) randomisations. In 2nd method, we increased number of transcripts from 12 to 24 and randomisations were decreased from 50,000 to 25,000 per transcript but the total number of randomisations remain same (600,000=24×25000).

In 1st method, expression values (dependent variable) were also shuffled instead of CRRs DHSs signal intensities (independent variable) and same procedure was followed but results were similar with the above method (where we shuffled CRRs). Therefore, we considered only one way which is shuffling the CRRs (explained in the above paragraph).

Empirical p values were calculated from the observed and random correlations (correlation between observed and predicted expression).

5.3.2.5 Multiple testing corrections

Empirical p values were adjusted by Benjamini and Hochberg [69] method to estimate the false discovery rate.

5.3.3 Results

5.3.3.1 LASSO

Here, we used LASSO penalty that permitted at most two or three non-zero coefficients (CRRs), which is generally more stringent than lambda.min because we preferred a conservative approach to the issue of possible overfitting. In genomic investigations it is important to assess the statistical significance of the results and we did this by randomisation as explained in the methods. Here, we have potential issue of overfitting, as each transcript (model) has many possible predictive variables (42 CRRs per transcript on average) and only small number of gene expression measures. Although LASSO is specifically designed to address this issue through penalization of the likelihood, we made a detailed study of the relationship of the LASSO parameters on the statistical significance of the models obtained. We also assessed the degree of penalization, investigating penalties leading to models with different number of CRRs, at least for globally expressed genes. The main concept is that the more predictive variables (CRRs) are allowed, the more likely overfitting. We assess the significance level in two coordinate model system (where only two CRRs were chosen) and in three coordinate model system (where only three CRRs were chosen) and we found that a larger quantity of models were significant in two coordinate model than three coordinate model. P values for two and three coordinate models are shown in Figure 5.5 A and B respectively.

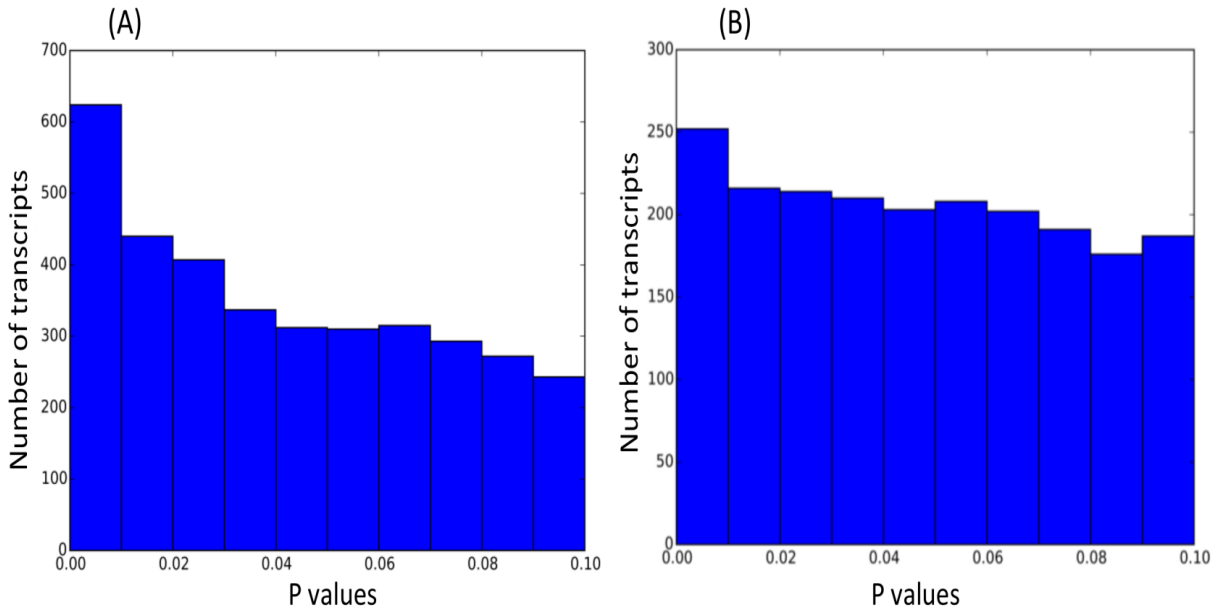


Figure 5. 5: Figure A and B show the p values of models for two coordinate and three coordinate models respectively. More than 600 models are significant (p values <0.01) in (A) and ~250 models are significant in (B), suggesting to consider two coordinate model.

5.3.3.1.1 LASSO method

As mentioned in the method section that LASSO was allowed to pick the best two CRRs as large number of significant correlated models (transcripts) are significant and total number of models are 17963. A summary of the results is given in the Table 5.4 and list of chosen CRRs along with their target genes is given in the additional file 1 (given in CD). We used randomisation to estimate the false positives. We only got 4 significant models (RPS4Y1 (ENST00000430575.1), KDM5D (ENST00000317961.4), RPS4Y1 (ENST00000250784.7), and SPARC (ENST00000520687.1)) after adjusting p values by Benjamini & Hochberg. Detailed results of LASSO method are mentioned in the Table 5.4**.** Chosen CRRs of models having correlation >=0.7 were divided into 1st chosen CRR and 2nd chosen CRR. It was observed that both chosen CRRs are located in upstream or downstream; and statistics of chosen CRRs location with respect to the TSS of transcript are detailed in the Table 5.5**.** Histograms showing distance between CRRs and TSS of their target genes are shown in Figure 5.6 (A & B).

We have only 4 significant models after correcting p value that suggests that we

can build significant models for individual genes, attempting this model building genome wide would lead to large false discovery rate. Therefore, we decided to build models for genes with high (top 25%) log of variance of gene expression and for genes with high (top 25%) log of coefficient of variance of gene expression separately. We got 6 significant models from each set of genes.

In this method, we have not enough number of significant models possibly because, we have given large number (on average 42) of CRRs for each transcript to LASSO. Therefore, we decided to filter the CRRs per transcript according to high TF binding and high evolutionary conservation score.

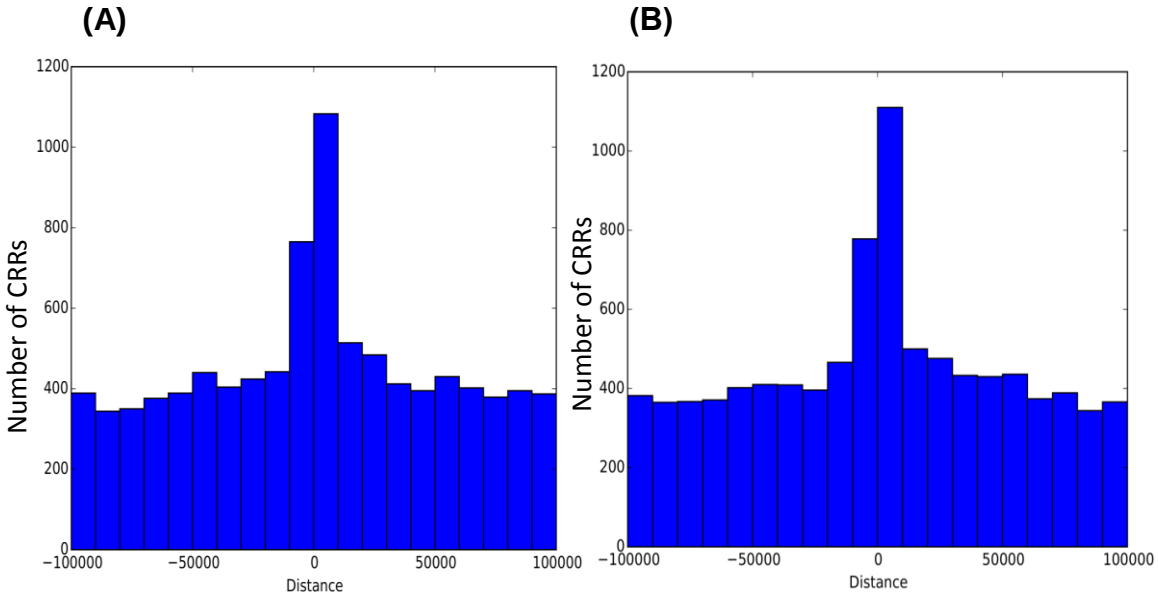


Figure 5. 6(A&B): These histograms show distances between the chosen CRR1 (A), CRR2 (B) and TSS. These CRRs are chosen by LASSO method.

5.3.3.1.2 LASSO with filtered input

Genomic features can be used to identify enhancers [147] and we can filter CRRs on the basis of biological features such as H3K27ac, TF binding and conservation score and we have explained this in Chapter 4.

Here, we have filtered 8 CRRs for each transcript (model), 4 were chosen on the basis of highest number of transcription factor binding sites and 4 were chosen on the basis of highest conservation score.

Both parameters TF binding and conservation score are important for identification of potential regulatory regions, these bound transcription factor have

some role and it was also observed in some studies that rate of transcription of gene would increase with the increase in TF binding. Here we are considering several factors such as conservation, TF occupancy and histone marks for filtering CRRs. We might miss some potential regulators by filtering them with the conservation criteria but they might will be picked according to high TF binding. Here again, LASSO was allowed to choose two best CRRs.

Table 5. 4: This table shows the statistics of model building by LASSO and LASSO with filtered input

Methods	Number of models built	Average number of CRRs per model	Average correlation between observed and predicted expression	Significant models after randomisation; p value <0.05	Standard deviation	Significant models after multiple testing correction
LASSO	16134	42	0.7105	1808	0.120	4
LASSO with filtered input	16056	8	0.5962	1979	0.14768	18

In this method, large number of models (transcripts) have high correlations and approximately 1321 have correlations>0.8. Distances of CRR1 and CRR2 with TSS for models with correlation>=0.7 are shown in Figure 5.7 A and B respectively. Chosen CRRs are located in upstream and downstream, or only in upstream or only in downstream of their target genes, and statistics of enhancers (chosen CRRs with positive correlation) are detailed in the Table 5.5.

Table 5. 5: Statistics of CRRs position with respect to their target genes

Methods	Genes (transcripts) regulated by upstream regulatory elements Threshold= correlation>=0.7	Genes regulated by downstream regulatory elements	Genes regulated by the upstream and downstream enhancers	Repressors
LASSO	2303	2838	4063	5022
Filtered LASSO	992	1321	1854	1958
Both shared	170	253	335	350

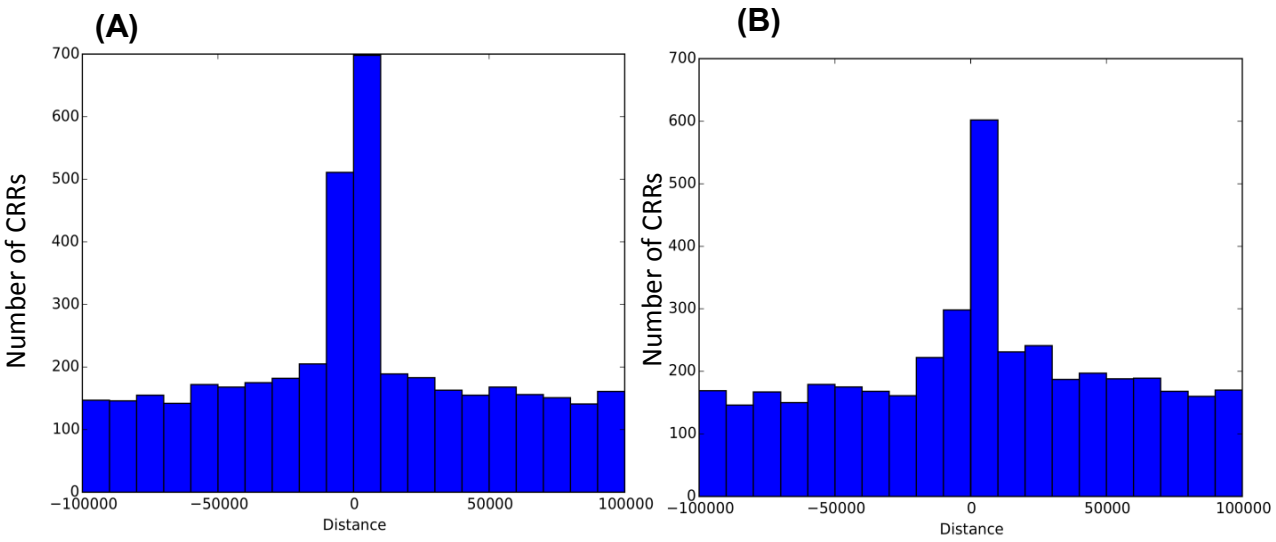


Figure 5. 7: These histograms show distances between chosen CRR1 (A), chosen CRR2 (B) and TSS for models with correlation> 0.7. These CRRs are chosen filtered LASSO method.

We used randomisation to estimate the false discovery rate. Empirical p values were calculated from observed correlations and random correlations, and statistics of results are detailed in the Table 5.4. P values of 1979 significant models after randomisation (p value<0.05) are shown in Figure 5.8 (A).

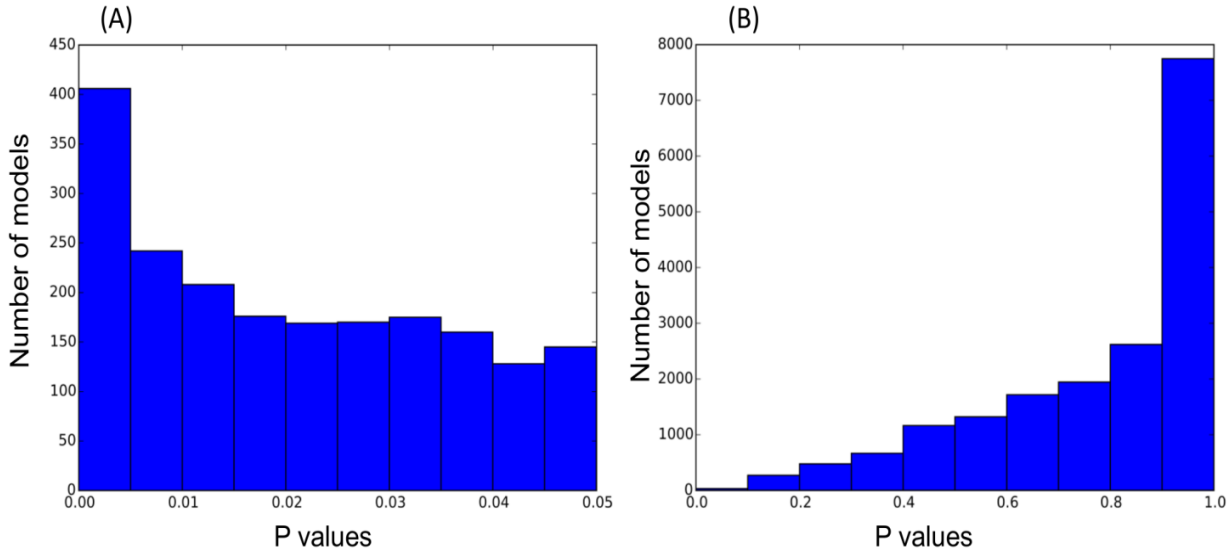


Figure 5. 8: (A) shows the distribution of p values (<=0.05) after randomisation from filtered LASSO method and (B) shows the distribution of p values after multiple testing correction from filtered LASSO method.

It is possible that significant p values are might be significant because of random by chance, so we applied Benjamini & Hochberg (Multiple testing correction) for correcting them. Only 18 significant models (p value <0.05) were significant after correction, these p values are shown in Figure 5.8 (B).

We have only 18 significant models after correcting p value that suggests that we can build significant models for individual genes, attempting this model building genome wide would lead to large false discovery rate, as discussed above. Therefore, we decided to build models for genes with high (top 25%) log of variance of gene expression [148] and for genes with high (top 25%) log of coefficient of variance of gene expression separately. There are 28 significant models from the log of variance of gene expression set and 48 models are significant from the log of coefficient of variance of gene expression set.

We build models for genes with high (top 10%) log of coefficient of variance and we built 74 significant models.

5.3.3.1.3 Comparison between LASSO method and filtered LASSO method

Question was asked that how many models (transcripts) have same chosen CRRs predicted in both methods. Both methods predicted both same CRRs for 1930 transcripts, and single same CRR for 4697 transcripts. However, there are 11336 transcripts where not a single CRR is shared in both methods. These numbers are mentioned in the Table 5.6.

These results show that CRRs chosen by LASSO method but not by filtered LASSO method indicates that these CRRs are possibly less conserved and less TF binding events occurred there.

Table 5. 6: This table shows the statistics about the chosen CRRs which were predicted by both methods to regulate the same gene

	Models	**Significant models, correlation>=0.7**	**Significant models, empirical p values<0.05**	**Significant models after p value correction**
Both shared CRRs	1930	796	452	48
Single shared CRR	4697	1954	981	27

Set of transcripts with both shared chosen CRRs (1930 transcripts) have higher correlations than other sets, correlations between observed expression and predicted expression for these models are shown in Figure 5.9 (A). These correlation values are considered from the LASSO method. Both methods would have similar correlations as same CRRs are chosen by both methods, hence, we can consider correlation from one of these two methods. One example of both shared CRRs is CNN3 shown in Figure 5.10. This figure is divided in 6 sub figures i.e., LASSO cross validation curve (A), predicted expression (B), chosen CRR1 correlation (C), chosen CRR2 correlation (D), rejected CRR correlation (E), and CNN3 transcriptional regulation (F). CNN3 transcript was mapped with the 72 CRRs and both methods have chosen both same CRRs that increase the significance of our methods. These chosen CRRs have high TF binding and are evolutionary conserved, and are predictive of gene expression.

If the CRRs were picked by both methods for the same transcripts and these models are also significant then these chosen CRRs can be considered potential cis regulatory regions because they have less probability to be picked in both methods.

We filtered those transcripts which were mapped to the lesser number of CRRs say 10 because if any transcript is mapped to 8 or less CRRs then obviously same CRRs will be picked up in both methods. This also can be narrated in this way that if certain transcript was mapped to the less number of CRRs and two of them are predictive of expression then we can consider these correlated CRRs as potential regulators of transcription for that transcript. Figure 5.9 (B) shows the histogram for distribution of model correlations (those transcripts which were mapped to at least 10 CRRs).

We further filtered the models and considered only those models which should be mapped to at least 20 CRRs; Figure 5.9 (C) shows the distribution of correlation for such models (transcripts).

Models which have at least one shared chosen CRR in both methods were also considered for further analysis, and we have 6627 (1930 + 4697) such models. Statistics of these numbers are shown in Table 5.6. In this set, 1419 models have

correlation>=0.8 and 3439 models have correlation>=0.7. Again here, we can filter them further by setting threshold such as transcript should be at least mapped to 10 CRRs or 20 CRRs. By considering 10 as a threshold for mapping, we got 1372 models have correlation>0.8 and 3258 models have correlation>0.7. By considering mapping threshold of 20 (those transcripts which map with at least 20 CRRs), we got 1176 models have correlation>0.8 and 2706 models have correlation>0.7.

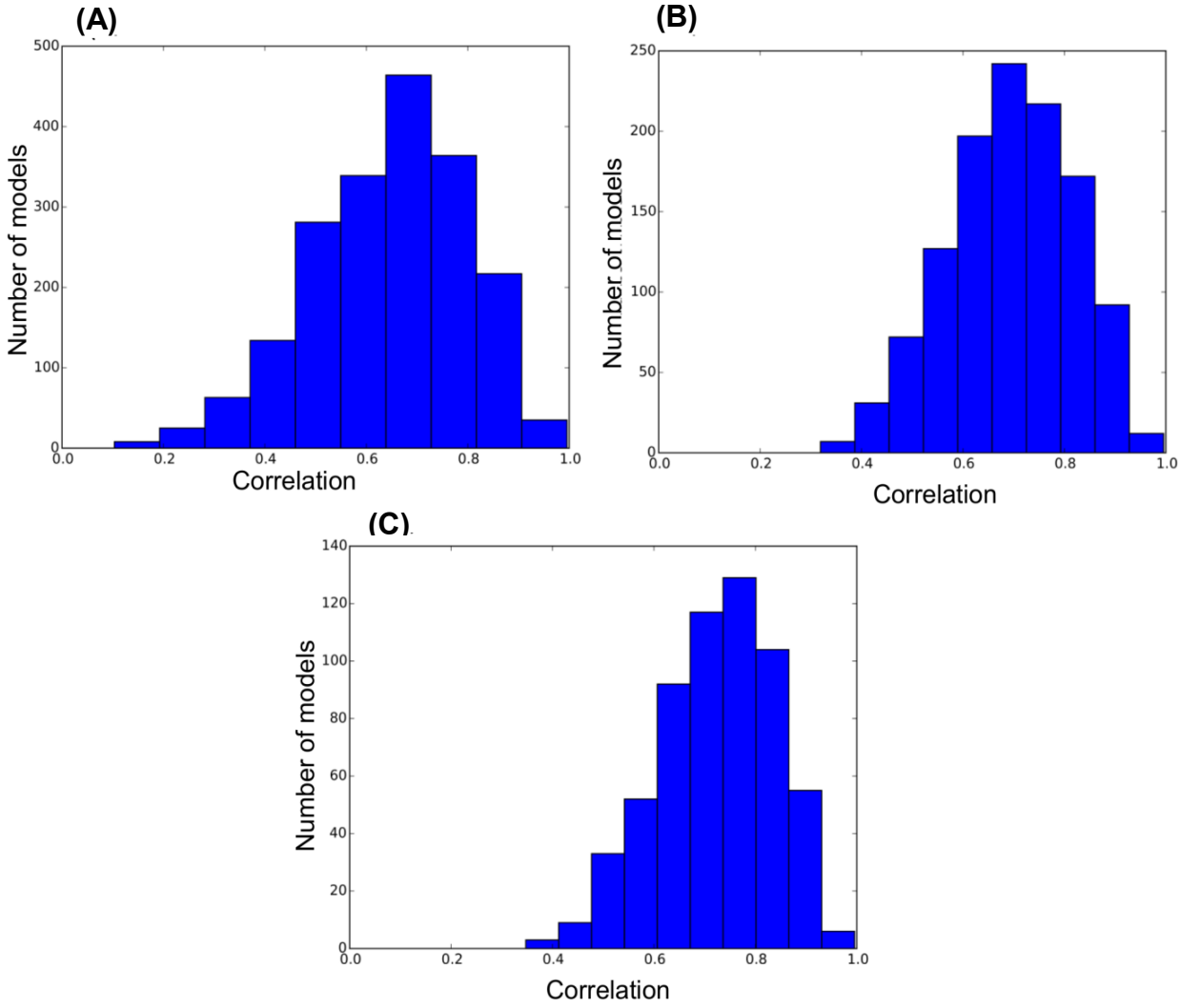


Figure 5. 9:(A) Distribution of correlations between observed and predicted expression from set of transcripts where both CRRs are shared in both methods, (B) Distribution of correlations from the transcripts which are mapped with at least 10 CRRs and have both shared CRRs in both methods, (C) Distribution of correlations from transcripts which are mapped with at least 20 CRRs and have both shared CRRs in both methods.

5.3.3.1.3.1 Statistical significance of models from both shared, single shared and non-shared set

There are 1930 models where both CRRs are same in both methods regulating same target transcripts, 48 of these models are significant after correcting p values (multiple testing correction). A total of 4697 models have only one shared CRR in both methods and 27 models are significant after correcting p values. A total of 11336 models do not share any CRRs for any transcript in both methods and one model is significant after adjusting p values from empirical p values (calculated by randomisation).

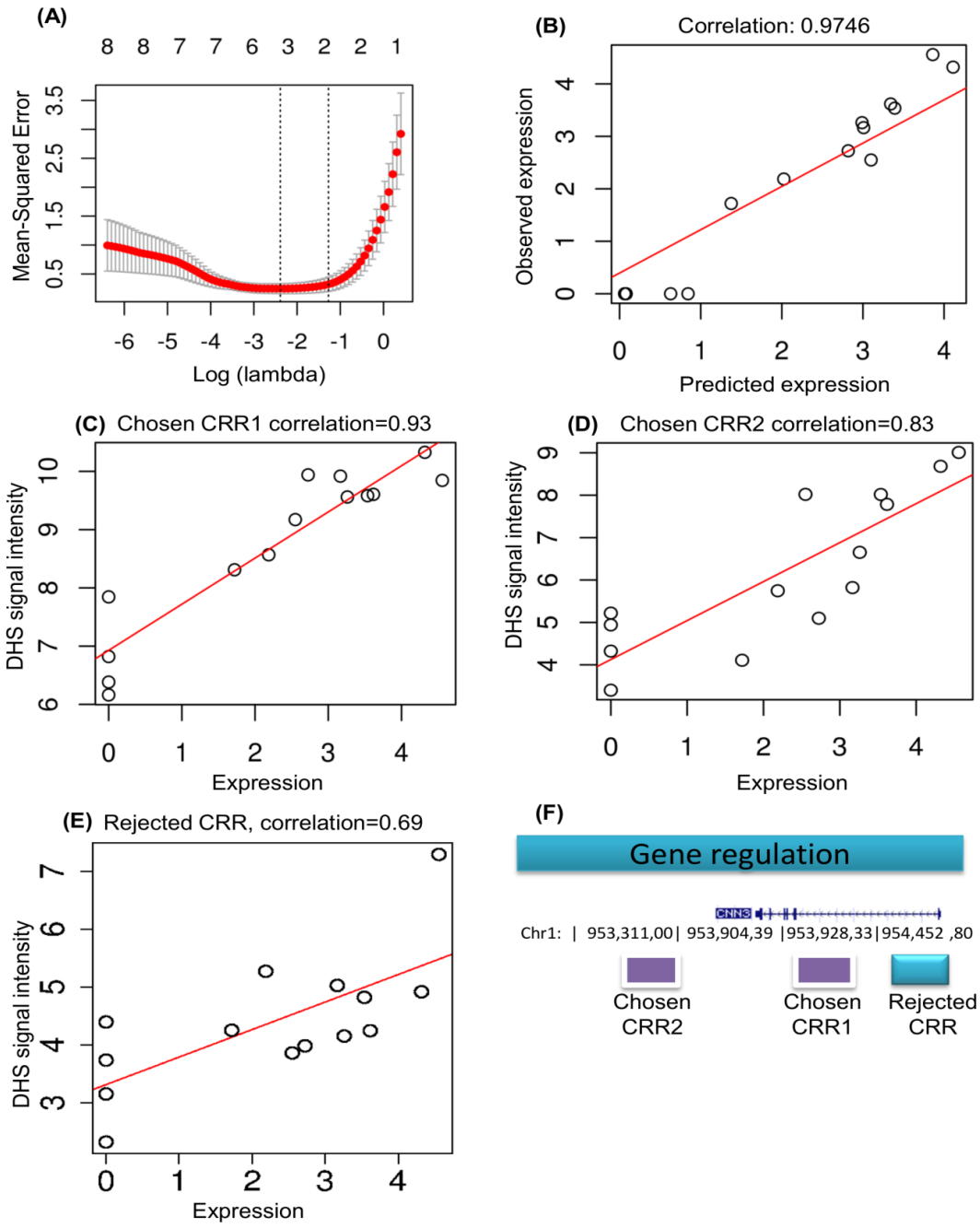


Figure 5. 10: Building an expression model for CNN3 (ENST00000370206.4). (A) shows the mean squared error against the log(λ) LASSO penalty parameter with numbers above the graph indicating the number of predictive variables (non-zero coefficients) in the corresponding LASSO model. Dotted lines show possible choices of λ at minimum mean-squared error (λ_{min}) and more conservatively at that value plus 1 standard error. This identifies models with 2 predictive variables as optimal. (B) Shows the correlation between observed expression and predicted expression from the model (empirical p value of correlation calculated after randomization: 2.83e-05). (C) and (D) show the correlation of DNase-seq signal intensities and expression for the two candidate regulatory elements (CRRs) chosen by the LASSO method. (E) Shows the correlation between DNase-seq signal intensities and expression for an example rejected CRR. (F) Shows the genomic location of the two chosen CRRs and one example of rejected CRR.

5.3.3.2 Analysis of chosen/identified CRRs

We have identified cis regulatory regions from above methods (LASSO and LASSO with filtered input) and we have compared results from both methods by dividing them into three sets, 1. Both methods have chosen both same CRRs (both shared set) for a particular transcript, 2. Both methods have chosen single same CRR for a particular transcript, and 3. Both methods have chosen both different CRRs for a particular transcript. Results from comparison of both methods are detailed in the Table 5.6.

There are 1930 models with both shared CRRs in both methods, and 796 models are well correlated (correlation>0.7) from both shared set. Our methods have identified already known regulators of some of the genes as discussed in the validation section of the each method. Only three chosen CRRs from both shared set are experimentally validated. This shows that not enough number of chosen CRRs are experimentally validated because we have considered only globally expressed genes and limited number of cell types.

WNT5A gene (ENST00000474267.1) is one of the example where our both methods have chosen same CRRs (both shared) and one of them is experimentally validated [145], and graphical representation for regulation of this gene is illustrated in the Figure 5.11. This figure contains both CRRs locations, their DHSs clusters, bound TFs and conservation. Our method (LASSO) has chosen two CRRs from all the CRRs given in Table 5.7, chosen CRRs are highlighted with the red colour. Both these regions are predictive of gene expression, have high TF binding ratio and high conservation score than other CRRs in the Table.

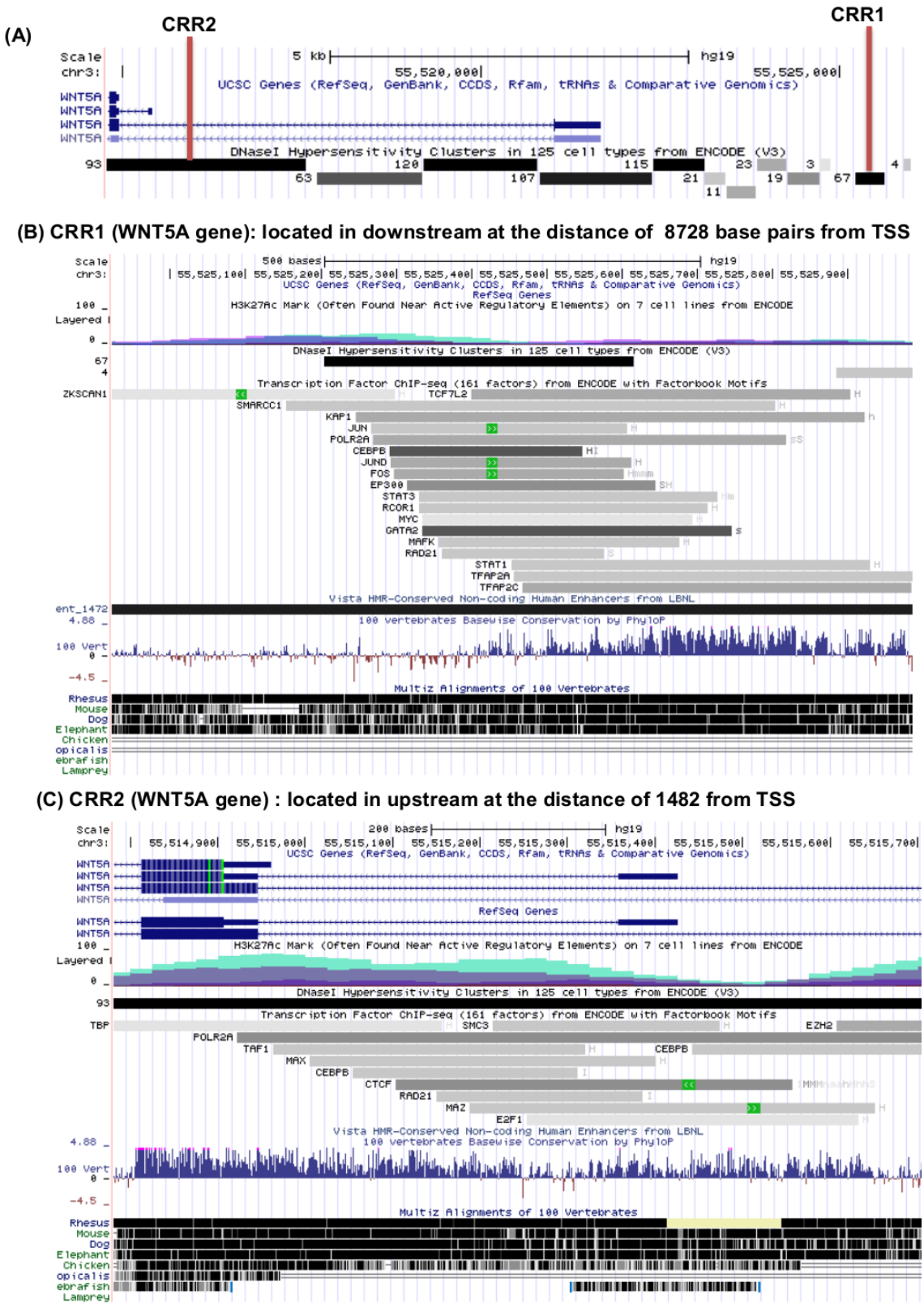


Figure 5. 11: (A) shows the location of chosen CRRs, mentioned as CRR1 and CRR2. B and C represent chosen CRR1 and CRR2 respectively. Both these figure panels (B & C) show H3K27AC signal, DHSs cluster, bound transcription factors, and conservation for CRR1 and CRR2. Location of chosen CRRs with respect to their TSS is also mentioned in their figure panels. CRR1 is experimentally validated and can be seen in Figure (B).

Table 5. 7: This table contains Candidate cis Regulatory Regions (CRRs) mapped to the WNT5A transcript (ENST00000474267.1) within 100kb. Two CRRs (highlighted with red colour) were chosen by LASSO.

Chr #	Start	End	Chr #	Start	End
chr3	55436931	55437292	chr3	55532597	55532897
chr3	55448600	55448750	chr3	55534680	55534830
chr3	55450111	55451016	chr3	55536040	55536190
chr3	55468808	55469379	chr3	55539125	55539506
chr3	55470708	55470994	chr3	55546360	55546530
chr3	55475595	55475824	chr3	55550805	55551144
chr3	55479621	55480062	chr3	55553240	55553390
chr3	55483563	55483842	chr3	55555840	55555990
chr3	55487956	55488219	chr3	55556207	55556536
chr3	55496060	55496470	chr3	55558551	55558870
chr3	55498770	55499145	chr3	55559000	55559150
chr3	55500583	55500902	chr3	55560431	55560793
chr3	55502420	55502570	chr3	55580207	55580446
chr3	55506132	55506407	chr3	55582061	55582310
chr3	55507978	55508926	chr3	55583284	55583525
chr3	55510516	55510879	chr3	55590043	55590318
chr3	55511460	55511610	chr3	55590480	55590582
chr3	55514782	55515707	chr3	55593748	55594027
chr3	55516142	55517571	chr3	55605488	55606410
chr3	55517891	55518515	chr3	55609397	55609660
chr3	55518525	55518675	chr3	55615018	55615257
chr3	55518985	55520824	chr3	55621467	55621830
chr3	55520847	55523137	chr3	55622367	55622778
chr3	55523189	55524273	chr3	55492414	55492754
chr3	55529043	55529386	chr3	55524923	55525986
chr3	55531415	55531694			

There are 4697 models (transcripts) where only one CRR is same in both methods (regulating the same transcript), and 1953 of them have correlations>=0.7, and statistics of this analysis are shown in the Table 5.6. A total of 23 chosen CRRs from single shared set are experimentally validated/known to regulate genes. PBX3 (ENST00000342287.4) is an example where one of the CRR is predicted by both methods and a CRR predicted by filtered LASSO is already known [145]. Therefore, three different CRRs are predicted by both methods. Three predicted regulators (chosen CRRs) are shown in the Figure 5.12, CRR chosen only by filtered LASSO is shown in Figure 5.12 (D) that is highly conserved. These predicted regulators are predictive of gene expression and they are predicted from a list of potential regulators detailed in the Table 5.8.

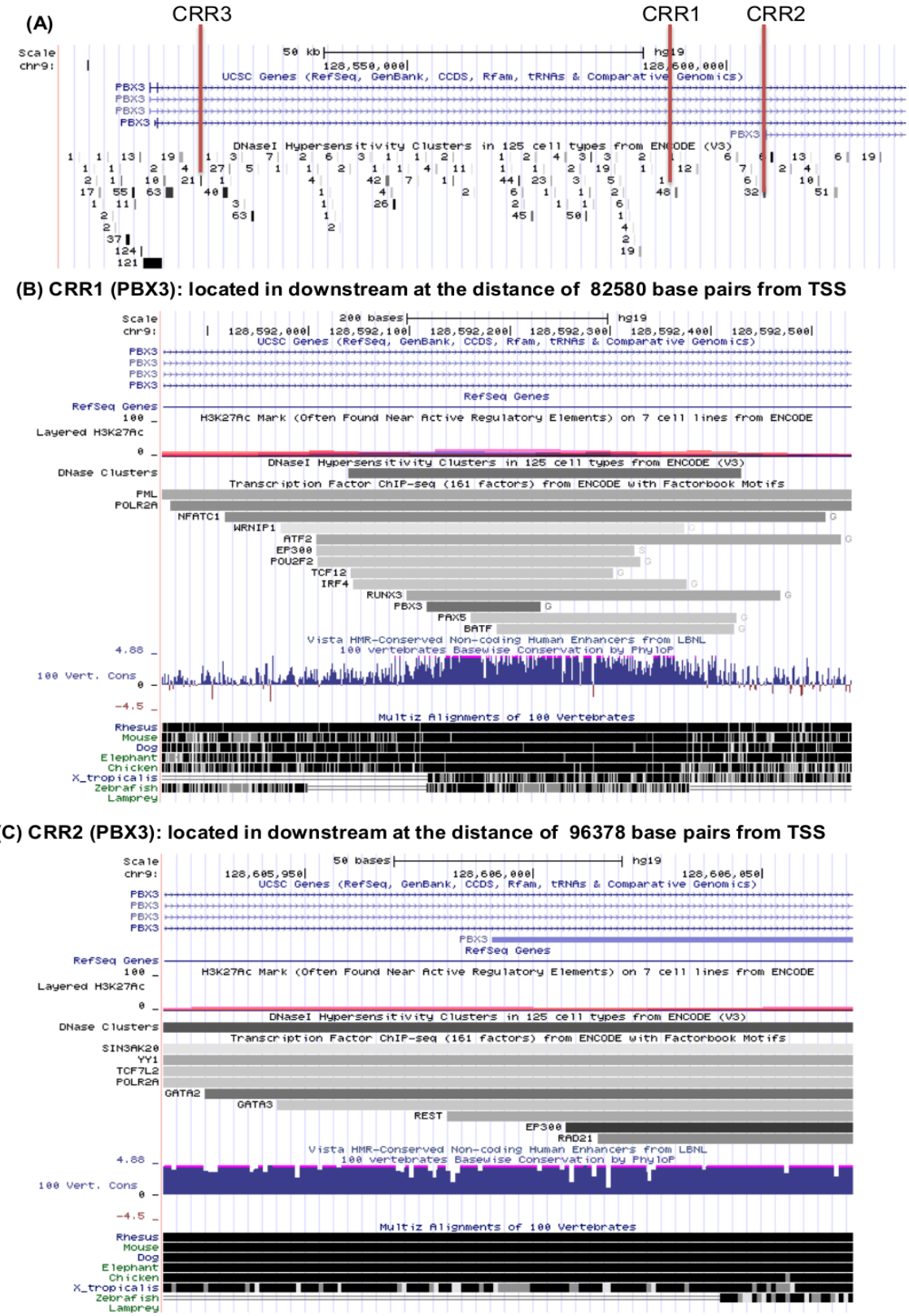


Figure 5. 12: (A) shows the location of chosen CRR1 (LASSO method), CRR2 (both methods), and CRR3 (filtered LASSO). B and C represent chosen CRR1 and CRR2 respectively. Both these figure panels (B & C) show H3K27AC signal, DHSs cluster, bound transcription factors, and conservation for CRR1 and CRR2. Location of chosen CRRs with respect to their TSS is also mentioned in their figure panels. CRR2 is highly conserved and also predictive of gene expression, therefore it was chosen by both methods.

D) CRR3 (PBX3)-Filtered LASSO: Located in downstream at the distance of 8268 base pairs from TSS

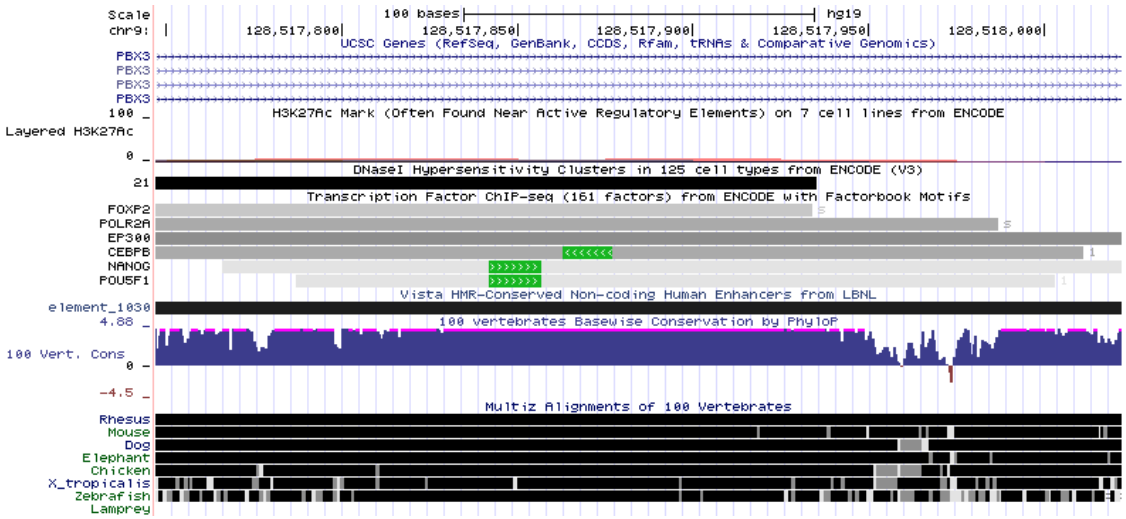


Figure 5.12 (D): This CRR is chosen by the filtered LASSO possibly because of conservedness. This chosen regulatory region is completely lies in the known enhancer that is 1438 bases long. This figure shows H3k27ac signal (weak signal), bound TFs, known enhancer from Vista enhancer database, and conservation score.

Table 5. 8: This table contains all the CRRs mapped to the PBX3 transcript (ENST00000342287.4) within 100kb. CRR1 was chosen by LASSO method that is highlighted with red colour. CRR2 was chosen by both methods and it is highlighted with yellow colour. CRR3 was chosen by filtered LASSO method is highlighted with green colour.

Chr #	Start	End	Chr #	Start	End
chr9	128411380	128411619	chr9	128507617	128510010
chr9	128411700	128411910	chr9	128510065	128511346
chr9	128412406	128413193	chr9	128512586	128513212
chr9	128414397	128414776	chr9	128517748	128518022
chr9	128416387	128416735	chr9	128521153	128522612
chr9	128422120	128422832	chr9	128528785	128529109
chr9	128425388	128425843	chr9	128531188	128531451
chr9	128437536	128438024	chr9	128535130	128535405
chr9	128441872	128442255	chr9	128546680	128546830
chr9	128442769	128443088	chr9	128547549	128547838
chr9	128444006	128444285	chr9	128551534	128551773
chr9	128445757	128446152	chr9	128553607	128553895
chr9	128446535	128446927	chr9	128560036	128560275
chr9	128451045	128451565	chr9	128567440	128567748
chr9	128452989	128453287	chr9	128569448	128569711
chr9	128456818	128457141	chr9	128572891	128573130
chr9	128457837	128458276	chr9	128574648	128575167
chr9	128458406	128458769	chr9	128577957	128578204
chr9	128462991	128464064	chr9	128579712	128579975
chr9	128466340	128466535	chr9	128583297	128584044
chr9	128466987	128467386	chr9	128585207	128585582
chr9	128468047	128468533	chr9	128586224	128586819
chr9	128468812	128469937	chr9	128591856	128592539
chr9	128494765	128495347	chr9	128596808	128597051
chr9	128498336	128498611	chr9	128605920	128606070
chr9	128505205	128505355	chr9	128607618	128607909
chr9	128506441	128506730	chr9	128417559	128417852
chr9	128507040	128507210			

ID1 (ENST00000376105.3) gene is also example of models where one of the predicted CRR is experimentally validated [145]. One of the two CRRs (single shared) for this gene is predicted by both methods, and here we have only presented CRRs predicted by LASSO method. Graphical representation of CRRs chosen by LASSO method for ID1 in Figure 5.13. Large number of TFs bind on both chosen CRRs as shown in Figure 5.13 B and C. These CRRs were chosen from list of CRRs given in Table 5.9, and chosen CRRs are highlighted with red colour.

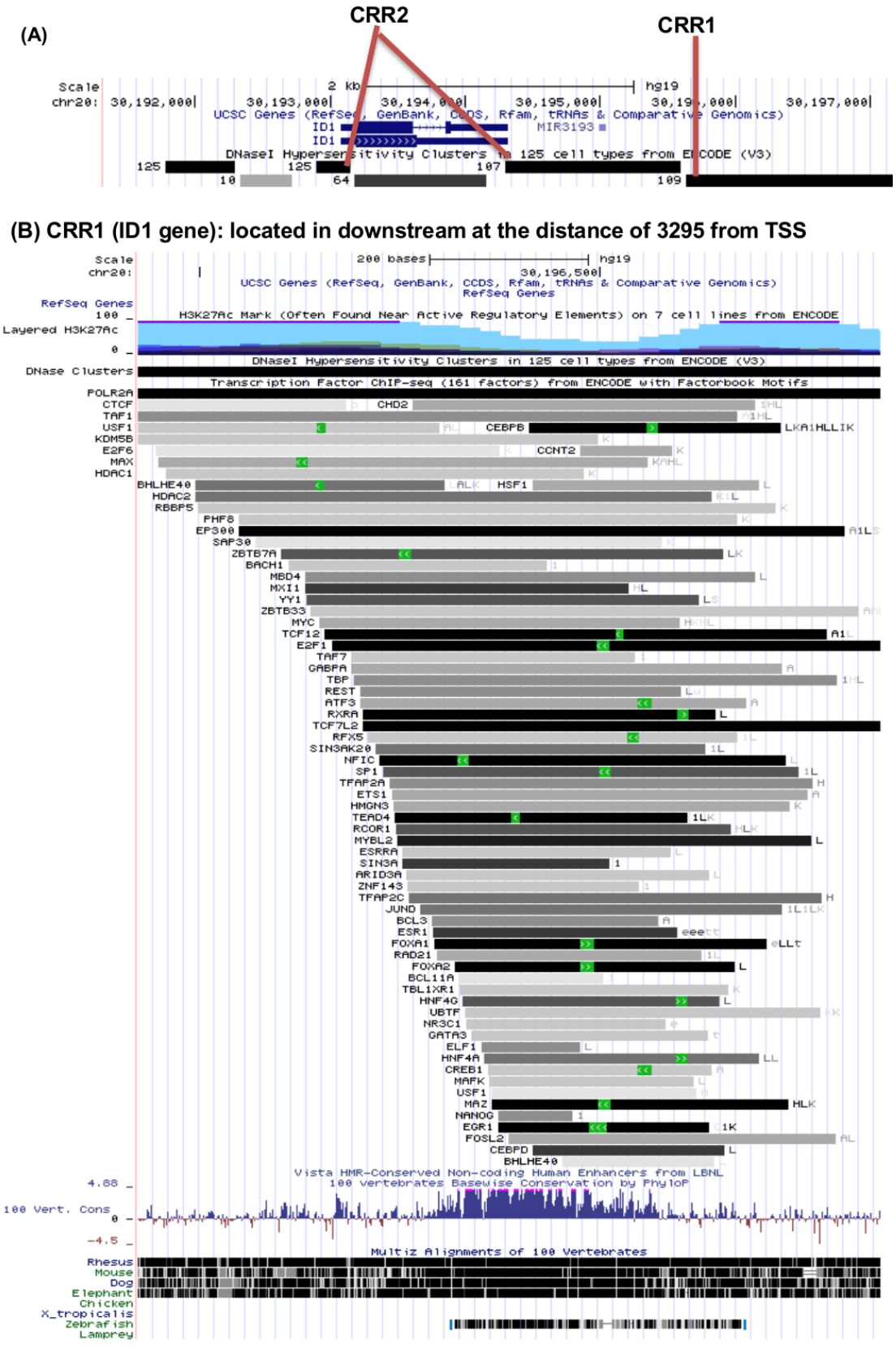


Figure 5. 13: (A) shows the location of chosen CRRs, mentioned as CRR1 and CRR2. Figure (B) represents chosen CRR1 and this figure shows H3K27AC signal, DHSs cluster, bound transcription factors, and conservation for CRR1. Location of chosen CRRs with respect to their TSS is also mentioned in the figure (B).

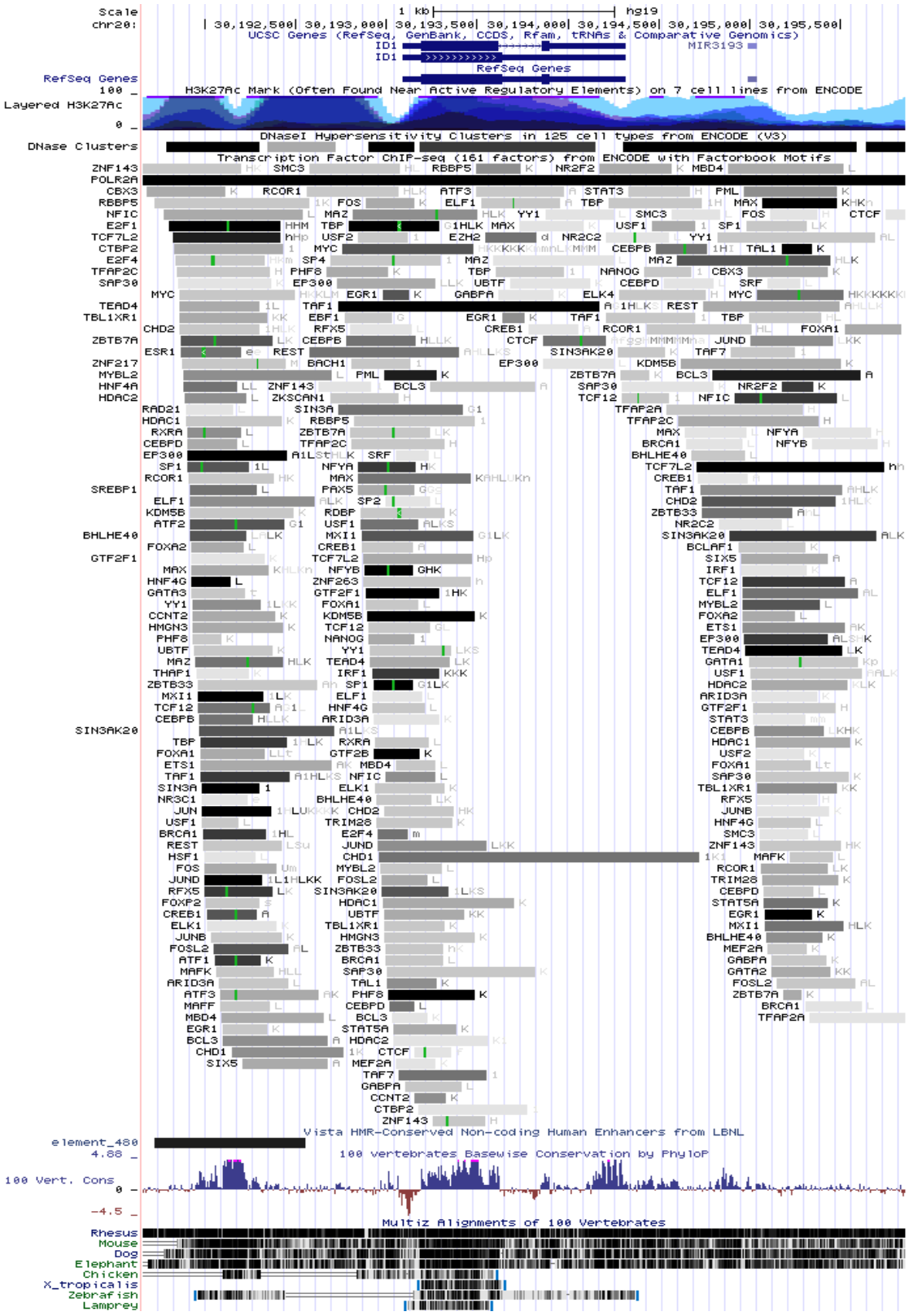


Figure 5.13 (C): This figure contains H3K27AC signal, DHSs clusters, bound TFs, and conservation. A region in the chosen CRR2 is known to regulate WNT5A and known region size is 838 bases, which completely lies within this chosen CRR and can be seen here.

Table 5. 9: This table contains all the CRRs mapped to the ID1 transcript (ENST00000376105.3) within 100kb. Two CRRs (highlighted with red colour) were chosen by LASSO.

Chr #	Start	End	Chr #	Start	End
chr20	30154621	30155285	chr20	30255165	30255440
chr20	30156001	30157732	chr20	30256257	30256500
chr20	30158625	30159259	chr20	30256776	30257051
chr20	30159280	30159430	chr20	30257617	30258136
chr20	30160137	30162027	chr20	30258429	30258883
chr20	30162043	30162574	chr20	30259832	30260513
chr20	30165620	30166145	chr20	30261594	30262342
chr20	30166226	30166449	chr20	30262520	30264543
chr20	30167803	30168078	chr20	30266189	30266818
chr20	30168543	30168905	chr20	30266973	30267242
chr20	30169207	30169635	chr20	30267268	30267557
chr20	30171949	30172278	chr20	30267613	30268600
chr20	30173474	30173864	chr20	30272430	30272816
chr20	30175248	30176967	chr20	30273179	30273712
chr20	30177540	30177690	chr20	30277880	30278030
chr20	30178546	30179306	chr20	30279453	30280049
chr20	30180669	30182208	chr20	30281119	30281522
chr20	30182490	30184554	chr20	30281665	30282432
chr20	30184907	30185308	chr20	30282518	30283067
chr20	30190465	30191151	chr20	30283765	30284448
chr20	30195923	30196852	chr20	30284862	30285405
chr20	30198296	30198992	chr20	30286480	30287531
chr20	30199250	30199983	chr20	30287794	30288449
chr20	30200482	30201144	chr20	30292092	30293126
chr20	30202260	30202535	chr20	30129957	30130472
chr20	30205369	30205512	chr20	30191655	30195860

LIMS1 (ENST00000480744.1) is the example, where one of the CRR is chosen by both methods (LASSO and filtered LASSO. Gene regulation of this gene and chosen CRRs along with multiple biological features are illustrated in the Figure 5.14. This is the unique example because CTCF has binding site in the chosen CRR1 as shown in Figure 5.14 (B), and CTCF can work as an insulator. However, CTCF binds on that region in some of the cell types, not all. Large number of TFs bind on the CRR2, so this region is possibly chosen by both methods. Table 5.10 contains list of CRRs mapped to the LIMS1 transcript.

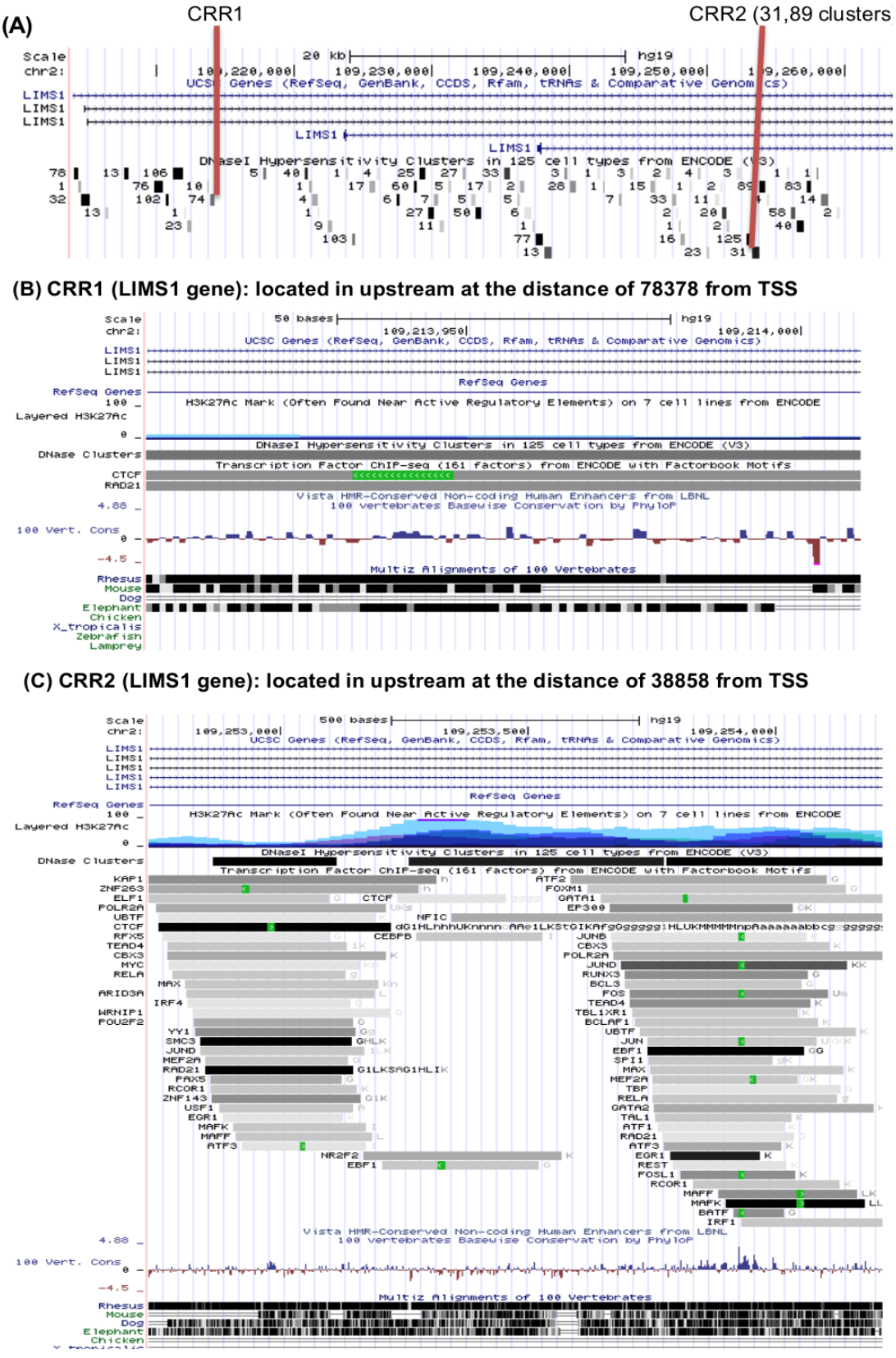


Figure 5. 14: (A) shows the location of chosen CRRs, mentioned as CRR1 and CRR2. B and C represent chosen CRR1 and CRR2 respectively. Both these figure panels (B & C) show H3K27AC signal, DHSs cluster, bound transcription factors, and conservation for CRR1 and CRR2. Location of chosen CRRs with respect to their TSS is also mentioned in their figure panels. Both chosen CRRs are not conserved and CRR1 does not have bound TFs except RAD21. However, CTCF has binding site in the CRR1 region.

Table 5. 10: This table contains all the CRRs mapped to the LIMS1 transcript (ENST00000480744.1) within 100kb. Two CRRs (highlighted with red colour) were chosen by LASSO.

Chr #	Start	End	Chr #	Start	End
chr2	109191426	109192419	chr2	109238194	109238764
chr2	109194851	109195140	chr2	109238806	109239081
chr2	109195260	109195536	chr2	109240062	109240337
chr2	109196073	109196336	chr2	109241759	109242142
chr2	109196539	109196814	chr2	109245713	109246381
chr2	109197322	109198032	chr2	109248727	109249070
chr2	109200372	109200853	chr2	109249185	109249474
chr2	109201920	109202150	chr2	109251100	109251250
chr2	109204140	109204290	chr2	109252735	109254216
chr2	109204470	109205341	chr2	109256140	109256290
chr2	109207494	109208161	chr2	109256560	109256710
chr2	109209805	109209955	chr2	109257240	109257450
chr2	109210060	109210290	chr2	109258274	109258553
chr2	109210546	109212127	chr2	109268760	109268910
chr2	109213903	109214009	chr2	109268938	109269611
chr2	109220675	109220938	chr2	109278542	109279137
chr2	109224121	109224640	chr2	109335372	109336463
chr2	109225325	109225920	chr2	109391721	109391960
chr2	109227242	109227878	chr2	109392054	109392329
chr2	109228482	109229608	chr2	109231726	109232526
chr2	109229951	109230464	chr2	109247855	109248668
chr2	109236887	109238154			

There are thousands of models where both methods predicted different CRRs for each gene, for an example TEAD3 (ENST00000338863.7). Graphical representation of TEAD3 regulation is illustrated in Figure 5.15. Only RFX5 binds on the CRR2, and both chosen CRRs are not conserved except their small regions as shown in Figure 5.15 (B&C). Table 5.11 contains list of CRRs mapped to the TEAD3 transcript within 100kb distance.

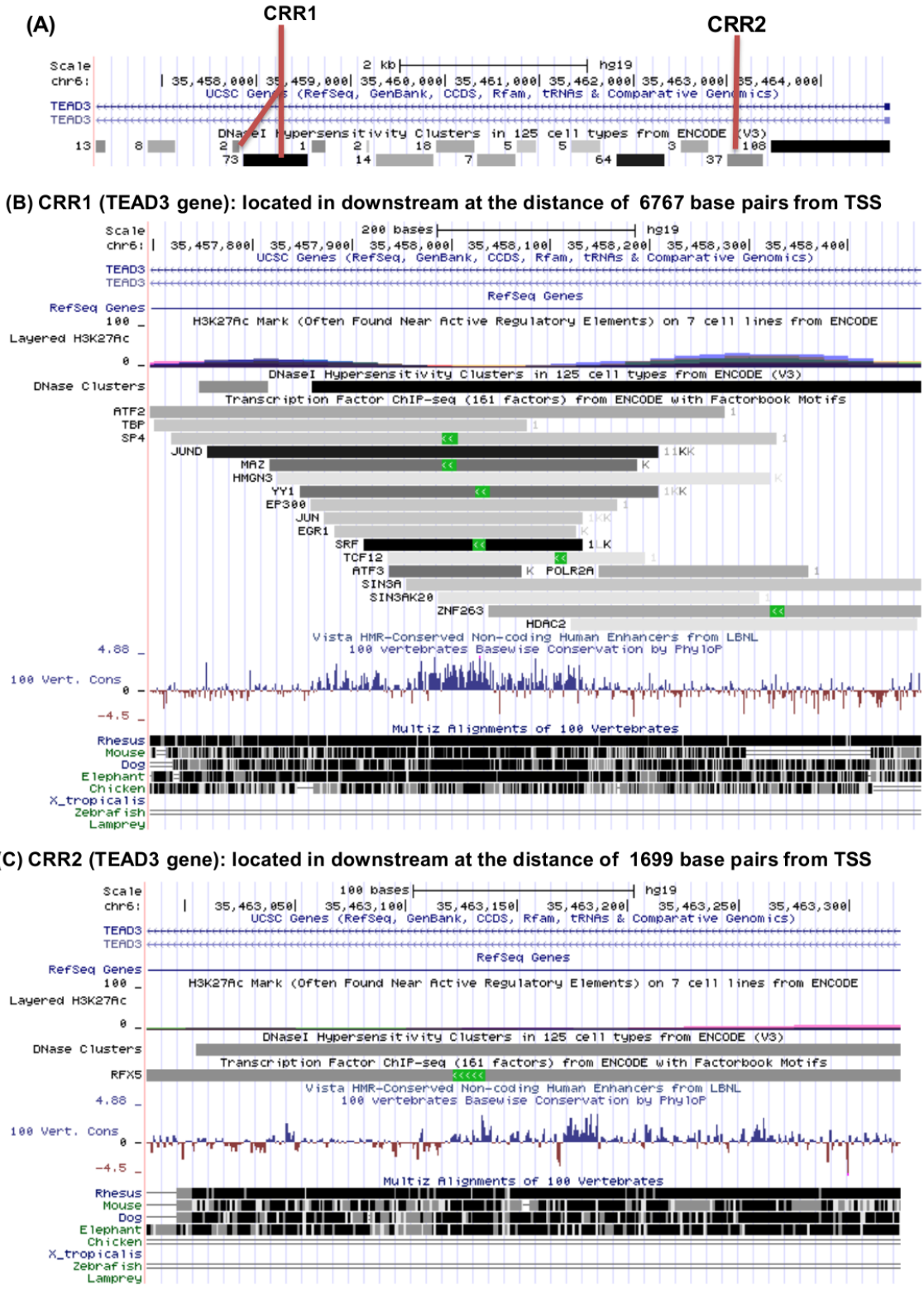


Figure 5. 15: (A) shows the location of chosen CRRs, mentioned as CRR1 and CRR2. B and C represent chosen CRR1 and CRR2 respectively. Both these figure panels (B & C) show H3K27AC signal, DHSs cluster (CRR1 has two DHSs clusters), bound transcription factors, and conservation for CRR1 and CRR2. Location of chosen CRRs with respect to their TSS is also mentioned in their figure panels. Only small regions in the CRR1 and CRR2 are conserved and CRR2 has only one bound TF.

Table 5. 11: This table contains all the CRRs mapped to the TEAD3 transcript (ENST00000338863.7) within 100kb. Two CRRs (highlighted with red colour) were chosen by LASSO.

Chr #	Start	End	Chr #	Start	End
chr6	35364880	35365030	chr6	35453129	35454615
chr6	35365366	35365921	chr6	35455438	35456111
chr6	35368780	35368990	chr6	35457697	35458474
chr6	35369799	35370784	chr6	35459223	35460004
chr6	35378836	35379028	chr6	35460752	35461037
chr6	35379985	35380135	chr6	35461144	35461772
chr6	35382902	35383509	chr6	35461945	35462215
chr6	35387288	35387677	chr6	35462984	35463323
chr6	35393061	35393402	chr6	35463485	35466635
chr6	35393957	35394160	chr6	35467381	35468158
chr6	35395880	35396070	chr6	35468562	35469038
chr6	35396145	35396850	chr6	35472089	35472378
chr6	35396967	35398053	chr6	35474263	35474702
chr6	35419443	35420881	chr6	35486791	35487086
chr6	35420985	35421472	chr6	35490267	35491164
chr6	35421667	35422062	chr6	35502220	35502370
chr6	35429140	35429290	chr6	35534091	35535128
chr6	35435502	35437866	chr6	35564286	35565356
chr6	35439526	35440428	chr6	35438132	35439364
chr6	35444841	35445180	chr6	35487335	35487838
chr6	35445447	35445602	chr6	35536679	35537198
chr6	35452691	35452966			

There are also examples where regulatory regions are known but or method has identified different regulatory regions, for an example, TCF4 has known regulatory region in Vista Enhancer [145], but our methods have chosen different CRRs. This is because, known regulatory region activity was observed in cells that we have not considered in our dataset.

5.3.4 Discussion

Gene expression is controlled by the cis regulatory regions such as enhancers and repressors. Here, we have identified regulatory regions by integrating ChIP-seq, DNase-seq and RNA-seq data as discussed in the method section. We started this chapter by obtaining Candidate cis Regulatory Regions (CRRs) from ChIP-seq and DNase-seq separately, and end up in combining both ChIP-seq and DNase-seq data sets together and then obtaining CRRs from them. Later on, we optimised method by limiting the choice of CRRs, not using fold change, and setting a mapping threshold of 100kb [149] (mapping of transcripts to potential regulators).

Two different strategies were adapted for LASSO followed by randomisation to estimate the false discovery rate. Significance of the models increases if both methods have chosen same CRRs/CRR for a particular transcript .These two different methods/strategies were used to study whether we can identify regulators by giving filtered input to the LASSO, else we allow LASSO to choose those CRRs which are predictive of gene expression from all CRRs mapped to the transcript within 100kb. CRRs were filtered on the basis of highest ratio of TF binding and highest conservation score. As we have discussed in the 1st chapter that TFs have important role in regulation of genes [150] and functional regions i.e., regulatory regions are less susceptible to change (remain conserved). However, non-conserved regions can also be functional regions [151]. Therefore, we also adapted strategy, where we gave input to the LASSO without filtering. Same CRRs were chosen by LASSO in both methods for large number of models, which suggest that several potential regulators regions were predictive of gene, expression, conserved and bound by several TFs. It increases the significance of our methods, as it is not possible for different methods to choose same CRRs from a large number of candidates.

We have discussed in the results section that we have limited number of significant models after correcting (multiple testing correction) p values that are generated after randomisation. Therefore, we concluded that models can be built for individual genes as genome wide model building lead to the large false

discovery rate. We can choose models according to biological features, for example log (coefficient of variance of gene expression) and different thresholds were considered. Models with high (top 10%) log (coefficient of variance of gene expression) were built, and we successfully built 74 significant models from this set.

We can assess the significance of our chosen CRRs by testing with some independent cancer data, knowing the fact that most of the cell types were considered for obtaining CRRs are cancer cell types. Recently published work suggested that cis regulatory regions contain significant number of somatic mutations and mutations in TF binding sites can drive different abnormalities including cancer [152].

Vista Enhancer database [145], contains 1790 enhancers for 1537 genes and 397 of these genes are in our set of LASSO models, which is less than quarter of total genes in this database. This shows that vista enhancers have identified enhancers for a specific set of genes and our method has predicted experimentally known enhancers for only 21 genes, which is a small number. It is possibly because, our method has identified ubiquitous regulatory regions and Vista Enhancer database contains mostly cell type specific enhancers. They have observed expression patterns for known enhancers in those cells which we have not considered for model building. We looked in experiments for known vista enhancers, and concluded that researchers have observed expression levels in neural tube, hindbrain, limb, cranial nerve, midbrain, forebrain, nose, tail, mesenchyme derived from neural crest, and heart. However, cell types considered for model building are mostly cancer cell types, and only SK-N-SH-RA (sknshra) is neuroblastoma cell line. Therefore, we decided to assess significance of predicted regulatory regions by independent cancer mutation dataset, and that is discussed in the next chapter.

Chapter 6

6 Mapping cancer somatic mutations to regulatory regions

6.1 Introduction

We have developed methods for prediction of cis regulatory regions in the previous chapters, but here we are pursuing with the models built by LASSO method. We have discussed in the previous chapter that, chosen CRRs are difficult to validate because we don't have enough existing knowledge about regulatory regions and it was also observed that regulatory regions (chosen CRRs) are prone to high false discovery rate. Therefore, we considered whether a completely independent dataset, known cancer somatic mutations could reveal further significance of these chosen CRRs. We test whether our chosen CRRs accumulate significant number of cancer somatic mutations than CRRs rejected by our method. Studying the relationship between mutational frequencies and variables such as replication timing and gene expression have allowed the identification of recurrently mutated regions and mutated protein and they are likely to be oncogenic drivers [153].

Mutations in regulatory regions could lead to the aberrant regulatory process, and this has a role in initiation and progression of cancer, such as constituent activation of transcription factors regulated by chromosomal re-arrangements [154].

Mutations in regulatory regions can lead to abnormal gene expression levels, which ultimately can result in uncontrolled cell growth that is one of the signs of cancer. Similarly, researchers have found in large percentages of cases in some cancer types that point mutations in TERT gene promoter are strongly linked to gene expression changes [155]. Developments in the field of whole genome sequencing, made possible for scientists to focus on mutations that occur in potential regulatory elements within the genome. Recently, it has been discovered that regulatory regions accumulate large numbers of the somatic mutations that have been observed in cancer cell genomes [156]. Mutations in regulatory regions are may be important in promoting survival and reproduction

of cancer cells, which is an evidence of positive selection for mutations in these regions

Weinhold et al., [157] have also observed recurrently mutated regulatory elements and discovered regions potentially regulating genes with known involvement in cancer. Fredriksson et al., [158] have developed method for the discovery of mutations that are strongly linked to expression levels of nearby genes in cancer samples. However, Identification of regulatory mutations driving cancer is a difficult process and existing methods have discovered only fraction of such mutations. Therefore, there is a clear need for new methods that would help us to understand the irregularities in gene regulation.

Understanding regulation of genes is challenging, as several factors influence the expression of genes, and each factor has its own dimensions. Large number of regulatory regions have been identified by recently discovered technologies and it was difficult to link them to their respective genes. However, we have successfully identified potential regulatory regions and link them to their respective genes by using correlative models, as discussed in the 5th chapter. We have identified regulatory regions mostly from cancer cell types and it was assumed that these regions may accumulate significant number of cancer somatic mutations. We have also compared the frequency of mutations in chosen regulatory regions for genes known to involve in cancer with the frequency of mutations in chosen regulatory regions of those genes which are not yet known to involve in cancer.

Here, we have mapped cancer somatic mutations on chosen and rejected CRRs that are identified by our method explained in the Chapter 5 and we investigated that whether chosen CRRs harbour significant number of mutations than rejected CRRs.

6.2 Methods

6.2.1 Regulatory regions

Here, we have considered chosen and rejected CRRs from LASSO method, which is explained in the 5th chapter. LASSO method based on correlative models chooses two CRRs (Candidate cis Regulatory Regions) that are predictive of gene expression and rejected all other CRRs mapped to the transcript within 100KB distance. R code for building LASSO models is given in **Appendix II**, this file as an additional file 4 is also available here:

https://static-content.springer.com/esm/art%3A10.1186%2Fs12943-016-0560-0/MediaObjects/12943_2016_560_MOESM4_ESM.r

6.2.2 Mapping cancer mutations to regulatory regions

Somatic cancer mutations were obtained from the COSMIC database V.76 [159] (Catalogue of somatic mutations in cancer). Approximately 2.3 million cancer somatic mutations were retrieved and mapped to the chosen and rejected CRRs. Duplicate mutations (mutations occurring at the same genomic location) were eliminated.

6.2.3 Statistical significance of differences in mutation counts

The statistical significances of differences in the counts of somatic mutations observed in chosen and rejected CRRs were tested in several ways. Differences in the average number of mutations per CRR were tested with two-sample t-tests, and also equivalent non-parametric Wilcoxon tests to account for possible non-normality. To account for other possible effects that might bias these considerations we also repeated these tests after first balancing the chosen and rejected sets to have the same distribution of any potential confounding variable. This was achieved by sampling the rejected set of CRRs randomly to match the distribution of a variable in the chosen set, which was enabled by the significantly larger size of the rejected set. The variables considered were replication timing, base pair composition, length of the CRRs and distance of the CRR to the

transcription start site (TSS). Replication timing and GC content data was downloaded from the UCSC website: the wavelet-smoothed signal of replication timing [160] for 9 cell types was obtained and we used the average signal. In each case data was binned in 4 equal bins and the process required a chosen CRR to be matched by a rejected CRR from the same bin. These all variables considered here may influence the CRRs choice by LASSO method and also can influence frequency of mutations in these regions. Replication timing is the order in which parts of genome are duplicated. It has been known that replication timing can influence the rate of mutation[153], and replication timing can also be influenced by the base pair composition [161]. Some of the base pairs are more prone to mutations and some are not, similarly proximal regions accumulate high frequency of mutations than distal and large CRRs can harbour large percentages of mutations. Therefore, we investigated that whether the chosen CRRs are accumulating significant number of somatic mutations considering these variables in mind.

As an alternative test of statistical significance which enabled us to model all potential effects on mutation counts together, we built generalised linear models using the glm function in R. Generalized linear model is a set of independent variables each with a distribution from the exponential family, and it has response variables, which are expected to share the same distribution from the exponential family, and it has a monotone link function [162].

We build the generalized linear model (glm), as explained below

Response or dependent variable (y) = Mutation counts

We assumed Poisson distribution where,

$$\log(E(y)) = \sum_i k_i x_i$$

Where independent variables are,

x1= distance of CRR from TSS

x2= Replication timing

x3=GC content

x4= Length of CRR

x5= Indicator variable for chosen/rejected CRRs

A log link function was used, first under the assumption of a Poisson distribution for the counts and then in cases of over-dispersion using the quasipoisson option in glm, which fits a dispersion parameter which is otherwise fixed at unity. The statistical significances of the effects of each variable were assessed from the standard Wald test statistics produced by glm.

6.2.4 Cancer census genes

It has been a central purpose of cancer research to identify genes involved in cancer. Futreal et al. [163], started compiling genes involved in cancer from literature. They included genes in the set of human cancer census genes if there are existing at least two independent reports showing mutations in primary patient material. As we have discussed previously that aberrant changes in regulation of genes cause cancer, and these changes in regulation are may be caused by mutations in regulatory regions. Therefore, we investigated the frequency of mutations in chosen CRRs associated with the cancer census genes.

A set of 533 cancer consensus genes were retrieved from the COSMIC database of which 304 entered our analysis (the others did not meet our modelling criterion of expressing in at least 7 cell types). These were analysed as a separate subset to investigate any possible specific effects for genes known to be directly involved in cancer.

6.3 Results

We investigated the possibility of a large-scale model building exercise for all genes/transcripts as we discussed in Chapter 5, and also in a restricted set of 533 genes known to be cancer associated [163]. Here, we have considered LASSO method chosen regulatory regions set, and most of the genes considered for model building are globally expressed genes as explained in the previous chapter. The relevant statistics of model building are shown in Table 6.1 (This table contains No. of CRRs without repetition i.e., 25025 CRRs). Models were successfully built for approximately 9000 genes (16000 transcripts) (This model building is discussed in detail in ordinary LASSO section), and 292 genes (650

transcripts) from the cancer set. It shows model building failed only for 12 cancer census genes. List of chosen CRRs for 9000 genes that also includes the cancer census genes along with the count of somatic mutations is given in additional file 1 (given in CD). Additional file 1 is also provided with our published paper [164] on Molecular cancer journal website:

https://static-content.springer.com/esm/art%3A10.1186%2Fs12943-016-0560-0/MediaObjects/12943_2016_560_MOESM1_ESM.xlsx

We have included the copy of published paper in **Appendix III**. We have discussed this in the previous chapter but just to remind that this scale of model building exercise leads to a significant false discovery rate. Therefore, individual models should be studied carefully. We built models genome wide for the identification of single set of chosen CRRs that are predictive of gene expression, covering substantial part of open chromatin region, and a complement set of CRRs with weaker relationships to gene expression. It should be noted that some of the CRRs were chosen for more than one transcript in both sets (All transcripts and cancer set transcripts), as detailed in the Table 6.1. Figure 6.1 illustrates three genes and their chosen CRRs, one of the CRR accumulating 4 COSMIC mutations was chosen as a potential regulator of NAB2 and STAT6, and both these genes are known to be involved in cancer. The transcription factors that bind on these chosen CRRs and number of somatic mutations harbour in these chosen CRRs are also mentioned in the figure.

Table 6. 1: Statistics of model building

	All transcripts	**Cancer set transcripts**
Number of models attempted	17963 transcripts from 9209 genes	731 transcripts (from 304 genes)
Number of models built	16134 (8670 genes)	654 (292 genes)
Average r,r^2	0.710, 0.519	0.718, 0.530
Range r^2	0.004-0.99	0.048-0.925
Total candidate elements	678020 (mean 42/transcript)	28844 (mean 44/transcript)
Chosen elements	25045 (2/transcript)	1140 (2/transcript)
Elements chosen for 1 transcript	20025	999
Elements chosen for >1 transcript	5020	141

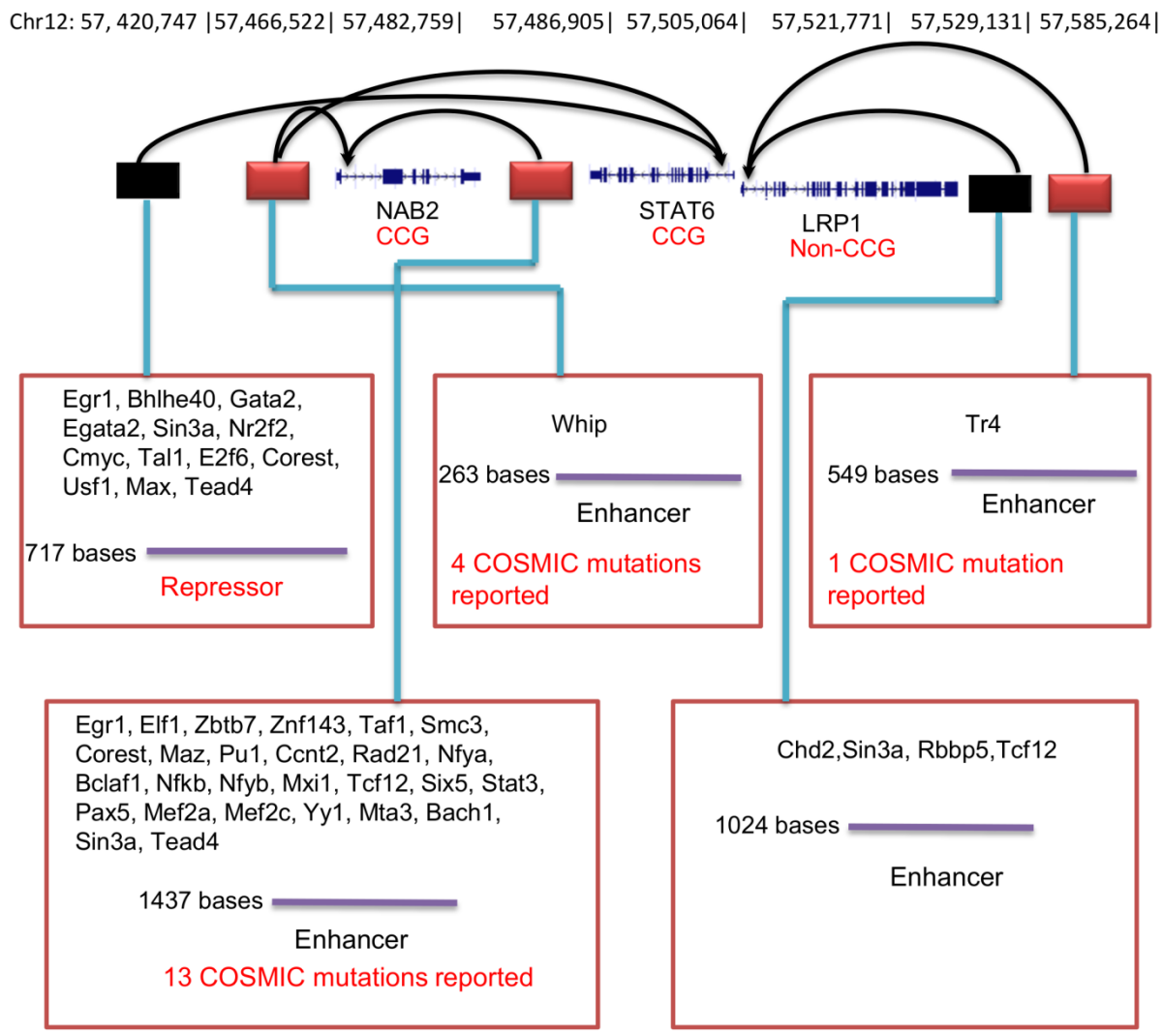


Figure 6. 1: The chosen CRRs for NAB2 (ENST00000342556.5), STAT6 (ENST00000300134.2) and LRP1 (ENST00000243077.2). Black arrows link CRRs to the transcripts for which they were chosen in expression models; note that one CRR was chosen for both STAT6 and NAB2. Details of the chosen CRRs are given red boxes, including the bound transcription factors, sizes of the CRRs and mutations mapped from the COSMIC database. CRRs are labelled as enhancers if they show positive correlation with expression and repressors if they show negative correlation. The chosen CRRs are marked as red boxes if there is at least one reported mutation in them, and black otherwise.

Overall 8% of these COSMIC mutations mapped to CRRs identified with the transcript set defined above, this percentage is small because we have restricted number of transcripts by considering only globally expressed transcripts.

Table 6.2 gives statistics showing the mutations mapped in chosen and rejected CRRs from the modelling exercise. These results show that a significant higher proportion of chosen CRRs are mutated at least once than the rejected CRRs, and that chosen CRRs harbour around 1.5 times more mutations than rejected CRRs. We repeated this exercise for cancer set and chosen CRRs have significant higher proportion of mutations than rejected CRRs. Proportion of somatic mutation in chosen CRRs is significantly higher than the proportion of mutations in rejected CRRs in most of the sets but smaller cancer genes set shows the same trends but with reduced levels of statistical significance. When limiting the analysis to CRRs only from higher quality expression models (confident models with r>0.7 and highly confident models with r>0.8) the effect size increases: mutations are enriched in chosen CRRs by a factor of 1.45 (=1.35/0.93) in all models and this rises to 1.81 in CRRs for highly confident models.

These results suggest that chosen regulatory regions could be functionally important for their respective linked genes, as they accumulate significant number of mutations. These mutations may have some influence on the gene expression levels as they are predicted from correlative models.

Table 6. 2: Mapping of somatic mutations from COSMIC to candidate regulatory regions (CRRs)

Title	All		Cancer census genes	
	Chosen CRRs	**Rejected CRRs**	**Chosen CRRs**	**Rejected CRRs**
Total number of CRRs	25045	158560	1140	7429
CRRs mutated at least once	3535 (14.11%)[1]	16241(10.24%)[1]	160 (14.03%)[2]	703 (9.46%)[2]
Mean mutations/CRR	1.35[3]	0.93[3]	1.51[4]	0.97[4]
Mean mutations/CRR (models with r>0.7)	1.40[3]	0.88[3]	1.63[5]	0.95[5]
Mean mutations/CRR (models with r>0.8)	1.50[3]	0.83[3]	1.55	0.92

[1]Proportion mutated in chosen set greater than in rejected set, $p<10^{-15}$ (Chi-squared and Fisher test)
[2]Proportion mutated in chosen set greater than in rejected set, $p<10^{-5}$ (Chi-squared and Fisher test)
[3]Mean mutations in chosen set greater than in rejected set, $p<10^{-23}$ (two sample t test), $p<10^{-8}$ (Wilcoxon test)
[4]Mean mutations in chosen set greater than in rejected set, $p<0.05$ (two sample t test and Wilcoxon test)
[5]Mean mutations in chosen set greater than in rejected set, $p<0.05$ (two sample t test), $p=0.06$ (Wilcoxon test).

It is known that DNA mutation frequencies are heterogeneous [153] over the genome, and are related to variables such as replication timing and GC content. In this analysis, average mutation frequencies within CRRs might be influenced by the length of the CRR and possibly the proximity to a transcription start site (TSS). We took two different approaches to investigate whether these effects could have biased the statistical considerations above. First we repeated the significance tests on the mean number of mutations per CRR, this time not using the entire set of rejected CRRs but by randomly choosing a set of equal size to the chosen set matched according to the variable concerned (e.g. matching each chosen set member with a rejected member falling in the same GC content bin). In the case of all variables (replication timing, GC content, length of CRR and proximity to a TSS) the effects reported above remained significant in all models set, but there are some cases of insignificance in cancer census genes set. Statistics (p values) of this analysis are shown in Table 6.3.

Table 6. 3: This table shows the statistical comparison (p values calculated from independent t-test) of frequency of mutations in chosen CRRs and rejected CRRs that match in replication timing, GC content, length of the CRRs and distance from TSS. Here CCGs is the Cancer Census Genes from cancer set and All is for all models.

	Replication timing		**GC content**		**Length of CRRs**		**Distance from TSS**	
	All	CCGs	All	CCGs	All	CCGs	All	CCGs
Mean mutations/CRR	1.5e-14	0.0028	2.9e-15	0.0005	1.8e-10	0.20	1.6e-36	0.017
Mean mutations/CRR (models with r>0.7)	2.6e-09	0.292	2.5e-17	0.379	3.6e-06	0.19	1.7e-23	0.0629
Mean mutations /CRR (models with r>0.8)	0.00085	0.025	1.594e-08	0.825	7.5e-07	0.77	7.7e-13	0.023

Second approach is to model all these potential effects simultaneously we built generalised linear models for the counts of mutations in CRRs, and the result showing correlation of frequency of mutations with all five variables are detailed in the Table 6.4. We found the counts to be over-dispersed with respect to a

Poisson distribution assumption, and modelled this with an additional dispersion parameter. The effect size for an indicator variable showing whether a CRR was chosen or rejected was 0.46+/-0.02 ($p<2x10^{-16}$,Wald test), revealing a highly significant effect on the (log) expected mutation counts consistent in size with observed differences in average mutation counts from Table 6.2.

Table 6. 4: This table shows the result of generalized linear model built by correlating frequency of mutations with five variables, only replication timing is not correlated.

Variables	**Estimate ± Std. error**	**P value**
Distance	$-(8.4 \pm 0.3) \times 10^{-6}$	<2e-16
Replication timing	$-(1.31 \pm 9.26) \times 10^{-4}$	0.887
GC content	$(7.4 \pm 0.095) \times 10^{-2}$	<2e-16
Length of the CRR	$(4.98 \pm 0.045) \times 10^{-4}$	<2e-16
Chosen CRR=1, Rejected CRR=0	$(4.57 \pm 0.2) \times 10^{-1}$	<2e-16

Chosen CRRs may be positively (enhancers) or negatively (repressors) correlated with the expression of the associated gene, and can be involved in the enhancing and repressing the gene expression levels. From our chosen CRRs, 32% showed negative correlations, and they can be called as repressors. We tested the difference in the average number of mutations between the enhancers and repressors, but there was no significant difference in rate of mutations between these two types of CRRs. This has been observed in all models/transcripts as well as in cancer set models, as mentioned in Table 6.5.

Table 6. 5: This table shows the average number of mutations per enhancer and repressors in all models as well as in cancer set models.

	All models		**Cancer census genes models**	
	Enhancers	Repressors	Enhancers	Repressors
Mean mutations	1.33	1.399	1.60	1.262
Two sample t test	p value =0.385		p value=0.44	

In other analysis, we divided CRRs into proximal and distal according to distance from the associated transcription start site (distal > 10kBases, proximal < 10kBases), and tested the difference in average number of mutations between these regions. We found that proximal CRRs have significantly higher tendency

to be mutated than the distal CRRs, as detailed in the Table 6.6. This difference in average number of mutations is more pronounced in CRRs identified with cancer associated genes.

Table 6. 6: Mapping of mutations to chosen CRRs proximal and distal to the transcription start site

	Proximal (<10kB from TSS)	**Distal (>10kB from TSS)**
Mean mutations/CRR (all models)	2.30[1]	1.25[1]
Mean mutations/CRR (cancer related	3.40[2]	0.99[2]

[1]Mean greater in proximal set, $p<10^{-39}$ (t-test), $p<10^{-17}$ (Wilcoxon)
[1]Mean greater in proximal set, $p<10^{-6}$ (t-test), $p<0.05$ (Wilcoxon)

6.4 Discussion

It has been reported by several researchers that somatic mutations in regulatory regions have important role in dis-regulation of genes [157] [165], as we have discussed in the results section, this dis-regulation ultimately leads to cancer, and those mutations which cause cancer or any abnormality are known as driver mutations. We have used the several cancer cell types in model building. Therefore, it was worthy to look for cancer somatic mutations in chosen CRRs and rejected CRRs, and we found that chosen CRRs accumulate significant number of somatic mutations than the rejected CRRs.

The work carried out in this chapter helped us to understand which somatic mutations in cancer cells drive the process of cancer progression and to identify underlying mechanisms of gene regulation.

Mutations in regulatory regions are an important feature of cancer, and the results reported here show that a set of candidate regulatory regions derived from simple correlative models preferentially harbour cancer somatic mutations compare to regions rejected by models. This suggest that chosen regions could be of functional significance in genetic regulation.

It is now known that mutations affecting regulatory regions are potentially as important in cancer progression as mutations in protein coding regions or those that directly alter functional RNA molecules. Further, we identified positive selection of mutations in these chosen regions and existing research now suggest that this positive selection of mutations might involve in cancer progression. The work reported here strongly suggests that modelling based on large data compendia like ENCODE can identify genomic regions which are potentially more strongly linked to gene expression, and propose links to the regulated genes. This could lead to more effective definition and prioritisation of mechanistic hypotheses for cancer somatic mutations, which will be accessible to confirmation or refutation with further detailed laboratory investigations.

Here, we investigated that whether different variables such as replication timing, GC content, length of the CRR, and distance of the CRR to the TSS (transcription start site) effect the rate of mutation. Existing studies suggested that replication

timing is correlated with the rate of mutation [166]. There are also existing studies which suggest that both variables: replication timing and base pair composition are correlated with the rate of mutation [153]. However, we found that replication timing is not correlated with rate of mutations in our set of potential regulatory regions, though other variables behave differently as shown in Table 6.5. Here, replication timing behaves differently than existing research possibly because we have limited set of regulatory regions that are associated with the globally expressed genes. In addition to that, we also considered all these four variables in comparing rate of mutation in chosen CRRs and rejected CRRs. We picked CRRs randomly from rejected CRRs set matched with chosen CRRs replication timing, GC content, length of the CRRs, and distance of the CRRs to the TSS. We found that still chosen CRRs accumulate significant high rate of somatic mutations than the rejected CRRs.

In our study, we also identified that proximal regions have significantly high rate of mutations than the distal regions [157]. We also analysed the rate of somatic mutations in the enhancers (positively correlated) and repressors (negatively correlated), but we didn't find any significant difference in rate of mutations between these regions.

We investigated that whether the chosen CRRs mapped to the cancer census genes behave differently than chosen CRRs mapped to all genes. We found that both set of CRRs behave similarly, and chosen CRRs accumulate significantly higher frequency of mutations than rejected CRRs in both sets. We tested the difference of mean mutations/CRR between chosen and rejected CRRs by using two different tests (Wilcoxon test and two sample t test), just to increase the accuracy of the results.

Chapter 7

7 Discussion and future work

Our research starts with the development of method for prediction of TF-TF mutual interactions from ENCODE ChIP-seq data. The ENCODE has mapped approximately 119 transcription factors out of 1800 transcription factors. It was challenging to develop any statistical method to predict the interactions as binding sites for a limited number of transcription factors have been mapped but we manage to develop method based on two statistical methods i.e., Poisson distribution and Randomisation, where Mu of randomisation and Lambda of Poisson distribution are similar. It was challenging to optimise the size of accessible genome, we considered accessible genome size by unifying binding sites from all ENCODE transcription factors, and this was leading to the large number of significant overlaps, suggesting that size of accessible genome should be optimised. Finally, we manage to optimize the size of accessible genome by considering certain percentage of accessible genome. Some of interactions predicted by our methods are already known, which signifies the accuracy of our method. This method can also be applied to the new generated data and it can help in understanding that how transcription factors form complex to regulate the expression of genes.

We also compared our method results from Gm12878 and K562 cell types with the results from ENCODE (K562) [34]. We found that, there are several common significantly overlapping TFs in our method and ENCODE method, even in different cell type (Gm12878).

We also studied the conservation of TF binding sites by asking different questions and we found that shared TF binding sites in multiple cell types for a particular TF are more significantly conserved than cell type specific binding sites of that TF, which supports the argument that functional binding sites are shared in multiple cell types and they are conserved [167].

In other analysis, we found that shared binding sites between TFs in a particular cell type are more significantly conserved than the non-shared (unique) binding

sites, this suggests us about the interaction of TFs through these shared and significantly conserved binding sites. We also looked for the distribution of conservation in shared and unique binding sites in a TF pair, six different examples from three cell types were analysed. We observed that distribution of conservation is bimodal in shared and unique binding sites set.

Transcription factor binding near the genes might influence the expression level of genes, here we identified that most of the genes mapped near the co-binding sites of TFs have high level of gene expression than those genes which are mapped to the single TF binding site. This tells us that co-binding sites of TFs enhance the expression in most of the cases but in some cases these sites can act as repressors of transcription as detailed in the Chapter 3.

In other study, we integrated ChIP-seq, DNase-seq and RNA-seq data for prediction of cis regulatory regions. In the beginning, ordinary linear regression was used to predict the cis regulatory regions and associate them with their target genes. We studied different factors which can be helpful in choosing CRRs. Two factors i.e., highest TF binding and highest conservation score were helpful and well correlated models were built from CRRs chosen by these factors. In the 4th chapter, we performed gene ontology enrichment analysis using DAVID, where we picked shared and unique set of well correlated models between three methods. We found that none of gene set was significantly involved in any biological process. However, most of the gene sets were significantly involved in protein binding and poly (A) RNA binding molecular function.

Later on (5th chapter), we used the LASSO instead of ordinary linear regression as LASSO can penalize the variables and select those CRRs which are predictive of gene expression. Previous studies and our analysis showed that regulatory regions are conserved, and have high frequency of TF binding and these factors are helpful in predicting cis regulatory regions [168]. Therefore, we considered biological features such as frequency of TF binding and conservation score to filter the candidate cis regulatory regions as each transcript was mapped to the approximately 42 CRRs. We allow LASSO to filter/penalize CRRs and we optimised it by considering only two best CRRs. We compared the two methods,

1. We filter candidate cis regulatory regions (CRRs) up to 8 according to highest TF binding and highest conservation score (4 CRRs by each factor), because LASSO should choose best two CRRs from a set of biological important CRRs; 2. LASSO was given all the CRRs as an input and it was restricted to choose best two cis regulatory regions. Results from these two methods were compared and both methods choose same cis regulatory elements for particular transcripts in large number of transcripts. Scientists have developed tools for prediction of cis regulatory regions in Drosophila such as jPREdictor [169], where this tool predicts the regulatory regions by taking sequence and motif as an input; but our method is different as explained above.

It has been observed in several studies that regulatory regions contain large number of cancer somatic mutations, as these regulatory regions have important role in regulation of genes. We analyzed that whether our chosen cis regulatory regions contain high frequency of mutations or rejected (rejected by LASSO) cis regulatory regions have high frequency of mutations, and we found that chosen CRRs accumulate significant number of mutations than rejected CRRs.

We also considered the biological parameters which may influence the frequency of mutations/count of mutations such as replication timing, base pair composition (%GC), length of the regulatory region, and distance of regulatory region with the transcription start site (TSS); despite of considering similar levels of these parameters for both sets, we found that chosen cis regulatory regions have significantly high frequency of mutations. We also built linear model to see whether these features are positively influencing the frequency of mutations or negatively and we found that all features significantly influence the frequency of mutation except replication timing. Though, it has been known that replication timing has significant influence on the frequency of mutations as discussed in the 6^{th} chapter. However, our results showed that replication timing does not have negative or positive influence on the frequency of mutations, possibly because we have considered restricted set of transcripts (globally expressed transcripts).

Our this method for prediction of cis regulatory regions can be used for newly generated data and important part of this method is that it has predicted regulatory regions for globally expressed genes, which are important for running the biological system.

Enhancers and promoters interact with each other through looping to control the expression levels of genes. DNA becomes flexible to bend for looping because of histone acetylation occurring at the site of bending. ENCODE has generated looping interactions data by 5C (Chromosome Conformation Capture Carbon Copy) technique [170]. Sean et al., have developed a tool called TargetFinder for predicting looping interactions by training models with the known interactions. They give signal profiles of RNA polymerase II, enrichment of H3K27ac, and depletion of mono-methylation of histone H3 at lysine 4 (H3K4me1) in regions flanking the TSS of interacting promoters [171]. Similarly, Bing et al., separated interacting enhancer-promoter pairs and non-interacting enhancer-promoter pairs by training the random forest classifier with four features i.e., Enhancer and target promoter activity profile correlation (EPC), Transcription factor and target promoter correlation (TPC), Coevolution of enhancer and target promoter (COEV), and Distance constraint between enhancer and target promoter (DIS). They developed a method called IM-PET (Integrated Methods for Predicting Enhancer Targets) [172].

There are insulators for example such as CTCF that can block the interaction between enhancers and promoters. So, if there is CTCF binding site between these two regulatory regions (enhancers and promoters) then these regions would not interact with each other. CTCF is the “CCCTC-binding factor” that binds to 55,000-65,000 sites in mammalian genome [173].
It is possible that these binding sites are located between our chosen CRRs for some of the models, as we have seen in the Chapter 5 that CTCF has binding site in the one of chosen CRR for the LIMS1 gene in some of the cell types, but not in all. Here, we have identified cis-regulatory regions for the globally expressed genes. However, 30-60% CTCF binding sites are cell type specific [174], therefore, they will be less likely to be located between our pair of chosen CRRs.

Few studies already exist where they differentiated promoters and enhancers, and also correlated enhancers with their associated genes, but in a different method than our method. Jason et al., confirmed that RNA polymerase II

(RNAPII) is highly enriched at the strong promoters and weakly enriched at the strong enhancers, therefore, this features differentiates the promoters and enhancers. These researchers also linked genes with their potential enhancers by correlating gene expression levels with the histone modifications (H3K4me1, H3K4me2 and H3K27ac) activity at the enhancer locations [175]. We have also considered only those CRRs which have H3K27ac signal, however, we have correlated DHSs signal intensities at the CRRs with the gene expression levels. Other researcher have used different biological features to identify cis-regulatory modules. Similarly, Ross et al., suggested in a review paper that cis regulatory modules can be identified by considering clusters of transcription factor binding sites motifs, conserved non-coding DNA, and biochemical marks associated with the regulatory regions, and these features can be used collectively or individually to predict the regulatory modules. Their validations results proposed that identifying cis regulatory modules through biochemical marks is more reliable than other features [176].

Here, we have predicted TF-TF mutual interactions, and identified enhancers and repressors of transcription along with their target transcripts that would help to understand the regulation of thousands of genes and to identify therapeutic targets for the cancer and other diseases. In my thesis, I have also discussed known cancer genes and our method also has predicted the regulatory region for large number of cancer genes. These methods can also be applied to the newly generated data.

7.1 Future Work

Experimentalists can use our set of predicted TF-TF interactions and chosen CRRs for experimental validation. Similarly, researchers can use chosen CRRs for interpreting newly identified mutations. Effect of somatic mutations lie in the TF binding sites motifs can be studied further to identify their influence on TF binding, ultimately, these mutations may alter genetic regulation. Transcription factor motifs can also be used to predict the TF binding sites.

Bibliography

1. Cooper, G.M., *The complexity of eukaryotic genomes.* 2000.
2. Kulaeva, O.I., et al., *Distant Activation of Transcription: Mechanisms of Enhancer Action.* Molecular and Cellular Biology, 2012. **32**(24): p. 4892-4897.
3. Blackwood, E.M. and J.T. Kadonaga, *Going the Distance: A Current View of Enhancer Action.* Science, 1998. **281**(5373): p. 60-63.
4. Mercurio, F. and M. Karin, *Transcription factors AP-3 and AP-2 interact with the SV40 enhancer in a mutually exclusive manner.* The EMBO Journal, 1989. **8**(5): p. 1455-1460.
5. Garvie, C.W. and C. Wolberger, *Recognition of Specific DNA Sequences.* Molecular Cell, 2001. **8**(5): p. 937-946.
6. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-151.
7. Spitz, F. and E.E.M. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-626.
8. Berman, B.P., et al., *Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(2): p. 757-762.
9. Dermitzakis, E.T. and A.G. Clark, *Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover.* Molecular Biology and Evolution, 2002. **19**(7): p. 1114-1121.
10. Stormo, G.D., et al., *Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli.* Nucleic Acids Research, 1982. **10**(9): p. 2997-3011.
11. Guigo, R. *An Introduction to Position Specific Scoring Matrices*. [cited 2017; Available from: http://bioinformatica.upf.edu/T13/MakeProfile.html.
12. Belton, J.-M., et al., *Hi-C: A comprehensive technique to capture the conformation of genomes.* Methods (San Diego, Calif.), 2012. **58**(3): p. 10.1016/j.ymeth.2012.05.001.
13. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8[thinsp]A resolution.* Nature, 1997. **389**(6648): p. 251-260.
14. Kouzarides, T., *Chromatin Modifications and Their Function.* Cell, 2007. **128**(4): p. 693-705.
15. Benevolenskaya, E.V., *Histone H3K4 demethylases are essential in development and differentiationThis paper is one of a selection of papers published in this Special Issue, entitled 28th International West Coast Chromatin and Chromosome Conference, and has undergone the Journal's usual peer review process.* Biochemistry and Cell Biology, 2007. **85**(4): p. 435-443.
16. Barski, A., et al., *High-Resolution Profiling of Histone Methylations in the Human Genome.* Cell, 2007. **129**(4): p. 823-837.
17. Rosenfeld, J.A., et al., *Determination of enriched histone modifications in non-genic portions of the human genome.* BMC genomics, 2009. **10**(1): p. 1.
18. Steger, D.J., et al., *DOT1L/KMT4 Recruitment and H3K79 Methylation Are Ubiquitously Coupled with Gene Transcription in Mammalian Cells.* Molecular and Cellular Biology, 2008. **28**(8): p. 2825-2839.
19. Koch, C.M., et al., *The landscape of histone modifications across 1% of the human genome in five human cell lines.* Genome Research, 2007. **17**(6): p. 691-707.
20. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(50): p. 21931-21936.
21. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Research, 2011. **21**(3): p. 381-395.

22. Holwerda, S.J.B. and W. de Laat, *CTCF: the protein, the binding partners, the binding sites and their chromatin loops.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2013. **368**(1620): p. 20120369.
23. Jabbari, K. and G. Bernardi, *Cytosine methylation and CpG, TpG (CpA) and TpA frequencies.* Gene, 2004. **333**: p. 143-149.
24. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes & Development, 2011. **25**(10): p. 1010-1022.
25. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-186.
26. Illingworth, R.S., et al., *Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome.* PLoS Genetics, 2010. **6**(9): p. e1001134.
27. Germain, P.-L., E. Ratti, and F. Boem, *Junk or functional DNA? ENCODE and the function controversy.* Biology & Philosophy, 2014. **29**(6): p. 807-831.
28. The, E.P.C., *A User's Guide to the Encyclopedia of DNA Elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.
29. The, E.P.C., *An Integrated Encyclopedia of DNA Elements in the Human Genome.* Nature, 2012. **489**(7414): p. 57-74.
30. Natarajan, A., et al., *Predicting cell-type–specific gene expression from regions of open chromatin.* Genome Research, 2012. **22**(9): p. 1711-1722.
31. Polak, P., et al., *Cell-of-origin chromatin organization shapes the mutational landscape of cancer.* Nature, 2015. **518**(7539): p. 360-364.
32. Valouev, A., et al., *Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data.* Nature methods, 2008. **5**(9): p. 829-834.
33. Landt, S.G., et al., *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.* Genome Research, 2012. **22**(9): p. 1813-1831.
34. *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
35. Hoffman, M.M., et al., *Integrative annotation of chromatin elements from ENCODE data.* Nucleic Acids Research, 2013. **41**(2): p. 827-841.
36. Neph, S., et al., *BEDOPS: high-performance genomic feature operations.* Bioinformatics, 2012. **28**(14): p. 1919-1920.
37. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
38. Crawford, G.E., et al., *Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).* Genome Research, 2006. **16**(1): p. 123-131.
39. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harbor protocols, 2010. **2010**(2): p. pdb.prot5384-pdb.prot5384.
40. Chu, Y. and D.R. Corey, *RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation.* Nucleic Acid Therapeutics, 2012. **22**(4): p. 271-274.
41. Ramsköld, D., E. Kavak, and R. Sandberg, *How to Analyze Gene Expression Using RNA-Sequencing Data*, in *Next Generation Microarray Bioinformatics*, J. Wang, A.C. Tan, and T. Tian, Editors. 2012, Humana Press. p. 259-274.
42. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.
43. Ponting, C.P. and R.C. Hardison, *What fraction of the human genome is functional?* Genome Research, 2011. **21**(11): p. 1769-1776.
44. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome research, 2005. **15**(8): p. 1034-1050.

45. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Research, 2005. **15**(8): p. 1034-1050.
46. Blow, M.J., et al., *ChIP-seq Identification of Weakly Conserved Heart Enhancers.* Nature genetics, 2010. **42**(9): p. 806-810.
47. Woolfe, A., et al., *Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development.* PLoS Biol, 2004. **3**(1): p. e7.
48. Pennacchio, L.A., et al., *In vivo enhancer analysis of human conserved non-coding sequences.* Nature, 2006. **444**(7118): p. 499-502.
49. Marshall, H., et al., *A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1.* Nature, 1994. **370**(6490): p. 567-571.
50. Evan, G.I. and K.H. Vousden, *Proliferation, cell cycle and apoptosis in cancer.* Nature, 2001. **411**(6835): p. 342-348.
51. Jemal, A., et al., *Cancer Statistics, 2008.* CA: A Cancer Journal for Clinicians, 2008. **58**(2): p. 71-96.
52. Zeng, Z.S., et al., *High level of Nm23-H1 gene expression is associated with local colorectal cancer progression not with metastases.* British Journal of Cancer, 1994. **70**(5): p. 1025-1030.
53. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer genes.* Nature, 2013. **499**(7457): p. 214-218.
54. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-1068.
55. The International Cancer Genome, C., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-998.
56. Bamford, S., et al., *The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.* British Journal of Cancer, 2004. **91**(2): p. 355-358.
57. Brown, M.P.S., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(1): p. 262-267.
58. Shipp, M.A., et al., *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nat Med, 2002. **8**(1): p. 68-74.
59. Libbrecht, M.W. and W.S. Noble, *Machine learning applications in genetics and genomics.* Nat Rev Genet, 2015. **16**(6): p. 321-332.
60. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
61. Kazemian, M., et al., *Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development.* Nucleic acids research, 2013. **41**(17): p. 8237-8252.
62. Neph, S., et al., *An expansive human regulatory lexicon encoded in transcription factor footprints.* Nature, 2012. **489**(7414): p. 83-90.
63. Gordân, R., A.J. Hartemink, and M.L. Bulyk, *Distinguishing direct versus indirect transcription factor–DNA interactions.* Genome Research, 2009. **19**(11): p. 2090-2100.
64. Li, Q., et al., *MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS.* The Annals of Applied Statistics, 2011. **5**(3): p. 1752-1779.
65. Zhang, Y., et al., *Model-based Analysis of ChIP-Seq (MACS).* Genome Biology, 2008. **9**(9): p. R137-R137.
66. Mao, Z., W. Cai, and X. Shao, *Selecting significant genes by randomization test for cancer classification using gene expression data.* Journal of Biomedical Informatics, 2013. **46**(4): p. 594-601.
67. Franken, P., *Haiqht, Fraxk A.: Handbook of the Poisson Distribution. (Publications in Operation Research No. 11.) John Wiley & Sons, Inc. New York, London, Sydney 1967. XI + 168 S.* Biometrische Zeitschrift, 1970. **12**(1): p. 66-67.
68. Audic, S. and J.-M. Claverie, *The significance of digital gene expression profiles.* Genome research, 1997. **7**(10): p. 986-995.

69. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
70. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Research, 2006. **34**(Database issue): p. D535-D539.
71. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Research, 2004. **32**(Database issue): p. D452-D455.
72. Stewart, A.J., S. Hannenhalli, and J.B. Plotkin, *Why Transcription Factor Binding Sites Are Ten Nucleotides Long.* Genetics, 2012. **192**(3): p. 973-985.
73. Kazuhiro, M., et al., *The physical size of transcription factors is key to transcriptional regulation in chromatin domains.* Journal of Physics: Condensed Matter, 2015. **27**(6): p. 064116.
74. Horton, N.J., E. Kim, and R. Saitz, *A cautionary note regarding count models of alcohol consumption in randomized controlled trials.* BMC Medical Research Methodology, 2007. **7**: p. 9-9.
75. Fitch, W.M., *Distinguishing Homologous from Analogous Proteins.* Systematic Biology, 1970. **19**(2): p. 99-113.
76. Sharan, R., et al., *Conserved patterns of protein interaction in multiple species.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(6): p. 1974-1979.
77. Lozzio, C. and B. Lozzio, *Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome.* Blood, 1975. **45**(3): p. 321-334.
78. Lozzio, B.B., et al., *A multipotential leukemia cell line (K-562) of human origin.* Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.), 1981. **166**(4): p. 546-550.
79. Kelly, K., et al., *Cell-specific regulation of the c-myc gene by lymphocyte mitogens and platelet-derived growth factor.* Cell, 1983. **35**(3): p. 603-610.
80. Schmeier, S., B. Jankovic, and V.B. Bajic, *Simplified Method to Predict Mutual Interactions of Human Transcription Factors Based on Their Primary Structure.* PLoS ONE, 2011. **6**(7): p. e21887.
81. Ben-Hur, A. and W.S. Noble, *Kernel methods for predicting protein–protein interactions.* Bioinformatics, 2005. **21**(suppl_1): p. i38-i46.
82. Berezikov, E., et al., *CONREAL: Conserved Regulatory Elements Anchored Alignment Algorithm for Identification of Transcription Factor Binding Sites by Phylogenetic Footprinting.* Genome Research, 2004. **14**(1): p. 170-178.
83. Eisermann, K., et al., *Evolutionary conservation of zinc finger transcription factor binding sites in promoters of genes co-expressed with WT1 in prostate cancer.* BMC Genomics, 2008. **9**(1): p. 1-15.
84. Whitfield, T.W., et al., *Functional analysis of transcription factor binding sites in human promoters.* Genome Biology, 2012. **13**(9): p. R50-R50.
85. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.
86. Nielsen, R., *Statistical methods in molecular evolution*. Vol. 6. 2005: Springer.
87. Dermitzakis, E.T., A. Reymond, and S.E. Antonarakis, *Conserved non-genic sequences [mdash] an unexpected feature of mammalian genomes.* Nat Rev Genet, 2005. **6**(2): p. 151-157.
88. Chan, Y. and R.P. Walmsley, *Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups.* Physical therapy, 1997. **77**(12): p. 1755-1761.
89. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Research, 2004. **32**(Database issue): p. D91-D94.

90. Yin, Y., S. Qiu, and Y. Peng, *Functional roles of enhancer of zeste homolog 2 in gliomas.* Gene, 2016. **576**(1, Part 2): p. 189-194.
91. Yano, K., et al., *Identification and Characterization of Human ZNF274 cDNA, which Encodes a Novel Kruppel-type Zinc-Finger Protein Having Nucleolar Targeting Ability.* Genomics, 2000. **65**(1): p. 75-80.
92. Nomura, N., et al., *Isolation of human cDNA clones of jun-related genes, jun-B and jun-D.* Nucleic Acids Research, 1990. **18**(10): p. 3047-3048.
93. Hess, J., P. Angel, and M. Schorpp-Kistner, *AP-1 subunits: quarrel and harmony among siblings.* Journal of Cell Science, 2004. **117**(25): p. 5965-5973.
94. Simon, A.L., E.A. Stone, and A. Sidow, *Inference of functional regions in proteins by quantification of evolutionary constraints.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(5): p. 2912-2917.
95. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Research, 2012. **22**(9): p. 1798-1812.
96. Hu, Z. and S.M. Gallo, *Identification of interacting transcription factors regulating tissue gene expression in human.* BMC Genomics, 2010. **11**: p. 49-49.
97. He, Q., et al., *High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species.* Nat Genet, 2011. **43**(5): p. 414-420.
98. Fatemi, M., et al., *Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level.* Nucleic Acids Research, 2005. **33**(20): p. e176-e176.
99. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(5): p. 1412-1417.
100. Ohler, U. and H. Niemann, *Identification and analysis of eukaryotic promoters: recent computational approaches.* Trends in Genetics, 2001. **17**(2): p. 56-60.
101. Pott, S. and J.D. Lieb, *What are super-enhancers?* Nat Genet, 2015. **47**(1): p. 8-12.
102. Martínez-Estrada, Ofelia M., et al., *The transcription factors Slug and Snail act as repressors of Claudin-1 expression in epithelial cells.* Biochemical Journal, 2006. **394**(Pt 2): p. 449-457.
103. Palstra, R.-J. and F. Grosveld, *Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux.* Frontiers in Genetics, 2012. **3**: p. 195.
104. Agarwal, S.K., et al., *Menin Interacts with the AP1 Transcription Factor JunD and Represses JunD-Activated Transcription.* Cell, 1999. **96**(1): p. 143-152.
105. Zhong, S., P. Salomoni, and P.P. Pandolfi, *The transcriptional role of PML and the nuclear body.* Nat Cell Biol, 2000. **2**(5): p. E85-E90.
106. Orphanides, G. and D. Reinberg, *A Unified Theory of Gene Expression.* Cell, 2002. **108**(4): p. 439-451.
107. Odom, D.T., et al., *Control of Pancreas and Liver Gene Expression by HNF Transcription Factors.* Science (New York, N.Y.), 2004. **303**(5662): p. 1378-1381.
108. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nat Rev Genet, 2013. **14**(4): p. 288-295.
109. Mendelson, C.R., *Role of transcription factors in fetal lung development and surfactant protein gene expression.* Annual Review of Physiology, 2000. **62**(1): p. 875-915.
110. Hemberg, M. and G. Kreiman, *Conservation of transcription factor binding events predicts gene expression across species.* Nucleic Acids Research, 2011. **39**(16): p. 7092-7102.

111. Bonn, S., et al., *Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development.* Nat Genet, 2012. **44**(2): p. 148-156.
112. Wilczynski, B., et al., *Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State.* PLoS Comput Biol, 2012. **8**(12): p. e1002798.
113. Shen, Y., et al., *A map of the cis-regulatory sequences in the mouse genome.* Nature, 2012. **488**(7409): p. 116-120.
114. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-318.
115. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biology, 2012. **13**(9): p. 1-22.
116. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.
117. O'Connor, T.R. and T.L. Bailey, *Creating and validating cis-regulatory maps of tissue-specific gene expression regulation.* Nucleic Acids Research, 2014. **42**(17): p. 11000-11010.
118. Koo, J., et al., *A GUS/luciferase fusion reporter for plant gene trapping and for assay of promoter activity with luciferin-dependent control of the reporter protein stability.* Plant and cell physiology, 2007. **48**(8): p. 1121-1131.
119. Wang, C., M.Q. Zhang, and Z. Zhang, *Computational Identification of Active Enhancers in Model Organisms.* Genomics, Proteomics & Bioinformatics, 2013. **11**(3): p. 142-150.
120. Pohl, A. and M. Beato, *bwtool: a tool for bigWig files.* Bioinformatics, 2014. **30**(11): p. 1618-1619.
121. MARIANI, T.J., et al., *A variable fold change threshold determines significance for expression microarrays.* The FASEB Journal, 2003. **17**(2): p. 321-323.
122. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(9): p. 5116-5121.
123. Li, Z., et al., *A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells.* Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8164-8169.
124. Patikoglou, G.A., et al., *TATA element recognition by the TATA box-binding protein has been conserved throughout evolution.* Genes & Development, 1999. **13**(24): p. 3217-3230.
125. Punnamoottil, B., et al., *Cis-regulatory characterization of sequence conservation surrounding the Hox4 genes.* Developmental Biology, 2010. **340**(2): p. 269-282.
126. Huang, D.W., et al., *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.* Genome Biology, 2007. **8**(9): p. R183-R183.
127. Schacht, T., et al., *Estimating the activity of transcription factors by the effect on their target genes.* Bioinformatics, 2014. **30**(17): p. i401-i407.
128. Cheng, C., et al., *Inferring activity changes of transcription factors by binding association with sorted expression profiles.* BMC Bioinformatics, 2007. **8**(1): p. 452.
129. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nat Rev Genet, 2012. **13**(1): p. 59-69.
130. Stadhouders, R., et al., *Transcription regulation by distal enhancers: Who's in the loop?* Transcription, 2012. **3**(4): p. 181-186.
131. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267-288.

132. Qabaja, A., et al., *Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions.* EURASIP Journal on Bioinformatics and Systems Biology, 2013. **2013**(1): p. 3-3.
133. Ghosh, D. and A.M. Chinnaiyan, *Classification and Selection of Biomarkers in Genomic Data Using LASSO.* Journal of Biomedicine and Biotechnology, 2005. **2005**(2): p. 147-154.
134. Lv, J., et al., *Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features.* Nucleic Acids Research, 2013. **41**(22): p. 10044-10061.
135. van der Ploeg, T. and E.W. Steyerberg, *Feature selection and validated predictive performance in the domain of Legionella pneumophila: a comparative study.* BMC Research Notes, 2016. **9**: p. 147.
136. Wang, Z., W. Xu, and Y. Liu, *Integrating full spectrum of sequence features into predicting functional microRNA–mRNA interactions.* Bioinformatics, 2015. **31**(21): p. 3529-3536.
137. Wu, T.T., et al., *Genome-wide association analysis by lasso penalized logistic regression.* Bioinformatics, 2009. **25**(6): p. 714-721.
138. Park, M.Y. and T. Hastie, *L1-regularization path algorithm for generalized linear models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2007. **69**(4): p. 659-677.
139. Tibshirani, R., *Regression shrinkage and selection via the lasso: a retrospective.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011. **73**(3): p. 273-282.
140. Dickel, D.E., et al., *Function-based identification of mammalian enhancers using site-specific integration.* Nat Meth, 2014. **11**(5): p. 566-571.
141. Blanchette, M., et al., *Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.* Genome Research, 2006. **16**(5): p. 656-668.
142. Su, J., S.A. Teichmann, and T.A. Down, *Assessing Computational Methods of Cis-Regulatory Module Prediction.* PLoS Computational Biology, 2010. **6**(12): p. e1001020.
143. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent.* Journal of statistical software, 2010. **33**(1): p. 1.
144. Blumenthal, T., *Gene clusters and polycistronic transcription in eukaryotes.* BioEssays, 1998. **20**(6): p. 480-487.
145. Visel, A., et al., *VISTA Enhancer Browser—a database of tissue-specific human enhancers.* Nucleic acids research, 2007. **35**(suppl 1): p. D88-D92.
146. Rajagopal, N., et al., *RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State.* PLoS Comput Biol, 2013. **9**(3): p. e1002968.
147. Chen, C.-y., Q. Morris, and J.A. Mitchell, *Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features.* BMC Genomics, 2012. **13**(1): p. 1-19.
148. Mar, J.C., et al., *Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease.* PLoS Genetics, 2011. **7**(8): p. e1002207.
149. Yaragatti, M., C. Basilico, and L. Dailey, *Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions.* Genome Research, 2008. **18**(6): p. 930-938.
150. Hobert, O., *Gene regulation by transcription factors and microRNAs.* Science, 2008. **319**(5871): p. 1785-1786.
151. Kellis, M., et al., *Defining functional DNA elements in the human genome.* Proceedings of the National Academy of Sciences, 2014. **111**(17): p. 6131-6138.
152. Melton, C., et al., *Recurrent Somatic Mutations in Regulatory Regions of Human Cancer Genomes.* Nature genetics, 2015. **47**(7): p. 710-716.

153. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
154. Rabbitts, T.H., *Chromosomal translocations in human cancer.* Nature, 1994. **372**(6502): p. 143-149.
155. Huang, F.W., et al., *Highly recurrent TERT promoter mutations in human melanoma.* Science (New York, N.Y.), 2013. **339**(6122): p. 957-959.
156. Melton, C., et al., *Recurrent somatic mutations in regulatory regions of human cancer genomes.* Nat Genet, 2015. **47**(7): p. 710-716.
157. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-1165.
158. Fredriksson, N.J., et al., *Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types.* Nat Genet, 2014. **46**(12): p. 1258-1263.
159. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Research, 2015. **43**(Database issue): p. D805-D811.
160. Hansen, R.S., et al., *Sequencing Newly Replicated DNA Reveals Widespread Plasticity in Human Replication Timing.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(1): p. 139-144.
161. Rhind, N. and D.M. Gilbert, *DNA Replication Timing.* Cold Spring Harbor perspectives in biology, 2013. **5**(8): p. a010132-a010132.
162. Dobson, A.J. and A. Barnett, *An introduction to generalized linear models*. 2008: CRC press.
163. Futreal, P.A., et al., *A CENSUS OF HUMAN CANCER GENES.* Nature reviews. Cancer, 2004. **4**(3): p. 177-183.
164. Shar, N.A., M. Vijayabaskar, and D.R. Westhead, *Cancer somatic mutations cluster in a subset of regulatory sites predicted from the ENCODE data.* Molecular Cancer, 2016. **15**(1): p. 76.
165. Piraino, S.W. and S.J. Furney, *Beyond the exome: the role of non-coding somatic mutations in cancer.* Annals of Oncology, 2016. **27**(2): p. 240-248.
166. Stamatoyannopoulos, J.A., et al., *Human mutation rate associated with DNA replication timing.* Nature genetics, 2009. **41**(4): p. 393-395.
167. Loots, G.G. and I. Ovcharenko, *rVISTA 2.0: evolutionary analysis of transcription factor binding sites.* Nucleic Acids Research, 2004. **32**(Web Server issue): p. W217-W221.
168. He, X., X. Ling, and S. Sinha, *Alignment and Prediction of <italic>cis</italic>-Regulatory Modules Based on a Probabilistic Model of Evolution.* PLoS Comput Biol, 2009. **5**(3): p. e1000299.
169. Fiedler, T. and M. Rehmsmeier, *jPREdictor: a versatile tool for the prediction of cis-regulatory elements.* Nucleic Acids Research, 2006. **34**(Web Server issue): p. W546-W550.
170. Dostie, J., et al., *Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements.* Genome Research, 2006. **16**(10): p. 1299-1309.
171. Whalen, S., R.M. Truty, and K.S. Pollard, *Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin.* Nature genetics, 2016. **48**(5): p. 488-496.
172. He, B., et al., *Global view of enhancer–promoter interactome in human cells.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(21): p. E2191-E2199.
173. Ong, C.-T. and V.G. Corces, *CTCF: An Architectural Protein Bridging Genome Topology and Function.* Nature reviews. Genetics, 2014. **15**(4): p. 234-246.
174. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-1245.

175. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-49.
176. Hardison, R.C. and J. Taylor, *Genomic approaches towards finding cis-regulatory modules in animals.* Nat Rev Genet, 2012. **13**(7): p. 469-483.

Appendix I

Narrow peak format

chrom	chromStart	chromEnd	name	score	strand	signalvalue	pvalue	qvalue	peak
chr1	713849	714369	.	358	.	56.11685	-1	4.723727	260
chr1	936119	936478	.	1000	.	219.308	-1	4.723727	177
chr1	948512	949032	.	261	.	40.93646	-1	4.723727	260
chr1	968260	968780	.	310	.	48.61145	-1	4.723727	260
chr1	999583	999866	.	647	.	101.2878	-1	4.723727	123
chr1	1243488	1244008	.	241	.	37.87323	-1	4.723727	260
chr1	1310568	1310702	.	382	.	59.91272	-1	4.723727	108
chr1	1342436	1342956	.	239	.	37.52594	-1	4.723727	260
chr1	1447124	1447644	.	385	.	60.39419	-1	4.723727	260
chr1	1609215	1609735	.	484	.	75.78422	-1	4.723727	260

1. **chrom:** Name of the chromosome
2. **chromStart:** The starting position of the feature in the chromosome
3. **chromEnd:** The ending position of the feature in the chromosome
4. **name:** Name given to a region. Use “.” if no name is assigned
5. **score:** Indicates how dark the peak will be displayed in the browser
6. **strand:** +/- to denote strand or orientation. Use “.” if no orientation is assigned
7. **signalValue:** Measurement of overall (usually, average) enrichment for the region
8. **pvalue:** Measurement of statistical significance (-log10). Use -1 if no pvalue is assigned.
9. **qvalue:** Measurement of statistical significance using false discovery rate (-log10). Use -1 if no qvalue is assigned.
10. **peak:** Point-source called for this peak; o-based offset from chromStart. Use -1 if no point source called.

(Taken from https://genome.ucsc.edu/FAQ/FAQformat#format12)

Appendix II

R code for building LASSO models

```
# CNN3.text is an input file
matrx=read.table("CNN3.text", header=TRUE)
matrx = as.matrix(matrx)
library (glmnet)
alpha=1
# DHSs signal intensities of all CRRs
x = matrx[,-1]
# Expression (FPKM) values
data = matrx[,1]
#fitting the model
fit=glmnet(x,data, family=c("gaussian"), alpha=alpha)
pdf("output.pdf")
plot(fit)
plot(fit, "lambda")
#Cross validation for glmnet
cvfit = cv.glmnet(x, data)
suma=summary (cvfit)
plot(cvfit)
# Adjusting the number of non-zero coefficients
lambdaFit = cvfit$glmnet.fit
df = lambdaFit$df
lambda = lambdaFit$lambda
fits=min(lambda[df==2])
# Below function makes predictions from cross validated glmnet model
predict_fit1<-predict(fit,x, s=fits)
sump=summary(predict_fit1)
cor(predict_fit1, data)
```

```
print (fits)

# Plotting the correlation between observed and predicted expression

title=paste("Correlation: ", cor(predict_fit1, data), "| Lambda = ", fits)

plot(predict_fit1,data, xlab="Predicted", ylab= "Observed", main=title)

x = as.matrix(coef(cvfit,s=fits))

nonZeroVecs = names(x[x!=0,])

#Plotting the two CRRs correlation between DHSs signal intensities and transcript expression (FPKM)

temp = as.data.frame(matrx)

for(i in 2:length(nonZeroVecs)){

   plot(temp$Expression,temp[,nonZeroVecs[i]], xlab="Expression", ylab = paste("DHS tag count in", nonZeroVecs[i]), main = nonZeroVecs[i] )

   cx  = cor(temp$Expression, temp[,nonZeroVecs[i]])

   cat(nonZeroVecs[i], "= ",cx,"\n")

}

dev.off()
```

Appendix III

“Cancer somatic mutations cluster in a subset of regulatory sites predicted from the ENCODE data”

Shar et al. Molecular Cancer (2016) 15:76
DOI 10.1186/s12943-016-0560-0

Molecular Cancer

RESEARCH Open Access

Cancer somatic mutations cluster in a subset of regulatory sites predicted from the ENCODE data

Nisar A. Shar[1,2], M. S. Vijayabaskar[1] and David R. Westhead[1*]

Abstract

Background: Transcriptional regulation of gene expression is essential for cellular differentiation and function, and defects in the process are associated with cancer. The ENCODE project has mapped potential regulatory sites across the complete genome in many cell types, and these regions have been shown to harbour many of the somatic mutations that occur in cancer cells, suggesting that their effects may drive cancer initiation and development. The ENCODE data suggests a very large number of regulatory sites, and methods are needed to identify those that are most relevant and to connect them to the genes that they control.

Methods: Predictive models of gene expression were developed by integrating the ENCODE data for regulation, including transcription factor binding and DNase1 hypersensitivity, with RNA-seq data for gene expression. A penalized regression method was used to identify the most predictive potential regulatory sites for each transcript. Known cancer somatic mutations from the COSMIC database were mapped to potential regulatory sites, and we examined differences in the mapping frequencies associated with sites chosen in regulatory models and other (rejected) sites. The effects of potential confounders, for example replication timing, were considered.

Results: Cancer somatic mutations preferentially occupy those regulatory regions chosen in our models as most predictive of gene expression.

Conclusion: Our methods have identified a significantly reduced set of regulatory sites that are enriched in cancer somatic mutations and are more predictive of gene expression. This has significance for the mechanistic interpretation of cancer mutations, and the understanding of genetic regulation.

Keywords: Cancer mutations, Cis regulation, Gene regulation, Modelling, Regulatory regions

Background

The majority of work on the somatic mutations that are found in cancer cell genomes has focussed on the analysis of protein coding exons. These regions have clear functional significance, and because they represent only a very small fraction of the genome are more amenable to systematic experimental investigation (e.g. in whole exome sequencing studies). Analysis of these data, taking account of the relationship between mutational frequencies and variables such as replication timing and gene expression, has allowed the identification of recurrently mutated regions and protein coding genes that when mutated are likely to be oncogenic drivers [1].

The role of aberrant genetic regulatory processes in the initiation and progression of cancer, for example the constituent activation of transcription factors driven by chromosomal re-arrangements [2], has been appreciated for many years. More recently, the discovery of point mutations in the TERT gene promoter that occur in large percentages of cases in some cancer types and are strongly linked to gene expression changes [3, 4], along with developments in whole genome sequencing, have focussed the field on mutations that occur in potential regulatory elements within the genome. It has been shown that regulatory regions harbour significant numbers of the somatic mutations that have been observed

* Correspondence: D.R.Westhead@leeds.ac.uk
[1]School of Molecular and Cellular Biology, Garstang Building, University of Leeds, Leeds LS2 9JT, UK
Full list of author information is available at the end of the article

in cancer cell genomes [5], and this work has also provided some evidence of positive selection for mutations in these regions, suggesting that regulatory mutations may be important in promoting survival and reproduction of cancer cells in the host. Other related work has examined recurrently mutated regulatory elements [6] and discovered regions potentially regulating genes with known involvement in cancer. Further, a method for the discovery of mutations that are strongly linked to expression levels of nearby genes in cancer samples has been developed [7]. However, given the complexity of genetic regulation in eukaryotic cells, it is likely that current work reveals only a fraction of the regulatory aberrations driving cancer, and there is a clear need for new methods that will reveal different insights.

New technologies for DNA sequencing have revolutionised our ability to map regulatory regions of the genome. For example, the ENCODE project [8] has mapped gene expression, transcription factor binding to DNA and other relevant variables such as DNaseI hypersensitivity and chromatin modifications on a whole genome scale in many laboratory cell lines, and more recent studies have examined the regulation of cellular differentiation [9, 10]. These studies and others have led to the development of databases, for example RegulomeDB [11], and these provide a rich source of information on potential regulatory elements. However, genetic regulation operates at multiple levels, and despite the volume of data now available it remains an unmet challenge to convert this data into more detailed mechanistic understanding of the regulation of individual genes. A large number of candidate regulatory elements are identified in the genome by these technologies, and the possibility that genes are regulated by elements that are relatively distant in the genome makes the process of assigning regulatory elements to genes very difficult. Nevertheless, these large data sets allow the development of correlative models whereby candidate regulatory elements may be identified, and used to develop regulatory networks linking them to the genes they control [12]. Similar work has used logistic regression [13], and Thurman and coworkers [14] introduced models that link DNaseI hypersensitivity data in promoter and distal sites to identify regulatory regions. While these methods are clearly useful, independent experimental knowledge of the links between genes and their regulatory regions is presently too limited for effective method comparison and validation.

A useful alternative view of the utility of correlative models of genetic regulation is to examine them in the context of relevant independent biological data, such as the somatic mutations observed in cancer genomes. Here we introduce our own model of genetic regulation based on ENCODE and examine the mapping of cancer mutations from the COSMIC [15] database to the regulatory regions it identifies. This integration of two large public sources of biological information through modelling, has the potential to improve our understanding both of genetic regulation and cancer.

Results

Figure 1 illustrates the process of building a simple correlative model of gene expression for a single transcript. As described in the Methods, candidate regulatory regions (CRRs) were identified as the union of all sites of transcription factor binding and the top 25% of DNase1 hypersensitive sites in all the ENCODE cell types considered. Each transcript was considered to be potentially regulated by any CRR within 100 kB [16] of the transcription start site (TSS), in this case 72 CRRs. Although genes can be regulated by enhancers up to 1 MB from the TSS, the figure of 100kB was chosen to encompass most regulatory elements, for example those of the leukaemia related oncogene Lmo2 [17]. The aim of our model was to predict the expression level of the gene in each of the cell types, as measured by RNA-seq experiments, from signal intensities in DNase1 hypersensitivity data, which we use as a crude measure of activity (e.g. transcription factor binding) at the CRR concerned.

Given the large number of CRRs relative to the number of cell types in which gene expression was measured, we adopted a penalised regression approach (LASSO) to identify a small set containing just those elements with the strongest relationships to gene expression. Analysis of the LASSO data indicated that the best supported models were based on just two candidate regulatory elements per transcript. We subsequently refer to these elements as the 'chosen' CRRs, and the remaining elements as 'rejected' CRRs. In the case of the transcript in Fig. 1 a convincing model was constructed, showing a (Pearson) correlation of observed to predicted expression values of 0.97 (Fig. 1b). We further assessed the statistical significance of this model using a randomisation approach, resulting in a p value of 0.0016 (see Statistical significance of models in Methods). Figure 1c and d show the correlation of DNaseI signal intensities and expression for the two CRRs chosen by the LASSO method, and Fig. 1e shows a example rejected CRR. The genomic location of the CRRs is shown in Fig. 1f.

We next investigated the possibility of a large-scale model building exercise for all genes/transcripts, and also in a restricted set of 533 cancer census genes from COSMIC [18]. We focussed on transcripts from GENCODE v7, and restricted the study to transcripts expressed in at least 7 cell types, which were more suitable for our regression based modelling techniques. Thus our study focused on genes expressed in a wider range of cell types, and we call these 'globally expressed' genes. The relevant statistics of model building are

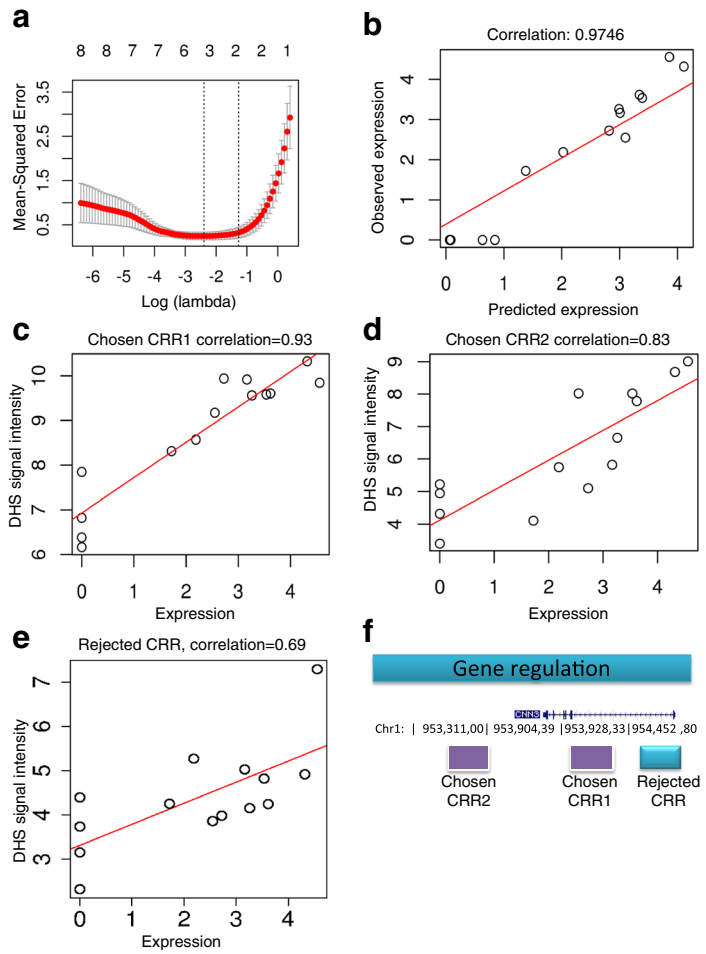


Fig. 1 Building an expression model for CNN3 (ENST00000370206.4). **a** shows the mean squared error against the log (λ) LASSO penalty parameter with numbers above the graph indicating the number of predictive variables (non-zero coefficients) in the corresponding LASSO model. Dotted lines show possible choices of λ at minimum mean-squared error (λ_{min}) and more conservatively at that value plus 1 standard error. This identifies models with 2 predictive variables as optimal. **b** shows the correlation between observed expression and predicted expression from the model. **c** and **d** show the correlation of DNaseI signal intensities and expression for the two candidate regulatory elements (CRRs) chosen by the LASSO method. **e** shows the correlation between DNaseI signal intensities and expression for an example rejected CRR. **f** shows the genomic location of the two chosen CRRs and one example rejected CRR

shown in Table 1, and a list of all chosen CRRs along with their target genes is included in Additional file 1: Table S1. Models were successfully built for approximately 9000 genes (16000 transcripts), and 290 genes (650 transcripts) from the cancer set. It should be noted that the scale of this model building exercise leads, after correction for multiple testing, to a significant false discovery rate. While any individual model should be considered carefully in this light, we treated the exercise as a means to the identification of a single set of CRRs covering a substantial proportion of the transcriptome that lead to the best supported models of gene expression (the chosen set), and a complement set of rejected CRRs with weaker relationships to gene expression. It should be noted that some elements were chosen for more than one transcript, and this is illustrated in Fig. 2, which also highlights the transcription factors known to bind in each CRR. As an illustration of the results for more genes, in Additional file 2: Figures S1 and S2 we include four examples (*WNT5A*, *ID1*, *LIMS1* and *TEAD3*) where predicted CRRs coincide with regulatory elements that are already known [19].

The COSMIC database [15] is a high-quality compilation of somatic mutations that have been observed in cancer cells. Mutations were downloaded from this database (a total of 2.3 million mutations) and mapped to the CRRs, as illustrated in Fig. 2. Overall 8% of these mutations mapped to CRRs identified with the transcript set defined above, and 14% of transcripts mapped to at least one mutated CRR. Table 2 gives statistics showing how these mutations are partitioned between chosen and rejected CRRs from the modelling exercise. This

Table 1 Statistics of model building

	All transcripts	Cancer set transcripts
Number of models attempted	17963 transcripts from 9209 genes	731 transcripts (from 304 genes)
Number of models built	16134 (8670 genes)	654 (292 genes)
Average r,r^2	0.710, 0.519	0.718, 0.530
Range r^2	0.004–0.99	0.048–0.925
Total candidate elements	678020 (mean 42/transcript)	28844 (mean 44/transcript)
Chosen elements	25045 (2/transcript)	1140 (2/transcript)
Elements chosen for 1 transcript	20025	999
Elements chosen for >1 transcript	5020	141

shows that a significantly higher proportion of chosen CRRs are mutated at least once compared rejected CRRs, and that chosen CRRs harbour around 1.5 times more mutations than rejected CRRs. This applies equally to all genes and to the cancer related subset. Within the all genes set all comparisons are highly statistically significant, while the smaller cancer genes set shows the same trends but with reduced levels of statistical significance. When limiting the analysis to CRRs only from higher quality expression models (confident models with $r > 0.7$ and highly confident models with $r > 0.8$) the effect size increases: mutations are enriched in chosen CRRs by a factor of 1.45 (=1.35/0.93) in all models and this rises to 1.81 in CRRs for highly confident models.

It is known that DNA mutation frequencies are heterogeneous [1] over the genome, and are related to variables such as replication timing and GC content. Equally, in the context of this analysis, average mutation frequencies within CRRs might be expected to be affected by the length of the CRR and possibly the proximity to a transcription start site (TSS). We took two different approaches to investigate whether these effects could have biased the statistical considerations above. First we repeated the significance tests on the mean

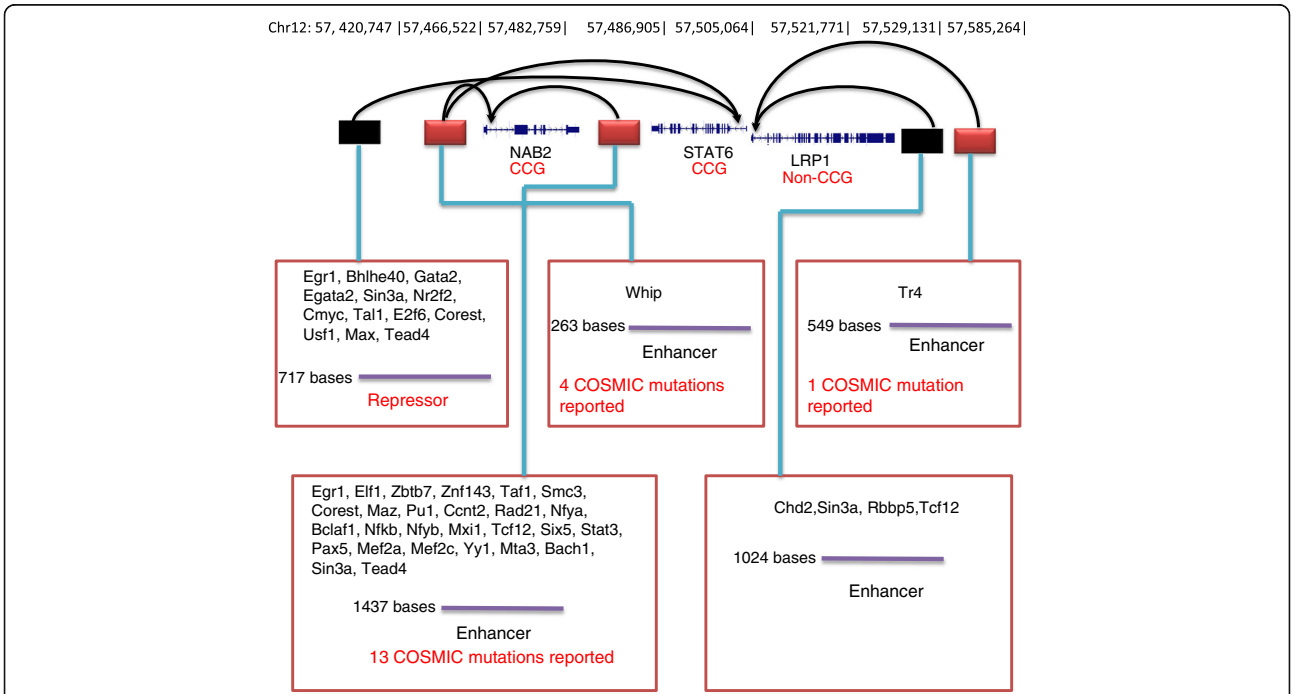


Fig. 2 The chosen CRRs for NAB2 (ENST00000342556.5), STAT6 (ENST00000300134.2) and LRP1 (ENST00000243077.2). *Black* arrows link CRRs to the transcripts for which they were chosen in expression models; note that one CRR was chosen for both STAT6 and NAB2. Details of the chosen CRRs are given *red* boxes, including the bound transcription factors, sizes of the CRRs and mutations mapped from the COSMIC database. CRRs are labelled as enhancers if they show positive correlation with expression and repressors if they show negative correlation. The chosen CRRs are marked as red boxes if there is at least one reported mutation in them, and black otherwise

Table 2 Mapping of somatic mutations from COSMIC to candidate regulatory regions (CRRs)

Title	All		Cancer census genes	
	Chosen CRRs	Rejected CRRs	Chosen CRRs	Rejected CRRs
Total number of CRRs	25045	158560	1140	7429
CRRs mutated at least once	3535 (14.11%)[1]	16241 (10.24%)[1]	160 (14.03%)[2]	703 (9.46%)[2]
Mean mutations/CRR	1.35[3]	0.93[3]	1.51[4]	0.97[4]
Mean mutations/CRR (models with $r > 0.7$)	1.40[3]	0.88[3]	1.63[5]	0.95[5]
Mean mutations/CRR (models with $r > 0.8$)	1.50[3]	0.83[3]	1.55	0.92

[1]Proportion mutated in chosen set greater than in rejected set, $p < 10^{-15}$ (Chi-squared and Fisher test)
[2]Proportion mutated in chosen set greater than in rejected set, $p < 10^{-5}$ (Chi-squared and Fisher test)
[3]Mean mutations in chosen set greater than in rejected set, $p < 10^{-23}$ (two sample t test), $p < 10^{-8}$ (Wilcoxon test)
[4]Mean mutations in chosen set greater than in rejected set, $p < 0.05$ (two sample t test and Wilcoxon test)
[5]Mean mutations in chosen set greater than in rejected set, $p < 0.05$ (two sample t test), $p = 0.06$ (Wilcoxon test)

number of mutations per CRR, this time not using the entire set of rejected CRRs but by randomly choosing a set of equal size to the chosen set matched according to the variable concerned (e.g. matching each chosen set member with a rejected member falling in the same GC content bin). In the case of all variables (replication timing, GC content, length of CRR and proximity to a TSS) the effects reported above remained significant, albeit with reduced levels of significance reflecting the reduction in size of the rejected set. Second, to model all these potential effects simultaneously we built generalised linear models for the counts of mutations in CRRs. We found the counts to be over-dispersed with respect to a Poisson distribution assumption, and modelled this with an additional dispersion parameter (see Methods). The effect size for an indicator variable showing whether a CRR was chosen or rejected was 0.46+/−0.02 ($p < 2 \times 10^{-16}$,Wald test), revealing a highly significant effect on the (log) expected mutation counts consistent in size with observed differences in average mutation counts from Table 2.

Finally, within the chosen set of CRRs we tested for differences in the average number of mutations in different types of CRR. Chosen CRRs may be positively or negatively correlated with expression of the associated gene, and hence tentatively identified with enhancing or repressing mechanisms. Of our chosen CRRs 32% showed negative correlations with expression, but there was no significant difference in mutation rates between these two types of CRR, whether considering all models or just those from cancer associated genes. On the other hand dividing CRRs into proximal or distal according to distance from the associated transcription start site (distal > 10 kB, proximal < 10 kB) showed a significant tendency for proximal CRRs to be mutated to higher levels, as shown in Table 3 and previously reported [6]. This effect seems to be more pronounced in elements identified with cancer associated genes.

Discussion

The recent revolution in DNA sequencing speed has allowed us to map multiple variables relevant to genetic regulation at genome-scale and sequence the genomes of many individual cancers. The work reported here is relevant to two important problems that arise from this data: the first is to move from a descriptive understanding of potential regulatory regions to a mechanistic understanding of the regulation of individual genes, and the second to understand which somatic mutations in cancer cells drive the process of cancer progression and to identify underlying mechanisms.

In respect of the problem of understanding genetic regulation, the genome scale data sets we have presently still represent relatively little data for each individual gene or transcript. The complexity of regulation in eukaryotic cells, involving the interactions of transcription factors and chromatin modifiers as well as miRNAs and lncRNAs, and the potential involvement of DNA regions (enhancers) distal to the transcript, mean that our present levels of mechanistic insight are limited. Based on the large scale data we have, the best that is possible is the building of simple correlative models, which aim to identify just those regions of the genome that seem most strongly influential on gene expression. As we have already commented, even this is subject to a significant false discovery rate when attempted at genome-scale.

Table 3 Mapping of mutations to chosen CRRs proximal and distal to the transcription start site

	Proximal (<10kB from TSS)	Distal (>10kB from TSS)
Mean mutations/CRR (all models)	2.30[1]	1.25[1]
Mean mutations/CRR (cancer related transcripts)	3.40[2]	0.99[2]

[1]Mean greater in proximal set, $p < 10^{-39}$ (*t*-test), $p < 10^{-17}$ (Wilcoxon)
[1]Mean greater in proximal set, $p < 10^{-6}$ (*t*-test), $p < 0.05$ (Wilcoxon)

Nevertheless, changes to genetic regulation are an important feature of cancer, and the results reported here show that a set of candidate regulatory regions derived from simple correlative models preferentially harbour cancer somatic mutations, suggesting that these regions are of functional significance in genetic regulation.

Conclusions

It is now recognised that mutations affecting regulatory regions are potentially as important in cancer progression as mutations in protein coding regions or those that directly alter functional RNA molecules. Here we have shown that somatic mutations that are found in cancer cells occur preferentially in those potential regulatory regions that are revealed by the ENCODE data to be more likely to be directly involved in the regulation of gene expression levels. This adds to the growing body of work in this area strongly suggesting that cancer progression involves positive selection for mutations with regulatory effects. This work also shows that modelling based on large data compendia like ENCODE can identify genomic regions which are potentially more strongly linked to gene expression, and propose links to the regulated genes. This could lead to more effective definition and prioritisation of mechanistic hypotheses for cancer somatic mutations, which will be accessible to confirmation or refutation with further detailed laboratory investigations.

Methods

Data sets and identification of candidate cis regulatory regions

Data sets were downloaded from ENCODE [8] for human genome version hg19 as shown in Table 4. Candidate Regulatory Regions (CRRs) were defined as all transcription factor binding sites (TFBS) found in the five cell types for which ChIP-seq data for transcription factors was available, plus the highest scoring 25% of DNaseI hypersensitive (DHS) sites for all 14 cell types, filtered to include only those with the H3K27ac active enhancer mark in at least one cell type. We used DHSs generated by the uniform processing pipeline of the ENCODE Analysis Working Group (AWG) for this study [8], and similarly TFBS were taken from the ENCODE standard data processing pipeline [8].

The DHS and TFBS were merged if they overlapped by at least 1 base pair using bedtools [20] and the resulting merged regions were considered as the full set of candidate regulatory regions (CRRs) for further analysis. DNaseI-seq signal intensities for each CRR in the 14 cell types (Table 4) were computed from the uniformly processed and normalised signal tracks using bwtool [21].

RNA-seq (whole-cell polyA+) transcript quantifications were downloaded from the ENCODE DCC portal of UCSC genome browser [22]. The expression for any transcript whose coordinates are defined by GENCODE (version 7) [23] is the average FPKM (Fragments Per Kilobase of transcript per Million sequenced reads) [24] of all the replicates, and they were filtered for IDR (Irreproducible Discovery Rate) < = 0.1. Further, only transcripts that were expressed (FPKM > =1) in at least 7 of the cell types defined in Table 4 were considered for all our analysis given below (such data is more suitable for our regression based modelling scheme). Our methodology is illustrated graphically in Fig. 3.

Table 4 ENCODE data sets used

S.No	Cell	ChIP-seq[a] (TFs)	DNaseI-seq	RNA-seq (FPKM)	ChIP-seq (H3K27ac)
1	K562	100	√	√	√
2	Gm12878	73	√	√	√
3	Hepg2	57	√	√	√
4	Helas3	54	√	√	√
5	H1hesc	47	√	√	√
6	A549		√	√	
7	Ag04450		√	√	
8	Bj		√	√	
9	Hsmm		√	√	
10	Huvec		√	√	
11	Mcf7		√	√	
12	Nhek		√	√	
13	Nhlf		√	√	
14	Sknshra		√	√	

[a]Total number of transcription factor ChIP-seq datasets considered, note that data sets of CTCF, CTCFL and RNA polymerase II were not used

Model

For applicability in the largest number of cell types, we based our model on DHS data and assumed a simple linear relationship between transcript expression (log (FPKM) values) and (log (signal intensity)) from the DNaseI data in each CRR.

$$y = k_0 + \sum_{i=1}^{n} k_i x_i$$

Here y is the expression value of the transcript, x_i the DNaseI signal intensity in the i^{th} CRR for that transcript and n is the number of CRRs within 100 kb of the transcription start site.

Since n is typically greater than the number of cell types for which data were available, model fitting demanded a penalised approach to limit the number of non-zero k_i coefficients. We chose LASSO regression implemented in the R glmnet package [25], which represents a least squares/maximum likelihood fit penalised with a term $\lambda \sum_{i=1}^{n} |k_i|$. We investigated a number of

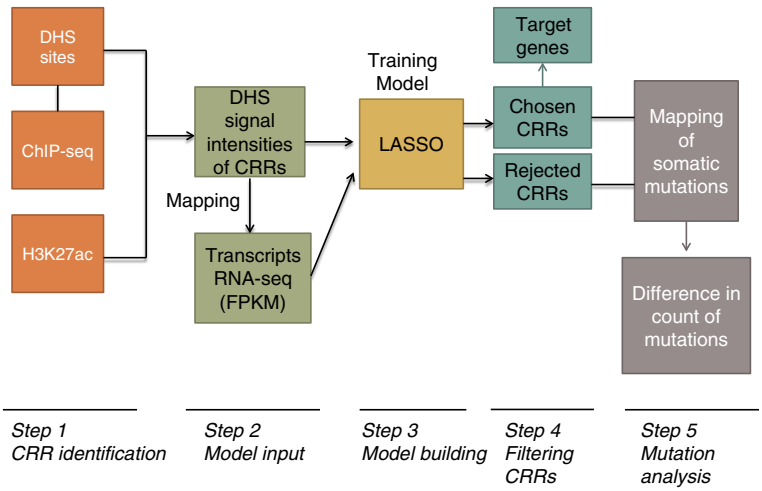


Fig. 3 Graphical summary of the methodology employed

different ways of determining appropriate values for the penalty scaling parameter λ, using a selection of example genes, and eventually chose conservatively so that two non-zero k_i parameters were determined for each model. As shown in Fig. 1, this is consistent with the glmnet package recommendation for choosing λ, as either λ_{min} (minimum mean square error) or this value plus one standard error. CRRs are subsequently referred to as 'chosen CRRs' if they appear with a non-zero coefficient in a LASSO model for at least one transcript, and rejected CRRs if they were considered in the analysis for any model but never associated with a non-zero coefficient. In the supplementary material we have included an input data file (Additional file 3) and R code (Additional file 4) to illustrate how the method can be implemented.

Statistical significance of models

The quality of the models was assessed through the (Pearson) correlation of predicted and observed gene expression values, using a leave-one-out cross validation scheme. To further assess statistical significance we generated models from randomly permuted data: we fixed the DHS data and generated 50000 random permutations of the gene expression values per transcript, calculating the empirical probability of obtaining a model from the random data showing a correlation at least as high as that for the model from the real data (using the same value of λ in each case). Since randomization is computationally expensive, we considered 12 models: 4 transcripts where the real model showed high correlation of predicted and actual expression (~0.9), 4 with moderate correlations (~0.5) and 4 where LASSO failed to find models. We found that the distribution of random model correlations was remarkably similar in all these cases and therefore used the distribution from these combined randomizations to generate p values for all models. When studying the generation of models for multiple genes we chose to control the false positive rate using the Benjamini-Hochberg method.

Mapping cancer mutations to regulatory regions

Somatic cancer mutations were derived from the COSMIC database [15] v76 (Catalogue of somatic mutations in cancer). 2.3 million somatic mutations were retrieved and mapped to the CRRs defined above. Duplicate/recurrent mutations were eliminated so only one mutation was considered at each genomic location.

Statistical significance of differences in mutation counts

The statistical significances of differences in the counts of somatic mutations observed in chosen and rejected CRRs were tested in several ways. Differences in the average number of mutations per CRR were tested with two-sample t-tests, and also equivalent non-parametric Wilcoxon tests to account for possible non-normality. To account for other possible effects that might bias these considerations we also repeated these tests after first balancing the chosen and rejected sets to have the same distribution of any potential confounding variable. This was achieved by sampling the rejected set of CRRs randomly to match the distribution of a variable in the chosen set, which was enabled by the significantly larger size of the rejected set. The variables considered were replication timing, base pair composition, length of the CRRs and distance of the CRR to the transcription start site (TSS). Replication timing and GC content data was downloaded from the UCSC website: the wavelet-smoothed signal of replication timing [26] for 9 cell types was obtained and we used the average signal. In

each case data was binned in 4 equal bins and the process required a chosen CRR to be matched by a rejected CRR from the same bin.

As an alternative test of statistical significance which enabled us to model all potential effects on mutation counts together, we built generalised linear models using the glm function in R. Mutation counts were modelled as a function of length of CRR, replication timing, GC content, shortest distance to a TSS and an indicator variable for chosen/rejected CRRs. A log link function was used, first under the assumption of a Poisson distribution for the counts and then in cases of over-dispersion using the quasipoisson option in glm, which fits a dispersion parameter which is otherwise fixed at unity. The statistical significances of the effects of each variable were assessed from the standard Wald test statistics produced by glm.

Cancer census genes

A set of 533 cancer consensus genes were retrieved from the COSMIC database of which 292 entered our analysis (the others did not meet our modelling criterion of expressing in at least 7 cell types). These were analysed as a separate subset to investigate any possible specific effects for genes known to be directly involved in cancer.

Additional files

Additional file 1: Identified regulatory sites containing somatic mutations. (XLSX 2006 kb)

Additional file 2: Examples of four genes, where predicted CRRs coincide with the regulatory elements that are already known. (PPTX 63 kb)

Additional file 3: Example of input data fi le for the predictive model. (TXT 14 kb)

Additional file 4: R code for LASSO model building. (R 1 kb)

Abbreviations

ChIP: Chromatin immunoprecipitation; COSMIC: Catalogue of somatic mutations in cancer; CRRs: Candidate cis regulatory regions; DHS: DNase I hypersensitive Site; ENCODE: Encyclopedia of DNA elements; FPKM: Fragments per kilobase of transcript per million sequenced reads; Glm: Generalized linear model; LASSO: Least absolute shrinkage and selection operator; TFBS: Transcription factor binding sites; TSS: Transcription start site

Funding

NAS acknowledges scholarship funding from NED University of Engineering & Technology, Karachi and University of Leeds. VMS and DRW acknowledge funding from BBSRC grant BB/I001220/1, and DRW acknowledges funding from MRC for the Leeds Medical Bioinformatics Centre.

Availability of data and materials

List of identified CRRs along with the mutation count will be provided on the journal website.

Authors' contributions

The study was conceived and designed by DRW. NAS carried out the study. VMS helped with data analysis, paper writing and supervision of the work. DRW and NAS wrote the paper and all authors approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publications

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]School of Molecular and Cellular Biology, Garstang Building, University of Leeds, Leeds LS2 9JT, UK. [2]Department of Biomedical Engineering, NED University of Engineering & Technology, University Road, Karachi 75270, Pakistan.

Received: 28 May 2016 Accepted: 15 November 2016
Published online: 25 November 2016

References

1. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer genes. Nature. 2013;499:214–8.
2. Rabbitts TH. Chromosomal translocations in human cancer. Nature. 1994; 372:143–9.
3. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly Recurrent TERT Promoter Mutations in Human Melanoma. Science. 2013; 339:957–9.
4. Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, et al. Frequency of TERT promoter mutations in human cancers. Nat Commun. 2013;4.
5. Melton C, Reuter JA, Spacek DV, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet. 2015;47:710–6.
6. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet. 2014;46:1160–5.
7. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet. 2014;46:1258–63.
8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
9. Goode DK, Obier N, Vijayabaskar MS, Lie ALM, Lilly AJ, Hannah R, Lichtinger M, Batta K, Florkowska M, Patel R, et al. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev Cell. 2016;36: 572–87.
10. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, et al. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell. 2012; 151:206–20.
11. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22: 1790–7.
12. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489:91–100.
13. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
14. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
15. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43:D805–11.
16. MacIsaac KD, Lo KA, Gordon W, Motola S, Mazor T, Fraenkel E. A quantitative model of transcriptional regulation reveals the influence of binding location on expression. PLoS Comput Biol. 2010;6, e1000773.
17. Landry JR, Bonadies N, Kinston S, Knezevic K, Wilson NK, Oram SH, Janes M, Piltz S, Hammett M, Carter J, et al. Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed

over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. Blood. 2009;113:5783–92.
18. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.
19. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35: D88–92.
20. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
21. Pohl A, Beato M. bwtool: a tool for bigWig files. Bioinformatics. 2014;30: 1618–9.
22. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
23. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
25. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;1:2010.
26. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. Sequencing Newly Replicated DNA Reveals Widespread Plasticity in Human Replication Timing. Proc Natl Acad Sci U S A. 2010;107:139–44.