A case-control approach to assess variability in distribution of distance between transcription factor binding site and transcription start site

by

Abdul Ragmaan Moos

Research Unit in Bioinformatics (RUBi)

Department of Biochemistry, Microbiology and Biotechnology

Rhodes University

A mini thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

of Rhodes University

2016

Abstract

Using the in-silico approach, with ENCODE ChIP-seq data for various transcription factors and different cell types; we systematically compared the distance between the transcription factor binding site (TFBS) and the transcription start (TSS). Our aim was to determine if the same transcription factor binds at a different position relative to the TSS in a normal and an abnormal cell type. We compare distribution of distance of binding sites from the TSS; to make description less verbose we call this "distance" where there is no possibility of confusion. We used a case-control methodology where the distance between the TFBS and the TSS in the normal, non-cancerous or untreated cell type is the control. The distance between the TFBS and the TSS in the cancerous or treated cell type is the case. We use the distance between the TFBS and the TSS in the control as the standard. We compared the distance between the TFBS and the TSS in the case and the control. If the distance between the TFBS and the TSS in the control was greater than the distance between the TFBS and the TSS in the case, we can infer the following. The transcription factor in the case binds closer to the TSS compared to the control. If the distance between the TFBS and the TSS in the control is smaller than the distance between the TFBS and the TSS in the case, we can infer the following. The TF in the case binds further away from the TSS compared to the control. Our method is a screening method whereby we compare ChIP-seq data to determine if there is a difference in the distribution distance between the TFBS and the TSS for normal and abnormal cell types. We used the R package ChIP-Enrich to compare the distribution of distance between ChIP-seq peak and the nearest TSS. ChIP-Enrich produces a histogram with the number of ChIP-seq peaks at a certain distance from the TSS. The results indicate for some transcription factors like GM12878-cMyc and K562-cMyc there is a difference between the distribution of distance between the TFBS and the nearest TSS. cMyc has more binding sites within a distance of 1kb from the TSS in GM12878 when compared to K562.

GM12878-CTCF and K562-CTCF have slight differences when comparing their distribution of distance from the TSS. This means CTCF binds almost the same distance from the TSS in both GM12878 and K562. A549-gr treated with dexamethasone is interesting because with increase dose of dexamethasone the distribution of distance from the TSS changes as well.

Acknowledgements

- Prof Philip Machanick – Gratitude and thanks for your guidance and support. You have always gone the extra mile to assist me.
- Dr Ozlem Tastan Bishop
- My wife-Su-aad
- Funding-Rhodes

  -NRF

Contents

Abbreviations

a30-Interferon alpha treatment for 30 min

a6hr-Interferon alpha treatment for 6hr

BED-Browser Extensible Data

ChIP-seq-chromatin immunoprecipitation followed by sequencing

SNP-single nucleotide polymorphism

dex-dexamethasone

dex 5nm-treatment dose of 5nm dexamethasone

dex 50nm-treatment dose of 50nm dexamethasone

dex 100 nm-treatment dose of 100nm dexamethasone

dex 500 pm-treatment dose of 500pm dexamethasone

g30-Interferon gamma treatment for 30 min

g6hr-Interferon gamma treatment for 6 hr

Ifn-Interferon

Haib-HudasonAlpha Institute for Biotechnology

NCBI-National Centre for Biotechnology Information

Sydh-Stanford/Yale/UCS/Harvard labs

TF-transcription factor

TSS-transcription start site

TFBS-transcription factor binding site

Uta-University of Texas, Arlington Lab

1 Introduction

Transcription factors (TF) are an important group of proteins that are essential for the process of transcribing DNA to RNA. Transcription factors either bind directly to DNA or bind with a host of other proteins to form a complex, which then binds to the DNA (Villard, 2004; Barrera and Ren, 2006).

How does a normal type become abnormal or cancerous? What is the difference between a normal and cancerous cell type? There are many reasons for a normal cell type to become cancerous and many differences between a normal and cancerous cell type. This study investigates the distance between the transcription factor-binding site (TFBS) and transcription start site (TSS). Could it be that a transcription factor binds closer or further away from the TSS, which causes the cell type to be abnormal. The aim of our study was to compare the distance between the TFBS and the TSS in a normal and abnormal cell type for the same transcription factor. We compare distribution of distance of binding sites from TSS; to make description less verbose we call this "distance" where there is no possibility of confusion.

Transcription factors bind at various distances from the transcription start site (TSS). However, no comprehensive study compared the distance between the TF binding site and the TSS in different cell types. We used an in-silico method with ChIP- seq data, and a case-control approach to compare the distance between the TF binding site and the TSS for the same TF but in two different cell types. Since we using an in-silico top- down approach, our starting point was the availability of ChIP-seq data for the control and case. We did not perform our own ChIP-seq experiments. We used ChIP-seq data from well- known established labs that is publicly available from ENCODE (Encode Consortium, 2011).

The centre of the ChIP-seq peak indicates the position of the TFBS. Our control is the distance between the TFBS and the nearest TSS in the normal or untreated cell type. Our case is always the distance between the TFBS and the TSS in abnormal or treated cell type. We use the same transcription factor in both the control and the case. For example, for the transcription factor CTCF, cell types GM12878 and K562. The distance between the TSS and the CTCF binding site in GM12878 was the control and the distance between the CTCF binding site and the TSS in K562 was the case. GM12878 is a normal cell type and K562 is a cancerous cell type.

The unit of measurement is the number of bases. We measure distance by counting the number of bases between the TFBS and the TSS.

The method for calculating the distance between the TFBS and the TSS is as follows:

Let us assume GM12878-CTCF has a ChIP-seq peak with the following arbitrary BED coordinates, chromosome 5, start position 1000 and end position 2000. We calculate the centre of the ChIP-seq peak by adding to the start position the difference between the end and start position and we add 1 to the total.

Centre of ChIP-seq peak (GM12878-CTCF) =Start + ((end-start)/2) +1

=1000 + (2000-1000)/2 + 1 = 1501

We use 1501 to indicate the position of the CTCF binding site in GM12878.

K562-CTCF with arbitrary BED coordinates chromosome 5, start position 1300 and end position of2500.

Centre of ChIP-seq peak (K562-CTCF) =Start + ((end-start)/2) +1

=1300 + (2500-1300)/2 + 1 = 1901

We used the refseq gene annotation data from University of California, Santa Cruz Bioinformatics (UCSC) to obtain the nearest transcription start site. We assume from the refseq data we find on chromosome 5 the nearest arbitrary TSS is at position 800.

We can calculate the distance between the TFBS and the nearest TSS as follows.

Distance between the TFBS and TSS (GM12878-CTCF) = Centre of ChIP-seq peak – TSS

=1501-800

=701

Distance between the TFBS and TSS (K562-CTCF) = Centre of ChIP-seq peak – TSS

=1901-800

=1101

We can now compare the distance between the TFBS and the TSS for GM12878-CTCF and K562-CTCF. In the example above, we notice with K562-CTCF the distance between the TFBS and the TSS is larger compare to GM12878-CTCF. We infer that CTCF binds further from the TSS in K562 as compared to GM12878.

The tool used to measure the genome wide distance between the TFBS and the TSS is ChIP-Enrich (Welch *et al.*, 2014) an R (R Development Core Team,2011) library which is part of Bioconductor (Doerge, 2006). ChIP-Enrich is a package specifically created to get the distribution of distance between ChIP-seq peak and the nearest TSS.

We use a case-control methodology to compare the distance between the TFBS and the TSS in normal and abnormal cell types.

The reason for using a case-control approach is that we need a standard for the distance between the TFBS and the TSS. The control gives us an idea as to what the normal distance between the TFBS and the TSS is. We compare the distance between the TFBS and the TSS in the case and the distance between the TFBS and the TSS in the control. Based on the measured distances we can determine if the transcription factors in the case binds at the same position, closer or further away from the TSS when compared to the control.

Studies have been done looking at the binding of TF relative to the TSS (Cheng *et al.*, 2011 ; Giannopoulou and Elemento, 2013); however a few short comings of these methods are the choice of control (Tallack *et al.*, 2012) or only a certain distance from the TSS is investigated (Cheng *et al.*, 2012).

Other studies also grouped many transcription factors together and then summarized their results (Wang *et al.*, 2012).

The approach of using a case-control methodology in comparing distance between the TFBS and the TSS in normal and abnormal cell types is novel and we hope to obtain meaningful results. We have tried to use ChIP-seq data from the same labs for both the control and case cell type for consistency and to keep the differences between control and case as negligible as possible.

Our aim of the project was to compare the distance between the TFBS and the TSS in control and case cell type and to determine whether there is a difference. We view the study as a screening method to identify whether distance between the TFBS and the TSS could be a factor that explains the differences between normal and abnormal cell type. If there is a difference between control and case, we can infer that the distance between the TFBS and the nearest TSS could possibly be one of many factors, which explain the difference between normal and abnormal cell types. If there is no difference between control and case, we can assume the distance between the TFBS and the nearest TSS is possible not a factor that could explain the difference between normal and abnormal cell type. If there is a difference between control and case distances, further investigation will be required. It is beyond the scope of this project to determine the reasons for the difference in distribution of distance between the TFBS and the TSS.

2 Hypothesis and Aims

2.1 Introduction

We compare the distance between the TFBS and the TSS in a normal cell type with the distance between the TFBS and the TSS in an abnormal cell type using a case-control methodology. The distance between the TFBS and the TSS in normal cell type is the control distance and the distance between the TFBS and the TSS in abnormal cell type is the case distance. Note we use the same TF in both the control and case distances and the only difference between the case and control is the cell type. We use a normal or untreated cell type for the control distances and an abnormal or treated cell type for the case distances.

2.2 Hypothesis

Our hypothesis is there is a difference between the distance of the TFBS and the TSS in the control cell type and distance between the TFBS and the TSS in the case type for the same transcription factor.

We tested our hypothesis by looking at the distribution distance between the ChIP-seq peak and the TSS and comparing the distribution distances between control and the case. Performing a genome wide comparison we looked at specific intervals or distance from the TSS when comparing the control and case distances.

2.3 Aims

Our aims are the following:

1. We use ChIP-seq data to measure the distance between the TFBS and nearest the TSS in a normal cell type. We defined this distance as our control distance.
2. We use ChIP-seq data to measure distance between the TFBS and the nearest TSS in an abnormal cell type. This distance we defined as our case distance.
3. Compare the distances between the control and the case. We have three scenarios for the difference in distance between the control and case.
    a. Control distance is greater than the case distance. The TF binds closer to the TSS in the case when compared to the control
    b. Control distance is smaller than the case distance. The TF binds closer to the TSS in the control as compared to the case.
    c. Control distance and case distance is the same. The TF binds at the same distance in control as compared to case.

3 Literature Review

3.1 Introduction

In this chapter, we review the current literature pertaining to transcription factors and the binding mechanism of TF with DNA, the manner in which TF activates genes. We also identify various sources of biological data. A particular section of this chapter deals with ENCODE because the source for all our data is the ENCODE project. Other sections in this chapter we cover include an overview of ChIP-seq and a discussion on cell types.

3.2 Overview of transcription

Chromosomes

Deoxyribonucleic acid (DNA) is the building block of nucleic material. DNA normally occurs as two long polymer chains. The chains consist of repeating units called nucleotides. A single polymer chain is a DNA strand. The DNA strand has a backbone, which consist of alternating phosphate and sugar residues. The sugar is a two deoxyribose, pentose sugar. The phosphate residue between the third and fifth carbon atoms of the adjacent ring, bind the sugars to from phosphodiester bonds, which gives the DNA strand polarity. The polarity in a DNA strand indicated by the 3' (hydroxyl) end and the 5' (phosphate) end.

Hydrogen bonds between base portion of the nucleotides and base stacking between the aromatic nucleobases stabilise and hold the DNA chains together. There are four bases found in DNA, namely thymine (T), adenine (A), quinine (G) and cytosine (C). Cytosine will pair with guanine and thymine with adenine (Watson and Crick, 1953; Yakovchuk *et al.*, 2006).

The DNA packaged with specific proteins such as histones into a structure called chromatin. The chromatin is further compressed into nucleosome which is DNA wrapped around 8 histone molecules. The nucleosome compressed further into a solenoid structure. The solenoidal structure is compressed further looping to form chromosomes.

For transcription, factors to bind to the transcription factor-binding site the DNA needs to be uncompressed and the correct binding site exposed for the transcription factor to bind to it. Chromatin remodelling is the process of uncompressing DNA and exposing the correct binding site for transcription factor. SWI/SNF complex acts in 3 different ways to make the DNA accessible. The first method is nucleosome remodelling whereby the structure of the nucleosome changed to allow transcription factors to bind to it. The second method, nucleosome sliding, the DNA is exposed by the nucleosome sliding along DNA. The last

method of nucleosome displacement involves nucleosome leaving and then binding to another nucleosome (Agalioti *et al.*, 2002).

DNA can also be uncompressed by histone modification, which results in a less tight chromatin structure. Histone modification occurs through acetylation, which is the addition of an acetyl group to free amino residue in the histone molecule, which reduces its net positive charge. Phosphorylation and methylation are also ways to open chromatin and expose DNA (Smale, 2001).

Genes

A gene is a small section of DNA, which codes for ribonucleic acid (RNA) which codes a particular protein. Genes consists of both introns and exons and a number of distinct regions e.g. promoter region, control element and transcription start site (TSS). The main purpose of the promoter region is for proteins, which has a DNA-binding motif to bind to the DNA. A transcription factor (TF) is a protein, which has a specific DNA-binding domain, which allows it to bind to DNA (Lemon and Tjian, 2000).

A gene has the following elements:

1. Gene promotor - the gene promotor or core promoter is the area upstream between the TSS and including the TATA box. Polymerase II and other proteins, which is essential for transcription, bind in this region. The TATA box is upstream next to TSS and it is an AT rich sequence found in many but not all genes. The function of the TATA box is to accurately position the start of the TSS. Genes, which do not have TATA box, may contain an initiator element, which is important for detecting the TSS.

2. Upstream promoter elements - The Sp1 box and CCAAT box are normally found in this region, which is upstream from the TATA box. These promotor elements are important for transcription and if they not present or damage then no transcription will take place (Lee *et al.*, 1987).

3. Regulatory elements - Glucocorticoid response element (GRE), Metal response elements (MRE). These elements allow genes to respond or react to specific stimuli (Halfon, 2006).

4. Enhances - Sequences which is located far away from a TSS, which enhances promotor activity. Enhancers can occur upstream, downstream or even within the transcription start sites. E.g. of enhances are AP1 and AP2 (Li *et al.*, 2006).

RNA Polymerase

RNA polymerase is an enzyme, which is responsible transcribing DNA. The types of RNA polymerase are RNA polymerase I, RNA Polymerase II and RNA Polymerase III. The three types of RNA polymerase are homologous to each other and their active site all occur in the second largest subunit. The three RNA polymerases are active on different genes and by using alpha-amanitin, which is a fungal toxin one can distinguish between the three different RNA polymerases. RNA polymerase cannot transcribe genes on its own. RNA polymerase forms stable complexes with other transcription factors, which allow transcription to form. Without stable complexes, no transcription can occur (Sentenac, 1985).

RNA polymerase I

RNA polymerase I transcribes genes encoding 5.8S, 18S and 28S ribosomal RNA. RNA polymerase I is not sensitive to alpha-amanitin. In Acanthamoeba, the transcription factor TF-1 binds to DNA at the promoter region and polymerase I will bind next to TF-1. Transcription can start once RNA polymerase I binds next to TF-1. As RNA polymerase I moves along DNA TF-1 will remain at the promoter region and be ready for next RNA polymerase molecule (Paule and White, 2000).

In vertebrates, the transcription factors required to form a stable complex is UBF (upstream binding factor), SL1 (TF-1 homolog) and RNA polymerase I. UBF binds to the promotor region and facilitates the binding of SL1 (Russell and Zomerdijk, 2005).

RNA polymerase II

RNA polymerase II transcribe all genes encoding for proteins and small nuclear RNA, which is involved in RNA splicing. RNA polymerase II is very sensitive to alpha- amanitin and inhibited by a dose of 1µg/ml.

The transcription factor TFIID binds to DNA in the promoter region. TFIIB binds to TFIID. TFIIB can bind to RNA polymerase II directly. The RNA polymerase II and TFIIF is recruited by TFIIB. Following the binding of RNA polymerase II the transcription factors TFIIE, TFIIH and TFIIJ binds to the complex. TFIIH opens the

double stranded which allows the transcription of the DNA (Zurita and Merino, 2003). TFIIH phosphorylates RNA polymerase II that allows transcription to begin. During transcription as the RNA polymerase II and TFIIH moves along DNA to copy it, TFIIA and TFIIH remains at the promoter region allowing other RNA polymerase II molecules to bind to it and start again with transcription (Hahn, 2004).

An alternative way in which RNA polymerase II transcription happen is as follows.

The transcription factors TFIIA and TFIID binds to DNA at promoter region. A preformed complex consisting of TFIIB, TFIIF, TFIIE, TFIIH, TFIIJ and RNA polymerase II forms. The preformed complex binds to TFIIA and TFIID and gene transcription can begin (Myer and Young, 1998).

RNA polymerase III

RNA polymerase III transcribe the 5S ribosomal RNA and transfer RNMA. RNA polymerase III is moderately sensitive to alpha-amanitin and inhibited by a dose of 10μg/ml.

RNA polymerase III consists of three classes, namely class I, class II and class III. All three classes require TFIIIB for transcription.

Class I require TFIIIB with additional factors, TFIIIA and TFIIIC.TFIIIA and TFIIIC can bind directly to DNA in the promoter region however, TFIIIB require TFIIIC. The polymerase recognises TFIIIB and binds next to it and transcription is ready to begin.

Class II and class III does not need TFIIIA for transcription. TFIIIB, TFIIIC and the polymerase is all that needed for transcription (Hernandez, 1993).

Transcription is the process by which DNA is copied to form messenger RNA (mRNA), or as the fundamental dogma of biology states, DNA produces RNA, which produces protein. The transcription process consists of three phases namely initiation, elongation and termination.

The process of transcription can be summarised in three phases as follows. The first phase is the initiation phase consists of mainly identifying the core promoter region of the gene, assembly of the various proteins required including polymerase as well as the transcription

factor to form a complex. Once the complex has been assembled the transcription factor will be the link which binds the assembled complex to the.

The second phase is the elongation phase involves transcribing of the DNA to RNA, replacing the thymine base with the uracil base.

The last phase is the termination phase which the completion of the transcription process.

3.3 Genomic data

The three major institutes that stores genomic data are as follow:

1) Genbank (Mrozek *et al.*, 2013)
2) European molecular and biology lab (EMBL)
3) National Institute of Genetics, Mishima

These three institutes generally share information amongst each other and many contain the same information. There are various tools available to access the information from these databases.

To access information stored in Genbank you can use the University of California, Santa Cruz Bioinformatics (UCSC) Genome browser, the NCBI and ENSEMBL browsers.

3.3.1 UCSC Genome browser

The University of California, Santa Cruz Bioinformatics group produced the UCSC genome browser. Information displayed in the form of a track. Tracks provide a host of different features e.g. genes, CpG islands (areas which high a concentration of CpG (cytosine and guanine nucleotides)), SNP's and predicted gene regulatory regions (Rosenbloom *et al.*, 2013).

3.3.2 NCBI

The National Centre of Biotechnoly information is part of the National Institutes of Health, which provide genomic information for organisms with complete genomic sequence assembly as well for organism who has no or little sequence information available. The

browser is linked to other NCBI tools e.g. Entrez, UniGene, Online Mendelian Inheritance in Man, dbSNP and dbSTS.

3.3.3 ENSEMBL

ENSEMBL is a joint venture between the Sanger institute and the European Bioinformatics Institute (EBI). Ensembl provides a genome browser and genome information on various species, which includes humans, mouse and some plant species. Ensembl provides gene annotation data, comparative genomic data, regulation information, variation information (Flicek *et al.,* 2010). Ensembl also provides information from ENCODE and data can be downloaded using a web browser or via application programming interface (API) using Perl scripts (Flicek *et al.*, 2011).

3.4 ENCODE

Encyclopaedia of DNA Elements (ENCODE) Project's aim is to study and understand the human genome in order to improve health (Encode Consortium, 2011). The ENCODE project was started in 2003 by the National Human Genome Research Institute (NHGRI) with the goal of identify and annotating all functional elements which is encoded by the human genome. The ENCODE project is an international collaboration amongst various researchers. The ENCODE website has a vast amount of information and we used primarily the ChIP-seq data from the ENCODE to do our investigations for the project.

3.5 Overview of ChIP-seq

An in-silico method to analyze genome-wide TF binding. ChIP-seq is a method for identifying the genome-wide binding location of a particular TF. The ChIP-seq identifies peaks, which approximate the binding location of a particular TF. The ChIP-seq peak data file when in bed format and contains the chromosome number, start and end coordinates for the ChIP-seq peak.

Protein groups interact differently with the genome depending on the type of protein. There is three classes of proteins namely point source factors, broad source and mixed source factors. Point source factors generate highly localized ChIP-seq signal, which includes most transcription factors and indicates binding which occurs when the chromatin is open. The ENCODE database contains ChIP-seq information mostly on point source factors. Broad

source factors bind to a wide genomic region e.g. chromatin marks. Mixed source factors bind in certain region as point source factors and other regions as broad source factors.

ChIP-seq data will combine with JASPAR or TRANSFAC motif database to identify transcription sites. TRANSFAC is a database, which contains information on eukaryotic DNA sequence elements as well as the transcription factor information (Wingender *et al.*, 1996; Matys, 2003; Qian *et al.*, 2006; Périer *et al.*, 2000; Stegmaier *et al.*, 2004), the transcription factor binding sites and DNA genome-wide binding information. The database has a relational architecture with two main tables namely, SITE and factor with an additional 50 other tables linked to each other

JASPAR is an open access database for eukaryotic transcription binding profiles (Sandelin *et al.*, 2004). According to Portales-Casamar *et al.* (2010), JASPAR is the leading open access database, which provides information on transcription factor binding

For a brief overview of ChIP-seq data that is available from ENCODE see Dunham *et al.* (2012). ChIP-seq allows one to do genome-wide analysis of transcription factor binding specificity (Hallikas *et al.*, 2006; Kaplan *et al.*, 2011). Numerous studies used ChIP-seq data from ENCODE (Encode Consortium, 2011) to identify transcription-binding sites (Zhang *et al.*, 2008; Chen *et al.*, 2012; Kulakovskiy *et al.*, 2013a). The main objective of these studies has been to identify and predict transcription binding sites, motifs, epigenetic marks (Thomas-Chollier *et al.*, 2012), transcription start sites and transcription binding specificity. ChIP-seq can be use to study evolutionary dynamics of transcription factor binding (Schmidt *et al.*, 2010).

In our study, one should be cognisant of the condensed state of the DNA. ChIP-seq data peaks indicate possible TF binding sites and binding to these sites occurred when the binding site was available or accessible which occurred generally occurred when the chromatin was in an open state. When looking at ChIP-seq data one must always be aware the ChIP-seq data might not represent all binding sites due to the inaccessibility of binding sites which might be due to the chromatin being in a closed or condensed state (Encode Consortium, 2011).

3.6 Motif based analysis of TF binding sites

Statistical computational methods used to obtain transcription factor binding site information.

These methods use ChIP-seq data and various algorithms to find motifs. The p-value is used to evaluate discovered motifs is accepted as a possible binding site or rejected as being not significant.

Below is a summary of the most common motif discovery tools.

MEME (Multiple Expectation Maximization for Motif Elicitation) used for discovering transcription factors binding sites as well as protein domains in a group of related DNA sequences or proteins (Bailey and Elkan, 1994). MEME is well known and accepted for its motif finding ability and by using MEME we can identify the motifs associated for a particular TF using ChIP-seq data (Bailey *et al.*, 2006).

Centrimo (Bailey and Machanick, 2012) is a statistical tool used to identify the distribution of the region of maximum enrichment within a set of binding regions. Centrimo output is a motif with an E-value, which indicates statistical significance of the centrally enriched region.

Centrimo can identify motifs that bind preferentially at or near the centre of the given binding region that occur at the peak or near the centre which indicates direct binding whilst motifs with no central bias which are located at the flanks of the peak normally indicate indirect or cooperative binding (Bailey and Machanick, 2012).

Centrimo will allow us to determine if the ChIP-seq data contain enriched motifs, binding specificity and directly or indirect TF binding.

Tomtom is part of the MEME suite and it iused to compare a known motif against a database of motifs (Gupta *et al.*, 2007).

Another useful tool is SpaMo, used to infer interactions between proteins (Whitington *et al.*, 2011). SpaMo takes a motif and searches a database of motifs to find neighbouring sites of enrichment.

diCHIPMunk (Kulakovskiy *et al.*, 2013b) is another tool which allows one to model transcription binding sites using ChIP-seq data. diCHIPMUNK transcription binding site model is based on an optimal dinucleotide position weighted matrices which takes into account correlating neighbouring nucleotides.

Another method uses the hidden Markov models (HMM)-based approach to model the dependence between adjacent nucleotide positions and show their method, which they called kmerHMM can also deduce multiple binding modes for a given TF. The authors claim the strength of the kmerHMM method is that it can distinguish distinct binding modes between a

DNA-binding protein and its target sequence, which could provide biological insights on the subtlety of the gene regulation (Wong *et al.*, 2013).

3.7 Cell Types

Cell types refer to the source of the DNA which particular tissue or blood used to obtain the cell to extract DNA. The ENCODE consortium divided the cell types into 3 tiers. The first tier has the highest priority and used first in experiments before the other tiers. Tier 2.5, the last tier has the lowest priority. Since tier 1 has highest priority, many data is available for cell types listed tier 1 (Rosenbloom *et al.*, 2013).

ENCODE displays the following information for each cell type.
Each cell has a name or identifier followed by a description of the cell followed by the lineage, the tissue of the cell. The karyotype specified indicating whether cell obtained is normal or cancerous. Information provided on the gender of the person of the cell, there is also information on the lab, and the specific method used to extract DNA.

If one is comparing transcription binding specificity for the same species but for different cell types one need to be aware of differences in transcription binding specificity across cell types. Studies of cell type specificities of transcription factor binding sites showed general binding differences increase as functional and evolutionary distances increase (Håndstad *et al.*, 2012). When comparing transcription-binding sites for various cell types Lee *et al.* (2012) showed that there is a variable overlap ranging from about 85% below 50%. The variable overlap as well as evolutionary differences in evolutionary distances between control cell type and abnormal cell type might cause a difference in the way the transcription factors binds to the different cell type.

We intend comparing the ChIP-seq peaks of different cell types to determine if there is a difference in distance between ChIP-seq peak and the TSS in a normal and abnormal cell type. We interested in finding out whether the distance between the TFBS and the TSS is a factor, which causes the difference between the normal and abnormal cell type. Cancer cells are known to multiply rapidly and the difference in metabolic state between the normal and cancerous cell can contribute difference in binding distribution.

We use the distance between the TFBS and the TSS in the normal non-cancerous cell type as a yardstick to which we compare the distance between the TFBS and the TSS in the cancerous cell type. We determining if the TFBS in the control is closer or further away from the TSS in the case.

3.8 Review studies which looked at TF-TSS binding distance

A number of studies looked at the TF-TSS binding distance. A study by Lee *et al.* (2012) compared the genome-wide transcription binding specificity of Myc, CTCF and RNA polymerase II. ChIP-seq data was used and to determine binding sites. The TSS was determined using the refseq gene annotation data from the UCSC genome browser. The Ensembl and Vega gene annotation set was used.

Once the TSS were determined, the data was divided into 3 regions based on the distance to the TSS. The first region defined as a distance ±2kb around TSS. This region named the promoter region. The second region, called upstream region that were defined as a distance of 2kb -20kb upstream from TSS. The last region named the intergenic region was defined a distance of > 20kb from TSS.

The result of the study indicated transcription factors positively correlated with gene density across the genome as well as a modest association between expression levels and increasing distance in the 2-20 kb region.

This study was very extensive and detailed but only looked at three transcription factors. Although 11 different cell types used, which included both normal and cancerous cell types the results of the study are not conclusive for binding distribution. Using three 3 TF and by looking at specific regions and clustering results gives a good overview of genome-wide TF binding but making specific conclusions is problematic. In addition, the study only looked at the upstream region for the regions greater than 2kb from TSS.

The study also looked at the occupancy of Myc, CTCF and RNA polymerase II however; the specific order of binding was not considered. They defined occupancy, as whether one, two or three of the TF was present.

Another study by Cheng *et al.* (2012) looked at transcriptional regulation by looking at TF-binding signals around TSS and the expression level of TSS measured by different technologies using 12 different cell types. This study tried to predict gene expression using the Distribution of Distance from ChIP-seq peaks to nearest TSS. This method used a neural network approach to predict gene expression using a region of 100 bp from the TSS.

The results of the study indicated gene expression could be calculated by using the distance of the TF from the TSS.

A major limitation of this study is the authors only looked at a narrow region of 100 base pairs around the TSS.

Another study by Wang *et al.* (2012) used 457 ChIP-seq data sets on 119 transcription factors in 72 cell types to characterise the sequence features of TF binding sites and to determine the local chromatin environment around them. MEME-ChIP used for motif discovery and 79 unique motifs were discovered. The ChIP-seq peaks of the TF were divided into those that were within 2kb (TSS-proximal) of TSS and those, which were > 2kb (TSS-distal) away from TSS. Motif finding performed separately on the top 500 TSS-distal and TSS-proximal peaks and consistent motifs found between the two sets. The result of the study found 79 unique motifs and from this, 67motifs were present in JASPAR or TRANSFAC. The remaining 12 motifs were highly significant and not present in JASPAR or TRANSFAC. They observed a general agreement between the strength of ChIP-seq signal and motif content in the same ChIP-seq data set. The study looked at various other issues besides motif discovery. The other factors, which they looked at, were:

Comparison of bound versus unbound motif sites.

Co-binding and tethered binding between different transcription factors.

Distance and orientation preference between the sites of co-binding transcription factors.

Cell type specific binding of sequence specific transcription factors.

Tethered binding of non-sequence specific transcription factors.

The ChIP-seq peaks of many transcription factors flanked by positioned nucleosomes.

Most transcription factors bind at GC-rich, nucleosome-depleted, and DNaseI accessible regions.

Chromatin structure around cell-line specific TF binding regions.

This study used motifs to characterise TF binding. They compared motifs and motif-to-peak distances for two regions namely a ± 150 bp and the flanking ± 300 bp regions. Unlike the previous studies, which looked specifically at the distribution of ChIP-seq peaks this study, focused mainly on motifs to investigate TF binding specificity.

A study by Tallack *et al.* (2012) used ChIP-seq data and mRNA expression data to investigate KLF1 binding. They also looked at the binding specificity of KLF1 relative to TSS. They found an overrepresentation of the CACC motif and CCAAT box. The distances between TSS and KLF1 ChIP-seq peaks for both KLF1 activated and repressed genes were calculated and compared to a control. The results for the repressed genes indicated the expression lowered and the ChIP-seq peak is closer to the TSS. The activated genes and the controls differed significantly but no mention was made by the authors on how this affected the KLF1-TSS distance. It is not clear how the authors obtained the controls and the method described when selecting samples for the controls indicated samples were not randomly selected but rather based on the Distribution of Distance from ChIP-seq peaks to nearest TSS. Generally, knock out genes are used in RNA expression studies as controls however knockout genes are not a suitable control for ChIP-seq studies because there is no ChIP-seq signal for the for the knocked-out genes.

Tallack *et al.* (2012) also looked at the KLF1 cooperation in vivo with GATA1, TAL1 and P300. They found KLF1 would bind with GATA1/TAL1 to recruit P300. They also compared the distance of ELK1-P300 to TSS with the distance of ELK1 to TSS and found that generally ELK1-P300 is closer to the TSS compared to ELK1 only. They do not mention which part of the ELK1-P300 complex is closest to TSS and whether the preferred binding distribution is upstream or downstream

Joshi et al. (2012) conducted the relationship between different types of breast cancer and the TSS. The authors looked at the promoter region, which they defined as the region, which has a distance -500bp to 100bp from the TSS They used the transcription factor-binding site (TFBS) matrix concept to study transcription factor binding in the promoter region. The TFBS matrix is a group of weighted matrices, which corresponds to the same family of transcription factors, or transcription factors, which has similar functions.

Genomatix

RegionMiner tool (Genomatix) used to calculate the TFBS family over represented motifs. The conclusion of the study was the promoter region for each cancer (or patient sub group) is distinct.

4 Methodology

4.1 Introduction

This chapter describes how we obtained the data and the methods used to calculate the TFBS and the TSS distance. We downloaded ChIP-seq data from ENCODE and added column headers to ChIP-seq files. ChIP-Enrich, which is a Bioconductor package, used to plot histograms of the distance between the TFBS and the TSS. We calculated the distance between the TFBS and the TSS separately for control and case.

4.2 Downloaded the ChIP-seq data from the ENCODE website

ChIP-seq data for various transcriptions factors for the GM12878 and K562 cell types downloaded. GM12878 is a B-lymphoblastoid (normal) cell type and K562 is chronic myelogenous/erythroleukemia (cancer) cell type. We downloaded ChIP-seq data for transcription factors, which had both GM12878, and K562 ChIP-seq data available. We also downloaded ChIP-seq data for transcription factors that were treated with interferon and dexamethasone. In some cases, no untreated ChIP-seq data is available but we included it in the study and we compared the treatment time or different doses to each other. The reason for including ChIP-seq data with no control was for interest sake and to see whether treatment caused a difference in distribution distance between the TFBS and the nearest TSS. We assumed since the data was obtained from ENCODE and from well-known established labs the ChIP-seq data had minimal errors.

All data downloaded from the following URL:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTFbsUniform/

4.3 Distribution of Distance from ChIP-seq peaks to the nearest TSS histogram

To calculate distance between the TFBS and the TSS we use ChIP-Enrich (Welch *et al.*, 2014), an R package which is part of Bioconductor. ChIP-Enrich has a function called plot_distance_to_tss, which calculates the distance between the TFBS and the nearest TSS for ChIP-seq peaks. The results summarized as a histogram plot, which shows the distribution of distance of ChIP-seq peaks to nearest TSS. One can use the original peak files was downloaded from ENCODE however if one is using a bed format file then the file must

have column headings. If using a 3-column bed format file then the column headings should be: “chrom”, “start” and “end”. If additional columns present in data set then one can name these columns with any name, the program only requires the first three columns to have the specified name of “chrom”, “start” and “end”.

The workflow for plotting a histogram using R and the ChIP-Enrich package is as follows:

1. Load chipenrich library
2. Load chipenrich data files
3. peak=read.table (file=path and name of input file,header=TRUE)
4. plot_dist_to_TSS (peaks=peak,genome ='hg19')

4.4 Conclusion

The package ChIP-Enrich is a straight forward and easy to use. The histogram output from ChIP-Enrich makes it easy to visually compare the genome wide distribution distance between the TFBS and the nearest TSS for both control and case distances. The histogram shows the distance between the TFBS and the TSS for various intervals, which makes it easier to compare and to see the percentage of the TFBS at a particular distance from the TSS.

5 Results

5.1 Introduction

In this chapter, we present the results for the data we analysed. The results of ChIP-Enrich is a histogram plot, which shows the distribution of distance from ChIP-seq peaks to nearest TSS. Although we are using a case-control methodology we have included ChIP-seq data where there is no control ChIP-seq data available, for example, K562-stat1 treated with interferon for 30min and 6 hrs. In this scenario, we do not have untreated K562-stat1 ChIP-seq data therefore; we compare the 30min and 6hrs ChIP-seq data to each other. We include these data for interest sake and to see if there is a notable difference between the distance of the TFBS and the TSS for the 30min and 6hr interferon treatment.

5.2 Gm12878-cMyc and K562-cMyc

5.2.1 Number of ChIP-seq peaks

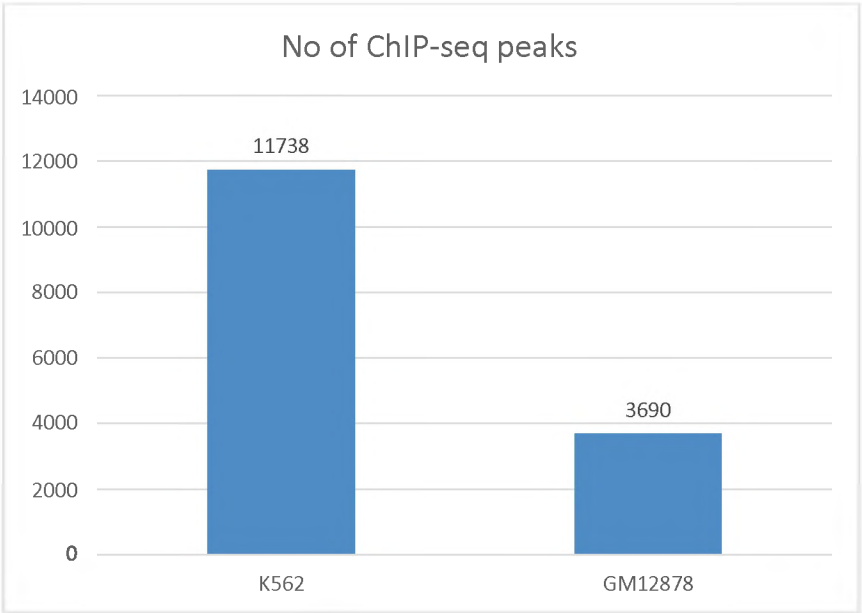


Figure 5-1 GM12878-cMyc (Control) and K562-cMyc (Case)

The values in figure 5-1 indicate there are almost 3 times more cMyc binding sites in the K562 cell type as compared to the GM12878 cell type.

5.2.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

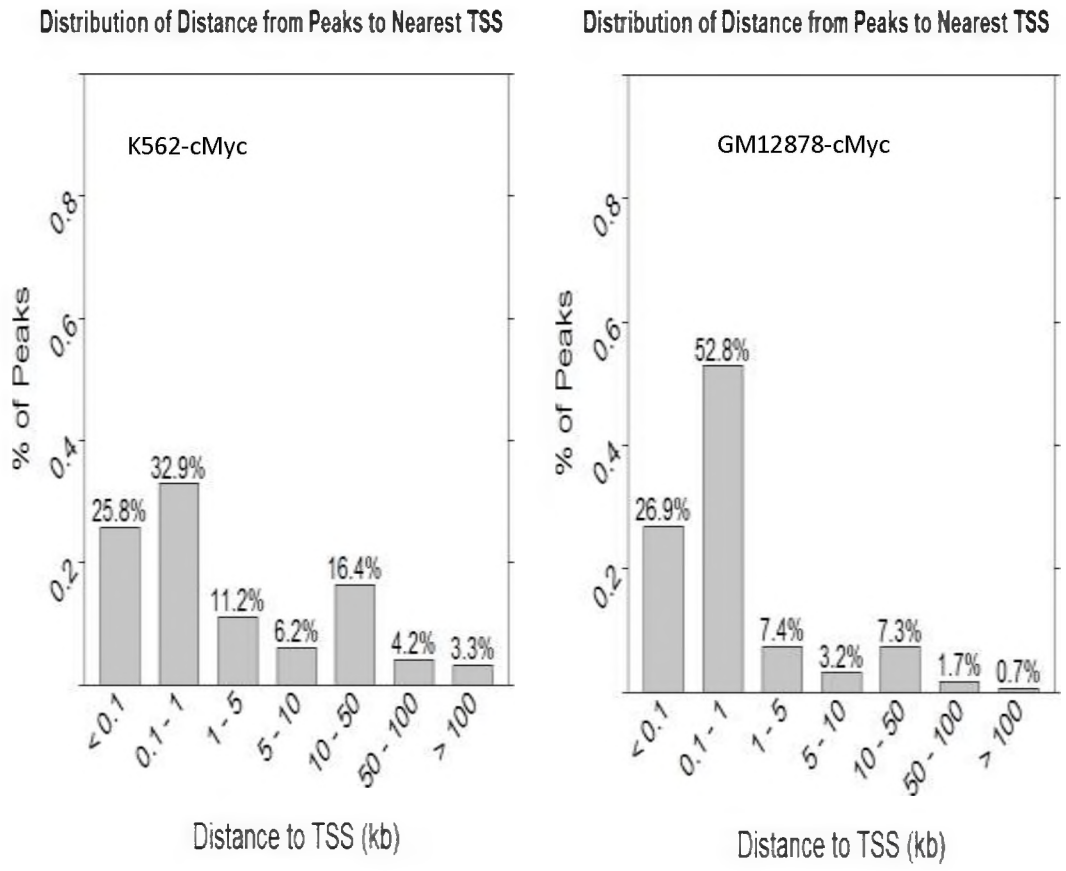


Figure 5-2 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Uta.

Figure 5-2 indicates GM12878-cMyc has a higher percentage of the TFBS in the region where distance between the TFBS and the TSS is less than 1kb compared to K562-cMyc. The difference between GM12878 and K562 is significant in the 0.1 to 1kb interval. Where the distance between the TFBS and the TSS is greater than 1kb, GM12878-cMyc has a lesser percent of the TFBS when compared to K562-cMyc. We conclude, cMyc in GM12872 has a large number of the TFBS close to the TSS and K562 has more TFBS further away from the TSS.

5.3 Gm12878-CTCF and K562-CTCF

5.3.1 Number of ChIP-seq peaks

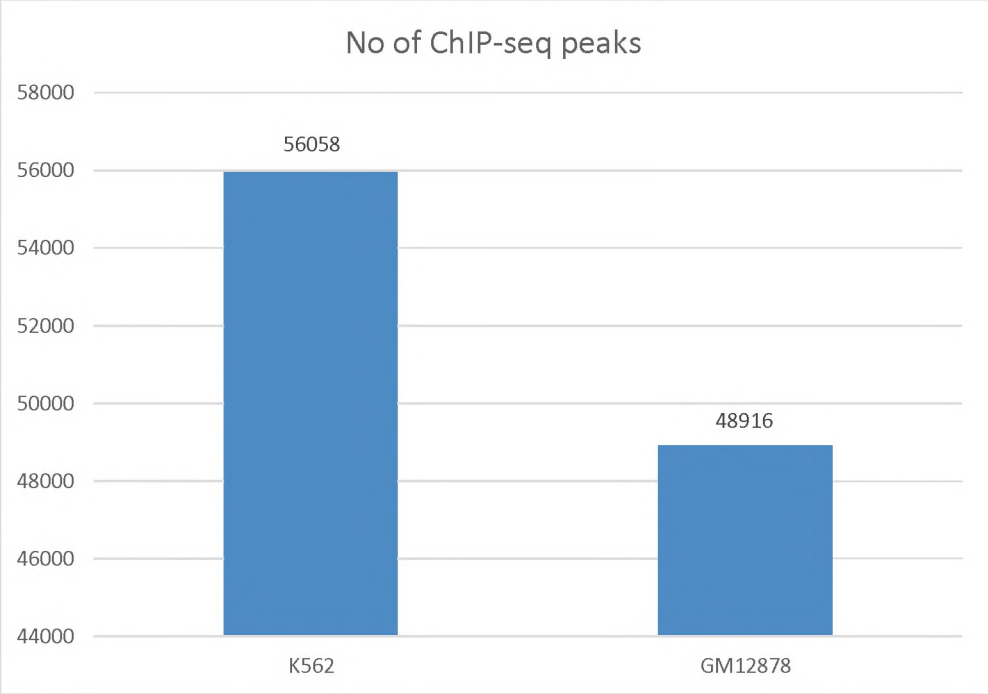


Figure 5-3 GM12878-CTCF (Control) and K562-CTCF (Case)

Figure 5-3 shows K562-CTCF has slightly more ChIP-seq peaks compared to GM12878-CTCF

5.3.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

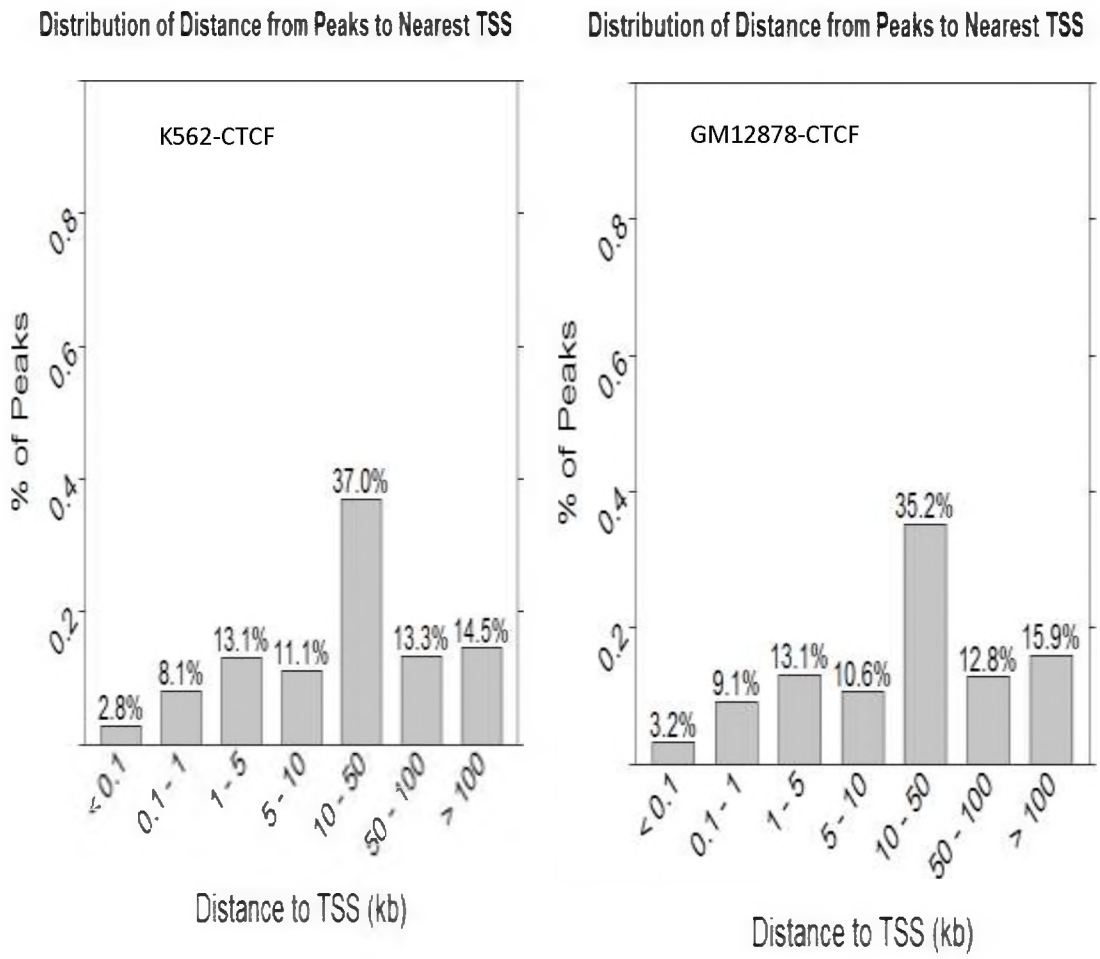


Figure 5-4 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Uta.

Figure 5-4 indicates for both GM12878-CTCF and K562-CTCF the distribution of distance from ChIP-seq peak to the nearest TSS is very similar.

5.4 K562-cJun treated with interferon alpha

We investigate the binding distribution of cJun in the K562 cell type, which was treated with interferon (Ifn) alpha for 30minutes and 6hr respectively. The distances between K562- cJun and the TSS that were not treated with interferon was used as the control and distances between K562-cJun and the TSS treated with interferon was used as case.

5.4.1 Number of ChIP-seq peaks

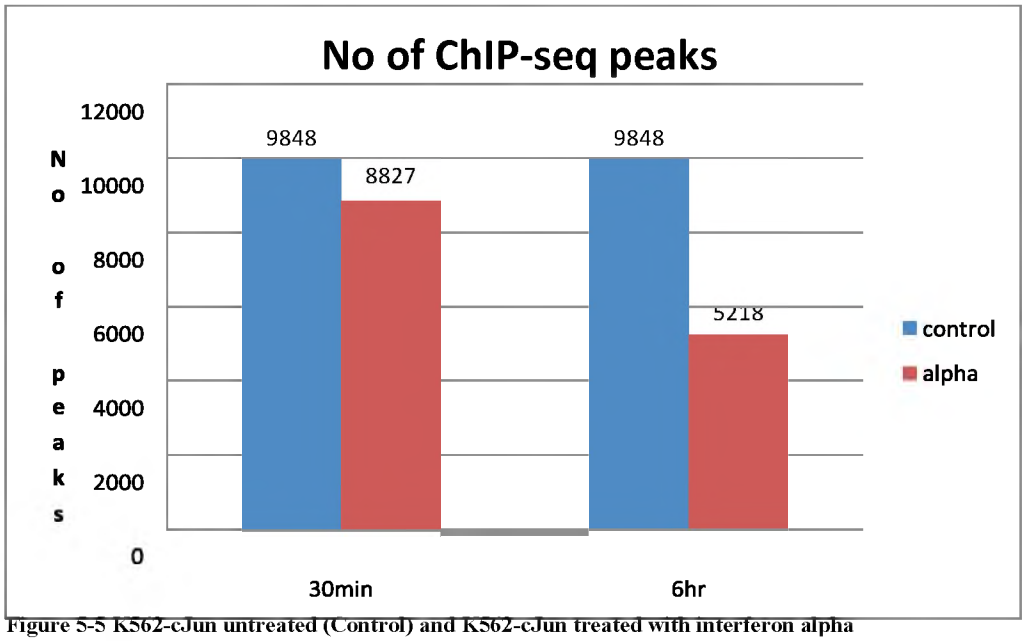


Figure 5-5 K562-cJun untreated (Control) and K562-cJun treated with interferon alpha for 30min and 6hrs (Case)

Figure 5-5 indicates K562-cJun treated with interferon alpha for 6hrs have fewer ChIP-seq peaks when compared to the untreated K562-cJun. The untreated K562-cJun cell type have almost the same number of ChIP-seq peaks as the K562-cJun treated for 30min.

5.4.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

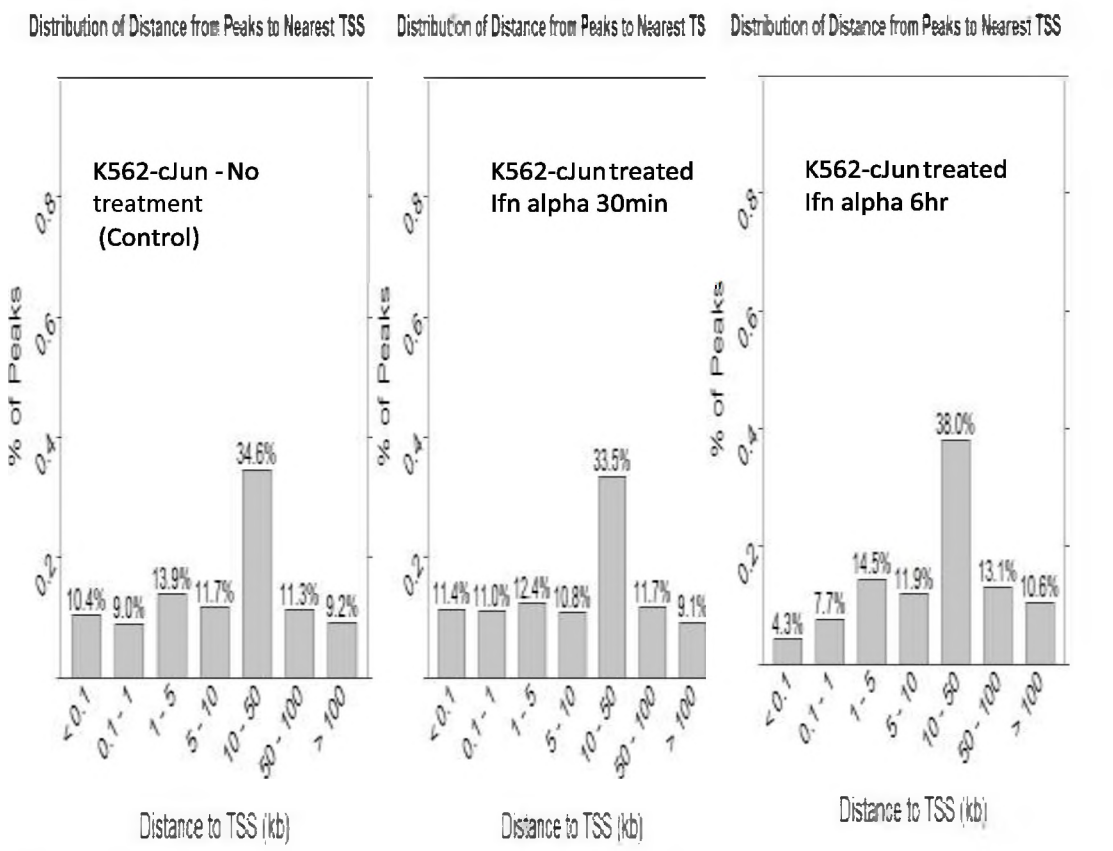


Figure 5-6 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-6 indicates K562-cJun, which is untreated, and the K562-cJun treated with interferon alpha for 30 min have a similar distribution of distance from the cJun peaks to the nearest TSS. When the distance between the cJun site and the TSS is less than 1kb, K562-cJun treated with interferon alpha for 6hrs has a lesser percent of peaks when compared to the untreated K562-cJun. The K562-cJun treated with interferon alpha has a greater percent of ChIP-seq peaks when the distance between the cJun site and the TSS is greater than 1kb when compared to the untreated K562-cJun.

5.5 K562-cJun treated with interferon gamma

We investigate the binding distribution of cJun in the K562 cell type, which has been treated with interferon gamma for 30minutes and 6hr respectively. The distance between the cJun site and the TSS for K562-cJun that were not treated with interferon gamma was used as the control. The distance between K562-cJun and the TSS, which has been treated with interferon gamma, was used as case.

5.5.1 Number of ChIP-seq peaks

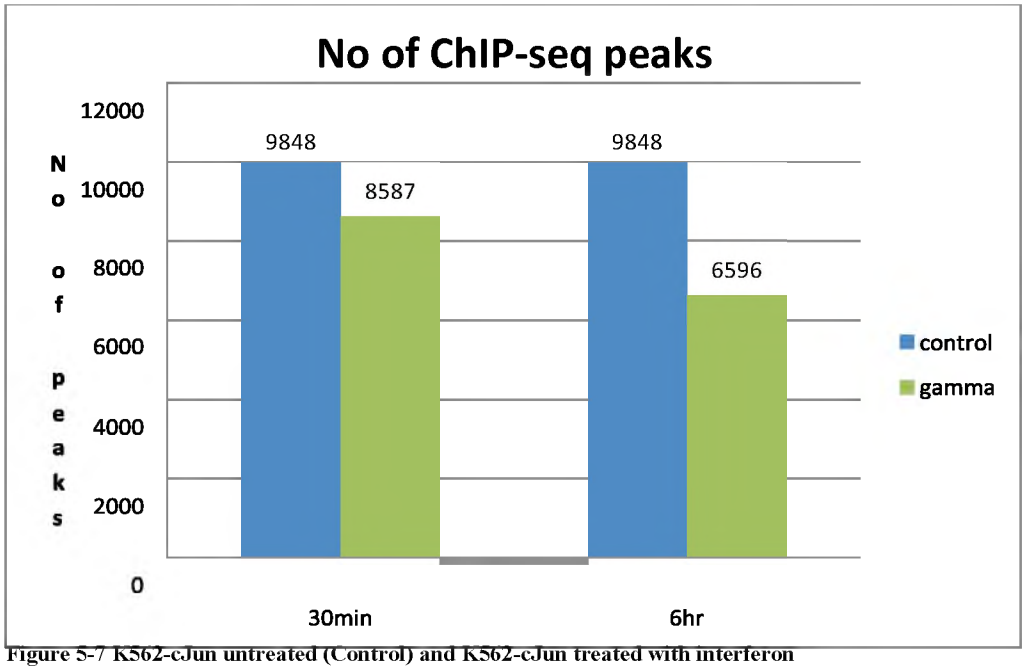


Figure 5-7 K562-cJun untreated (Control) and K562-cJun treated with interferon gamma for 30min and 6hrs (Case)

Figure 5-7 indicates K562-cJun treated with interferon gamma for 30mins have a similar ChIP-seq distance distribution as the untreated K562-cJun. The K562-cJun 6hr gamma interferon treated cells have a fewer number of ChIP-seq peaks when compared to the untreated K562-cJun.

5.5.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

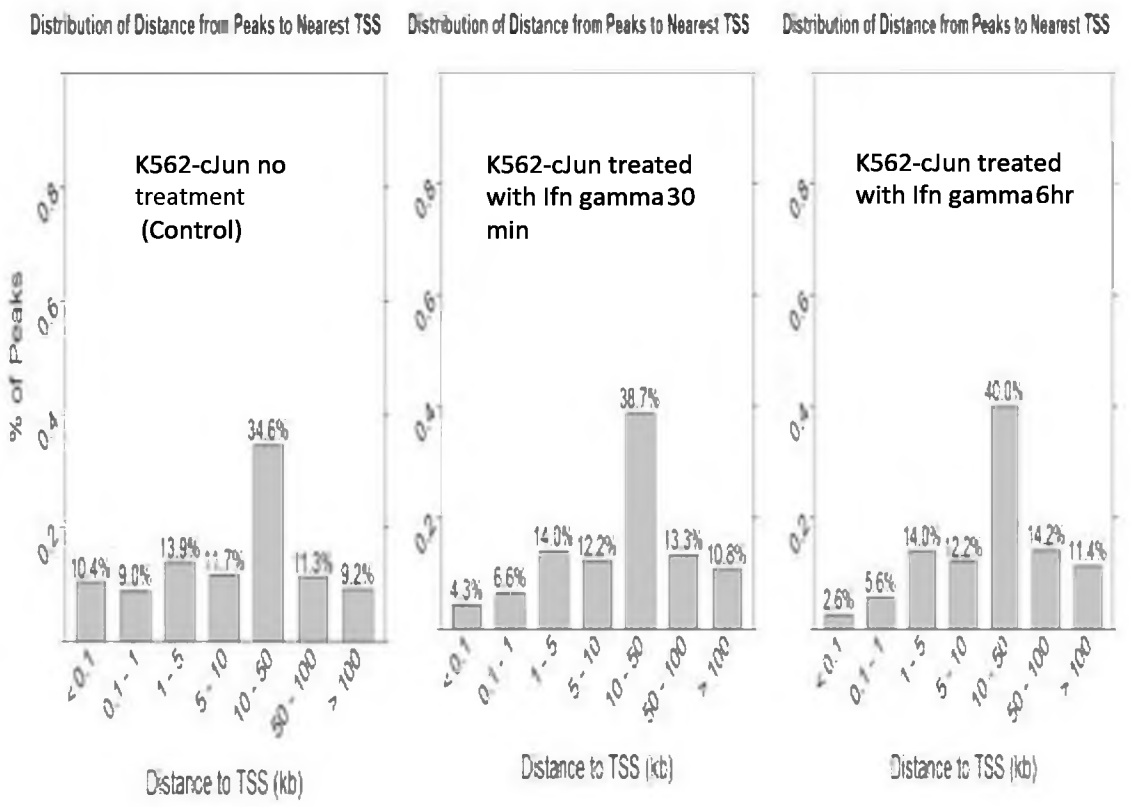


Figure 5-8 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-8 indicates the K562-cJun treated for 30min have fewer TFBS to the TSS when the distance between cJun site and the nearest gene is less than 1kb. K562-cJun treated with interferon gamma for 6hrs have significant fewer TFBS when the distance between cJun site and nearest gene is less than 1kb. When the distance between cJun site and the nearest TSS is greater than 1kb one finds more cJun binding sites further away from a TSS with increase treatment time.

5.6 K562-cMyc treated with Ifn alpha for 30min and 6hrs

5.6.1 Number of ChIP-seq peaks

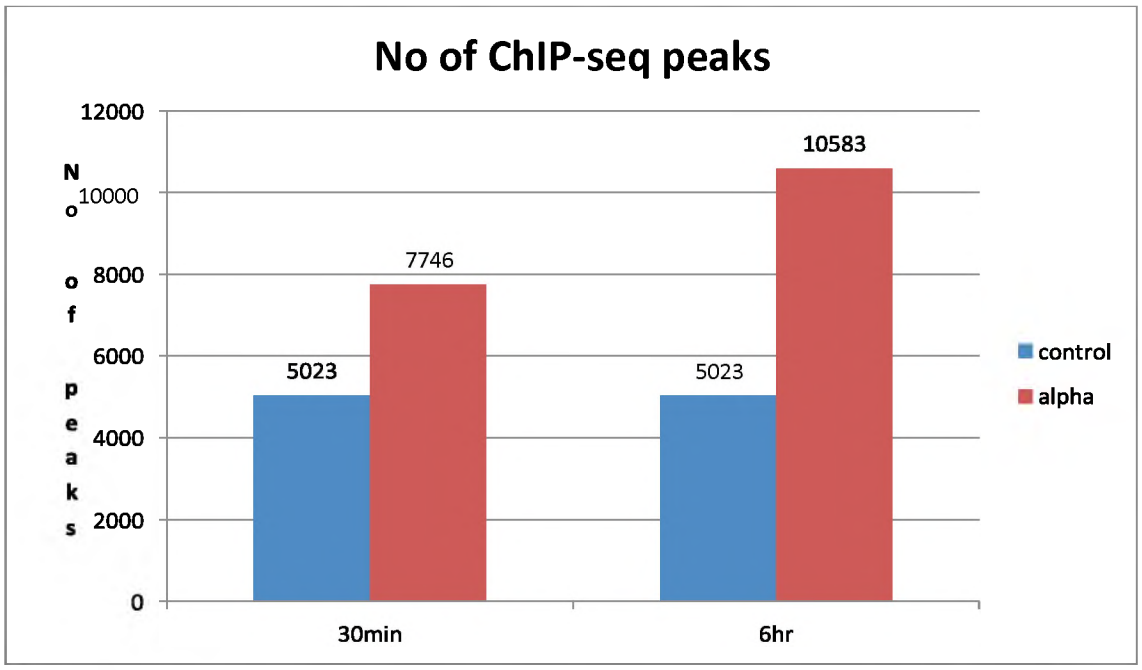


Figure 5-9 K562-cMyc untreated (Control) and K562-cMyc treated with Ifn alpha for 30 min and 6 hrs (Case)

Figure 5-9 indicates the number of ChIP-seq peaks increases with increase interferon treatment time when compared to the untreated K562-cMyc. K562-cMyc treated for 6hrs with interferon alpha has double the amount of ChIP-seq peaks when compared to the untreated K562-cMyc.

5.6.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

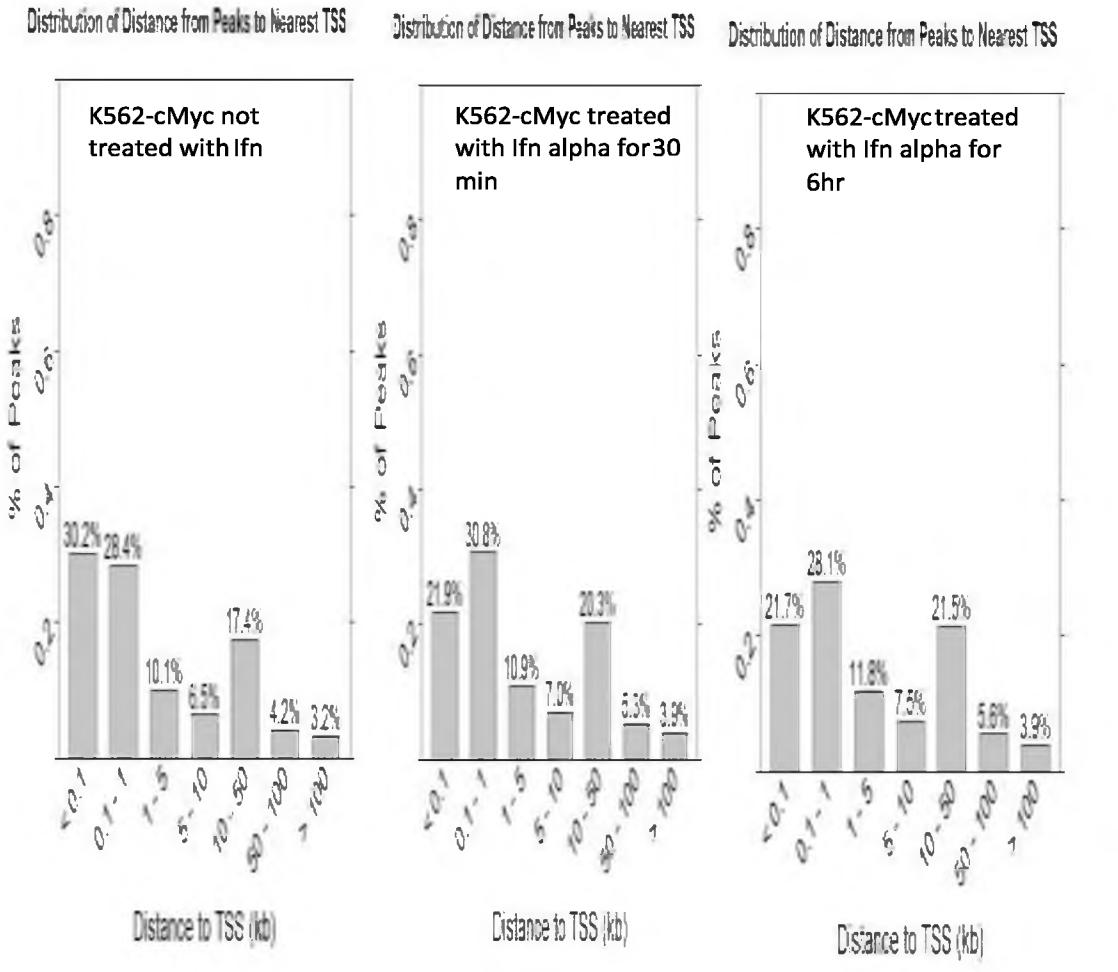


Figure 5-10 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

In figure 5-10, we see the untreated K562-cMyc has a higher percentage of the TFBS when the distance between the TFBS and nearest gene is less than 0.1kb when compared to the 30min and 6hrs interferon alpha treated K562-cMyc. When the distance between the TFBS and the TSS > 0.1kb the distribution of distance of cMYc to the nearest TSS for untreated K562-cMyc is similar to both 30min and 6hrs interferon alpha K562-cMyc treated

5.7 K562-cMyc treated with Ifn gamma for 30min and 6hrs

5.7.1 Number of ChIP-seq peaks

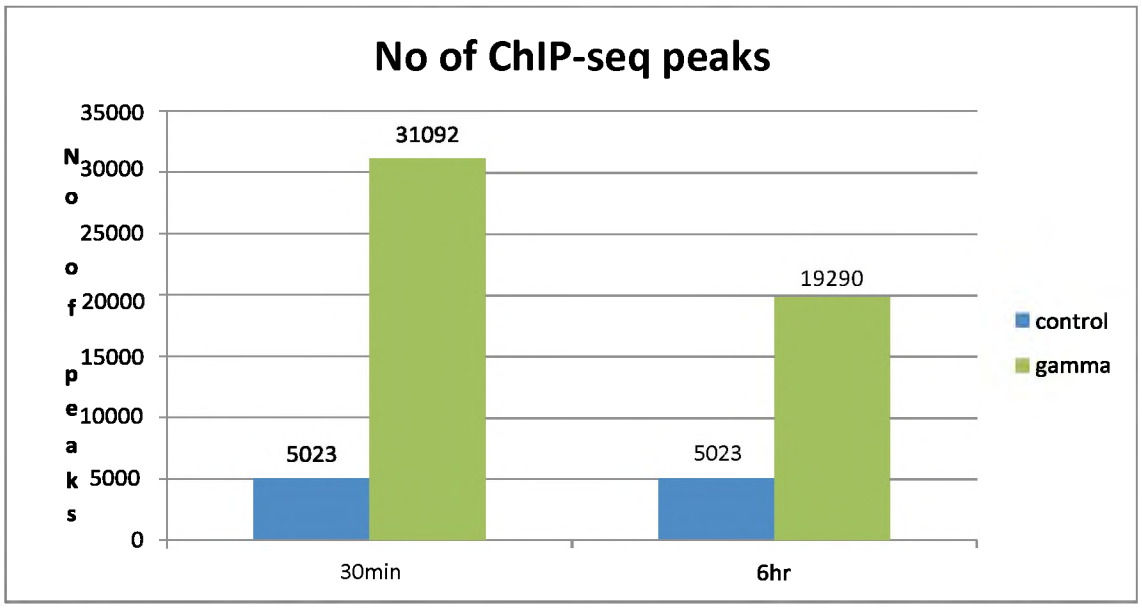


Figure 5-11 K562-cMyc untreated (Control) and K562-cMyc treated with Ifn gamma for 30min and 6hrs (Case)

Figure 5-11 indicates K562-cMyc interferon gamma treated have more TFBS compared to the K562-cMyc which did not receive any treatment. The amount of ChIP-seq peaks decrease with increase treatment time.

5.7.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

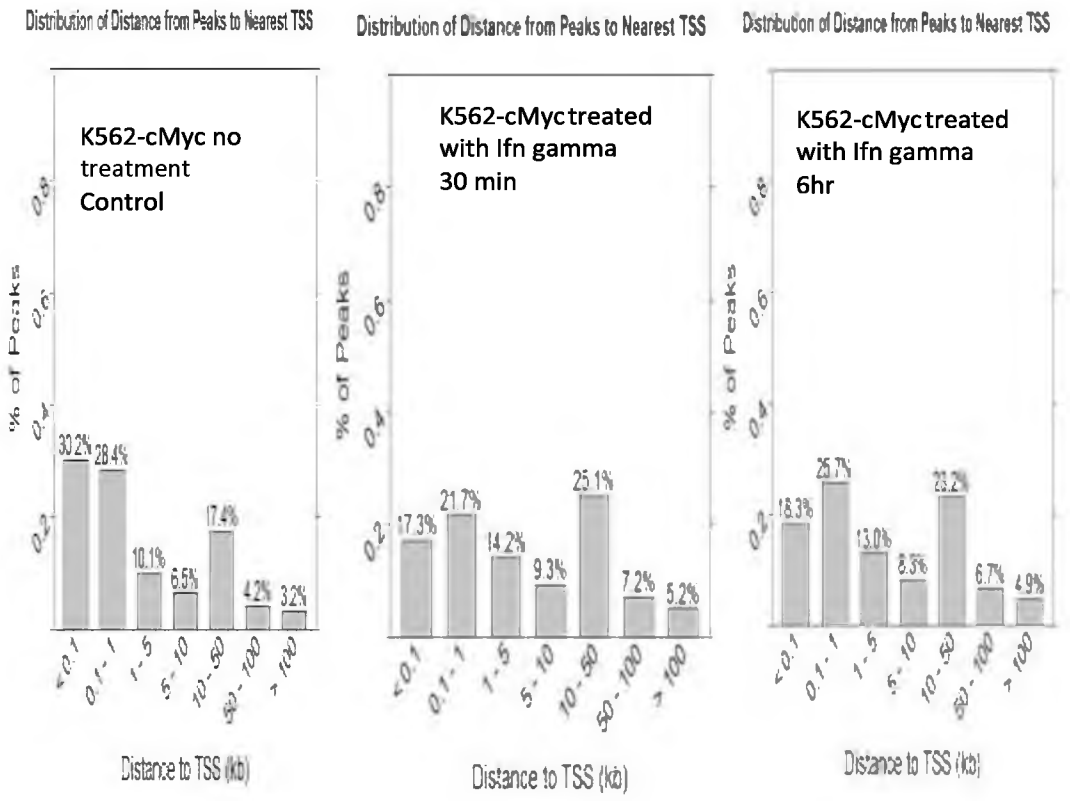


Figure 5-12 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-12 indicates when the distance between cMyc and the TSS is less than 1kb, the untreated K562-cMyc has a higher percentage of ChIP-seq peaks than the 30min and 6hrs interferon gamma treated K562-cMyc. When the distance between cMyc and the TSS of nearest gene is greater than 1kb, K562-cMyc treated with interferon gamma has a higher percentage of the TFBS when compared to the untreated K562-cMyc.

5.8 K562-Irf1 treated with Ifn alpha and gamma 30min and 6hrs

5.8.1 Number of ChIP-seq peaks

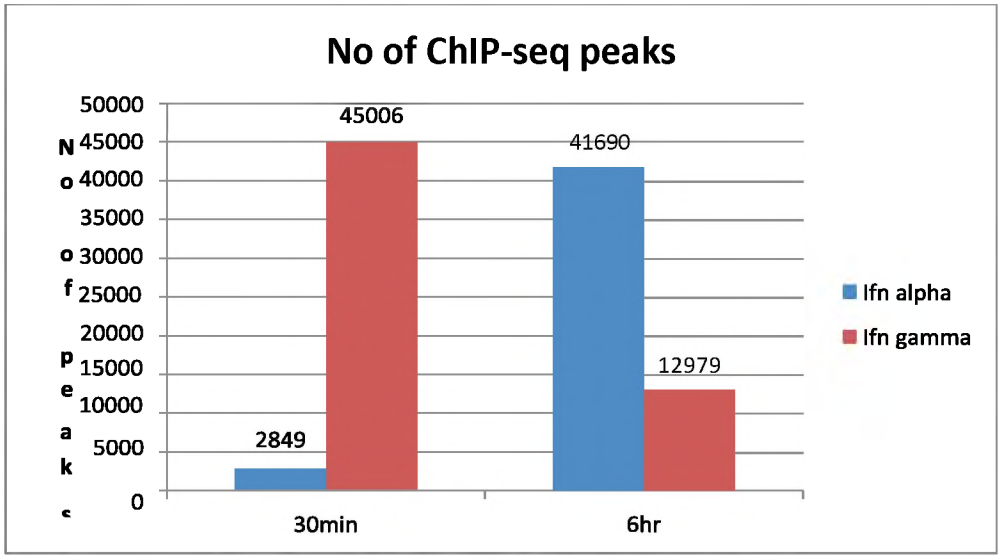


Figure 5-13 K562-Irf1 treated with Ifn alpha and gamma for 30min and 6hrs

Figure 5-13 indicates an inverse relation between K562-Irf1 alpha and K562-Irf1 gamma treatment. The increase interferon alpha treatment time for K562-Irf1 cause's increase in the number of ChIP-seq peaks and with interferon, gamma treatment the increase treatment time causes a decrease in the number of ChIP-seq peaks.

5.8.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

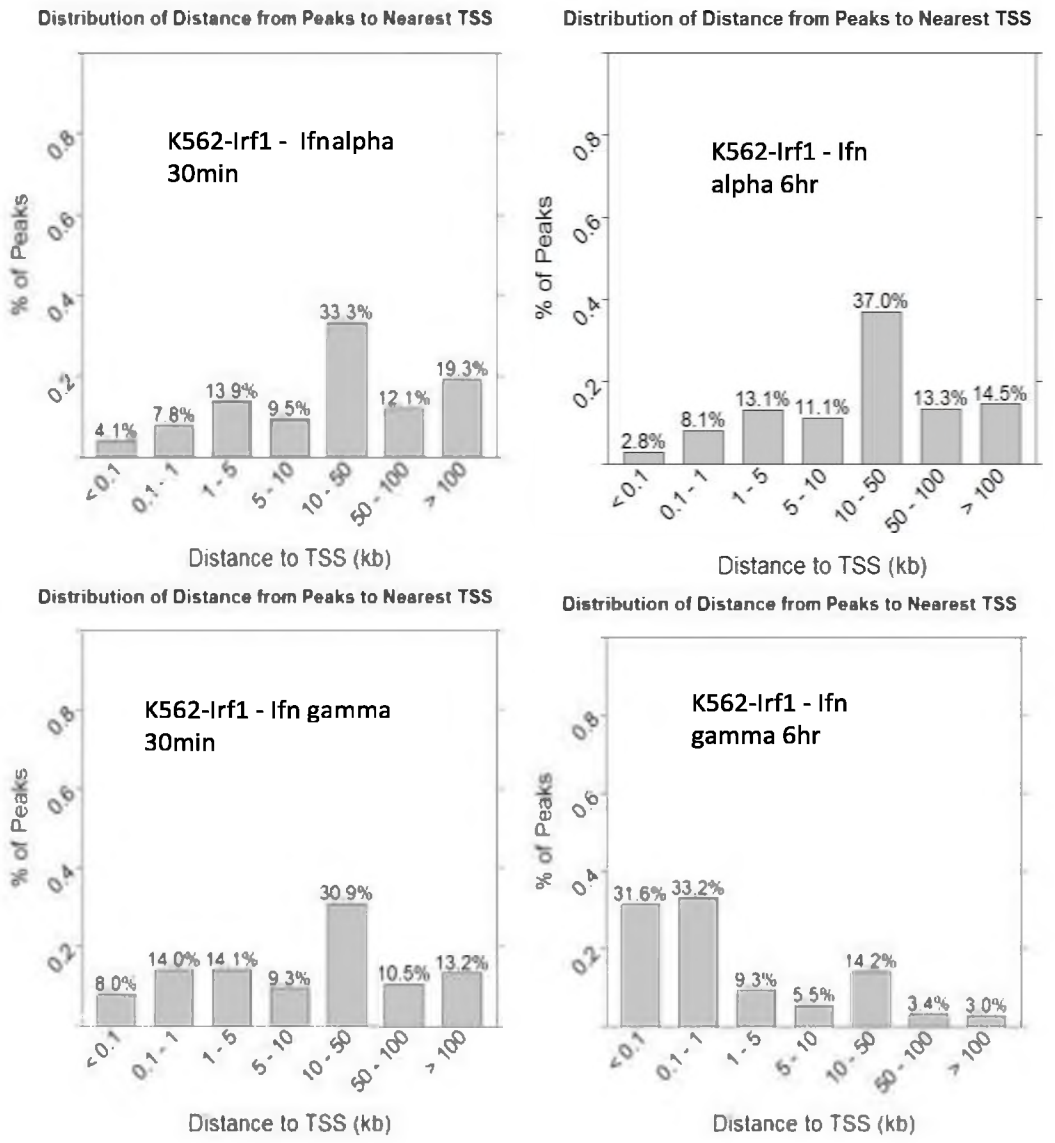


Figure 5-14 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-14 indicates there is not much difference between the distributions of distance from Irf1 to the nearest TSS for the K562-Irf1 interferon alpha treated. There is a significant difference in distribution of distance of Irf1 to the nearest TSS for K562-Irf1 treated with interferon gamma. When the distance between Irf1 and the nearest TSS is less than 1kb, K562-Irf1 treated for 6hrs with interferon gamma has a higher percentage of ChIP-seq peaks in this region, when compared to K562-Irf1 30min interferon gamma treated. When distance between Irf1 and nearest gene is between 10 and 50kb the 30min interferon gamma, treated cells have more TFBS in this region when compared to the 6hrs. The distribution distance for

the 30min and 6hrs alpha interferon treatment and the 30min gamma interferon treatment looks similar.

5.9 A549-Gr treated with Dexamethasone (dex)

5.9.1 Number of ChIP-seq peaks

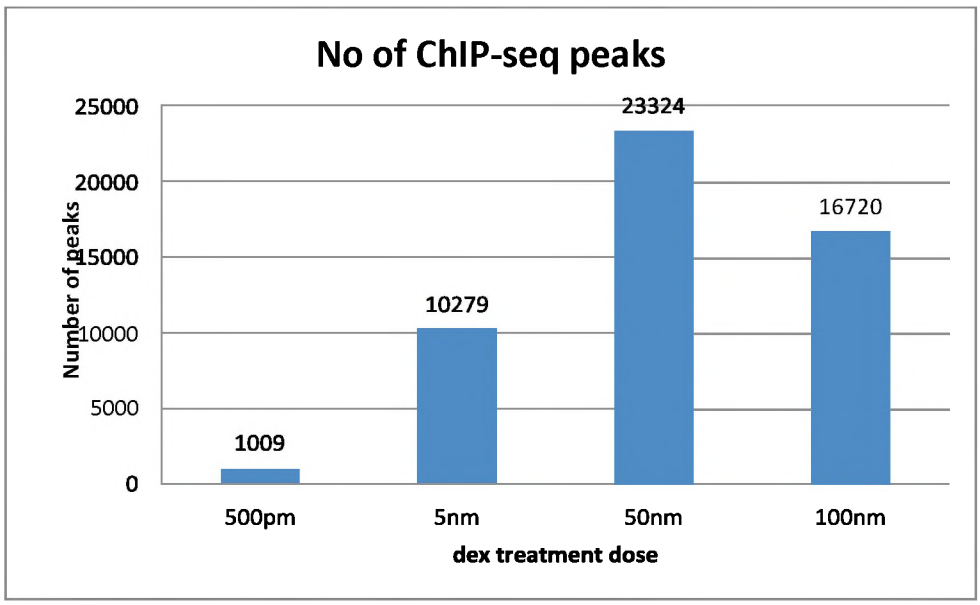


Figure 5-15 A-549 treated with various doses of dexamethasone

Figure 5-15 shows as the dose increase the amount of TFBS increase and then after the 50nm dose the TFBS decrease which could mean after 50nm saturation is reached and the increase dose no longer have any effect.

5.9.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

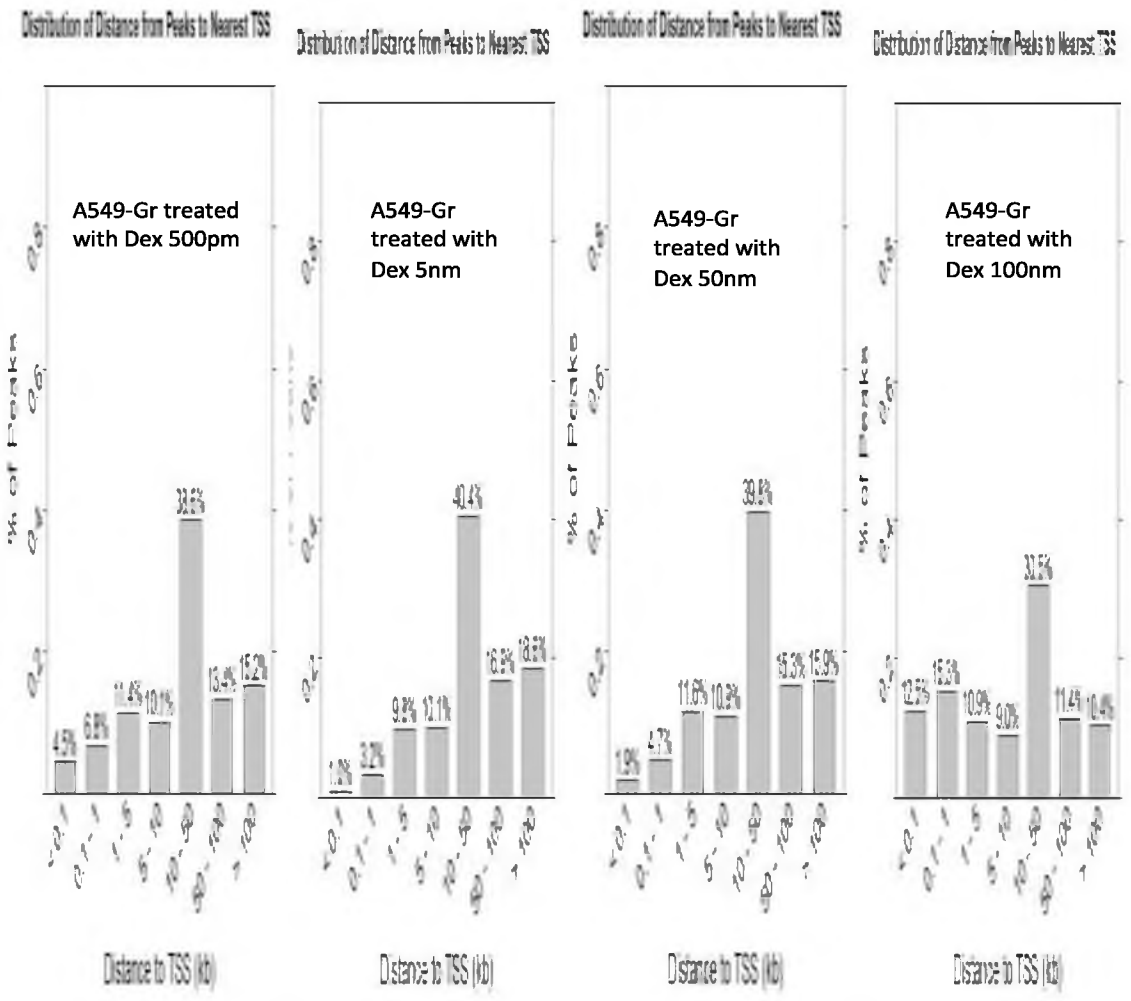


Figure 5-16 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Haib.

Figure 5-16 shows as the dose of dexamethasone increase more TFBS occur further away from the TSS with A549-Gr treated with dexamethasone 100nm being the exception. A549-Gr treated with dexamethasone 100nm has a higher percentage of the TFBS closer to a TSS within a distance of 1kb compared to the others. Interestingly most of the TFBS occurs far away from a TSS at a distance of greater than 5Kb.

5.10 K562-stat1 treated with Ifn alpha and gamma

5.10.1 Number of ChIP-seq peaks

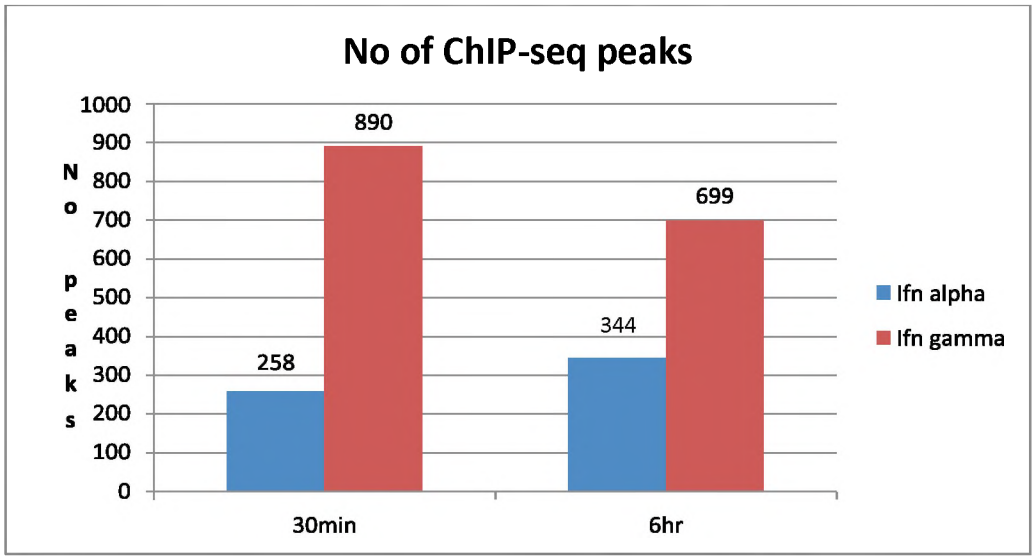


Figure 5-17 K562-stat1 treated with Ifn alpha and gamma for 30min and 6hrs

Figure 5-17 shows there is not much difference between the 30min and 6hrs interferon treatment. The number of ChIP-seq peaks are small for both alpha and gamma interferon treatment.

5.10.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

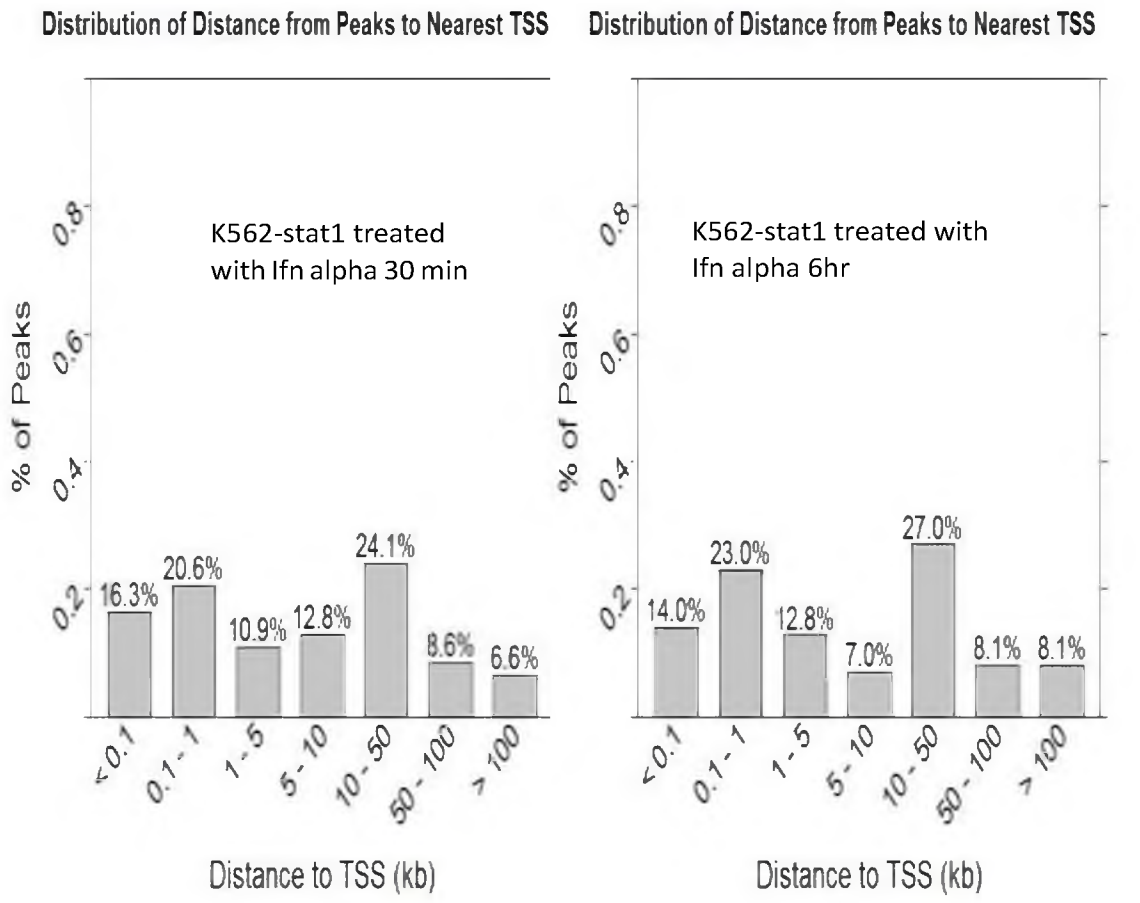


Figure 5-18 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-18 shows similar distribution of distance between ChIP-seq peaks and the nearest TSS for both K562-stat1 treated with interferon alpha for 30min and 6hrs.

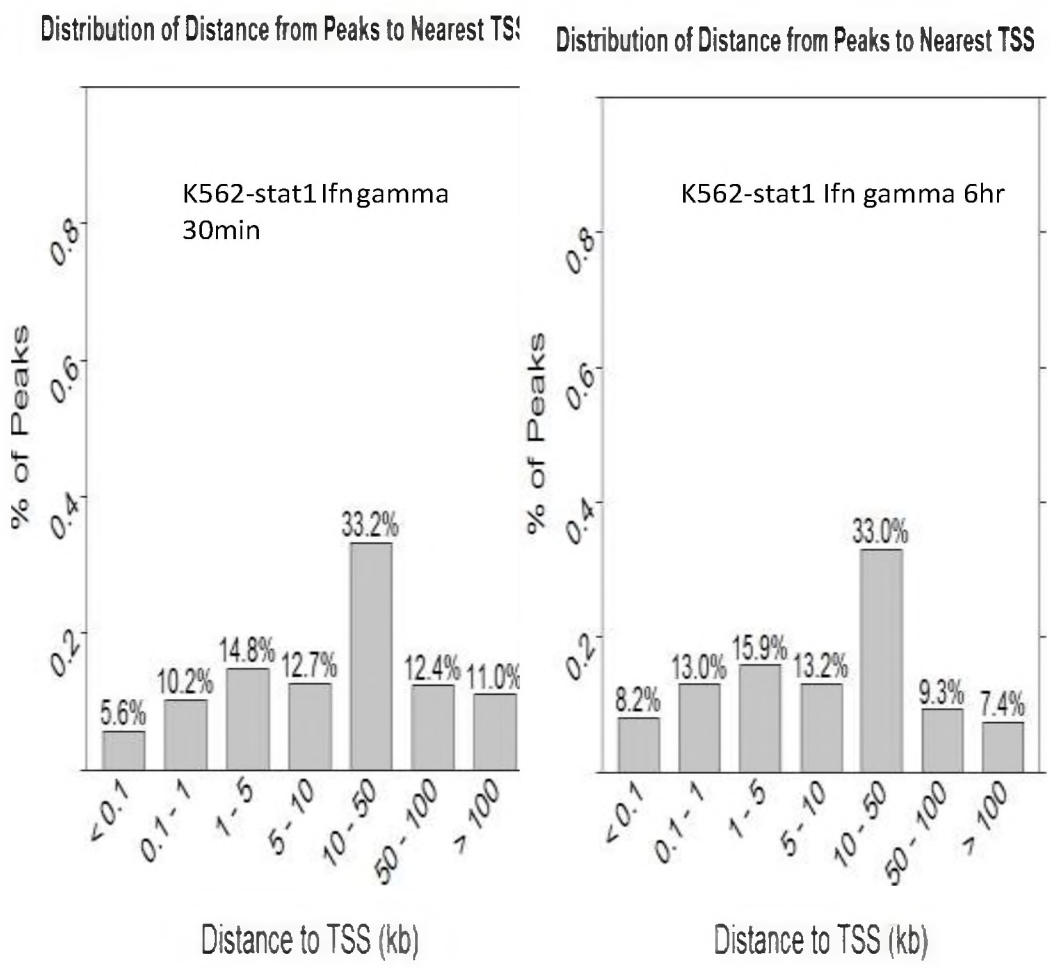


Figure 5-19 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-19 shows similar distribution of distance between ChIP-seq peaks and the nearest TSS for both K562-stat1 treated with interferon alpha for 30min and 6hrs.

5.11 K562-stat2 treated with Ifn alpha

5.11.1 Number of ChIP-seq peaks

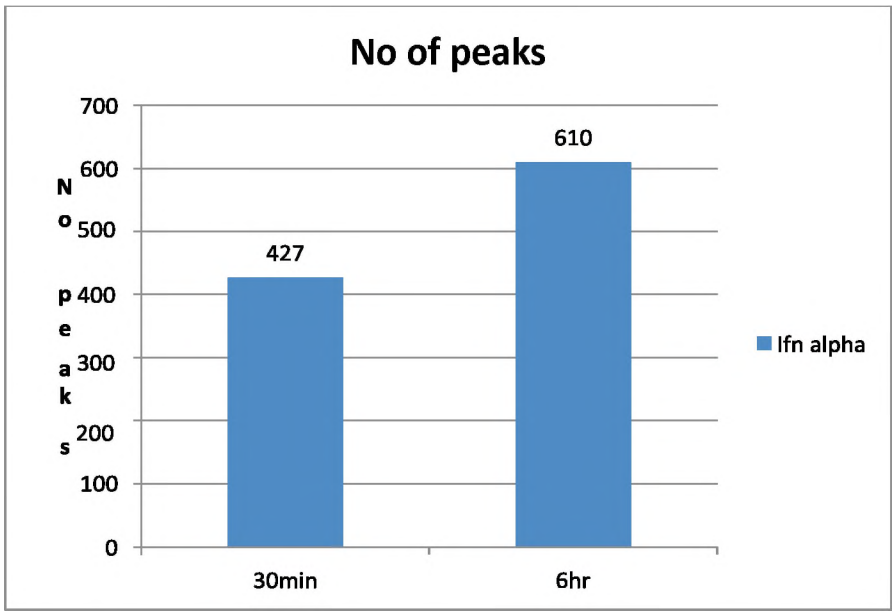


Figure 5-20 K562-stat2 treated with Ifn alpha for 30min and 6hrs

Figure 5-20 shows increasing the treatment time of interferon alpha causes an increase in the number of ChIP-seq peaks. The number of ChIP-seq peaks is relatively small the increase peaks with increase treatment time might not be significant.

5.11.2 Distribution of Distance from ChIP-seq peaks to the nearest TSS

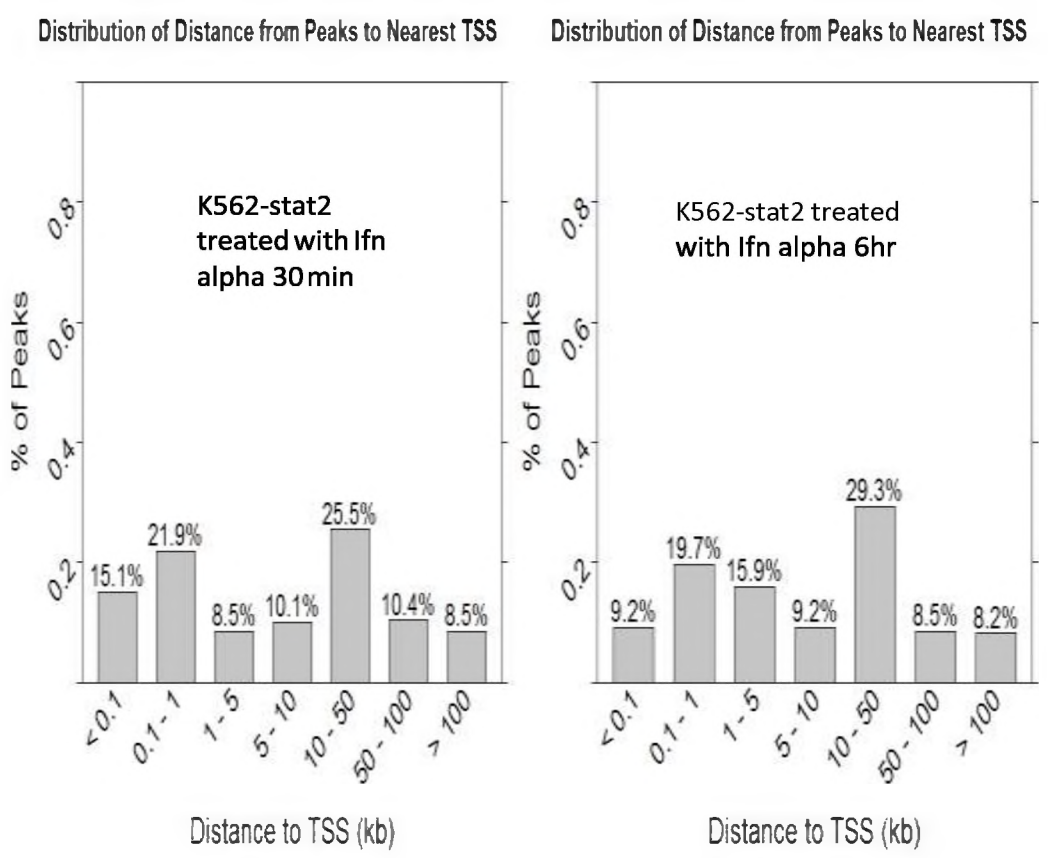


Figure 5-21 distance between the TFBS and the TSS. Percentage binding in the various regions. Data from Sydh.

Figure 5-21 shows the distribution of distance between the 30min and 6hrs is similar.

6 Discussion

6.1 Introduction

In this chapter, we discuss and analyse the results obtained from chapter 5. As we are a top-down approach, we now try and make sense of the results we obtained in the previous chapter. We reviewed the literature specifically looking at the distance the TF binds from the TSS and provide a brief overview of the various transcription factors. Finally, we compared our results with results obtained from the literature however, as mentioned before none of the studies used a case-control approach but where results were available for some of the cell types and transcription factors we used it.

6.2 Gm12878-cMyc and K562-cMyc

6.2.1 Background

Myc, formerly known as cMyc is a transcription factor which is known to be involved in cell differentiation, apoptosis, progression ,cell growth, chromosomal instability, DNA repair and metabolism (Dang, 2012; Gardner *et al.*, 2002; Lee *et al.*, 2012). Myc plays a role in tumour initiation as well tumour maintenance.

Knockout of Myc in cancer cells caused a reduction in proliferation and in some instances, apoptosis was induced. Myc has been linked to T-cell leukaemia via the Notch1 pathway. The Myc protoncogene is regulated both upstream by the far upstream element (FUSE) and downstream by TCF.

According to Lee *et al.* (2012), Myc had 45 % to 75 % of its binding sites within a distance of 2KB around the TSS and Myc found in promoter rich CpG islands, which indicates Myc regulates genes, by binding to proximal promoters to a diverse group of gene targets across various cell types. Myc also had an overrepresentation of bidirectional promoters in the 2kb interval around the TSS. Myc when bound to the bidirectional promoter will activate both genes irrespective of the distance of the Myc binding from the TSS.

Myc caused an increased in target gene expression levels when binding occurred upstream only and Myc binding did not bind to the promoter region of target gene. Myc binding at both promoter and upstream region of target gene resulted in decreased expression level of target gene.

6.2.2 Analysis

From figure 5.2, we see at a distance of less than 1kb away from a TSS we find more GM12878-cMyc peaks when compared to K562-cMyc. GM12878-cMyc have about 75% of its TFBS within a distance of 1kb from a TSS and K562-cMyc have about 56% of its TFBS in the same region. K562-cMyc have more TFBS where distance between the TFBS and the TSS is greater than 1kb when compared to GM12878-cMyc.

Also of interest is K562-cMyc has about 3 times more ChIP-seq peaks compared to GM12878-cMyc as shown in figure 5-1.

Our results for the binding sites around 1kb region around TSS are in line with Lee *et al.* (2012) which found 45% to 75 % of cMyc binding sites within a distance of 2kb from the TSS.

6.3 Gm12878-CTCF and K562-CTCF

6.3.1 Background

cMyc and CTCF have similar function i.e. cell growth and development and chromatin organisation (Macquarrie *et al.*, 2011).

According to Filippova *et al.* (2002), CTCF binds to the cMyc promoter region that is 2kb upstream from the TSS.

It is known human CTCF can bind to 4 different promoter regions of cMyc namely 2kb and 0.6 kb from a TSS upstream and two promoter regions downstream from a TSS (Fiorentino and Giordano, 2012).

The cMyc-CTCF interaction can be summed up as follows. The expression of cMyc activates CTCF. Once CTCF activates, it will negatively regulate cMyc. CTCF binds to the promoter region of cMyc, which restrict the expression of cMyc, and if CTCF binds to other regions excluding promoter region then CTCF will increase cMyc gene expression.

6.3.2 Analysis

Figure 5-4 indicates GM12878-CTCF and K562-CTCF have similar distribution of distance from ChIP-seq peaks to nearest TSS.

GM12878-CTCF and K562-CTCF have more than 50% of its binding sites at a distance greater than 10kb form the TSS. In the 1kb distance from the TSS both GM12878-CTCF and K562-CTCF, have approximately 10% of their binding sites. GM-CTCF binds to cMyc promoter region therefore we need to look at CTCF in conjunction with cMyc.

The number of ChIP-seq peaks for GM12878-CTCF and K562-CTCF as shown in figure 5-3 is very similar (56058 vs. 48916).

CTCF activated by cMyc and negatively regulates cMyc; we need to look at CTCF and cMyc.

K562-CTCF has approximately 5 times more ChIP-seq peaks compared to K562-cMyc. GM12878-CTCF has approximately 13 times more ChIP-seq peaks when compared to GM12878-cMyc.

cMyc has more than 50% of its binding sites within 1kb from TSS while CTCF has approximately 10%. CTCF has more binding sites compared to cMyc we can assume that CTCF will bind to cMyc in the promoter region as indicated by (Fiorentino and Giordano, 2012).

Of interest is the 50% CTCF binding sites that occur at a distance of greater than 10kb from the TSS. cMyc does not have such a large number of binding sites at a distance greater than 10kb from the TSS.

6.4 Interferon, JAK, STAT, cJun and Irf1

6.4.1 Background

Interferon is a member of the cytokines family and involved in anti-viral cellular response and with cell growth. Interferon is classified into two categories namely Type I and Type II. Interfon alpha (Ifn α) is classified as a type I interferon and interferon gamma (Ifn γ) is a type II interferon (Vera *et al.*, 2011; Pitha, 1998).

Ifn α is used for the treatment of various cancers especially blood related cancers and Ifn γ is used to treat acute infections.

The difference between type I and type II interferon is the particular cell receptor to which the interferon binds. Ifn α binds to the interferon alpha receptor and Ifn γ binds to the interferon gamma receptor. Interferon alpha and gamma receptor have subunits which interact with a specific member of Janus activated kinase (JAK).

Type I interferon cell receptor units will bind to either Tyrosine Kinase-2 or JAK1 and type II interferon cell receptors will bind to JAK1 and JAK2. The binding of interferon to cell receptors signals the phosphorylation and activation of the applicable JAK member.

Once JAK has been activated this signals the activation and phosphorylation of various STAT (Signal Transducers and Activators of Transcription). The STAT form dimers, which initiate transcription by binding to the promoter region of interferon-stimulated genes. Type I interferon form an ISGfr3 (interferon stimulated gene factor 3) complex which consists phosphorylated STAT1, STAT2 and Irf9.

The purpose of ISGrf3 is to bind to the interferon stimulated response element, which is present in the promoter region of the interferon-stimulated genes. The binding of ISGR3 to the interferon stimulated response element initiates transcription in these genes (Khodarev *et al.* 2012).

Interferon alpha also causes a prolonged activation of c-Jun NH2-terminal kinases, which causes a potential decrease in the mitochondrial membrane. Interferon alpha activated c-Jun NH2-terminal kinases through the activation of PKC-δ signaling which resulted in TRAIL

apoptosis (Yanase *et al.*, 2012).

ISGrf3 can also bind to an interferon gamma activated site, which is also present in the promoter regions of interferon-stimulated genes which also initiates activation of the gene when binding occur.

The Type II interferon cell receptors will bind to JAK1 and JAK2 to STAT1 homodimers, which will bind to Ifn gamma, activated sites on the promoter of interferon-stimulated genes. The Type II interferon mechanism also regulates also the phosphorylation of STAT1 (Andrews *et al.*, 1987).

Interferon gamma activation of c-Jun can be summarized by the following pathway:

Ifn gamma -> Jak1/2 -> MEK 1/2 -> Erk 1/2 -> c-Jun -> AP-1 -> iNos

The above pathway indicates c-Jun is important in the formation of AP-1 and the activation of c-Jun is not dependent on STAT1; however, c-Jun is dependent on interferon gamma (Gough *et al.*, 2007; Chesler and Reiss, 2002).

Irf1 (Interferon release factor 1) is activated by Interferon alpha and gamma by STAT1 and NF-KB through the JAK/STAT pathway (Krämer and Heinzel, 2010).

6.4.1.1 K562-cMyc treated with interferon.

cMyc is known to be involved in cell proliferation and the deregulated expression of cMyc leads to uncontrolled cell proliferation and generally cancer has an increase in cell proliferation cancer one would expect to find an increase in the expression cMyc (Dang, 2012).

Interferon regulates cMyc at the RNA level and mRNA and treatment with interferon α and interferon β resulted in the decreased expression of cMyc-mRNA. Interferon γ treatment decreased the v-cMyc-Max heterodimer complex level, which directly affects the transcription mechanism of cMyc (Hu *et al.*, 2005; Chesler & Reiss, 2002).

According to Dani *et al.* (1985), Daudi cells were treated with interferon α2 and β did not alter the rate of transcription cMyc but altered the half-life of cMyc-mRNA, which caused decreased expression levels of cMyc-mRNA.

6.4.2 K562-cJun treated with interferon alpha and gamma

6.4.2.1 K562-cJun treated with Ifn alpha for 30min and 6hrs

Figure 5-5 indicates an increase in the interferon alpha treatment time of K562-cJun of 30 minutes and 6hrs causes a drop in the number of ChIP-seq peaks. The K562-cJun treated for 6hrs has 5218 ChIP-seq peaks compared to K562-cJun untreated, which has 9848.

Figure 5-6 indicates in the less than 1kb distance, the control (K562-cJun untreated) has 19.4% TFBS, the 30min Ifn alpha treatment has 22.4% and the 6hr treatment has 12% TFBS. K562-cJun has more than 33% of its TFBS in the 10-50kb interval for the control and treated samples. At a distance of greater than 10kb from the TSS the control has 55.1%, the 30min treated sample has 54.3% and the 6hr sample has 61.7% TFBS. The increased treatment time of 6hrs not only causes less ChIP-seq peaks but also fewer TFBS occur in the less than 1kb distance. However, in the greater than 10kb distance the 6hr sample has more TFBS compared to the control and 30min.

cJun regulates cell progression, apoptosis and therefore the down regulation of K562-cJun by Ifn cause the cancer cells not to progress normally (López-Bergami *et al.*, 2005).

6.4.2.2 K562-cJun treated with Ifn gamma for 30minutes and 6hrs

Figure 5-7 indicates an increase in the interferon gamma treatment time of K562-cJun of 30 minutes and 6hrs causes a slight drop in the number of ChIP-seq peaks. The control has 9848, the 30min 8587 and 6hr has 6596 ChIP-seq peaks.

Figure 5-8 indicates in the less than 1kb distance from TSS, the control (K562-cJun untreated) has 19.4% binding sites, the Ifn gamma 30min has 10.9% and 6hr Ifn gamma treated 8.2%. More than 34% binding sites occur in the 10-50kb distance from the TSS for the control and Ifn gamma treated samples. In the greater than 10kb distance, the number of TFBS for the control is 55.1%, 30min Ifn gamma treated 62.8% and 6hr treated 65.6%.

We conclude an increase in Ifn gamma treatment time cause cJun to have fewer binding sites close to the TSS in the 1kb interval. An increase in Ifn gamma treatment time causes binding away from promoter region with more TFBS at a distance of greater than 10kb from the TSS.

K562-cJun without treatment and with treatment generally have more than 55% of its binding sites at distance of 10kb from the TSS. The increased treatment time causes a decrease in the promoter region binding and an increase in the distal binding from the TSS.

cJun regulates cyclin which is required by cells for cell progression and therefore the down regulation of K562-cJun by Ifn cause the cancer cells not to progress (López-Bergami *et al.*, 2005). We assume the Ifn gamma treatment causes cJun to bind further away from the promotor region, which causes the cancer cells not to progress.

6.4.3 K562-cMyc treated with Ifn alpha and gamma for 30min and 6hrs

6.4.3.1 K562-cMyc treated with Ifn alpha 30min and 6hrs

Figure 5-9 indicates an increase in the interferon alpha treatment time of K562-cMyc from 30 minutes to 6hrs causes a corresponding increase in the number of ChIP-seq peaks. The control (K562-cMyc-untreated) has fewer ChIP-seq peaks when compared to the 30min and 6hr Ifn alpha treated K562-cMyc. The 6hr treated sample has double the amount of peaks compared to the control.

From figure 5-10, the number of TFBS in the less than 1kb for the control is 58.6%, the 30min treated has 52.7% and the 6hr treated has 49.8%. In the greater than 1kb region the number of binding sites for the control is 41.4%, the 30min treated 47.7% and the 6hr treatment has 50.3%.

Increase in treatment of Ifn alpha times causes increase in the number of ChIP-seq peaks and a decrease in the number of TFBS in the promoter region, as well as an increase in the number of TFBS that binds further away from the TSS.

According to Andrews *et al.* (1987), Ifn alpha inhibits cell proliferation and therefore the increase in the number of ChIP-seq peaks will prevent proliferation of K562 cells.

6.4.3.2 K562-cMyc treated with Ifn gamma 30min and 6hrs

Figure 5-11 indicates an increase in the interferon alpha treatment time of K562-cMyc from 30 minutes to 6hrs causes a corresponding increase in the number of ChIP-seq peaks. The control (K562-cMyc-untreated) has fewer ChIP-seq peaks when compared to the 30min and 6hr Ifn alpha treated K562-cMyc. The 6hr treated sample has double the amount of peaks compared to the control.

From figure 5-10, the number of TFBS in the less than 1kb for the control is 58.6%, the 30min treated has 52.7% and the 6hr treated has 49.8%. In the greater than 1kb region the number of binding sites for the control is 41.4%, the 30min treated 47.7% and the 6hr treatment has 50.3%.

Increase in treatment of Ifn alpha times causes increase in the number of ChIP-seq peaks and a decrease in the number of TFBS in the promoter region, as well as an increase in the number of TFBS that binds further away from the TSS.

6.4.4 K562-Irf1 treated with Ifn alpha 30min and 6hrs

6.4.4.1 K562-Irf1 treated with Ifn alpha 30min and 6hrs

Untreated ChIP-seq data not available therefore compared the 30min data to the 6hr data.

Figure 5-13 indicates for Ifn alpha the number of peaks increase with increase treatment time. #0min treated Ifn alpha has 2849 ChIP-seq peaks and 6hr has 41690 ChIP-seq peaks. Ifn gamma treatment causes a decrease in the number of ChIP-seq peaks with increase treatment time.

Figure 5-14 shows in the less than 1kb distance from the TSS, the 30min Ifn alpha has 11.9% and the 6hr Ifn alpha has 10.9% TFBS. The 30min Ifn gamma has 22% and the 6hr has 64.8% of its TFBS in the less than 1kb distance from the TSS.

In the greater than 10kb distance from the TSS, Ifn alpha treated for 30min has 64.76% and the 6hr has 64.8% of its TFBS. In the greater than 10kb distance from the TSS, Ifn gamma treated for 30min has 54.6% and the 6hr has 20.6% of its TFBS.

The big difference between Ifn alpha 30min and 6hr treatment is the number of ChIP-seq peaks. The distribution of distance between ChIP-seq peak and the TSS is similar for the Ifn alpha 30min and 6hr treatment.

With Ifn gamma treatment there is a difference in the number of ChIP-seq peaks between the 30min and 6hr treatment. The 30min Ifn gamma treatment has fewer TFBS close to the TSS compared to the 6hr treated sample. The 30min Ifn gamma treated sample has more TFBS in the greater than 10kb distance from the TSS compared to the 6hr treatment.

6.4.5 K562-stat1 treated with Ifn alpha and gamma for 30min and 6hrs

6.4.5.1 K562-stat1 treated with Ifn alpha and gamma for 30min and 6hrs

We do not have untreated K562-stat1 ChIP-seq data and therefore we will compare the 30min data to the 6hr data.

Looking at Figure 5-17, we notice the number of ChIP-seq peaks are very little. K562-stat1 treated with Ifn alpha causes a slight increase in the number of ChIP-seq peaks with increase treatment time. Ifn treatment corresponds to a slight decrease in the number of ChIP-seq peaks with increased treatment time.

Figure 5-18 indicates in the less than 1kb distance from the TSS, Ifn alpha 30min has 36.9% and Ifn alpha 6hr has 37 % TFBS. Ifn alpha in the greater than 10kb distance from the TSS, the 30min has 39.3% and the 6hr has 43.2% TFBS.

Figure 5-19 indicates in the less than 1kb distance from the TSS, Ifn gamma 30 min has 15.8% and the 6hr has 22.2% TFBS. In the greater than 10kb distance from the TSS, Ifn gamma 30min has 56.6% and the 6hr has 49.7% TFBS.

Ifn alpha activates Stat1 because it's part of the Ifn alpha JAK stat pathway (Li, 2008). Ifn alpha treatment has more TFBS closer to the TSS compared to the Ifn gamma treatment. Ifn gamma has more TFBS at distance greater than 10kb from the TSS compared to Ifn alpha treatment.

6.4.5.2 K562-stat2 treated with Ifn alpha for 30minutes and 6hrs

Untreated ChIP-seq data was not available therefore; we compared the 30min data to the 6hr data.

Looking at Figure 5-20, we see K562-stat2 treated with Ifn alpha for 6hr has slightly more peaks compared to the 30min. Increase treatment time corresponds to slight increase in ChIP-seq peaks.

Figure 5-21 shows in the less than 1kb distance from the TSS, Ifn alpha 30min has 37% and the 6hr has 28.9% TFBS. In the greater than 10kb distance from the TSS Ifn alpha treated 30min has 44.4% and the 6hr has 46% TFBS.

We can deduce that more than 50% stat2 binding occurs in the less than 10kb distance from the TSS.
Ifn gamma does not activate STAT2 because it's not part of the Ifn gamma -JAK-STAT pathway (Li, 2008).

6.5 A549-Gr treated with Dexamethasone (dex)

6.5.1 Background

Glucocorticoids used for the treatment of inflammatory disease. The glucocorticoid receptor (GR) is a receptor to which glucocorticoids bind. The immunosuppressive and anti-inflammatory response is induced when glucocorticoid bind to GR which alters the transcription of numerous genes in leukocytes (Liberman *et al.*, 2007).

The GR are located in the cytosol of the cell and when a glucocorticoid bind to the GR, it initiates two processes. The first process is the GR- glucocorticoids complex, which cause an

increase in the expression of anti-inflammatory proteins from the nucleus and the second process is the down-regulation of pro-inflammation proteins in the cytosol.

According to So *et al.* (2007), the glucocorticoids, which bind to GR, are generally located far from their target genes and they distributed equally upstream and downstream from their target genes.

GR in humans occurs in two isoforms namely Grβ and Grα.

Dexamethasone is a synthetic glucocorticoids analog which is produced by the pharmaceutical industry to treat various disease e.g. inflammation (Liberman *et al.*, 2007). When cells treated with dexamethasone, the dexamethasone will bind to GR, which then trigger the cells immunosuppressive and anti-inflammatory response. Dexamethasone is poorly metabolised and stays in the plasma longer when compared to the endogenous hormones. The endogenous hormones generally released in circadian cycle and are highly pulsatile whereas dexamethasone administered at a constant rate during treatment.

Dexamethasone treatment is also related to apoptosis and reducing the amount of T cells which is migrated to an inflammatory region.

According to Reddy *et al.* (2009), genes which were activated by the dexamethasone treatment, on the target genes the GR-dex complex were 11kb away from the TSS and for gene which were repressed the GR-dex complex was 146kb away from the TSS, both upstream relative to the target gene TSS.

For GR not treated with dexamethasone the ChIP-seq signal was very weak and the treatment with dexamethasone resulted in an 11-fold increase in GR ChIP-seq signal (Reddy *et al.* 2009).

6.5.2 Analysis

ChIP-seq data for A549-Gr, which received no dose, is not available therefore; we compared the ChIP-seq data for the various doses with each other. No information is available on the

duration of the treatment and what time after treatment data was collected. It's known that small amount of dexamethasone treatment can activate genes (Reddy *et al.*, 2012).

Figure 5-15 indicates as the dose increase the number of ChIP-seq peaks increases. At a dose of 100nm, the number of ChIP-seq peaks is fewer compared to the 50nm dose.

From figure 5-16 we see in the less than 1kb distance from the TSS, dex 500pm has 11.3%, dex 5nm has 4.2%, dex 50nm has 6.6% and dex 100nm has 27.8% of its TFBS. At a distance greater than 10kb from the TSS, dex 500pm has 67.2%, dex 5nm has 75.9%, dex 50nm has 71% and dex 100nm has 52.3% of its TFBS. It is noted for all samples, most of the binding occurred in the interval greater than 10kb from TSS. It seems A549-Gr treated with dexamethasone bind far away from the TSS.

Reddy *et al.* (2009) found 70% of genes had no Gr binding within a distance of 10kb from the TSS. Our results differ slightly where found fewer binding sites within the 10kb distance from the TSS and most of the TFBS at a distance greater than 10kb form the TSS.

7 Conclusion

The case-control methodology for comparing the distance of the TFBS to the nearest TSS is very useful and informative. ChIP-Enrich is quick and easy to use and the histogram plots give gives a simple and quick overview of the distribution of distances of ChIP-seq peaks to the nearestTSS.

We observed that in the case of K562-cMyc there were a lesser percentage of TFBS, when the distance between the TFBS and the TSS was less than 1kb, compared to GM12878-cMyc. In the case of GM12878-CTCF and K562-CTCF, there was not much difference in the distribution of distances between ChIP-seq peaks and the nearest TSS. Thus our hypothesis of that there is a difference between the distance of the TFBS and the TSS in normal and abnormal cell type is partially correct.

In the case of K562-Irf1 treated with interferon gamma significant differences between the distribution of distance between the ChIP-seq peak and the TSS for the 30min and 6hr samples were seen.

In the case of dexamethasone, we can use the distance between the TFB and the TSS to determine the ideal dose.

ChIP-Enrich could be used as a screening method to get an overview of distribution of distance to nearest transcription factor binding. If there is a difference between the distributions of distances from ChIP-seq peak to nearest gene one could go further by looking at specific distances from a TSS where the differences exist. MEME could also be used to identify motifs and compare the case and control motifs. The next step would be to identify specific genes, which could explain the difference between the normal and case.

A shortcoming of our study was we did not look at the reasons as to why there is was a difference between the control distance and case distance binding.

If there is no difference between distributions of distance from ChIP-seq peak to the nearest TSS then one can assume the difference between the case and control is not caused by the distance between the TFBS and the nearest TSS.

The quality and accuracy of the ChIP-seq data is vital important because poor ChIP-seq data will give false results.

References

Agalioti, T., Chen, G. and Thanos, D. (2002) Deciphering the transcriptional histone acetylation code for a human gene. *Cell* ,**111**, 381-392.

Andrews *et al., (*1987) Effect of recombinant alpha- interferon on the expression of the BCR-ABL fusion gene in human chronic myelogenous human leukemia cell types. *Cancer Res.,* **47**:6629-32.

Bailey *et al*., (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Bailey *et al.*, (2012) The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* 2010, **11**,179

Bailey,T.L. and Elkan,C. (1994) Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Menlo Park, California. American Association for ArtificialIntelligence.

Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, 1-10.

Barrera,L.O. and Ren,B. (2006) The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr. Opin. Cell Biol.*, **18**, 291-8.

Chen *et al.,* (2012) DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep.*, **2**, 1197-206.

Cheng *et al.,* (2011) TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics*, **27**, 3221-7.

Cheng *et al.,* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658-67.

Chesler,D. and Reiss,C.S. (2002) The role of IFN-gamma in immune responses to viral infections of the central nervous system. *Cytokine Growth Factor Rev.*, **13**, 441-54.

Dang,C. V (2012) Myc on the path to cancer. *Cell*, **149**, 22-35.

Dani *et al.,* (1985) Increased rate of degradation of c-myc mRNA in interferon-treated Daudi cells. *Proc. Natl. Acad. Sci. U. S. A.*, **82**, 4896-9.

Doerge,R. (2006) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Biometrics*, **62**, 1270-1271.

Dunham *et al.,* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

The ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

Filippova *et al.,* (2002) Tumor-associated Zinc Finger Mutations in the CTCF Transcription Factor Selectively Alter Its DNA-binding Specificity Advances in Brief Tumor-associated Zinc Finger Mutations in the CTCF Transcription Factor. *Cancer Res,*62,48-52.

Fiorentino,F.P. and Giordano,A. (2012) The tumor suppressor role of CTCF. *J. Cell. Physiol.*, **227**, 479-92

Flicek *et al* (2010) Ensembl's 10th year. *Nucleic Acids Res.,* **38**, D557-D562.

Flicek *et al.,* (2011) Ensembl 2012. *Nucleic Acids Res.*, **40**, 84-90. 15.

Gardner *et al.,* (2002) The c-Myc oncogenic transcription factor. *Encycl. Cancer*, 1-13.

Giannopoulou,E.G. and Elemento,O. (2013) Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome research*, **23**,1295-1306.

Gough *et al.,* (2007) A novel c-Jun-dependent signal transduction pathway necessary for the transcriptional activation of interferon gamma response genes. *J. Biol. Chem.*, **282**, 938-46.

Gupta *et al., (*2007) Quantifying similarity between motifs. *Genome Biol*, **8,**R24.

Hahn, S. (2004) Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural and Molecular Biology,* **11**, 394-403.

Halfon, M.S. (2006) (Re)modelling the transcriptional enhancer. *Nature Genetics*, **38**, 1102-1103.

Hallikas *et al.,* (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47-59.

Håndstad *et al.,* (2012) Cell-type specificity of ChIP-predicted transcription factor binding sites. *BMC Genomics*, **13**, 372.

Hernandez, N. (1993) TBP, a universal eukaryotic transcription factor. *Genes and Development,* **7**, 1291-1308.

Hu *et al.,* (2005) Interferon beta increases c-Myc proteolysis in mouse monocyte / macrophage leukemia cells. **29**, 1307-1314.

Joshi *et al.,* (2012) Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics*, **13**, 199.

Kaplan *et al.,* (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet.*, **7**, e1001290.

Khodarev *et al.,* (2012) Molecular pathways: interferon/stat1 pathway: role in the tumor resistance to genotoxic stress and aggressive growth. *Clin. Cancer Res.*, **18**, 3015-21.

Krämer,O.H. and Heinzel,T. (2010) Phosphorylation-acetylation switch in the regulation of STAT1 signaling. *Mol. Cell. Endocrinol.*, **315**, 40-8.

Kulakovskiy *et al.,* (2013a) From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.

Kulakovskiy *et al.,* (2013b) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, D195-202.

Lee *et al*., (1987) Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40. *Nature,* **325**, 369–372.

Lee *et al.,* (2012) Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome research.* , **22**, 9-24.

Lemon,B. and Tjian,R. (2000) Orchestrated response : a symphony of transcription factors for gene control. *Nucleic Acids Res.*, **25**, 302-3.

Li *et al*., (2006) Chromatin looping and the probability of transcription. *Trends in Genetics*, **22**, 197–202.

Liberman *et al.,* (2007) Glucocorticoids in the regulation of transcription factors that control cytokine synthesis. *Cytokine Growth Factor Rev.*, **18**, 45-56.

López-Bergami *et al.,* (2005) RACK1 mediates activation of JNK by protein kinase C [corrected]. *Mol. Cell*, **19**, 309-20.

Macquarrie *et al.,* (2011) Genome-wide transcription factor binding:beyond direct target regulation. *Trends Genet*, **27**, 141-148.

Mandelkern *et al.,* (1981) The dimension of DNA in solution. *J Mol Biol.*, **152**, 153-161.

Matys,V. (2003) TRANSFAC (R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374-378.

Mrozek *et al.,* (2013) search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information. *BMC Bioinformatics*, **14**, 73.

Myer, V.E. and Young, R.A. (1998) RNA polymerase II holoenzymes and subcomplexes. *Journal of Biological Chemistry,* **273**, 27757-27760.

Paule, M.R. and White, R.J. (2000) Transcription by RNA polymerases I and III. *Nucleic Acids Research*, **28**, 1283-1298.

Périer *et al.,* (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.*, **28**, 302-3.

Pitha,P.M. (1998) Role of the interferon regumatory factors (IRFs) in virus-mediated signaling and regulation of cell growth. *Biochimie.,* **80**, 651-658.

Portales-Casamar *et al.,* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105-10.

Qian *et al.,* (2006) Characterization of binding sites of eukaryotic transcription factors. *Genomics. Proteomics Bioinformatics*, **4**, 67-79.

R Development Core Team,R. (2011) R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.*, **1**, 409.

Reddy *et al.,* (2012) and Expression of Circadian Genes Period 1 and Expression of Circadian Genes. *Mol. Cell. Biol.*, **32**.

Reddy *et al.,* (2009) unexpected mechanisms of gene regulation Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Inform.*, **19**, 2163-2171.

Rosenbloom *et al.,* (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56-63.

Russell, J. and Zomerdijk, J.C.B.M. (2005) RNA-polymerase-I-directed rDNA transcription, life and works. *Trends in Biochemical Sciences,* **30**, 87-96.

Sandelin *et al.,* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91-4.

Schmidt *et al.,* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036-40.

Sentenac, A. (1985) Eukaryotic RNA polymerases. *CRC Critical Reviews in Biochemistry*,**1**, 31-90.

Smale, S.T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes and Development,* **15**, 2503-2508.

So *et al.,* (2007) Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor. *PLoS Genet.*, **3**, e94.

Stegmaier *et al.,* (2004) Systematic DNA-binding domain classification of transcription factors. *Genome informatics Int. Conf. Genome Informatics*, **15**, 276-286.

Tallack *et al.,* (2012) Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq. *Genome Res.*, **22**, 2385-98.

Thomas-Chollier *et al.,* (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.

Vera *et al.,* (2011) Systems biology of JAK-STAT signalling in human malignancies. *Prog. Biophys. Mol. Biol.*, **106**, 426-34.

Villard,J. (2004) Transcription regulation and human diseases. *Swiss Med. Wkly.*, **134**, 571-9.

Wang *et al.,* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798-812.

Watson, J.D. and Young, F.H. (1953) A structure for Deoxyribose Nucleic Acid. *Nature,* **171**, 737-738.

Welch *et al., (*2014) ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucl. Acids Res.* **42** (13):e105.

Whitington *et al*., (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.*, **39 (**15): e98.

Wingender *et al.,* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238-241.

Wong *et al*., (2013) DNA motif elucidation using belief propagation. *Nucleic Acids Res*., **23**, 1–12

Yakovchuk *et al.,* (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix . *Nucleic Acids Res.*, **34**, 567-574.

Yanase *et al.,* (2012) PKC-δ mediates interferon-α-induced apoptosis through c-Jun NH_2-terminal kinase activation. *BMC Cell Biol.*, **13**, 7.

Zhang *et al.,* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Zurita, M. and Merino, C. (2003) The transcriptional complexity of the TFIIH complex. *Trends in Genetics,* **19**, 578-584.

Appendix A – obtaining data and compare distance and binding order of transcription factors to the TSS.

In this section, we illustrate how to obtain and annotate data. We show how to calculate the distance between the transcription factor-binding site and transcriptions start site. We compare the binding order of the case and control binding site relative to the TSS.

1. Obtain data

1.1 Download the ChIP-seq data from the ENCODE website

All the data were downloaded from the following URL:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/

1.2 Extract and uncompressing the compressed tar files

The command used to uncompressed and extract compressed tar files:

tar –zxvf filename.tar.gz

1.3 Create a bed file from uncompressed narrow peak files

Using a python script to create a bed file, which contains 3 columns.

The first column will be the chromosome name, second column will be centre of the peak and the third column will be centre of peak + 1. We using the method which (Tallack *et al.*, 2012) used. We are using the centre of the ChIP-seq peak as the point from which we will be measuring the distance to the TSS. The Bedtools (Quinlan and Hall, 2010) we will be using to join case with controls and with annotation data requires bed file to have the format of at least, chromosome, start and end coordinates.

2. Gene annotation data

The UCSC table browser was use to download the gene annotation data in a bed format file for the entire human genome. The file contained Information on the gene name, transcription start coordinate (TSS), TSS end coordinate and strand.

The gene annotation data was edited using a python script where only genes which corresponded to the + strand was selected. The new gene annotation bed file contained the

chromosome name, TSS start, TSS start + 1. The reason for selecting only TSS and TSS + 1 was because our main goal was to calculate the distance between the centre of the ChIP-seq peak and the TSS. We used the TSS start point and the centre of the ChIP-seq peak as our reference points when calculating distance between TSS and TF.Gene annotation data for the + strand was only selected because the ChIP-seq data only had Information for the + strand.

Procedure for obtaining gene annotation data in particular the TSS Information is as follows:

1. Go to: https://genome.ucsc.edu
2. Click on table browser
3. The table should have the following:

   group – Genes and Gene prediction Tracks

   table – refGene

   output format – selected fields from primary and related tables

   output file – enter a name for output file
4. Click – get output
5. Select fields you want e.g. name, chrom, txStart, txEnd
6. Click – get output

2.1 mergeBed – joining cancer-control cell types

mergeBed is a program which is part of the Bedtools suite of programs. The purpose of mergeBed is to merge overlapping entries into a single interval reducing redundancy and improving the quality of the data. A requirement of mergeBed is the data needs to be sorted first based on chromosome then start. The general syntax for using sortBed and mergeBed is as follows.

2.2 sortBed –i inputFile.bed | mergeBed > out_sorted_merged.bed

The input bed file was sorted using the default setting for sortBed and then piped to mergeBed and the output file will be a sorted file with overlapping entries merged.

2.3 closestBed – joining case-control cell types

To join the case cell type with the control closestBed, was used. The closestBed program which will find the closest region between two bed files

2.4 closestBed -a cancerCell –b controlCell – d – t first > cancer_control.bed

-a indicates the name of the first file

-b indicates the name of second file

-d calculates the distance between the case and control cell, meaning calculate the distance between the first file (-a argument) and the second file (-b argument)

- t first – return the first match in file b and ignore any other subsequent matches which might correspond to the same region in the first file.

The input to closestBed is two bed files each contain 3 column (chromosome, start, end) and the output of closestBed is a bedfile contains 7 columns. The first 3 columns will be ChIP-seq peak of cancer cell, the next 3 columns will be the closet control cell type ChIP-seq peak and the last column will contain distance between the centre of the cancer ChIP-seq peak and the centre of the closest control ChIP-seq peak. The distance is measured by the number bases between the two peaks.

2.5 closestBed – joining case/control cell types with TSS

The procedure in step 5 joined the cancer cell type with the control cell type and to find the closest TSS one need to join the case/control output from step 5 with the gene edited gene annotation data. closestBed is used again this time the first input file argument, –a case/control.bed (output from step 5) and the second input file argument, – b will be the

edited TSS data from step 4 above. The program closestBed will find the closest regions from the first three columns in the first input file and match the corresponding first three columns in the second input file. The output will consist of the closest TSS data for the case files because the first three columns from first input file correspond to the case coordinated.

The syntax for the closestBed command:

closestBed –a case_control.bed –b edited_TSS.bed –d –t first > cancer_control_TSS.bed

The output of the above closestBed will have 11 columns with the additional columns the nearest TSS to the cancer cell_type and the last column will be the distance between the centre of the ChIP-seq peak of the cancer cell type and TSS, which is measured in number of bases.

3. psT Distribution

To gain a better understanding of the binding of the case cell type vs. control cell type vs. TSS we introduce the psT distribution nomenclature.

The psT distribution allow us to determine the binding order of the case cell type relative to the control cell type as well as the binding order of the case/control cell type relative to the TSS. We define upstream as the TFBS that occur before the TSS and downstream as TFBS that occur after the TSS.

The psT distribution naming convention is as follows:

p – TF used in first (primary) argument in closestBed (correspond –a value in closestBed)

s- TF used in second (secondary) argument in closestBed (correspond –b value in closestBed)

T – Refers to the TSS

The various combinations for p, s and T are as follows.

spT, psT, Tsp, Tps, pTs, sTp

3.1 Binding order of p, s and T

When looking at the binding order we would like to know when p is in front of s, and when s is in front of p. When we used the word "in front" we are comparing the centre of the ChIP-seq peaks of s and p; p in front of s means on a linear scale the location of the CHiP-seq peak of p has a smaller coordinate than s. It is a given p and s should be on the same chromosome.

E.g., if p had bed coordinates chr 10 100(peakcentre) and s bed coordinates chr 10 450 then p would be in front of s because a 100 is less than 450. When looking at the order (binding of p relative to s) we also take into account whether the binding is upstream or downstream. We consider the binding of p and s relative to the TSS. Upstream defined as the location in front of the TSS with coordinates less than the TSS. Downstream defined as the location after the TSS, those with coordinates greater than the TSS.

3.2 p, s and TSS binding combinations

The binding of cases cell type and control cell type relative to TSS can occur in the following combination. Let assume p = case cell type, s= control cell type and T= TSS and the same TF is used in both cell type. Since the TF is the same for both cell types, we will refer to the cell types to avoid confusion or where the TF is different between the cell types we will refer to the TF.

psT – s (control cell type) binds closer to the TSS and p (case cell type) binds further away from TSS. Relative to the TSS s is in front of p. Both case and control binds upstream of

TSS with control binding closer to TSS and cancer cell type further away from TSS

spT - both case and control binds upstream of TSS with caser binding closer to TSS and control cell type further away from TSS

Tps - both cancer and control binds downstream of TSS with cancer binding closer to TSS and control cell type further away from TSS

Tsp - both cancer and control binds downstream of TSS with control binding closer to TSS and cancer cell type further away from TSS

pTs – cancer cell type binds upstream and control is binding downstream with TSS in the middle

sTp – control cell type binds upstream and cancer is binding downstream with TSS in the middle

We can also infer from the psT binding order whether cancerous cell type or normal type tend to bind more closely to the TSS and whether its preferred binding is upstream or downstream relative to the TSS.

A summary of the binding order of p and s relative to TSS shown below.

binding order	Upstream (before TSS)	Downstream (after TSS)
p in front of s	psT or pTs	Tps
s in front of p	spT or sTp	Tsp

Table 1 p and s upstream and downstream binding order relative to TSS.

A summary of the closeness of p and s relative to T shown below

closeness to TSS	Upstream (before TSS)	Downstream(after TSS)
p closer to T	spT or pTs	Tps or sTp
s closer to T	psT or sTp	Tsp or pTs

Table 2 closeness of p and s relative to the TSS

Dividing data into intervals

To compare bindings we normalized the two datasets.

After normalization, we summed all the values in the table and in each interval; we expressed each value as a percentage of the total, which allowed us to compare values.

The final dataset from closestBed was divided into 3 groups.

1. psT peak distribution

The data was divided into intervals based on the distance between the two peaks i.e. the centre of the peak of the first input file (p) and the centre of the peak of the second input file (s).

We divide this group further into two intervals. The first interval is where the distance between the peaks of p and s is less than 2kb and the second interval is where the distance between the peaks of p and s is greater or equal to 2kb. In this group, we only focussing on the distance between the peaks of p and s.

E.g., psT peak distances less 2kb would mean the distance between the peaks of p and s is less than 2000 bases.

2. psT TSS distribution

The data was divided into intervals based on the distance between the TSS and the peak of p or s.

We divide this group further into two intervals. The first interval is where the distance between the TSS and peaks of p or s is less than 2kb and the second interval is where the distance between the TSS and the peaks of p or s is greater or equal to 2kb. In this group, we only focussing on the distance between the TSS and the peak of p or s.

E.g., psT TSS-TF distance less or equal to 2kb would mean the distance between either the centre of the peak of p or s and the TSS is less than 2000 bases.

3. psT Global distribution

The psT global distribution looks at the genome wide ordering of p and s. The data is not divided into any intervals.

4. Example of Results

4.1 K562-Gm12878- cMyc

The binding distribution K562 and GM12878 cell types with particular reference to the TF cMyc.

4.2 Number of cMyc peaks for K562 and GM12878

A summary of the number of peaks found for K562 and GM12878 cell types for the TF cMyc shown.

	cMyc(no of peaks)
K562	11738
GM12878	3690

Table 3 K562 and GM12878 number of cMyc peaks

A graphical display of the number of peaks shown below.

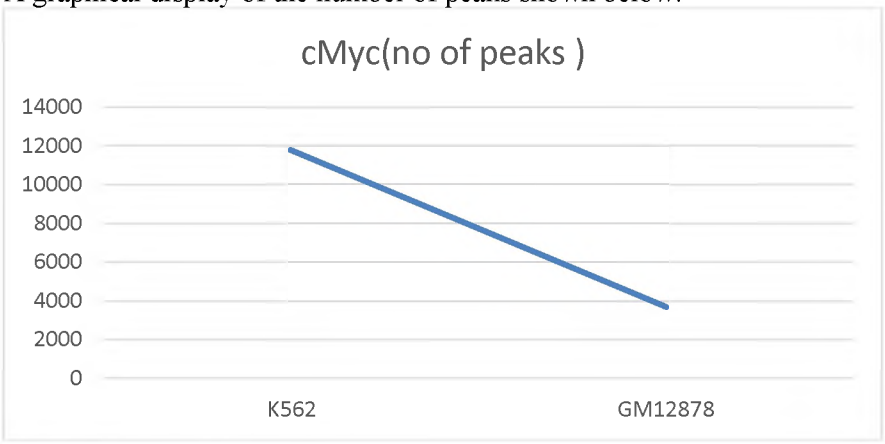


Figure 1 Number of cMyc peaks

Figure 1 indicates there is almost 3 times more cMyc binding sites in the K562 cell type as compared to the GM12878 cell type.

4.3 Genome-wide distribution for K562-cMyc and GM12878-cMyc

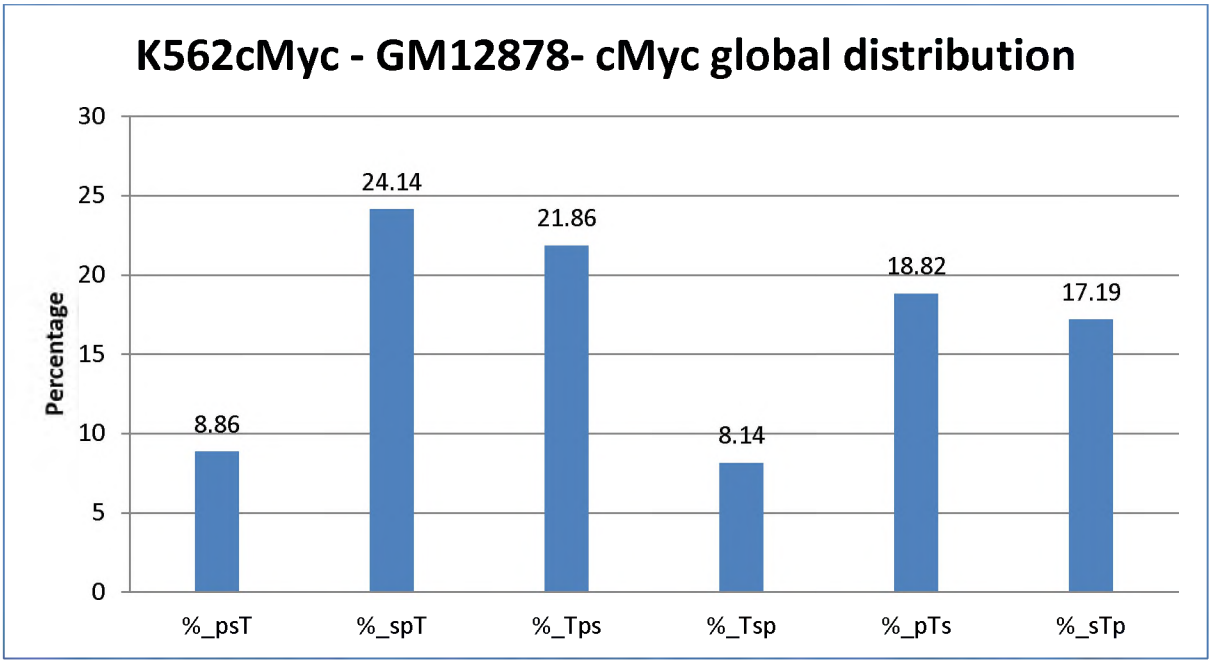


Figure 2 global psT distribution for p and s relative to TSS. All values in percentage. p - K562cMyc, s- GM878-cMyc

p / s	p in front s		s in front p		p close to T		s close to T	
	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream
UtaGm12878Cmyc/ UtaGm12878Cmyc	48.6	51.4	0.0	0.0	0.0	51.4	48.6	0.0
UtaK562Cmyc/ UtaK562Cmyc	52.9	47.1	0.0	0.0	0.0	47.1	52.9	0.0
UtaK562Cmyc/ UtaGm12878Cmyc	28.7	21.9	41.3	8.1	32.1	28.5	19.0	20.4

Table 4 global psT Genome distribution (p-K562, s-GM12878, T-TSS) for cMyc (values in percentage)

Discussion on Table 4

UtaGm12878Cmyc_UtaGm12878Cmyc

The global psT distribution for GM12878-cMyc can be summarized by psTps.
In this case p equals s which is equal to GM12878-cMyc (p=s=GM12878-cMyc). We can see that 48.6 % of the time in GM12878 cMyc binds upstream and the 51.4% downstream with respect to TSS.

UtaK562Cmyc_UtaK562Cmyc

Since p=s=K562-cMyc the psT pattern would be pTp. We see from Table cMyc in K562 binds 52.9% upstream and 47.1 % downstream with respect to the TSS.

UtaK562Cmyc_UtaGm12878Cmyc

The psT distribution for K562-GM12878-cMyc can be summarized by**:** spTps
This means cMyc in the K562 cell type tend to bind closer to the TSS both upstream and downstream compared to cMyc in the GM12878 cell type.
When comparing the binding of K562-cMyc and GM12878-cMyc we notice K562-cMyc

has more binding sites upstream compared to GM12878cMyc.

4.4 psT distribution where TF (p)-TF(s) distance less or equal to 2kb

P / S	p in front s		s in front p		p close to T		s close to T	
	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream
UtaGm12878Cmyc/ UtaGm12878Cmyc	48.6	51.4	0.0	0.0	0.0	51.4	48.6	0.0
UtaK562Cmyc/ UtaK562Cmyc	52.9	47.1	0.0	0.0	0.0	47.1	52.9	0.0
UtaK562Cmyc/ UtaGm12878Cmyc	29.5	20.2	32.3	17.9	29.4	23.9	22.3	24.4

Table 5 psT distribution peak distance less 2kb (p-K562, s-GM12878, T-TSS) for cMyc (values in percentages)

Discussion on Table 5

UtaGm12878Cmyc_UtaGm12878Cmyc

The psT distribution for GM12878 is pTp with 48.6% binding occurring upstream and 51.4% of binding occurring downstream.

UtaK562Cmyc_UtaK562Cmyc

Since p=s=K562-cMyc the psT pattern would be pTp. We see from Table Table cMyc in K562 binds 52.9% upstream and 47.1% downstream with respect to the TSS.

UtaK562Cmyc_UtaGm12878Cmyc

The psT distribution for K562-GM12878-Cmyc can be summarized by**:** spTps

This means cMyc in the K562 cell type tend to bind closer to the TSS both upstream and downstream compared to cMyc in the GM12878 cell type.

4.5 psT distribution where TF (p)-TF(s) distance greater than 2kb

P / S	p in front s		s in front p		p close to T		s close to T	
	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream
UtaGm12878Cmyc UtaGm12878Cmyc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
UtaK562Cmyc UtaK5 62Cmyc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
UtaK562Cmyc UtaGm12878Cmyc	28.1	22.9	47.0	1.9	33.5	30.8	17.3	18.4

Table 6 psT distribution peak distance greater 2kb (p-K562, s-GM12878, T-TSS) for cMyc (values in percentage)

Discussion on Table 6

UtaGm12878Cmyc_UtaGm12878Cmyc

In the interval greater than 2kb we expect the distance between the peaks to be 0 because p=GM12878–cMyc and s=GM12878-cMyc.

UtaK562Cmyc_UtaK562Cmyc

In the interval greater than 2kb we expect the distance between the peaks to be 0 because

p=K562-cMyc and s=K562-cMyc.

UtaK562Cmyc_UtaGm12878Cmyc
The psT distribution summary for K562-GM12878-cMyc is: spTps
This is in agreement with the results obtained thus far of K562-cMyc binding closer to TSS both upstream and downstream.

4.6 psT TSS-TF distance less or equal to 2kb

p	p in front s		s in front p		p close to T		s close to T	
s	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream
UtaGm12878Cmyc/ UtaGm12878Cmyc	42.4	57.6	0.0	0.0	0.0	57.6	42.4	0.0
UtaK562Cmy/ UtaK562Cmyc	52.3	47.7	0.0	0.0	0.0	47.7	52.3	0.0
UtaK562Cmyc/ UtaGm12878Cmyc	32.1	20.2	38.5	9.2	31.4	26.8	18.7	23.1

Table 7 psT distribution TSS-peak distance less 2kb (p-K562, s-GM12878, T-TSS) for cMyc (values in percentage)

Discussion on Table 7
UtaGm12878Cmyc_UtaGm12878Cmyc
The psT distribution for GM12878 is pTp with 42.4 % binding occurring upstream and 57.6 % of binding occurring downstream.

UtaK562Cmyc_UtaK562Cmyc
Since p=s=K562-cMyc the psT pattern would be pTp. We see from Table Table cMyc in K562 binds 52.3% upstream and 47.7 % downstream with respect to the TSS.

UtaK562Cmyc_UtaGm12878Cmyc
The psT distribution for K562-GM12878-Cmyc can be summarized by**:** spTps
This means cMyc in the K562 cell type tend to bind closer to the TSS both upstream and downstream compared to cMyc in the GM12878 cell type.

4.7 psT TSS-TF distance greater than 2kb

p	p in front s		s in front p		p close to T		s close to T	
s	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream	Up stream	Down stream
UtaGm12878Cmyc/ UtaGm12878Cmyc	55.7	44.3	0.0	0.0	0.0	44.3	55.7	0.0
UtaK562Cmyc/ UtaK562Cmyc	53.3	46.7	0.0	0.0	0.0	46.7	53.3	0.0
UtaK562Cmyc/ UtaGm12878Cmyc	25.8	23.3	43.8	7.2	32.7	30.1	19.3	18.0

Table 8 psT distribution TSS-peak distance greater 2kb (p-K562, s-GM12878, T-TSS) for cMyc (values in percentage)

Discussion on Table 8
UtaGm12878Cmyc_UtaGm12878Cmyc

The psT distribution for GM12878 is pTp with 55.7 % binding occurring upstream and 44.3 % of binding occurring downstream.

UtaK562Cmyc_UtaK562Cmyc
Since p=s=K562-cMyc the psT pattern would be pTp. We see from Table Table cMyc in K562 binds 53.3% upstream and 46.7 % downstream with respect to the TSS.

UtaK562Cmyc_UtaGm12878Cmyc
The psT distribution for K562-GM12878-Cmyc can be summarized by**:** spTps

Appendix B – Quality control of CHiP-seq data

1. Summary statistics of raw sequence data- GM12878-cmyc (control)

gm12878_cmyc_sample.fastq FastQC Report

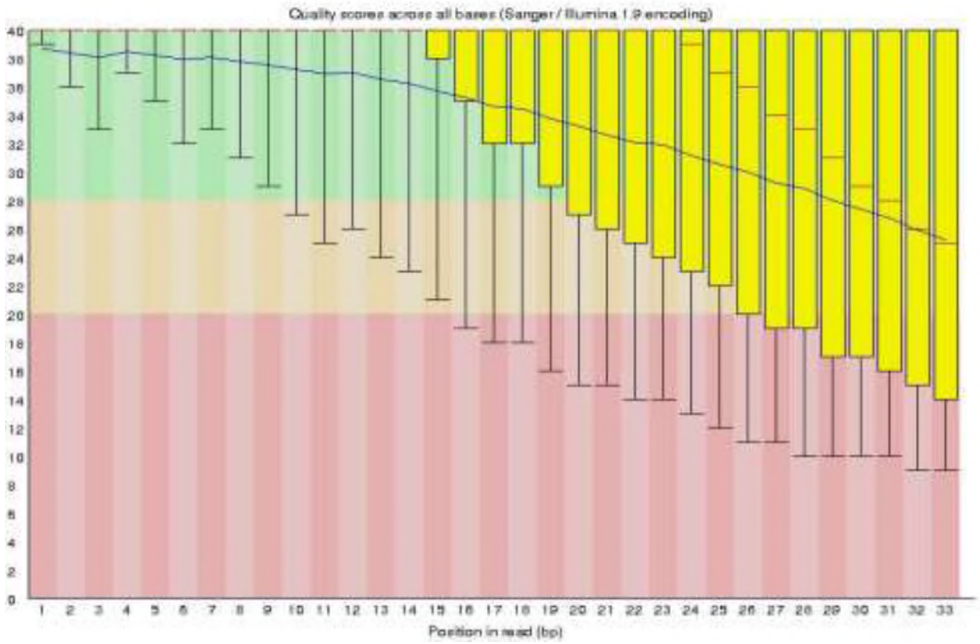
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

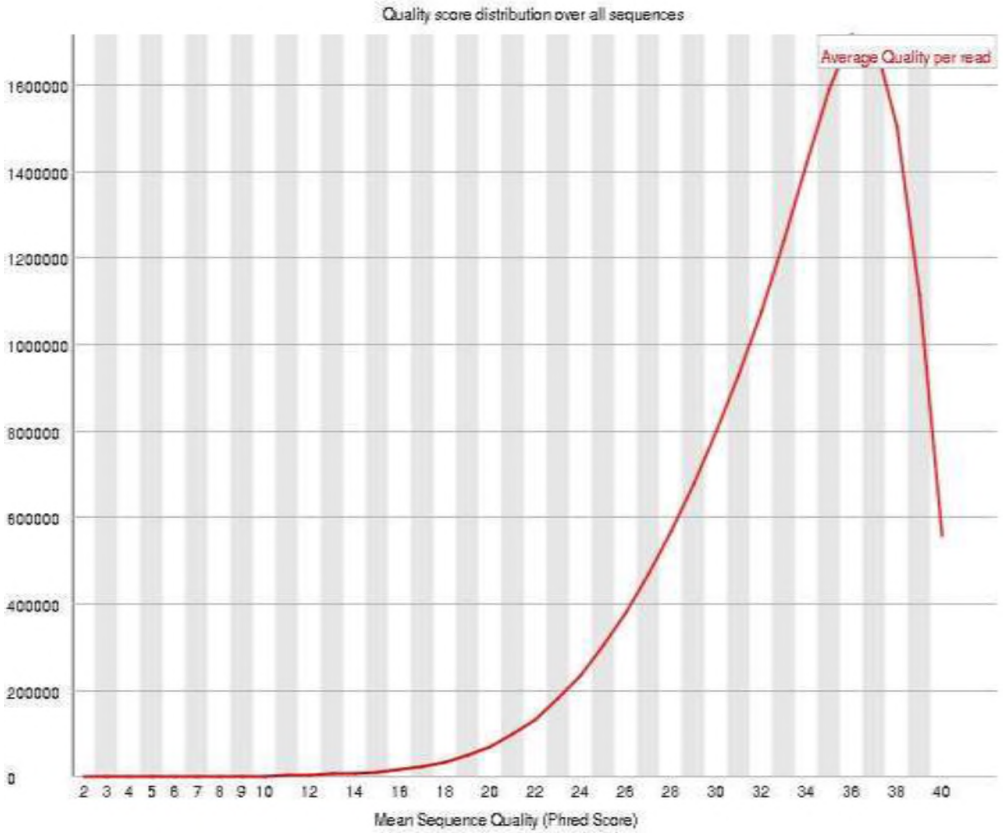
Basic Statistics

Measure	Value
Filename	gm12878_cmyc_sample.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Totl Sequences	16954037
Sequences flagged as poor quality 0 Sequence length	33
%GC	41

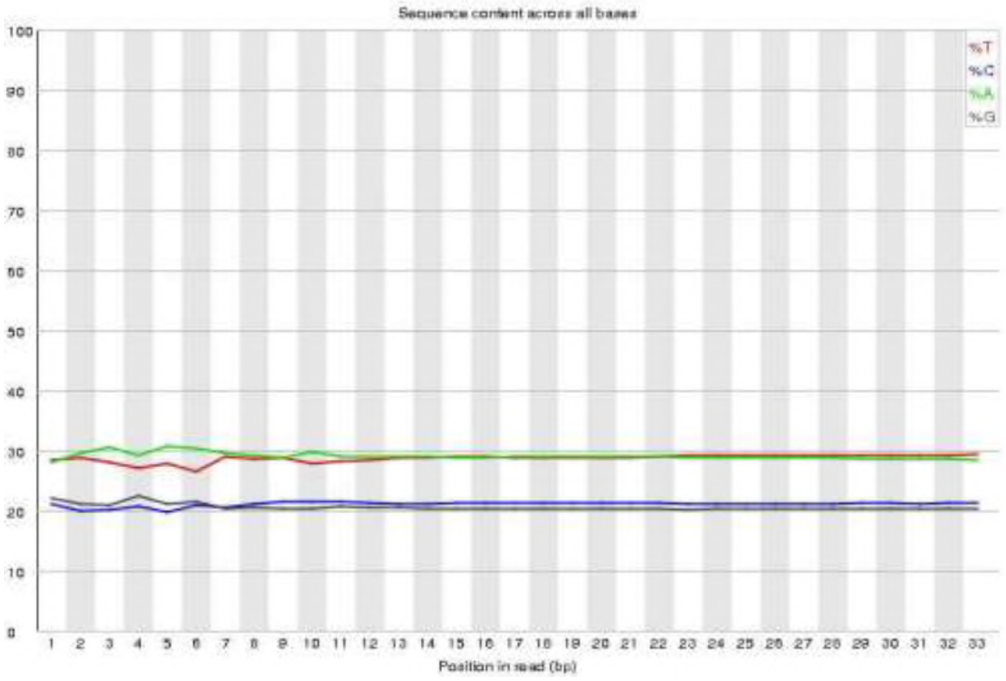
Per base sequence quality



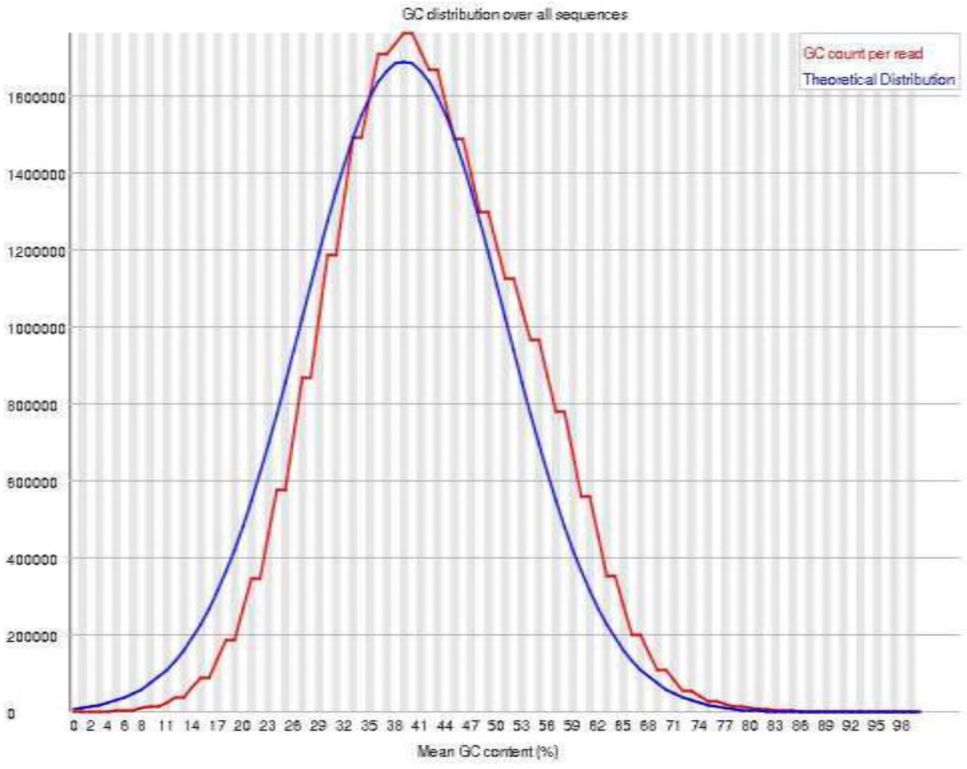
Per sequence quality scores



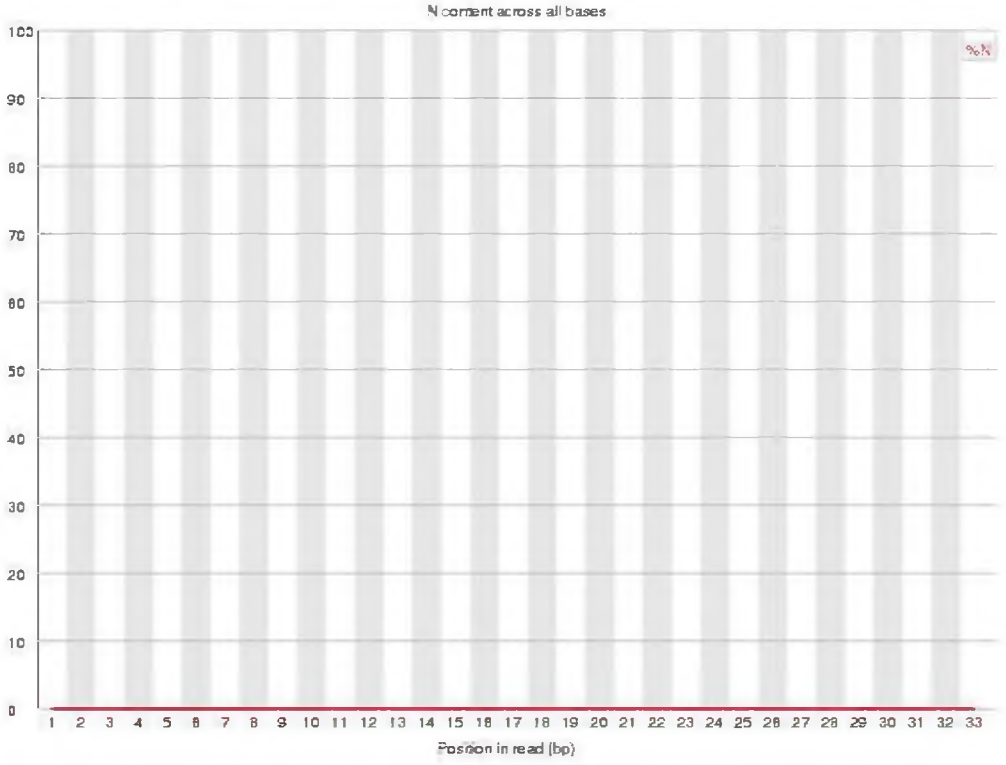
Per base sequence content



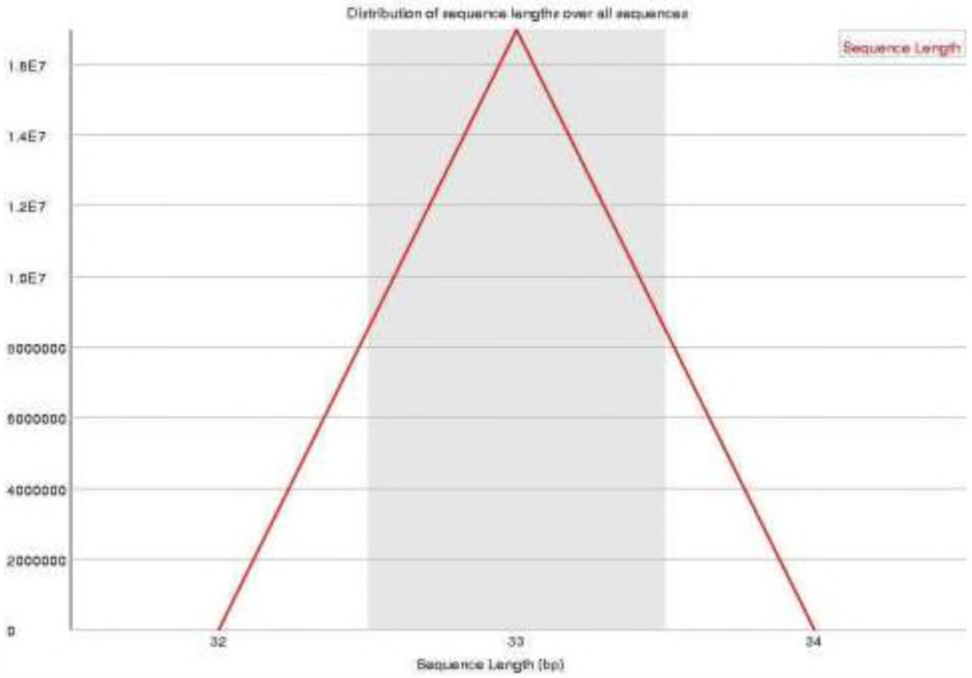
Per sequence GC content



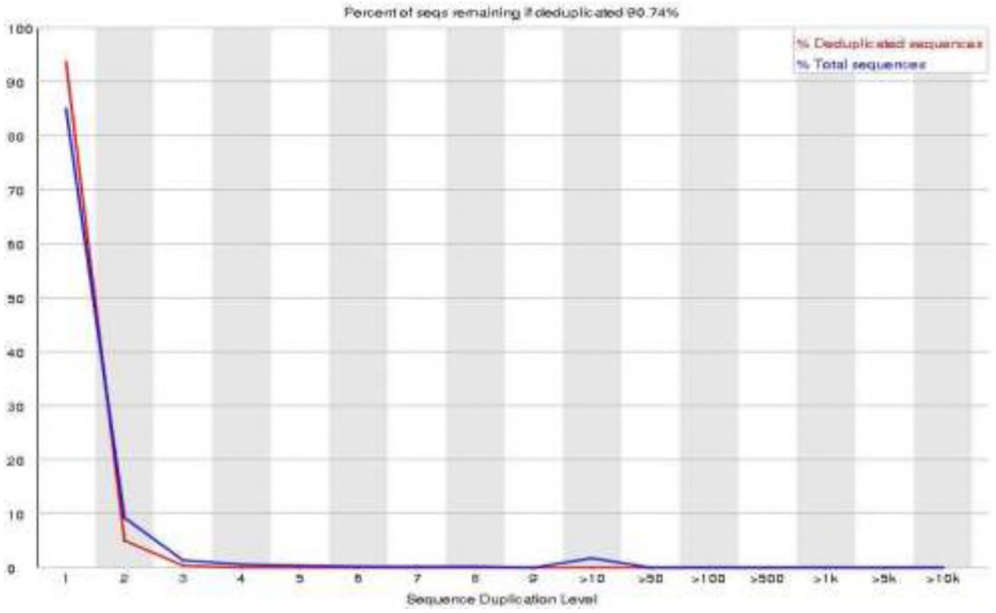
Per base N content



Sequence Length Distribution



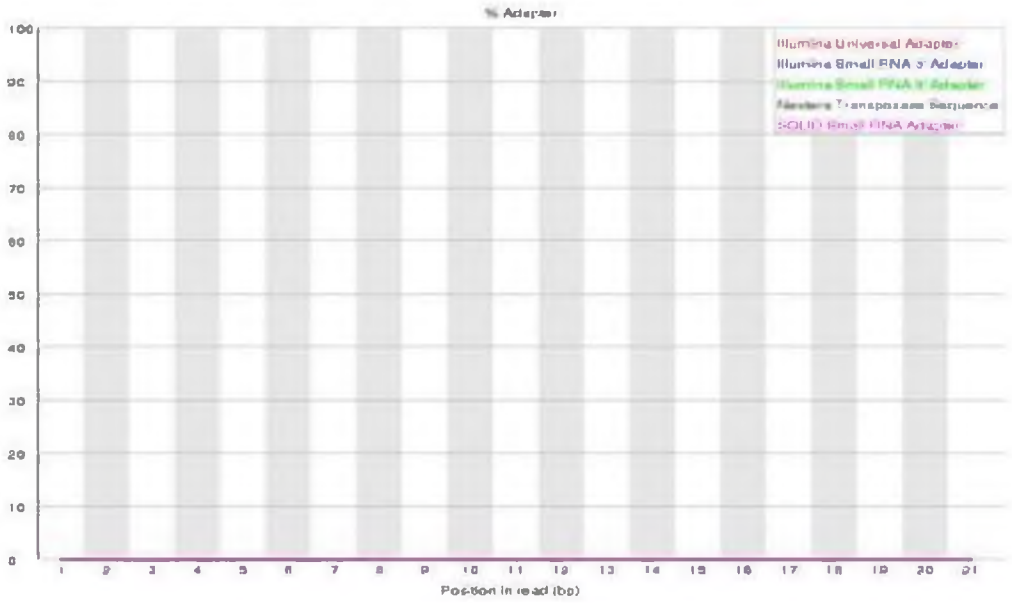
Sequence Duplication Levels



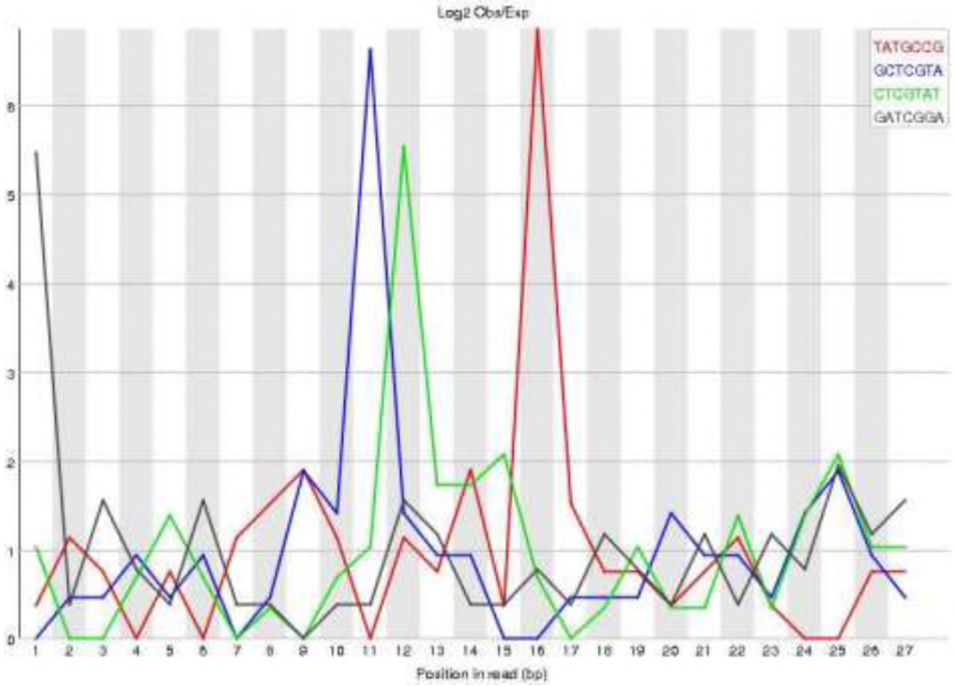
Overrepresented sequences

No overrepresented sequences

Adapter Content



Kmer Content



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TATGCCG	355	1.0552722E-6	6.847477	16
GCTCGTA	285	1.28397E-4	6.6327167	11
CTCGTAT	390	2.0644582E-4	5.539706	12
GATCGGA	345	0.001500593	5.4769382	1

Produced by FastQC (version 0.11.5)

2. Summary statistics of raw sequence data- K562-cMyc (case)

k562_cmyc_sample.fastq FastQC Report

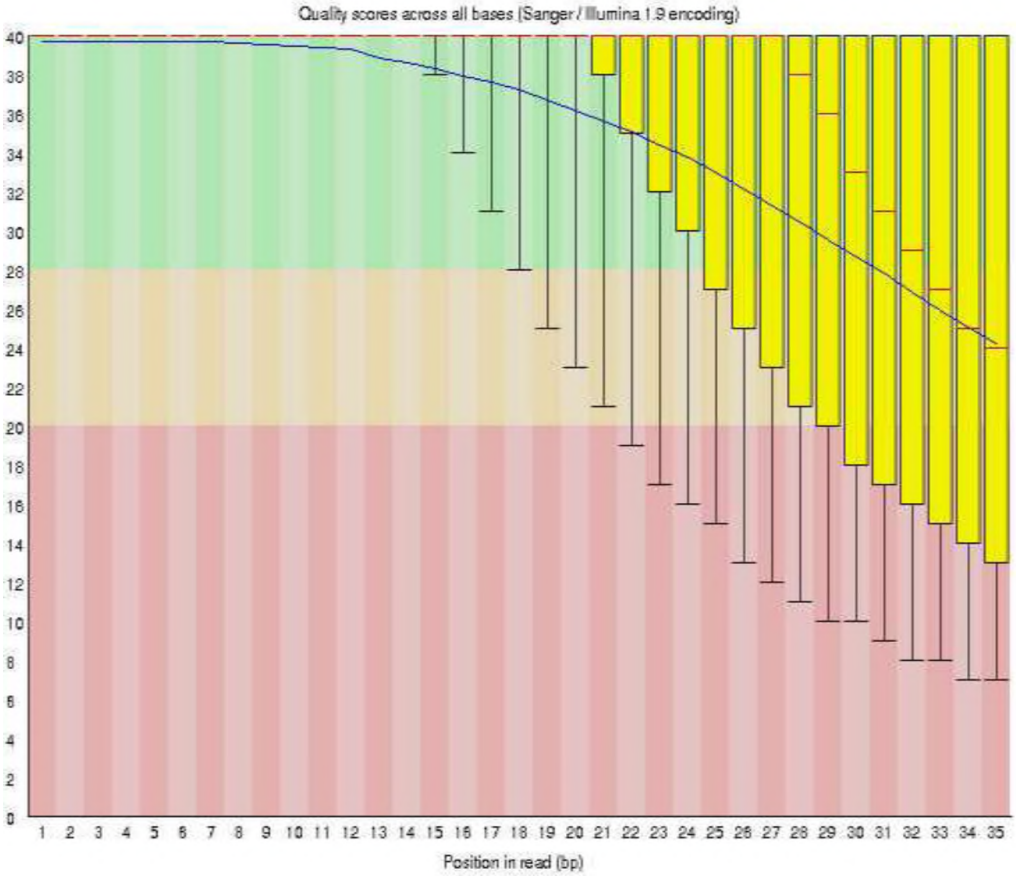
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

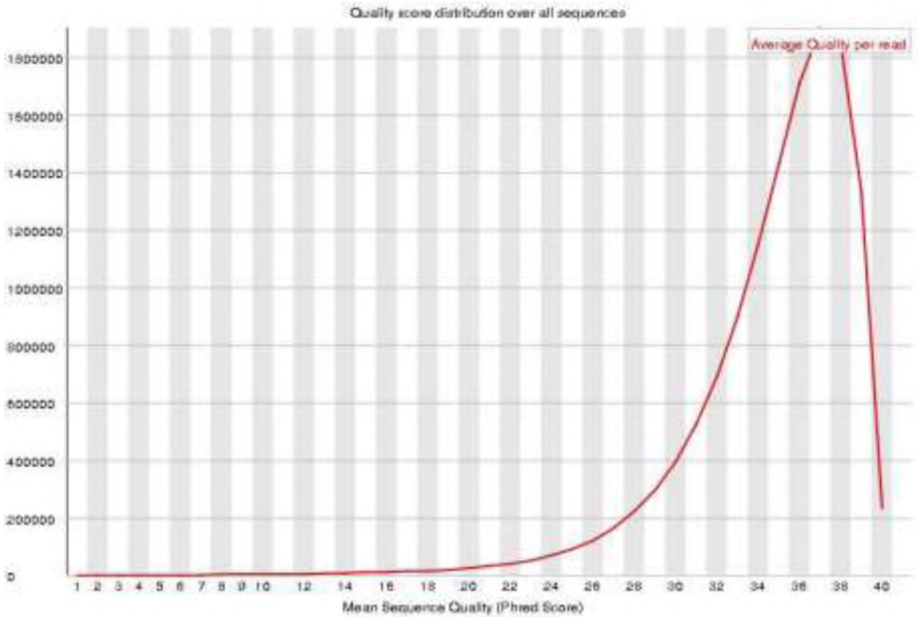
Basic Statistics

Measure	Value
Filename	k562_cmyc_sample.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	13388470 Sequences flagged as poor quality 0 Sequence
length	35
%GC	44

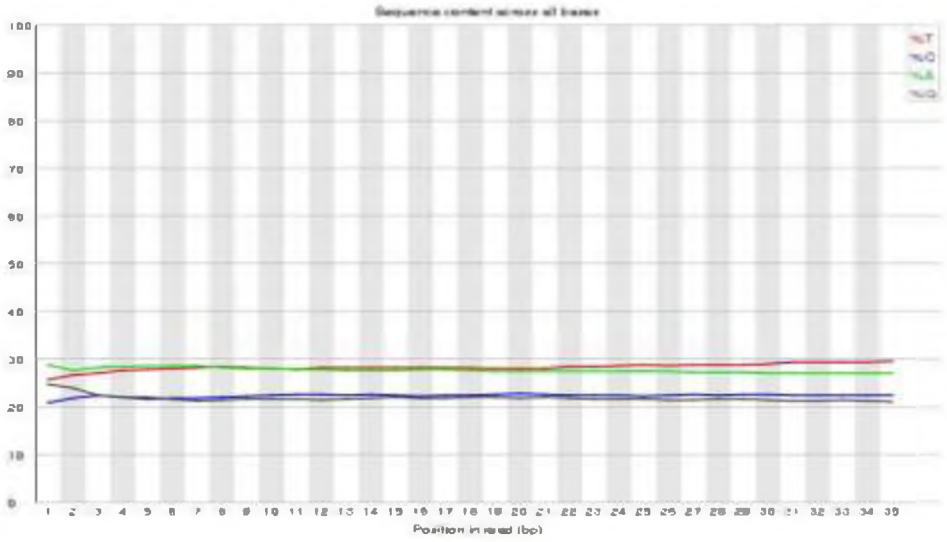
Per base sequence quality



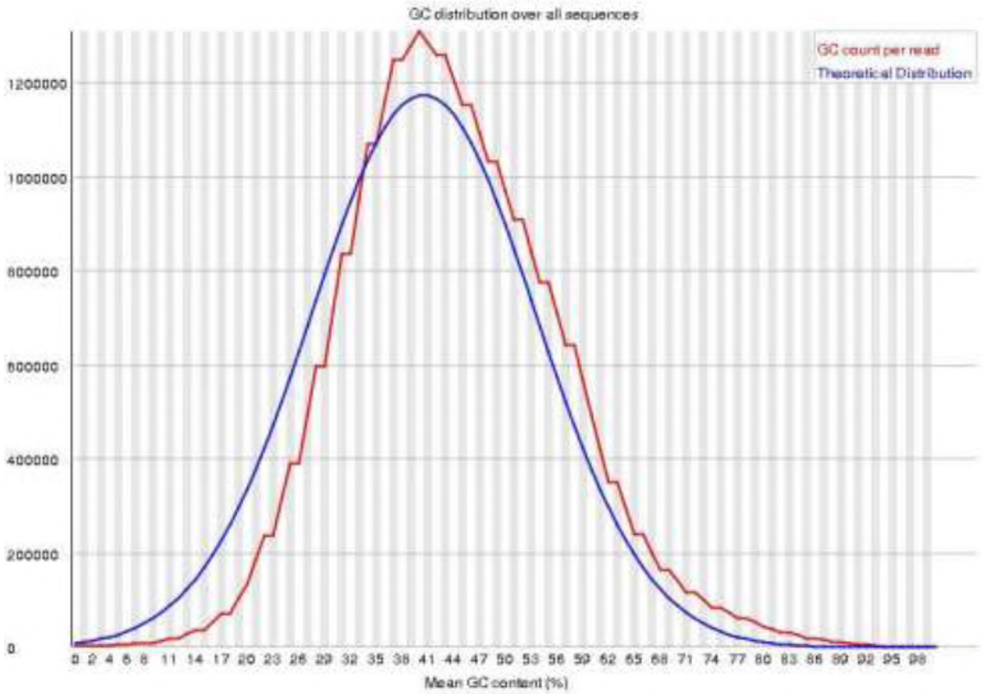
Per sequence quality scores



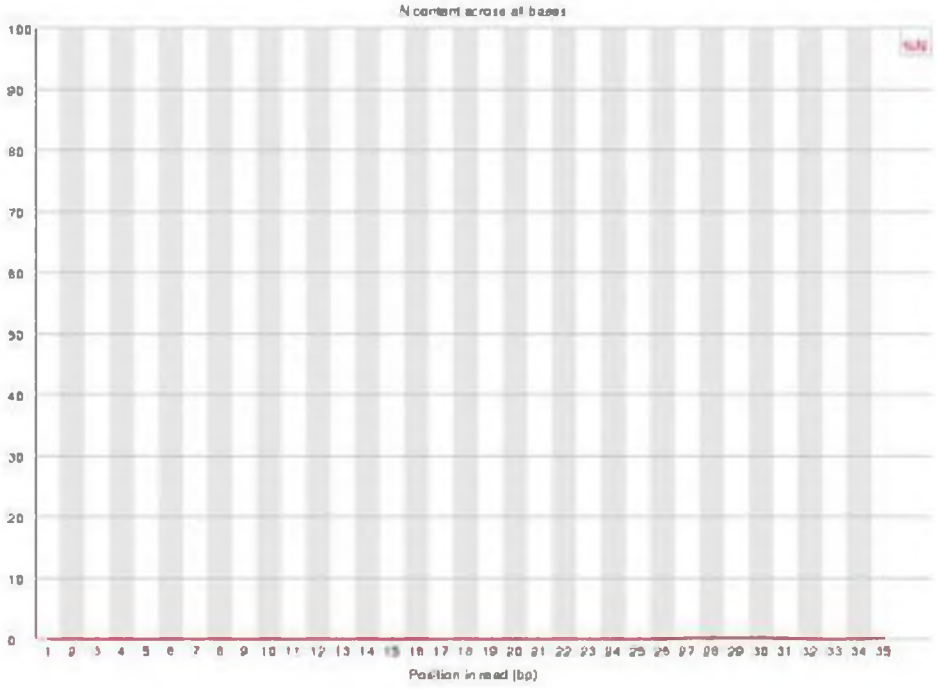
Per base sequence content



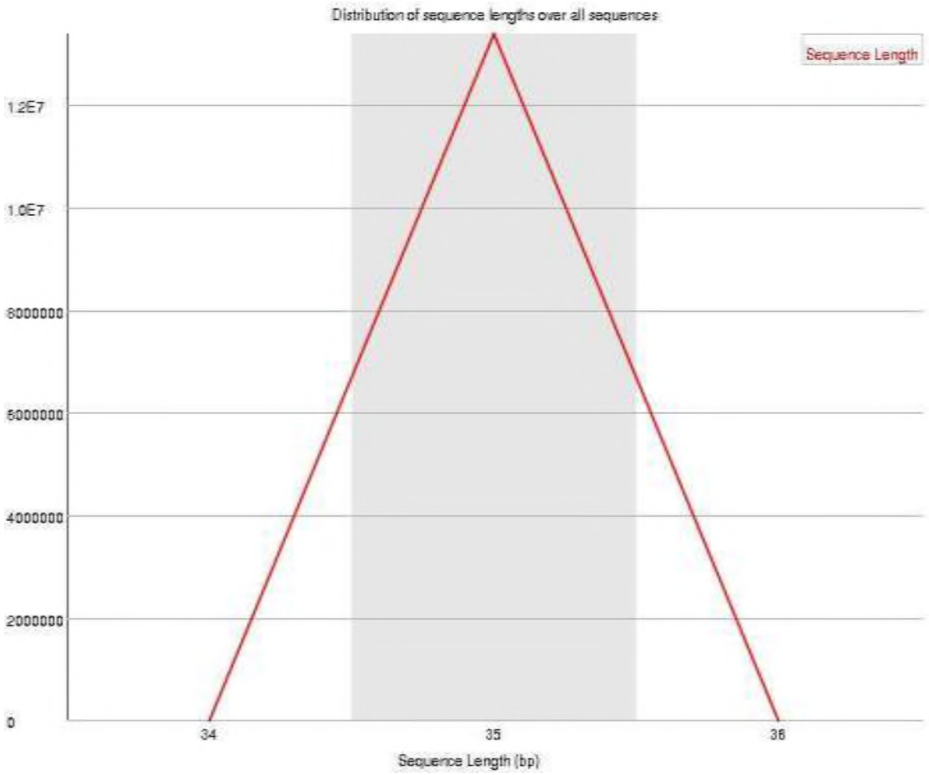
Per sequence GC content



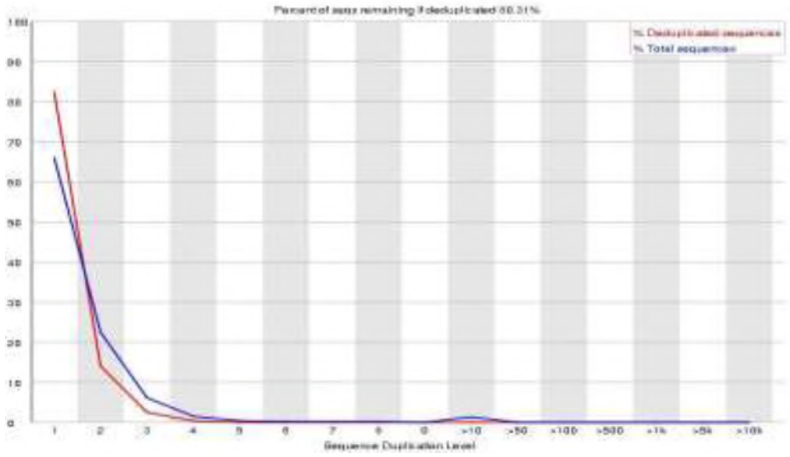
Per base N content



Sequence Length Distribution



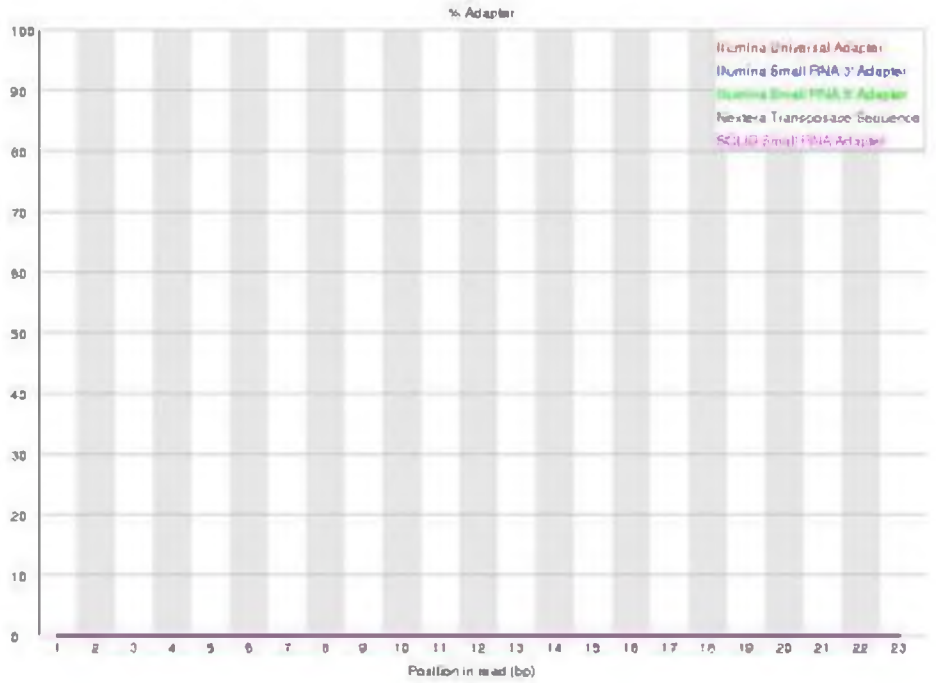
Sequence Duplication Levels



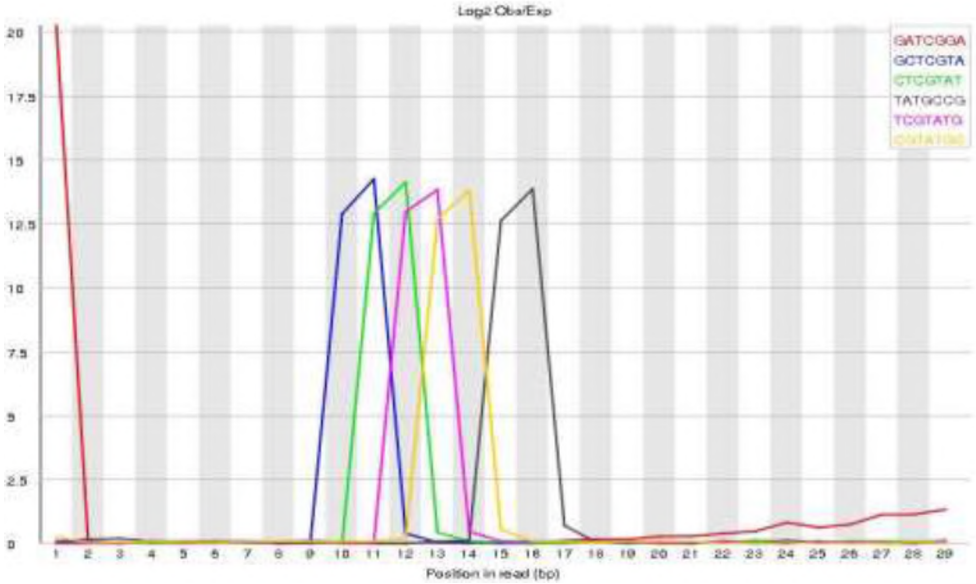
Overrepresented sequences

No overrepresented sequences

Adapter Content



Kmer Content



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GATCGGA	3930	0.0	20.222609	1
GCTCGTA	5315	0.0	14.255909	11

CTCGTAT	5355	0.0	14.151802	12
TATGCCG	5455	0.0	13.879352	16
TCGTATG	5440	0.0	13.825324	13
CGTATGC	5500	0.0	13.811094	14
AGCTCGT	5480	0.0	13.63799	10
ATGCCGT	5560	0.0	13.305397	17
GCCGTCT	5550	0.0	13.303135	19
TGCCGTC	5660	0.0	12.941986	18
GAGCTCG	5870	0.0	12.829101	9
CCGTCTT	5830	0.0	12.817245	20
CGTCTTC	5850	0.0	12.458263	21
GTATGCC	6320	0.0	12.2975235	15
CGGAAGA	6445	0.0	12.218935	4
TCGGAAG	6560	0.0	12.203339	3
ATCGGAA	6535	0.0	12.161415	2
GAAGAGC	8015	0.0	9.62678	6
AGAGCTC	8085	0.0	9.367284	8

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GTCTTCT	8680	0.0	8.7527685	22

Produced by FastQC (version 0.11.5)

3. Comparing K562-cMyc and GM12878-cMyc signal intensity at various intervals around the TSS

Below is a summary of the signal intensity and count of the ChIP-seq peaks for K562-cMyc and GM12878 at various intervals around the TSS.

Summary of results

G12878-cMyc

Before TSS

Interval	Ave signal intensity	Count
Average signal intensity in (0-0.1]	13.691	6
Average signal intensity in (0.1-1]	6.926	54
Average signal intensity in (1-5]	2.617	240
Average signal intensity in (5-10]	2.359	300
Average signal intensity in(10-20]	2.282	600

After TSS

Interval	Ave signal intensity	Count
Average signal intensity in (0-0.1]	15.484	6
Average signal intensity in (0.1-1]	9.042	54
Average signal intensity in (1-5]	2.845	240
Average signal intensity in (5-10]	2.579	300
Average signal intensity in(10-20]	2.44	600

K562-cMyc

Before TSS

Interval	Ave signal intensity	Count
Average signal intensity in (0-0.1]	3.045	6
Average signal intensity in (0.1-1]	3.297	54
Average signal intensity in (1-5]	2.503	240
Average signal intensity in (5-10]	2.21	300
Average signal intensity in(10-20]	2.133	600

After TSS

Interval	Ave signal intensity	Count
Average signal intensity in (0-0.1]	2.99	6
Average signal intensity in (0.1-1]	3.042	54
Average signal intensity in (1-5]	2.785	240
Average signal intensity in (5-10]	2.461	300
Average signal intensity in(10-20]	2.333	600

GM12878-cMyc– from above intervals. Averages for both upstream and downstream

Interval	Ave signal Intensity	Count

Average signal intensity in (0-0.1]	14.5875	12
Average signal intensity in (0.1-1]	7.984	108
Average signal intensity in (1-5]	2.731	480
Average signal intensity in (5-10]	2.469	600
Average signal intensity in(10-20]	2.361	1200

K562-cMyc– from above intervals. Averages for both upstream and downstream

Interval	Ave signal Intensity	Count
Average signal intensity in (0-0.1]	3.0175	12
Average signal intensity in (0.1-1]	3.1695	108
Average signal intensity in (1-5]	2.644	480
Average signal intensity in (5-10]	2.3355	600
Average signal intensity in(10-20]	2.233	1200

Global – [0-20kb] distance from the TSS

Before TSS

	Ave signal Intensity	Count
G12878-cMyc -Global Average signal intensity	2.634	1200
K562-cMyc -Global Average signal intensity	2.283	1200

After TSS

	Ave signal Intensity	Count
G12878-cMyc -Global Average signal intensity	2.918	1200
K562-cMyc -Global Average signal intensity	2.491	1200

Global-average intensity in the [-20kb-20kb] interval from the TSS

	Ave signal Intensity	Count
G12878-cMyc -Global Average signal intensity	2.776	2400
K562-cMyc -Global Average signal intensity	2.389	2400

4. Welch, z-Test and Wilcox statistics

As shown in section 3, we calculated the signal intensity from the TSS for both K562-cMyc and GM12878 in the [0-0.1kb], (0.1-1kb], (1-5kb], (5-10kb] and (10-20kb] interval from the TSS. The left hand side of the interval correspond to the minimum distance between peak and the TSS. The right hand side correspond to the maximum distance between the peak and the TSS.
For interest sake, we used the z-Test, Welch and Wilcox test to compare the mean between the two samples to determine if the mean between to the samples are statistically different.

4.1 Welch statistics

Upstream:

1. 0-0.1kb upstream

```
Welch Two Sample t-test

data:  gm12_1m and K562_01m
t = 3.3945, df = 23.787, p-value = 0.00241
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.520214 6.241762
sample estimates:
mean of x mean of y
 6.925720  3.044732
```

2. 0.1-1 kb upstream

```
Welch Two Sample t-test

data:  gm12_1m and K562_1m
t = 4.0272, df = 60.996, p-value = 0.0001587
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.826764 5.429881
sample estimates:
mean of x mean of y
 6.925720  3.297398
```

3. 1-5 kb upstream

```
Welch Two Sample t-test

data:  gm12_5m and K562_5m
t = 1.2298, df = 477.97, p-value = 0.2194
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06876499  0.29884180
sample estimates:
mean of x mean of y
 2.617382  2.502343
```

4. 5-10 kb upstream

```
Welch Two Sample t-test

data:  gm12_10m and K562_10m
t = 2.5005, df = 597.66, p-value = 0.01267
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 0.0319532 0.2658498
sample estimates:
mean of x mean of y
 2.358797  2.209896
```

5. 10-20 kb upstream

```
Welch Two Sample t-test

data:  gm12_20m and K562_20m
t = 3.8649, df = 1196.6, p-value = 0.0001171
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0733555 0.2246175
sample estimates:
mean of x mean of y
 2.281613  2.132626
```

Downstream:

1. 0-0.1 kb downstream

```
Welch Two Sample t-test

Welch Two Sample t-test

data:  gm12_01m and K562_01m
t = 1.9173, df = 5.1825, p-value = 0.1113
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.477667 24.770342
sample estimates:
mean of x mean of y
13.691070  3.044732
```

2. 0.1-1kb downstream

```
Welch Two Sample t-test

data:  gm12_1m and K562_1m
t = 4.0272, df = 60.996, p-value = 0.0001587
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.826764 5.429881
sample estimates:
mean of x mean of y
 6.925720  3.297398
```

3. 1-5 kb downstream

```
Welch Two Sample t-test

data:  gm12_5m and K562_5m
t = 1.2298, df = 477.97, p-value = 0.2194
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06876499  0.29884180
sample estimates:
mean of x mean of y
 2.617382  2.502343
```

4. 5-10 kb downstream

```
Welch Two Sample t-test

data:  gm12_10m and K562_10m
t = 2.5005, df = 597.66, p-value = 0.01267
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0319532 0.2658498
sample estimates:
mean of x mean of y
 2.358797  2.209896
```

5. 10-20 kb downstream/8

```
Welch Two Sample t-test

data:  gm12_20m and K562_20m
t = 3.8649, df = 1196.6, p-value = 0.0001171
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0733555 0.2246175
sample estimates:
mean of x mean of y
 2.281613  2.132626
```

4.2 z-Test

1.0 0-0.1 kb upstream (small number of samples therefore did t-Test)

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	13.69106965	3.0447323
Variance	181.6832121	3.316008289
Observations	6	6
Hypothesized Mean Difference	0	
Df	5	
t Stat	1.917303653	
P(T<=t) one-tail	0.056659895	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	0.113319789	
t Critical two-tail	2.570581836	

2. 0.1-1 kb upstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	6.92572045	3.297398215
Known Variance	40.74137	34.36008
Observations	54	54
Hypothesized Mean Difference	0	
Z	3.07665337	
P(Z<=z) one-tail	0.00104669	
z Critical one-tail	1.64485363	
P(Z<=z) two-tail	0.00209339	
z Critical two-tail	1.95996398	

3. 1-5kb upstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	2.617382	2.50336227
Known Variance	1.042287	1.087176
Observations	240	240
Hypothesized Mean Difference	0	
Z	1.210457	
P(Z<=z) one-tail	0.113052	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.226104	
z Critical two-tail	1.959964	

4. 5-10kb upstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	2.358797	2.209896
Known Variance	0.544556	0.519225
Observations	300	300
Hypothesized Mean Difference	0	
Z	2.500539	
P(Z<=z) one-tail	0.0062	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.0124	
z Critical two-tail	1.959964	

5. 10-20kb upstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	2.281613	2.132626
Known Variance	0.460907	0.430704
Observations	600	600
Hypothesized Mean Difference	0	
Z	3.864869	
P(Z<=z) one-tail	5.56E-05	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.000111	
z Critical two-tail	1.959964	

Downstream

1. 0-0.1 kb downstream (small number of samples therefore did t-Test)

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	15.48434	2.989824
Variance	242.5302	3.082639
Observations	6	6
Hypothesized Mean Difference	0	
df	5	
t Stat	1.952852	
P(T<=t) one-tail	0.054137	
t Critical one-tail	2.015048	
P(T<=t) two-tail	0.108274	
t Critical two-tail	2.570582	

2. 0.1-1 kb downstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	9.041993	3.041723
Known Variance	77.36758	2.667685
Observations	54	54
Hypothesized Mean Difference	0	
z	4.928639	
P(Z<=z) one-tail	4.14E-07	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	8.28E-07	
z Critical two-tail	1.959964	

3. 1-5kb downstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	2.844631	2.784962181
Known Variance	1.759337	1.817623
Observations	240	240
Hypothesized Mean Difference	0	
z	0.488758	
P(Z<=z) one-tail	0.312507	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.625013	
z Critical two-tail	1.959964	

4. 5-10kb downstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	2.578851	2.461215
Known Variance	1.053939	1.064203
Observations	300	300
Hypothesized Mean Difference	0	
z	1.399987	
P(Z<=z) one-tail	0.080759	
z Critical one-tail	1.644854	
P(Z<=z) two-tail	0.161517	
z Critical two-tail	1.959964	

5. 10-20kb downstream

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	2.440163	2.33338
Known Variance	0.803378	0.851494
Observations	600	600
Hypothesized Mean Difference	0	
z	2.033266	
P(Z<=z) one-tail	0.021013	

z Critical one-tail	1.644854
P(Z<=z) two-tail	0.042026
z Critical two-tail	1.959964

4.3 Wilcox Test

Upstream

1. 0-0.1kb upstream

```
Wilcoxon rank sum test

data:  aa$gm_0.1m and aa$k562_0.1m
W = 28, p-value = 0.132
alternative hypothesis: true location shift is not equal to 0
```

2. 0.1-1kb upstream

```
        Wilcoxon rank sum test with continuity correction

data:  aa$gm_1m and aa$k562_1m
W = 2079, p-value = 0.0001375
alternative hypothesis: true location shift is not equal to 0
```

3. 1-5kb upstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_5m and aa$k562_5m
W = 34125, p-value = 0.000458
alternative hypothesis: true location shift is not equal to 0
```

4. 5-10kb upstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_10m and aa$k562_10m
W = 58063, p-value = 7.622e-10
alternative hypothesis: true location shift is not equal to 0
```

5. 10-20kb upstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_20m and aa$k562_20m
W = 231370, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Downstream

1. 0-0.1kb downstream

```
Wilcox on rank sum test
```

```
data:  aa$gm_0.1p and aa$k562_0.1p
W = 28, p-value = 0.132
alternative hypothesis: true location shift is not equal to 0
```

2. 0.1-1kb downstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_1p and aa$k562_1p
W = 2175, p-value = 1.07e-05
alternative hypothesis: true location shift is not equal to 0
```

3. 1-5kb downstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_5p and aa$k562_5p
W = 31923, p-value = 0.03988
alternative hypothesis: true location shift is not equal to 0
```

4. 5-10kb downstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_10p and aa$k562_10p
W = 55282, p-value = 1.281e-06
alternative hypothesis: true location shift is not equal to 0
```

5. 10-20kb downstream

```
Wilcoxon rank sum test with continuity correction

data:  aa$gm_20p and aa$k562_20p
W = 221490, p-value = 4.772e-12
alternative hypothesis: true location shift is not equal to 0
```