Application of ChIP-Seq data analysis softwares in study of gene regulation

Application of ChIP-Seq data analysis softwares in study of gene regulation

MASTER THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF BIOLOGY

BY
JATIN TALWAR

Department of Anthropology and Human genetics- Faculty of Biology II, Ludwig-Maximilians-Universität München under the joint supervision of:

PROF. DR. WOLFGANG ENARD,
PROF. DR. REINHARD FAESSLER &
DR. BIANCA HABERMANN

Table of contents Table of contents

TABLE OF CONTENTS 2

TABLE OF FIGURES 4

ABBREVIATIONS 5

ABSTRACT 6

1. INTRODUCTION 8

1.1 ChIP-Seq 8
1.1.1 ChIP-Seq data analysis 10

1.2 Transcription factor binding 11
1.2.1 C-MYC 12
1.2.2 GATA1 15

2. METHODS 18

2.1 Raw Data procurement 18

2.2 Tools for analyzing raw data 22
2.2.1 RNA-Seq Pipeline 22
2.2.1.1 FastQC 25
2.2.1.2 Quality trimming: Cutadapt 25
2.2.1.3 Mapping: TopHat 25
2.2.2 Bam to BED conversion: Bedtools 26

2.3 ChIP-Seq analysis software (peak finding and visualization) 27
2.3.1 CoPrA (Comparative Profile Analyser) 27
2.3.2 EpiCenter 27
2.3.3 AnnoMiner 29

2.4 Data Visualization: UCSC Genome Browser 29

2.5 Enrichment analysis (GO-Elite) 30

3. RESULTS 32

3.1 Raw data analysis **32**
3.1.1 Quality Check 32
3.1.2 Quality trimming 33
3.1.3 Mapping 34

3.2 Finding differential peaks (CoPrA) **35**
3.2.1 Preprocessing part I 35
3.2.2 Preprocessing part II 35
3.2.3 Main CoPrA run 36

3.3 Further predicting differential peaks (EpiCenter) **37**
3.3.1 Human TF differential binding for MYC & GATA1 37
Sample 1: K562 vs. GM12878 (Human)
3.3.1.1 Annotating peaks to genes (AnnoMiner) 39
3.3.1.2 Gene List 40
3.3.1.3 Gene Ontology and pathway enrichment analysis 41
3.3.1.4 KEGG associations 43
3.3.1.5 Common genes and differential binding conclusions 45
3.3.1.6 peak data visualization 46

3.3.2 Mouse TF differential binding for MYC & GATA1 47
Sample 1: MEL (K562 Analogue) vs. CHX.12 (GM12878 analogue) (Mouse)
3.3.2.1 Annotating peaks to genes (AnnoMiner) 47
3.3.2.2 Gene List 47
3.3.2.3 Gene Ontology and pathway enrichment analysis 48
3.3.2.4 KEGG associations 52
3.3.2.5 peak data visualization 54

3.3.3 Common differential binding across Human And Mouse 54

4. Conclusions and Discussions **57**

5. References **60**

APPENDIX **66**

ACKNOWLEDGEMENTS **68**

STATEMENT OF ORIGINALITY **69**

Table of Figures

Figure 1.1 Experimental protocol for ChIP-Seq (Furey 2012) 5
Figure 1.2 C-MYC protein family architecture 6
Table 1.1 MYC regulation and its targets involved in transformation 7
Figure 1.3 MYC targeting Chromatin (Tansey 2014). 8
Figure 1.4 Functional aspects of GATA1 (Shimizu, Engel, and Yamamoto 2008) 9
Figure 1.5 GATA1 transcription factor gene architecture (Hitzler 2005) 10
Table 2.1 : Sample information and sources link 14
Figure 2.1 Schematics of RSAP workflow 15
Figure 2.2 TopHat Pipeline. The RNA-seq reads are mapped first to the whole genome (Trapnell, Pachter, and Salzberg 2009b) 25
Figure 2.3 Illustration for Epicenter analysis approach for ChIP-Seq data analysis (Huang et al. 2011) 26
Figure 2.4 An exemplar manual visualization session for peaks visualized using UCSC Genome browser 32
Figure 2.5 GO-Elite workflow and sources. 34
Figure 3.1 Per Base Sequence Quality- FastQC 35
Figure 3.2 Cutadapt output. 37
Figure 3.3 EpiCenter workflow for desired results 38
Figure 3.4 EpiCenter sample output 45
Figure 3.5 Base Pair Coverage of gene-intervals graph AnnoMiner. 43
Table 3.1 Table for GO terms enriched in K562_MYC sample. 45
Table 3.2 Table for GO terms enriched in GM12878_MYC sample 45
Table 3.3 Table for KEGG associations 46

Figure 3.6 Manual benchmark session UCSC genome browser (K562_MYC_vs_GM12878_MYC) 47
Figure 3.7 Base Pair Coverage of gene-intervals graph Annominer 49
Table 3.4 GO enrichment analysis for MEL-MYC mouse sample. 51
Table 3.5 GOelite results for CHX12-MYC sample 34
Table 3.6,3.7 KEGG associations for MEL_MYC & CHX12 sample 35
Figure 3.8 Manual benchmark session UCSC genome browser (MEL_MYC_vs_CHX12_MYC). 36
Table 3.8 Common genes across Human and Mouse samples that had MYC occupancy 45
Table 3.9 Common Genes GO enrichment 48

Abstract

Background:
Two pivotal factors regulating gene expression are transcription factors and epigenetic processes like histone modifications. ChIP-Seq is a powerful technique for cell specific detection and understanding of these regulatory interactions. Most of the previous studies in this regard have focused on single tissue, ignoring binding across multiple tissue types. In order to completely understand the transcriptional network, we need to focus studies on transcription factor (TF) binding events across cell types/time points. For this purpose technologies like ChIP-Seq & DNase-seq are monumental in studying binding of these TFs. Extending analysis across multiple species will help resolve evolutionarily constraint TF binding regions. It will also help delineate species-specific TF binding patterns.

Method:
Most of the tools available to date fail to consider replicates and control samples for accurate peak comparison. This study utilized EpiCenter, an algorithm for comparative profile analysis, which employs 3 independent statistical tests and sequence coverage normalization, along with a sliding window approach to detect differential binding events across samples REFERENCE!. *EpiCenter* was employed to study ChIP-Seq profiles of two transcription factors: MYC & GATA1. The binding events were studied across two different kinds of cell types: K562 & GM12878 both os which are immortalized cell lines derived from Human. Two different cell types from mouse MEL & CHX12 were also studied for TF binding. These profiles for MYC and GATA1 were also compared between Human and Mouse.

Results:
EpiCenter is a comparative profile analyzer for detection of differential binding events across a whole genome. Since it performs three different statistical tests, it exhibits higher precision in detection of differential binding events across samples. This study further tested its robustness in predicting differential TF binding. Gene ontology and KEGG was used to gain information about the function of genes targeting by MYC Comparison of enriched pathways provided information regarding tissue specific functionality of MYC. Finally comparison across species helped delineate the conserved binding domains for MYC and the species specific regions. Similar studies could turn out to be a big step towards understanding genome wide

TF regulation of complex transcriptional networks.

1 Introduction

1.1 ChIP-Seq

DNA-binding proteins are molecules that play a role in cellular processes like transcription, translation, splicing, replication and DNA repair. One major class are transcription factors, which bind specifically to motifs in DNA and regulate gene expression. Identifying the binding regions of these transcription factors is necessary to understand the regulatory functions carried out. Chromatin Immunoprecipitation coupled with oligonucleotide hybridization tiling array (ChIP-chip) (Kharchenko, Tolstorukov, and Park 2008; Valouev et al. 2008) and with ultra high-throughput sequencing (ChIP-Seq) has become a widely used technology to study transcription factor binding for the entire genome as well as the chromatin state of a cell (Histone modifications). ChIP-Seq is the most efficient way to identify binding sites for a single transcription factor or the location of histones(Furey 2012).

In a ChIP-Seq experiment the cells are initially treated with formaldehyde in order to cross-link proteins (transcription factors or histones) associated with DNA. The DNA is then sheared randomly (using endonucleases or sonication) in order to generate sub-kilo base double strands. A specific antibody directed towards the protein of interest is collected using the immunoprecipitation (IP) process Followed by amplification of selected fragments using PCR. These are called ChIP-fragments. In the ChIP-Seq protocol, adapters are ligated to both sides of these ChIP fragments to produce a library, which are then sequenced using massively parallel manner by next-generation sequencing machines.

ChIP-Seq offers multiple advantages over other techniques such as ChIP-chip. It has capability of single nucleotide resolution, higher coverage, exponentially reducing costs, relatively low amount of DNA requirement and has the possibility of multiplexing (Park 2009) (Ho et al. 2011). The biggest advantage of ChIP-Seq is that it provides the possibility of high-resolution profiling on a genome-wide level. Due to these reasons it has become a principle tool for gene-regulatory network profiling and the interaction

between transcriptome and the epigenome.

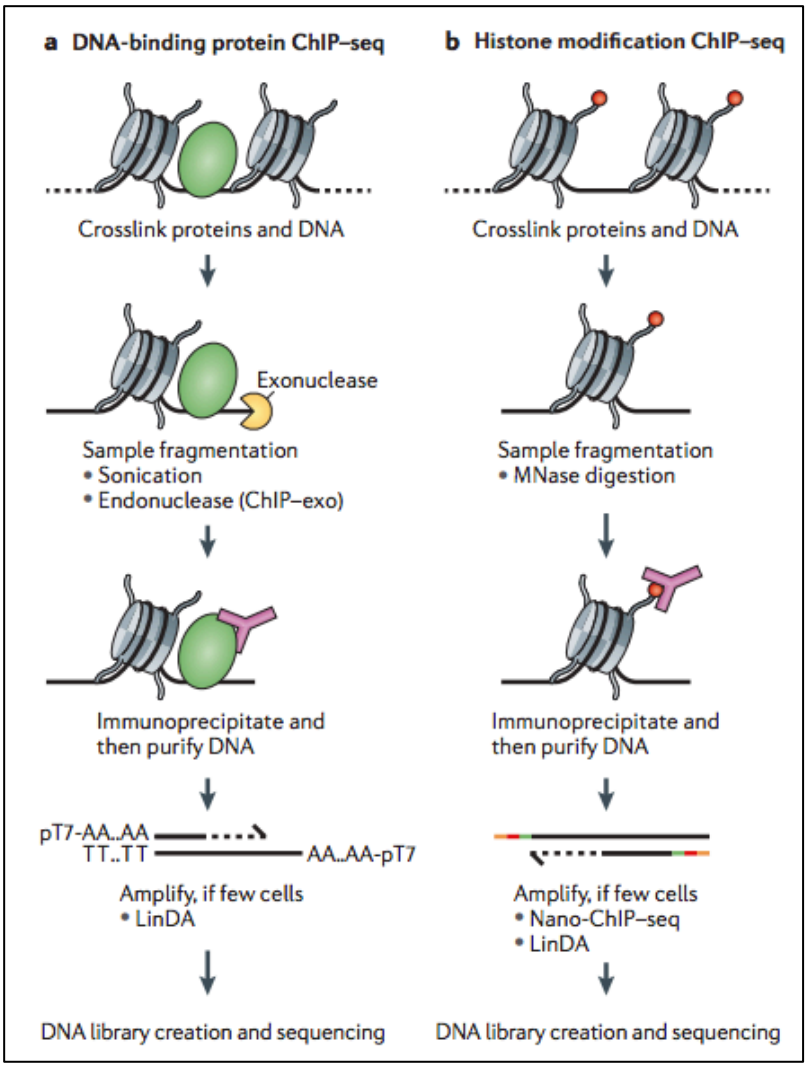

Figure 1.1 Experimental protocol for ChIP-Seq. Experimental procedures to detect DNA-binding proteins (transcription factors & histone modifications) are shown**. a|** Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) for transcription factors. Many recent advances in the technology have made this technique more robust and less prone to contaminating DNA**. b|** ChIP-Seq for histone modification uses the same strategy as ChIP-Seq for transcription factors except using micrococcal nuclease (MNase) ro fragment DNA. Image courtesy: (Furey 2012).

1.1.1 ChIP-Seq data analysis

Analysis of ChIP-Seq data requires computational tools, which can identify the differences in peaks accurately across samples. This can be challenging, given the fact that multiple biases arise from factors like sequencing depth, background normalization and proper statistics. Multiple tools have been developed recently, which interpret ChIP-Seq data like MACS (Feng, Liu, and Zhang 2011), MMDiff (Schweikert et al. 2013), diffReps (Shen et al. 2013). The primary aim of these tools is to predict the genomics regions that contain enriched read counts ("peaks"), where more sequences have been aligned than would be expected by chance. Many of these programs also used control normalization to remove background noise. However, the primary goal of ChIP-Seq studies is to compare data or binding across multiple conditions: for example assaying the binding of a transcription factor across two different samples to study cell-type specific response. Simply comparing the reads from two samples can be inaccurate due to inherent sequencing biases.

EpiCenter (Huang et al. 2011) aims to remove the sequencing biases while reporting regions differing across conditions in in the density of protein bound DNA while keeping the false discovery rate (FDR) at minimum. EpiCenter claims to perform multiple normalizations using their novel "parsimony" method for adjusting read coverage depths between samples. In order to achieve this, EpiCenter performs a series of statistical tests starting from filtering out the background regions, followed by two tests (exact ratio test and Z-test) for detecting significant changes. This makes EpiCenter efficient and robust in detecting differential peaks across samples.

1.2 Transcription factor binding

Cell fate and complex body functions are carried out by a succession of signals that are a part of complex and precise pattern of gene expression. Transcription factors are one of the major components of this process. Transcription factors along with co-factors form complexes that regulate the transcription of genes. Transcription factors that are sequence-specific identify consensus sequences on DNA (enhancer or promoter regions) for binding and initiating transcription.

Numerous diseases arise from the disruption of the transcriptional regulatory machine. For example, the overexpression of certain transcription factors can cause cancer (Furney et al. 2006). Moreover OMIM suggests that more than one third of all diseases have dysfunctional transcription factors associated to them (Hamosh et al. 2005). Furthermore, tissue specific binding of transcription factors and other binding variation might be a source of phenotypic diversity and evolutionary adaptation (De, Lopez-Bigas, and Teichmann 2008) (Lopez-Bigas, De, and Teichmann 2008).

In order to model and construct transcriptional regulatory networks, we need to study genome wide binding sites of transcription factors. ChIP-Seq has been extensively used for this purpose. Comparison between ChIP-Seq experiments can provide insights into differences in protein binding and histone modifications (Ji et al. 2013) (Ross-Innes et al. 2012). (Follows et al. 2003) recently carried out a comparison of chromatin structure and transcription factor occupancy at the human and mouse c-FMS loci. They showed that even though the distribution of chromatin modification and chromatin remodeling across both loci is highly similar, the transcription factor composition at the two-gene locus is different, suggesting a conservation of regulatory features between the mouse and human c-FMS locus.
Tissue specific transcription factor binding seems to be essential in regulating the temporal and spatial expression of genes. Integrating information from various studies of transcription factor binding will help shed light on specific transcriptional factor occupancy in different cell types or developmental stages.

1.2.1 C-MYC

The MYC family of proteins are basic-helix-loop-helix-leucine zipper transcription factors (Lüscher and Larsson 1999). They are one of the most studied proteins and are involved in cancer (Dang 2012). These proteins are majorly overexpressed in malignant tumors driving cell proliferation, growth, metabolism, DNA replication, cell cycle progression, adhesion and metastasis (Table 1.1). They are known to be deregulated in multiple cancer types via insertional mutagenesis, chromosomal translocations and gene amplifications (Meyer and Penn 2008). C-MYC functions as a direct regulator of gene expression via transcription factor activity and DNA replication (Adhikary and Eilers 2005; Cole and Cowling 2008; Lin et al. 2012) (Fig 1.1). One group showed that MYC together with its interacting partner MAX form a heterodimer and bind a CACGTGE-box sequence with high affinity. This binding in turn can activate transcription via multiple mechanisms (Blackwood and Eisenman 1991). MYC shows also increased transcription by recruiting RNA Polymerase II and promoting elongation through the PTEFb (positive transcription elongation factor) complex (Eberhardy and Farnham 2001, 2002).

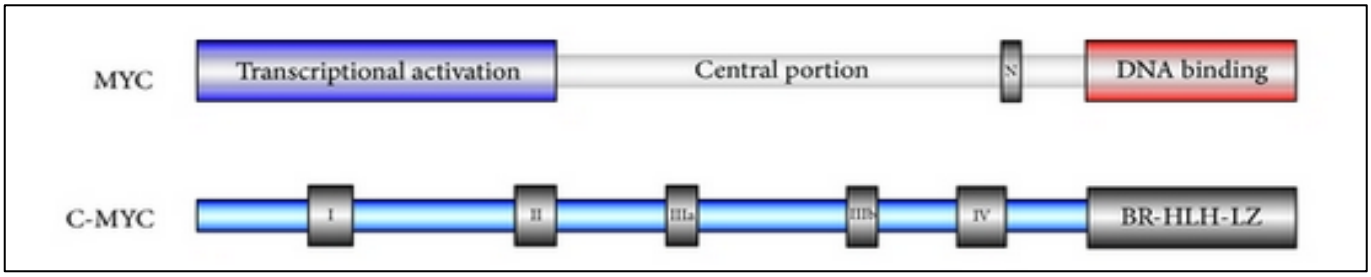

Figure 1.2 C-MYC protein family architecture. Transcriptional activation domain followed by a central position and a canonical nuclear localization sequence. C-MYC has a total of 439 amino acids. Image courtesy: (Tansey 2014)

Functional class	Description of function	Examples of responsible genes*
Cell cycle	MYC–ER activation drives quiescent cells to enter and transit through the cell cycle; primary cells from conditional knockout mice arrest in the absence of MYC expression	Cyclin D2, CDK4 (induced); p21, p15, GADD45 (repressed)
Differentiation	Deregulated MYC blocks differentiation of many cell systems; MYC accelerates epidermal differentiation	CEBP (repressed)
Cell growth, metabolism and protein synthesis	MYC expression levels are associated with body size owing to regulation of cell size and cell number	Lactate dehydrogenase, CAD, ODC, ribosomal proteins, EIF4E, EIF2A (induced)
Cell adhesion and migration	MYC drives tumorigenesis in part by allowing for anchorage-independent growth	N-cadherin, integrins (both repressed)
Angiogenesis	MYC induces angiogenesis in a wide range of tissues	IL1β, miR-17–92 microRNA cluster (induced), thrombospondin (repressed)
ROS, DNA breaks and chromosomal instability	MYC can contribute to instability, trigger telomere aggregation and increase ROS production	MAD2, TOP1, BUBR1, cyclin B1, MT-MCI
Stem cell self-renewal and/or differentiation	Ectopic MYC can potentiate induced pluripotent stem cells; MYC can control the balance between stem cell self-renewal and differentiation	To be determined, potentially genes associated with cell cycle, immortalization, adhesion and migration
Transformation	MYC can drive focus formation and anchorage-indepenent growth *in vitro* and full tumorigenesis *in vivo*; MYC is often deregulated in primary human cancers	Multiple targets are thought to contribute to transformation

Table 1.1 MYC regulation and its targets involved in transformation. Table courtesy: (Dang et al. 2006)

There have been several attempts to identify target genes for MYC (Bello-Fernandez, Packham, and Cleveland 1993). However, identifying gene targets via cycolhexamide treatment (Patel et al. 2004) were labor intensive and slow . With the advent of microarray expression studies, large-scale MYC-regulated genes could be analyzed at once. Even then the poor signal-to-noise ratio of microarray analysis exacerbated the target prediction. Only in recent years, Chromatin Immunoprecipitation (ChIP) followed by next generation sequencing (Seq) has allowed researchers to predict true targets of MYC. ChIP-Seq has enabled the researchers to look at genome wide targets of MYC with high sensitivity and specificity (Perna et al. 2012).Through early ChIP-Seq binding studies of MYC, it has been deciphered that MYC binds to approximately 10-15% (24,000 genomic sites) of the genomic locations unlike any other transcription factor (Dang et al. 2006). Most of the target genes were involved with cell cycle regulation (CDKs), protein synthesis, cell adhesion, metabolism, and RNA-binding factors (Lee and Dang 2006). Not only that, MYC was also shown to transcriptionally regulated non-protein coding genes like miRNAs.

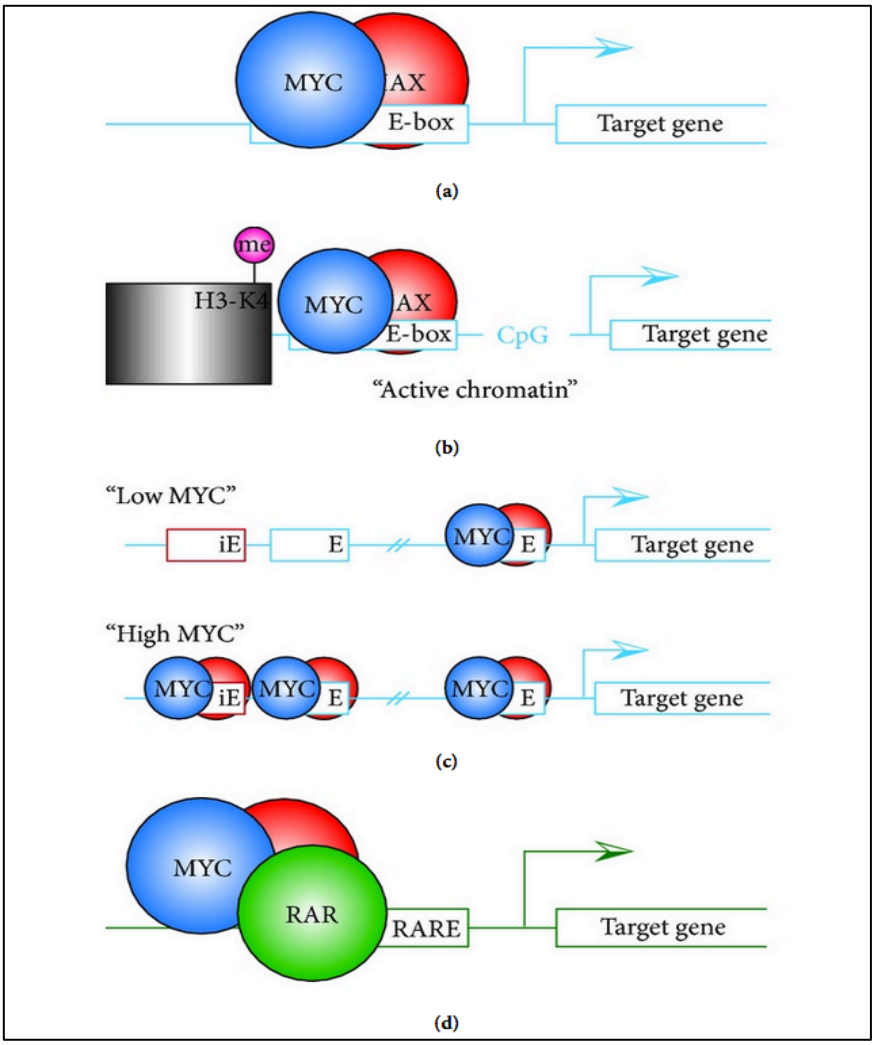

Figure 1.3 MYC targeting Chromatin. The four possible ways MYC targets genomic sites. (a) Binding by MYC/MAX is induced by sequence similarity to the E-Box. (b) in this case MYC/MAX dimers only bind to E-Box under certain chromatin states such as CpG islands. (c) Dosage specific binding by MYC. Under low MYC levels, MYC binds to promoter proximal to E-boxes (B-HLH-LZ motifs). At high levels of MYC, MYC/MAX dimer binds to consensus E-boxes but also bind to imperfect E-boxes ("iE"). (d) Recruitment of MYC via other transcription factors. Like in this case where MYC is recruited by retinoic acid receptor – α (RAR) to the DNA element (RARE) to regulate possibly new set of genes. Image courtesy (Tansey 2014).

Given the importance of MYC and its involvement in diverse and crucial cellular functions it is necessary to discover and study novel MYC targets. They could be used to study the gene expression pattern of multiple cancers.

1.2.2 GATA1

GATA1, also known as erythroid transcription factor, is a member of GATA transcription factor family. In mammals, the GATA family is composed of six members that are divided into two subfamilies depending on their expression and overall gene structure (Cantor and Orkin 2002; Lowry and Atchley 2000; Patient and McGhee 2002). Some of the members of this family (GATA1, GATA2, GATA3) are expressed specifically in the hematopoietic lineages (Shimizu and Yamamoto 2005).

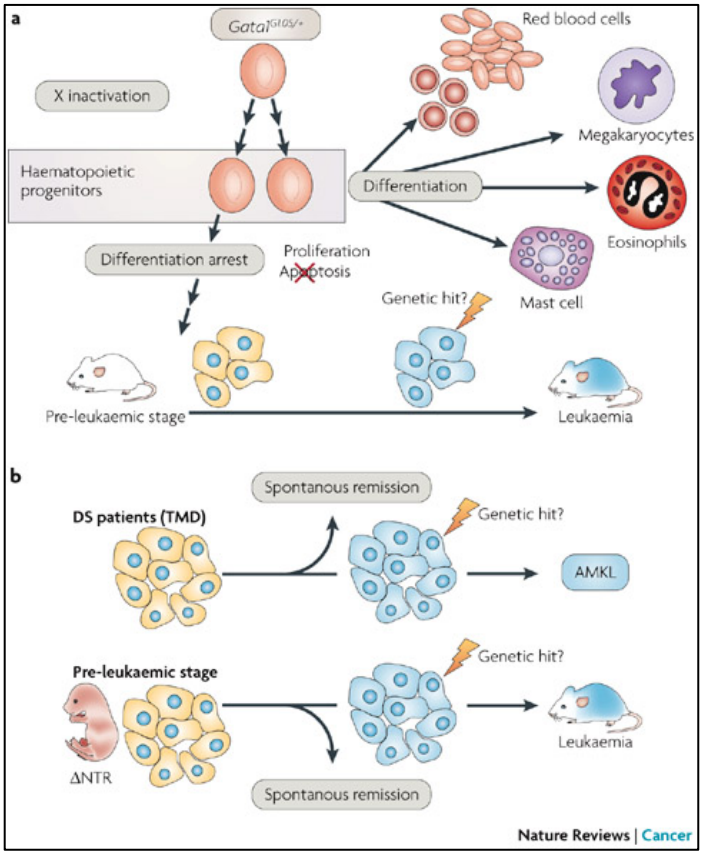

Figure 1.4 Functional aspects of GATA1. Two types of leukemia can be caused by GATA1. (a) Quantitative deficits of GATA1 in mice. (b) Qualitative defects of GATA1 in human Down syndrome (DS) patients. (Blue acute megakaryoblastic leukemia (AMKL). Image courtesy : (Shimizu, Engel, and Yamamoto 2008)

The GATA-1 protein contains multiple domains (C-finger, N finger, and N terminus) that

work as transcriptional activators (Ferreira et al. 2005). The C-finger domain mediates zinc finger sequence specific DNA binding. GATA1 Protein is typically expressed in differentiated erythrocytes, megakaryocytes, and eosinophil. The expression profile of GATA during erythroid cell differentiation shows a distinct pattern (Bresnick et al. 2005). GATA1 is known to interact with other transcription factors in regulating the expression of lineage specific genes (Rylski et al. 2003). Given the importance of GATA1 in hematopoiesis, mutations in this gene can cause defects that lead to hematopoietic disorders such as leukemia. GATA1 knock down embryos die in a very early stage due to anemia caused from impaired maturation of erythroid cells (Fujiwara et al. 1996). It has also been observed that children with Down syndrome, who develop acute megakaryoblastic leukemia, harbor GATA-1 mutations that reduce its expression (Muntean and Crispino 2005). However, there is still limited knowledge of the extent of damage caused by mutations or mis–expression of GATA1.

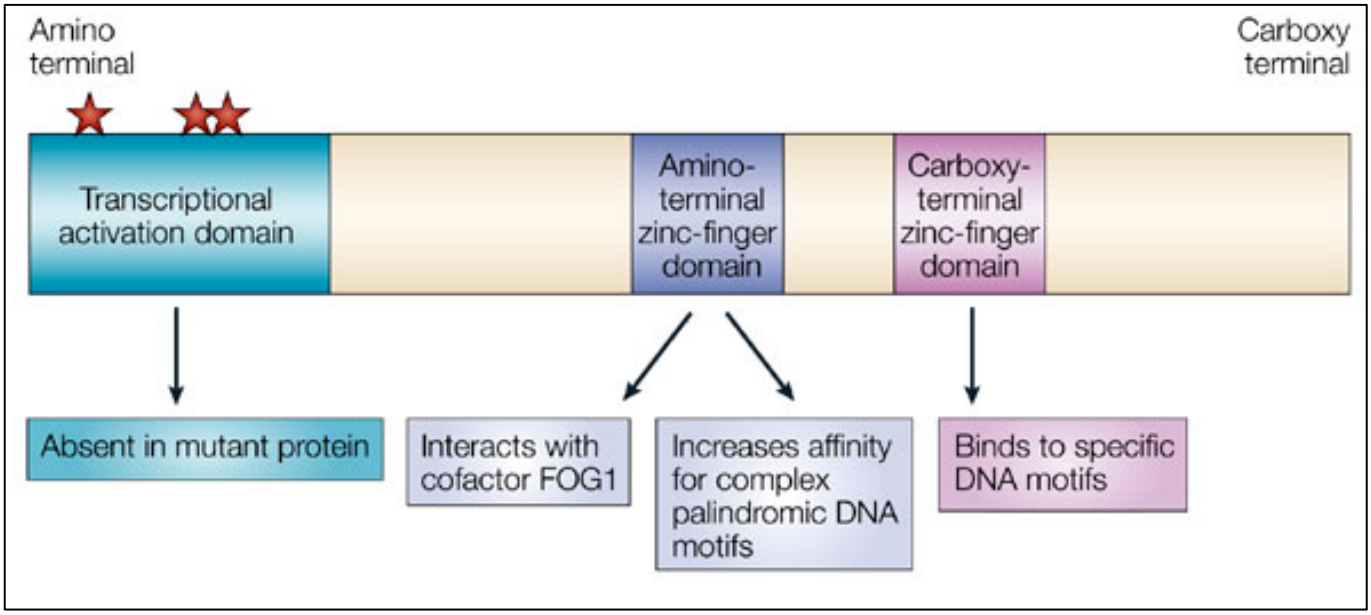

Figure 1.5 GATA1 transcription factor gene architecture. The gene contains a transcriptional activation domain. Stars represent the mutations in this region is absent in mutant protein due to presence of a premature stop codon. The gene also contains a carboxy-terminal zinc-finger domain that is required for binding to specific DNA motifs. The amino-terminal zinc-finger domain interacts with cofactors such as FOG1 and increases affinity for complex palindromic DNA morifs. Image courtesy (Hitzler and Zipursky 2005)

GATA family proteins recognize and bind to the consensus sequence (A/T)GATA(A/G) by two characteristic C4 (Cys-X2-Cys-X17-Cys-X2-Cys) zinc-finger motifs specific to this family (Ko and Engel 1993). These consensus sequences can be found in many regions of the genome. Specifically GATA-1 is known to target these genes: α- and β-globins (Evans, Reitman, and Felsenfeld 1988), Heme biosynthesis enzymes (Rylski et al. 2003), Erythropoietin receptor (EpoR) (Mitchell J. Weiss, Keller, and Orkin 1994), Bcl-Xl (Silva et al. 1996), other hematopoietic transcription factors like GATA2, MaFK and p45 NF-E2 (Shirihai et al. 2000), as well as cell cycle components and proliferation related genes like Cdks. Furthermore, GATA1 induced expression of growth inhibitors, including Btg2, Hipk2, JunB, and Crep and down regulated the expression of genes with mitogenic properties such as MYC, MYB, and Nab2 (M J Weiss, Yu, and Orkin 1997).

This study sets out to find differential binding of two transcription factors: GATA1 & MYC across two different cell types (K562 & GM12878 immortalized cell lines) in two species (Human and Mouse). Results show that the transcription factors bind to non-identical regions across the two tissues and the numbers of regions they bind to also differ. Moreover, the binding targets show tissue specific enrichment of biological processes. The results were further corroborated after integrating pathway information. Differential transcription factor occupancy could help delineate cell specific mechanisms that might play crucial roles in various phenotypes. It might lead to understand the behavior of these transcription factors and their transcriptional regulatory network.

2 Materials and Methods

2.1 Raw data procurement

In order to test the differential transcription factor binding across different conditions using CoPrA and EpiCenter, two exemplary datasets each for Human and Mouse were chosen from ENCODE. Each data set consisted of two sub-sets of ChIP-Seq data, derived from different cell lines were chosen for comparison. Since of the major features of CoPrA and EpiCenter is to take into account the replicates and control files for each data set, the respective files were also downloaded from the ENCODE website. As different transcription factors might have different binding sites across the same cell line, their binding profiles and thus the ChIP-Seq peaks can very. To test the ability of CoPrA to detect the differences in binding of TFs, two ChIP-Seq experiments were used from two different TFs: GATA1 & C-MYC (MYC).
The peaks of GATA1 binding is typically near the TSS and is represented as clear narrow peaks. In contrast, MYC is involved in cancer formation (Surget, Khoury, and Bourdon 2013) and functions as an oncogene. It has also been implicated as an important protein in regulating the genome regulation. MYC peaks are generally near the TSS of genes involved in DNA repair and cell cycle genes such as CDKs. Both the peaks are typically near +- 1000 bp of the TSS.
The data for the GATA1 and MYC was retrieved from two different cell lines:

1. **K562**- an immortalized cell line, derived from a female patient with chronic myelogenous leukemia (CML). The corresponding cell lines in mouse are also called **MEL.**
2. **GM12878**- an immortalized cell line produced from blood of a female donor with EBV transformation. It has a normal karyotype and develops well. Its mouse analogue is called **CH12**.

The advantages of using ENCODE datasets are the high quality standard for biological experiments, the availability of replicates, multiple experiments from the same lab, and consistent standards for processing of data.

Table 2.1 : Sample information and sources link

Transcription factor	Experimental condition & Link
GATA1	
GSM935601_Human_K562_ChIP-seq_input_RAW **Target:** Control,Input DNA **one replicates, 36 M reads for sample** **Project**: ENCODE **Data type :** Raw (fastq files) **Bed file name :** GSM935601_Human_K562_Input.bed	**Control** http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM935601 http://www.ncbi.nlm.nih.gov/sra?term=SRX150680
GSM1003608_Human_K562_ChIP-seq_GATA_RAW **Target: GATA1 ChIP-Seq** **Project**: ENCODE **1 replicate, 26 M reads each** **Data type :** Raw (fastq files) **Bed file name :** GSM1003608_Human_K562_ChIP-Seq_GATA1.bed	**Experiment** http://www.ncbi.nlm.nih.gov/sra?term=SRX186613
GSM912894_Mouse_K562_ChIP-seq_input_RAW **Target:** Control,Input DNA **one replicates, 56 M reads for sample** **Project**: ENCODE **Data type :** Raw (fastq files) **Bed file name :** GSM912894_Mouse_K562_Input.bed	**Control** http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM912894 http://www.ncbi.nlm.nih.gov/sra?term=SRX140357
GSM912907_Mouse_K562_ChIP-seq_GATA1_RAW **Target: GATA1 ChIP-Seq** **Project**: ENCODE **1 replicate, 23M & 28M reads** **Data type :** Raw (fastq files) **Bed file name :** GSM912907_Mouse_K562_ChIP-Seq_GATA1.bed	**Experiment** http://www.ncbi.nlm.nih.gov/sra?term=SRX140370 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM912907
GSM923581_Mouse_G1Erythrocytes_ChIP-	**Experiment**

seq_GATA1_RAW	
Target: GATA1 ChIP-Seq **one replicates, 32 M reads for single sample** **Project**: Individual study **Data type :** Raw (fastq files), mapped files **Bed file name :** GSM923581_Mouse_G1Ecells_GATA1.bed	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM923581 https://www.encodeproject.org/experiments/ENCSR000DIC/
Input ChIP-Seq: **one replicates, 20 M reads** **Data type : Raw files** **Bed_file name :** GSM946538_Mouse_G1E_Input_ChIP-Seq.bed	**Control** http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM946538 https://www.encodeproject.org/experiments/ENCSR000DJA/
MYC **Mouse**	
GSM912906_Mouse_K562_MYC_ChIP-Seq	**Experiment**
Target: MYC ChIP-Seq **0 replicates, 7 M reads** **Project**: Individual study **Data type : BAM** Files **Bed file name :GSM912906_Mouse_K562_MYC_ChIP-Seq.bed**	https://www.encodeproject.org/experiments/ENCSR000EUA/
GSM1003747_Mouse_K562_Input_ChIP-Seq	**Control**
Target: Input ChIP-Seq **0 replicates, 7 M reads** **Project**: Individual study **Data type : BAM** Files **Bed file name** **:GSM1003747_Mouse_K562_Input_ChIP-Seq.bed**	https://www.encodeproject.org/experiments/ENCSR000ADN/
GSM912906_Mouse_CHMX12_MYC_ChIP-Seq	**Experiment**
Target: MYC ChIP-Seq **0 replicates, 7 M reads** **Project**: Individual study **Data type : BAM** Files **Bed file name** **:GSM912906_Mouse_CHMX12_MYC_ChIP-Seq.bed**	https://www.encodeproject.org/experiments/ENCSR000ERN/

GSM912917_Mouse_CHMX12_Input_ChIP-Seq	**Control**
Target: INput ChIP-Seq **one replicates** **Project**: Individual study **Data type : BAM** Files **Bed file name** **:GSM912917_Mouse_CHMX12_Input_ChIP-Seq.bed**	https://www.encodeproject.org/experiments/ENCSR000ERS/
Human	
GSM935516_Human_K562_MYC_ChIP-Seq	**Experiment**
Target: MYC ChIP-Seq **one replicates** **Project**: Individual study **Data type : BAM** Files **Bed file name :GSM935516_Human_K562_MYC_ChIP-Seq.bed**	https://www.encodeproject.org/experiments/ENCSR000EGJ/
GSM935618_Human_K562_Input_ChIP-Seq	**control**
Target: Input ChIP-Seq **one replicates** **Project**: Individual study **Data type : BAM** Files **Bed file name** **:GSM935618_Human_K562_Input_ChIP-Seq.bed**	https://www.encodeproject.org/experiments/ENCSR000EHI/
GSM822290_Human_GM12878_MYC_ChIP-Seq	**Experiment**
Target: MYC ChIP-Seq **one replicates** **Project**: Individual study **Data type : BAM** Files **Bed file name** **:GSM822290_Human_GM12878_MYC_ChIP-Seq**	https://www.encodeproject.org/experiments/ENCSR000DKU/
GSM822292_Human_GM12878_Input_ChIP-Seq	**Control**
Target: Input ChIP-Seq **one replicates** **Project**: Individual study **Data type : BAM** Files **Bed file name** **:GSM822292_Human_GM12878_Input_ChIP-Seq**	https://www.encodeproject.org/experiments/ENCSR000DKW/
data procurement from : http://cistrome.org/db/#/	

All files were obtained from the ENCODE database in *SRA* format. The files were then converted to *fastq* format using the Bedtools package (Quinlan and Hall 2010). Since the transcription factor binding data is still in infancy we could only obtain a single replicate for most of the samples.

NOTE: GATA1 ChIP-Seq data was not of comparable quality to MYC ChIP-Seq data, having errors in the peak length (greater than 2000 bp) possibly due to experimental errors during Immunoprecipitation step (IP). Later it was checked that the authors of the data agreed to the substandard quality of the data. GATA1 was hence dropped from any further analysis.

2.2 Tools for analyzing raw data

2.2.1 NGS data analysis pipeline

The analysis of sequencing data for this project involved large sequencing files which needed to be analyzed first before using them for differential peak finders. In order to make the analysis faster and automated, a RNA-Seq Pipeline (RSAP) was generated, which takes fastq files as input and performs complete analysis on it with minimum input from the user. The workflow is user friendly with enough freedom for changing multiple parameters.

The script is able to check the quality of the data (FastQC & Cutadapt) (Andrews 2010)(Martin 2011), perform mapping (TopHat & STAR) (Dobin et al. 2013; Trapnell, Pachter, and Salzberg 2009b), quantification (HTSeq & FeatureCounts) (S. Anders, Pyl, and Huber 2014; Liao, Smyth, and Shi 2014) and differential expression analysis (DESeq) (Simon Anders and Huber 2010). At each step the data is stored in a directory form with individual specific folders. The RSAP workflow is an easy to use shell script for analyzing large scale NGS data. It is a semi automatic pipeline, which allows the user to access bioinformatics resources and tools without bioinformatics and IT skills.
The results can be easily browsed, exported or transferred to various resources. This

pipeline can be run from a local machine and does not need to be installed as a program or software. The user needs to provide the path for the pipeline in a shell environment. The user also needs to provide the parameter files for each step after which the user can stop and check results at each of the individual steps. The broad variety of options and parameters selection makes this pipeline very useful for analyzing data not only for this study but also for NGS data in general.

The pipeline is available as a set of scripts in a zip file downloadable from: https://github.com/jatintalwar/RNA-Seq-Analysis-Pipeline-/tree/master/Roman_data_new_analysis). The implementation is in a shell environment and the user can run the pipeline with a simple shell command (*e.g.: sh pipeline.sh*). The user needs to allocate the desired amount of space for the results files to be stored. Once the user chooses a particular location in the system/server, the pipeline will create the directory structure at the same location and run from that location from there on. The pipeline will require user input wherever needed and will prompt for the same inside the terminal. The script stops after each analysis step for the user to check the results produced by the last step and to decide about the parameters for the next. The script also asks the user to provide the parameter file in a particular format (specified in the pipeline). If the user wishes to quit the pipeline at moment, it can be done via a simple *exit* command. A sample output of the script run is available at: https://github.com/jatintalwar/RNA-Seq-Analysis-Pipeline-/blob/master/Roman_data_new_analysis/Sample_outputs.txt. More detail about the tools implemented by the pipeline is available in the following section.

The raw data for the CHIP-Seq (Table 2.1) was analyzed using this pipeline. Files downloaded from ENCODE were fed into the pipeline for Quality trimming of low quality reads, followed by mapping to reference genome. After mapping the data, the files were converted into appropriate format (*.BED*) and compared across samples for differential transcription factor binding.

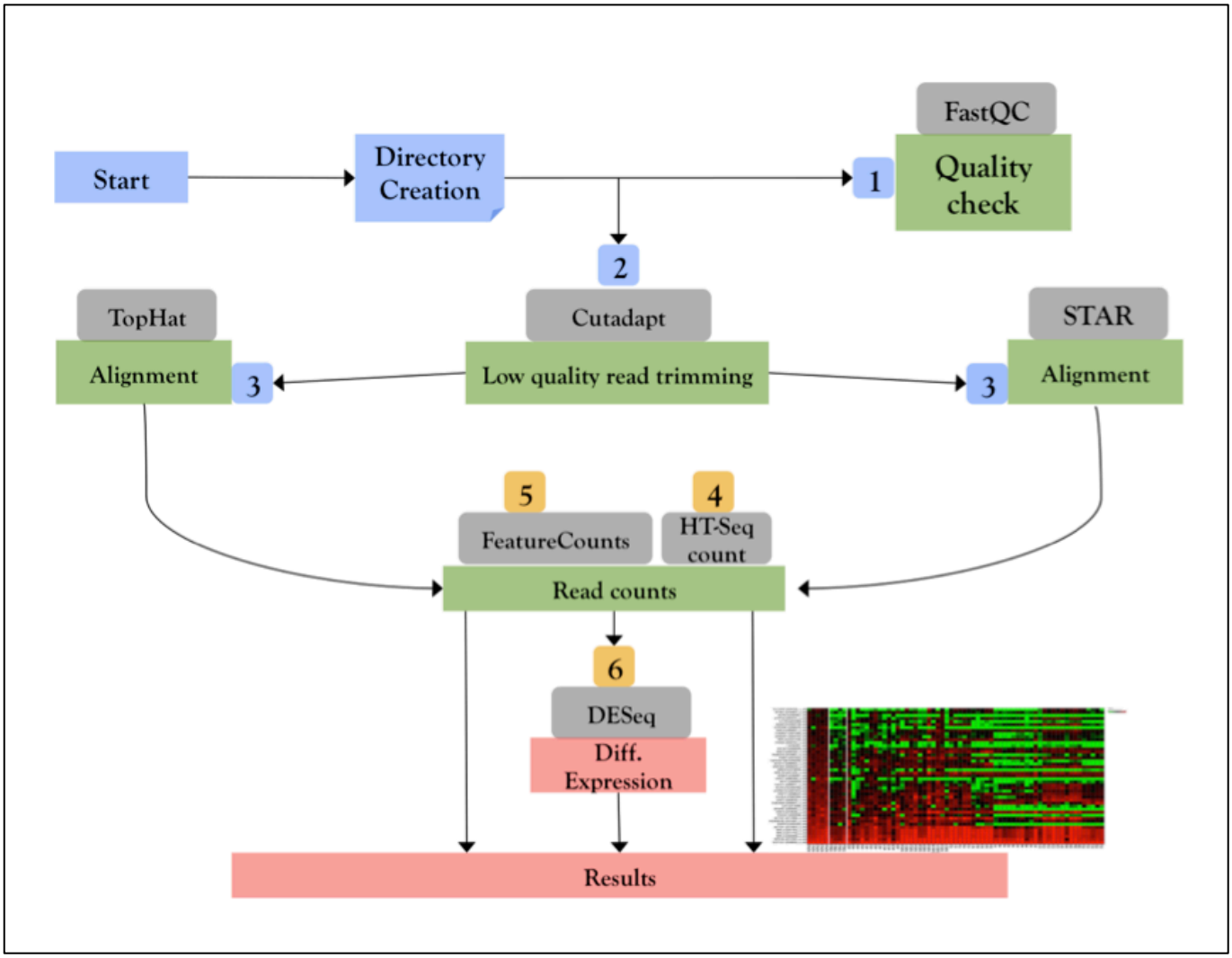

Figure 2.1 Schematics of NGS data analysis Pipeline NGSdp workflow. After directory creation the first step of the pipeline is quality check using FastQC (step 1). Then pipeline performs low quality read trimming (step 2). Mapping is step 3 (TopHat & STAR). Reads that are mapped are quantified using HTSeq & FeatureCounts (step4 &5). Differential expression is performed using DESeq (step 6). The heatmaps, PCA plots and MA plots are also generated during DESeq run.

2.2.1.1 FastQC:

The first critical step in data analysis is always an effective and reliable data Quality

check. FastQC (Andrews 2010) is a widely used tool for NGS data quality testing. It provides users with multiple options to check the sequencing qualities such as: per base quality, duplication levels, per sequence GC content and sequence length distributions. The tool also allows removing low quality reads and other contaminants. High quality reads are then used for further analysis steps.

2.2.1.2 Quality trimming: Cutadapt

The second step in the analysis pipeline involves trimming the reads for quality. For this purpose Cutadapt (Martin 2011) was used, In order to trim the sequencing reads. Cutadapt provides multiple parameters to choose from. For example, the minimum length of reads, trimming from both ends, trimming sequencing adapter contamination etc. The tool was used with default settings and the low quality read filter was set to Phred score of 25 and a minimum sequence length of 20 bp.

2.2.1.3 Mapping: TopHat

Mapping was performed with eukaryotic genomes in mind. Hence this study employed a mapping tool, which could predict and annotate splicing events. TopHat (Trapnell, Pachter, and Salzberg 2009b)(Trapnell et al. 2012)(Trapnell, Pachter, and Salzberg 2009a) was chosen for this purpose. TopHat is a widely used, tested and reliable aligning tool. The tool provides the option of splice junction detection. Multiple mapping hits was set to 3 and a GTF files (*igenomes*) was used while mapping files to human and mouse genomes to increase the specificity and sensitivity of the mapping.

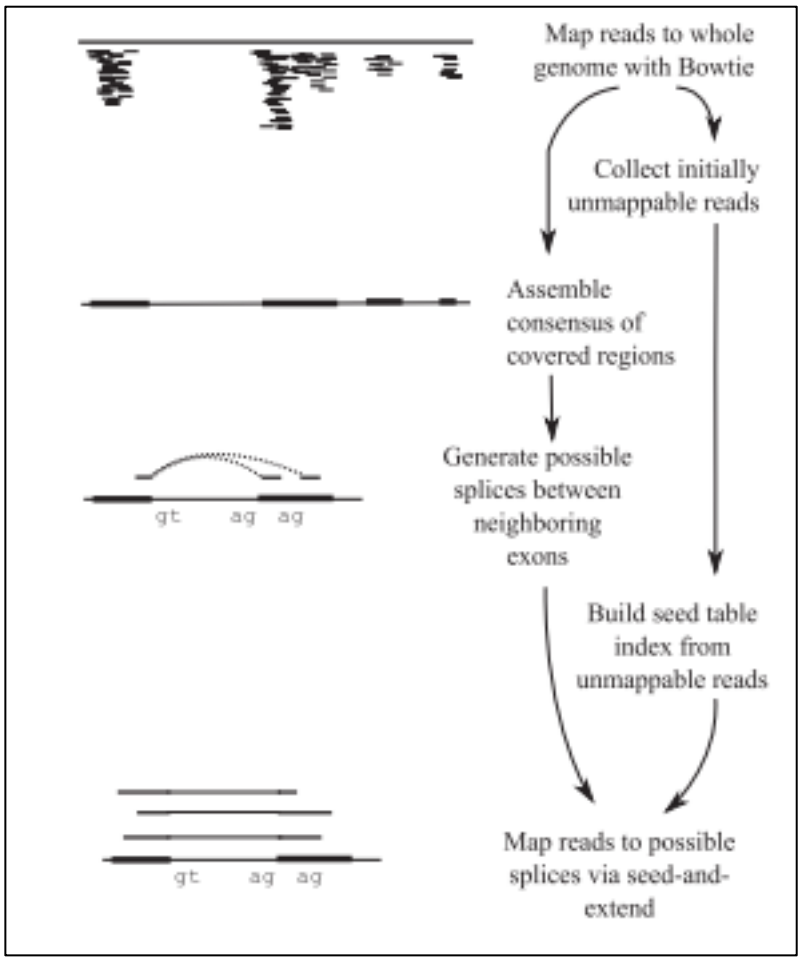

Figure 2.2 TopHat Pipeline. The RNA-seq reads are mapped first to the whole genome. The reads that do not align are set aside to be mapped later. Sequencing that flank donor/acceptor splice sites are then joined to form splice junctions. The initially unmapped reads are then aligned to these splice junction sequences. Image courtesy (Trapnell, Pachter, and Salzberg 2009b)

2.2.2 Bam to BED conversion: Bedtools

BAM files generated after mapping were directly converted to BED files using the Bedtools package (Quinlan and Hall 2010) with a simple *BamToBed* command. Mapping coverage of BAM files was checked using the *genomcov* command.

2.3 ChIP-Seq analysis software (peak finding & visualization)

2.3.1 CoPrA (Comparative Profile Analyzer):

CoPrA (Corinna Klein et al., unpublished) is a python based differential peak finder for studying chromatin states between two samples of ChIP-Seq experiments. It is designed to overcome the deficits like false prediction of peaks (peak accumulation in certain regions), taking into account the background files by preprocessing replicates and control data. It is a novel algorithm for comparative profile analysis, which employs a peak calling independent, sliding window approach to detect differential binding events across samples. In CoPrA, ideas from stock market analysis have been adapted to compare two peak profiles. In order to minimize false-positive predictions, CoPrA also takes into account the replicates and control samples for each condition. More details about the tool is beyond the scope of this study as the tool is yet to be published. CoPrA provides freedom in selecting parameters like: selecting cutoff values used to filter the raw difference, values determining minimal difference region length, and significance level of corrected P-Values.

2.3.2 EpiCenter

In order to test for biases in the analysis from CoPrA, the study also employed EpiCenter to compare the results obtained and to further detect any differences in peaks across the samples. EpiCenter (Huang et al. 2011) is able to detect differential changes in epigenetic marks by comparing the profiles of two ChIP-Seq samples. Its complex algorithms can account for signals from histone modifications and transcription factor binding events. It provides different normalization procedures for the user to choose from. It also provides with three different statistical procedures (*Z-test, Bonferroni correction, and exact ratio test*) to reduce FDR by also minimizing background noise. It provides the option to choose from a fixed-size

window, semi-dynamic window and a full dynamic window for the analysis. It also allows the user to select the maximum allowed gap distance between two reads. Epicenter accepts a BAM file as an input for the analysis and requires no further preprocessing.

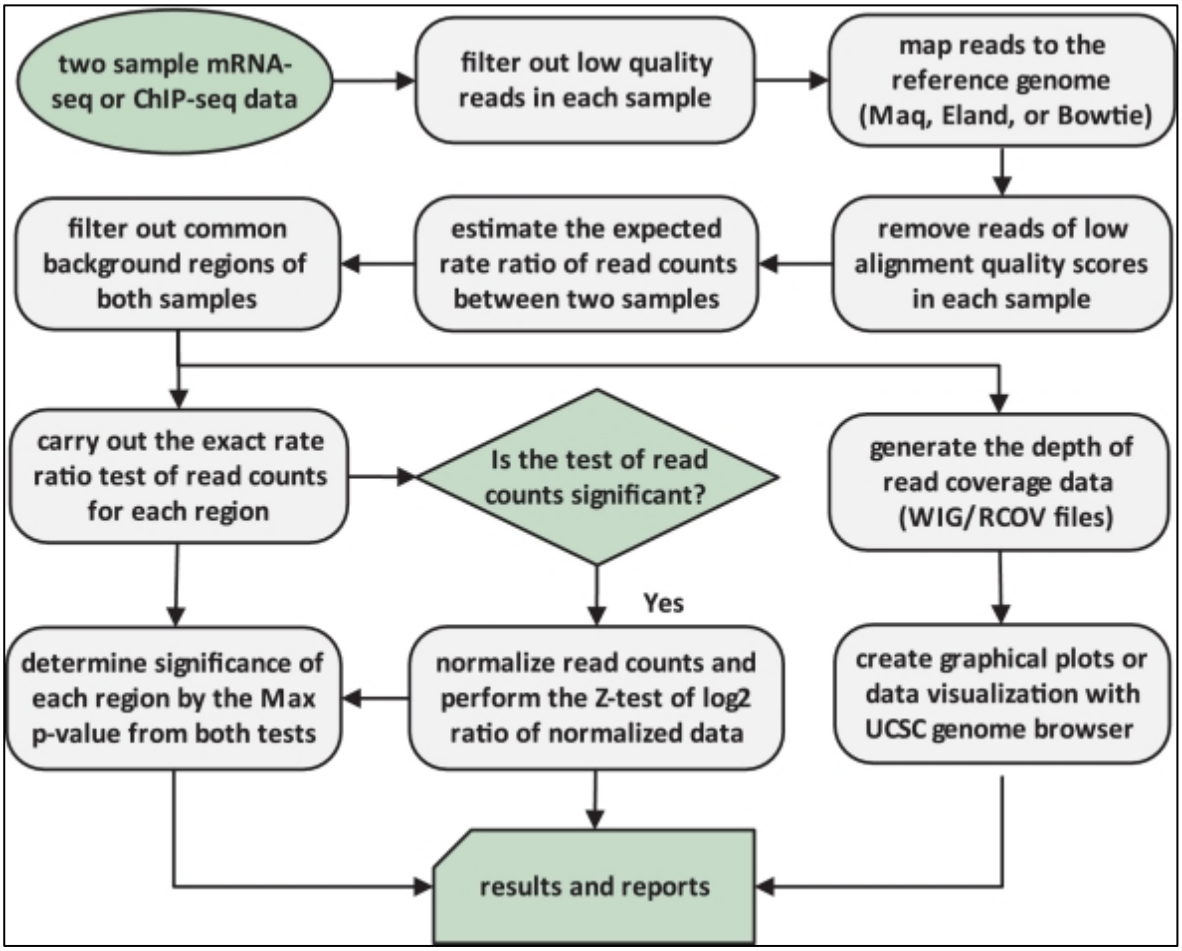

Figure 2.3 Illustration for Epicenter analysis approach for ChIP-Seq data analysis. Image courtesy : (Huang et al. 2011).

2.3.3 AnnoMiner

The output of both the differential peak finders (*CoPrA & EpiCenter*) was a BED file with first three rows representing the location of peaks on the genome (*Chr, start, stop*) and the last column being the numerical values for the peaks (*intensity*). However, it is imperative to know to which gene a peak belongs to. For this purpose Annominer (Arno Meiler, et al., unpublished) was employed. AnnoMiner accepts a BED file and the organism information and provides a list of genes the peaks might be associated with. The user can select the expected window size of the peaks as well as a window for upstream and downstream regions from TSS. The user can also choose between transcript based or a gene based search.
AnnoMiner can be found at: http://vm2-annominer:8080/AnnoMiner/
It is an in-house developed tool and is still under testing. It will be published later this year via Sourceforge.net

2.4 Data visualization: UCSC genome Browser

The UCSC genome browser is an open-source web based genome browser for visualizing genomic data. The user can upload its own data for visualization. A typical ChIP-Seq file consists of read density, indicating accumulation of mapped reads (“peaks”). In order to improve the visualization, the control files were also uploaded. After adding the required tracks to the session from UCSC browser, the session for comparison across samples was then completed.

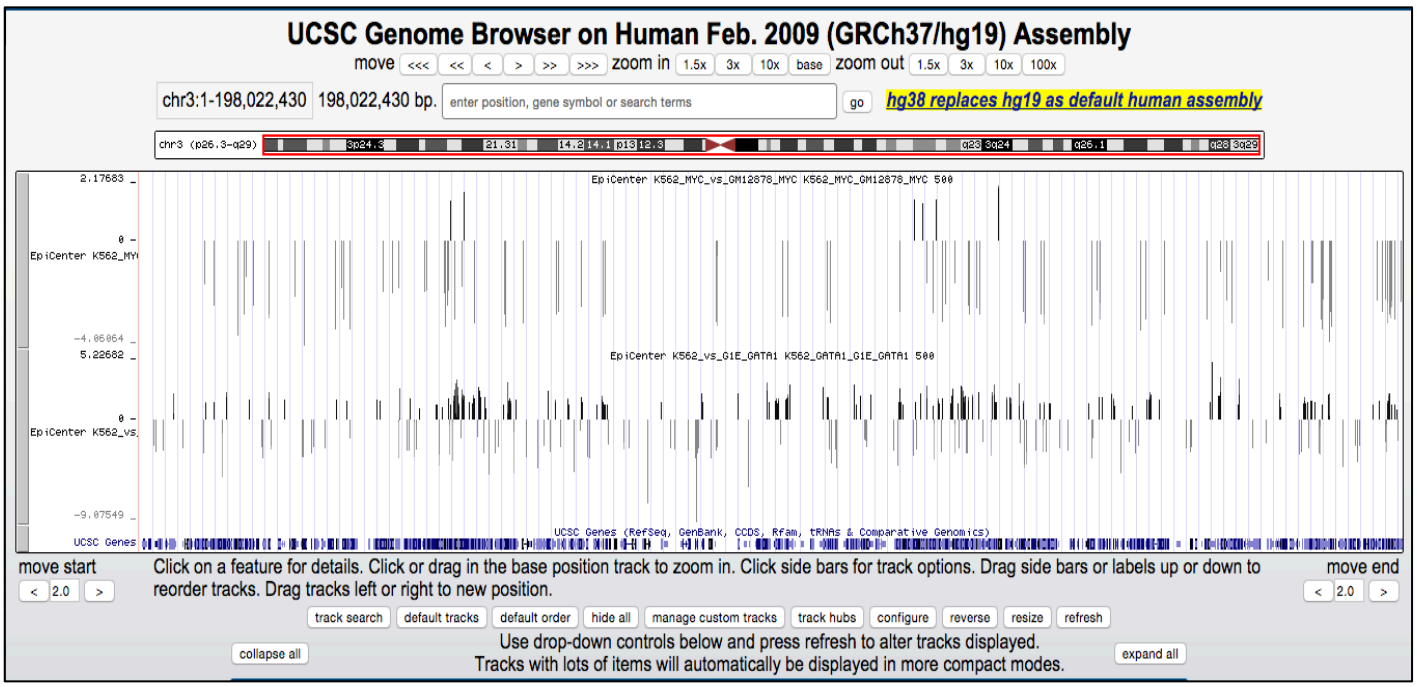

Figure 2.4 An exemplar manual visualization session for peaks visualized using UCSC Genome browser. The output tracks from Epicenter were loaded along with UCSC genes to view the genome-wide peak coverage. Chromosome 3 in this case

2.5 Enrichment analysis (GOelite)

In order to compare the gene lists from the different samples for enrichment of specific KEGG Pathways and Gene ontology terms, GO-Elite (Zambon et al. 2012) was used for enrichment analysis.

GO-Elite identifies a non-redundant set of biological ontology terms or pathways to describe the set containing multiple genes. Its diverse resources (ontology databases, WikiPathways, KEGG, Pathway Commons, microRNA target database and cellular biomarkers) make it a valuable tool for enrichment studies. It provides the freedom to run the analysis via a GUI (graphical user interphase) or through the command line. GO-Elite requires one input file and a denominator file. The input file must contain the identifiers (IDs) to be examined for enrichment, along with a system ID code. The denominator file must contain “ALL” IDs along with a system ID code. GO-Elite performs

over representation analysis (ORA). It also provides an option for pruning the results.

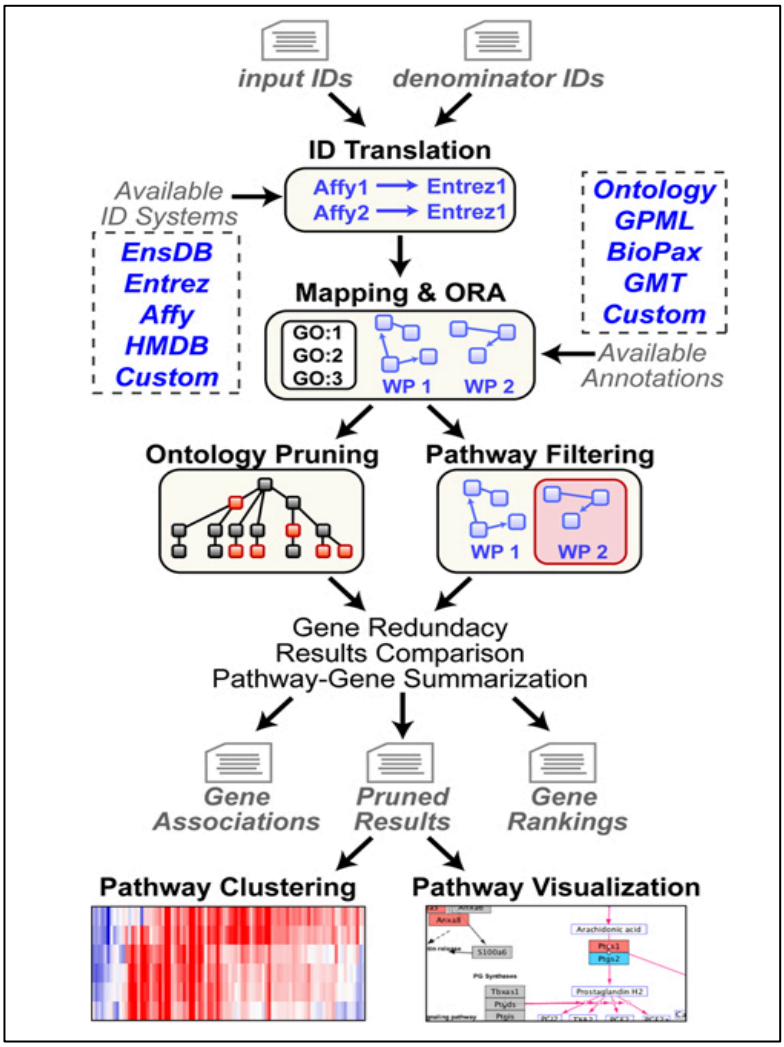

Figure 2.5 GO-Elite workflow and sources. Two text files (Input & Denominator) are provided by the user to begin the analysis. The Ids in these files are mapped to the system IDs in the databases (ENSEMBL, EntrezGene,). ORA is performed in the gene-set and the filtered pathways information is generated as output files (Gene-associations, Pruned results, Gene rankings). The pathways can be viewed on Go-Elite with a plugin or viewed on other external platforms. Image courtesy: (Zambon et al. 2012)

3 Results

3.1 Raw data analysis

3.1.1 Quality check

To check the sequence read quality, FastQC (Andrews 2010) was run on all the raw fastq files. The overall quality was of the data was very good, most of the reads having a PHRED score >= 28. It is very well known that the read quality decreases along the 5' end of the read. Still the read quality never went below Phred 28.

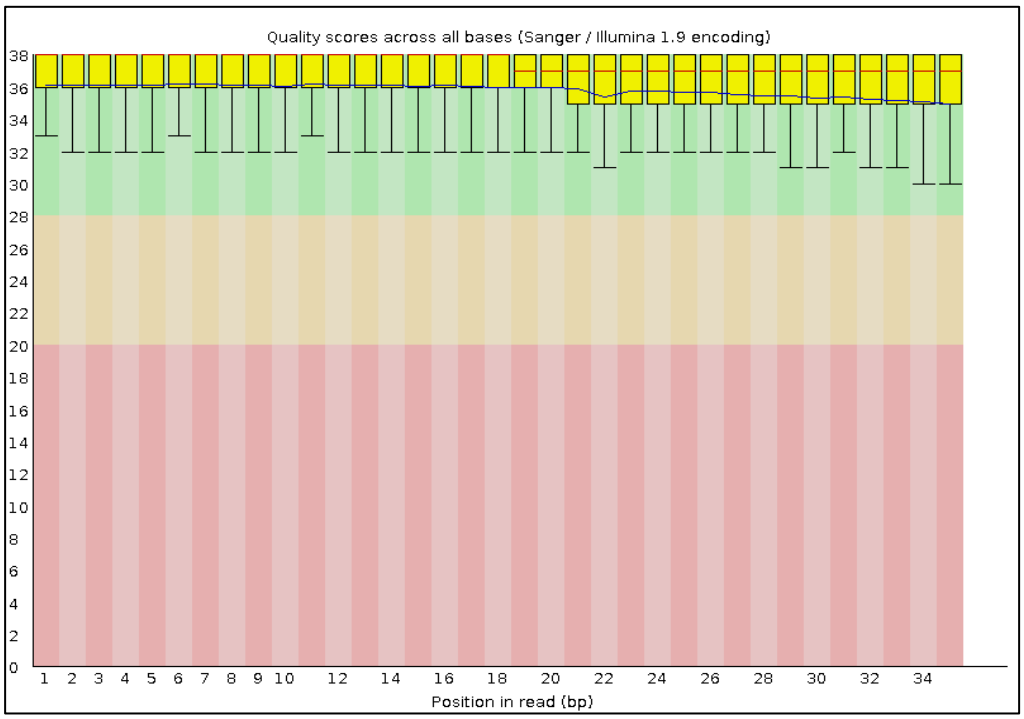

Figure 3.1 Per Base Sequence Quality- FastQC. A sample representation of quality of data as shown by FastQC. *GSM647222_Mouse_ESCells_Input_ChIP-Seq*. Notice how all the bases have a PHRED score of >=28 (green zone) indicating that they are of good quality and can be mapped without the need of further trimming.

Similarly, other quality checks like per sequence quality scores, per base sequence

content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, over represented sequences & Kmer content were within the acceptable limits.

FastQC was run with default parameters and all the input files were run from the RNA-Seq pipeline generated for the study. They can be found in the Appendix section. The results of other FastQC runs can be found on the web link for the supporting information.

3.1.2 Quality Trimming

Read quality trimming is one of the most used preprocessing procedures during analysis. Trimming aims at removing the low quality region of a read while preserving the longest high quality part. Trimming has been shown to increase quality and reliability of NGS analysis (Del Fabbro et al. 2013) It also saves time and computational power. The study employed Cutadapt (Martin 2011) for quality trimming of reads. Cutadapt was made a part of the RNA-Seq pipeline and the user has to provide parameters like quality cutoff and minimum sequence length as a part of a parameter file readable by the pipeline. Reads (30-40 M reads) for all the samples were trimmed successfully until the whole read had a PHRED score of >=28. Since Cutadapt automatically searches for the presence of Illumina Adapter, no input adapter sequence for trimming was provided. Error rates were kept at 10.0% and the tool was run in single end mode (see methods).

Summary	
Total Reads processed:	32,993,502
Reads with adapters:	0 (0.0%)
Reads that were too short	146,997 (0.4%)
Reads written (passing filter):	32,846,505 (99.6%)
Total base pairs processed:	1,154,772,570 bp
Quality-trimmed:	8,818,803 (0.8%)
Total written (filtered):	1,144,788,616 bp (99.1%)

Figure 3.2 Cutadapt output. A sample output from Cutadapt for *GSM647222_Mouse_ESCells_Input_ChIP-Seq.* Summary

view provides insight into the run and output results from Cutadapt. The quality-trimmed bases (0.8%) signal the good quality of data.

3.1.3 Mapping

Mapping was performed using TopHat (Trapnell, Pachter, and Salzberg 2009b) which is implemented in the RNA-Seq pipeline. Maximum multi-hits were kept at 3, and a GTF file along with replicates if any was provided for accurate alignment. The files were mapped to the reference genome with an accuracy of 95% for most of the samples. After the mapping, the BAM files were further tested for coverage using the BedTools *genomcov* option. The samples showed uniform distribution of reads across the whole genome without any biases at particular locations. Mapped files (BAM) along with mapping summary are available as supplementary information at this link: https://drive.google.com/drive/folders/0B_MVmsAk2E6MUXA3WWVCdW1SeEE

3.2 Finding differential peaks (CoPrA)

3.2.1 Preprocessing step I

In order to remove biases like sequencing depth and number of reads across samples, CoPrA performs a preprocessing step. The preprocessing step I generates files, which are in BED format with the first three columns specifying chromosome, start and stop position of the read respectively. The last two columns are the random representation/name of the read and the directionality respectively. This file is filtered for baseline values. The file generated will look like:
A.*GSM981238_ESCells_Input_ChIP_Seq_Human.bed_longReads_uniq.bed_singleFiltered.bed.*
B. GSM1003608_K562_ChIP-Seq_GATA1_Human.bed_longReads_uniq.bed.
Once the files have been generated the next (preprocessing Step II) step can be run.

3.2.2 Preprocessing step II

This step creates the file that will be used by CoPrA for the main comparisons.
It uses the *GSM981238_ESCells_Input_ChIP-Seq_Human.bed_longReads_uniq.bed_singleFiltered.bed* File as an input and generates a file with read locations in each chromosome. Each file is thus a txt file with two columns, the first one specifying location in the chromosome and the next one a value with the number of reads for that location. This step also generates a coverage frequency file, which contains the information about the number of reads for each score value. All of these files are later used by CoPrA when comparing binding events for two samples.

3.2.3 Main CoPrA run

CoPrA uses the files from preprocessing steps as input along with values for few other parameters:
1. Config_file: A text file with the file names for sample to be analyzed, given sample name, background filter value (this value is decided after studying the coverage files from preprocessing step II).
2. Chromosome information: this will be a file with chromosome names and their sizes (same file as used in preprocessing steps)
3. Step size: this is probably one of the most important parameters. The use provides a discretization step size for the samples.
4. Region length filter value: Value that determines the minimal difference region length, used for filtering the difference regions
5. Alpha 1 values: Significance level of the corrected p-value (=q-value) of difference regions up to which results should be reported.
Sample CoPrA run command: *python CoPrA_with_Control.py -i config.txt -d ../../ -o ../../Copra_out_s_50_r_100_corinna_code/ -f ../../human.hg19.genome -s 50 -l A549_vs_AG04449_hahn_data -r 100 -a 0.05*

NOTE: Due to errors in software, CoPrA could not produce result files. It was dropped from the study and other differential peak analysis was used for comparisons.

3.3 Predicting differential peaks using EpiCenter Run

3.3.1 Human TF differential binding for MYC

Sample 1: K562 vs. GM12878

EpiCenter identifies genome-wide epigenetic changes or TF binding events across various scenarios. It also provides multiple normalization methods and a series of 3 statistical tests for estimating background regions and allowing adjustment for multiple testing to control False Discovery rate. In the previous studies Epicenter performed better (Olivier Hahn et al. unpublished) than other differential peak finders.
In order to test for differential TF binding of C-MYC, EpiCenter was employed. The study found that there were considerable differences in binding of MYC across the two different cell lines (K562 & GM12878). The following steps were performed in order to get the desired results:

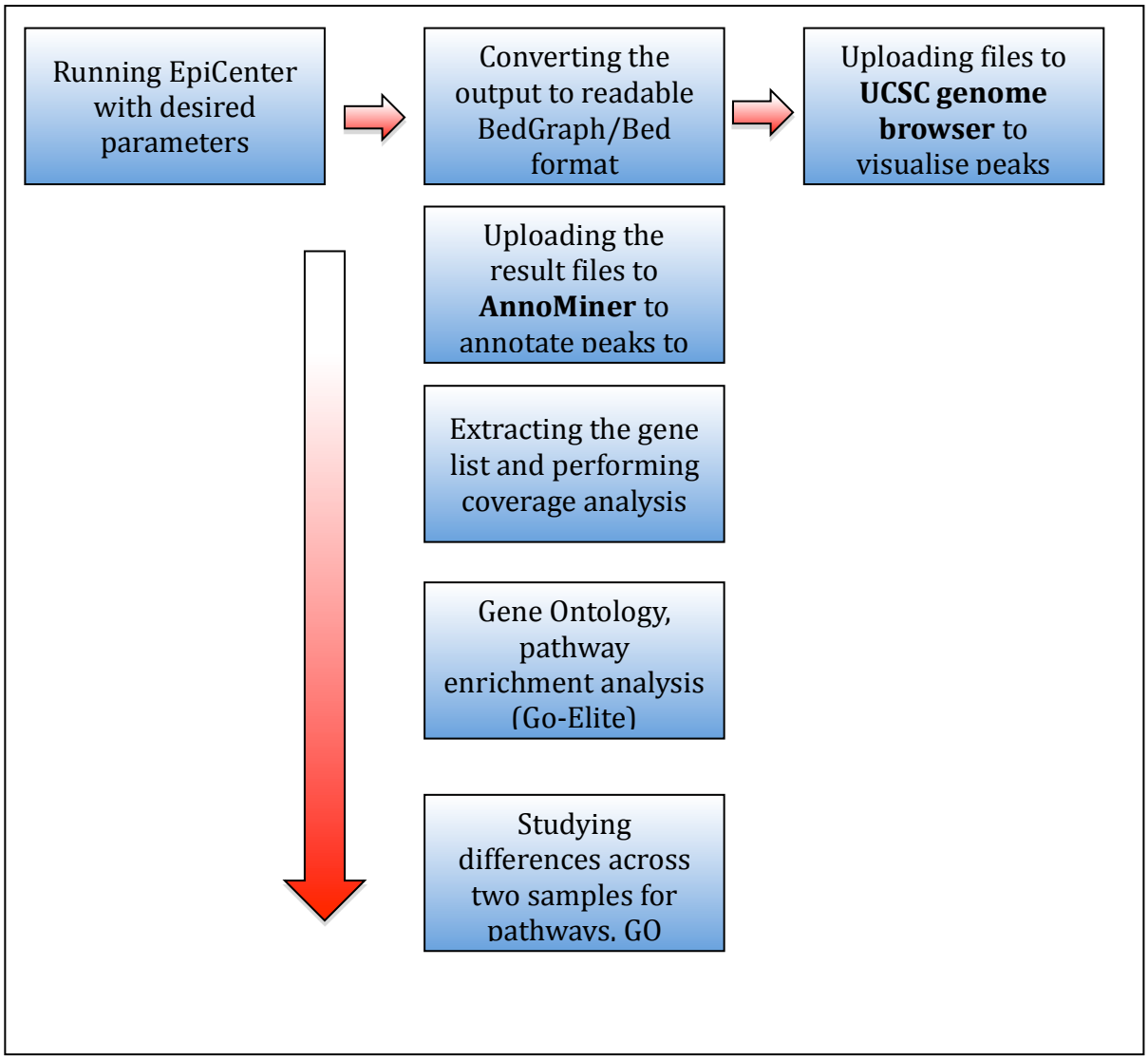

Figure 3.3 EpiCenter workflow for desired results

MYC binding data for K562 and GM12878 was compared with Epicenter (semi-dynamic window size of 500). A total of 162704 tests were performed across the two samples, giving an indication already for the presence of significant peaks to compare. The number of data points used for the estimation of standard deviation for Log2ratio null distribution were: 24555.

```
Number of tests:    162704
Cutoff False Discovery Rate (based on Benjamini-Hochberg Method):   0.05
FWR control for the exact rate ratio test (p_rr)
  Significant level:    :0.05
  Bonferroni correction:    3.07307e-07
  Sidak correction: 3.15255e-07
The estimated SD of log2ratio distribution: 1.20922
The estimated SD of log2ratio NULL distribution:    0.655466
    The number of data points for the estimation:   24555
The numbers of significant genes/regions (based on the FDR cutoff 0.05 ) are:
    the exact rate ratio test: 1965
output file is ./GSM935516_Human_K562_MYC_ChIP-Seq_GSM822290_Human_GM12878_MYC_ChIP-Seq-500.tscan
```

Figure 3.4 EpiCenter sample output. EpiCenter performs 3 statistical tests (bonferroni correction, Sidak correction & exact ration tests) for removing any variation/bias from background. The out put files (.tscan) can easily be converted into bed/bedGraph format with simple python script. Epicenter was run with various test parameters to choose the best ones for requirement of this study.

3.3.1.2 Annotating peaks to genes (AnnoMiner)

Output result files from EpiCenter were uploaded to AnnoMiner in order to annotate peak location to nearby genes. AnnoMiner needs a *.bed* file as input and provides gene associations to potential peaks. AnnoMiner contains the complete database information from ENSEMBL. Since all the further analysis is gene based, this option was selected for this analysis. Peak regions was kept to 150bp (transcription factor binding footprint) and gene-flanking region interval was set to 500 bp.

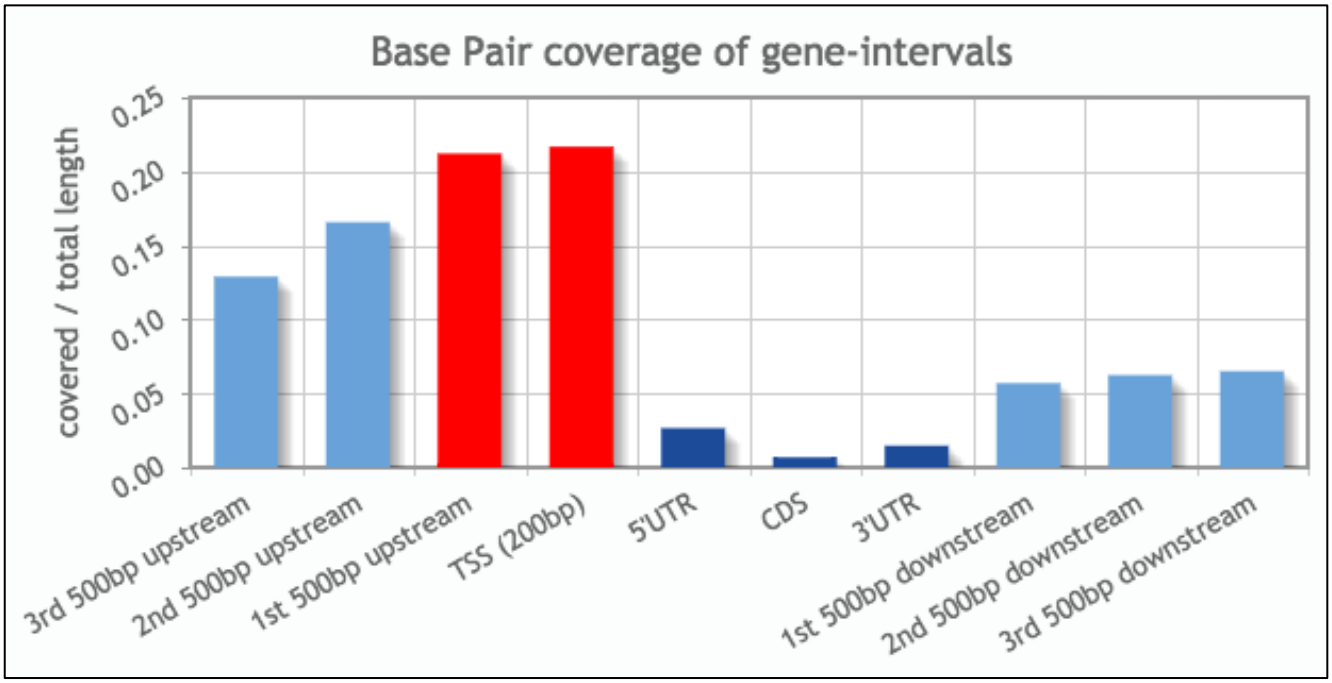

Figure 3.5 Base Pair Coverage of gene-intervals graph Annominer. EpiCenter results from MYC ChIP-Seq data for GM12878 were uploaded to AnnoMiner to check for gene associations. X-axis represents different gene-regions. The intensity of the bars is an average representation of peaks. Each bars represents the percentage of peaks falling in that part of genes. As was expected highest number of binding events are around TSS (200bp-500 bp Upstream). The red highlighted bars represent the regions for which further analysis was performed. Since the rest of the regions were not of relevance. A total of 208 genes were extracted from the list.

3.3.1.3 Gene List

AnnoMiner provides a gene list as an output (*.tsv format*), which contains the peak information (*start, stop*) and associated gene (*Chr, start, stop, geneId*). For final analysis this output file was combined with Epicenter output file and a final file with peak values integrated was generated. This file contained all the peaks, their associated genes, and the peak values for each peak. (See appendix)

3.3.1.4 Gene ontology & Pathway Enrichment analysis

Gene List generated from AnnoMiner contained the complete gene information and the respective peaks. This file was used as an input for GO-Elite. Denominator files were also provided, which contained all gene Ids from ENSEMBL. Go-Elite compares the input file with Denominator file and performs enrichment/overrepresentation analysis. It maps the gene identifiers to databases to look for GO associations and Pathway information (KEGG). The study found following GO and Pathways from the MYC binding in K562:

Many of the pathways enriched in K562_MYC sample were associated with cell cycle progression, and translation (initiation, termination). Since MYC is already known to be a transcriptional regulator, this result confirmed that predicting already known targets were identified. However, many novel pathways were also among those that were enriched.

GO TERM	FREQUENCY OF GENES ASSOCIATED
ATP_binding	66
RNA_binding	60
Gene_expression	48
Translation	30
ncRNA_metabolic_process	28
Cell_cycle_phase	27
Mitochondrial_inner_membrane	26
Viral_reproduction	24
ATPase_activity	23
Soluble_fraction	21
Ribosome	20
Tanscription_cofactor_activity	20
Translational_initiation	19
Structural_constituent_of_ribosome	18
Mitotic_cell_cycle	17
Respiratory_electron_transport_chain	17
SRP-dependent_protein_targeting_to_membrane	15
Translational_elongation	15

Table3.1 Histogram for GO terms enriched in K562_MYC sample. The highest number of genes belongs to "ATP binding" (66 genes) indicating that MYC is responsible for regulating transcription for multiple genes associated with mitochondrial pathways. Then was "RNA-binding" which points to MYC targeting genes with RNA-binding capability. These were genes like "*MARS, RLP14, EIF4G3*". Cytokine mediated cell-signaling pathway, cell cycle progression pathways are the ones that MYC has been extensively studies to be associated with. All of these have been previously associated with genes transcriptionally regulated by MYC. Note: only top 20 GO terms are shown here. For complete table see appendix.

GO TERM	FREQUENCY OF GENES ASSOCIATED
Membrane	20
Cytoplasm	16
Plasma_membrane_part	10
Regulation_of_molecular_function	8
Regulation_of_response_to_stimulus	8
Phosphorylation	7
Cytokine-mediated_signaling_pathway	6
Kinase_activity	6
Phosphotransferase_activity, alcohol_group_as_acceptor	6
Regulation_of_phosphorus_metabolic_process	6
Cell_projection	5
Defense_response	5
Enzyme_regulator_activity	5
Negative_regulation_of_macromolecule_metabolic_process	5
Regulation_of_cell_proliferation	5
Regulation_of_protein_modification_process	5
Small_GTPase_mediated_signal_transduction	5
Cell_activation	4

Table 3.2 Histogram for GO terms enriched in GM12878_MYC sample. The most enriched terms included Regulation of metabolic function, phosphorylation, cytokine binding, cytokine-mediated cell signaling, regulation of cell proliferation, regulation of molecular function, and kinase activity. Note: only top 20 GO terms are shown here. For complete table see appendix.

For example: “ATP binding” (GO:0005524) was highly enriched with 65 predicted target genes associated with this GO term. MYC has previously been thought to dictate the transcriptional regulation of ATP binding genes in leukemic samples (Porro et al. 2011) and these targets were never confirmed via follow up studies. This study accurately predicted 65 target genes associated with ATP binding with high confidence (avg. P-Value: 0.04). Similarly the GO term “RNA binding” (GO:0003723) was second most enriched term. Recently (David et al. 2010) showed that C-MYC regulates hnRNP1 & hnRNP2 which are two well established RNA binding proteins. This study also found C-MYC regulating ‘HNRNPA3’ (P-Value: 0.000472444) further confirming the association between the two. The study also found RNA-binding proteins (See appendix for the complete list), which are not yet known to interact with C-MYC according to literature. Follow up research for these factors will further shed light on the extent of regulation of RNA binding proteins by C-MYC.

Other enriched pathways included the mitochondrial processes “mitochondrial_DNA_metabolic_process, mitochondrial_inner_membrane, mitochondrial_matrix, mitochondrial_nucleoid, and NADH_dehydrogenase_activity”. Previous reports of C-MYC regulating mitochondrial genes (Li et al. 2005; Yu et al. 2008) have been limited to mitochondrial biogenesis genes (TFAM & NRF). However, in this study multiple genes involved in mitochondrial processes were enriched (See Appendix). The family of genes with highest associations to a GO terms were RPL (Ribosomal Protein Family) 13 genes with a total of 275 GO terms.

GM12878_MYC sample had fewer predicted binding sites (Table 3.2) compared to K562_MYC (230 vs. 1737 peaks resp.) hence the following enrichment analysis using GOelite also produced fewer GO terms (For complete list, see Appendix). The most enriched terms included phosphorylation, cytokine binding, cytokine-mediated cell signaling, regulation of cell proliferation, regulation of molecular function, regulation of metabolism, and kinase activity.

3.3.1.5 KEGG associations:

GO-Elite has the option of providing pathway information along with GO terms. A plot was generated for all the enriched KEGG terms enriched in each of the sample. As a result for the K562_MYC sample, a number of pathways associated with mental illness

were enriched: Alzheimer's disease, Parkinson's disease, and Huntington's disease. When observed closely the list of genes that were enriched in these pathways it was found that most of the genes (70%) were mitochondrial genes (MT-CO & MT-ATP genes). Consequently the other pathway highly enriched was 'Oxidative phosphorylation' (Table 3.3 and 3.4).

KEGG TERM	FREQUENCY OF GENES ASSOCIATED
Parkinson's_disease	18
Oxidative_phosphorylation	16
Ribosome	14
Huntington's_disease	12
RNA_transport	12
Alzheimer's_disease	11
Purine_metabolism	10
Pyrimidine_metabolism	9
Aminoacyl-tRNA_biosynthesis	7
DNA_replication	5
Homologous_recombination	4

Table 3.3 Histogram for KEGG associations in K562_MYC sample. Pathways related to mental illnesses were highly enriched. Most of the genes involved in these pathways were mitochondrial (MT-CO, MT-ATP etc).

KEGG TERM	FREQUENCY OF GENES ASSOCIATED
Cytokine-cytokine_receptor_interaction	4
Regulation_of_actin_cytoskeleton	4
Tuberculosis	4
Calcium_signaling_pathway	3
Jak-STAT_signaling_pathway	3
Natural_killer_cell_mediated_cytotoxicity	3
Neurotrophin_signaling_pathway	3

Table 3.4 Histogram for KEGG associations in GM12878_MYC sample. Two pathways were of specific interest here which involve apoptosis or immune response: "Jak-STAT signaling pathway" & "Natural killer cell mediated cytotoxicity". Genes from the InterLeukin family (IL21R, llL21b) were seen enriched.

3.3.1.6 Common genes & Differential binding conclusions

Furthermore, in order to check the overlap (if any) between the binding regions for the two samples, genes with common binding were selected and analyzed for binding value of MYC. It was interesting to find only two genes that were found for this overlap: CENPM (Centromere Protein M) (K562 value: 1.734, GM12878 binding value: 1.329) & DEPDC5 (DEP domain containing protein 5) (K562 value: 1.781, GM12878 binding value: 2.8). The fact that only few genes had overlap across the two tissues, there might be a mechanism of tissue specific binding for MYC.

3.3.1.7 Peak data visualization (UCSC)

In addition to comparing peak information via EpiCenter and further by Gene Ontology, peak data from Epicenter was also visualized for manually checking the binding regions for MYC across K562 & GM12878. The output files from EpiCenter were converted to '*.BedGraph'* format (using a python script). These files were then uploaded to UCSC genome browser for a manual benchmark session.

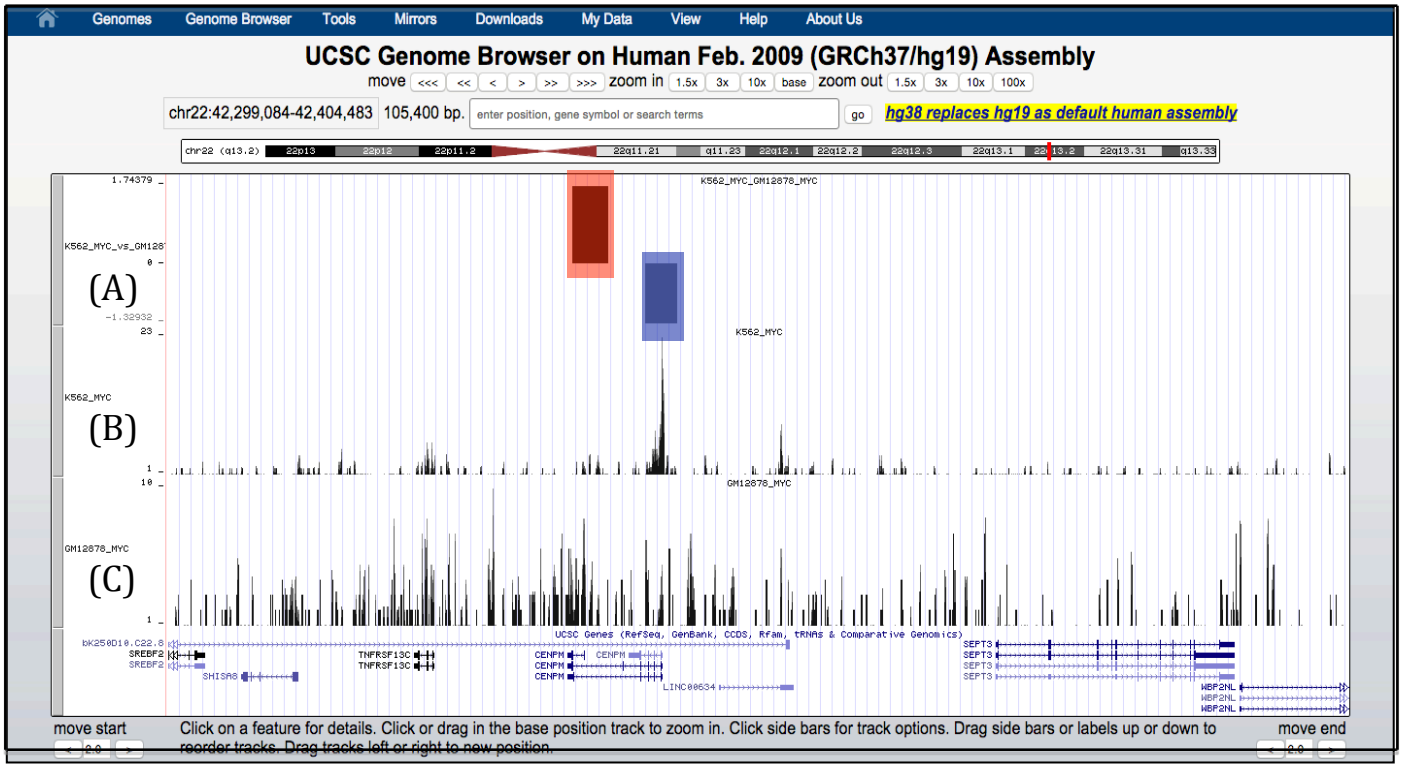

Figure 3.6 Manual benchmark session UCSC genome browser (K562_MYC_vs_GM12878_MYC). One sample benchmarking session of EpiCenter output (A) for comparison of MYC binding across K562 (B) And GM12878 (C) distributed over gene sites (CENPM in this case). For reference, the raw bed files (pre normalization) were also added to the session to compare the raw read counts for a particular reads. The Black rectangular bar on top (highlighted red) represents the peak value in GM12878 (1.743) and the grey bar (highlighted purple represents the peak value in K562 (1.3293) sample respectively. The peak can be seen as the average of read counts (intensities) for each sample minus the noise. Epicenter does automatic normalization for read depth hence the final comparisons (A) are without any sequencing biases.

3.3.2 Mouse TF differential binding MYC

Sample 2: MEL (K562 analogue) VS CHX.12 (GM12878 analogue)

To compare the differential binding of transcription factors across species, Epicenter was run with mouse cell lines for binding of MYC in two different types of immortal cell lines (MEL (Human K562 analogue) & CHX12 (Human GM12878 analogue)). Since the two cell lines were similar to the cell lines used for human samples, we could ideally compare the binding for MYC across two cell lines to look for either conserved binding domains or novel binding area. Two mouse samples (MEL & CHX12) were run with EpiCenter to study genome-wide binding of MYC. The same procedure was followed as with the human samples to get the desired results (see section 3.3.1).

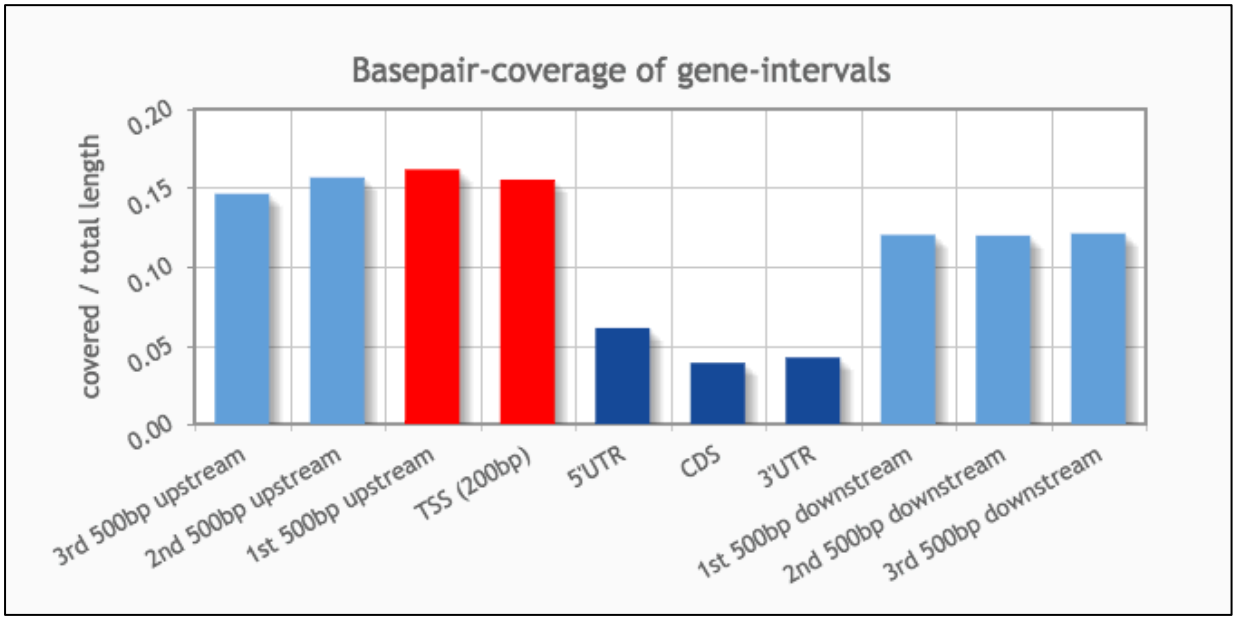

Figure 3.7 Base Pair Coverage of gene-intervals graph Annominer. EpiCenter results from MYC ChIP-Seq data for MEL were uploaded to AnnoMiner to check for gene associations. Each bars represents the percentage of peaks falling in that part of gene region. As was observed in Human samples, highest number of binding events is around TSS (200bp-500 bp Upstream). The red highlighted bars represent the regions for which further analysis was performed.

Since the rest of the regions were not of relevance. A total of 2974 genes were extracted from the list (A 10 fold increase from Human samples).

Compared to the human samples, mouse samples (MEL & CHX12) had higher number of significant peaks and thus genes associated with them (2974 divided among 1548 genes & 3080 peaks divided among 1171 respectively, P-value < 0.05). A higher number of peaks could direct to a higher genome wide occupancy of MYC in Mouse samples. (For complete tables see appendix).

3.3.2.2 Gene ontology & Pathway Enrichment analysis

Similar to the human samples, the output gene lists provided by AnnoMiner were submitted to GOelite for studying the enrichment of GO terms and Pathways in the respective samples. The following results for MEL_MYC sample were obtained:

The highest enriched pathways were: positive regulation of metabolic process (142 genes), regulation of signal transduction (140 genes), and regulation of cell death (95 genes). Most of these pathways were also enriched in the human samples. Immune system process (120 genes) & regulation of immune system process (100 genes) were two pathways, which were specifically enriched in mouse samples (Table 3.6). The reason for enrichment of immune system related processes still remains elusive. Not much is known about the direct relation of MYC in regulating genes involved in the immune system. However, given the diverse role of MYC and its involvement in multiple cancers, it is possible that MYC might be regulating genes, which regulate the immune system and maintain a stable T-Cell homeostasis. Furthermore, few mitochondrial GO terms were present in this particular sample (as compared to human K562) indicating a cell and species-specific binding of MYC to mitochondrial genes. Complete GO term information in appendix.

GO TERM	FREQUENCY OF GENES ASSOCIATED
Positive_regulation_of_metabolic_process	142
Regulation_of_signal_transduction	139
Immune_system_process	123
Regulation_of_cell_communication	109
Regulation_of_localization	106
Regulation_of_cell_death	105
Regulation_of_immune_system_process	100
Anatomical_structure_morphogenesis	93
Cell_projection	85
Cytosol	83
Enzyme_binding	78
Regulation_of_hydrolase_activity	83
Multicellular_organismal_development	67
Positive_regulation_of_signaling	63
Protein_domain_specific_binding	62
Defense_response	61
Protein_dimerization_activity	61
Negative_regulation_of_response_to_stimulus	58

Table 3.6 GO enrichment analysis for MEL-MYC mouse sample. Contrary to human samples, the mouse samples showed a higher number of peaks and thus a higher number of genes associated with those peaks. The highest enriched terms were: "regulation of signal transduction (139 genes), positive regulation of metabolic processes (142 genes), immune system process (100 genes), regulation of cell communication (106 genes) and regulation of cell death (95 genes). A total of 4012 genes were associated with a total of 256 GO processes. The gene with highest ontology associations was 'tlr4' (45 GO terms). For complete table see appendix

GO TERM	FREQUENCY OF GENES ASSOCIATED
Catabolic_process	90
Enzyme_binding	73
Identical_protein_binding	71
Chromatin_organization	52
Macromolecular_complex_subunit_organization	52
Endoplasmic_reticulum_part	48
Cell_cycle	42
DNA_metabolic_process	39
Soluble_fraction	35
Nucleoplasm	31
Ligase_activity,_forming_carbon-nitrogen_bonds	26
Generation_of_precursor_metabolites_and_energy	24
ncRNA_metabolic_process	21
Response_to_oxidative_stress	21
Lysosome	20
Proteolysis_involved_in_cellular_protein_catabolic_proce Ts	20
Transferase_activity	18
Endosomal_part	17

Tables 3.7 GOelite results for CHX12-MYC sample. List of gene associated with Go terms was provided to GOelite. Catabolic processes (90 genes), identical protein binding (71 genes), chromatin organization (52 genes), cell cycle (42 genes), and enzyme binding (73 genes) were the highest enriched terms. Ontology terms associated with apoptosis and mitochondria were also enriched however, with fewer gene linkage. There were a total of 1938 genes associated with

191 GO terms. The gene with maximum number of ontology association sin the sample was 'AKT1' (15 associations). For complete table see appendix.

3.3.2.3 KEGG associations

The same list of genes was used to search for enriched pathways with GOelite. It was found that most of the enriched pathways were associated with intracellular signaling (12 in total) (Table 3.8). MYC seems to be regulating multiple targets that are involved in major signaling pathways for the cell. Out of these JAK-STAT, MAPK and T,B-cell signaling were the most enriched. Due to the enrichment of multiple signaling pathways, Nfkb1 & Pik3ca were the genes with highest KEGG associations (21 & 19 resp.). Unlike the human samples, fewer mitochondrial pathways were enriched.

KEGG TERM	FREQUENCY OF GENES ASSOCIATED
Cytokine-cytokine_receptor_interaction	34
Pathways_in_cancer	29
Tuberculosis	27
MAPK_signaling_pathway	25
Cell_adhesion_molecules_(CAMs)	23
Toxoplasmosis	22
Osteoclast_differentiation	21
T_cell_receptor_signaling_pathway	21
B_cell_receptor_signaling_pathway	20
Jak-STAT_signaling_pathway	20
Phagosome	20
Leishmaniasis	19
Chagas_disease_(American_trypanosomiasis)	18
Chemokine_signaling_pathway	18
Measles	17
Neurotrophin_signaling_pathway	17
Natural_killer_cell_mediated_cytotoxicity	16
Toll-like_receptor_signaling_pathway	16
Autoimmune_thyroid_disease	15

Table 3.8 KEGG associations for MEL_MYC sample. The highest enriched pathways were: cytokine-cytokine receptor interaction (43 genes) , pathways in cancer (29 genes), MAPK signaling (25 genes), tuberculosis (27 genes), other signaling pathways (JAK-STAT, neurotrophin, chemokine, B-cell, and apoptosis). See complete table in appendix.

KEGG TERM	FREQUENCY OF GENES ASSOCIATED
Insulin_signaling_pathway	18
Hepatitis_C	13
Parkinson's_disease	12
Adipocytokine_signaling_pathway	9
Bile_secretion	9
ErbB_signaling_pathway	9
Renal_cell_carcinoma	9
Proteasome	7
SNARE_interactions_in_vesicular_transport	7
Fructose_and_mannose_metabolism	6
Mineral_absorption	5
Pentose_phosphate_pathway	5
Porphyrin_and_chlorophyll_metabolism	5
Galactose_metabolism	4
Fatty_acid_elongation_in_mitochondria	3

Table 3.9 KEGG associations for CHX12 sample. The highest enriched pathways were: insulin signaling pathways (18 genes) and Parkinson's disease (12 genes), Hepatitis C (13 genes).

CHX12_MYC sample had completely different enriched pathways. Highest being insulin signaling pathway (17 genes) followed by Parkinson's disease. This sample also showed fewer pathways than MEL sample although the input number of genes for both was almost similar (1080 vs. 1370 genes resp.) one of the possible explanations could be that fact that the MYC binds to different targets across the two cell types. Leading to a highly variable degree of tissue-specific regulation.

3.3.2.4 Peak data visualization

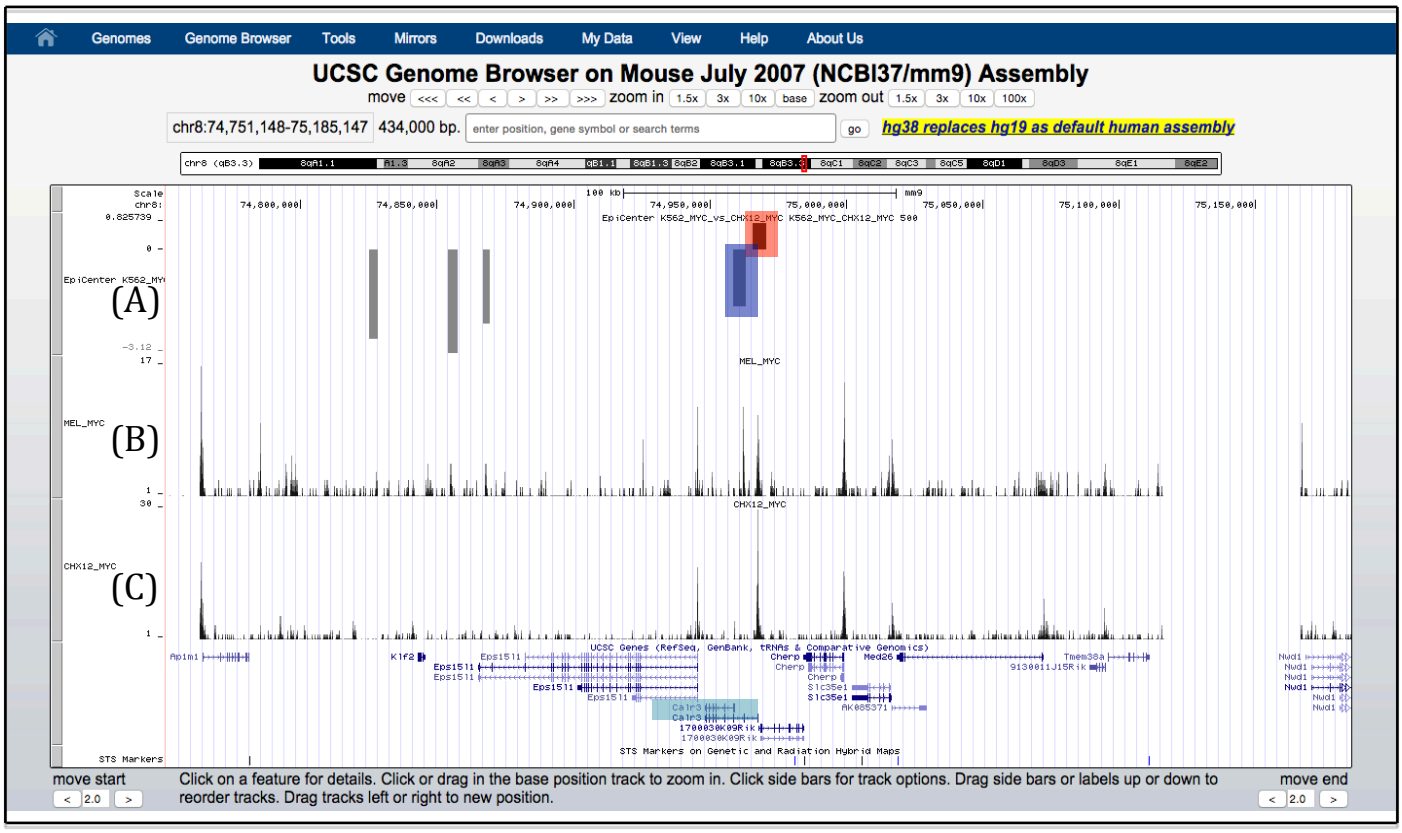

Figure 3.8 Manual benchmark session UCSC genome browser (MEL_MYC_vs_CHX12_MYC). Manual peak visualization session of EpiCenter output (A) for comparison of MYC binding across MEL (B) And CHX12 (C) for mouse samples distributed over gene sites (CALR3 in this case). For reference, the raw bed files (pre normalization) were also added to the session to compare the raw read counts for a particular reads. The Black rectangular bar on top (highlighted red) represents the peak value in CHX12 (0.84) and the grey bar (highlighted purple) represents the peak value in MEL (1.728) sample respectively. The peak can be seen as the average of read counts (intensities) for each sample minus the noise. Epicenter does automatic normalization for read depth hence the final comparisons (A) are without any sequencing biases

3.3.1.3 Common differential binding across Human and Mouse

Many transcription factors are believed to have conserved binding sites across promoter regions, which have been conserved during evolution. In order to test this hypothesis, EpiCenter results were combined across human and mouse samples to look for

conserved binding sites. *Bedtools* package's *Intersect* command was used for this purpose. A total of 4,516,557 peaks had overlap for the K562 sample across human and mouse and a total of 1,535,480 peaks had overlap for GM12878 sample. The average length of the overlap was 9 bp. The list was fed into *PAVIS* *(http://manticore.niehs.nih.gov/pavis2/)* for peak annotation (AnnoMiner was not able to handle such large files). 764587 of 1535480 (49.79%) of the loci were associated with genes (Fig. 3.9). Note: Upstream length was set to 5000 and Downstream length was set to 1000.

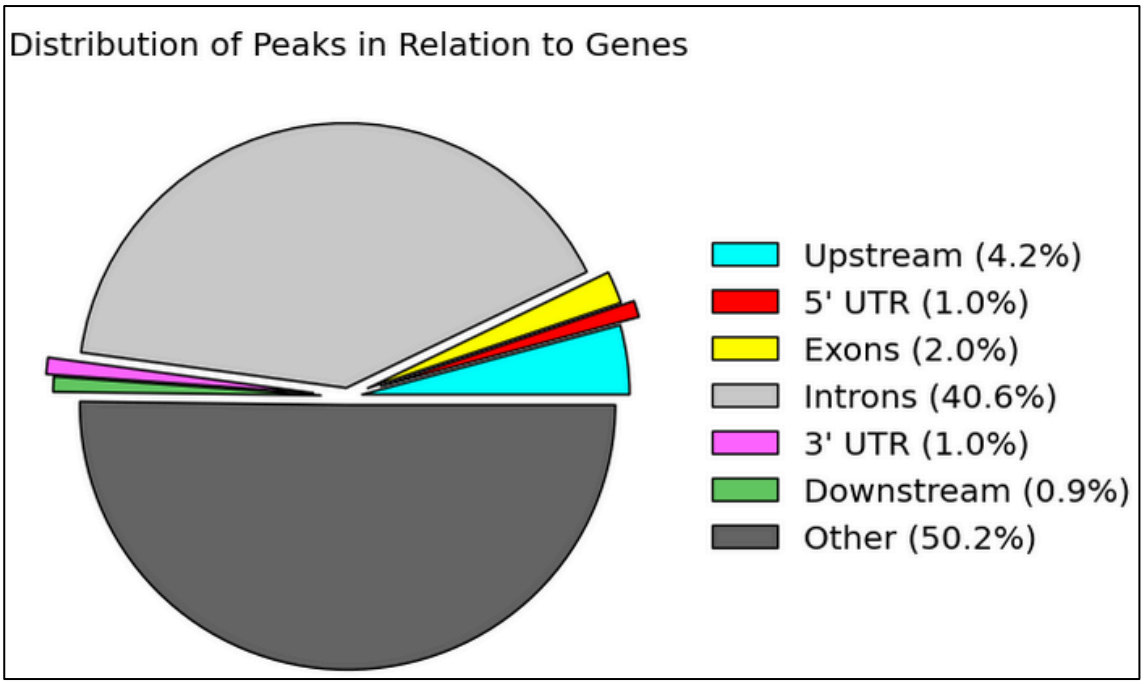

Figure 3.9 Distribution of common peaks between human and mouse samples (GM12878).

Most of the peaks were present in the intronic region for GM12878_MYC sample (40.6%) and fewer were associated with exonic regions (2.0%). This could be due to the longer intronic size (PAVIS does not normalize for the region length). However, the most interesting regions were the upstream region (4.2%) as that is the location where most of the TFs binding occur. This does point to the fact there are conserved binding regions at least for MYC. A Further in depth analysis was required to decipher which genes does these promoters belong to and their functional relevance. For this purpose, the gene list

(3517 genes) from *PAVIS* was fed into GOelite to study the processes that were enriched across these genes.

Highest 30 enriched GO terms are listed in Table 3.11. The highest being membrane bound organelles (1563 genes), cellular metabolic process (1215 genes) and its regulation (754 genes), Primary metabolic process (1201 genes) & its regulation (722 genes), and catalytic activity (984 genes) (Table 3.10). These processes seemed to be the most conserved for MYC binding across human and mouse.

GO TERM	FREQUENCY OF GENES ASSOCIATED
membrane-bounded_organelle	1563
cellular_metabolic_process	1215
primary_metabolic_process	1201
cytoplasmic_part	1179
intracellular_organelle_part	1179
catalytic_activity	984
cytoplasm	925
regulation_of_cellular_metabolic_process	779
regulation_of_primary_metabolic_process	754
regulation_of_macromolecule_metabolic_process	722
non-membrane-bounded_organelle	588
nucleotide_binding	436
small_molecule_metabolic_process	415
organelle_membrane	386
intracellular	370
organelle_organization	321
zinc_ion_binding	314
regulation_of_catalytic_activity	279
intracellular_transport	216
enzyme_binding	206
RNA_binding	185
cell_cycle_process	174
cell_cycle	164
regulation_of_cell_cycle	164
macromolecular_complex_assembly	159

gene_expression	153
ribonucleoprotein_complex	139
cellular_macromolecular_complex_subunit_organization	128
enzyme_linked_receptor_protein_signaling_pathway	118

Table 3.11 GO enrichment for peaks belonging to common genomic regions (human and mosue).

Using the Gene list provided by Pavis, the genes with highest number of peaks near the TSS (+- 500 bp) were selected. These genes are listed in table 3.11. The highest conservation in the binding regions seems to be in genes presented here. This alluded to the fact that there is indeed conserved binding by MYC across human and mouse genomic regions.

Gene Name	Frequency of common peaks
DQ582201	643
JA429504	462
DQ582265	459
BC018860	451
TVAS5	445
JA040725	287
AF079515	273
JA429830	247
UBA2	142
PDIA4	141
CEP135	133
RWDD3	130
SNHG5	113
TMEM97	107

Table 3.12 Common genes across Human and Mouse samples that had MYC occupancy. A total of 14 genes showed similar occupancy of MYC across human and mouse samples. The values represent peak values generated using Intersect option from Bedtools and which were near TSS (+-500 bp). For complete table refer to appendix.

4 Conclusions and Discussions

Transcription factor and histone modifications are two key factors that mediate gene regulation. TF-binding data and histone modification data capture the gene expression to a high extent (Cheng et al. 2012) (Fig. 4.1). Chromatin Immunoprecipitation (ChIP) followed by Sequencing (Seq) has become the primary method to identify these large-scale transcription factor binding, histone marks and modifications, and mechanisms of differential gene regulation (Furey 2012). We here show that using ChIP-Seq data, comparative studies for transcription factor binding can be performed across multiple samples. Using Gene ontology (combined with pathway) enrichment, we could confer the processes that MYC targets specific to particular tissue types. We show that this type of in-depth analysis for transcription factor binding can be instrumental in understanding the extent of regulation for specific transcription factors. Comparing binding across two species (human and mouse in our case) can also help elucidate the conserved binding regions and hence conserved processes targeted by transcription factors.

A variety of ChIP-Seq analysis packages are available that perform peak detection for multiple samples. However they fail to acknowledge the importance of replicates files, background normalizations and the biases caused by sequencing depth. Since accurate identification of real differential binding sites is likely to rely more on biological replicates and sequencing depth than number of absolute reads, a software tool for taking in to account these factors is crucial in study of ChIP-Seq data. Therefore this study employed the software EpiCenter (Huang et al. 2011) to compare transcription factor binding across differential cell types and between species (human and mouse). Unlike the existing methods, EpiCenter uses a combination of two statistical tests (exact ratio test, and z test on log2ratio of read counts) for determining differential regions between samples and thus controls for a lower FDR.

Using EpiCenter for two samples (K562 & GM12878) in humans and their counterparts in mouse (MEL & CHX12) to study the genome-wide binding of C-MYC, It was found that C-MYC seems to have tissue specific binding to an extent. The number of binding sites also varied across the samples (K562_human: 1737 peaks, GM12878_Human: 231 peaks) & (MEL_Mouse: 21857 peaks, CHX12_Mouse: 4132 peaks). One critical reason for this variation might be due to the fact that these experiments were not from the same group/lab and thus technical variation in sample and library preparation might affect the overall read count. Also errors in the immunoprecipitation step can create high sample variance. After analyzing the gene lists for ontology and pathway enrichment, our study could predict diverse cellular processes that are associated with MYC targets. Moreover the enrichment differences were highly variable across the samples. For example: K562_human had high enrichment in ATP binding genes, other mitochondrial pathways and apoptosis – all of which are indicative of a cell in a state of high energy demand. GM12878_human, on the other hand, had enrichment in processes related to cytokine signaling, cellular proliferation and phosphorylation, which indicated the cell being in a proliferating and signaling mode.

Similarly, pathway information from KEGG provided information about enriched pathways in each sample. The primary finding from this was the fact that for the K562_human sample, the majority of the target genes for MYC were associated with mental diseases (Alzheimer's, Parkinson's, and Huntington's disease). Many of these target genes were mitochondrial (MT-COA1, 2, 3, 4 & ATP enzymes). This information might be relevant for researchers studying mental disorders. C-MYC might be playing an important role in regulating the expression levels of these genes.

Studying differential binding of transcription factors is pivotal in understanding the different cellular phenotypes. Moreover, accurate prediction of these 'differential' sites is necessary in order to produce reliable results. This study provides a framework for testing the differential binding of transcriptional factors across multiple samples to study gene regulation.

6 References

1. Adhikary S, Eilers M. Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol*. 2005;6(8):635-645. doi:10.1038/nrm1703.
2. Anders S, Pyl PT, Huber W. *HTSeq A Python Framework to Work with High-Throughput Sequencing Data*.; 2014. doi:10.1101/002824.
3. Anders S. Analysing RNA-Seq data with the DESeq package. *R Man*. 2012:1-28. papers2://publication/uuid/8ABCEEC5-4D98-4858-AA40-8D54B601FBFF.
4. Anders S, Huber W. DESeq: Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106. doi:10.1186/gb-2010-11-10-r106.
5. Anders S, Huber W. Differential expression of RNA-Seq data at the gene level – the DESeq package. *Bioconductor Packag Vignette*. 2013.
6. Anders S, McCarthy DJ, Chen Y, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*. 2013;8(9):1765-1786. doi:10.1038/nprot.2013.099.
7. Andrews S. FastQC: A quality control tool for high throughput sequence data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. 2010.
8. Bello-Fernandez C, Packham G, Cleveland JL. The ornithine decarboxylase gene is a transcriptional target of c-Myc. *Proc Natl Acad Sci U S A*. 1993;90(16):7804-7808. doi:10.1073/pnas.90.16.7804.
9. Blackwood EM, Eisenman RN. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science*. 1991;251(4998):1211-1217. doi:10.1126/science.2006410.
10. Bresnick EH, Katsumura KR, Lee HY, Johnson KD, Perkins AS. Master regulatory GATA transcription factors: Mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Res*. 2012;40(13):5819-5831. doi:10.1093/nar/gks281.
11. Bresnick EH, Lee HY, Fujiwara T, Johnson KD, Keles S. GATA switches as developmental drivers. *J Biol Chem*. 2010;285(41):31087-31093. doi:10.1074/jbc.R110.159079.
12. Bresnick EH, Martowicz M, Pal S, Johnson KD. Developmental control via GATA factor interplay at chromatin domains. *J Cell Physiol*. 2005;205(1):1-9. doi:10.1002/jcp.20393.
13. Cantor AB, Orkin SH. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*. 2002;21(21):3368-3376. doi:10.1038/sj.onc.1205326.
14. Cole MD, Cowling VH. Transcription-independent functions of MYC: regulation of translation and DNA replication. *Nat Rev Mol Cell Biol*. 2008;9(10):810-815. doi:10.1038/nrm2467.
15. Dang C V. MYC on the path to cancer. *Cell*. 2012;149(1):22-35. doi:10.1016/j.cell.2012.03.003.
16. Dang C V., O'Donnell KA, Zeller KI, Nguyen T, Osthus RC, Li F. The c-Myc target gene network. *Semin Cancer Biol*. 2006;16(4):253-264. doi:10.1016/j.semcancer.2006.07.014.

17. David CJ, Chen M, Assanah M, Canoll P, Manley JL. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature*. 2010;463(7279):364-368. doi:10.1038/nature08697.
18. De S, Lopez-Bigas N, Teichmann SA. Patterns of evolutionary constraints on genes in humans. *BMC Evol Biol*. 2008;8:275. doi:10.1186/1471-2148-8-275.
19. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One*. 2013;8(12). doi:10.1371/journal.pone.0085024.
20. Dobin A, Davis C a, Schlesinger F, et al. RNA-STAR : ultrafast universal spliced sequences aligner : Supplementary materials. *Bioinformatics*. 2012:1-7. doi:10.1093/bioinformatics/bts635.
21. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635.
22. Eberhardy SR, Farnham PJ. c-Myc Mediates Activation of the cad Promoter via a Post-RNA Polymerase II Recruitment Mechanism. *J Biol Chem*. 2001;276(51):48562-48571. doi:10.1074/jbc.M109014200.
23. Eberhardy SR, Farnham PJ. Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter. *J Biol Chem*. 2002;277(42):40156-40162. doi:10.1074/jbc.M207441200.
24. Evans T, Reitman M, Felsenfeld G. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc Natl Acad Sci U S A*. 1988;85(16):5976-5980. doi:10.1073/pnas.85.16.5976.
25. Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011;Chapter 2:Unit 2.14. doi:10.1002/0471250953.bi0214s34.
26. Follows GA, Tagoh H, Lefevre P, Morgan GJ, Bonifer C. Differential transcription factor occupancy but evolutionarily conserved chromatin features at the human and mouse M-CSF (CSF-1) receptor loci. *Nucleic Acids Res*. 2003;31(20):5805-5816. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=219482&tool=pmcentrez&rendertype=abstract. Accessed September 25, 2015.
27. Fujiwara Y, Browne CP, Cunniff K, Goff SC, Orkin SH. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc Natl Acad Sci U S A*. 1996;93(22):12355-12358. doi:10.1073/pnas.93.22.12355.
28. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*. 2012;13(12):840-852. doi:10.1038/nrg3306.
29. Furney SJ, Higgins DG, Ouzounis CA, López-Bigas N. Structural and functional properties of genes involved in human cancer. *BMC Genomics*. 2006;7:3. doi:10.1186/1471-2164-7-3.
30. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(DATABASE ISS.). doi:10.1093/nar/gki033.
31. Hitzler JK, Zipursky A. Origins of leukaemia in children with Down syndrome. *Nat Rev Cancer*. 2005;5(1):11-20. doi:10.1038/nrc1525.
32. Ho JWK, Bishop E, Karchenko P V, Nègre N, White KP, Park PJ. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*. 2011;12(1):134. doi:10.1186/1471-2164-12-134.

33. Huang W, Umbach DM, Vincent Jordan N, Abell AN, Johnson GL, Li L. Efficiently identifying genome-wide changes with next-generation sequencing data. *Nucleic Acids Res*. 2011;39(19). doi:10.1093/nar/gkr592.
34. Ji H, Li X, Wang Q, Ning Y. Differential principal component analysis of ChIP-seq. *Proc Natl Acad Sci U S A*. 2013;110(17):6789-6794. doi:10.1073/pnas.1204398110.
35. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007;316(5830):1497-1502. doi:10.1126/science.1141319.
36. Kharchenko P V, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*. 2008;26(12):1351-1359. doi:10.1038/nbt.1508.
37. Ko LJ, Engel JD. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol*. 1993;13(7):4011-4022. doi:10.1128/MCB.13.7.4011.Updated.
38. Lee LA, Dang C V. Myc target transcriptomes. *Curr Top Microbiol Immunol*. 2006;302:145-167.
39. Li F, Wang Y, Zeller KI, et al. Myc stimulates nuclearly encoded mitochondrial genes and mitochondrial biogenesis. *Mol Cell Biol*. 2005;25(14):6225-6234. doi:10.1128/MCB.25.14.6225-6234.2005.
40. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656.
41. Lin CY, Lovén J, Rahl PB, et al. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*. 2012;151(1):56-67. doi:10.1016/j.cell.2012.08.026.
42. Lopez-Bigas N, De S, Teichmann SA. Functional protein divergence in the evolution of Homo sapiens. *Genome Biol*. 2008;9(2):R33. doi:10.1186/gb-2008-9-2-r33.
43. Lowry JA, Atchley WR. Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. *J Mol Evol*. 2000;50(2):103-115. doi:10.1007/s002399910012.
44. Lüscher B, Larsson LG. The basic region/helix-loop-helix/leucine zipper domain of Myc proto-oncoproteins: function and regulation. *Oncogene*. 1999;18(19):2955-2966. doi:10.1038/sj.onc.1202750.
45. Marcu KB, Bossone SA, Patel AJ. myc function and regulation. *Annu Rev Biochem*. 1992;61:809-860. doi:10.1146/annurev.bi.61.070192.004113.
46. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011;17(1):10. doi:10.14806/ej.17.1.200.
47. Meyer N, Penn LZ. Reflecting on 25 years with MYC. *Nat Rev Cancer*. 2008;8(12):976-990. doi:10.1038/nrc2231.
48. Meyer N, Penn LZ. Reflecting on 25 years with MYC. *Nat Rev Cancer*. 2008;8(12):976-990. doi:10.1038/nrc2231.
49. Muntean AG, Crispino JD. Differential requirements for the activation domain and FOG-interaction surface of GATA-1 in megakaryocyte gene expression and development. *Blood*. 2005;106(4):1223-1231. doi:10.1182/blood-2005-02-0551.
50. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10(10):669-680. doi:10.1038/nrg2641.
51. Patient RK, McGhee JD. The GATA family (vertebrates and invertebrates). *Curr Opin Genet Dev*. 2002;12(4):416-422. doi:10.1016/S0959-437X(02)00319-2.

52. Pevny L, Lin CS, D'Agati V, Simon MC, Orkin SH, Costantini F. Development of hematopoietic cells lacking transcription factor GATA-1. *Development*. 1995;121(1):163-172.
53. Porro A, Iraci N, Soverini S, et al. c-MYC Oncoprotein Dictates Transcriptional Profiles of ATP-Binding Cassette Transporter Genes in Chronic Myelogenous Leukemia CD34+ Hematopoietic Progenitor Cells. *Mol Cancer Res*. 2011;9(8):1054-1066. doi:10.1158/1541-7786.MCR-10-0510.
54. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033.
55. Ross-Innes CS, Stark R, Teschendorff AE, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. 2012. doi:10.1038/nature10730.
56. Rylski M, Welch JJ, Chen Y-Y, et al. GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol*. 2003;23(14):5031-5042. doi:10.1128/MCB.23.14.5031-5042.2003.
57. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC Genomics*. 2013;14(1):826. doi:10.1186/1471-2164-14-826.
58. Shen L, Shao N-Y, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One*. 2013;8(6):e65598. doi:10.1371/journal.pone.0065598.
59. Shimizu R, Engel JD, Yamamoto M. GATA1-related leukaemias. *Nat Rev Cancer*. 2008;8(4):279-287. doi:10.1038/nrc2348.
60. Shimizu R, Yamamoto M. Gene expression regulation and domain function of hematopoietic GATA factors. *Semin Cell Dev Biol*. 2005;16(1):129-136. doi:10.1016/j.semcdb.2004.11.001.
61. Shirihai OS, Gregory T, Yu C, Orkin SH, Weiss MJ. ABC-me: a novel mitochondrial transporter induced by GATA-1 during erythroid differentiation. *EMBO J*. 2000;19(11):2492-2502. doi:10.1093/emboj/19.11.2492.
62. Silva M, Grillot D, Benito A, Richard C, Nuñez G, Fernández-Luna JL. Erythropoietin can promote erythroid progenitor survival by repressing apoptosis through Bcl-XL and Bcl-2. *Blood*. 1996;88(5):1576-1582.
63. Surget S, Khoury MP, Bourdon JC. Uncovering the role of p53 splice variants in human malignancy: A clinical perspective. *Onco Targets Ther*. 2013;7:57-67. doi:10.2147/OTT.S53876.
64. Tansey WP. Mammalian MYC Proteins and Cancer. *New J Sci*. 2014;2014:1-27. doi:10.1155/2014/757534.
65. Trapnell C, Pachter L, Salzberg SL. TopHat Manual. *Bioinformatics*. 2009;25:1105-1111. doi:10.1093/bioinformatics/btp120.
66. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120.
67. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120.
68. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562-578. doi:10.1038/nprot.2012.016.
69. Uribesalgo I, Buschbeck M, Gutiérrez A, et al. E-box-independent regulation of transcription and differentiation by MYC. *Nat Cell Biol*. 2011;13(12):1443-1449. doi:10.1038/ncb2355.

70. Valouev A, Johnson DS, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*. 2008;5(9):829-834. doi:10.1038/nmeth.1246.
71. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484.
72. Weiss MJ, Yu C, Orkin SH. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol*. 1997;17(3):1642-1651. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=231889&tool=pmcentrez&rendertype=abstract. Accessed September 25, 2015.
73. Weiss MJ, Keller G, Orkin SH. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1- embryonic stem cells. *Genes Dev*. 1994;8(10):1184-1197. doi:10.1101/gad.8.10.1184.
74. Yu Y, Niapour M, Zhang Y, Berger SA. Mitochondrial regulation by c-Myc and hypoxia-inducible factor-1 alpha controls sensitivity to econazole. *Mol Cancer Ther*. 2008;7(3):483-491. doi:10.1158/1535-7163.MCT-07-2050.
75. Zambon AC, Gaj S, Ho I, et al. GO-Elite: A flexible solution for pathway and ontology over-representation. *Bioinformatics*. 2012;28(16):2209-2210. doi:10.1093/bioinformatics/bts366.
76. Cheng C, Alexander R, Min R, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. 2012;22(9):1658-1667. doi:10.1101/gr.136838.111.
77. 1. Ferreira R, Ohneda K, Yamamoto M, Philipsen S. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol*. 2005;25(4):1215-1227. doi:10.1128/MCB.25.4.1215-1227.2005.
78. 2. Perna D, Fagà G, Verrecchia A, et al. Genome-wide mapping of Myc binding and gene regulation in serum-stimulated fibroblasts. *Oncogene*. 2012;31(13):1695-1709. doi:10.1038/onc.2011.359.
79. 3. Patel JH, Loboda AP, Showe MK, Showe LC, McMahon SB. Opinion: Analysis of genomic targets reveals complex functions of MYC. *Nat Rev Cancer*. 2004;4(7):562-568. doi:10.1038/nrc1393.
80. 4. Cheng C, Alexander R, Min R, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. 2012;22(9):1658-1667. doi:10.1101/gr.136838.111.
81. 5. Maher B. ENCODE: The human encyclopaedia. *Nature*. 2012;489(7414):46-48. doi:10.1038/489046a.

Appendix

Supplementary data and result files can be found at:
https://drive.google.com/drive/folders/0B_MVmsAk2E6MUXA3WWVCdW1SeEE

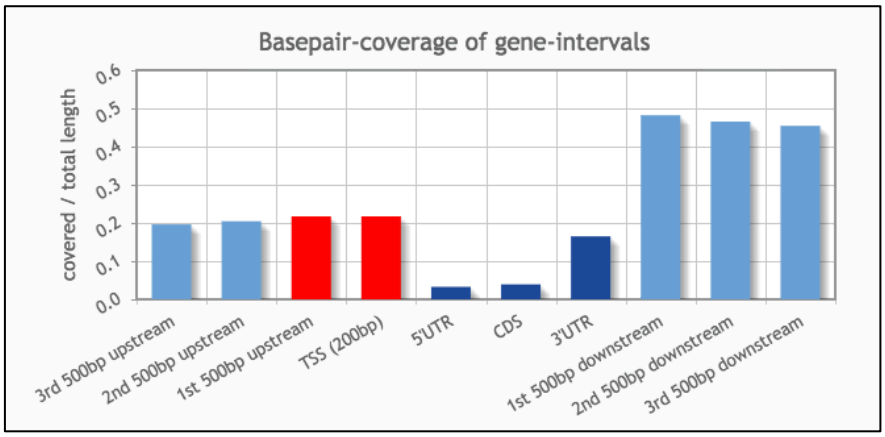

Figure : Annominer Coverage for G1E_GATA1 samples. The sample was not of good quality. Most of the peaks belonged to downstream regions instead of near TSS.

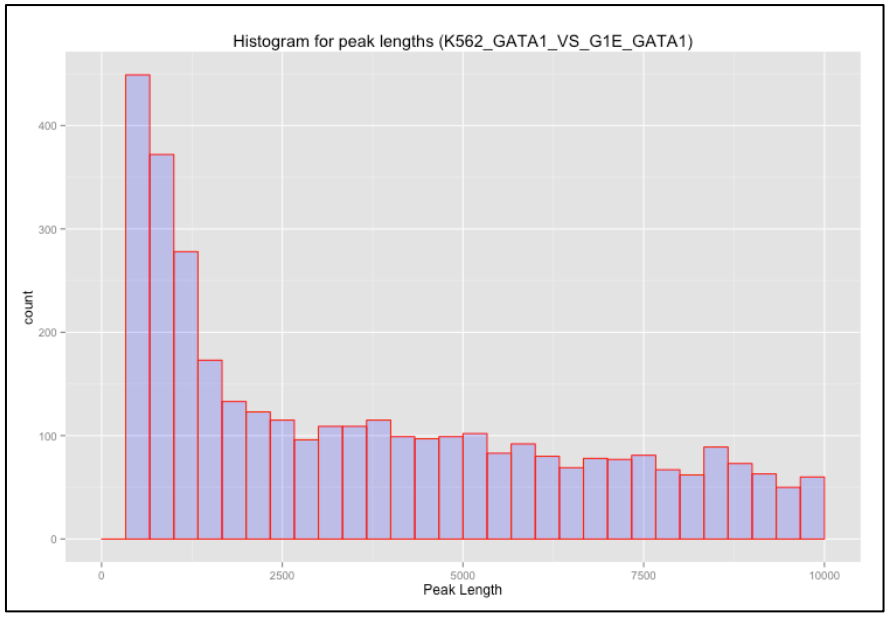

Figure: Peak length distribution for K562_GATA1_VS_G1E_GATA1. The peaks were larger than expected (TF binding is generally 150 bp (+- 500 bp). Improper shearing of DNA during sample preparation could cause this or nonspecific antibody binding.

Acknowledgements

I would like to thank everyone who provided his or her support to me throughout my thesis.

Firstly, I would like to thank Dr. Bianca Habermann for giving me this opportunity to work in her group. She was an inspirational supervisor with whom I always had insightful discussions about the project. She helped me in my times of difficulty by providing guidance and encouraging words.

I would also like to take this opportunity to thank the whole Habermann group for my wonderful time there. I thank everyone in the group for making this a great learning experience for me. Special thanks goes to Assa and Michael, who helped me cross-countless hurdles in the project.
I would also like to thank Prasanna for unlimited interesting discussions we had over our lunch.

A special thanks also goes to my dearest friends Kanishk, Abhijeet, Sonal, Victor and to other countless people who inspired me some or the other way.

Finally, this thesis would not have been possible without the support of family back home and rest of friends all around the world.

Statement of originality

Erklärung zur Masterarbeit

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden. Diese Erklärung erstreckt sich auch auf in der Arbeit enthaltene Graphiken, Zeichnungen, Kartenskizzen und bildliche Darstellungen.

Master's thesis statement of originality

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgements. This applies also to all graphics, drawings, maps and images included in the thesis.

Munich, October 6, 2015 .....
Ort und Datum/ Place and date Unterschrift/Signature

This form must be filled out and signed by your supervisor. It is to be included in the final version of the Masters Thesis as the last page in the copy for the Examination Board.

Name (last, first): **Talwar, Jatin**

Title of Thesis: **Application of ChIP-Seq softwares in study of gene regulation**.

Group in which the work was performed: Research Group of Dr. Bianca Habermann,

Group leader, bioinformatics.
Max Planck Institute of Biochemistry (MPI-Biochemie), Martinsried, Germany

Examiner/Supervisor: **Prof. Dr. Wolfgang Enard**

Signature (supervisor): _____

Starting date: 14.04.2015

Submission date: 06.10.2015

_____

Signature, Chair of the Examination Board or authorized representative