

PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

**An integrative approach to identify binding partners of Myc
using (epi)genomics data in the 3T9^{MycER}, Eμ-*myc* and tet-MYC
systems**

Pranami Bora

Center for Genomic Science of

IIT@SEMM, Milan

Matricola n. R10337

Supervisor:

Dr. Bruno Amati

Center for Genomic Science of IIT@SEMM, Milan

Added Supervisor:

Dr. Marco Morelli

Center for Genomic Science of IIT@SEMM, Milan

Anno accademico 2016-2017

Table of Contents

List of abbreviations	4
Figures index	5
Abstract	8
Chapter 1	10
Introduction	10
1.1 Epigenetics	10
1.1.1 Acetylation and methylation	10
1.1.2 Transcription factors.....	12
1.1.3 Myc dependent gene regulation	17
1.1.4 Next generation sequencing methods.....	18
1.1.5 Public datasets.....	26
1.2 Objectives	26
1.2.1 Identify possible binding partners of Myc in cellular growth and lymphomagenesis	26
1.2.2 Choosing a DNase-seq footprint caller	27
1.2.3 Developing a pipeline and methods for the analysis of low depth and high depth	
DNase-seq data.....	27
Chapter 2	29
Materials and methods	29
2.1 R Functions and methods:	31
Chapter 3	35
Results	35
3.1 Benchmarking study of footprint callers	35

3.1.1	Introduction.....	35
3.1.2	Materials and methods.....	36
3.1.3	Results	37
3.2	DNase-seq analysis pipeline	42
3.2.1	Introduction.....	42
3.2.2	Materials and methods.....	43
3.2.3	Results	49
3.3	Integration of 3T9^{MyceR} and Eμ-<i>myc</i> DNase-seq data with CHIP-seq and RNA-seq	58
3.3.1	Introduction.....	58
3.3.2	Materials and methods.....	59
3.3.3	Results	60
3.4	Single feature classification.....	75
3.4.1	Introduction.....	75
3.4.2	Materials and methods.....	75
3.4.3	Results	78
3.5	Random forest classification	81
3.5.1	Introduction.....	81
3.5.2	Materials and methods.....	84
3.5.3	Results	86
Chapter 4	98
Discussion	98
References	108

List of abbreviations

AUC	Area under the curve
bp	base pair(s)
bHLH-Zip	basic helix-loop-helix leucine zipper
ChIP	Chromatin immunoprecipitation
DHS	DNase I hypersensitive site
FP	footprint
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
NGS	Next generation sequencing
PWM	Position weight matrix
RF	Random forest
ROC	Receiver operating characteristic
TF	Transcription factor
TSS	Transcription start site
VI	Variable importance

Figures index

Figure 1. Venn diagram showing the MYC-induced (left) and MYC repressed (right) genes in tet-MYC/LAP-tTA.	17
Figure 2. The basic steps involved in (a) transcription factor ChIP-seq experiments (b) Histone mark ChIP-seq. The main difference between the two is that the antibody used in the first case is targeted against the transcription factor while in the second it is targeted against the histone modification.....	20
Figure 3. The basic steps in a RNA-seq experiment.....	21
Figure 4. The basic steps in a FAIRE-seq experiment.	22
Figure 5. High depth DNase seq and identification of transcription factor footprints.	24
Figure 6. Example of the structure of one element of a DHS class.....	32
Figure 7. Example showing the nearest footprint to a summit.....	33
Figure 8. Number of footprints called by DNaseR and Wellington at different stringency levels.	38
Figure 9. (A) Receiver-Operator Characteristic (ROC) curves for the predictions provided by the binding motifs alone. (B–D) ROCs for the sets of footprints obtained by DNaseR, Wellington and for the set used in (Neph, et al. 2012).....	40
Figure 10. Heatmap showing the edge-to-edge correlation among the TF-TF networks reconstructed with the sets of footprints obtained with DNaseR, Neph, Wellington in three different cell lines (K562, SkMC, HepG2).	42
Figure 11. DNase-seq pipeline.....	46
Figure 12. Snapshot of the HTML page containing the list of PWMs collected from various sources.....	48
Figure 13. Screenshot of the genome browser displaying the low depth DNase-seq (in blue) and FAIRE-seq (in red) and the input (black) tracks in 3T9 ^{MycER} 4h A sample.....	50
Figure 14. Distribution of DHSs (from high-depth DNase-seq experiments) on promoters, intergenic regions and genebody.....	51
Figure 15. The overlap of low depth and high depth DHSs in 0h and 4h 3T9MycER.....	52
Figure 16. Tracks showing the signals of DNase-seq reads in 4h 3T9 ^{MycER} , corresponding to the DHS, per base signal and the footprints.....	53
Figure 17. Aggregate per base DNase I cleavage patterns for three transcription factors in the 3T9 ^{MycER} cells in untreated (0h) and treated (4h) conditions for (a-b)(L-R) Myc 0h and 4h (c-d) Ctf 0h and 4h (d-e) Sp1 0h and 4h.	55
Figure 18. (a). Heatmap showing the percentage of overlap among the DHSs of 3T9 ^{MycER} (4h/0h) and Eμ-myc (C, P and T). (b) Heatmap showing the percentage of overlap among the footprints of 3T9 ^{MycER} (4h/0h) and Eμ-myc (C, P and T).....	57
Figure 19. Heatmap showing the presence or absence of peaks corresponding to all the promoter regions (-2kb, +1kb from the TSS) in chromosome 1 for histone marks (H3K4me1, H3K4me3 and H3K27ac),	

transcription factor (*Myc* and *Ctcf*), and *Pol II* ChIP-seqs and *DNase-seq* of (a) *Eμ-myc* (C, P and T) and (b) $3T9^{MycER}$ at 0h and 4h..... 62

Figure 20. (a) Number of E-boxes identified by FIMO (cut-off 10^{-5}) on promoters of *Myc* peaks under the summit using varying peak widths (200bp, 400bp) and complete peak in $3T9^{MycER}$ at 0h and 4h time-points ... 63

Figure 21. Distribution of the up, down and no-deg genes into 3 categories based on the footprint data (4h): (1) footprint of the canonical E-box binding proteins, (2) footprint of other proteins and (3) no footprints. 64

Figure 22. Boxplots showing the enrichment of promoter and distal *Myc* peaks divided into three subgroups, those with no footprint (no FP), with a footprint that does not have the E-box sequence (with FP other than E-box) and finally the group of *Myc* peaks that have a footprint that has a E-box (with E-box FP) at 0h and 4h in $3T9^{MycER}$ and in the C, P and T conditions of *Eμ-myc*..... 66

Figure 23. Boxplot showing the enrichment of *Myc* peaks in 4h $3T9^{MycER}$ samples binding to upregulated and downregulated genes..... 68

Figure 24. Run times of parallel MEME, CUDA-MEME and DREME. DREME is the fastest of the three although it can only be used for short motifs..... 69

Figure 25. The list of TFs whose binding sites were identified by DREME in the (A) $3T9^{MycER}$ up-regulated (B) $3T9^{MycER}$ down-regulated (C) *Eμ-myc* T up-regulated (D) *Eμ-myc* T down-regulated..... 71

Figure 26. Enrichment of TF motifs in (left to right) upregulated genes vs. no-deg genes and down-regulated vs no-deg in (top-bottom) $3T9^{MycER}$ 4h, *Eμ-myc* P and *Eμ-myc* T samples. 74

Figure 27. An example of a Random forest. Random forests are a type of ensemble method which consists of many trees to classify the data. 83

Figure 28. Main steps in the machine learning based classification of the up, down and no-deg genes using different features obtained from NGS data..... 85

Figure 29. Top features obtained from the model with the highest AUC from random forests based classifications ranked by their variable importance (a-b) $3T9^{MycER}$ up/no-deg and down/no-deg (c-d) *Eμ-myc* T vs.C up/no-deg and down/no-deg and (e-f) tet-MYC *Myc*-dependent up/no-deg and down/no-deg (g-h) tet-MYC *Myc* independent. 91

Figure 30. Heatmap showing the normalized frequencies (footprinted only) of the top 20 predictive features (PWMs) from the random forest classification in $3T9^{MycER}$ (up/no-deg and down/no-deg), *Eμ-Myc* (up/no-deg and down/no-deg) and tet-MYC (*Myc* dependent up/no-deg and down/no-deg and *Myc* independent up/no-deg and down/no-deg) in the different gene subsets (up, down, *Myc*-dependent down etc.). 92

Figure 31. (a) The overlap of all the CH12 *Myc* peaks with the *Myc* peaks in the *Eμ-myc* C, P and T samples (b) The overlap of only the promoter *Myc* peaks in CH12 with the promoter *Myc* peaks in the *Eμ-myc* C, P and T samples. 95

Figure 32. Heatmap of the percentage of overlap of ChIP-seq peaks of different transcription factors in the CH12 cell line with the Myc bound promoters in Eμ-myc C, P and T samples. 96

Abstract

The c-MYC oncogene encodes the transcription factor Myc, which regulates a large number of biological processes and is overexpressed in a large number of cancers. When overexpressed, Myc binds to almost all open promoters but only regulates specific subsets of genes. We investigated this issue in three systems where Myc is overexpressed: 3T9^{MycER} fibroblasts, E μ -myc B cells and tet-MYC liver cells, through an approach integrating different types of next generation sequencing data, such as DNase-seq footprinting, ChIP-seq and RNA-seq, with motif analysis and machine learning methods (random forest). In particular, the DNase-seq technique can detect genome-wide open chromatin regions (DNase hypersensitive sites or DHSs) and sites where a transcription factor (TF) is bound (footprints). In order to analyse the DNase-seq footprinting data in our systems, we developed a novel pipeline that carries out a step-by-step analysis of the raw DNase-seq data, and outputs DHS and TF footprints. To select the best footprint caller for the pipeline we carried out a benchmarking study comparing two footprint calling algorithms DNaseR and Wellington on ENCODE data. The Wellington algorithm, scored consistently best both in terms of specificity and sensitivity and therefore it was chosen for our pipeline. We overlapped genome wide the footprints identified by the pipeline with matches of a PWM library, obtaining a list of footprinted PWMs. Then, we used this list as a series of features to carry out pairwise classifications of the upregulated, downregulated and not-deregulated subsets of genes in the three systems. A PWM that classifies the data with a large enough Area Under the Curve (AUC) pointed to a TF possibly selectively binding with Myc in a subset of genes only. We first applied a single feature classifier assessing the performance of each of the PWMs one by one, and we found that single PWMs only provided a limited classification of the gene subsets. We then turned to a random forest classifier that considers

combinations of all the features. This strategy provided a good separation of the data sets (AUC>0.7) and identified some candidates, such as Nrf1/Nrf2 (E μ -*myc* T up), Tead factors (E μ -*myc* T and tet-*MYC* up), E2f4 (E μ -*myc* T up) and E2f1 (E μ -*myc* T and tet-*MYC* up), that could potentially act with Myc in regulating specific subsets of genes.

Chapter 1

Introduction

1.1 Epigenetics

The eukaryotic DNA forms complexes with proteins called “chromatin”. These proteins are mostly histone proteins which are of five major types: H1, H2A, H2B, H3 and H4 to form DNA-histone complexes termed the “chromatin”. Additionally, chromatin also contains many different types of nonhistone chromosomal proteins which carry out a number of different activities, including DNA replication and gene expression (Cooper, 2000). Changes to these histones or the DNA, such as covalent modification of the histone (addition of acetyl, methyl or phosphate groups) or methylation of the cytosines are collectively known as ‘epigenetic marks.’ These modifications control the structure of the chromatin and its accessibility by interacting with various binding proteins, transcription factors, and chromatin remodelling complexes that affect and regulate transcription (Jaenisch and Bird, 2003). The field of Epigenetics involves the study of these changes in the structure of the chromatin or the DNA that do not affect the DNA sequence itself.

1.1.1 Acetylation and methylation

Among the different histone modifications known, the two most widely studied are acetylation and methylation. Histone acetylation of the lysine residues at the N terminus of

histone proteins is commonly known to mark open and active chromatin regions (Taylor et al., 2013; Verdone et al., 2005). The histone acetyltransferase (HATs) and deacetylase enzymes (HDACs) are responsible for this addition of the acetyl group to the lysine residues. Histone lysine methylations on the other hand, are found on histones H3 and H4 which can be added by methyltransferases (HMTs) and removed by demethylases. These modifications can have different effects depending on the position of the methylated amino acid and the number of methyl groups added. For example, histone 3 lysine 4 methylation (H3K4me) and histone 3 lysine 4 trimethylation (H3K4me₃) are usually associated with gene activation, whereas histone 3 lysine 9 di- and tri-methylation (H3K9me₂ and H3K9me₃) and histones 3 lysine 27 tri-methylation (H3K27me₃) have been associated with gene inactivation (Peters et al., 2002; Vakoc et al., 2005; Zhang et al., 2012). Therefore, studying the distribution of these modifications can be useful in determining transcriptional activity of these regions.

Acetylation and methylation marks can also be used to identify regulatory regions such as promoters and enhancers. Promoters are cis-regulatory elements that define when the transcription of a gene takes place and are found directly upstream to the transcription start site (TSS). Enhancers too are cis-regulatory elements but are found at varying distances from the TSS of the gene they regulate, either up- or down-stream. When enhancers are bound by transcription factors they can enhance the activation of associated gene (Shlyueva et al., 2014). Promoters are usually marked by H3K4me₃ while enhancers are mainly marked by H3K4me₁, and both are also marked by H3K27ac when activated (Bonn et al., 2012; Dunham et al., 2012; Rada-Iglesias et al., 2011). H3K9me₃ is usually found on transcriptionally silent heterochromatin regions (Peters et al., 2002) whereas H3K27me₃ is found on euchromatin regions (Simon and Kingston, 2009). H3K27me₃ can mark both promoters and enhancers and does not co-occur with H3K27ac on the same histone. Many

studies have used histone mark profiles to predict enhancers positions (Arnold et al., 2013; Bonn et al., 2012; Ernst et al., 2011; Kharchenko et al., 2011; Rada-Iglesias et al., 2011; Shen et al., 2012) and these predictions have been shown to agree well with enhancer activity assays (Arnold et al., 2013; Bonn et al., 2012; Heintzman et al., 2007).

Given the important role that these histone marks play in the regulation and expression of genes it is not surprising that aberrant patterns of these marks have been reported in many different cancers (Dawson and Kouzarides, 2012; Halkidou et al., 2004; LeRoy et al., 2013; Müller et al., 2013; Song et al., 2005). Many HATs such as, CBP, p300 and MOZ and MORF can form chimeric fusion proteins that arise from chromosomal translocations often associated with leukaemia (Yang, 2004).

HDAC4 is frequently downregulated in gastric tumors, HDAC1 somatic mutations have been detected in some dedifferentiated human liposarcomas (Taylor et al., 2011), and a frame-shift mutation leading to a dysfunctional HDAC2 expression has been observed in human epithelial cancers (Ropero et al., 2006). Aberrant targeting of HDACs is therefore, thought to be involved in the silencing of tumour-suppressor genes (West and Johnstone, 2014).

The dysregulation of HMTs too can contribute to the pathogenesis of different types of cancers by causing aberrant histone methylations (Michalak and Visvader, 2016). For example, deregulation of the HMT, enhancer of zeste homologue 2 (EZH2) leads to aberrant patterns of the H3K27me3 which has been linked to poor patient outcome (Oh et al., 2014).

1.1.2 Transcription factors

Epigenetics also involves the study of the transcription factor (TF) binding and nucleosome positioning on the DNA and their effect on gene-regulation. Many transcription factors bind to their target genes and recruit co-regulators by recognizing a specific short sequence called the ‘binding motif’ (Jaenisch and Bird, 2003) to regulate the expression of the genes. Transcription factors can bind to both promoter and enhancer regions and interact with other bound transcription factors and recruit RNA polymerase II. They can also act as pioneer factors to open the chromatin, thus making the DNA accessible (open chromatin) to other proteins and transcription factors (Soufi et al., 2015).

1.1.2.1 Role of the transcription factor Myc in cancer

A cell is governed by regulatory pathways composed by a large number of proteins and genes interacting in a complex, combinatorial manner. Large-scale analyses of these regulatory networks have statistically characterized their structure and led to the identification of ‘master regulators’ and conserved functional modules (recurring regulation patterns) (Alon, 2007). One of these master regulators is the *c-myc* oncogene, which plays a key role in the development of many types of cancers (Ciriello et al., 2013) where a number of regulatory pathways are disrupted. The *c-myc* gene encodes the transcription factor Myc, that is overexpressed in many cancers (Gabay et al., 2014; Land et al., 1983). The Myc transcription factor regulates the normal proliferation, development and apoptosis, functions that become aberrantly regulated when Myc is overexpressed (Meyer and Penn, 2008). For years, it was known that Myc acts as a general amplifier of transcription by targeting all active promoters and enhancers. However, recent studies have shown that (Sabò et al., 2014; Walz et al., 2014) Myc specifically activates and represses transcription of distinct gene sets, leading to changes in cellular state that can in turn lead to a global increase in gene expression.

The deregulation of MYC usually occurs through 3 main mechanisms: insertional mutagenesis, chromosomal translocations and gene amplification. These mechanisms are described below:

Insertional mutagenesis

The expression of MYC can be activated through retroviral promoter insertion. This is a frequently observed oncogenic mutation in retrovirally induced tumours, and mainly induces haematopoietic tumours like erythroleukaemias and T-cell lymphomas (Meyer and Penn, 2008; Payne et al., 1982).

Chromosomal translocation

In Burkitt's lymphomas, the MYC gene is translocated to the immunoglobulin Ig heavy chain locus leading to its overexpression. This translocation event was modelled in the E μ -*myc* mouse model that develop B-cell lymphoma by Adams *et al.* (Adams et al., 1985).

Amplifications

The MYC gene has been found to be amplified in many cancers such as colon cancer and leukaemia where the cancer cells can contain multiple copies of MYC (Alitalo et al., 1983; Collins and Groudine, 1982; Dalla-Favera et al., 1982). In contrast to chromosomal translocations in haematopoietic cancers, activation of the MYC genes by amplification is commonly detected in solid human tumours (Meyer and Penn, 2008).

In addition to the mechanisms described above, de-regulation of Myc can also occur through mutations of upstream regulators. For example, adenomatous polyposis coli (APC) which forms a part of the Wnt/ β -catenin pathway is mutated at a very high frequency in sporadic

colorectal cancer. The silencing of APC leads to the overexpression of Myc which leads to uncontrolled cell division (Sioson et al., 2014; Renoll and Yochum, 2014).

Among the different systems have been developed to study the mechanisms by which the deregulation of MYC leads to tumour development, in this thesis, we consider the systems described below:

i. 3T9^{MycER} model

The 3T9^{MycER} are a mouse fibroblasts cell line containing the mycER transgene which encodes for a chimeric protein made of the human MYC and the hormone-binding domain of the estrogen receptor. This protein is constantly expressed in the 3T9^{MycER} cells, but only becomes active upon addition of the synthetic estrogen OHT (4-hydroxy tamoxifen) to the medium. Upon OHT binding, the fusion proteins translocate into the nucleus, where Myc binds to DNA and changes the expression of its target genes (Eilers et al., 1989; Littlewood et al., 1995). The activation of the mycER takes place very rapidly and a high amount of Myc can be detected bound to the DNA by Chromatin Immunoprecipitation (ChIP) after only 4 hours of treatment with OHT (Sabò et al., 2014). This system provides a very convenient way to study the direct regulation of target genes by Myc as it is not affected by the complex layers of interactions that occur in tumour cells that make it difficult to study the Myc specific responses.

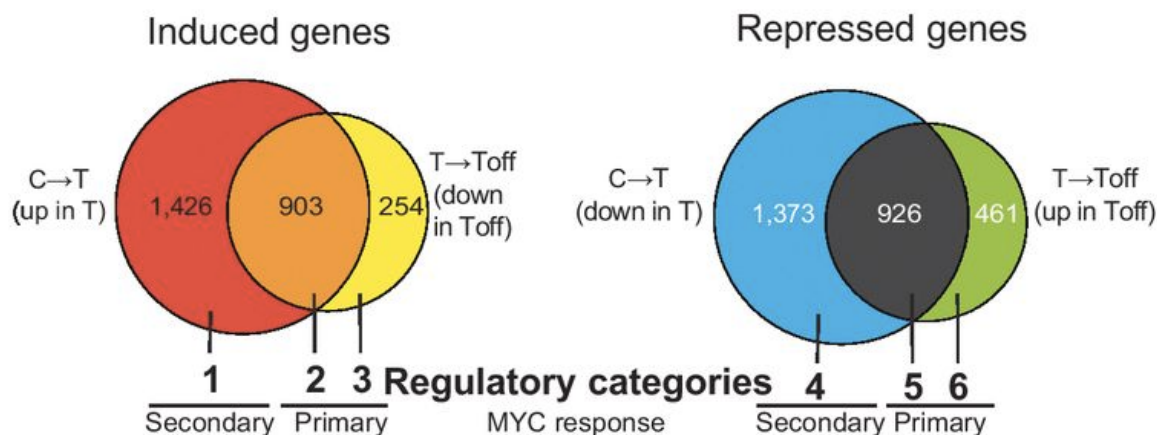
ii. E μ -myc mouse model

In almost all cases of Burkitt's lymphoma, the MYC gene is translocated to the immunoglobulin heavy chain locus. Adams et al. 1985, constructed a transgenic mouse model that reproduces this translocation *in vivo* called the E μ -myc mouse model. The c-MYC

oncogene in these B-cells is placed under the control of the intronic enhancer of the Ig heavy chain, and is therefore constitutively expressed in the B cell lineage. The transgenic mice carrying this system develop B-cell lymphomas with a mean latency of 12-16 weeks of age (Adams et al., 1985). This E μ -*myc* system therefore serves as a model to study the effects of Myc overexpression during B-cell lymphoma progression.

iii. tet-MYC

The tet-MYC/LAP-tTA transgenic mice conditionally express the c-MYC proto-oncogene in liver cells (Kistner et al., 1996). In these mice, the liver activator protein (LAP) promoter drives the expression of the tetracycline-controlled trans-activating protein (tTA) in the liver in absence of doxycycline and is combined with a tet-regulated c-MYC transgene. These Myc-expressing mice develop oncogene-addicted liver tumours but these tumours regress rapidly upon the silencing of MYC due to the administration of doxycycline (Cairo et al., 2008; Kress et al., 2016; Shachaf et al., 2004). Kress et al., 2016, used this reversible system to identify the Myc dependent regulatory events in hepatocellular carcinoma. They were able to identify distinct sets of genes that were activated and repressed by Myc in carcinomas that were no longer deregulated when Myc was turned off (Figure 1). The genes that were deregulated in T->Toff are defined as primary MYC-response or MYC dependent genes. Whereas, the genes that were deregulated in C->T but not deregulated in T-> Toff are defined as secondary MYC-response or MYC independent genes.



Kress et al., 2016

Figure 1. Venn diagram showing the MYC-induced (left) and MYC repressed (right) genes in tet-MYC/LAP-tTA system. were calculated by comparing tumours with control samples (C→T), or tumours after to before tet-MYC inactivation (T→Toff). The genes that are deregulated in T→Toff are defined as primary MYC-dependent DEG categories. Genes deregulated in C→T but not deregulated in →Toff are defined as secondary MYC-response genes.

1.1.3 Myc dependent gene regulation

The Myc transcription factor contains a basic helix-loop-helix leucine zipper (bHLHZip) domain that binds to DNA in correspondence to the E-box element CACGTG (known as the canonical E-box) with high affinity, or to its variants (CANNTG non-canonical E-box) with lower affinity. In order to bind to the DNA, Myc forms heterodimers with another bHLHZip transcription factor, called Max (Blackwood and Eisenman, 1991). Myc is also known to form complexes with the transcription factor Miz1 to down-regulate expression of target genes (Varlakhanova et al., 2011; Walz et al., 2014; Wiese et al., 2013). Myc was also shown to coimmunoprecipitate with NF-Y (Izumi et al., 2001) and the peaks of Myc and NF-Y were found to overlap significantly in ENCODE data (Fleming et al., 2013). Recently, Li et al. 2016, showed that the FOXR2 transcription factor forms a stable complex with Myc and Max to regulate cell proliferation. Myc can therefore bind to different transcription factors in addition to heterdimerization with Max to activate and silence its target genes.

Although Myc is known to preferentially binds E-box containing sequences, many studies have shown that many Myc-bound promoters lack an E-box element (Kim et al., 2008; Li et al., 2003), thus raising the possibility that Myc could bind to the promoters of its target genes in other ways as well, such as (i) indirect binding through to another protein that bind directly to the DNA also called “piggy backing” or through (ii) non-specific binding to the DNA backbone.

1.1.4 Next generation sequencing methods

In the last few years, “next-generation”, high-throughput sequencing technologies have revolutionized epigenetic studies and made it possible to gather data at a genome-wide scale. Nowadays, these techniques allow the mapping of the epigenetic modifications across the entire genome with minimum hands-on time (Ku et al., 2011). Among the different types of next generation sequencing experiments in this thesis we will be using the ones described below:

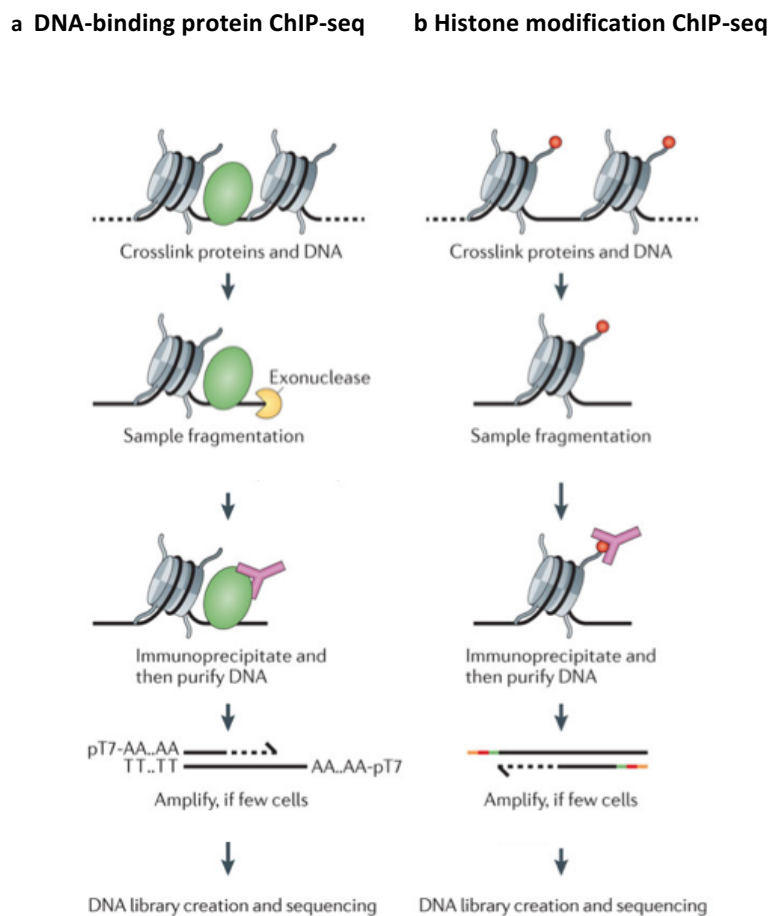
1.1.4.1 ChIP-seq (Chromatin ImmunoPrecipitation-Sequencing)

The ChIP-seq technique is used to identify histone modifications and the interaction of proteins to DNA and is based on the sequencing of the genomic DNA fragments co-immunoprecipitated with a protein of interest or modified histones. ChIP-seq of transcription factors can be used to identify all the binding sites of a transcription factor of interest in the genome. ChIP-seq of histone marks on the other hand, are used to study the genome-wide patterns of epigenomic modifications in the cells. As described before, these patterns can reveal information about the state (active or repressed) and the function of the regulatory regions (enhancer or promoters) that they mark.

The first step of a ChIP-seq experiment depends on the type of the protein under study (Figure 2). In the case of TFs and Pol II the first step is usually to carry out formaldehyde crosslinking of the protein with the DNA so that their interaction does not break during immunoprecipitation. However, in the case of histone modifications this step is not necessary. In the case of chromatin-remodelling enzymes such as HDACs or HATs, an additional cross-linking step (using disuccinimidyl glutarate) can be included, to preserve protein-protein complexes before cross-linking with formaldehyde. After cross-linking, the chromatin is fragmented into pieces of about 150 to 500 bp by sonication.

After fragmentation, the next step is immunoprecipitation, using a specific antibody against the protein of interest. The success of a ChIP-seq experiment depends crucially on strong enrichment of the chromatin specifically bound by the protein under study. Only antibodies that give consistently high enrichment of DNA at a known binding site when compared with the DNA immunoprecipitated by a nonspecific control antibody such as anti-IgG and no enrichment at negative control sites should be chosen.

Once a satisfactory enrichment is achieved, the material is sequenced. Finally, the short sequenced reads are computationally mapped to the reference genome and regions where these reads accumulate are identified: this step is known as peak calling. The number of peaks that are identified depends on the algorithm, and in particular the significance threshold chosen. The peaks are the regions where the transcription factor binding (in case of TF ChIP-seq) or the histone modification (in case of histone mark ChIP-seq) is most likely to have occurred.



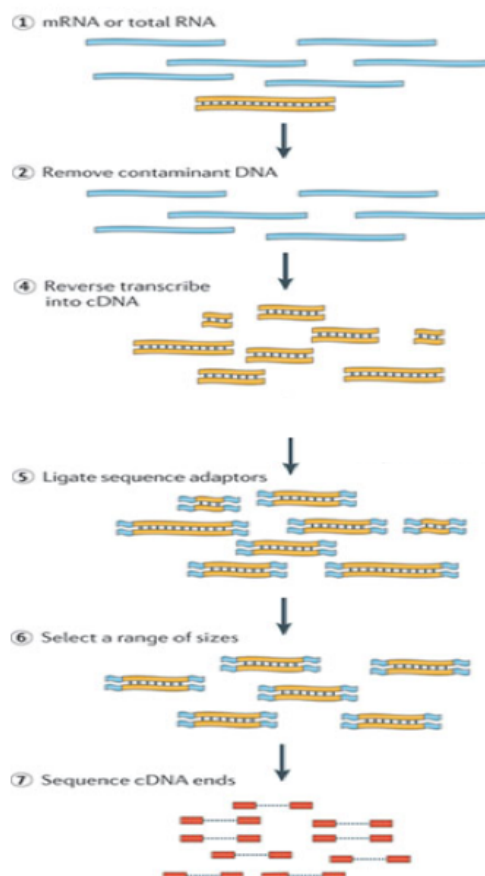
Adapted by permission from Nature Publishing Group: [Furey 2012](#), copyright 2012

Figure 2. The basic steps involved in (a) transcription factor ChIP-seq experiments (b) Histone mark ChIP-seq. The main difference between the two is that the antibody used in the first case is targeted against the transcription factor while in the second it is targeted against the histone modification.

1.1.4.2 RNA-seq

The RNA-Seq technique is a NGS technology that allows measuring the RNA expression levels in a population of cells. In a typical RNA-seq experiment (Figure 3), the first step of the procedure is usually the isolation and purification of RNA from a population of cells which is then reverse transcribed to create a collection of cDNA fragments. These cDNA fragments are then ligated to adapters and sequenced by a high-throughput sequencing

method giving short sequences from one end (single-end sequencing) or both ends (paired-end sequencing). The sequence reads so obtained are usually 30–400 bp in length, depending on the sequencing technology used. After the sequencing step, the raw reads are aligned to a reference genome creating a genome-scale transcription map that provides information on both the transcriptional structure and the level of expression for each gene (Nagalakshmi et al., 2008; Wang et al., 2009).

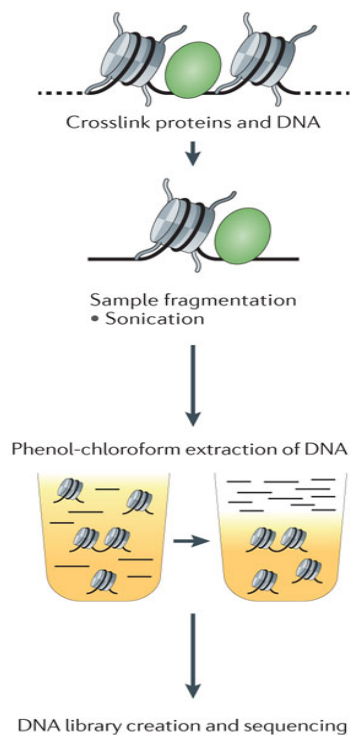


Adapted by permission from Nature Publishing Group: [Martin & Wang 2011](#), copyright 2011

Figure 3. The basic steps in a RNA-seq experiment. mRNA is extracted from the cells of interest, contaminant DNA is removed and the DNA is reverse transcribed into cDNA. These cDNA are then ligated with adaptor sequences and sequenced by high throughput sequencing technology.

1.1.4.3 FAIRE seq

FAIRE-seq or Formaldehyde-Assisted Isolation of Regulatory Elements is a technique that is used to identify open or nucleosome depleted chromatin regions. Nucleosome disruption leading to the opening of chromatin is a well-known hallmark of active regulatory chromatin in the eukaryotic cells. The procedure for carrying out FAIRE was first demonstrated in *Saccharomyces cerevisiae* (Nagy et al., 2003) and has since been applied to human and other mammalian samples too.



Adapted by permission from Nature Publishing Group: [Furey 2012](#), copyright 2012

Figure 4. The basic steps in a FAIRE-seq experiment. First, proteins are crosslinked to the DNA followed by sonication to obtain smaller fragments of DNA bound to the proteins (mostly histones). Finally, the DNA fragments are separated from the proteins (mostly histones) using a phenol-chloroform extraction method.

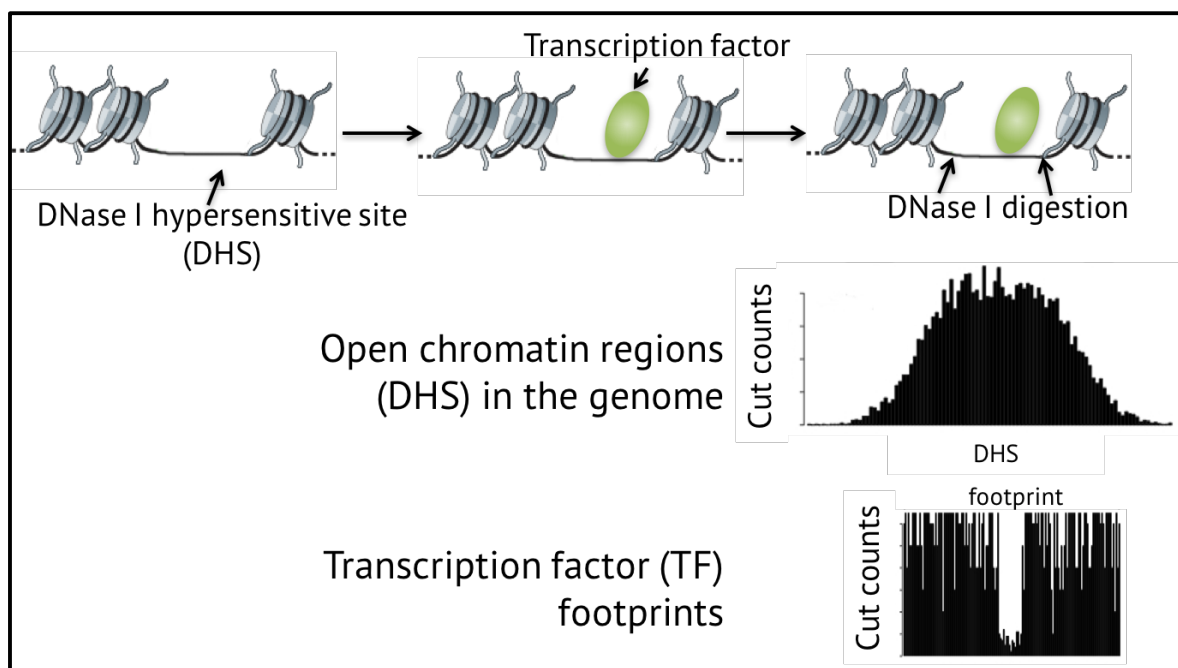
In this approach (Figure 4), phenol-chloroform extraction method is used on formaldehyde-crosslinked chromatin. This leads to the segregation of the open chromatin regions into the aqueous phase as these regions are less packed by the histones proteins (Brutlag et al., 1969; Solomon and Varshavsky, 1985). The DNA fragments are then extracted from the aqueous phase to create a cDNA library which is then sequenced. The FAIRE technique can create a map of the nucleosome distribution throughout the genome. In addition, the promoters of actively transcribed yeast genes were more highly enriched by FAIRE than promoters of genes with lower transcription initiation rates (Nagy et al., 2003). However, this technique suffers from a very low signal to noise ratio (Tsompana et al., 2014), an issue that will be also described in chapter 3.2.

1.1.4.4 DNase-seq and DNase- footprinting

Similarly to the FAIRE-seq technique, DNase-seq identifies genomic regions where chromatin is open and the DNA is accessible to the transcription machinery and the binding of transcription factors (TFs) (Song and Crawford, 2010). This method (Figure 5) combines the traditional approach of mapping DNase I hypersensitive (DHS) sites with high throughput next-generation sequencing: DNase I is used to selectively digest nucleosome-depleted DNA, whereas closed chromatin is more resistant. These regions of open chromatin are called DNase hypersensitive sites (DHSs), owing to their susceptibility to DNase I digestion. DHSs often contain active regulatory elements, including promoters, enhancers, silencers, insulators, and locus control regions (Song and Crawford, 2010; Wu et al., 1979).

The first step of the procedure is usually the isolation and purification of the nuclei from a population of cells, to which DNase I is added to digest the DNA. The DNA is then extracted

using a standard phenol–chloroform extraction protocol. A small amount of the digested DNA of each digested sample is checked on an agarose gel for the appearance of a smear of slightly digested DNA. The DNA fragments are then size selected on an agarose gel or sucrose gradient ultracentrifugation and sequenced (Boyle et al., 2008; Hesselberth et al., 2009). The 5' end of a sequence tag generated by DNase-seq corresponds to the site of a DNase I cut, and regions of enrichment or DHS sites so identified can contain binding sites of multiple factors within it.



Adapted by permission from Nature Publishing Group: [Furey 2012](#), copyright 2012

Figure 5. High depth DNase seq and identification of transcription factor footprints. This technique uses the DNase I enzyme to cut the chromatin regions that are sensitive to the enzyme (open chromatin or DNase hypersensitive sites). The resulting fragments are then enriched using size selection followed by cDNA library creation and high throughput sequencing. If the sequencing depth is high enough, DNase-seq can also reveal the binding sites transcription factor in the DNA.

When sequenced at high enough depth (usually >200 M reads) this technique can also identify small regions (35-50bp) DNase I protection corresponding to the binding sites of transcription factors. DNase I footprinting has been applied widely to study the dynamics of transcription factor occupancy and cooperativity within regulatory DNA regions of individual genes, and to identify cell- and lineage-selective transcriptional regulators. For the ENCODE project, researchers applied the DNase-seq technique to study the chromatin accessibility and transcription factor binding in various cell types. Thurman et al. 2012, carried out DNase-seq in 125 cell and tissue types to identify ~ 2.9 million DHS that contained almost all known cis-regulatory elements. Analysis of these DHSs revealed novel relationships between chromatin accessibility, transcription, DNA methylation, and regulatory factor occupancy patterns. Neph, Vierstra, et al. 2012 carried out high depth DNase-seq to identify 45 million transcription factor footprints across 41 diverse cell and tissue types. These footprints were used to create an extensive core human regulatory network of 475 sequence-specific TFs (Neph et al., 2012b). We will be comparing these footprints with the footprint calls from other algorithms in our benchmarking study described in Chapter 3.1.

Another technique that is commonly used for identifying open chromatin regions is ATAC-seq. This is a rapid and sensitive method to identify open chromatin sites (Buenrostro et al., 2013). However, this method cannot be used to identify TF footprints. DNase-seq, on the other hand, that can potentially identify the transcription factor binding sites on the entire genome. ChIP-seq experiments can only identify the binding sites of a single transcription factor at a time. As mentioned earlier, one of the main goals of my PhD project was to identify the binding partners of Myc. In this thesis, we will demonstrate that the DNase-seq technique can be used to obtain the complete transcription factor binding map in the genome

of the cells of our interest. Although the resolution is not as good as ChIP-seq data, it can still provide a good idea about the most important TF-TF interactions in the system (Barozzi et al., 2014).

1.1.5 Public datasets

As described above, the different next generation sequencing technologies that have been developed in the recent years make it possible to study the epigenome of an organism in detail. The ENCODE Project (Consortium, 2004; ENCODE Project Consortium et al., 2007) used these approaches with the goal to create a catalogue of the regulatory elements in human cells, studying the epigenomic signatures of cells grown in culture. This was followed by the Roadmap Epigenomics Project (Romanoski et al., 2015) that build on the ENCODE project to utilizes different next generation technologies to map DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems that are often found to be involved in human disease. These projects created a reference collection of normal epigenomes that can be used for comparison and integration with various other studies. For this thesis, we will be using some of these datasets that will be described in the Results chapters.

1.2 Objectives

The main objectives of this PhD thesis are:

1.2.1 Identify possible binding partners of Myc in cellular growth and lymphomagenesis

As described earlier, the transcription factor Myc is an oncogene that is overexpressed in many different cancers (Amati et al., 1993; Blackwood and Eisenman, 1991; Lüscher and Larsson, 1999; Walhout et al., 1997). When Myc is overexpressed it binds to almost all open promoters. Still, it only up- or down-regulates a specific subset of genes (Sabò et al., 2014). Moreover, the precise mechanism by which Myc regulates gene expression leading to tumorigenesis is not yet fully understood. Therefore, the main goal of this PhD project was to identify possible binding partners of Myc, either directly or indirectly (“piggy backing”) involved in DNA binding and regulation of gene expression. We apply an approach that integrates different NGS data such as high-depth DNase-seq, RNA-seq and ChIP-seq data to identify possible binding partners of Myc involved in gene-regulation in the 3 Myc models described before: MycER, E μ -myc and tet-MYC (see sections 3.3-3.5).

1.2.2 Choosing a DNase-seq footprint caller

To identify the transcription factor binding footprints from a high depth DNase I seq experiment we need to use specialized algorithm called footprint callers. To the best of our knowledge, at the time of this study no study had been done to compare the available footprint callers. As the footprint calling method play a crucial role in the accuracy of the footprint identified, we carried out a benchmarking study to identify the best footprint caller that we could use to analyse our DNase-seq data. We tested these methods on ENCODE data and assessed the consequences of different footprint calls on the reconstruction of TF-TF regulatory networks (see section 3.1).

1.2.3 Developing a pipeline and methods for the analysis of low depth and high depth DNase-seq data

Although many studies had been done including the ENCODE project, at the time of the study there was no pipeline for the automation of the steps for identifying transcription factor

footprints from high depth DNase seq data. Therefore, an intermediate goal of this PhD project was to develop a pipeline that can carry out step by step analysis of the DNase seq data from the raw sequencing reads which was essential to achieve the major goal of the project (see section 3.2).

Chapter 2

Materials and methods

i. Benchmarking study of footprint callers

Digital footprinting (DGF) data and ChIP-seq datasets for TFs in K562 (human myeloid erythroleukemia), HepG2 (liver hepatocellular carcinoma) and SkMC (skeletal muscle) cell lines were downloaded from the ENCODE project (Consortium, 2004; Thurman et al., 2012). The FIMO (Grant et al., 2011) tool was used to match footprints to the corresponding transcription factors PWMs. Genomic coordinates of the footprints published in (Neph et al., 2012a) in K562, HepG2, and SkMC cell lines, based on the same DGF data and obtained with the FOS metric (Neph et al., 2012a), were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/footprints/. Thresholds on footprint calls for DNaseR (Madrigal, 2013) and Wellington v.0.1.0 were chosen in order to obtain a number of footprints comparable to (Thurman et al., 2012). Only footprints contained in DHSs were considered. Network reconstruction was performed according to the procedure described in (Neph et al., 2012a). For each TF, a window of 10 kbps centered on the RefSeq TSSs was scanned for matches of PWMs in TRANSFAC (Matys et al., 2006) using FIMO and overlapped with footprints using BEDOPS. Receiver-Operator Characteristics (ROCs) and Areas Under the Curve (AUCs) were generated using the ROCR package (Sing et al., 2005). The igraph R package (Cs'ardi, 2006) was used to compute large-scale properties of the inferred networks and to generate random networks.

ii. DNase-seq analysis pipeline

The pipeline was completely written in R/Bioconductor (R Development Core Team, 2013). We used the FASTX toolkit (Pearson et al., 1997) for pre-processing raw FASTQ files. Alignment of the reads to the reference genome is carried out using the Burrows Wheeler Aligner (BWA) (Li and Durbin, 2009), with standard parameters. DNase-hypersensitive sites are identified (DHSs) are identified using the MACS peak caller (Feng et al., 2012) with a p-value cut-off of 10^{-5} and mfold cut-off of 7.3. The Wellington algorithm (Piper et al., 2013) is used for calling footprints on the DHS regions obtained from the DHS-calling block of the pipeline. The FIMO tool is used for matching footprints to transcription factors. Motif over-representation analysis is carried using Pscan (Zambelli et al., 2009). *De novo* motif analysis using MEME-chip (Bailey and Elkan, 1994; Machanick and Bailey, 2011).

iii. Integration of MycER and E μ -myc DNase-seq data with ChIP-seq and RNA-seq

We used ChIP-seq samples and RNA-seq samples obtained in 3T9^{MycER} and E μ -myc systems. The time-points used in 3T9^{MycER} are 0h and 4h after OHT treatment. In E μ -myc system, the samples used are: B-cells from control mice or (C), B-cells from pretumoral E μ -myc mice at 5-6 weeks of age (P) and B-cells from E μ -myc tumour mice at 12-16 weeks of age (T) (Sabò et al., 2014). To generate qualitative heatmaps of the overlaps of DHSs with ChIP-seq peaks we used the ‘compEpitools’(Kishore et al., 2015) R package. For motif analysis, we used MEME (parallel version (Machanick and Bailey, 2011), CUDA-MEME (Liu et al., 2010), STEME (Reid and Wernisch, 2014) and the command-line version of DREME (Bailey, 2011).

iv. Single feature and Random forest classification

Single feature classification of the data was carried out using the ROCR package (Sing et al., 2005). List of up, down and no-deg genes were obtained from RNA-seq experiments (time-points 0h and 4h for 3T9^{MycER} and C, P, T for Eμ-*myc*). While for the random forest classification we used the ‘cforest’ function from the ‘party’ R package (Hothorn, 2005; Strobl et al., 2007). The predictive performance of the models was evaluated using a k-fold (k=10) classification approach and the average AUCs were calculated using the ROCR package. The relative importance of the features was calculated using the variable importance function included in the ‘RandomForest’ function.

2.1 R Functions and methods:

To provide a way to store the DHS and the related footprint information in one place we developed a new S4 R class called the ‘DHS’ class (Figure 6). Each element of an object of this class contains 3 slots that are described below:

DHS slot: a Grange object containing the position of a i^{th} DHS in the genome. The length of this GRnage in this slot is always 1.

FP slot: a GRange object containing all footprints that overlap with the i^{th} DHS. The length of this GRange can be more than 1 because a DHS can often contain multiple footprints.

PWM slot: a list of GRanges containing the PWM matches to each of these footprints (given by FIMO). The length of this list is equal to the length of the GRange in the FP slot, where, the i^{th} element in the PWM slot corresponds to the PWMs that are matched by FIMO to the i^{th} footprint in the FP slot.

Element 1					
Slot "dhs"					
chr	start	end	score		
chr1	4847263	4848996	682.72		
Slot "fp"					
chr	start	end	score		
[1] chr1	4847656	4847696	-56.51		
[2] chr1	4847729	4847754	-10.16		
Slot "PWM"					
[[1]]					
chr	start	end	PWM	p-value	
[1] chr1	4847677	4847688	Myf	7.46e-05	
[2] chr1	4847683	4847695	RFX2	6.44e-05	
[[2]]					
chr	start	end	PWM	p-value	
[1] chr1	4847730	4847739	SP1	1.43e-05	
[2] chr1	4847731	4847740	SP1	7.81e-05	
[3] chr1	4847730	4847737	KLF6	6.67e-06	

Figure 6. Example of the structure of one element of a DHS class. It consists of 3 slots: the first slot (top) is the "dhs" slot contains is a GRange with the position of the i^{th} DHS on the genome. The second slot is the "fp" slot, containing a GRange with the positions of the footprints that overlap with the i^{th} DHS. Finally, the third slot is the "PWM" slot, which is a list of GRanges containing the PWM matches to each of the footprints in the "fp" slot.

'DHS class' specific methods:

We also developed specific methods for the DHS class that are listed below:

fpUnderSummit: returns the footprint closest to the summit of a given ChIP-seq peak (Figure 7). This information can be useful in two ways: first, to estimate the efficiency of the footprinting experiment. Most of these footprints under a summit are expected to contain the motif of the ChIP-ed transcription factor. Second, if the transcription factor

that was ChIP-ed forms dimers or trimers with other transcription factors, we can identify the motif of its partner transcription factors in these footprints.

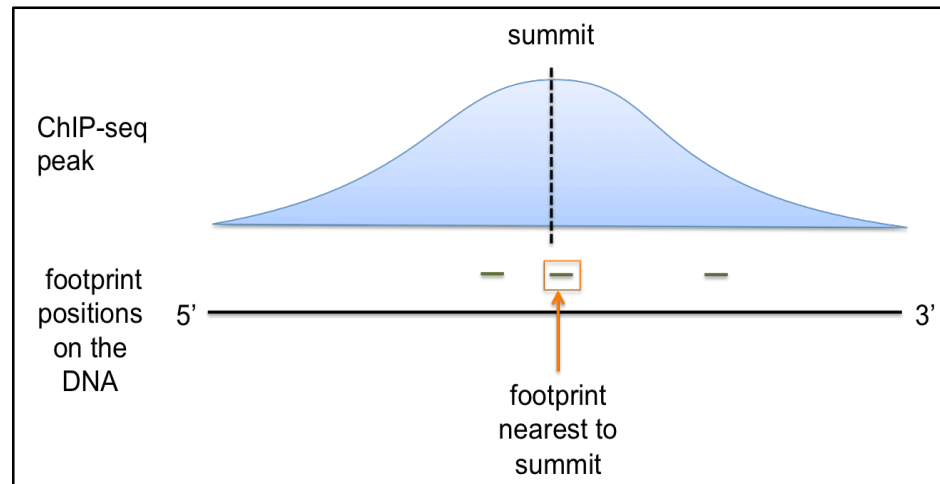


Figure 7. Example showing the nearest footprint to a summit

calOpenChromatin: calculates the percentage of open chromatin given a DHS.

separatePromoters: returns all the elements containing DHSs that are on promoters.

separateDistals: returns the elements containing DHSs that are on distal regions.

calFootprintsOverlapGR: given a list of GRanges and a DHS object, returns a data frame, where each column corresponds to a single PWM class, and the rows the number of FP matches to the corresponding PWM class for each GRRange.

In addition to the footprint and DHS specific functions, during this PhD I also developed functions to facilitate motif analysis. These functions are described below:

submitFIMOjob: given a set of Granges and the specified TF or TFs, runs FIMO (with the p-value cut-off 10^{-4}) on all the given set of sequences using the PWMs that

correspond to the TFs, reads the output and annotates the result with information about the family, class and consensus and returns the output as a GRanges object.

FIMotoGR: takes in the path of a FIMO output file in the .txt form return the results in a Granges format.

motifToSeqLogoHTML: given a list of PWM names, create a HTML page containing the list of PWMs and their corresponding sequence logos using the seqLoGo (<https://github.com/carushi/seqLoGo>) and R Markdown.

methodPscan: Given two sets of sequences, one positive set and the other a negative or background set in the GRanges format, the method converts them to the FASTA format, submits them to PScan to run discriminative motif analysis, carries out multiple testing correction (Benjami-Hochberg) and filters the resulting list based on the user provided corrected p-value cut-off (default 0.01) and writes it in a .xlsx or .txt format. In case of ChIP-seq peaks it also provides the user the option to reduce the size of sequences to a small region around the summit for running the analysis.

More details about the materials and methods can be found in the Results chapters 3.1-3.5.

Chapter 3

Results

3.1 Benchmarking study of footprint callers

3.1.1 Introduction

High depth DNase-seq experiments can be used to identify genome-wide regions of local DNase I protection (footprints) created by the binding of transcription factors to the DNA. The detection of these transcription factor footprints from high depth DNase-seq data involves the identification of a specific signature (sharp trough) in the read density and requires dedicated algorithms for its detection. One of the first approaches for footprint detection was developed and applied to *Saccharomyces cerevisiae* (Hesselberth et al., 2009): this method detects short regions of reduced DNase I cleavage density compared to the immediately flanking regions. Subsequently, a method was also developed for mammalian genomes based on a five-state hidden Markov model (HMM) (Boyle et al., 2011). However, a software implementation of the method was not released. This was followed by other methods such as CENTIPEDE (Pique-Regi et al., 2011), that uses hierarchical Bayesian mixture model to identify genome-wide transcription factor binding sites: first, it matches PWMs from a database to a set of genomic sequences and then classifies the matches as bound or not-bound based on read counts from DNase-seq data around the PWMs. This method therefore only provides the possibility to search for footprints of transcription factors

with an already identified PWM. For the ENCODE project, Neph et al. 2012 adapted the previous method on yeast (Hesselberth et al., 2009) to carry out digital DNase-seq footprinting (high depth DNase-seq) data in human samples. However, the algorithm was not released at the time of this study.

The approaches described above were reviewed and compared by Piper et al., 2013, who introduced Wellington, an algorithm for footprint detection, which leverages on a characteristic pattern of strand imbalance in the sequenced fragments surrounding the protein-DNA binding sites. There, Wellington scored best against the previously published tools. Another tool, called DNaseR (Madrigal P., 2014) and published on Bioconductor, utilizes the Skellam distribution to detect the imbalance between sequencing reads on the two strands, thus representing a potential alternative to Wellington.

At the time of developing our DNase-seq pipeline, there were no studies comparing Wellington to DNaseR. Therefore, in order to select the best footprint caller for our pipeline, we carried out a detailed comparison of the accuracy of the footprint calls given by the two methods, as well as those obtained using the FOS (Neph et al., 2012c) and their effects on the resulting TF-TF regulatory networks. The results of benchmarking study have been published in Barozzi *et al.*, 2014.

3.1.2 Materials and methods

The performances of DNaseR and Wellington footprint calls were compared using DGF data from the K562 (chronic myelogenous leukaemia) cell line. The footprints from Footprint Occupancy Score (FOS) on the same cell line (Neph et al., 2012a) were also included in the

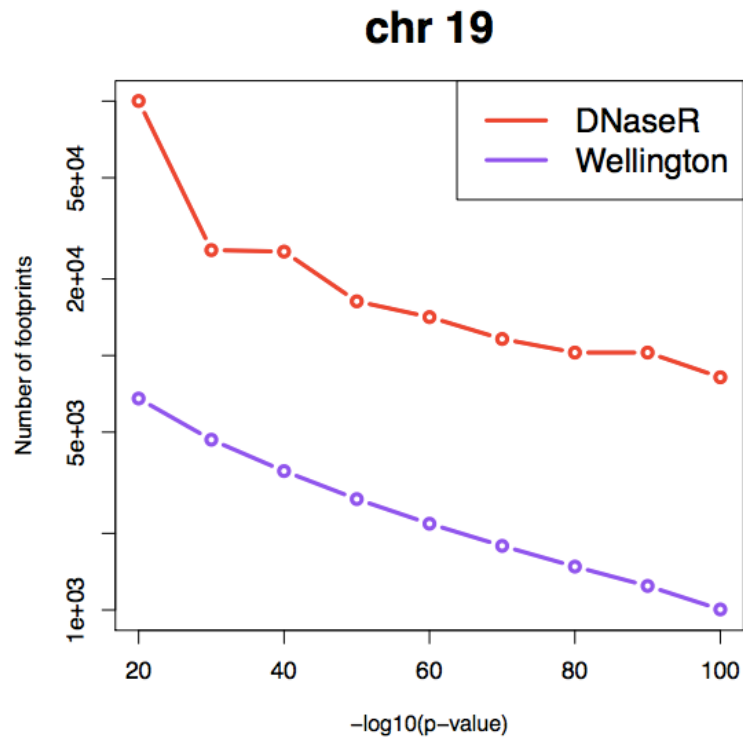
comparison. Thresholds on footprint calls for DNaseR (Madrigal, 2013) and Wellington v.0.1.0 were chosen in order to obtain comparable number of footprints using the two methods (DNaseR: 1,075,979; Wellington: 1,833,281), which was also comparable to the number reported by Neph et al. (498,683). ChIP-seq data corresponding to 11 transcription factors in K562 cells were downloaded from ENCODE to validate the footprint predictions obtained with DNaseR and Wellington. For each TF, a window of 10 kbps centered on the RefSeq TSSs was scanned for matches of PWMs in TRANSFAC (Matys et al., 2006) using FIMO (Grant et al., 2011) and overlapped with footprints using BEDOPS. Receiver-Operator Characteristics (ROCs) and Area Under the Curve (AUCs) were generated using the ROCR package (Sing et al., 2005). ROC curves are used to visualize the performance of a binary classifier; by plotting the false positive rate against the true-positive rate. Consequently, the AUC of a ROC curve is a measure of how well a classifier can distinguish between the two classes. Network reconstruction was performed according to the procedure described in (Neph et al., 2012a). The igraph R package (Csardi, 2006) was used to compute large-scale properties of the inferred networks and to generate random networks.

3.1.3 Results

3.1.3.1 Comparison of footprint callers

Using the digital footprinting datasets in K562 cell line from ENCODE, we followed the approach used by Piper et al. 2013 to compare footprints obtained by DNaseR, Wellington and Neph, et al. 2012. First, we extracted the footprints calls on the DHSs in the K562 cell line with DNaseR and Wellington and compared them to the set of footprints obtained using the FOS metric. While Wellington runs only on the genomic coordinates of the DHSs, DNaseR looks for footprints on the entire genome. Hence, to make the tools comparable we

restricted the footprint search of DNaseR to the DHSs only. DNaseR consistently identified more footprints than Wellington at comparable stringency levels.



Barozzi et al, Front. Genet. 2014

Figure 8. Number of footprints called by DNaseR and Wellington at different stringency levels.

ChIP-seq experiments are used to identify the binding sites of transcription factors to the DNA genome-wide. Therefore, the presence of a ChIP-seq peak overlapping a footprint can be represent a validation of the presence of a transcription factor on the footprint. We used 17 binding patterns from ChIP-seq experiments corresponding to 11 TFs in K562 to validate the 3 sets of footprint calls obtained using Wellington, DNaseR and FOS. For these 11 TFs, we computed the ROCs for the predictions generated by the binding motifs only (Figure 9A) and for the three sets of footprint calls mentioned above (Figure 9 B-D). A footprint that

overlapped with one of the known binding motifs of a specific TF was considered as a prediction corresponding to an actual binding event of the TF.

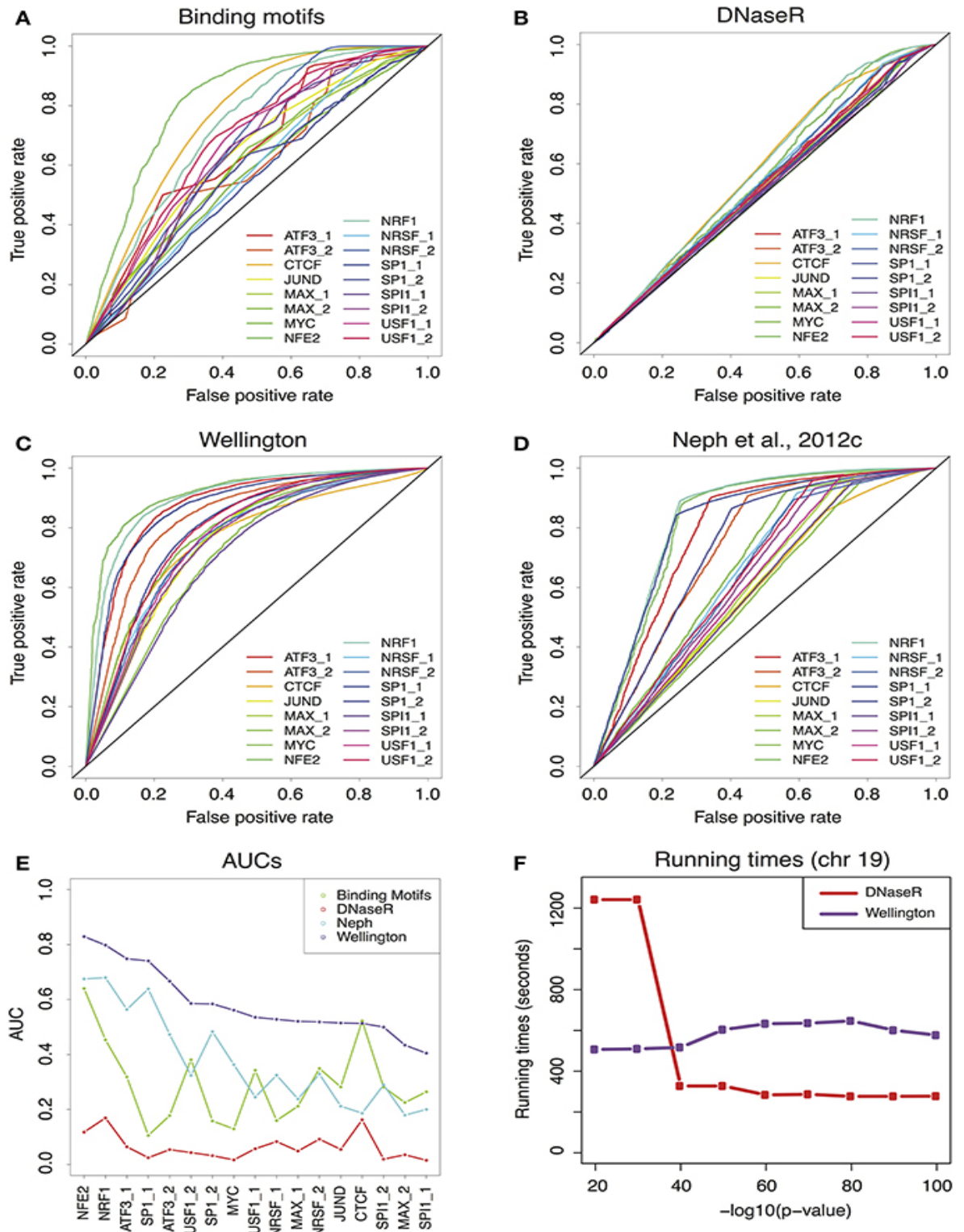


Figure 9. (A) Receiver-Operator Characteristic (ROC) curves for the predictions provided by the binding motifs alone. (B–D) ROCs for the sets of footprints obtained by DNaseR, Wellington and for the set used in (Neph, et al. 2012). The kinks in the ROCs of the Neph footprints indicate the absence of the data at those cut-offs. (E) AUCs of ROCs obtained using binding motifs alone, DNaseR, Wellington and the set used in Neph et al. 2012 for the 17 binding patterns from ChIP-seq experiments corresponding to 11 TFs in K562 cell line. The Wellington tool shows the highest predictive power irrespective of the TF considered. (F) The running times of DNaseR and Wellington with respect to decreasing p-value cut-offs. DNaseR ran considerably slower at lower p-value cut-offs.

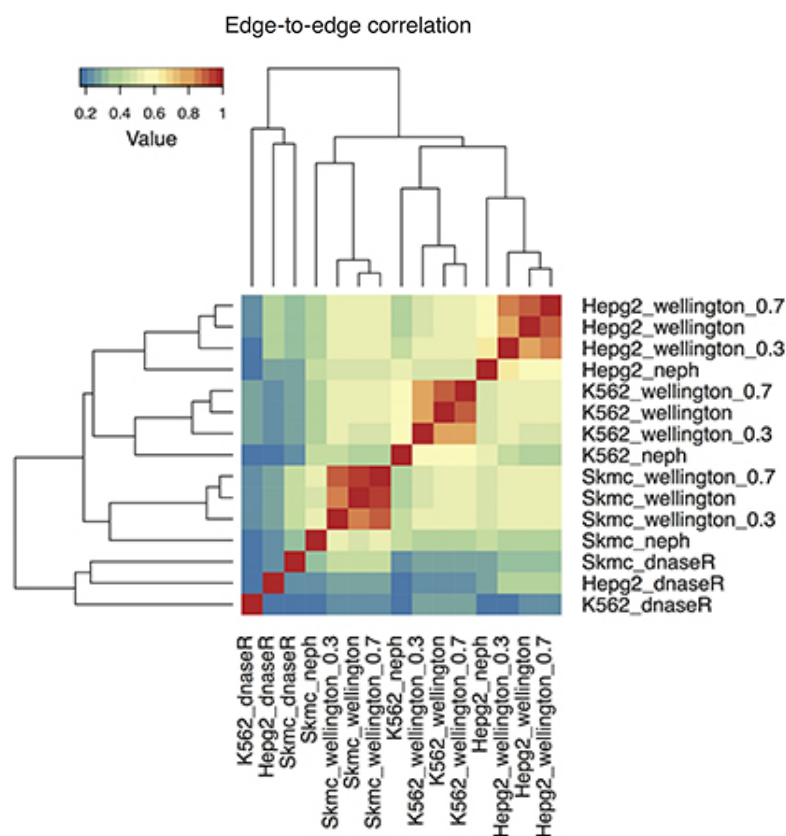
In figure 9E, we show the global performances of the methods by the AUC of each of the ROCs. Irrespective of the considered TF, the method with the highest predictive power is always Wellington. However, it must be noted that the AUC calculated using the FOS score (Neph, et al. 2012) might have been underestimated, as we could not perform a more permissive footprint call, due to the lack of the required software. Remarkably, for many TFs, the overlap of their motifs with DHS coordinates already provided a considerably good prediction of their binding sites without the need to consider any footprint information. For some of these TFs the prediction was comparable to (USF, NRSF, SPI1, MAX, JUND) or better (CTCF) than the footprints calculated in (Neph et al., 2012b). For factors, like CTCF, due to its 11 zinc fingers has a highly specific motif and almost always binds to regions where its motif is present. On the other hand, the performance of DNaseR remained consistently poorer than the other methods, indicating that the majority of the DNaseR footprints do not correspond to validated binding sites of a TF. Moreover, the DNaseR calls were adding new footprints to the sets of footprints for different significance thresholds that did not overlap well with each other; on the other hand, Wellington showed the expected behaviour.

We compared the running times (Figure 1F) of Wellington and DNaseR on chromosome 19 for varying significance thresholds. While Wellington consistently ran at approximately the same speed, DNaseR ran much slower for permissive calls.

3.1.3.2 Robustness and characteristics of the inferred networks

We used the two sets of footprints to reconstruct the TF-TF regulatory network on K562, skeletal muscle cells (SkMC) and HepG2 data as done in Neph et al. (2012). In addition, we studied the impact of sequencing depth on the network reconstruction by running Wellington on progressively down-sampled alignment files for the three cell lines, and reconstructing the corresponding TF-TF networks. The regulatory networks thus obtained were compared against each other by counting how often a specific edge is present between each pair of nodes. The heatmap in figure 2 displays the edge-to-edge correlation between all pairs of samples.

The network reconstruction using footprints called by Wellington or using the FOS score provides a better separation of cell types compared to the network obtained from DNaseR footprints. In addition, the networks obtained using Wellington with decreasing sequencing depth remain very similar. This indicates that most of the footprints that are not identified at lower depths are weak footprints that do not correspond to real interactions between TFs with their annotated binding preferences. Based on these results we choose Wellington as the tool of choice for footprint calling.



Barozzi et al, Front. Genet. 2014

Figure 10. Heatmap showing the edge-to-edge correlation among the TF-TF networks reconstructed with the sets of footprints obtained with DNaseR, Neph, Wellington in three different cell lines (K562, SkMC, HepG2). This heatmap compares the relationship between the networks obtained in the 3 cell lines that were studied (K562, Hepg2 and SkMC). The suffix 0.7 and 0.3 in the sample name indicates the networks obtained by down-sampling the reads to use 70% and 30% of the original reads respectively. Irrespective of the read density, the network reconstruction using footprints called by Wellington or using the FOS score provides a better separation of cell types compared to the networks obtained from DNaseR footprints.

3.2 DNase-seq analysis pipeline

3.2.1 Introduction

As stated before, the main goal of this thesis was to identify binding partners of Myc in the 3T9^{MycER} and E μ -myc using DNase-seq footprinting data. The DNase-seq technique identifies genome-wide open chromatin regions and if the sequencing depth is high enough

(usually more than 100,000,000 reads) it can also identify small regions of DNase I protection (around 30-50bp) due to the binding of a protein such as a transcription factor. These regions are called transcription factor footprints. However, at the time of this study there were no pipelines that allowed the automation of all the steps from processing raw sequencing data to the identification of open chromatin regions (DNase I Hypersensitive Sites, or DHS) and footprints, followed by motif analysis. To address this issue, we developed a pipeline that automates all data analysis steps for both high-depth and low-depth DNase-seq data. The pipeline can also be used to analyze FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements), which is another technique used to identify open chromatin regions.

3.2.2 Materials and methods

3.2.2.1 DNase-seq pipeline

The pipeline is written in R/Bioconductor and step-by-step transforms raw FASTQ files obtained from Illumina sequencing into DHSs and TF footprints in the standard BED (Browser Extensible Data) and bigBed format. In addition, it also provides options for running motif analysis on the DHSs and footprints. The pipeline is divided into five blocks that are detailed below.

Pre-processing block

The pre-processing block offers three different options for filtering FASTQ files: reads with low quality scores can be 1) discarded, 2) trimmed and/or 3) nucleotides with low quality scores can be masked using the FASTX toolkit (Hannonlab.cshl.edu/fastx_toolkit/). The filtered reads are then passed on the alignment block.

Alignment block

This block aligns the reads to the reference genome using the Burrows Wheeler Aligner (BWA) (Li and Durbin, 2009), with default parameters. The output is stored in the BAM format containing the information regarding the alignment of each read.

DHS-calling block

This block of the pipeline carries out DHS-calling to identify regions of open chromatin from the BAM files using the Model-based Analysis of ChIP-seq (MACS2) peak caller (Feng et al., 2012) with the user selected p-value cut-off (default 10^{-4}). The outputs of this block are the genomic coordinates of the DHSs in the BED and bigBed format. The bigBed files can be exported to a genome browser for visualization.

Footprint calling block

The footprint calling block of the pipeline uses the Wellington algorithm (Piper et al., 2013) based the benchmarking to choose the best footprint caller study (chapter 3.1). We use the user selected p-value cut-off (default 10^{-10}) to call footprints on the DHSs obtained from the previous block. The pipeline outputs the identified footprints in the BED and bigBed format.

Motif analysis block

The Motif analysis block of the pipeline provides three options, namely:

a. finding over-represented motifs using Pscan (Zambelli et al., 2009)

This option allows the user to search for motifs that are enriched in a set of sequences compared to a background set. The input required for this analysis is a PWM (Position

Weight Matrix) library, a set of sequences of interest and a background set of sequences in the FASTA format. We use the command line version of the Pscan tool to run this analysis.

b. *de novo* motif analysis using MEME and DREME

This option carries out *de novo* motif enrichment analysis in a set of sequences of interest. MEME motif search is based on expectation maximization (EM) algorithm which allows parameter estimation in probabilistic models with incomplete data (Bailey and Elkan, 1994; Do and Batzoglou, 2008). Although MEME can discover complex motifs, it may require long processing times (>24h) depending on the size of the motif and the size of the sequences. DREME on the other hand, uses a non-probabilistic regular expression search and is optimized to search short motif sequences (4-8 nt long). It can therefore, search for many small motifs on an entire set of DHSs in a relatively short time (< 1h).

The motifs identified by these two tools are matched to a library of known PWMs using TOMTOM (Gupta et al., 2007). We then create a union of the motifs found by the two tools and the output is stored in an .xlsx or .txt document containing the identified motif and the transcription factor matched to the motif.

c. *Matching footprints to a known motif using FIMO*

The footprints are matched to a corresponding transcription factor using the FIMO (Grant et al., 2011) tool: FIMO scans for PWMs from a given PWM library in the footprint sequences and returns the matches in a BED format.

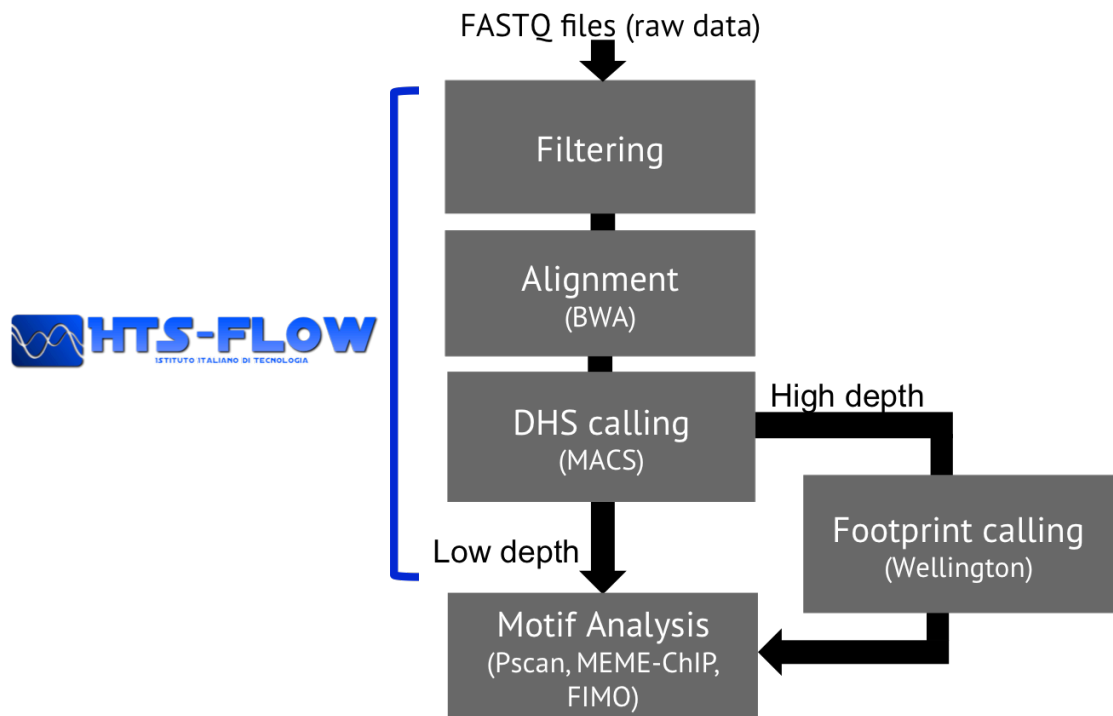


Figure 11. DNase-seq pipeline. The pipeline is divided into 5 blocks. In the first block the raw FASTQ files from the sequencer are filtered using the FASTX toolkit (Hannonlab.cshl.edu/fastx_toolkit/). Next, in the alignment block, the filtered reads are aligned to the reference genome using the BWA (Li and Durbin, 2009) aligner. This is followed by the identification of the enriched DNase-hypersensitivity sites (DHSs) on the genome using MACS (Feng et al., 2012). If the sequencing has sufficient depth, these DHSs are then passed on to the footprint identification block to search for footprints of transcription factors using the Wellington (Piper et al., 2013) tool. Finally, the footprint sequences are then matched to their corresponding transcription factors, by FIMO (Grant et al., 2011). The user can also choose to run de novo motif analysis with MEME-ChIP (Machanick and Bailey, 2011) or motif over-representation analysis using Pscan (Zambelli et al., 2009) on the DHSs and footprints. The final output of the pipeline are the DHSs and footprints matched to PWMs in the BED and bigBed formats. For low depth DHS data, the footprint-calling block is skipped and the DHSs can be sent directly to the motif analysis block. All the analysis blocks up to the footprint calling block has been integrated in a framework for NGS data analysis and management called HTS-flow (Bianchi et al., 2016).

The motif analysis methods outlined above require a database containing the PWMs corresponding to transcription factor binding sites. To create a comprehensive set of PWMs for footprint matching by FIMO, we collected 2433 PWMs from different databases like

JASPAR (Mathelier et al., 2014), UniPROBE (Newburger and Bulyk, 2009), HOCOMOCO (Kulakovskiy et al., 2016), SwissRegulon (Pachkov et al., 2013) as well as other published PWM sets as those from Bucher, 1990, Berger et al 2008, Hallikas et al 2006, Wei et al. 2010, Jolma et al., 2010, Jolma et al., 2013 and Jolma et al., 2015 (includes composite motifs). Many transcription factors bind together as a complex and as a result recognize new composite sites which vary largely from their individual motifs (Jolma et al., 2015). Jolma et al. 2015 applied a technique called consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX) to identify TF-TF pairs that bind to the DNA together. They identified 315 TF-TF pairs that recognize 618 heterodimeric motifs, which are referred to as “composite motifs”.

It is crucial to note that different transcription factors can bind the same or very similar PWMs and the same transcription factor can be associated to similar PWMs. Hence, when matching genomic regions to PWMs from multiple databases, we often found multiple transcription factors associated to the same genomic region. To remove this redundancy and to allow better summarization of results, we calculated the similarity between pairs of PWMs in our collection using a Pearson coefficient-based function from the TFBSTools R package (Ge, 2015). If the correlation between two PWMs is ≥ 0.9 , they are placed into the same class. PWMs that correspond to the same transcription factor are automatically put in the same class, irrespective of their similarity score. Using this approach, we classified the 1815 non-composite motifs into 445 classes and the 618 composite motifs into 295 classes. We created an easy to read HTML page containing the name of the PWM, its corresponding class (based on similarity), the protein family (based on common evolutionary origin) the transcription actor belongs to and the sequence logo for all the 2482 motifs that we collected. Figure 4 shows a snapshot of the HTML page.





PWM	TF	Class	Family	Consensus
RUNX1	RUNX1	Class1	Runt-related_factors	
HC_RUNX1_f1	RUNX1	Class1	Runt-related_factors	
ZEB1	ZEB1	Class2	HD-ZF_factors	
NR4A2	NR4A2	Class2	NGF-B-related/receptor(NR4)	

Figure 12. Snapshot of the HTML page containing the list of PWMs collected from various sources. The first column contains the name of the PWM, the second contains the name of the corresponding transcription factor, the third contains the class of the transcription factor based on the similarity between factors, the fourth column contains the name of transcription factor family and the fifth column contains the sequence logo corresponding to the PWM.

All the blocks of the pipeline up to the footprint-calling block has been added to HTS-flow workflow management system (Bianchi et al., 2016), a framework for NGS data analysis and management.

We tested the pipeline on the DNase I seq samples from 3T9^{MycER} system at low depth (2 replicates each A and B) and high depth (without replicates) in 0h and 4h after OHT treatment and in the E μ -myc system from control (C, non-transgenic mice), pre-tumoral (P, an intermediate stage before the development of full-fledged tumours) and the tumoral stage (T). The inputs used for the DNase-seq analyses were sonicated genomic DNA samples.

3.2.3 Results

3.2.3.1 Application of the DNase-seq pipeline to 3T9^{MycER} and E μ -myc

i. Low depth DNase-seq and FAIRE-seq

We tested the pipeline on low depth DNase-seq and FAIRE-seq data (~ 60M raw reads) in the 3T9^{mycER} samples. We obtained an average of 35M uniquely aligned reads in DNase-seq and 40M in FAIRE-seq samples (using sonicated genomic DNA as a control). We called an average of 36,000 peaks/DHSs in the DNase-seq samples (spanning ~1.7% of the mouse genome) and an average of 14,000 peaks for FAIRE-seq (spanning ~0.8% of the mouse genome). More than 60% of the peaks called by FAIRE-seq experiment were also included in the matched DNase-seq peaks, indicating that the two methods have a significant overlap. However, the FAIRE-seq signal was consistently weaker, displaying a lower signal to background ratio and thus providing a lower resolution when compared to the DNase-seq signal. Due to this poor performance of the FAIRE-seq in comparison to DNase-seq, we chose to focus on DNase-seq experiments for our further analysis on Myc dependent gene-regulation.

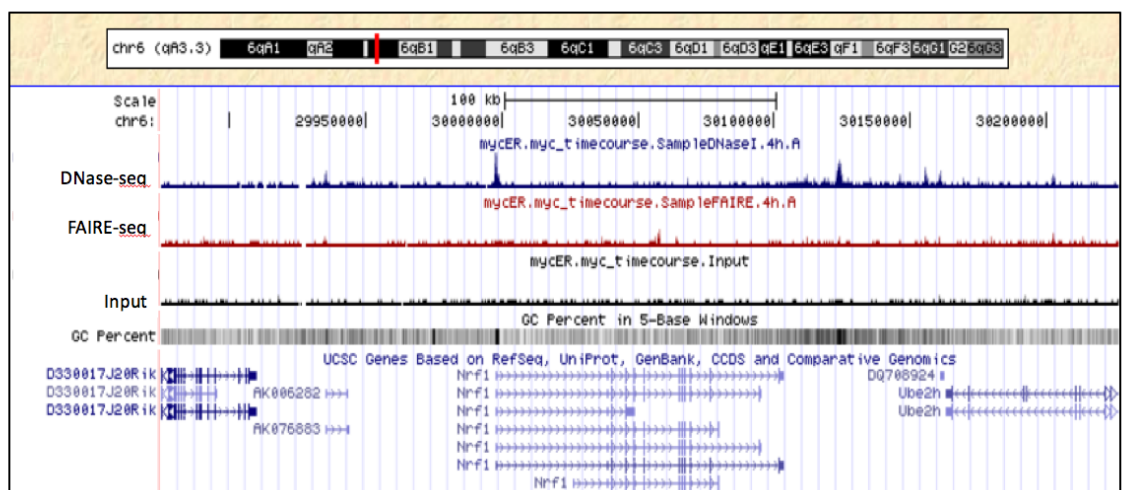


Figure 13. Screenshot of the genome browser displaying the low depth DNase-seq (in blue) and FAIRE-seq (in red) and the input (black) tracks in 3T9^{MycER} 4h A sample.

ii. High-depth DNase-seq (DNase-footprinting)

As the low depth DNase-seq experiments showed promising results, we decided to follow up these experiments with high depth DNase-seq analysis that could identify genome-wide footprints of transcription factors. We analysed high-depth DNase-seq sequencing (>400M sequencing reads) in 3T9^{MycER} at 0h (before Myc activation) and 4h (after Myc activation) time-points and in Eμ-*myc* on the Control (C), Pre-tumoral (P) and Tumoral (T) samples. Using the DNase-seq pipeline (see Chapter 3.2), the number of reads we aligned were 356M in 3T9^{MycER} 0h, 357M in 3T9^{MycER} 4h, 312M in Eμ-*myc* C, 281M in Eμ-*myc* P and 325M in Eμ-*myc* T. We identified 120,626 and 118,334 DHSs in 0h and 4h 3T9^{MycER} samples, respectively. In Eμ-*myc*, the number of DHSs identified was comparable between C (47,542) and P (48,414) and slightly increased to 53,593 DHSs in T sample. Previous ChIP-seq and RNA-seq results in our lab (Sabò et al., 2014) in the Eμ-*myc* system showed that there is a progressive increase in number of Myc binding sites and its mRNA levels from C to P to T. The number of DHSs seems to correlate with this increase in this over-expression of Myc. It is possible that the over-expression of Myc in the P and T samples indirectly causes the opening of new chromatin regions. We characterized the location of these DHSs on the genome (Figure 14) and found that the majority of the DHSs in the 3T9^{MycER} samples were on genebody and intergenic regions. The DHSs in Eμ-*myc* samples C and P are distributed almost equally between the genebody, intergenic and promoter regions while in the sample T there were more DHSs on the intergenic and genebody regions, similar to the 3T9^{MycER} samples. This shows that when Myc is expressed at low levels it binds more to open promoters but at high level it starts invading also the distal regions (Sabò et al., 2014).

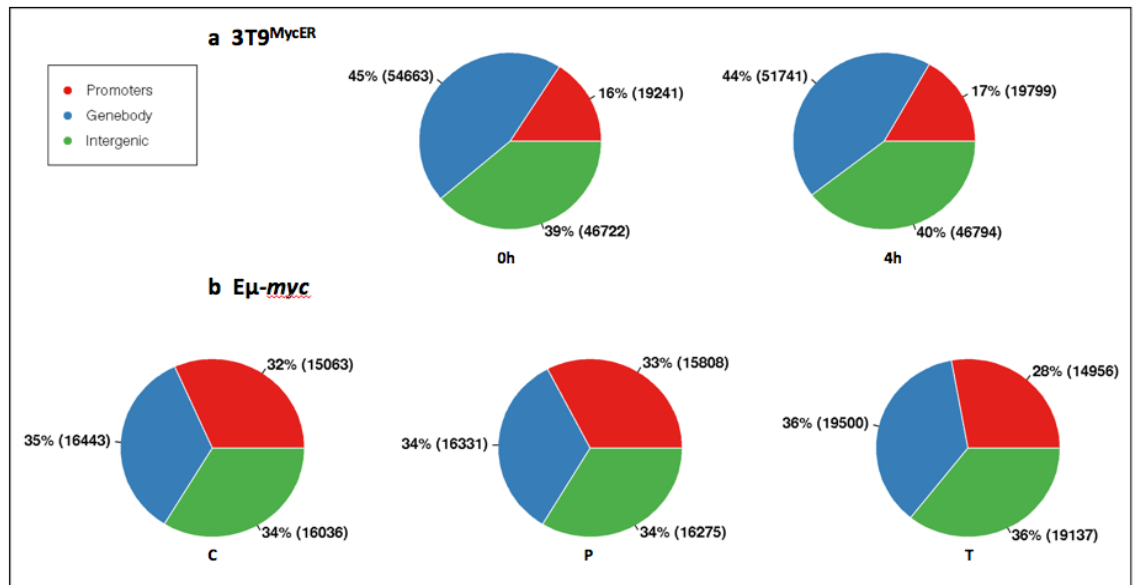


Figure 14. Distribution of DHSs (from high-depth DNase-seq experiments) on promoters, intergenic regions and genebody for (a) 3T9^{MycER} 0h and 4h (left to right) and (b) Eμ-*myc* C, P and T (left to right). In the 3T9^{MycER} samples we detect a higher percentage of peaks on genebody and intergenic regions compared to the Eμ-*myc* samples.

Most of the DHSs that we identified earlier in the low depth DNase-seq samples are also found in the high depth samples (Figure 15).

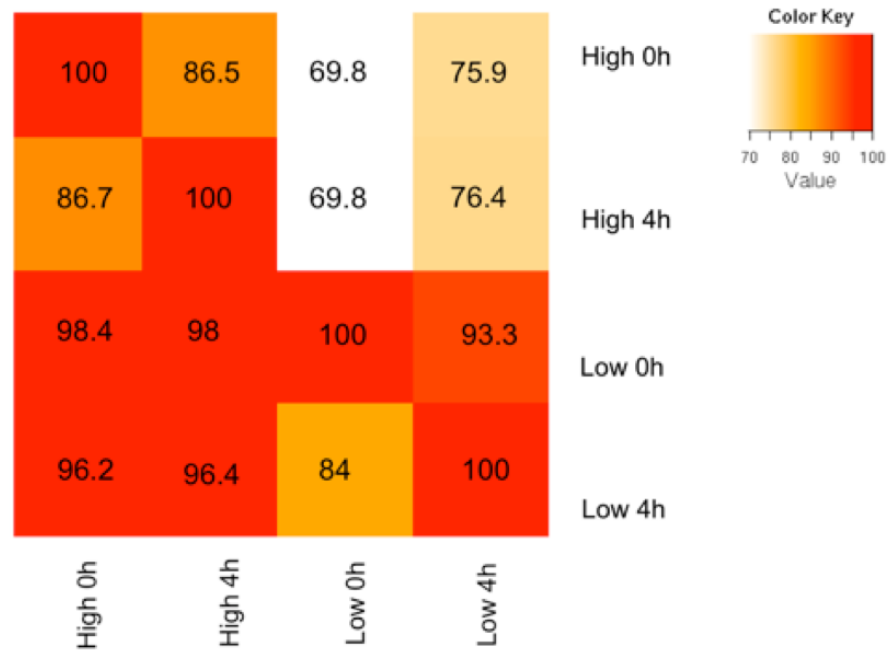


Figure 15. The overlap of low depth and high depth DHSs in 0h and 4h 3T9MycER. The majority of the low depth DHSs (>96%) are also present in the high depth data.

We obtained 528,137 footprints in 3T9^{MycER} at 0h and 328,715 at 4h. In E μ -myc, we identified 90,452 and 85,841 footprints in C and P respectively, and almost thrice the number in T (286,041 footprints). The sample T as mentioned before, is a tumour condition, and therefore expected to have more DHSs and footprints compared to the other samples. Figure 16 shows the tracks of the complete and the per base (reads shortened to a single nucleotide base from the 5' end) signals. The footprints identified by Wellington show a good correspondence with the per base signal (Figure 16).

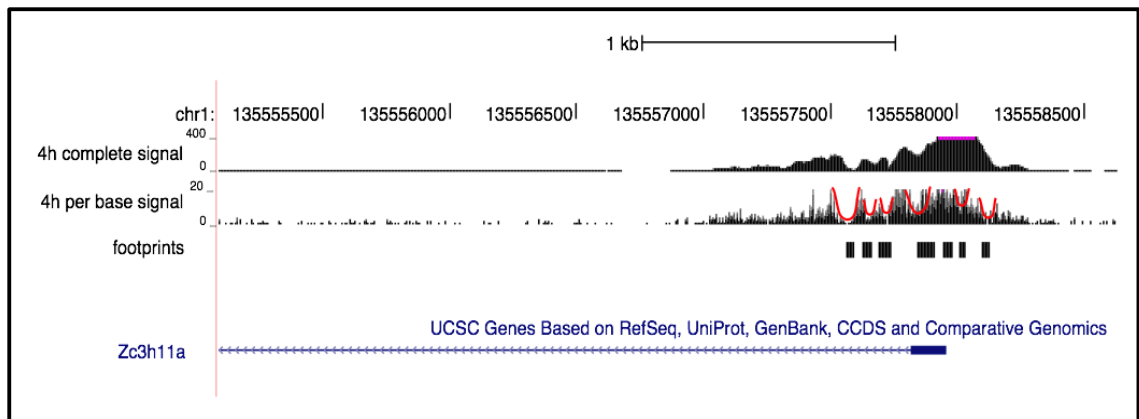
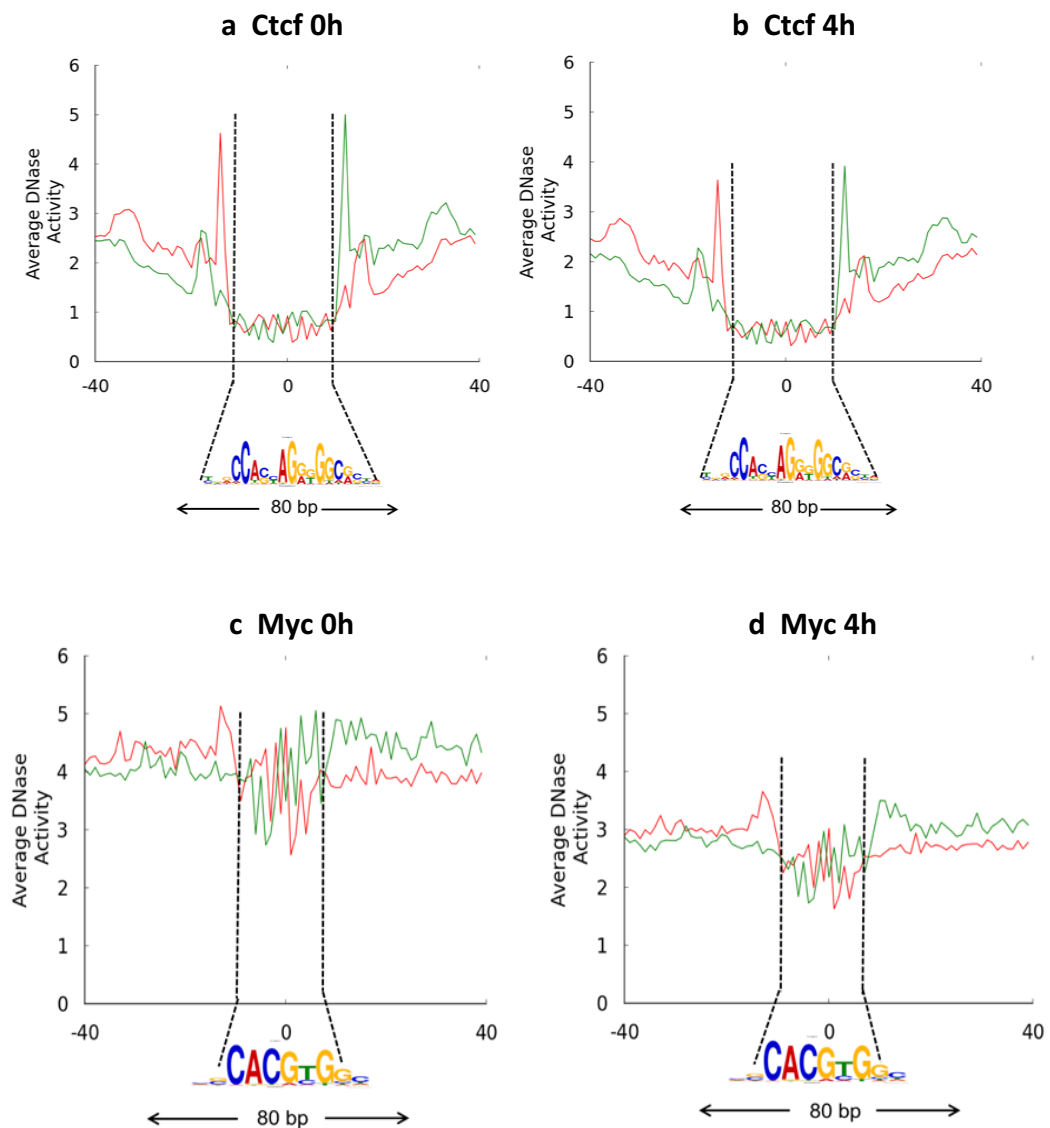


Figure 16. Tracks showing the signals of DNase-seq reads in 4h 3T9^{MycER}, corresponding to the DHS, per base signal and the footprints. The red troughs highlight the drop in the DHS signal corresponding to the position of the footprint.

Recently, it has been shown that the footprinting efficiency of transcription factors is dependent on their residence time on the DNA (Sung *et al.*, 2015), thus raising questions about the use of footprinting data in identifying TF binding sites. In order to test this observation, we computed the digestion profiles of three transcription factors Myc, Ctf and Sp1 around their motifs at low Myc (0h) and high Myc (4h) conditions (figure 8). Ctf, with its 11 zinc fingers, forms a very strong bond resulting in a longer residence time and thus leaving stronger footprint irrespective of the Myc levels as evident from its profiles. Sp1 has only one zinc finger and therefore forms a weaker bond with the DNA, resulting in low DNase I protection. Although weaker than Ctf, the profile of Myc unlike Sp1 shows a clear region of DNase I protection, indicating that its footprints can be trusted to be real. As expected, the profiles of both Ctf and Sp1 remained unchanged during the transition from 0h to 4h. The profile of Myc at 4h however, shows a visible difference when compared to its profile at 0h, indicating that the DNase I protection profiles are affected by the concentration of the transcription factor inside the nucleus. Moreover, this analysis showed that indeed some transcription factors such as Sp1 (owing to their structure and the site they

recognize) might form bonds not strong enough to form a footprint, factors such as Myc and Ctfc do create identifiable footprints. Hence, the DNase-seq technique can be confirmed to be a useful tool for identifying the binding of the transcription factors that leave a footprint.



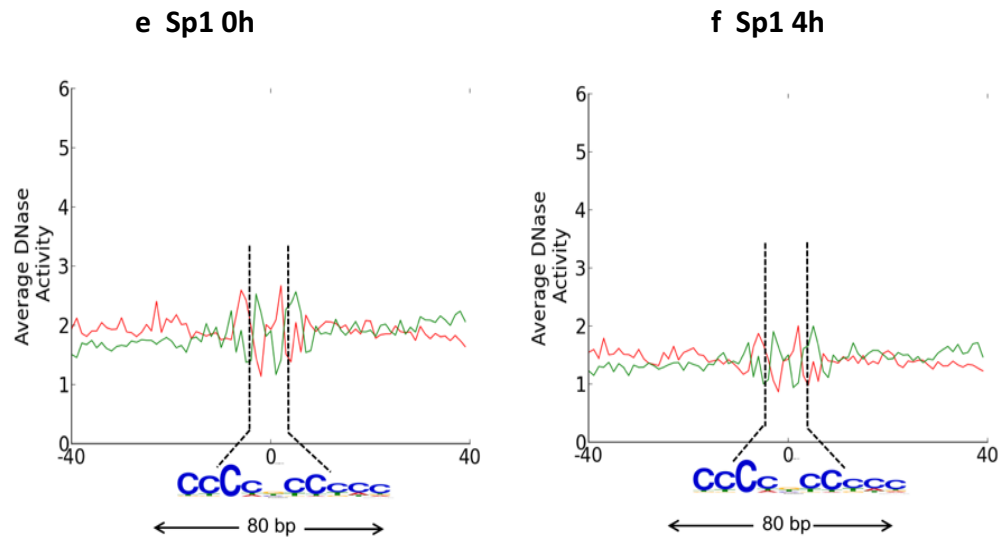


Figure 17. Aggregate per base DNase I cleavage patterns for three transcription factors in the $3T9^{MycER}$ cells in untreated (0h) and treated (4h) conditions for (a-b)(L-R) Myc 0h and 4h (c-d) Ctf 0h and 4h (d-e) Sp1 0h and 4h. The profiles show the expected trend of DNase I digestion for the three factors. Ctf with a complex motif has well defined DNase I protected regions (footprints), Myc shows an intermediate level of digestion and the DNase I protection increases after 4h. Sp1 shows the lowest level of DNase I protection with poor footprints.

Next, we compared the differences between the two systems in relation to DHSs. Figure 18 shows the overlap among the all DHSs in all conditions. Around 73% of the DHSs in 0h $3T9^{MycER}$ overlap with those in 4h. In $E\mu-myc$, the overlap of the DHSs between conditions ranged from 82% for C with P to 81% for P with T and finally to 78% for C with T. This indicates that the regions of open chromatin do not change drastically during the transition from 0h to 4h and from C to P to T and therefore do not depend completely on the Myc levels (Sabò et al., 2014). However, the overlap of the DHSs of $E\mu-myc$ with those of $3T9^{MycER}$ for all the conditions was only around 58%, indicating that while there is a common set of DHSs, there are also many DHSs specific to each system. This is expected, since the two systems $3T9^{MycER}$ (fibroblasts) and $E\mu-myc$ (B-cell lymphoma) are biologically very

different. A similar trend is seen also with the footprints, with ~68% of the C and P footprints overlapping with the T footprints. In 3T9^{MycER} too, there is a good overlap of the footprints of 0h with that of 4h (>70%).

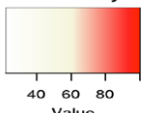
a DHS overlaps

100	82	78	59	58	euMyc_C_DHS
84	100	81	59	59	euMyc_P_DHS
67	69	100	57	57	euMyc_T_DHS
23	22	25	100	73	mycER_0h_DHS
24	23	27	78	100	mycER_4h_DHS
euMyc_C_DHS	euMyc_P_DHS	euMyc_T_DHS	mycER_0h_DHS	mycER_4h_DHS	

b Footprint overlaps

100	39	68	51	49	euMyc_C_footprints
41	100	68	47	45	euMyc_P_footprints
25	24	100	43	42	euMyc_T_footprints
17	15	39	100	70	mycER_0h_footprints
17	15	37	71	100	mycER_4h_footprints
euMyc_C_footprints	euMyc_P_footprints	euMyc_T_footprints	mycER_0h_footprints	mycER_4h_footprints	

Color Key



40 60 80
Value

Figure 18. (a). Heatmap showing the percentage of overlap among the DHSs of 3T9^{MycER} (4h/0h) and Eμ-*myc* (C, P and T). The overlap among the DHSs of the same system is higher than the overlap between the two systems is low (>67% vs. ~58%). (b) Heatmap showing the percentage of overlap among the footprints of 3T9^{MycER} (4h/0h) and Eμ-*myc* (C, P and T).

The footprints obtained from the pipeline were passed on to the motif analysis block. Around 50% of the identified footprints were matched to a TF PWM class in our PWM library. The remaining footprints are most likely generated by transcription factors whose motif is not yet known. This approach is therefore limited by the number transcription factors with a known PWM.

3.3 Integration of 3T9^{MycER} and E μ -myc DNase-seq data with CHIP-seq and RNA-seq

3.3.1 Introduction

The Myc transcription factor acts as a master regulator of cellular growth (Dang, 2013) and has been shown to drive differential gene-expression of specific subsets of genes (Sabò et al., 2014). As mentioned in Chapter 1, Myc can bind together with other transcription factors forming complexes to control the expression of other genes to activate and repress its target genes (Kress et al., 2016; Varlakhanova et al., 2011; Wiese et al., 2013). The most likely binding scenarios are: (i) Myc dimerizing with another member of the BHLH family to bind the same sequence of DNA as it does with Max, (ii) Myc binding close to another TF with a small gap in between the recognition sites of the two (iii) Myc binding indirectly to the DNA through another TF (“piggy backing”) (iv) Myc binding aspecifically to the DNA without recognizing a specific motif. Our goal was to identify TFs binding with Myc to the DNA by the second and third mechanisms, in order to better understand the process of Myc dependent gene regulation using DNase footprinting data in the 3T9^{MycER} and E μ -myc systems.

We first integrated data from DNase-seq experiments with other NGS data, such as CHIP-seq and RNA-seq, in the 3T9^{MycER} and E μ -myc systems in order understand the characteristics of the Myc bound promoters in terms of histone and DNase I signatures. Next, we carried out motif analysis on subsets of Myc bound promoters defined by RNA-seq expression data as up, down or no-deg genes. As Myc is known to bind to a specific (“canonical”) form of the E-box, namely CACGTG with high affinity and to the non-

canonical (CANNTG) forms with lower affinity (Walhout et al., 1997), the presence of the E-box in our sequences, therefore, serves as a control for our analysis. If Myc binds to the DNA through the second mode described above, we should find the motif of another factor very close to the E-box motif. In this scenario the ‘composite motif’ set (see Introduction) can be very useful: if we find the composite motif of E-box with another TF, under the footprints (close to the summit of the Myc peak), this would strongly indicate the presence of a binding complex. Instead, in the case of a ‘piggy backing’ scenario we would instead expect to find the binding motif of another TF and no E-box under the footprint (close to the summit of the Myc peak).

3.3.2 Materials and methods

For the integrative analysis we used ChIP-seq samples for the transcription factors Myc, Ctf, Miz1, Pol II, histone marks H3k4me3, H3k4me1 and H3k27ac, and RNA-seq samples obtained in both systems in 3T9^{MycER} (0h and 4h) and E μ -myc (C, P and T) systems (Sabò et al., 2014). Qualitative heatmaps for studying the overlaps of DHSs with ChIP-seq peaks were generated using the ‘compEpitools’(Kishore et al., 2015) R package. Enrichment of the ChIP-seq samples were calculated using the ‘GRenrichment’ function in the ‘compEpitools’.

For motif analysis, we compared MEME (parallel version (Machanick and Bailey, 2011), CUDA-MEME (Liu et al., 2010), STEME (Reid and Wernisch, 2014) and the command-line version of DREME (Bailey, 2011). CUDA-MEME is a version of MEME (see chapter 3.2) that harness the power of the highly parallelized CUDA computing platform using

graphics processing unit to increase the computing speeds (Nickolls et al., 2008). STEME applies suffix trees, a data structure for efficiently storing and indexing a set of sequences (strings) to the Expectation Maximization (EM) algorithm, in order to increase the computation speed. DREME on the other hand, uses a non-probabilistic regular expression search and is optimized to search short motif sequences (4-8 nt long). For motif over-representation analysis we used the command line version of the Pscan tool (see Chapter 3.2).

3.3.3 Results

i. Myc binds to already open DHSs

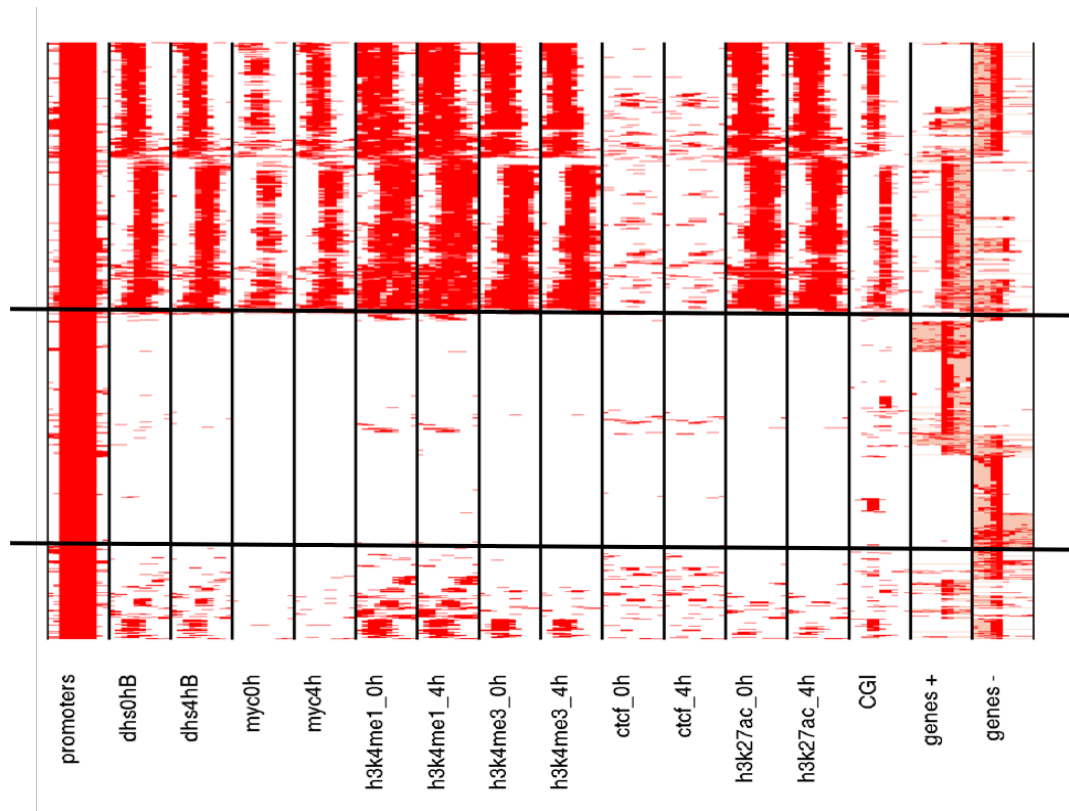
We visualized the overlap of DHSs from high depth DNase-seq experiments (Figure 19) with CHIP-seq peaks of Myc, Ctf, Miz1 and DNA Polymerase II (Pol II) and histone marks in the 3T9^{MycER} and E μ -myc systems respectively using heatmaps on the promoters in chromosome 1. In both systems, more than 70% of the Myc peaks overlap with a DHS, indicating that Myc binds to regions of DNA that are already open and does not act as a pioneer transcription factor, in agreement with previous studies (Soufi et al., 2012, 2015). In addition, more than 95% of the DHSs overlap with an H3K4me3, H3K27ac and a Pol II peak, indicating that these Myc-bound peaks are actively transcribed regions.

In both heatmaps, we see three distinct clusters of peaks/DHSs. The first cluster contains Myc bound regions with high H3K4me1, H3K27ac, H3K4me3 and CpG content that also overlap highly with DHSs indicating active promoter regions. The second cluster contains mostly regions with no Myc, DHSs, H3K4me1, H3K27ac, H3K4me3 and CpG indicative of

closed and inactive regions. Finally, the third cluster contains regions that are not bound by Myc but contain DHSs, high levels of H3K4me1, and low levels of H3K4me3 and H3K27ac indicative of open but inactive promoters. Thus, the heatmaps show that the DHSs change very little between the time-points in 3T9^{MycER} or between conditions in E μ -myc and both systems from similar clusters in the heatmaps.

a 3T9^{MycER}

Chr 1



b Eμ-myc

Chr 1

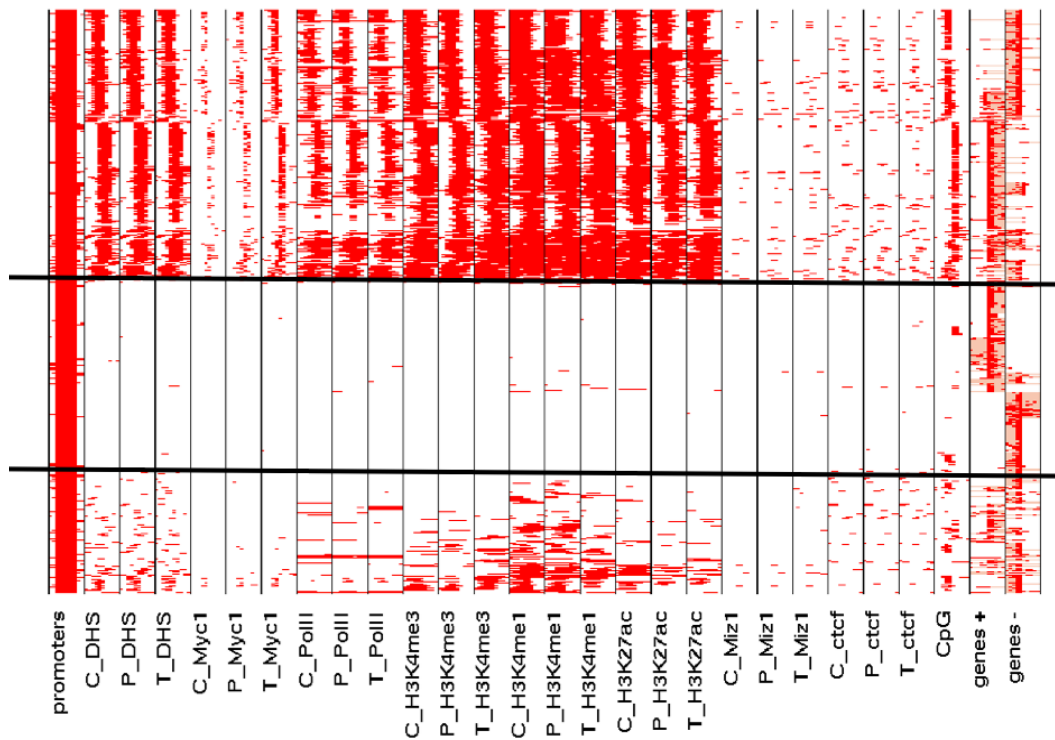


Figure 19. Heatmap showing the presence or absence of peaks corresponding to all the promoter regions (-2kb, +1kb from the TSS) in chromosome 1 for histone marks (H3K4me1, H3K4me3 and H3K27ac), transcription factor (Myc and Ctf), and Pol II ChIP-seqs and DNase-seq of (a) Eμ-myc (C, P and T) and (b) 3T9^{MycER} at 0h and 4h. The heatmap also shows regions corresponding to CpG islands (CGI) and genes (+ and - strand).

ii. E-box footprints are found close to the summit of the Myc peak

Many E-boxes are present on the Myc bound promoters and under a Myc peak but they are not all recognized and bound by Myc: the sites recognized by a TF are usually close to the summit of its ChIP-seq peak. Therefore, if we carry out motif analysis on the complete ChIP-seq peak the results can be often confusing, as the motif analysis tools may pick up several motifs which are not real binding sites of the TF. To limit this effect, it is better to choose a smaller region around the summit of the ChIP-seq peak. To choose the best length of the peak for motif analysis we used FIMO to search for all e-boxes under the summit of Myc

peaks using 200bp (-100/+100 from summit) and 400bp (-200/+200 from summit) and the full peak (Figure 20). Although we identified many E-boxes on the original width of the Myc peaks, only a smaller proportion of these are footprinted. In comparison 200bp and 400bp regions have a higher percentage of footprinted E-boxes indicating that although a promoter can contain multiple E-boxes, the E-box that Myc identifies is most likely to be very close to the summit. Hence, in the absence of DNase-footprinting data, it would be more effective to carry out downstream analysis (motif analysis) on a smaller region around the summit such as 400bp or 200bp region. However, the 200bp regions also contain fewer E-boxes. Therefore, in order to be not be too stringent, we used a region of 400bp (+200/-200bp around the summit) for motif analysis.

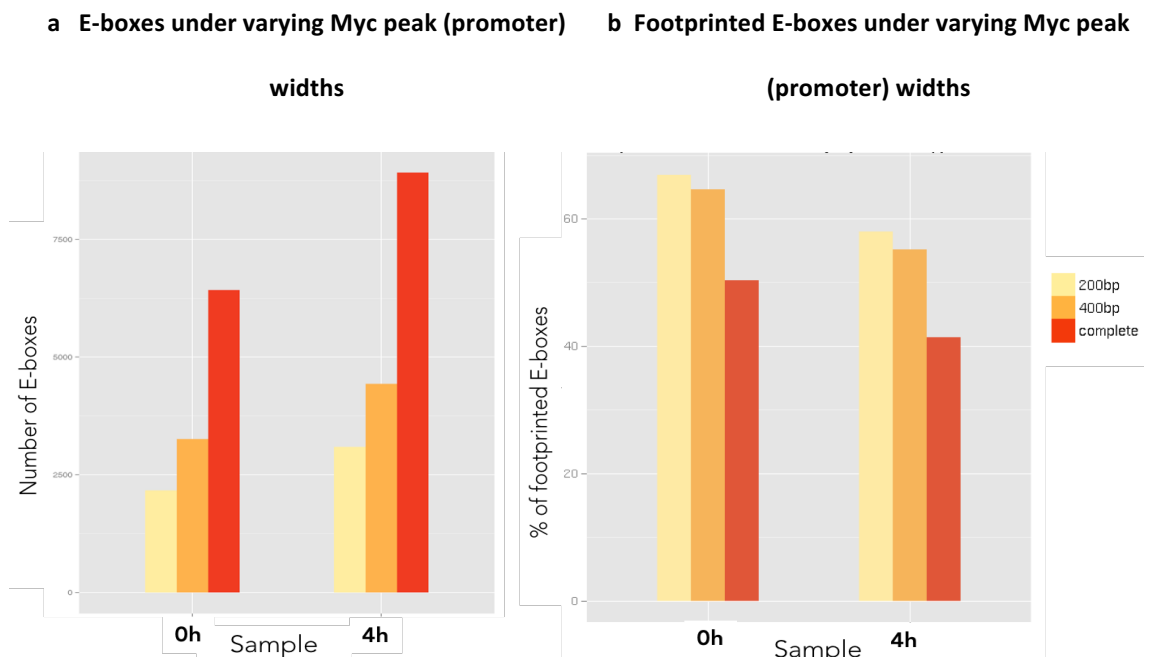


Figure 20. (a) Number of E-boxes identified by FIMO (cut-off 10^{-5}) on promoters of Myc peaks under the summit using varying peak widths (200bp, 400bp) and complete peak in 3T9^{MycER} at 0h and 4h time-points. The full length peaks contain more E-boxes compared to the 200bp and 400 bp regions. (b) Percentage of E-boxes that are footprinted under the summit of the Myc peak using varying peak widths (200bp, 400bp) and complete peak in 3T9^{MycER} at 0h and 4h time-points.

A higher percentage of E-boxes are footprinted in the 200bp regions compared to the 400bp and the full peak.

iii. Up-regulated genes contain more footprinted E-boxes compared to not-deregulated and down-regulated genes

We identified the footprints in the proximity of the promoters of the Myc-bound up, down and no-deg genes in the two systems. The number of footprints corresponding to the canonical E-box is considerably lower in down genes compared to up and no-deg genes (Figure 21 a and b).

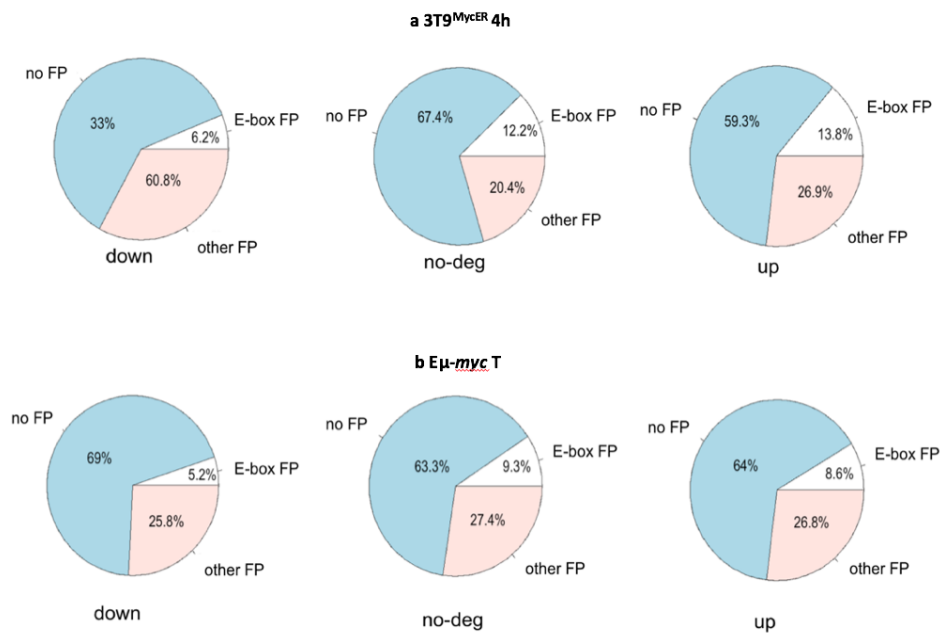


Figure 21. Distribution of the up, down and no-deg genes into 3 categories based on the footprint data (4h): (1) footprint of the canonical E-box binding proteins, (2) footprint of other proteins and (3) no footprints. (a) In 3T9^{MycER} (4h/0h) the percentage of footprinted canonical E-boxes is similar in the up, and no-degs and least in the down genes. (b) In E μ -myc (T/C) the percentage of footprinted canonical E-boxes is highest in the up and no-degs followed by the down genes.

This indicates that the interaction of Myc with the DNA is more frequently direct in up genes as compared to the down genes. It is possible that up-regulation of genes is caused by direct interaction of Myc with other transcription factors while down-regulation could be caused by secondary effects or Myc binding indirectly to the DNA through another transcription factor. Myc is known to bind with the transcription factor Miz1 to down-regulate gene-expression but in our data we could not find the Miz1 motif among the enriched motifs in the down-regulated genes.

iv. Myc peaks with footprinted E-boxes have higher signal enrichment than those with other footprints or no footprints.

We compared the enrichment of the Myc peaks with an E-box motif overlapping a footprint overlapping (E-box FP) under the summit (+200/-200bp), the enrichment of Myc peaks with an overlapping footprint but no E-box (other FP) and the enrichment of Myc peaks without any overlapping footprints (Figure 22). For both promoter and distal Myc peaks we see a similar trend in the two systems. The enrichment of peaks is highest in the peaks with E-box footprint, followed by the peaks with other footprints and finally the peaks without any footprints. This indicates that Myc binding is strongest in peaks with an E-box and supports the finding that Myc binds directly. On the other hand, lower binding intensity in peaks without an E-box could possibly indicate an indirect binding scenario, where Myc is associated to the promoters of the target genes through another transcription factor. Finally, the lowest binding intensities in peaks without any footprints could indicate either not enough resolution to see a footprint or a specific binding of Myc.

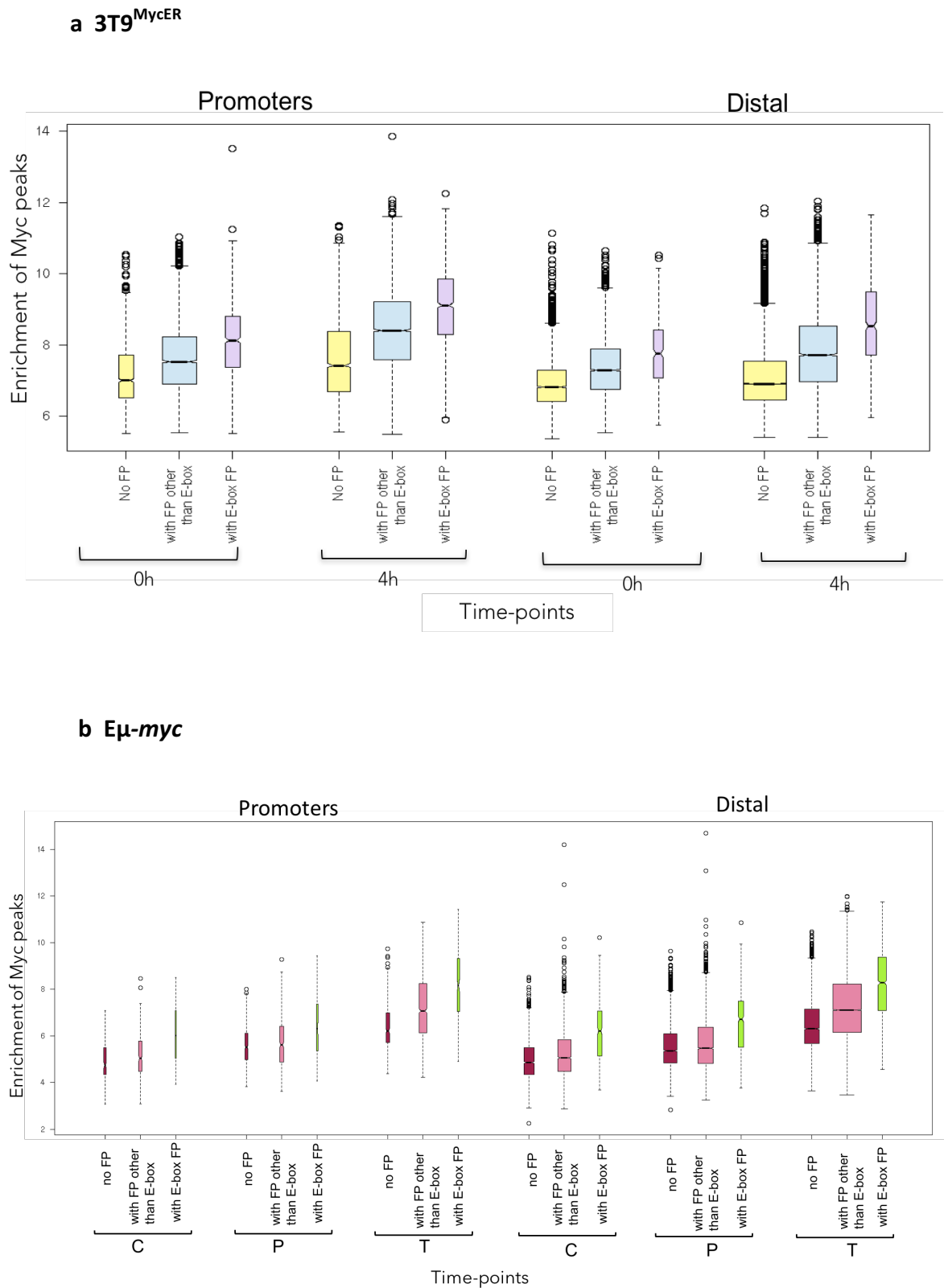


Figure 22. Boxplots showing the enrichment of promoter and distal Myc peaks divided into three subgroups, those with no footprint (no FP), with a footprint that does not have the E-box sequence (with FP other than E-box) and finally the group of Myc peaks that have a footprint that has an E-

box (with E-box FP) at 0h and 4h in 3T9^{MycER} and in the C, P and T conditions of E μ -myc. The variable widths of the boxes indicate the number of peaks in the category.

Following this, we also compared the enrichment of the Myc peaks in 3T9^{MycER} at 4h on the Myc-bound up, down and no-deg genes (Figure 23). The up and down genes were further divided into sub categories based on their log₂ fold change ratio as: (i) less than 0.5, (ii) greater than or equal to 0.5 and less than 1 (iii) greater than or equal to 1. We see that the upregulated genes have highest enrichment of Myc peaks while the downregulated genes have the lowest enrichment. Moreover, higher fold change did not imply higher Myc binding intensities, instead the binding intensities were highest for the group of Myc peaks binding to DEGs (both up and down) with a fold change less than 0.5. These results are in line with observation that the upregulated genes have more both (canonical and non-canonical) e-boxes compared to the no-deg and downregulated genes (Figure 21).

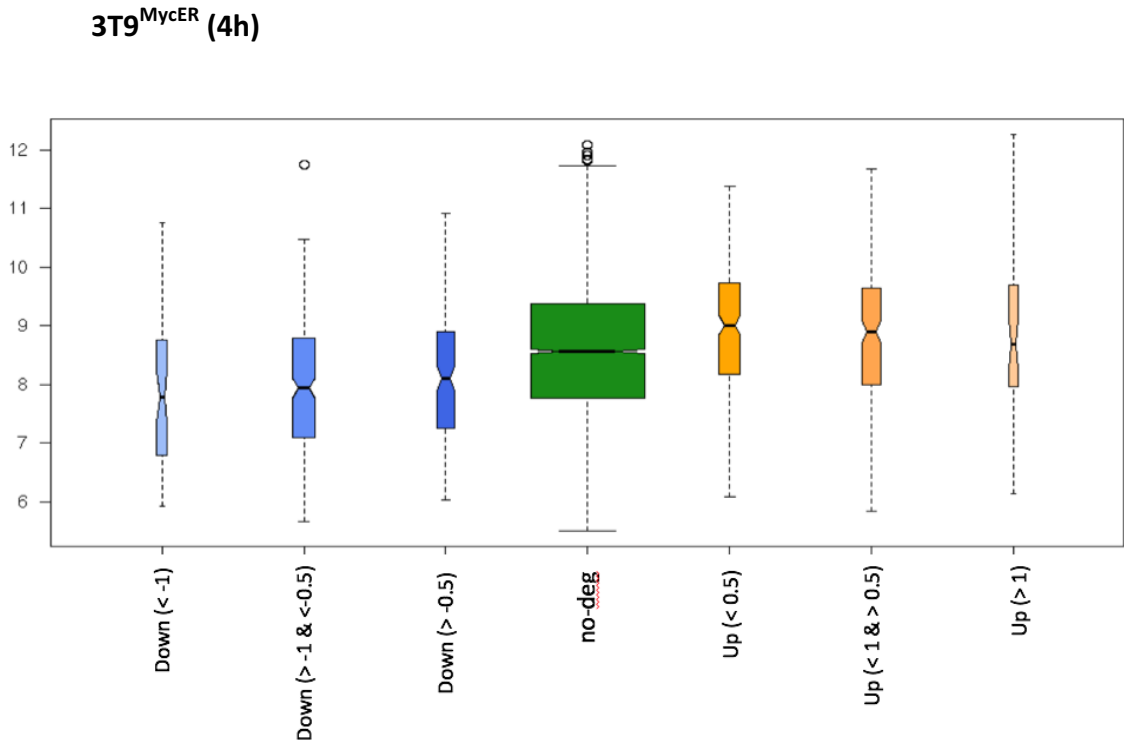


Figure 23. Boxplot showing the enrichment of Myc peaks in 4h 3T9^{MycER} samples binding to upregulated and downregulated genes (divided into three categories by their log₂ fold change ratio values: less than 0.5, higher than or equal to 0.5 and less than 1, higher than or equal to 1).

v. The E-box motif is one of the most enriched motif in de novo motif analysis

To identify possible binding partners of Myc in the 3T9^{MycER} (4h) and E μ -myc (P and T) systems we carried out motif analysis on the Myc bound DHS regions. We integrated the data from high depth DNase I seq, CHIP-seq and RNA-seq experiments to create functional subsets consisting of Myc-bound up, down-regulated and no-deg (not deregulated) genes in the two systems. We divided the list by gene expression to identify if Myc interacts with specific binding partners to drive up- and downregulation of genes. Therefore, we selected the set of sequences under the footprints that overlap with a Myc peak (-200/+200bp around the summit) corresponding to the up, down and no-deg genes (-2000/+1000 from TSS) for *de novo* motif analysis.

We used two widely used *de novo* motif analysis tools MEME and DREME. One of the main limiting factors of *de novo* motif analysis is its extremely long run times, which can be affected by three different factors: the length of the sequences, the total number of sequences submitted and the motif size. MEME especially can take a very long time to run depending on the factors mentioned above. Therefore, several new versions of MEME have been proposed to improve the run times, namely, parallel-MEME, cuda-MEME and STEMME with DREME which is known to be relatively faster but can only search for short motifs (Reid and Wernisch, 2014). We compared the run times of all the four tools on sets of 1000, 2000, 3000, 5000 sequences corresponding to the top scoring peaks of the 3T9^{MycER} 0h low-depth DNaseI-seq sample (see 3.3.2).

Our results (Figure 24), show that DREME has the shortest run-time and it is able to detect all the motifs or parts of the complete motifs detected by MEME on the same set of sequences. MEME run times become too long (>24h) when used on >5000 sequences. STEME run-times are extremely slow (>24h) even with 2000 sequences and hence we did not consider it for further analysis. Based on these results we chose DREME followed by TOMTOM for our further analysis.

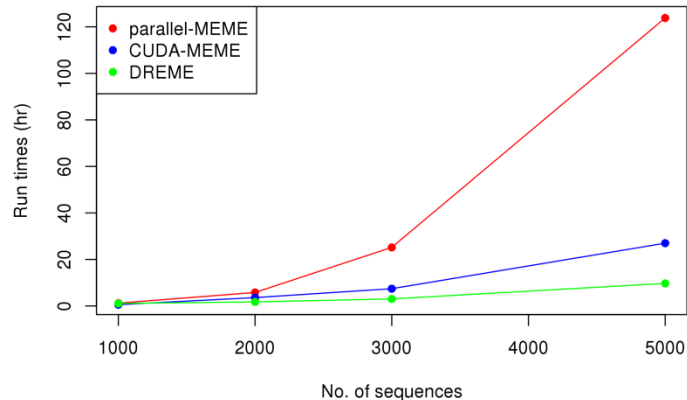


Figure 24. Run times of parallel MEME, CUDA-MEME and DREME. DREME is the fastest of the three although it can only be used for short motifs. CUDA-MEME is considerably faster than its parallel counterpart while offering the same options.

We ran DREME+TOMTOM on the subsets of sequences described above. As expected, the E-box (which is our positive control) was enriched in all the Myc-bound up and down-regulated subsets (Figure 25, below). In the 3T9^{MycER} 4h and E μ -myc P, the canonical E-box was found on the upregulated genes while the non-canonical E-box was found in the downregulated ones. In contrast in E μ -myc T, both the canonical and non-canonical E-boxes were found in the up- and the downregulated genes. Most of the other motifs we identified such as NRF1, ETS family and POU family factors, are found in both the up and down lists

and therefore are not unique to any specific list. This is because the DREME searches for enriched motifs in the single set of sequences compared to the background but these motifs might not be specific to that set only. Hence, it is not easy to identify motifs specific to a particular set of sequences compared to another set. This analysis therefore, needs to be followed by a motif over-representation analysis which identifies the motifs that are specific to a given set of sequences compared to another. However, this DREME+TOMOTM motif analysis can be used as a quick way to validate the data of transcription factor CHIP-seq experiments. The binding motif of the CHIP-ed transcription factor should be one of the most enriched motifs identified using *de novo* motif analysis on the CHIP-seq peaks.

a 3T9^{MycER} up

Motif logo	Transcription factor matched by TOMTOM
	Dbx1, Tbp, Sox14, Tlx2
	HDAC1, PATZ1, EGR4
	NFY{A,B,C}, PBX3, FOXI1, CEBPZ
	SIX5, ZNF143
	SP{1,2,4}, KLF{11,5}, Zfp281
	NRF1
	IRF4, Zfp105
	YY1, YY1
	POU domain factors
	MAX, MYC, USF1, BHLH family
	ETS family factors
	GFI1

b 3T9^{MycER} down

Motif logo	Transcription factor matched by TOMTOM
	Zfp105, Dbx1, Sox7
	SIX5
	ETS family factors
	IRF4, SPI1
	SP{1,2,3,4,8}, ZNF148, ZBTB7B, GC_box, KLF{4,5,7,11,14,16}
	IRF{1,2,8}, PRDM1
	YY1, YY1
	JDP2, TFEB, CREB{3,5}, GDNF, E4F1, ATF5_CREB3, CREB3L1, JUN, ARNTL, ATF{1,7}, XBP1
	POU domain factors
	BHLH_family, MITF, MYCL2, MAX, MYC::MAX,
	SNAI{1,2}, HTF4, MYF6, NHLH1
	TFAP4
	CEBPZ, NFY{A,B,C}, FOXI1
	NRF1

c Eμ-myc P up

Motif logo	Transcription factor matched by TOMTOM
	Yy2
	Cebpz, Nfyb
	Klf13, Klf14, Sp4, Klf12
	Ahr, Arnt, Arnt2
	BHLH_family, Myc::Max, Usf1, Max
	NRF1
	Spi1, Spic, Spib, Gabpa, Etv6
	Sp1, KLF4, Sp3, Maz, Patz1, Znf740, Wt1, Klf16,

d Eμ-myc P down

Motif logo	Transcription factor matched by TOMTOM
	Elf2
	Mitf, BHLH_family, Mycl2, Tcf15, Mycn, Myc::max
	Sp1, Klf7, Patz1, Klf4, Sp4, Sp3, Klf12

e. Eμ-myc T up

Motif logo	Transcription factor matched by TOMTOM
	SNAI1, HC_HTF4_f1
	POU domain factors
	YY2, YY1
	TFEB, ARNT, BHLHE41, SREBF2r, BMAL1, Srebf1
	bHLH_family, USF1, Mycn, MYC::MAX
	KLF13, AHR, ARNT
	GLIS1
	Klf {4,7,12,16}, SP{1,3,4}, WT1, PATZ1, ZNF148
	HC_EGR4_f1
	Ets family factors
	NFY{A,B,C}, CEBPZ, FOXI1, PBX3
	NRF1
	MAZ, EP300, MZF1
	SPDEF

f. Eμ-myc T down

Motif logo	Transcription factor matched by TOMTOM
	IRF{1,3,4,8,9}
	YY1
	ZNF238
	ETS family factors
	SPIB ,TEAD1
	POU domain factors
	MITF, bHLH_family, MYCL2, TCF15, , Mycn, MYC::MAX
	bHLH_family, USF1, Mycn, MYC::MAX
	Sp1, SP3, Klf4
	NHLH1, MYF family
	CREB3, JDP2, ATF5_CREB3, JUN
	CEBPZ, NFY{A,B,C}, FOXI1
	NRF1
	MAZ
	KLF{12,14,16}, SP{3,4,8}
	MEF2{A,B,C,D}
	RUNX{1,2}

Figure 25. The list of TFs whose binding sites were identified by DREME in the (A) 3T9^{MycER} up-regulated (B) 3T9^{MycER} down-regulated (C) Eμ-myc T up-regulated (D) Eμ-myc T down-regulated. In 3T9^{MycER} the non-canonical E-box (CANNTG, highlighted in green) was identified in the down-regulated lists while the canonical E-box (CACGTG, highlighted in orange) was identified in the up-regulated lists. In Eμ-myc, the E-box was identified in both the up- and downregulated lists.

vi. The E-box motif is over-represented in the up-regulated genes

In order to identify motifs that are unique to a particular set, we carried out motif over-representation analysis using the Pscan (Zambelli et al., 2009) tool. However, this tool did not accept sequences less than 100 bases in length so we could not use it on the footprint sequences (which are ~35-50bp in length). We resorted to Pscan to search for over-represented motifs in the region surrounding the summit of the Myc peaks (-200/+200) corresponding to up- and downregulated genes (positive set) against the no-deg genes (background or negative set) in the 3T9^{MycER} and E μ -myc (both in P and T) systems.

From the Pscan results (Figure 26) we see that the E-box motif is enriched in the upregulated subset but not in the downregulated subset, when compared to the no-deg subset (background) in all the 3 samples (3T9^{MycER} 4h, E μ -myc P and T). These results could explain our previous observation (Figure 23) that the enrichment of Myc peaks is higher in the upregulated genes compared to the no-deg and downregulated genes where the higher number of E-boxes could lead to stronger Myc binding.

Most of the motifs we found to be enriched in the upregulated list are CG rich motifs like the Sp1, and the core promoter motifs BRE and MTE sequences (Figure 26 a, c and e) while in the downregulated list we find more AT rich motifs like core promoter motifs TATA box and Inr (Figure 26 b, d and f). Sp1 is a constitutive transcription activator of housekeeping genes and other TATA-less genes (Vizcaino et al., 2015) and hence it is not surprising that it is found to be enriched in the up-regulated genes. The analysis also revealed some other interesting motifs such as Bcl6b that is overrepresented in all the downregulated lists. Bcl6b (also known as Bazf) is a transcriptional repressor that works in association with Bcl6 and is involved in early B-cell development (Takenaga et al., 2003) and key role in many cancers

(Li et al., 2015; Wang et al., 2015). However, since this analysis was carried out on the Myc peaks and not on the footprints, it is difficult to say whether the motifs we identified from this analysis were actually bound by the corresponding transcription factors.

Although the motif over-representation analysis once again confirmed that the upregulated subset of genes has more E-box motifs compared to the no-deg and downregulated we could not define a conclusive list of binding partners of Myc involved in gene-regulation in these systems. Therefore, we need an approach that would allow use to the footprints data to identify motifs that are specific to up and downregulated Myc bound genes in our systems.

a 3T9^{MycER} up

Motif logo	Enriched motifs
	BRE sequence
	MTE
	Ascl2
	E-box
	E2F2, E2F3
	Egr1, AHR, Arnt::Ahr, Pax2
	Gmeb1, Gmeb2
	Hnf4a
	Mtf1
	Myf6, Nhlh1
	Plag1

b 3T9^{MycER} down

Motif logo	Enriched motifs
	Inr
	TATA_box
	Arid3a, Hoxa3
	Arid5a
	Bbx
	Bcl6b
	Eomes, Brac, Tbx5, Mga, Tbx15, Tbx19, Tbx1
	Esrra, Err1, Err2, Err3, Nr5a2, Nr6A1, Stf1, Rora, Esrrb
	Foxa2, Foxa1
	Foxj1

c Eμ-myc P up

Motif logo	Enriched motifs
	BRE sequence
	E-box
	E2F5
	Hes1
	Hif1A, Hif1A::Arnt
	Arnt::Ahr
	E2f1
	Spdef

d Eμ-myc P down

Motif logo	Enriched motifs
	Inr sequence
	Foxa2, Foxa1
	Foxj1
	Myf6
	Six6, Six3
	Egr3
	Elf3
	Foxf1
	Myod1, Tgif1
	Onec2

e Eμ-myc T up

Motif logo	Enriched motifs
	E2f2, E2f3, E2f1, E2f4, E2f1
	E-box
	Mtf1
	Zfp161
	E2F5
	E2f7, E2f7, E2f8
	Hif1A, Hif1A::Arnt
	Tfdp1
	Zn423
	Arnt::Ahr

f Eμ-myc T down

Motif logo	Enriched motifs
	Inr sequence
	TATA_box
	Arid3a, Hoxa3
	Arid5a
	Ascl2
	Bbx
	Bcl6b
	Elf3, ElfF2, Elf5, Ets1, Etv4, Etv5, Fev, Spib, Fev, Spi1, Spic
	Eomes, Brac, Mga, T-box factors
	Foxa1, Foxa2

Figure 26. Enrichment of TF motifs in (left to right) upregulated genes vs. no-deg genes and down-regulated vs no-deg in (top-bottom) 3T9^{MycER} 4h, Eμ-myc P and Eμ-myc T samples. The canonical E-box sequence ‘CAGGTG’ is identified as an over-represented motif the upregulated vs. no-deg list but absent from the downregulated vs. no-deg list in all the 3 samples.

3.4 Single feature classification

3.4.1 Introduction

As shown in the previous chapter, using DREME and Pscan we could not obtain a conclusive list of binding partners of Myc. We therefore classified the up, down and no-deg genes with another approach that leverages on the different of types data available in our systems of interest such as enrichments of histone modification profiles, transcription factor binding signatures obtained from footprints and ChIP-seq experiments. The rationale behind this was that, if a particular transcription factor A is a binding partner of Myc in up-regulating a set of genes, we can expect the PWM of that transcription factor to be a discriminating feature of the up-regulated against the no-deg genes. We assessed the ability of each of these features in classifying the genes in pairwise combinations (up/down, up/no-deg and down/no-deg). If any of these features are enriched in a particular category, it should be possible to classify the data into the categories considered with good accuracy. We also applied this approach on data from the tet-MYC, a hepatocellular carcinoma model in which genes which are directly and indirectly deregulated by Myc have already been identified (see Introduction).

3.4.2 Materials and methods

We computed the ratios of the enrichment of ChIP-seq data on TFs and histone marks (Sabò et al., 2014) (4h compared to 0h in 3T9^{MycER}, T compared to C and P compared to C in Eμ-*myc* and tumour to normal in tet-MYC). In the 3T9^{MycER} and Eμ-*myc* systems, we extracted the region under the summit of the Myc peaks (+250/-250bp around the summit) overlapping with the promoters (within -2kb/+1kb from TSS) of the up, down and no-deg genes with an overlapping DHS. Next, we used FIMO to search on these regions the binding sites of all the 2433 PWMs present in our database (see section 3.2). Then we used DNase-footprinting

information to keep only the motifs with an overlapping footprint (allowing a gap of 15 bp on both sides). These are motifs which are most likely bound by a transcription factor. We counted the number of times each of these footprinted PWMs are found on the promoter of each gene in the up, down and no-deg lists (Table 1). These footprinted PWM counts and enrichment ratios from the ChIP-seqs were all tested for their discriminative power using the ‘roc’ function of the ‘pROC’ (Robin et al., 2011) R package. We used this function to carry out binary classifications of up/down, down/no-deg and up/no-deg classes by plotting the ROC curves for each classification. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The complete list of features tested are listed in table 2. In addition to the data listed in Table 2, we also used the presence or absence of a CpG island on the promoter of the genes as another feature.

Upregulated genes	PWM 1 (number of binding sites)	PWM 2 (number of binding sites)	PWM N (number of binding sites)
Gene 1	1	0		2
Gene 2	0	1		0
Gene 3	2	1		0
Gene 4	1	2		0
Gene 5	0	1		1

Table 1: Example showing how the transcription factor binding information is calculated for use as a feature for single feature classification.

In the tet-MYC system we did not have DNase-seq data, so we used the FIMO matches of PWMs in the -450/+50bp region around the TSS as putative transcription factor binding sites. We preferred to use the region around the summit of the Myc peaks for this system as the number of Myc ChIP-seq peaks in the Myc-off samples were very few.

Experiment	System	Type	Feature
ChIP-seq	3T9 ^{MycER}	Histone mark enrichment ratios	H3K27ac, H3k4me3, H3K36me2, H4K16ac, H3K36me3, H3K4me1, H3K27me3
ChIP-seq	tet-MYC	Histone mark enrichment ratios	H3K4me3, H3K4me1, H3K27ac, H3K79me2
ChIP-seq	Eμ- <i>myc</i>	Histone mark enrichment ratios	H3K27ac, H3k4me3, H3K4me1
ChIP-seq	3T9 ^{MycER}	Transcription factor enrichment ratios	Myc, Ctf and Miz1
ChIP-seq	Eμ- <i>myc</i>	Transcription factor enrichment ratios	Myc, Ctf and Rad21
ChIP-seq	tet-MYC	Transcription factor enrichment ratios	Myc and Ctf
ChIP-seq	All systems	Transcription mark enrichment ratios	Pol2
DNase-seq	3T9 ^{MycER} and Eμ- <i>myc</i>	Enrichment ratio	-
DNase-footprinting	3T9 ^{MycER} and Eμ- <i>myc</i>	Binding sites identified by FIMO + footprints under Myc peak summit	2433 PWMs

FIMO analysis	tet-MYC	Binding sites identified by FIMO only (no footprints) around the TSS	2433 PWMs
---------------	---------	---	-----------

Table 2: List of data used in the single feature classification approach.

3.4.3 Results

We calculated the performance of each feature in separating the data between pairwise combinations: up vs. no-deg, down vs. no-deg and up vs. down, and computed the AUCs corresponding to the ROCs of each of these classifiers. The overall top scoring feature and the top scoring PWM class for all the models are listed in table 3.

In 3T9^{MycER}, the maximum AUC corresponds to the enrichment ratio of the histone mark H3K27ac in all the 3 models. This is expected since H3K27ac is an activation mark (Tie et al., 2009) and therefore, more likely to be enriched in active genes than repressed genes. In the Eμ-*myc* system, the top scoring classifier was the RNA polymerase II ratio with an AUC higher than 0.8 in all the classifications (both for the T/C and P/C comparisons), suggesting that the Pol II signal alone can distinguish the up, down and no-deg gene categories with a high accuracy. In tet-MYC, the most predictive feature for Myc-dependent up vs. no-deg was the RNA polymerase II enrichment ratio (AUC 0.87) and for Myc-dependent down vs. no-deg was H3K4me3 (AUC of 0.74). The H3K4me3 is another mark that is associated with active promoters and hence can be expected to have different distributions in up, down and no-deg genes. For the Myc-independent up vs. no-deg and Myc-independent down vs. no-deg classifications, none of the features reached an AUC higher than 0.6. In all the other

classification too only a few features had an AUC of more than 0.7. All other features had much lower AUCs (less than 0.7) and therefore were not predictive. Among the PWMs, none were able to classify the data with an AUC higher than 0.6 in any of the systems. Thus, except the obvious features such as H3K27ac and Pol II none of the other features had enough predictive power to classify the data. On the other hand, this predictive power could arise from a combination of features, which together could classify the data better.

Model	Overall top AUC (single feature)	Top AUC among PWMs only (single feature)
Up/Down (3T9 ^{MycER} 4h)	0.87 H3K27ac	0.52 SP4
Up/No-deg (3T9 ^{MycER} 4h)	0.76 H3K27ac	0.507 Klf 11
Down/No-deg (3T9 ^{MycER} 4h)	0.73 H3K27ac	0.52 Sp4
Up/Down (E μ -myc T)	0.94 Pol II	0.53 SP4
Up/No-deg (E μ -myc T)	0.84 Pol II	0.51 NFYA
Down/No-deg (E μ -myc T)	0.89 Pol II	0.52 SP4
Up/Down (E μ -myc P)	0.90 Poll II	0.53 TFAP2C_HES7
Up/No-deg (E μ -myc P)	0.75 Poll II	0.52 SP4
Down/No-deg (E μ -myc P)	0.73 Poll II	0.50 TFAP2C_HES7

Myc-indep Up/No-deg (tet-MYC)	0.59 H3K79me2	0.58 E2F1_ELK1
Myc-indep Down/No-deg (tet-MYC)	0.59 H3K4me3	0.58 E2f1_ELk1

Table 3: Table showing the highest AUC of the ROCs obtained using the single feature classification approach. The histone mark H3K27ac and Pol II features were the best discriminative features overall. Among the PWMs the discrimination power was very low with the AUC in the range of 0.5 to 0.53.

3.5 Random forest classification

3.5.1 Introduction

Our objective was to identify features that can successfully separate pairs of classes of genes in the three systems under study. In particular, looking at the PWMs that are more enriched in a gene set will help us to pinpoint transcription factors that could be possible binding partners of Myc. We saw in chapter 3.4 that single PWMs could only classify the data with limited accuracy (see 3.4). It is however possible that the classification will improve if we take into account combinations of features. We explore this possibility with a machine learning algorithm called Random Forest (RF), which combines many features together to create a robust classifier and provides a ranking of the importance of the features in the classification.

Machine learning approaches give computers the ability to learn associations without being explicitly programmed (Sherwood et al., 2014) and are therefore well suited for automatically identifying complex patterns in large datasets. They can be divided into two categories: unsupervised and supervised. Unsupervised methods are similar to pattern discovery, where the machine tries to learn some patterns or rules in a given dataset without any prior knowledge or training set. On the other hand, supervised learning methods require a training set from which they can learn the rules for classifying the data. A trained machine can then be applied to a new dataset for which the rules are not known. Among the different types of supervised machine learning algorithms, Ensemble methods combine several features to give a more robust classification compared to single feature based methods. Random forest (Breiman and Leo, 2001) is an Ensemble method that learns binary classifications of data using an ensemble of binary classification trees, and provides a

measure that ranks the features used for the classification. This measure can be used to filter out uninformative features and keep only those that classify the data best.

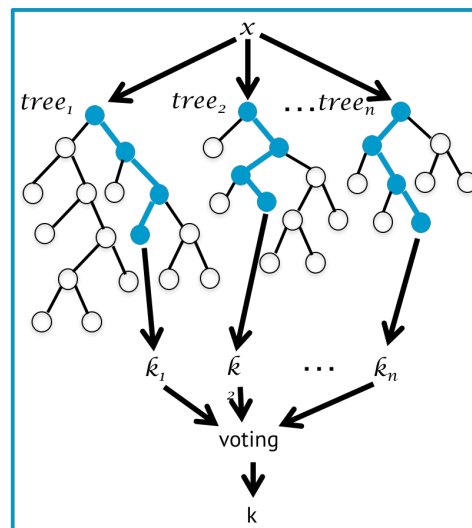
We used the ‘cforest’ function of the ‘party’ package that applies a random forest algorithm to classify the genes on the same features that were also previously employed in the single-feature classifications. One of the main advantages of using Random forest algorithms is that they can handle a large number of variables of different types with many missing values and therefore well suited for our problem (we have 2433 PWMs, in addition to other features). These algorithms use a combination of several decision trees (Figure 27) or ensemble of trees to classify the data. The basic unit, or the base learner, of a random forest is a binary tree constructed using recursive partitioning typically grown using the CART (Classification and Regression Trees) method (Breiman, 1984), where a tree is recursively partitioned by binary splits into homogeneous or near homogeneous terminal nodes. Homogeneity is defined by the Gini index that provides an indication of how “pure” the nodes are. If a dataset T contains examples from n classes, gini index,

$$G = \sum_{k=1}^M (P_k * (1 - P_k))$$

Where, G is the Gini index, P_k is the frequency of label k at a node and M is the number of unique labels. A node that has all classes of the same type (perfect class purity) will have $G=0$, whereas a node that has a 50-50 split of classes (worst purity) will have a $G=0.5$. A good binary split pushes data from a parent tree-node to its two daughter nodes so that the resulting purity of the daughter nodes is higher than that of the parent node. The tree is divided recursively until no further improvement of homogeneity can be made. RFs are often made of a collection of hundreds to thousands of such trees and each tree is grown using a

bootstrap sample of the original data and about one-third of the cases are left out of the bootstrap (out of bag or OOB data) in the construction of the k^{th} tree. Each case of this OOB data is used as a test set to get a classification from the k^{th} tree. In this way, a classification is obtained for each case in about one-third of the trees. At the end of the run, the class with the most votes is chosen as the final class j of the case n every time it was ‘OOB’. The proportion of times that j is not equal to the true class of n averaged over all cases gives the OOB error estimate. OOB data are also used to estimate importance of variables.

RF trees also add an additional step of randomization: instead of splitting a tree node using all variables, at each node of each tree, a random subset of variables is selected and only these variables are used as candidates to find the best split for the node. This two-step randomization de-correlates trees lowering the variance of the forest ensemble.



Adapted from Nguyen et al., 2013

Figure 27. An example of a Random forest. Random forests are a type of ensemble method which consists of many trees to classify the data.

We classified our datasets using a random forest algorithm in the 3 systems of interest and then validated some of the top features identified by approach in the $E\mu$ -myc system using ChIP-seq data from the ENCODE repository in the CH12 cell line.

3.5.2 Materials and methods

The random forest implementation of the party package ‘cforest’ function from the ‘party’ R package was used to carry out pairwise classifications of our datasets (as in Chapter 3.4). We used the same features described in Chapter 3.4 for single feature classification in $3T9^{MycER}$, $E\mu$ -myc and tet-MYC systems. The ROCs of the classifications were generated using the ‘ROCR’ R package. A $k=10$ -fold cross validation was used to measure the performance of the models. In this method, the dataset is divided into 10 subsets and each time one of the k subsets is used as the test set and the other 9 subsets are combined to form a training set. Finally, the average error across all k trials is computed.

The variable importance (VI) function was used to compute the importance of a feature in the classification. We used the standard version of the ‘varimp’ function in the ‘party’ package to calculate the VI. With this method, each predictor variable is randomly permuted to break its original association with the response Y . When this permuted variable X_j is used together with the remaining non-permuted predictor variables to predict the response for the OOB observations, the prediction accuracy should decrease substantially if the original variable X_j was associated with the response. The difference in prediction accuracy before and after permuting X_j , averaged over all trees, gives the variable importance.

We used the random forest approach to classify the same pairwise combinations (up/down, up/no-deg, down/no-deg) as in single-feature classification. Features that would rank highest in these classifications can be regarded as distinguishing features of the classes considered.

Next, we calculated the variable importance of all the features from the model with the best AUC obtained using $k=10$ cross-validation. Only the features that had a positive variable importance were selected to obtain the final list of the top-most predictive features. After we obtained the lists of the top-most predictive features, we calculated their enrichment in the positive set compared to the negative set (for example, in the up/no-deg, down/no-deg classification, up and down are the positive sets and no-deg is the negative set).

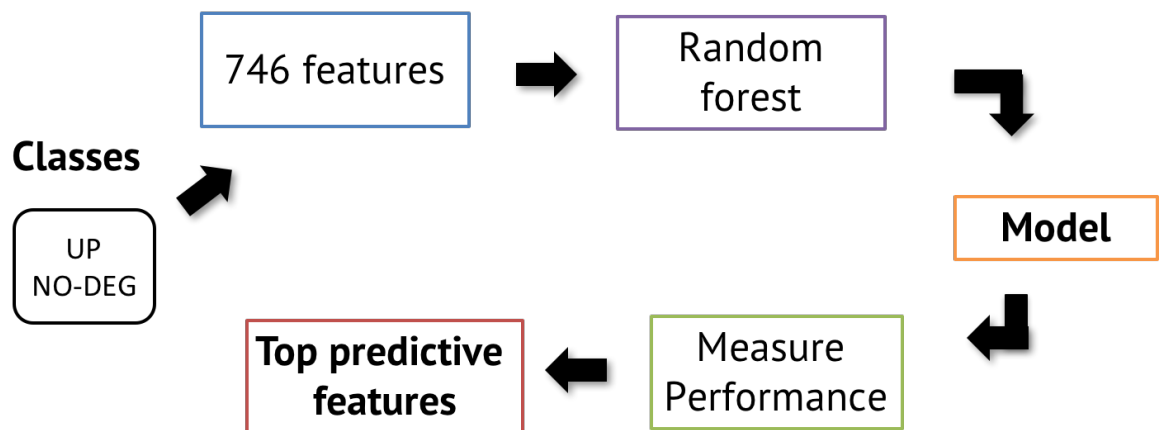


Figure 28. Main steps in the machine learning based classification of the up, down and no-deg genes using different features obtained from NGS data. A label is attached to each gene based on the RNA-seq data as up, down or no-deg. A random forest algorithm is used to classify the data into up/down, up/no-deg and down/no-deg categories using the different features such as ChIP-seq enrichment ratios and PWM counts on the gene. Performance of the data is measured by a k -fold ($k=10$) cross-validation method. Finally, the features with the the top-most variable importance in the model with the best AUC are used as the list of best predictive features in that classification.

3.5.3 Results

We calculated the AUCs under the ROCs of each of the classifications obtained using random forest algorithm. Table 4 shows a comparison of these new AUCs with the highest AUCs obtained using the single feature approach (chapter 3.4). In the 3T9^{MycER} system, the AUCs improved for all the three classifications (up/no-deg, down/no-deg and up/down). The highest improvement of AUC was observed for down vs. no-deg classification, where it increased from 0.73 to 0.79. Similar increases of AUCs were also observed for the other two classifications in 3T9^{MycER} (table 4). Instead, in the E μ -myc system, the overall AUCs decreased in all the three models by about 3-4% compared to the top AUC obtained by the Pol II enrichment ratio in single feature approach. As random forest algorithms only use a subset of features at each node for dividing the tree, it is likely that the Pol II feature is not considered at every node. Still the best AUCs obtained using 10-fold cross-validation of ‘cforest’ for all the E μ -myc classifications were higher or equal to 0.79.

Model	Average AUC using all features with single feature classification	Best AUC using all features with cforest
Up/Down (3T9 ^{MycER})	0.88 (H3K27ac 4h/0h ratio)	0.91
Up/No-deg (3T9 ^{MycER})	0.76 (H3K27ac 4h/0h ratio)	0.84
Down/No-deg (3T9 ^{MycER})	0.72 (H3K27ac 4h/0h ratio)	0.79
Up/Down (E μ -myc)	0.95 (Pol II T/C ratios)	0.92
Up/No-deg (E μ -myc)	0.84 (Pol II T/C ratios)	0.80
Down/No-deg (E μ -myc)	0.83 (Pol II T/C ratios)	0.79

Myc-dep Up/No-deg (tet-MYC)	0.87 (Pol II tumour/control ratio)	0.87
Myc-dep Down/No-deg (tet-MYC)	0.74 (H3K4me3 tumour/control ratio)	0.76
Myc-indep Up/No-deg (tet-MYC)	0.59 (H3K79me2 tumour/control ratio)	0.67
Myc-indep Down/No-deg (tet-MYC)	0.59 (H3K4me3 tumour/control ratio)	0.66

Table 4: Table showing the comparison of the best AUC obtained with single feature classification and the AUC obtained using random forests. A significant improvement is seen for all the models.

The major advantages of using the RF method are not limited to improvements of the AUCs but extend to the variable importance function, that provides an estimate of the importance of the features in the classification. The Figure 29 shows the top features obtained using this approach ranked by their variable importance. The acetylation marks H4K16ac (Taylor et al., 2013), H3K27ac (Tie et al., 2009) and the transcription mark Pol II were among the top features in all the systems based on the variable importance function (Figure 29). Since these are activation marks, they are expected to be more enriched in actively transcribed genes, consequently in our predictions we find them to be more enriched in the up-regulated genes.

In contrast to the results of the single feature approach, we found several PWMs among the list of top predictive features from the random forest method. In all the systems, for the up vs. no-deg classification, the footprints corresponding to the canonical E-box were among the top scoring features, once again confirming our previous observation that the canonical E-box is more enriched in the up genes compared to down genes.

In addition to the E-box (MAX motif), in the up/no-deg classification of 3T9^{MycER}, we also found the composite motif CLOCK_BHLHA15, of the type ‘canonical-E-box_non-canonical-E-box’ (CACGTG_CANNTG), indicating the presence of a footprint of a complex of two TFs, most likely Myc with another E-box binding factor (Figure 29 a). However, as many TFs (such as Usf1, BHLHA family members, Arntl, Hey2 etc.) can bind to the non-canonical E-box it is difficult to pinpoint which TF is actually binding. On the other hand, in the down/no-deg classification most of the top features identified are under-represented in the down genes and more enriched in the no-degs.

In E μ -myc T, for the up/no-deg classification we identified several composite motifs that indicate the presence of a footprint of Myc binding very close to another TF, such as E2f1 (E2F1_HES7 motif, where the HES7 binding site is a canonical E-box), Tfp2c (TFAP2c_MAX motif, where the MAX binding site is a canonical E-box), E2f1 (E2F1_NHLH1 motif, where the NHLH1 binding site is a canonical E-box) and TEAD family members (TEAD4_CLOCK, where the Clock motif is a canonical E-box) (Figure 29c). The TEAD factors can form a complex with YAP/TAZ to interact with many cell cycle regulators, including Myc (Leone et al., 2001). This list also contained the E2F4 motif in agreement with the Pscan analysis (Figure 26) which identified this motif as enriched in upregulated list. Myc can induce the expression of E2F factors which are a component of Myc dependent control of cell proliferation and cell fate decisions pathways (Alvaro-Blanco et al., 2009) and it is possible that these factors play a role in Myc dependent gene activation in the E μ -myc system. We also found the NRF1/NRF2 motif among the top-features in the up/no-deg list. The Nrf1 transcription factor is a master regulator of the mitochondrial biogenesis mechanism respiratory chain and Myc can mediate its apoptotic function by binding to the Nrf1 target genes (Virbasius et al., 1993). The presence of these factors

could indicate a possible ‘piggy backing’ scenario (see Chapter 1) where Myc binds to the DNA indirectly by binding to these factors. Similar to 3T9^{MycER} system, very few features were enriched (Figure 29) in the down/noddeg classification. Among these were the transcription factors Cdca7l, Tfp2b, Hen1 (binds to the non-canonical E-box) and Nr0b1. These top features in the up/no-deg and down/no-deg lists can be considered as candidate binding partners of Myc in gene-regulation.

a MycER 4h up/no-deg

Feature	Motif	Direction
H4K16ac enr ratio	NA	+
Pol II enr ratio	NA	+
H3K4me3 enr ratio	NA	+
H3k36me3 enr ratio	NA	+
Myc enr ratio	NA	+
DHS enr ratio	NA	+
H3K27ac enr ratio	NA	+
H3K36me2 enr ratio	NA	-
H3K4me1 enr ratio	NA	-
CLOCK_BHLHA15		+
ETV2_DLX3		-
CGI state	NA	+
ETS2		-
ETV5_FOXI1		-
FOXO1_ELK1		-
MAX		+
ETV2_HOXA2		-
GABP1, GABP2		-
CTCF		+

b MycER 4h down/no-deg

Feature	Motif	Direction
H4K16ac enr ratio	NA	-
H3K36me3 enr ratio	NA	+
H3K4me3 enr ratio	NA	-
Myc enr ratio	NA	-
H3K27ac enr ratio	NA	-
Pol2 enr ratio	NA	-
H3K36me2 enr ratio	NA	+
H3K4me1 enr ratio	NA	-
DHS enr ratio	NA	+
SP4		-
NRF1		-
CGI state	NA	-
ETV2_RFX5		-
ZFX		-
ERF_HES7		-

c $E\mu$ -myc T up/no-deg

Feature	Motif	Direction
Poll II enr ratio	NA	+
H3k4me3 enr ratio	NA	+
Rad21 enr ratio	NA	+
H3k27ac enr ratio	NA	+
CTCF enr ratio	NA	+
H3k4me1 enr ratio	NA	+
E2f1_Hes7		+
E2f4		+
Tfap2c_Max		+
E2f1_Nhlh1		+
Tead4_Clock		+
Gcm1_Pitx1		+
Nrf1, Nrf2		+

d $E\mu$ -myc T down/no-deg

Feature	Motif	Direction
H3k4me3 enr ratio	NA	-
Poll II enr ratio	NA	-
Rad21 enr ratio	NA	-
H3k27ac enr ratio	NA	-
Cdca7l		+
Tfap2b		+
Hen1		+
Ets2, Ets1		-
Tead_Clock		-
Elk3		-
Nr0b1		+

e tet-MYC Myc-dependent up/no-deg

Feature	Motif	Direction
enPol II	NA	+
enH3K27ac	NA	+
enH3K36me3	NA	+
enH3K4me3	NA	+
enH3k79me2	NA	+
enH3K4me1	NA	-
MYCL2, MAX, MYC		+
ZNF740		-
SP2, SP3, ZBT7B, KLF5, SP4		-
GCM1_MAX		+
E2F1		+
NFYA, FOXI1, NFYB		+
enMyc	NA	+
TEAD4_HES7		+

f tet-MYC Myc-dependent down/no-deg

Feature	Motif	Direction
enH3K4me3	NA	-
enPol II	NA	-
enH3K27ac	NA	-
ETV2_RFX5		-
enH3K36me3	NA	-
TFAP2C_HES7		-
HINFP1		-
E2F3_HES7, E2F1_HES7		-
HES1		-
NFIX		+
ETV2_FOXI1		-
TFAP4_DLX3		+

g tet-MYC Myc-independent up/no-deg

Feature	Motif	Direction
enH3K79me2	NA	+
ETV2_RFX5		-
ETV2_HES7		-
enPol II	NA	+
enH3K36me3	NA	+
HINFP1		-
E2F1_ELK1		-
ETV5, EHF, ELK1, GABPA		-
YY2		-

h tet-MYC Myc-independent down/no-deg

Feature	Motif	Direction
E2F2		-
ELK1_HOXA1		-
E2F1_EOMES		-
ETV2_HES7		-
SIX5		-
YY2		-
SPZ1		-
E2F1_HES7		-
enH3K79me2	NA	+
ETV5_FOXI1		-

Figure 29. Top features obtained from the model with the highest AUC from random forests based classifications ranked by their variable importance (a-b) $3T9^{MycER}$ up/no-deg and down/no-deg (c-d) $E\mu$ -myc T vs.C up/no-deg and down/no-deg and (e-f) tet-MYC Myc-dependent up/no-deg and down/no-deg (g-h) tet-MYC Myc independent. The red boxes around a motif indicate that the motif contains a canonical E-box sequence while a green box indicates a non-canonical E-box sequence. The direction indicates whether a feature is more enriched (+) or less enriched (-) in the positive set compared to the negative set. The E-box sequence is among the top most predictive features in all the lists and is always more enriched in the upregulated genes.

In the tet-MYC system, we once again found the TEAD4_HES7 motif as a top predictive feature in the Myc dependent up/ no-deg classification (similar to the TEAD4_CLOCK motif found in the top most predictive features of the up/no-deg classification of $E\mu$ -myc T vs. C) (Figure 29 c). Conversely, both in the Myc independent up/no-deg and down/no-deg, all the top motif features that were identified were only enriched in the no-degs and not in the up or down genes (Figure 29 d).

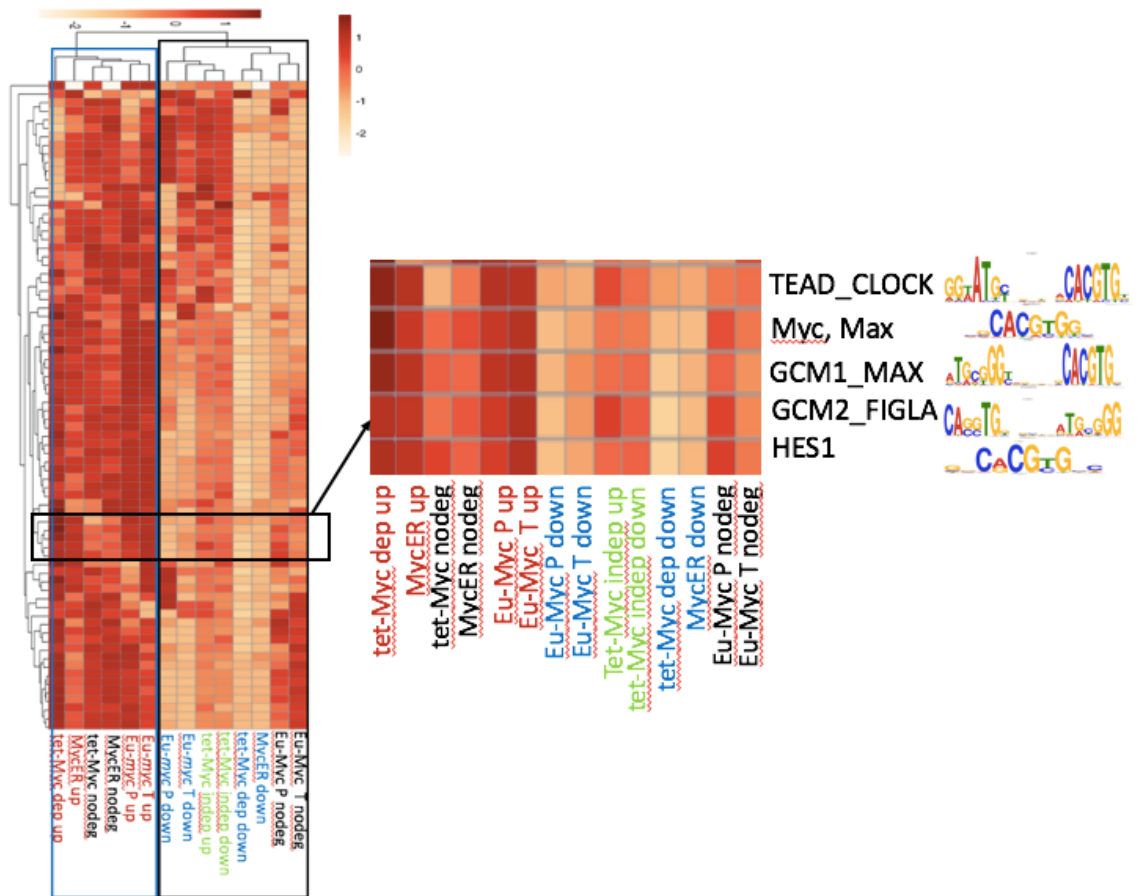


Figure 30. Heatmap showing the normalized frequencies (footprinted only) of the top 20 predictive features (PWMs) from the random forest classification in $3T9^{MycER}$ (up/no-deg and down/no-deg), $E\mu$ -Myc (up/no-deg and down/no-deg) and tet-MYC (Myc dependent up/no-deg and down/no-deg and Myc independent up/no-deg and down/no-deg) in the different gene subsets (up, down, Myc-dependent down etc.). The frequencies are normalized first by the number of genes in the subset and the by the row-wise z-score. All the upregulated genes (highlighted by the blue colour box) cluster separately from the downregulated genes (highlighted by the black colour box), while the no-degs mix with the both clusters. We also identify a small cluster of PWMs that are highly enriched in only the upregulated gene subsets.

To compare the overlap of the top motifs across the 3 systems, we considered all the motifs that appear among the top 20 features from the different comparisons (mycER up/down, up/no-deg, etc.). Next, we calculated their enrichment in each of the categories (mycER down, mycER up etc.). For $3T9^{MycER}$ and $E\mu$ -myc, we calculated this enrichment by adding up the total number of times a motif is found under the footprints overlapping a promoter of

the genes in a MYC, particular set, divided by the total number of genes in that set. For tet-MYC, where we do not have footprint data, we considered the motifs that were present within -450/+50 of the TSS. We then plotted a heatmap showing the enrichment of motifs in the up, down (Myc-dep and Myc independent for tet-MYC) and the no-deg categories from all the 3 systems ($3T9^{\text{mycER}}$, $E\mu\text{-myc}$, tet-MYC) (Figure 30).

This heatmap allows us to extract the distribution of the PWMs across samples. We found that the down and up categories separate quite well, while, the no-degs do not form a distinct cluster and tend to mix with the ups. The tet-MYC Myc-dependent up, $3T9^{\text{mycER}}$ up, tet-MYC no-deg, $3T9^{\text{mycER}}$ no-deg, $E\mu\text{-myc}$ up (P and T) cluster together indicating that they have similar PWM enrichments and most of them are more enriched in this cluster compared to the other two clusters. Both the Myc-independent up and down categories cluster together as expected. We also found that many of the top features are shared across the systems. For example, TEAD_CLOCK motif (where CLOCK motif is an E-box) is highly enriched in the up-regulated genes of all the systems.

Finally, for the $3T9^{\text{MycER}}$ system, we tried to classify the gene categories using only the footprint information (motif counts) in order to test whether the PWM features alone were able to classify the data into the different categories. We found that the performances of the classifiers were limited, with a maximum AUC of 0.62 in the up/down model while for the other models the AUC was lower (Table 5). One of the explanations for this decrease in the performance is that while the histone marks are general features for all up-regulated and down-regulated genes, Myc binding to the DNA with binding partners might occur only on specific subsets of genes within the up and down genes. Therefore, the overall dataset would be too heterogeneous to be classified based on PWMs alone but these PWM features can

still be picked up by the random forest classification approach, albeit with a lower but significant variable importance.

Model	AUC	AUC	AUC
	All features all genes	PWMs only all genes	PWMs only high fold change genes
Up/Down	0.91	0.62	0.68
Up/No-deg	0.84	0.51	0.54
Down/No-deg	0.79	0.62	0.65

Table 5: Table showing the AUCs of the 3 models up/down, up/no-deg and down/no-deg using different features in 3T9^{MycER}. The models perform best when using both ChIP-seq ratios and PWMs as features. Removal of ChIP-seq features leads to a drastic drop in the AUCs. Some improvement is observed when using only PWMs as features to classify only the genes with high-fold change (>0.5 for Up and <1 for Down).

3.5.3.1 Validation using CH12 ChIP-seq datasets

To validate the results obtained with the random forest method, we used publicly available ChIP-seq datasets from the CH12 cell line, a B-cell lymphoma cell line, similar to E μ -myc P and T cells. The CH12 cell line is a mouse B-cell lymphoma cell line. ChIP-seq data-sets are available from the ENCODE project of the transcription factors E2f4, Ets1, Nrf2, Hcfc1, Zkscan1, Nelfe, Gcn5, Znf384, Maz, Ctcf, Chd1, Mxi1, Bhlhe40, Rad21, Sin3A and p300. In order to check the similarity between the two systems, especially with regards to Myc we calculated the overlap of the CH12 Myc ChIP-seq peaks with the ChIP-seq peaks of Myc in the E μ -myc. Although the overlap was moderate (46% in E μ -myc T) when considering all the peaks (Figure 31 a), the peaks on promoters of E μ -myc T overlapped well (72%, Figure

31 b). This observation allowed us to study the overlap the Myc peaks (on promoters) in $E\mu$ -*myc* T sample with the available ChIPs of other transcription factors in this system. If the peaks of a TF overlap well with the Myc peaks, it could indicate a potential interaction with Myc. We were particularly interested in the overlap of the Nrf2 and E2f4 peaks with the Myc peaks in $E\mu$ -*myc* T samples, as their motifs were among the top scoring motifs in our random forest classification for up vs. no-deg classification in $E\mu$ -*myc* T.

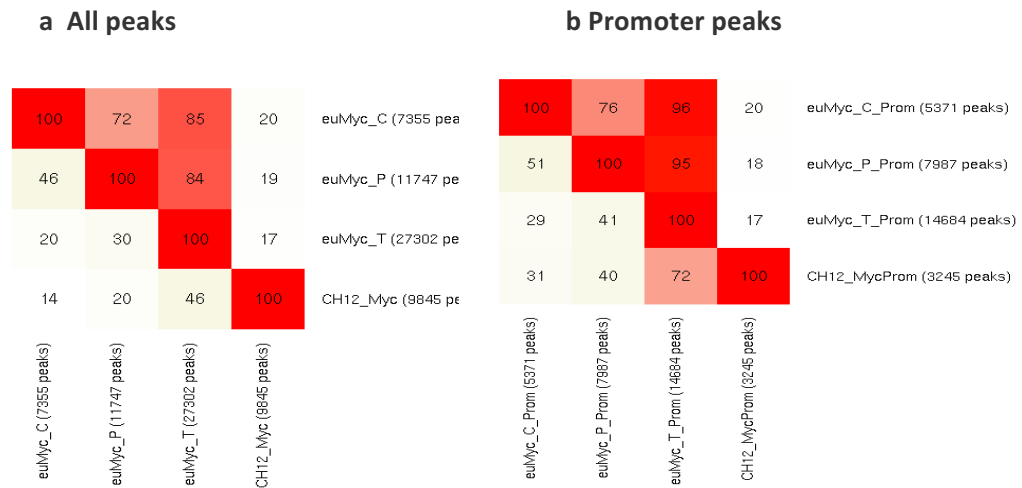


Figure 31. (a) The overlap of all the CH12 Myc peaks with the Myc peaks in the $E\mu$ -*myc* C, P and T samples (b) The overlap of only the promoter Myc peaks in CH12 with the promoter Myc peaks in the $E\mu$ -*myc* C, P and T samples. The promoter Myc peaks of $E\mu$ -*myc* T sample have a significant overlap (72%) with the CH12 Myc peaks.

We downloaded ChIP-seq experiment data for transcription factors in the CH12 cell line present in the ENCODE data repository and calculated the percentage of overlap of all these TF peaks with the Myc peaks in $E\mu$ -*myc* C, P, T (Figure 32). We found that indeed the peaks of the transcription factors Nrf2 and E2f4 overlap well with the Myc peaks in T sample along with the TFs Hcfc1, Nelfe, Gcn5, Maz, Mxi1 and Sin3A (>70% overlap) suggesting that these TFs could bind close to Myc in this system. Hcfc1 (Lundberg et al., 2016), Gcn5 (Martínez-Cerdeño et al., 2012), Nrf2 (Levy and Forman, 2010), Mxi1 (Armstrong et al., 2013) and Sin3A (Garcia-Sanz et al., 2014) have been reported to interact with Myc in other

systems. Hence, it is possible that these factors could be possible binding partners of Myc in driving the de-regulation of genes in the $E\mu$ -myc system, among these, Nrf2 and E2f4 are particularly well suited candidates as their footprinted PWMs were among the top-predictive features in the $E\mu$ -myc T up vs. no-deg classifications.

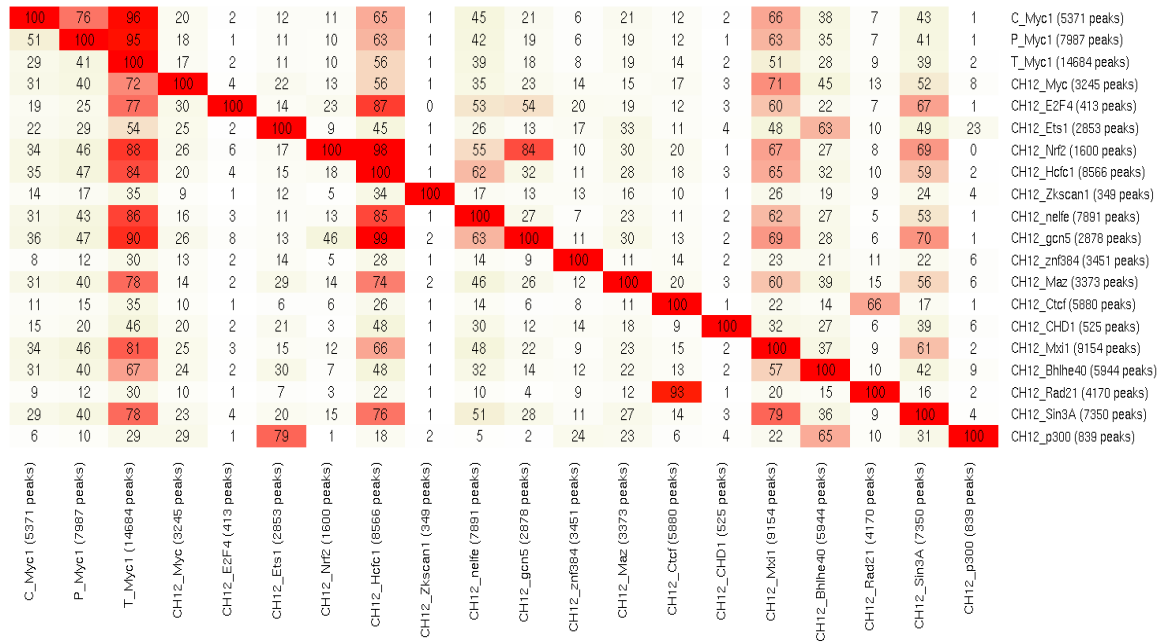


Figure 32. Heatmap of the percentage of overlap of CHIP-seq peaks of different transcription factors in the CH12 cell line with the Myc bound promoters in $E\mu$ -myc C, P and T samples. E2f4, Nrf2, Hcfc1, Nelfe, Gcn5, Maz and Sin3a peaks show a high overlap (greater than or equal to 74%) with the Myc peaks in $E\mu$ -myc T.

Very few of the enriched PWMs identified by Pscan scored high in the random forest classifications. The PWMs of E2F family factors such as E2F2, E2F3 and E2F4 ($E\mu$ -myc T up) were among the few that were identified by both the approaches. The Bcl6b PWM that was over-represented in down-regulated genes in $E\mu$ -myc T had very low variable importance in the random forest classification. We also identified many new features in the list of top discriminative features that were not identified by the Pscan analysis. Moreover, unlike the DREME results, the features we identified using this random forest approach were specifically enriched in only one of the two categories being classified. Therefore, the

random forest method allowed us to identify features that could be potential binding partners of Myc in the three systems.

Chapter 4

Discussion

In this thesis, we describe several approaches to identify binding partners of the transcription factor Myc in gene-regulation in three models where Myc is overexpressed: 3T9^{MycER}, Eμ-*myc* and tet-MYC. The most promising approach is based on high-depth DNase-seq data which reveals genome-wide transcription factor binding sites. However, to the best of our knowledge there was no pipeline to automatically carry out of all the steps of DNase-seq data analysis. We addressed this issue by developing a pipeline that integrates state-of-the-art tools to automatize the analysis of raw DNase-seq data (see chapter 3.2). We also demonstrated how the processed DNase-seq data can be integrated with other types of NGS data (RNA-seq and ChIP-seq) to identify key binding partners in the systems considered (chapter 3.5). To select the best footprint caller to be inserted into our pipeline, we carried out a benchmarking study on the available tools that use high-depth DNase-seq data to call TF footprints, using data from the ENCODE data repository (see chapter 3.1). The study identified the Wellington as the algorithm performing best both in terms of specificity and sensitivity. Hence, we used the Wellington as the footprint caller of choice in our pipeline. With the footprints obtained from Wellington, we constructed TF-TF interaction networks at different sequencing depths by down sampling the original datasets at 100%, 70% and 30% of the original reads. Although the depth of footprinting is important for identifying TF

footprints, the most important interactions in the resulting TF-TF network are still retained even using lower-depth data.

The DNase-seq analysis pipeline that we developed provides an easy and automated way to carry out DNase-seq data for both high depth and low depth experiments. The final outputs of the pipeline are the lists of DHSs and footprints in bed and bigBed format. The pipeline also provides a motif analysis block with options to carry out *de novo* motif analysis using MEME and /or DREME, motif over-representation analysis using Pscan and motif searching using FIMO. *De novo* motif analysis can be used to identify the transcription factors regulating the genes in the system of interest. Motif over-representation analysis can instead be used to transcription factors that are over or underrepresented in a particular set of genes compared to another.

To perform the motif analyses mentioned above, we compiled a list of 2433 different PWMs from various databases and published studies (see Chapter 3.2). We also propose a way to simplify this list of PWMs in order to reduce their redundancy and to combine similar transcription factors together into separate PWM classes. We classified all the PWMs based on their similarities into 445 classes containing all the 1815 non-composite motifs and 295 classes containing the 618 composite ones. This piece of information was stored in a database containing the name of each PWM, the corresponding PWM class and the family the transcription factor belongs to. Such information can be extremely useful in interpreting the results of motif analysis: sometimes, an enriched motif can be associated by a motif matching tool such as TOMTOM to a particular TF from a database, yet it could be bound by another very similar TF. The database that we created (see 3.2) helps to resolve these scenarios as we can check the list of TFs that bind to a particular motif before drawing a

conclusion. For example, the E-box is bound by many different transcription factors such as Max, Clock, Hey2 and Usf1. However, in our systems we know that when Myc is over-expressed it binds to almost all the open promoters. Therefore, when we find a motif associated to Clock or Hey2 to be enriched under the summit of a Myc peak or on footprints under the summit of a Myc peak, it is most likely that it is Myc that is binding there.

By carrying out motif analysis on the DHSs with DREME, we confirmed the presence of the E-box under the ChIP-seq peaks of Myc in open chromatin in both 3T9^{MycER} and Eμ-*myc*. Then, using Pscan, we identified the over-represented motifs in the up and the down genes compared to the no-degs. The results showed that the E-box is more enriched in the up genes compared to the no-degs. Next, using footprints identified with high depth DNase-seq we restricted the region of interaction of the TF to the DNA to improve the specificity of the motif analysis. We confirmed the presence of the E-box motif in the footprinted sequences and found that the E-box footprints have highest enrichment compared to the footprints of other TFs (Figure 22). By comparing the distribution of footprinted E-boxes on the up, down and no-deg genes we found that up and no-deg genes contain more footprinted E-boxes compared to down genes. This represents an indication that in down genes Myc binds to the DNA either mostly indirectly through another TF by “piggy backing” or aspecifically without recognizing a specific motif.

As DREME only looks for the motifs enriched in a single set of sequences it was not possible to identify motifs specific to one subset of genes as compared to another (e.g up vs. no-deg). On the other hand, Pscan could identify over-represented motifs in a set of sequences compared to another but it was not possible to run this analysis on footprint sequences as it did not accept short sequences (less than 100bp). Hence, it was difficult to determine whether

the over-represented motif identified by Pscan were actually bound by a TFs. In fact, many of the motifs identified by Pscan, such as Bcl6b (Figure 26), did not score high in the random forest classification. The most probable reason for this could be that although the motif of Bcl6b is enriched in the down genes of Eμ-*myc* T, it is not bound by the TF and hence not footprinted. Alternatively, it is possible that the TF does not footprint (due to low residence times), but the footprint profile comparison in chapter 3.2 makes this scenario seems unlikely as the motif of Bcl6b is 12 bp long with many highly conserved residues.

We developed a new R class, called ‘DHS’, and several methods to store and carry out integrative analysis of DNase-seq data and footprints. The new class we created is called ‘DHS’, and has three slots for storing all the DHSs, the footprints identified within these DHSs, and the PWM matched by FIMO to the footprint, respectively, from one sample, all in one object. We also provide specific methods for this class that extract which footprints fall under the summit of the Myc peak, separate the DHSs which are on promoters and distal elements. In addition, we also developed specific functions and methods for motif analysis including a method to run Pscan and return the results in an easily readable table in .xlsx or .txt format. This method takes as input two sets of sequences in the FASTA format, a positive set and a negative set, or background, which are passed to Pscan. The output file obtained from the Pscan analysis is corrected for multiple testing using the Benjamini-Hochberg method, the PWMs are filtered based on the corrected p-value (default cut-off 0.01) and the final output is written to a .xlsx or .txt file. Lastly, also wrote a function to look for PWMs in our database in a given set of sequences using FIMO and return the output in a GRange format.

Although with the motif analyses we identified some possible binding partners of Myc, this was not a conclusive list as DREME only finds motifs enriched in a single set of genes and Pscan could not be run on footprint regions. So, we chose to use approaches that can combine all the pieces of information that we have in these systems to classify the up, down and no-deg genes. Our main goal here was not to classify the data but to identify the features that can give the best classification of the gene subsets. We searched for all possible binding sites of transcription factors using FIMO on comprehensive list of PWMs collected from multiple published resources (described in section 3.2). We then selected the sites overlapping a footprint and bound by Myc (overlap with a Myc peak) and counted the number of times each PWM is footprinted on our genes of interest (i.e. belonging to up, down and no-deg lists). Finally, we used this piece of information and the ChIP-seq enrichment ratios to classify the up, down and no-deg categories. If a feature is able to classify the data well, it would mean that the feature is discriminative and has different distributions in the two categories considered.

We first applied this approach using one feature (PWM features, presence of absence of CpG islands and ChIP-seq enrichment ratios) at a time and calculated its predictive power in classifying the data in pairwise-combinations. Although the ChIP-seq features such as H3K27ac and Pol II were able to classify the data well with AUCs higher than 0.8, the same could not be said of the PWM, which classified the data with very poor AUCs. This indicated that this approach is not suitable for our problem and most likely, a single feature is not responsible alone in driving the differences between the two categories, but a combination of features separates the two categories. When we used the same features but with a random forest classifier that combines all the features together, we obtained a much improved result. The random forest classifier not only gave a good separation of data but also identified some

interesting PWMs features among the top predictive list of features (based of variable importance) that could potentially be binding partners of Myc.

The presence of the enrichment ratios of H3K27ac, H4K16ac and Myc in the list of top discriminating features from the random forest classifiers in all the three systems can be considered a positive control confirming that this approach is able to identify features that are known to discriminate these gene sets. Since both H3K27ac and H4K16 are activation marks they are expected to be highest in the activated genes and lowest in the repressed genes.

In the 3T9^{MycER} system, in the up/no-deg classification (Figure 29 a), we found the footprinted complex motif of canonical E-box with a non-canonical E-box indicating that two E-box binding transcription factors are present on the promoters of those genes, one of which is most likely Myc. In addition, we also found other motifs, such as the CTCF which has been linked to the activation of c-MYC expression (Gombert and Krumm, 2009; Klenova et al., 1993). On the other hand, for the down/no-deg classification we found some PWM in the list of top predictive features, but all of them were enriched in the no-deg genes and not in the down genes. Therefore, the TFs binding to these motifs could not be considered as binding partners of Myc in downregulation.

In the E μ -myc system, some of the PWMs that we identified using the random forest algorithm for the up/no-deg classification among others are the NRF1/NRF2, TEAD4, the TFAP2 family factors and ETS transcription factors. All of these TFs have connections with Myc in gene-regulation (Morrish et al., 2003; Roussel et al., 1994; Virbasius et al., 1993; Wasylyk et al., 1998). We also found the complex motif of activation factor 2 (TFAP2) and

the E-box (TFAP2c_MAX) in the up/no-deg classification: the Tfp2 family consists of 5 different homologous transcription factors (TFAP2a-e) which contain a highly conserved C-terminal helix-span-helix motif required for dimerization. Nrf2 is an oncogene that has been reported to interact with Myc (Levy and Forman, 2010) and is suggested to be up-regulated due to the tumorigenic activity of c-Myc (DeNicola et al., 2011). Moreover, Myc is also reported to form a ternary complex with Nrf2 and p-c-Jun to regulate drug-metabolism (Levy and Forman, 2010). In this system too, for the down/no-deg classification we found few significant PWMs enriched in the down genes: Hen1, Cdca7l and Tfp2b. The core of the Hen1 motif is a non-canonical E-box (CAGCTG), indicating that the downregulated genes have more non-canonical E-boxes compared to the no-degs. Cdca7l (also known as Jpo2 or R1) is known to be a transcriptional repressor (Chen et al., 2005) and is reported to play an important oncogenic role in mediating the full transforming effect of Myc in medulloblastoma cells (Huang et al., 2005). Given this connection of Cdca7l to Myc in the development of medulloblastoma, it is possible that it also plays a role in Myc-dependent gene regulation in lymphomagenesis. The TEAD_E-box is another top predictive feature of the up/no-deg classification in the $E\mu$ -myc which suggests that the footprint of the Tead family members binding in a complex with Myc is enriched in the up-regulated genes. The Tead proteins are transcription factors involved in development and are also known to play a role in cancer development (Knight et al., 2008; Landin Malt et al., 2012; Skotheim et al., 2006).

The addition of the tet-MYC system to our analysis helped to separate the Myc dependent response from the Myc-independent response. Although we do not have footprints in this system, the motif matches under the summit of the Myc peak can still be used to guess the transcription factors that are involved in the Myc-dependent and Myc-independent gene regulation. We once, again found the TEAD4_E-box motif in the up/no-deg classification

indicating that this feature is not only more enriched in the upregulated genes but is specific for the Myc-dependent upregulated genes as we do not find it enriched in the Myc-independent genes. Moreover, in the heatmap integrating the top features of all the classifications (Figure 30), TEAD4_E-box is more enriched in all the upregulated list compared to the no-deg and downregulated lists, thus, indicating a possible direct interaction of Myc with Tead in Myc-dependent upregulation of genes.

In the case of the Myc independent genes in the tet-MYC system, none of top features that were identified can be called 'discriminative' because the AUCs for both the Myc-independent up/no-deg and down/no-deg were low (0.62 and 0.64). These values could mean that there are no specific transcription factors that drive the de-regulation of these genes together with Myc. The histone marks and ChIP-seq ratios too do not separate these classes well. Therefore, these genes are most likely to be deregulated by indirect effects.

The three systems that we studied were different in terms of cell types but had one common feature: the overexpression of Myc. Therefore, it was not surprising that they shared some common PWM features as shown by the heatmap (Figure 30) displaying the enrichments of the top PWM features from all the classifications in the different gene subsets (up, down, no-deg in 3T9^{MycER} and Eμ-*myc*, Myc-dependent up and down, Myc-independent up and down in tet-MYC). All the upregulated lists (except for the Myc independent list in tet-MYC) were present in the same cluster along with the no-deg lists, while the downregulated lists formed a separate cluster indicating that these systems share many common transcription factors that drive the regulation of the up and down genes. In all the systems, in down-regulated lists fewer features were enriched indicating once again the down-regulated genes are likely to be regulated by indirect effects.

Applying the random forest method, we were able to identify some features corresponding to TF binding motifs, that correspond to possible binding partners of Myc in the systems that we studied. Although the variable importance of transcription factor footprints was lower than the histone marks, their presence among the top discriminating features is nevertheless intriguing. The lower predictive power of the TF footprints instead may indicate that Myc does not regulate all the up or down-regulated genes in the same manner, but binds to the DNA with different binding partners in different subsets of genes to up- or down-regulate their expression. If these subsets make up only a small percentage of all the up or down genes it would explain the lower variable importance of the corresponding PWM class feature.

Only a few of the enriched PWMs, such as those belonging to the E2F family factors E2f2, E2f3 and E2f4 identified by Pscan scored high in the random forest classifications. Many of the promising candidate binding partners of Myc identified by Pscan, such as Bcl6b, scored low in the random forest classifications indicating that the mere presence of an enriched motif is not a sufficient piece of evidence. We identified a large number of new features that we did not find using Pscan most likely because we could not run analysis on use it on footprints regions. This analysis also helped to overcome the issue that we had with DREME in distinguishing the TFs specific to the up and down genes, for example, the motif of Nrf1(similar to Nrf2 motif) TF was found by DREME in both $E\mu$ -myc T up and down. Using random forest, we found that this PWM was specific to the up vs. no-deg classification and not the down vs. no-deg classification. Moreover, by the calculating the enrichment of the top most predictive PWMs on the up, down and no-degs genes we confirmed that this motif is enriched in the up genes and not in the down genes in this system.

Our results show that using a random forest approach to classify the Myc-regulated up/down, up/no-deg and down/no-deg genes can identify features that discriminate these gene sets. This might otherwise prove difficult to do with qualitative or visual approaches such as heatmaps and boxplots. The random forest based approach identified some TFs that could be potential binding partners of Myc involved in gene-regulation. The advantage of this approach lies in its ability to handle hundreds of features and extract the most informative ones while ignoring the rest. However, when using footprint calls as a proxy for TF binding it should be kept in mind that this approach would fail to identify TFs which have transient binding times (Yardımcı et al., 2014) as well as TFs without any known motifs.

The random forest approach that we applied can also be used for a variety of other problems which requires the separation of two classes based on many features, for example, a set of genes known to be involved in a particular pathway compared to a background set. In the future, we plan to extend this approach to other models and identify features that differentiate the up- down and no-deg categories in these models. Clearly, these results need to be validated using experimental approaches, for example using knock out (of a candidate TF) mice to test the effect of the knock out on tumour progression in the E μ -myc C, P, T model.

References

- Adams, J.M., Harris, A.W., Pinkert, C.A., Corcoran, L.M., Alexander, W.S., Cory, S., Palmiter, R.D., and Brinster, R.L. (1985). The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature* 318, 533–538.
- Alitalo, K., Schwab, M., Lin, C.C., Varmus, H.E., and Bishop, J.M. (1983). Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* 80, 1707–1711.
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461.
- Alvaro-Blanco, J., Martínez-Gac, L., Calonge, E., Rodríguez-Martínez, M., Molina-Privado, I., Redondo, J.M., Alcamí, J., Flemington, E.K., and Campanero, M.R. (2009). A novel factor distinct from E2F mediates C-MYC promoter activation through its E2F element during exit from quiescence. *Carcinogenesis* 30, 440–448.
- Amati, B., Brooks, M.W., Levy, N., Littlewood, T.D., Evan, G.I., and Land, H. (1993). Oncogenic activity of the c-Myc protein requires dimerization with Max. *Cell* 72, 233–245.
- Armstrong, M.B., Mody, R.J., Ellis, D.C., Hill, A.B., Erichsen, D.A., and Wechsler, D.S. (2013). N-Myc Differentially Regulates Expression of MXI1 Isoforms in Neuroblastoma. *Neoplasia* 15, 1363–1370.
- Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* (80-.). 339.
- Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27, 1653–1659.

- Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.
- Barozzi, I., Bora, P., and Morelli, M.J. (2014). Comparative evaluation of DNase-seq footprint identification strategies. *Front. Genet.* 5, 278.
- Bianchi, V., Ceol, A., Ogier, A.G.E., de Pretis, S., Galeota, E., Kishore, K., Bora, P., Croci, O., Campaner, S., Amati, B., et al. (2016). Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions. *Front. Genet.* 7, 75.
- Blackwood, E., and Eisenman, R. (1991). Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* (80-). 251, 1211–1217.
- Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E.E.M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* 44, 148–156.
- Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322.
- Breiman, L. (1984). *Classification and regression trees* (Chapman & Hall).
- Breiman, L., and Leo (2001). Random Forests. *Mach. Learn.* 45, 5–32.
- Brutlag, D., Schlehuber, C., and Bonner, J. (1969). Properties of formaldehyde-treated nucleohistone. *Biochemistry* 8, 3214–3218.
- Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212, 563–578.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenlead, W.J., (2013). Transposition of native chromatin for fast and sensitive epeigenomic profiling of open

- chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*, 1213-1218.
- Cairo, S., Armengol, C., De Reyniès, A., Wei, Y., Thomas, E., Renard, C.-A., Goga, A., Balakrishnan, A., Semeraro, M., Gresh, L., et al. (2008). Hepatic stem-like phenotype and interplay of Wnt/beta-catenin and Myc signaling in aggressive childhood liver cancer. *Cancer Cell* *14*, 471–484.
- Chen, K., Ou, X.-M., Chen, G., Choi, S.H., and Shih, J.C. (2005). R1, a Novel Repressor of the Human Monoamine Oxidase A. *J. Biol. Chem.* *280*, 11552–11559.
- Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* *45*, 1127–1133.
- Collins, S., and Groudine, M. (1982). Amplification of endogenous myc-related DNA sequences in a human myeloid leukaemia cell line. *Nature* *298*, 679–681.
- Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* *306*, 636–640.
- Cooper, GM (2000). *The Cell: A Molecular Approach*. 2nd edition. Sunderland (MA): [Sinauer Associates](#).
- Cs'ardi, G. and T.N. (2006). The igraph software package for complex network research. *Complex Syst.*
- Dalla-Favera, R., Wong-Staal, F., and Gallo, R.C. (1982). Onc gene amplification in promyelocytic leukaemia cell line HL-60 and primary leukaemic cells of the same patient. *Nature* *299*, 61–63.
- Dang, C. V (2013). MYC, metabolism, cell growth, and tumorigenesis. *Cold Spring Harb. Perspect. Med.* *3*, a014217-.
- Dawson, M.A., and Kouzarides, T. (2012). Cancer Epigenetics: From Mechanism to Therapy. *Cell* *150*, 12–27.

- DeNicola, G.M., Karreth, F.A., Humpton, T.J., Gopinathan, A., Wei, C., Frese, K., Mangal, D., Yu, K.H., Yeo, C.J., Calhoun, E.S., et al. (2011). Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature* 475, 106–109.
- Do, C.B., and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nat. Biotechnol.* 26, 897–899.
- Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C. a, Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
- Eilers, M., Picard, D., Yamamoto, K.R., and Bishop, J.M. (1989). Chimaeras of myc oncoprotein and steroid receptors cause hormone-dependent transformation of cells. *Nature* 340, 66–68.
- ENCODE Project Consortium, E., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.
- Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* 7, 1728–1740.
- Fleming, J.D., Pavesi, G., Benatti, P., Imbriano, C., Mantovani, R., and Struhl, K. (2013). NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Res.* 23, 1195–1209.
- Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and

- characterize protein–DNA interactions. *Nat. Rev. Genet.* *13*, 840–852.
- Gabay, M., Li, Y., and Felsher, D.W. (2014). MYC Activation Is a Hallmark of Cancer Initiation and Maintenance. *Cold Spring Harb. Perspect. Med.* *4*, a014241–a014241.
- Garcia-Sanz, P., Quintanilla, A., Lafita, M.C., Moreno-Bueno, G., García-Gutierrez, L., Tabor, V., Varela, I., Shio, Y., Larsson, L.-G., Portillo, F., et al. (2014). Sin3b interacts with Myc and decreases Myc levels. *J. Biol. Chem.* *289*, 22221–22236.
- Ge, T. (2015). TFBSTools: Software Package for Transcription Factor Binding Site (TFBS) Analysis. R Packag. Version 1.8.0 <http://jas>.
- Gombert, W.M., and Krumm, A. (2009). Targeted deletion of multiple CTCF-binding elements in the human C-MYC gene reveals a requirement for CTCF in C-MYC expression. *PLoS One* *4*, e6109.
- Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* *27*, 1017–1018.
- Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. *Genome Biol.* *8*, R24.
- Halkidou, K., Gaughan, L., Cook, S., Leung, H.Y., Neal, D.E., and Robson, C.N. (2004). Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer. *Prostate* *59*, 177–189.
- Hannonlab.cshl.edu/fastx_toolkit/ FASTX-Toolkit.
- Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* *39*, 311–318.
- Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA

- interactions in vivo by digital genomic footprinting. *Nat. Methods* 6, 283–289.
- Hothorn, T. (2005). Survival ensembles. *Biostatistics* 7, 355–373.
- Huang, A., Ho, C.S.W., Ponzielli, R., Barsyte-Lovejoy, D., Bouffet, E., Picard, D., Hawkins, C.E., and Penn, L.Z. (2005). Identification of a novel c-Myc protein interactor, JPO2, with transforming activity in medulloblastoma cells. *Cancer Res.* 65, 5607–5619.
- Izumi, H., Molander, C., Penn, L.Z., Ishisaki, A., Kohno, K., and Funa, K. (2001). Mechanism for the transcriptional repression by c-Myc on PDGF beta-receptor. *J. Cell Sci.* 114, 1533–1544.
- Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33, 245–254.
- Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527, 384–388.
- Kharchenko, P. V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 471, 480–485.
- Kim, J., Lee, J., and Iyer, V.R. (2008). Global identification of Myc target genes reveals its direct role in mitochondrial biogenesis and its E-box usage in vivo. *PLoS One* 3, e1798.
- Kishore, K., de Pretis, S., Lister, R., Morelli, M.J., Bianchi, V., Amati, B., Ecker, J.R., and Pelizzola, M. (2015). methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. *BMC Bioinformatics* 16, 313.
- Kistner, A., Gossen, M., Zimmermann, F., Jerecic, J., Ullmer, C., Lübbert, H., and Bujard, H. (1996). Doxycycline-mediated quantitative and tissue-specific control of gene expression in transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10933–10938.
- Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H.,

- Neiman, P.E., and Lobanenkov, V. V (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol. Cell. Biol.* *13*, 7612–7624.
- Knight, J.F., Shepherd, C.J., Rizzo, S., Brewer, D., Jhavar, S., Dodson, A.R., Cooper, C.S., Eeles, R., Falconer, A., Kovacs, G., et al. (2008). TEAD1 and c-Cbl are novel prostate basal cell markers that correlate with poor clinical outcome in prostate cancer. *Br. J. Cancer* *99*, 1849–1858.
- Kress, T.R., Pellanda, P., Pellegrinet, L., Bianchi, V., Nicoli, P., Doni, M., Recordati, C., Bianchi, S., Rotta, L., Capra, T., et al. (2016). Identification of MYC-Dependent Transcriptional Programs in Oncogene-Addicted Liver Tumors. *Cancer Res.* *76*, 3463–3472.
- Ku, C.S., Naidoo, N., Wu, M., and Soong, R. (2011). Studying the epigenome using next generation sequencing. *J. Med. Genet.* *48*, 721–730.
- Kulakovskiy, I. V, Vorontsov, I.E., Yevshin, I.S., Soboleva, A. V, Kasianov, A.S., Ashoor, H., Ba-Alawi, W., Bajic, V.B., Medvedeva, Y.A., Kolpakov, F.A., et al. (2016). HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.* *44*, D116-25.
- Land, H., Parada, L.F., and Weinberg, R.A. (1983). Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature* *304*, 596–602.
- Landin Malt, A., Cagliero, J., Legent, K., Silber, J., Zider, A., and Flagiello, D. (2012). Alteration of TEAD1 expression levels confers apoptotic resistance through the transcriptional up-regulation of Livin. *PLoS One* *7*, e45498.
- Leone, G., Sears, R., Huang, E., Rempel, R., Nuckolls, F., Park, C.-H., Giangrande, P., Wu, L., Saavedra, H.I., Field, S.J., et al. (2001). Myc Requires Distinct E2F Activities to Induce S Phase and Apoptosis. *Mol. Cell* *8*, 105–113.
- LeRoy, G., DiMaggio, P.A., Chan, E.Y., Zee, B.M., Blanco, M., Bryant, B., Flaniken, I.Z.,

-
- Liu, S., Kang, Y., Trojer, P., et al. (2013). A quantitative atlas of histone modification signatures from human cancer cells. *Epigenetics Chromatin* 6, 20.
- Levy, S., and Forman, H.J. (2010). C-Myc is a Nrf2-interacting protein that negatively regulates phase II genes through their electrophile responsive elements. *IUBMB Life* 62, 237–246.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Li, X., Yu, J., Brock, M. V., Tao, Q., Herman, J.G., Liang, P., and Guo, M. (2015). Epigenetic silencing of BCL6B inactivates p53 signaling and causes human hepatocellular carcinoma cell resist to 5-FU. *Oncotarget* 6, 11547–11560.
- Li, X., Wang, W., Xi, Y., Gao, M., Tran, M., Aziz, K.E., Qin, J., Li, W., and Chen, J. (2016). FOXR2 Interacts with MYC to Promote Its Transcriptional Activities and Tumorigenesis. *Cell Rep.* 16, 487–497.
- Li, Z., Van Calcar, S., Qu, C., Cavenee, W.K., Zhang, M.Q., and Ren, B. (2003). A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8164–8169.
- Littlewood, T.D., Hancock, D.C., Danielian, P.S., Parker, M.G., and Evan, G.I. (1995). A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins. *Nucleic Acids Res.* 23, 1686–1690.
- Liu, Y., Schmidt, B., Liu, W., and Maskell, D.L. (2010). CUDA-MEME: Accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units.
- Lundberg, S.M., Tu, W.B., Raught, B., Penn, L.Z., Hoffman, M.M., Lee, S.-I., Au, S., Belsley, D., Kuh, E., Welsch, R., et al. (2016). ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. *Genome Biol.* 17, 82.
- Lüscher, B., and Larsson, L.G. (1999). The basic region/helix-loop-helix/leucine zipper

- domain of Myc proto-oncoproteins: function and regulation. *Oncogene* 18, 2955–2966.
- Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697.
- Madrigal, P. (2013). DNaseR: DNase I footprinting analysis of DNase-seq data.
- Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682.
- Martínez-Cerdeño, V., Lemen, J.M., Chan, V., Wey, A., Lin, W., Dent, S.R., Knoepfler, P.S., Grant, P., Duggan, L., Cote, J., et al. (2012). N-Myc and GCN5 Regulate Significantly Overlapping Transcriptional Programs in Neural Stem Cells. *PLoS One* 7, e39456.
- Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142-7.
- Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108-10.
- Meyer, N., and Penn, L.Z. (2008). Reflecting on 25 years with MYC. *Nat. Rev. Cancer* 8, 976–990.
- Michalak, E.M., and Visvader, J.E. (2016). Dysregulation of histone methyltransferases in breast cancer – Opportunities for new targeted therapies? *Mol. Oncol.*
- Morrish, F., Giedt, C., and Hockenbery, D. (2003). c-MYC apoptotic function is mediated by NRF-1 target genes. *Genes Dev.* 17, 240–255.
- Müller, B.M., Jana, L., Kasajima, A., Lehmann, A., Prinzler, J., Budczies, J., Winzer, K.-J., Dietel, M., Weichert, W., Denkert, C., et al. (2013). Differential expression of histone

- deacetylases HDAC1, 2 and 3 in human breast cancer - overexpression of HDAC2 and HDAC3 is associated with clinicopathological indicators of disease progression. *BMC Cancer* *13*, 215.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* *320*, 1344–1349.
- Nagy, P.L., Cleary, M.L., Brown, P.O., and Lieb, J.D. (2003). Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc. Natl. Acad. Sci. U. S. A.* *100*, 6364–6369.
- Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012a). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* *489*, 83–90.
- Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J.A. (2012b). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* *150*, 1274–1286.
- Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* *37*, D77-82.
- Nguyen, C., Wang, Y., and Nguyen, H.N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Sci. Eng.* *6*, 551–560.
- Nickolls, J., Buck, I., Garland, M., and Skadron, K. (2008). Scalable parallel programming with CUDA. *Queue* *6*, 40.
- Oh, E.J., Yang, W.I., Cheong, J.-W., Choi, S.-E., and Yoon, S.O. (2014). Diffuse large B-cell lymphoma with histone H3 trimethylation at lysine 27: another poor prognostic phenotype independent of c-Myc/Bcl2 coexpression. *Hum. Pathol.* *45*, 2043–2050.

-
- Pachkov, M., Balwierz, P.J., Arnold, P., Ozonov, E., and van Nimwegen, E. (2013). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.* *41*, D214-20.
- Payne, G.S., Bishop, J.M., and Varmus, H.E. (1982). Multiple arrangements of viral DNA and an activated host oncogene in bursal lymphomas. *Nature* *295*, 209–214.
- Pearson, W.R., Wood, T., Zhang, Z., and Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics* *46*, 24–36.
- Peters, A.H.F.M., Mermoud, J.E., O’Carroll, D., Pagani, M., Schweizer, D., Brockdorff, N., and Jenuwein, T. (2002). Histone H3 lysine 9 methylation is an epigenetic imprint of facultative heterochromatin. *Nat. Genet.* *30*, 77–80.
- Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* *41*, e201.
- Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* *470*, 279–283.
- R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing.
- Reid, J.E., and Wernisch, L. (2014). STEME: A Robust, Accurate Motif Finder for Large Data Sets. *PLoS One* *9*, e90735.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., Swets, J., Pepe, M., Sonego, P., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* *12*, 77.
- Romanoski, C.E., Glass, C.K., Stunnenberg, H.G., Wilson, L., and Almouzni, G. (2015). Epigenomics: Roadmap for regulation. *Nature* *518*, 314–316.

- Ropero, S., Fraga, M.F., Ballestar, E., Hamelin, R., Yamamoto, H., Boix-Chornet, M., Caballero, R., Alaminos, M., Setien, F., Paz, M.F., et al. (2006). A truncating mutation of HDAC2 in human cancers confers resistance to histone deacetylase inhibition. *Nat. Genet.* *38*, 566–569.
- Roussel, M.F., Davis, J.N., Cleveland, J.L., Ghysdael, J., and Hiebert, S.W. (1994). Dual control of myc expression through a single DNA binding site targeted by ets family proteins and E2F-1. *Oncogene* *9*, 405–415.
- Sabò, A., Kress, T.R., Pelizzola, M., de Pretis, S., Gorski, M.M., Tesi, A., Morelli, M.J., Bora, P., Doni, M., Verrecchia, A., et al. (2014). Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature* *511*, 488–492.
- Shachaf, C.M., Kopelman, A.M., Arvanitis, C., Karlsson, Å., Beer, S., Mandl, S., Bachmann, M.H., Borowsky, A.D., Ruebner, B., Cardiff, R.D., et al. (2004). MYC inactivation uncovers pluripotent differentiation and tumour dormancy in hepatocellular cancer. *Nature* *431*, 1112–1117.
- Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature* *488*, 116–120.
- Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., and Gifford, D.K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* *32*, 171–178.
- Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* *15*, 272–286.
- Simon, J.A., and Kingston, R.E. (2009). Mechanisms of Polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* *10*, 697.

-
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* *21*, 3940–3941.
- Skotheim, R.I., Autio, R., Lind, G.E., Kraggerud, S.M., Andrews, P.W., Monni, O., Kallioniemi, O., and Lothe, R.A. (2006). Novel genomic aberrations in testicular germ cell tumors by array-CGH, and associated gene expression changes. *Cell. Oncol.* *28*, 315–326.
- Solomon, M.J., and Varshavsky, A. (1985). Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc. Natl. Acad. Sci. U. S. A.* *82*, 6470–6474.
- Song, L., and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* *2010*, pdb.prot5384.
- Song, J., Noh, J.H., Lee, J.H., Eun, J.W., Ahn, Y.M., Kim, S.Y., Lee, S.H., Park, W.S., Yoo, N.J., Lee, J.Y., et al. (2005). Increased expression of histone deacetylase 2 is found in human gastric cancer. *APMIS* *113*, 264–268.
- Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* *151*, 994–1004.
- Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell* *161*, 555–568.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* *8*, 25.
- Takenaga, M., Hatano, M., Takamori, M., Yamashita, Y., Okada, S., Kuroda, Y., and Tokuhisa, T. (2003). Bcl6-dependent transcriptional repression by BAZF. *Biochem.*

Biophys. Res. Commun. 303, 600–608.

Taylor, B.S., DeCarolis, P.L., Angeles, C. V, Brenet, F., Schultz, N., Antonescu, C.R., Scandura, J.M., Sander, C., Viale, A.J., Socci, N.D., et al. (2011). Frequent alterations and epigenetic silencing of differentiation pathway genes in structurally rearranged liposarcomas. *Cancer Discov.* 1, 587–597.

Taylor, G.C.A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M.M., and Bickmore, W.A. (2013). H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res.* 23, 2053–2065.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82.

Tie, F., Banerjee, R., Stratton, C.A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M.O., Scacheri, P.C., and Harte, P.J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes *Drosophila* Polycomb silencing. *Development* 136, 3131–3141.

Tsompana, M., Buck, M.J., Luger, K., Mader, A., Richmond, R., Sargent, D., Richmond, T., Richmond, T., Davey, C., Kornberg, R., et al. (2014). Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 7, 33.

Vakoc, C.R., Mandat, S.A., Olenchok, B.A., and Blobel, G.A. (2005). Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol. Cell* 19, 381–391.

Varlakhanova, N., Cotterman, R., Bradnam, K., Korf, I., and Knoepfler, P.S. (2011). Myc and Miz-1 have coordinate genomic functions including targeting Hox genes in human embryonic stem cells. *Epigenetics Chromatin* 4, 20.

Verdone, L., Caserta, M., and Di Mauro, E. (2005). Role of histone acetylation in the control of gene expression. *Biochem. Cell Biol.* 83, 344–353.

- Virbasius, C.A., Virbasius, J. V, and Scarpulla, R.C. (1993). NRF-1, an activator involved in nuclear-mitochondrial interactions, utilizes a new DNA-binding domain conserved in a family of developmental regulators. *Genes Dev.* 7, 2431–2445.
- Vizcaíno, C., Mansilla, S., and Portugal, J. (2015). Sp1 transcription factor: A long-standing target in cancer chemotherapy. *Pharmacol. Ther.* 152, 111–124.
- Walhout, A.J., Gubbels, J.M., Bernardis, R., van der Vliet, P.C., and Timmers, H.T. (1997). c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene. *Nucleic Acids Res.* 25, 1493–1501.
- Walz, S., Lorenzin, F., Morton, J., Wiese, K.E., von Eyss, B., Herold, S., Rycak, L., Dumay-Odelot, H., Karim, S., Bartkuhn, M., et al. (2014). Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature* 511, 483–487.
- Wang, W., Huang, P., Wu, P., Kong, R., Xu, J., Zhang, L., Yang, Q., Xie, Q., Zhang, L., Zhou, X., et al. (2015). BCL6B expression in hepatocellular carcinoma and its efficacy in the inhibition of liver damage and fibrogenesis. *Oncotarget* 6, 20252–20265.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.
- Wasylyk, B., Hagman, J., and Gutierrez-Hartmann, A. (1998). Ets transcription factors: nuclear effectors of the Ras-MAP-kinase signaling pathway. *Trends Biochem. Sci.* 23, 213–216.
- West, A.C., and Johnstone, R.W. (2014). New and emerging HDAC inhibitors for cancer treatment. *J. Clin. Invest.* 124, 30–39.
- Wiese, K.E., Walz, S., von Eyss, B., Wolf, E., Athineos, D., Sansom, O., and Eilers, M. (2013). The role of MIZ-1 in MYC-dependent tumorigenesis. *Cold Spring Harb. Perspect. Med.* 3, a014290.
- Wu, C., Wong, Y.-C., Elgin, S.C.R., Ashburner, M., Bellard, M., Gannon, F., Chambon, P.,

- Bloom, K.S., Anderson, J.N., Butler, M.J., et al. (1979). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* *16*, 807–814.
- Yang, X.-J. (2004). The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Res.* *32*, 959–976.
- Yardımcı, G.G., Frank, C.L., Crawford, G.E., and Ohler, U. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* *42*, 11865–11878.
- Yin, Y.-W., Jin, H.-J., Zhao, W., Gao, B., Fang, J., Wei, J., Zhang, D.D., Zhang, J., and Fang, D. (2015). The Histone Acetyltransferase GCN5 Expression Is Elevated and Regulated by c-Myc and E2F1 Transcription Factors in Human Colon Cancer. *Gene Expr.* *16*, 187–196.
- Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.* *37*, W247-52.
- Zhang, X., Wen, H., and Shi, X. (2012). Lysine methylation: beyond histones. *Acta Biochim. Biophys. Sin. (Shanghai)*. *44*, 14–27.