ANALYSIS OF PREDICTIVE POWER OF BINDING AFFINITY OF PBM-DERIVED SEQUENCES

RHODES UNIVERSITY

Where leaders learn

A mini-thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

In

Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

By

Lavious Tapiwa Matereke

Department of Biochemistry and Microbiology

January 2015

Supervisor: Prof Philip Machanick

Department of Computer Science, Rhodes University

Abstract

A transcription factor (TF) is a protein that binds to specific DNA sequences as part of the initiation stage of transcription. Various methods of finding these transcription factor binding sites (TFBS) have been developed. *In vivo* technologies analyze DNA binding regions known to have bound to a TF in a living cell. Most widely used *in vivo* methods at the moment are chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) and DNase I hypersensitive sites sequencing. *In vitro* methods derive TFBS based on experiments with TFs and DNA usually in artificial settings or computationally. An example is the Protein Binding Microarray which uses artificially constructed DNA sequences to determine the short sequences that are most likely to bind to a TF. The major drawback of this approach is that binding of TFs *in vivo* is also dependent on other factors such as chromatin accessibility and the presence of cofactors. Therefore TFBS derived from the PBM technique might not resemble the true DNA binding sequences.

In this work, we use PBM data from the UniPROBE motif database, ChIP-seq data and DNase I hypersensitive sites data. Using the Spearman's rank correlation and area under receiver operating characteristic curve, we compare the enrichment scores which the PBM approach assigns to its identified sequences and the frequency of these sequences in likely binding regions and the human genome as a whole. We also use central motif enrichment analysis (CentriMo) to compare the enrichment of UniPROBE motifs with *in vivo* derived motifs (from the JASPAR CORE database) in their respective TF ChIP-seq peak region. CentriMo is applied to 14 TF ChIP-seq peak regions from different cell lines. We aim to establish if there is a relationship between the occurrences of UniPROBE 8-mer patterns in likely binding regions and their enrichment score and how well the *in vitro* derived motifs match *in vivo* binding specificity.

We did not come out with a particular trend showing failure of the PBM approach to predict *in vivo* binding specificity. Our results show Ets1, Hnf4a and Tcf3 show prediction failure by the PBM technique in terms of our Spearman's rank correlation for ChIP-seq data and central motif enrichment analysis. However, the PBM technique also matched the *in vivo* binding specificities of FoxA2, Pou2f2 and Mafk. Failure of the PBM approach was found to be a result of variability in the TF's binding specificity, the presence of cofactors, narrow binding specificity and the presence ubiquitous binding patterns.

Declaration

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from 15 July 2014 to 31 January 2015 under the supervision of Prof Philip Machanick.

I, Lavious Tapiwa Matereke, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature ……..

Date …….

Dedication

This thesis is dedicated to my parents, family members, relatives and friends for the love, encouragement and support.

Acknowledgements

I would like to thank my supervisor and instructor, Prof Philip Machanick for the ideas, advice and support required in this work. I would also like to thank Prof Ozlem Tastan Bishop, our course coordinator, for our good learning environment and making sure that we had all what was required for the course throughout the whole year.

I wish to thank all my MSc classmates for the year 2014 for your moral support and encouragement. I also acknowledge Ngonidzashe Faya, Caleb Kibet and all the RUBi members for company and inspirational talks.

I would like to also thank the Rhodes University Council Research and National Research Foundation (NRF) for the financial support.

Finally, I wish to thank my family, relatives and friends for everything and the one above, God Almighty.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

AlignACE	Aligns Nucleic Acids Conserved Elements
AUROC	Area under Receiver Operating Characteristic Curve
BED	Browser Extensible Data
BEST	Binding-site Estimation Suite of Tools
B-ZIP	Basic Region Leucine Zipper
bp	base pair
CentriMo	Centrality of Motifs
CMEA	Central Motif Enrichment Analysis
ChIP	Chromatin Immunoprecipitation
ChIP-ed TF	Transcription factor used to create binding regions
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
DBD	DNA Binding Domain
DNA	Deoxyribonucleic Acid
DNase I	Deoxyribonuclease I
DREME	Discriminative Regular Expression Motif Enrichment
ENCODE	Encyclopedia of DNA elements
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements-sequencing
GH	Growth Hormone
HESC	Human Embryonic Stem Cell
HLH	Helix-Loop-Helix
HS	Hypersensitive Sites
HTH	Helix-Turn-Helix

HT-SELEX	High Throughput SELEX
Irf4	Interferon regulatory factor 4
MaMF	Mammalian Motif Finder
MEA	Motif Enrichment Analysis
MEME	Multiple EM for Motif Elicitation
mRNA	messenger RNA
PBM	Protein Binding Microarrays
PCR	Polymerase Chain Reaction
PWM	Position Weight Matrix
RNA	Ribonucleic Acid
SpaMo	Spaced Motif Analysis
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SELEX-SAGE	SELEX-Serial Analysis of Gene Expression
TAD	Transcriptional Activation Domain
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSH	Thyroid Stimulating Hormone
UCSC	University of California, Santa Cruz
ZIP	Leucine Zipper

CHAPTER 1

1.1 Introduction

Transcription can be defined in simple terms as the process whereby mRNA is synthesized from DNA by the enzyme RNA polymerase. This is the first step in gene expression and it occurs in three stages; initiation, elongation and termination. The most important stage is initiation whereby RNA polymerase attaches itself to the promoter region of DNA. Elongation is the addition of nucleotides to the 3' end of a growing polypeptide chain with the formation of a covalent bond between two subsequent nucleotides. Recognition of the terminator sequence by RNA polymerase results in the release of mRNA and RNA polymerase distances itself from DNA. Transcription Factors (TFs) are proteins that bind to DNA as part of the initiation stage of transcription. Binding of transcription factors either activates or represses transcription. Activation occurs when binding of the TF facilitates RNA polymerase-DNA binding. Repression occurs when binding of the TF blocks RNA polymerase-DNA binding. Unlike other associated proteins that play a role in transcription, eukaryotic TFs interact with DNA with some sequence specificity, affinity and other factors such as cell type and cellular contexts.

1.2 Transcription Factors and DNA binding

The binding of a transcription factor to a specific DNA sequence results in a subsequent transcriptional activation or repression. The mechanism through which transcription factors and DNA interact involves the structural properties of DNA and other unknown properties (Fukue *et al*., 2005). According to Carey and Smale (2000), a promoter can be defined as a regulatory region of DNA in close proximity to the transcription start site and enhancer is another regulatory region located at a greater distance from the transcription start site, and can be located at either side of the gene locus. Transcription factors basically have three functional domains, each playing a different role which can be transcription activation, repression or a completely different role from the two (Yanagisawa, 2001). The DNA binding domain interacts with specific DNA sequences known as response elements and it helps the transcription activation domain in initiating transcription (Kranthi *et al*., 2009). Transcriptional Activation Domain (TAD) and Transcriptional Repressor Domain are for transcription activation and repression respectively, either directly or through other proteins (co-repressors and co-activators) (Kranthi *et al*., 2009).

1.3 Transcription Activators

Activators are DNA-binding proteins that bind to promoters or enhancers stimulating the transcription of a gene. They stimulate transcription by neutralizing the action of repressors, facilitating changes in chromatin structure or directly interacting with the transcription machinery. Most transcription factor activators are acidic. The three highly acidic amino acids (glutamine, aspartic acid and proline) dominate in their structures (Titz *et al*., 2006) and the activation domain is connected to the DNA binding domain (Lodish *et al*., 2000). Binding of the activator to DNA moves the activator to a position from which the activation domain can facilitate transcription activation (Barberis & Petrascheck, 2003). Eukaryotic activators can be inactivated by binding of another protein to the activation domain. This disrupts interactions necessary for the activation process (Ptashne & Gann, 2002; Barberis & Petrascheck, 2003). Another basic mechanism of inactivating these activators is simply by blocking their movement into the nucleus (Chi *et al*., 2001).

1.4 Transcription Repressors

A wide variety of factors also regulate transcription by inhibiting the transcription of specific genes. Repression can be direct (active) or indirect (passive) and chromatin structure is also said to play a role in the repression mechanism (Gaston & Jayaraman, 2003). Direct or active occurs when certain proteins that contain specific functional domains interact (Cooper & Hausman, 2007). The interaction involves direct contact between the repressor and transcription apparatus (Gaston & Jayaraman, 2003). As a result, an inactive preinitiation complex or no preinitiation complex is formed at all (Hanna-Rose & Hansen, 1996). Indirect or passive repression involves occupation of the binding site by the repressor thus inhibiting the activator from accessing the binding site or the repressor interacting with the activator in solution and preventing its DNA binding. The repressor masks the binding site simply by binding to the same site as the activator but due to lack of the activation domain in the repressor, it fails to activate transcription. The Sp3 factor competes for the same binding site with Sp1. When Sp3 binds to the site it cannot activate transcription; it therefore blocks Sp1 binding site thus preventing activation (Latchman, 2004).

1.5 Transcription Factors and Diseases

According to Villard (2004), genetic disorders in transcription factors are responsible for a number of human diseases. As of 2004, most human developmental disorders were due to mutations in transcription factors that play a role in development (Latchman, 2004). An example is the POU family transcription factor, Pit-1 (GHF-1), responsible for combined

pituitary hormone deficiency (Andersen & Rosenfeld, 1994). Combined pituitary hormone deficiency is lack of the three hormones, thyroid stimulating hormone (TSH), growth hormone (GH) and prolactin, which results in mental and growth retardation (Engelkamp & van Heyningen, 1996). Eye defects can also result from mutations in the genes encoding the Pax family transcription factors (Pax3 and Pax4). In a study by Zheng and Blobel (2010), GATA1 mutations were found to be linked to leukemia and GATA3 was found to inhibit breast cancer. Acute myeloid leukemia has been found to be associated with more than one transcription factor with diagnosed patients expressing GATA1, GATA2 and EKLF (Ayala *et al*., 2012). Alterations in Hnf4a binding sites are also linked to human diseases such as hemophilia (Bolotin *et al*., 2011).

1.6 Problem Statement

Although *in vitro* techniques for finding TF binding sites have found great acceptance and popularity in scientific research, the question of how their models predict *in vivo* binding still needs to be answered. A PWM is a matrix of scores (weights) representing the chances of each base occurring at that position on the binding site and these are usually visualized as sequence logos (Siggers & Gordan, 2014). Although PWMs are most widely used in modeling TF-DNA binding, their assumption that binding positions are independent in practice is not true in several cases (Zhong *et al*., 2013). Therefore using PWMs only may result in loss of information. Weirauch *et al*., (2013) concluded that when testing on PBM data, *k*-mer based models perform better than PWMs. Besides, TF-DNA binding *in vivo* is also governed by other conditions such as chromatin accessibility and the presence of cofactors (Zhong *et al*., 2013; Spitz & Furlong, 2012).

In vivo technologies such as ChIP-chip and ChIP-seq are highly dependent on the antibody used which needs to be specific to the epitope and the procedure is only applicable for one TF per experiment. The procedures for these techniques are discussed in section 2.5 of this work. Although these methods are widely used, modelling TF-DNA binding using their data only can be tricky since some TFs bind to DNA in the presence of cofactors or first bind to another TF, then the complex binds to DNA (cooperative or indirect binding). *In vitro* data derived from the Protein Binding Microarray (PBM) technique can be found from the UniPROBE motif database (Bulyk & Newburger, 2009). UniPROBE motifs usually represent transcription factor binding sites (TFBS) but it is not clear whether motifs over-represented in likely binding regions represent true binding sites or they are just there for support during cooperative binding or indirect binding or they do not participate in binding at all

(background). Each TF in the UniPROBE motif database also contains 8-mer patterns with their enrichment scores. The enrichment score (E-score) was computed using the seed-and-wobble algorithm and it is a measure of the binding specificity. The value ranges from -0.5 to 0.5 with 0.5 indicating a high binding specificity. We seek to determine whether the PBM approach scores these 8-mer patterns highly in some cases where they have low relative prevalence in likely binding regions or vice-versa by checking for the occurrence of these patterns in likely binding regions (DNase I hypersensitive sites and ChIP-seq peak regions) as well as in the whole genome (human genome).

1.7 Research Question

Does the seed-and-wobble algorithm used in PBM approach score sequences highly in some cases where they have low relative prevalence in likely binding regions or vice versa?

1.8 Aims and Objectives

We expect that *k*-mers that are over-represented in DNase data represent background rather than binding affinity, since DNase data is not specific to a particular TF. However it is also possible that some TFs have ubiquitous binding, and separating out these two cases is a challenge; we start with the following objectives:

1. To determine whether PBM scoring generally correlates with the probability that a *k*-mer participates in binding
2. To compare the PBM E-scores for *k*-mers with the relative prevalence of *k*-mers in likely binding regions
3. To gain an insight on how the difference between the E-score and DNase or ChIP-seq based frequency of that same *k*-mer is biologically relevant using literature
4. To evaluate the accuracy and performance of the seed-and-wobble algorithm used by the PBM technique in finding transcription factor binding motifs based on *k*-mer scores as compared to other currently used tools

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Understanding transcription factors-DNA interactions has always been complicated due to the DNA structure and transcription factors' binding specificity and affinity. In our work, affinity refers to the extent to which a transcription factor is attracted to its binding site and specificity is how well the transcription factor distinguishes its binding site from a wide range of potential binding sites. Transcription is depended on other environmental factors such as the cooperative binding of transcription factors and chromatin structure. Identification of transcription factor binding sites on the genome therefore becomes a challenge since most experiments cannot accommodate all these factors. The accuracy of these experiments (*in vitro* methods) in predicting transcription binding sites remains a question. In this chapter, we discuss the experiments used in predicting transcription factor binding sites, their advantages and disadvantages as compared to each other and the most widely used approach. We also describe in short how chromatin structure affects transcription factor-DNA binding. The chapter also contains information about how TFBS can be modeled using a number of different computational approaches and where the data can be found.

2.2 Chromatin structure and Transcription

Chromatin is an association of DNA and histone proteins and it has been found to have major effects on gene expression. Negatively charged DNA is attracted to positively charged histone proteins thus a strongly tight bond is formed. Chromatin's basic unit, a nucleosome consists of DNA (147bp long) which coils around the histone proteins (Cooper & Hausman, 2007). Other than histone proteins, chromatin also associates with several other proteins including highly mobile group proteins (Carey & Smale, 2000). The result is that the transcriptional machinery is inaccessible to transcription factors and RNA polymerase. In order for gene expression to occur, changes in chromatin structure (chromatin remodeling) must occur and these are histone methylation, acetylation or phosphorylation (Ralston & Brown, 2008). Histone modification by acetylation has been found to play a major role in chromatin remodeling. During acetylation, the enzyme histone acetyltransferase attaches acetyl groups (--$COCH_3$) from acetyl-coenzymeA to the amino terminal tails of histone proteins (Brooker *et al*., 2010). This weakens the positive charge of histones resulting in

relaxed chromatin since this also changes the overall structure of nucleosome (Sterner & Berger, 2000). Chromatin remodelers also have a higher affinity for acetylated histone tails than non-acetylated ones. Both transcriptional activators and repressors target histone acetylation. Other methods of histone modifications are phosphorylation of lysine residues, methylation of lysine and arginine residues and the addition of ubiquitin to lysine residues. Some protein complexes also enable chromatin remodeling without removing or covalently modifying the histones.

2.3 Transcription Factors and DNA binding specificity

Most motif discovery tools use position weight matrices (PWMs) to model TF-DNA binding specificity (Slattery *et al*., 2014). However, the assumption of PWMs that positions are independent is not always the case (Zhong *et al*., 2013) and some positions have been found to be interdependent. Therefore using PWMs only may result in loss of information. With TFs having different DNA binding modes, their binding specificities vary and these can be represented easily using *k*-mers (Badis *et al*., 2009). Zhong *et al*., (2013) developed a *k*-mer based method for analyzing PBM data.

2.4 Determining Transcription Factor Binding Sites

Different technologies have been developed to study transcription factor binding sites across the genome. *In vivo* technologies analyse DNA binding regions known to have bound to a TF in a living cell. These include ChIP-chip (Lieb *et al*., 2001), ChIP-seq (Robertson *et al*., 2007) and the recent ChIP-exo (Serandour *et al*., 2013). Orenstein and Shamir (2014) have pointed out that *in vivo* binding is affected by additional factors such as chromatin structure, nucleosome positioning and cofactors. *In vitro* methods derive DNA binding sites based on experiments with TFs and DNA usually in artificial settings or computationally (in silico). An example is the protein binding microarray (PBM) (Berger & Bulyk, 2009) which uses artificially constructed DNA sequences to determine short DNA sequences that are most likely to bind to a TF. The major drawback to this approach is that the artificially constructed DNA sequences may not resemble the available DNA sequences with high frequency for binding. It is also not clearly understood how accurate the models derived from *in vitro* technologies are in predicting in vivo binding (Orenstein & Shamir, 2014).

2.5 *In vivo* Technologies

2.5.1 ChIP-chip

Chromatin immunoprecipitation combined with DNA microarrays is one of the first discovered methods used to investigate *in vivo* protein-DNA interactions. According to Buck and Liab (2004), the first successful ChIP-chip technique identified binding sites for individual TFs in *Saccharomyces cerevisiae*. The first step in this technique is *in vivo* fixation of protein-DNA interactions with a crosslinker. The most widely used crosslinker is formaldehyde which has the ability to diffuse across cell membranes. Formaldehyde crosslinks proteins to each other primarily between the amino group of lysine residues and an adjacent peptide bond (Buck & Lieb, 2004). DNA is then subjected to sonication which shears the chromatin to small fragments.

Immunoprecipitation of the extract with a specific antibody (usually coupled to either agarose or magnetic beads) against the protein of interest follows (Waldiminghaus & Skarstad, 2010). After the purification step, DNA amplification (PCR-based are widely used) usually follows due to low DNA yields from immunoprecipitation (Zeitlinger *et al*., 2007). Quantified DNA is fluorescently labelled for hybridization. Total DNA before immunoprecipitation is also labelled with a different dye, a fluorescent one for control purposes. After hybridization, a genome-wide map of protein-DNA interactions can be constructed since the location of background DNA oligomers is known. Increasing costs, its failure to analyse half of the human genome and detection of false positive target sites saw the technique being overtaken by ChIP-seq (Gilfillan *et al*., 2012). ChIP-seq also offers higher resolution and much more accurate locality (plus or minus 50 bp if it works well as compared to 1000 bp for ChIP-chip) (Park 2009; Gilfillan *et al*., 2012)

2.5.2 ChIP-seq

ChIP-seq is similar to ChIP-chip but differs in combining deep sequencing with chromatin immunoprecipitation instead of hybridization of DNA fragments in an array as done in ChIP-chip. ChIP-seq uses next generation sequencing and employs computational mapping of the sequenced DNA which then identifies the genomic locations of bound DNA-protein interactions (Bailey *et al*., 2013). Bailey *et al*., (2013) also point the reduction of background signals, high resolution and high genomic coverage as the advantages which led to the widespread use of ChIP-seq over ChIP-chip. ChIP-seq experiments for more than one hundred and forty different TFs across different cell types and organisms have been performed by the ENCODE and modENCODE consortia (ENCODE 2004, 2012). A major

drawback of this technique is high cost (machine depreciation and reagent cost) and availability (Park 2009). Several different algorithms have been developed for analysis of ChIP-seq data. Sequencing errors occur towards the end of each read, but the problem can be solved using computational analysis. There is also a bias towards GC rich regions in fragment selection (Park 2009). The percentage or proportion of Guanine and Cytosine bases could be associated with the TF's functionality leading to a bias in the generation of reads.

2.5.3 ChIP-exo

ChIP-exonuclease (ChIP-exo) is a modification of ChIP-seq with improved resolution of binding sites. Although Bailey *et al*., (2013) pointed out high resolution as one of the advantages of ChIP-seq over ChIP-chip, recent studies (Serandour *et al*., 2013) reveal that this resolution is not adequate for modelling TF binding specificity using different motifs within binding sites. The protocol for ChIP-exo is similar to that of ChIP-seq except that before sequencing, protein-bound DNA is degraded by an exonuclease (enzyme that degrades DNA in 5'- 3' direction), primer extended and ligated with a sequencing adaptor (Rhee & Pugh, 2012). Figure 2-1 illustrates the major steps for carrying out a ChIP-exo experiment.

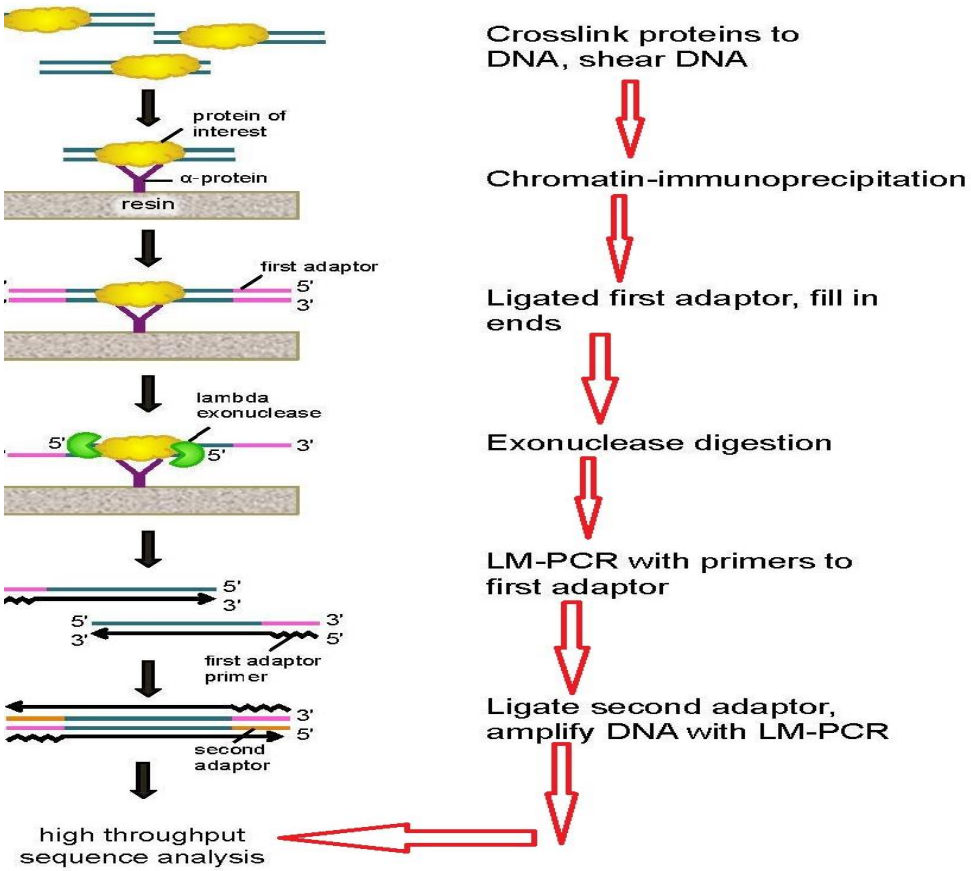

Figure 2-1: Schematic outline of the ChIP-exo method

(*http://commons.wikimedia.org/wiki/File:ChIP-exo_process_diagram.pdf*)

This technique is time-consuming. According to Rhee and Pugh (2012), the protocol can be completed in three and half days. The experimental method has many complicated steps that may fail for technical reasons.

2.5.4 Finding open chromatin sites

DNase I hypersensitive sites can be defined as DNA regions that are nucleosome free, also described as open chromatin sites that can be identified after digestion of DNA by DNA nuclease I (DNase I) which prefers nucleosome depleted regions of DNA (Boyle *at al*., 2008). Nucleosome packaging organises DNA in such a way that enables or hinders protein binding (including TFs). Experimental protocols for determining open chromatin sites are DNase I hypersensitive sites sequencing (DNase-seq) and Formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq). DNase-seq involves DNase I digestion DNA followed by high-throughput sequencing to identify open chromatin sites *in vivo* in a given cell type (Boyle *et al*., 2011). The basis of the experiment is that transcription regulatory sites are located within open chromatin sites while the closed chromatin sites (DNA wrapped around histone proteins) are protected from DNase I cleavage. TF-bound locations are also protected from DNase I. Thus DNase I footprints give an idea of the TF binding sites. A footprint is a short site (8 bp) with little or no DNase cleavage sandwiched by DNase I cleaved sites. As compared to ChIP-seq, DNase-seq has the ability of mapping binding sites of all TFs, very high nucleotide resolution and it doesn't require high quality expensive antibodies. Formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq) also identifies open chromatin regions based on the differences in cross-linking of DNA with formaldehyde. DNA is cross-linked with formaldehyde in order to bind proteins, therefore protein-unbound DNA reflects failure to cross-link with formaldehyde. According to Boyle *et al*., (2008), Faire-seq has been shown to be highly associated with DNase hypersensitive sites and other chromatin marks due to its successfulness on multiple eukaryotic cells and tissues.

A number of limitations have been found to be associated with using either DNase-seq or Faire-seq. Krajewski and Madrigal (2012) pointed to the need for Faire-seq users to test their experiment using a proper statistical model. DNase I has been shown to be sequence biased, when cutting DNA: it prefers some specific sequences as compared to others. Hence the cleavage patterns can be similar yet they are expected to differ according to whether DNA is protein bound or naked. The major weakness is that we cannot identify what has been bound

to these open chromatin sites; neither can you infer the function of the bound protein using data from both techniques. It is for this reason that DNase footprinting studies are combined with ChIP-seq data which can provide the precise positions of TF binding sites.

2.6 The ENCODE Project

The Encyclopedia of DNA Elements (ENCODE) project is an international consortium of investigators funded to analyse the human genome with an aim of creating a database of functional elements (Rosenbloom *et al*., 2010). It began as a small scale project in 2003 focusing on a small percentage (1%) of the human genome but it has grown exponentially focusing on other species such as *mus musculus* (ENCODE, 2012). The project is aimed at identifying functional elements (proteins or non-coding RNA) in the human genome using quite a number of experimental approaches (ENCODE, 2004). Of the 1800 known TFs, ENCODE had sampled 119 by 2012 involving 147 different cell types and they hope to enlarge their analysis using additional factors and other cell types. In our study, we chose to use ChIP-seq and DNase-seq data because the two techniques outperform their rivals (ChIP-chip and FAIRE-seq) in terms of the current number of experiments performed by the ENCODE employing these techniques (Table 2-1).

Table 2-1: Summary of some of the ENCODE datasets as of December 2014 (extracted from the ENCODE UCSC website)

Source = (*https://www.encodeproject.org/search/?type=experiment*)

Data Type	Description	Number of Experiments
ChIP-seq	TF and polymerase binding, histone marks by ChIP	2466
DNase-seq	DNase I digestion of DNA followed by sequencing	265
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements and sequencing	37
RNA-seq	RNA sequencing	695
Methy-seq	DNA methylation by restriction enzymes and sequencing	103

2.7 *In Vitro* Technologies

2.7.1 Protein Binding Microarrays (PBM)

PBM is an *in vitro* technology which determines binding specificities of individual DNA binding proteins using artificial DNA sequences confined to microarrays with potential DNA

binding sites. Summarizing the procedure as described by Bulyk (2006), arrays containing double-stranded DNA are incubated with epitope tagged purified transcription factor. Protein-bound DNA is then washed to remove non-specifically bound transcription factor. The protein-bound array is labelled with a fluorophore-conjugated antibody specific to the epitope, and the fluorescence provides an estimate of the TF-DNA binding affinity (Berger & Bulyk, 2009). Binding specificity is commonly modelled using position weight matrix (PWM) and statistical analysis is used to assess the likelihood of that motif being the DNA binding motif of the given protein (Bulyk, 2007). Berger and Bulyk (2009) also described another distinct representation of the DNA binding specificity of a TF, a table for the relative preferences for all *k-mers* (a *k-mer* is a short DNA sequence of length k). This gives a full picture of TF binding, since it involves both low and high affinity sequences. DNA binding affinity is measured using enrichment scores derived from PBMs for these *k-mers*. PWMs provide a complete knowledge of a TF's binding specificity and most motif discovery tools allow the use of PWMs as input (Berger and Bulyk, 2009). According to Orenstein and Shamir (2014), although there is a widespread use of PBMs currently, it is unclear how accurate these models are in predicting *in vivo* binding. Several studies have shown that using PWMs to predict *in vivo* binding can be misleading since TF-DNA binding in a living tissue also depends on other different factors (Orenstein *et al*., 2012; Weirauch *et al*., 2013; Orenstein & Shamir, 2014).

2.7.2 SELEX (Systematic Evolution of Ligands by Exponential Enrichment)

SELEX is an *in vitro* strategy to study properties of DNA molecules that bind proteins. A purified TF is added to a library of DNA sites. After incubation, bound sites are separated from unbound ones by various means such as gel filtration, followed by amplification of the bound sites to allow rebinding and this stage can be repeated until a significant quantity of the bound sites is reached (Stormo & Zhao, 2010). The first modification of the standard SELEX combines ligation of sites and sequencing, known as SELEX-serial analysis of gene expression (SELEX-SAGE) but it is time-consuming (Roulet *et al*., 2002; Stormo & Zhao, 2010). A newer technology, high throughput SELEX (HT-SELEX) further sequences a sample of the amplified bound sites and feeds them to the next cycle (Orenstein & Shamir, 2014).

2.8 The UniProbe Motif Database

The Universal PBM Resource for Oligonucleotide-Binding Evaluation (UniProbe) (Bulyk & Newburger, 2009) is an online database that contains *in vitro* data derived from universal

PBM technology. Initially, the database was only centralized on providing PBM data in the form of possible *k-mers*, PWMs and graphical sequence logos of the *k-mer* data (Bulyk & Newburger 2009). For each TF in the UniProbe database, 8-mer patterns are provided each with its respective enrichment score, derived from a test statistic (Wilcoxon-Mann-Whitney) which ranges from -0.5 (least favoured k-mer) to 0.5 (most favoured k-mer) (Bulyk & Newburger, 2009). According to Bulyk and Robasky (2011), the updated version of the same database now has more than one hundred and thirty percent expansion of the initial database content, including a protein blast (BLASTp) tool which searches for protein sequence homologs in a protein database using a protein sequence as a query and the introduction of UniProbe Accession numbers. The database also contains data for many proteins not included in other databases such as the curated JASPAR (Bryne *et al*., 2008) and TRANSFAC (Matys *et al*., 2003). The database also has the Tomtom (Gupta *et al*., 2007) program, an analysis tool which compares motifs against PWMs in the database. Another feature on the database allows the user to enter a FASTA file containing up to thirty DNA sequences and it scans user-supplied input DNA sequences for TF binding sites (Bulyk & Newburger, 2009).

2.9 Motif Discovery, Enrichment and Analysis Tools

Finding TFBSs is an expensive and time consuming process. Many different computational approaches for the identification motifs in biological sequences have been developed but correct prediction of these eukaryotic TFBSs has become a great challenge in computational biology. This has led to the development of numerous tools for this task with different definitions of motifs, the correct way of representing a motif statistically and to find statistically overrepresented motifs (Tompa *et al*., 2005). Below are some of the tools used and some short descriptions of them.

2.9.1 AlignACE

AlignACE (Aligns Nucleic Acid Conserved Elements) finds sequence elements conserved in a set of DNA sequences using a Gibbs sampling strategy and an iteration procedure (Hughes *et al*., 2000). Previously, it was used to find DNA motifs in *Saccharomyces cerevisiae* genome (Hughes *et al*., 2000, Roth *et al*., 1998). AlignACE also enables the user to quickly determine whether the resulting hypothetical motifs have homologs or not. It uses a motif comparison algorithm to search for motifs that are similar to the query motif. Fu and Weng (2005) compared AlignACE with GLAM (Frith *et al*., 2004), MEME (Bailey & Elkan, 1994) and Motifsampler (Thijs *et al*., 2001) in constructing PWMs using TFBS data from the

TRANSFAC database. Their results showed that Motifsampler, MEME and AlignACE produced worse PWMs compared to those from TRANSFAC.

2.9.2 W-AlignACE

W-AlignACE (Chen *et al*., 2008b) is a modified version of AlignACE, which uses the same algorithm as AlignACE (Gibbs sampling) which takes into consideration the strength of binding sites' binding ratios in ChIP experiments thus true motifs have strong binding ratios (Chen *et al*., 2008b). Kurata *et al*., (2013) reported the use of W-AlignACE in identifying ModE-binding consensus sequences. Yamamoto *et al*., (2011) also used W-AlignACE to identify MntR binding sequences and it identified 26 motifs, a result which matched the previously proposed binding sites according to the literature.

2.9.3 Consensus

Consensus (Hertz & Stormo, 1999) models motifs using position weight matrices, targeting those with the highest information content (Tompa *et al*., 2005). According to the authors, it generates a motif with the greatest information content by pairing sequences with more informative motifs. Hon and Jain (2006) compared their MaMF (Mammalian Motif Finder) algorithm with Consensus and AlignACE in predicting motifs using TRANSFAC dataset. MaMF performed better than both AlignACE and Consensus. AlignACE found no motifs at all and Consensus found one out of eight motifs. Che *et al*., (2005) incooperated Consensus with AlignACE and MEME in their Binding-site Estimation Suite of Tools (BEST) which compares different motif finders for TFBS prediction.

2.9.4 MotifSampler

MotifSampler (Thijs *et al.*, 2001) identifies motifs and conserved region in DNA or protein sequences using a Gibbs sampling approach and a higher order Markov model (background) for comparison (Tompa *et al*., 2005). Their Gibbs sampling method assumes the lengthy of the motif is known and the motif occurs once (one real instance) in each input sequence. According to the authors, the higher order background model strengthened the performance of their algorithm in the presence of highly noisy data. MotifSampler together with MEME, Consensus, GLAM and AlignACE is part of the Toolbox of Motif Discovery (Tmod) (Sun *et al*., 2010), which has 12 integrated motif discovery programs. Fauteux *et al*., (2008) compared the performance of their Seeder (discriminative seeding DNA motif discovery) algorithm with MotifSampler and MEME but surprisingly they did not show how MotifSampler performed.

2.9.5 MEME

MEME (Multiple EM for Motif Elicitation) (Bailey & Elkan, 1994) uses expectation-maximization (EM) which is also a statistical iterative method. The algorithm estimates the occurrence of each motif in a single sequence in the dataset, aligns the motifs and outputs the alignment (Bailey & Elkan, 1994). Candidate motifs returned by each EM run are ranked according to their E-values (Tanaka *et al*., 2014). However, MEME has been found to be slow on very large sequence sets. There are high chances that true motifs can be rejected due to MEME's approximation of the E-value (Nagarajan *et al*., 2005, Tanaka *et al*., 2014). It is for this reason that Tanaka *et al*., (2014) proposed a "two-tiered significance analysis" as a substitute for MEME's E-value to select the best EM-generated motifs as well as to assign an overall statistical significance so as to improve MEME's performance.

2.9.6 DREME

Due to the slow speed of MEME on very large sequence data sets, Bailey (2011) presented DREME (Discriminative Regular Expression Motif Enrichment), another motif discovery algorithm which can find short motifs (4-8bp) of eukaryotic TFs and analyze very large ChIP-seq datasets at a much faster rate. It uses regular expressions to create a pattern out of words that occur often and refines the regular expressions to find those most statistically significant. The user enters DNA sequences and a significance threshold. According to the author, DREME successfully discovered motifs and many cofactors in mouse embryonic stem (ES) cells, mouse erythrocyte and human cell lines data sets among other used data sets.

2.9.7 TOMTOM

TOMTOM (Gupta *et al*., 2007) is a motif comparison algorithm that finds motif homologs after thorough searching in a database of known motifs. For every match between two motifs, TOMTOM allocates a numerical score and its statistical significance. TOMTOM also outputs logos, showing the alignment between the query and target motif.

2.9.8 CentriMo

CentriMo (Bailey & Machanick, 2012) is a motif enrichment analysis and visualization tool which measures central enrichment of motifs in ChIP-seq peak regions. The user inputs genomic sequences of equal length and selects motif databases: usually UniPROBE and JASPAR motifs are used. The length of these genomic sequences is of great importance. Longer genomic sequences more likely include bound sites and also enhance an effective binomial enrichment test. CentriMo plots a site probability curve indicating the probability of the occurrence of a predicted binding site at that position and the width of the centrally

enriched region for each motif according to a statistical test (P-value). A very sharp peak at the center of the graph with a narrow window (<100bp) and a very low p-value (10^{-1000} or even less) indicates direct binding by a single factor or it's a likely binding motif. Less sharp peaks (in site-probability curves) indicate cooperative binding while lack of distinct peaks can be a result of failure of the ChIP-seq experiment or indirect binding. The web-based CentriMo algorithm now performs differential motif enrichment analysis, a new methodology for identifying sequence motifs that are differentially enriched in one set of DNA or RNA sequences relative to another set (Lesluyes *et al*., 2014). The later version of CentriMo is capable of finding motifs in any region of ChIP-seq peak sequences as well as measuring their enrichment with respect to another set of sequences.

CHAPTER 3

MATERIALS AND METHODS

3.1 Introduction

In this chapter, we talk about the methods and data used in carrying out our investigations. We employ two statistical/mathematical approaches (Spearman's rank correlation and area under receiver operating characteristic curve) and central motif enrichment analysis tool, CentriMo. Using the statistical methods, we check if there is a relationship between UniPROBE 8-mer enrichment scores and the occurrences of the 8-mer patterns in likely binding and nonbinding regions. We compare the results from our different methods in terms of their authenticity and do the same again for the different data sets. Using CentriMo, we compare the enrichment of UniPROBE and JASPAR motifs for a TF in its respective ChIP-seq peak region. We also take into consideration whether binding is direct, indirect or cooperative according to the CentriMo site probability curves.

3.2 Summary of methods

Our method searches for the occurrences of PBM 8-mers in open chromatin sites, *in vivo* ChIP-seq peaks and in the human genome as a whole. The procedure counts the single occurrence of each 8-mer in each DNA sequence. To evaluate this method, we used 14 TFs for which both ChIP-seq and PBM data are available. For each PBM data set, only data from 8-mers top enrichment text files were used. The file contains top enriched 8-mers with an E-score threshold above 0.25. From these files, the best 50 scoring 8-mers usually with an E-score threshold above 0.49 and the lowest 50 scoring 8-mers were selected. Using Linux commands and bash scripting, we counted the occurrence of these 8-mers in ClusteredDNase data, human genome and in the ChIP-seq peak of that same TF downloaded from ENCODE. We compared the E-scores for these 8-mers and their occurrence (counts) in those four different datasets using two different approaches, Spearman rank's correlation and area under receiver operating characteristic (AUROC) as these have been used in evaluating a number of motif discovery algorithms. If the *in vitro* PBM technique matches *in vivo* binding specificity, we expect to see significant correlations between the 8-mer E-scores and 8-mer counts in

likely binding regions. We also expect higher AUROC when using ClusteredDNase and single ChIP-seq peak data since these are more representative of *in vivo* binding as compared to human genome counts. We also use central motif enrichment analysis (CentriMo). Our interpretation is that if *in vitro* motifs predict *in vivo* binding well, then UniPROBE motifs will be at least as well centrally enriched with highly significant p-values and narrower centrally enriched region widths as compared to JASPAR CORE motifs. We also study which comparison method is better between the Spearman's rank correlation and AUROC although a much broader comparison of the data is required for concluding the results.

3.3 ChIP-seq data: ENCODE

A total of 10 TFs ChIP-seq peaks were downloaded from the ENCODE project plus four mouse embryonic stem cells from the Chen *et al*., (2008a) data sets (Esrrb, Klf4, Oct4 and Sox2). The most studied cell lines, GM12878 and K562, were given the first preference, but peaks for all the 14 TFs could not be found within those two cell lines only so we resorted to other cell lines as well. Most of these were from the HudsonAlpha and Stanford Labs. For preparation, we extracted 500 bp sequences centered on each peak in FASTA format from the human genome version 19 (Hg19) using the FastaFromBed and Bed-widen tools. The data sets used in this research are freely available on the ENCODE project webpage. Table 3-1 gives a brief summary of the ChIP-seq data used and Table 3-2 gives little information on the selected cell lines.

Table 3-1: ChIP-seq peaks downloaded from ENCODE. *(*Source = *http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform)*

TF	Cell line	Treatment
Hnf4a	HepG2	None
FoxA2	HepG2	None
Mafk	H1-HESC	None
Sp4	H1-HESC	None
Gata3	T-47D	DMSO_0.02pct
Ets1	K562	None
Irf4	GM12878	None
Max	GM12878	None
Tcf3	GM12878	None
Pou2f2	GM12878	None

3.4 DNaseI Hypersensitivity Clusters and the Human Genome: ENCODE

DNase clusters (version 3) and human genome (hg19) were downloaded from the ENCODE project. FASTA sequences were obtained from these two data sets using the FastaFromBed tool. For hg19, the whole genome was used and this consists of both coding and noncoding parts. DNase clusters can be described as combined open chromatin sites across 125 ENCODE cell types based on DNase-seq data. According to the ENCODE (2014), the number of cell types has been increased from the initial 41 cell types. The rationale for using DNase-seq is that TFBS are found within open chromatin sites. All these data sets are also freely available on the ENCODE project webpage.

Table 3-2: Variations across selected cell lines.

Source = (*http://genome.ucsc.edu/ENCODE/cellTypes.html*)

Cell line	Description	Lineage	Tissue	Karyotype
GM12878	lymphoblastoid	Mesoderm	Blood	Normal
K562	Leukemia	Inner cell mass	Blood	Cancer
H1-HESC	Embryonic stem cells	Inner cell mass	Embryonic stem cell	Normal
HepG2	Hepatocellular carcinoma	Endoderm	Liver	Cancer
T-47D	Epithelial cell line derived from mammary ductal carcinoma		Breast	Cancer

3.5 PBM data: UniPROBE

We downloaded PBM data for 14 mouse TFs from the UniPROBE database which contains *in vitro* DNA-binding specificity data. From each PBM data set, we extracted the 8-mer top enrichment text file. Enrichment scores (E-scores) are a measure of the binding specificity and usually 8-mers with E-score greater than 0.49 represent high binding specificity thus we selected the first 50 from the original PBM data file. Those with an E-score less than 0.35 are generally considered to have a low specificity according to Jiang *et al* (2013). We also selected the last 50 8-mers from the original PBM data file with an E-score of at least 0.25. Where both version 1 and 2 were available, the latter was used.

Table 3-3: PBM data from the UniPROBE motif database

Source = (*http://the_brain.bwh.harvard.edu/uniprobe/*)

Protein/TF	**UniPROBE Accession Number**	**Domain**
Esrra	UP00079	ZnF_C4
Ets1	UP00414	ETS
FoxA2	UP00073	Fork_head
Gata3	UP00032	GATA
Hnf4a	UP00066	ZnF_C4
Irf4	UP00018	IRF
Klf7	UP00093	Zf-C2H2
Mafk	UP00044	BRLZ
Max	UP00060	HLH
Pou2f1	UP00254	Homeo, POU
Pou2f2	UP00191	Homeo, POU
Sox12	UP00101	HMG_box
Sp4	UP00002	Zf-C2H2
Tcf3	UP00058	HMG_box

3.6 CentriMo

CentriMo takes as input equal length DNA sequences and then finds the best site (of known length) in each sequence that matches a given motif. CentriMo adds the number of sequences where the best binding site occurs at the same position with respect to the 5'-end of the sequences. The quotient from the number of sequences with the binding site at the same position and the total number of sequences is the estimated probability of the best binding site occurring at that position in a ChIP-seq peak region. CentriMo finally plots a site-probability curve with the width of the centrally enriched region for each motif according to a statistical test (P-value). A very sharp peak at the center of the graph with a narrow window (<100bp) and a very low p-value (10^{-1000} or less) shows that there is high likelihood of the motif in consideration to be the true binding motif. The CentriMo algorithm is capable of performing more than one type of motif enrichment analysis (Lesluyes *et al*., 2014). The user is also now free to display motifs of his or her choice, together with their motif logos and to create quality

images of the plots. In this research, we use UniPROBE and JASPAR CORE motifs in CentriMo to compare their enrichment in the respective TF ChIP-seq peak regions. We expect that if *in vitro* motifs predict *in vivo* binding well, then UniPROBE motifs will be more centrally enriched with highly significant p-values and shorter region widths as compared to JASPAR CORE motifs.

3.7 Area under Receiver Operating Characteristic Curve (AUROC)

The AUROC is a method used to measure how good a classifier is (Fawcett 2005). It is a measure of how well the classifier can distinguish two data sets. The predictions are classified into true positives and false positives. From these two, true positive rate (true positives divided by positives) also called sensitivity and false positive rate (false positives divided by negatives) can be calculated (Primer 2005). The receiver operating characteristic curve is a two-dimensional plot of the true positive rate on the y-axis against the false positive rate on the x-axis. The area under this curve is the AUROC, which is the probability of a true positive having a higher score than a true negative (Hanczar *et al*., 2010). One is the maximum value for the AUROC, indicating a perfect discrimination of data and an area of 0.5 indicates no discrimination (50% sensitive and 50% specific), represented by a straight diagonal line from left corner to top right corner (Fawcett 2005). In this study, AUROC is used to measure our classification of the highest 50 scoring UniPROBE 8-mers and lowest 50 scoring 8-mers according to their counts in DNase I Hypersensitivity Clusters, Human Genome (Hg19) and ChIP-seq peak regions.

3.8 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is a non-parametric measure (makes no assumption on the population distribution or sample size) of the strength of a relationship between paired data (Pianadosi *et al*., 2007). The paired data need to be monotonically related. A coefficient of one reflects a perfect positive monotonic relationship and negative one reflects a negative monotonic relationship (Bonett & Wright, 2000). When applying the Spearman's rank correlation, the null hypotheses is that there is no correlation (relationship) between the samples. In our research, using R-studio, we use the Spearman's rank correlation coefficient to check if there is a relationship between UniPROBE 8-mers and their frequency (occurrence) in DNase I Hypersensitivity Clusters, Human Genome (Hg19) and ChIP-seq peak regions.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

As described briefly in 3.8, the Spearman's rank correlation coefficient is a value that indicates or is used to measure how paired data are related monotonically with the null hypothesis that there is no correlation (relationship) between the paired data. UniPROBE 8-mer patterns are provided with their enrichment scores. This enrichment score is an indicator of the TF's binding affinity and specificity towards that pattern as compared to other 8-mers. Therefore the assumption when using PBM-derived data (*in vitro*) is that there is a relationship between the enrichment score and binding affinity. We expected large correlation coefficients between the UniPROBE 8-mer enrichment scores and their occurrence in likely binding regions (DNase I hypersensitive sites and their respective ChIP-seq peaks) as compared to their occurrence in the whole genome (Hg19).

The AUROC is a method used to measure how good a classifier is in discriminating two data sets. After allocating a score to each datum, the binary classification predicts the true positives and false positives using a varying threshold. One is the maximum value for the AUROC, indicating a perfect discrimination of data and an area of 0.5 indicates no discrimination (50% sensitive and 50% specific). In this research, AUROC was used to measure the performance of our classification for the highest 50 scoring UniPROBE 8-mers (E-score approximately equal to 0.5) and lowest 50 scoring 8-mers (E-score approximately equal to 0.25) according to their counts in DNase I Hypersensitivity Clusters, Human Genome (Hg19) and ChIP-seq peak regions. We consider AUROC greater than or equal to 0.8 as a high quality discrimination and anything close or equal to 0.5 as no discrimination.

We used CentriMo to find out which motifs are centrally enriched within the TFs' ChIP-seq peak regions between UniPROBE and JASPAR CORE motifs. UniPROBE motifs are derived from universal PBM technology. JASPAR database contains curated motifs derived from *in vitro* and *in vivo* experiments (ChIP-seq, Chip-chip experiments, PBM and SELEX). This chapter focuses more on the correlations using ChIP-seq data since it is specific. With ChIP-seq data, we know the bound TF unlike DNase I hypersensitive sites which are likely binding regions but the bound protein is not known. We also make references to correlations for ClusteredDNase and Hg19 data and the AUROC for all the three data sets. Our

interpretation is that if the UniPROBE 8-mer patterns match *in vivo* specificity, then a significant positive correlation for ChIP-seq data should be accompanied by more central enrichment of UniPROBE motifs with highly significant p-values and shorter region widths as compared to JASPAR motifs in the TF's respective ChIP-seq peak region. A significant negative correlation therefore means the opposite is true. We also compare the nucleotide content of the highly scored 8-mer pattern and the centrally enriched motifs. If there is no similarity we assume the PBM technique might have failed to match *in vivo* specificity but the reasons are yet to be found.

4.2 Only two negative correlation coefficients: ChIP-seq data

Table 4-1: Spearman's rank correlation between UniPROBE 8-mer E-scores and the 8-mer counts in the TF's ChIP-seq peak

TF	Correlation Coefficient	P-value
Esrra	-0.736	$2.2e^{-16}$
Ets1	0.626	$3.213e^{-12}$
FoxA2	0.709	$2.2e^{-16}$
Gata3	0.626	$4.161e^{-12}$
Hnf4a	0.395	$4.713e^{05}$
Irf4	-0.239	0.017
Klf7	0.752	$2.2e^{-16}$
Mafk	0.763	$2.2e^{-16}$
Max	0.628	$2.749e^{-12}$
Pou2f1	0.293	0.003
Pou2f2	0.421	$1.299e^{-05}$
Sox12	0.614	$1.122e^{-11}$
Sp4	0.757	$2.2e^{-16}$
Tcf3	0.284	0.004

In our study, the sample size was quite large (100) so we assume 0.5 and above is a large correlation, 0.3 is intermediate and small is anything smaller than or equal to 0.1. As expected, positive significant correlations were observed in 12 out of the 14 TFs used (Table 4-1). Using the counts only (see appendix for results), most of the 8-mer patterns with high enrichment scores (E-score => 0.49) occurred more frequently in the ChIP-seq peak of that

same TF than the ones with low enrichment scores. Although it is tempting to say that the most frequent 8-mer is the true TF binding site, this is not always the case. It can be some sequence necessary for TF-DNA binding (background). For example, A-T regions are necessary in the vicinity of binding. Esrra and Irf4 have negative correlation coefficients but these are also highly significant with p-values as low as $2.2e^{-16}$. We will discuss the results for this section together with the CentriMo results in the next sections.

4.3 Low significant correlations: DNase I hypersensitive sites data

Table 4-2: Spearman's rank correlation between UniPROBE 8-mer E-scores and the 8-mer counts in DNase I hypersensitive sites.

TF	Correlation Coefficient	P-value
Esrra	0.266	0.007
Ets1	-0.447	$3.172e^{06}$
FoxA2	0.254	0.011
Gata3	0.366	0.000
Hnf4a	0.242	0.015
Irf4	-0.569	$6.29e^{-10}$
Klf7	0.175	0.082
Mafk	0.385	$7.63e^{-05}$
Max	-0.249	0.012
Pou2f1	0.122	0.226
Pou2f2	0.333	0.001
Sox12	0.237	0.018
Sp4	0.169	0.09
Tcf3	0.335	0.001

Our results for the occurrence of UniPROBE 8-mers in DNase I hypersensitive sites show that only three (Ets1, Irf4 & Max) out of the fourteen TFs have negative correlation coefficients, two of which are quite significant (Table 4-2). Although we were expecting to see positive correlation coefficients in this section, the values were low (less than 0.4) with only Mafk having a highly significant coefficient (p-value=$7.63e^{-05}$). We recall that DNase I hypersensitive sites are open chromatin sites that can be identified after digestion of DNA by DNA nuclease I (DNase I) which prefers these nucleosome depleted regions. The rationale is

that TFBS are located within these open chromatin sites since DNA wrapped around histone proteins is inaccessible for binding (Section 2.4.4). Therefore these hypersensitive sites only give an idea of where binding is likely to occur but the bound protein is not known.

The fact that the top-scored UniPROBE 8-mers occur less in DNase I hypersensitive sites can also mean that the PBM approach scores true TFBS more highly than the background. The positive correlation coefficients are expected since DNase I hypersensitive sites are open chromatin sites, genome regions were TFs are likely to bind. This is also a clear indication that the occurrences of these 8-mers are almost similar (see appendix for the occurrences) although these were from different TFs (experiments). This could be a result of the bias of DNase I as reported by He *et al*., (2014). They noticed some similar cleavage patterns after experimenting with naked DNA and TF-bound DNA. The rationale for DNase-seq is that protein-bound DNA is protected from DNase I digestion. However, Neph *et al*., (2012) argued that not all protein-bound DNA is completely guarded against DNase I cleavage. Assuming this to be true could also be another reason for the low occurrences of highly scored UniPROBE 8-mers in these likely binding regions.

4.4 Mixed negative and positive correlations: Hg19 data.

The human genome (Hg19) was used as a control dataset. Our aim was to check if there is a difference in the occurrences of UniPROBE 8-mers in likely binding regions (ChIP-seq peaks and DNase I hypersensitive sites) and the human genome in general which contains both binding and nonbinding regions. From our results, the correlations coefficients for Hg19 data are a mixture of negative and positive values (Table 4-3). They are different from those for DNase I hypersensitive sites and ChIP-seq data. Irf4 is an exception with its highly significant negative correlation (-0.575) almost similar to the one for itself again in DNase I hypersensitive sites (-0.569). Another exception is Ets1 which also has a significant negative coefficient (-0.472) almost the same coefficient for DNaseI hypersensitive sites again (-0.447). Comparing values for tables 4-1 and 4-3, Tcf3 and Pou2f1 have almost the same values in both tables i.e. the correlation coefficient for Tcf3 in table 4-1 (0.284) is in the same range with its correlation coefficient in table 4-3 (0.228). Similar correlation coefficients from these two tables means that the TFs have almost similar occurrences of their UniPROBE 8-mer patterns in likely binding regions and the human genome as a whole.

Table 4-3: Spearman's rank correlation between UniPROBE 8-mer E-scores and the 8-mer counts in Hg19.

TF	Correlation Coefficient	P-value
Esrra	-0.162	0.107
Ets1	-0.472	$6.968e^{-07}$
FoxA2	0.161	0.110
Gata3	0.242	0.016
Hnf4a	0.041	0.685
Irf4	-0.575	$4.012e^{-10}$
Klf7	-0.445	$3.44e^{-06}$
Mafk	0.059	0.557
Max	-0.366	0.0002
Pou2f1	0.478	$4.98e^{-08}$
Pou2f2	0.515	$4.152e^{-08}$
Sox12	0.333	0.0007
Sp4	-0.376	0.0001
Tcf3	0.228	0.023

4.5 High AUROC values: ChIP-seq data

Our results for ChIP-seq data show that 10 out of 14 TFs have an AUROC greater than or equal to 0.8 and only Irf4 has a value close to 0.4 with the remaining (Pou2f1, Pou2f2 and Tcf3) having almost the same value (Figure 4-1). Higher AUROC values means that our method could discriminate the high scoring UniPROBE 8-mer patterns and the low scoring ones based on their counts from the TF ChIP-seq peak region. Pou2f1 and Pou2f2 have almost the same value. The two are members of the POU family transcription factors. Although the value is below our threshold for high quality discrimination, this shows some similarity in the occurrences of their 8-mer patterns which might be a good prediction of their *in vivo* binding since members of the same TF family normally have a particular binding pattern although the affinity might differ. Looking at the AUROC values (Figure 4-1) and the Spearman's rank correlations (Table 4-1) for the same data set (ChIP-seq), we can actually see some similar pattern. Esrra, FoxA2, Klf7, Sp4 and Mafk which have large significant

correlation coefficients are also the leading ones interms of their AUROC values whereas Hnf4a and the POU TFs have intermediate values in both datasets and Irf4 remains an outlier.

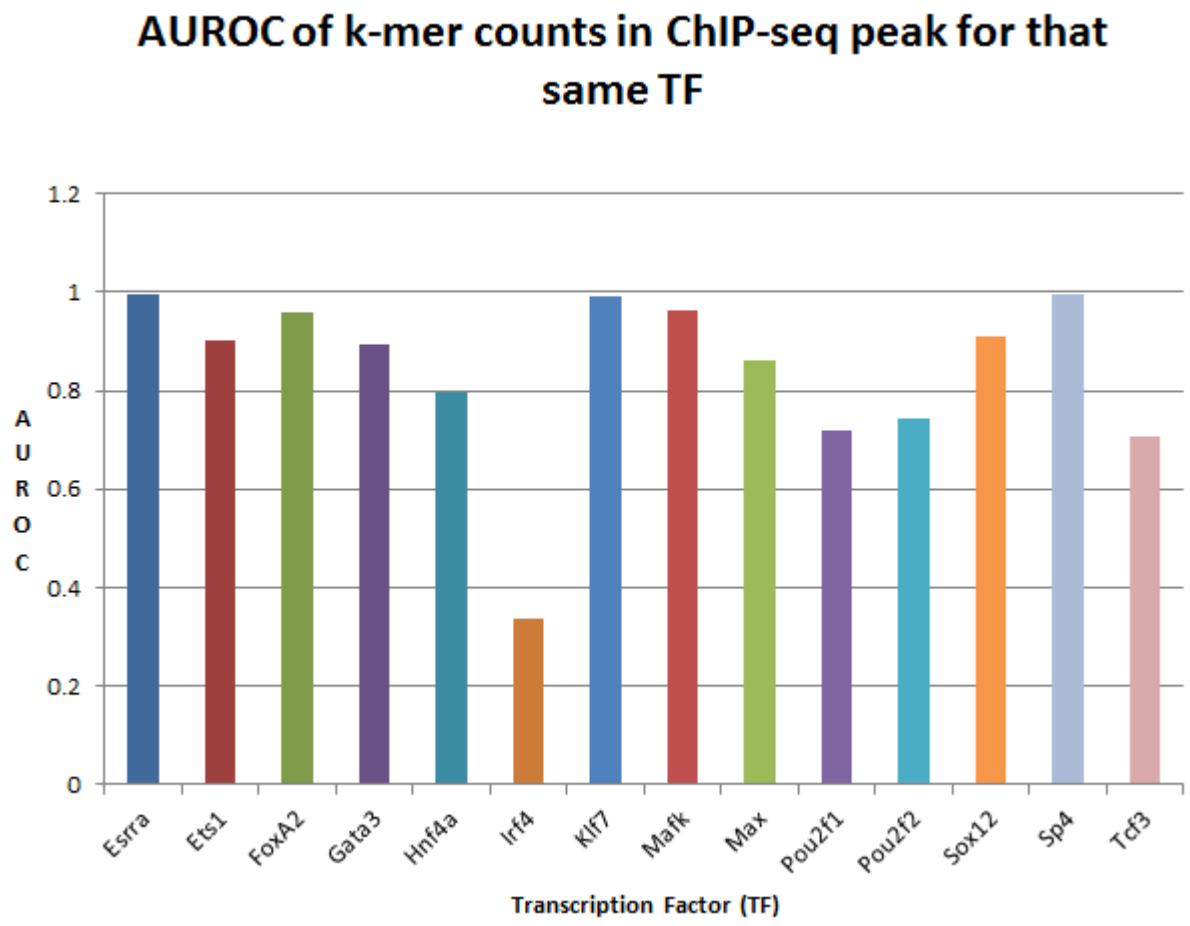

Figure 4-1: The AUROC values for ChIP-seq data

4.6 Intermediate AUROC: DNase I hypersensitive sites data

Using DNase I hypersensitive sites, our results show that none of the 14 TFs we used has an AUROC above our threshold (0.8) and only two (Mafk and Tcf3) have AUROC greater than 0.7 (Figure 4-2). The rest are ranging from 0.6 to 0.7 except for our usual outlier Irf4 but this time it has found two partners (Max and Ets1). Comparing these results and the Spearman rank correlations (Table 4-2) for the same data set, although the pattern looks different, Irf, Max and Ets1 are the only ones with significant negative correlation coefficients. POU TFs have AUROC values not so close to each other as was the case in their correlation coefficients. Our results from both approaches (AUROC and Spearman's rank correlation) do not provide solid evidence to link the occurrences of UniPROBE 8-mer patterns in DNase I hypersensitive sites with their enrichment scores.

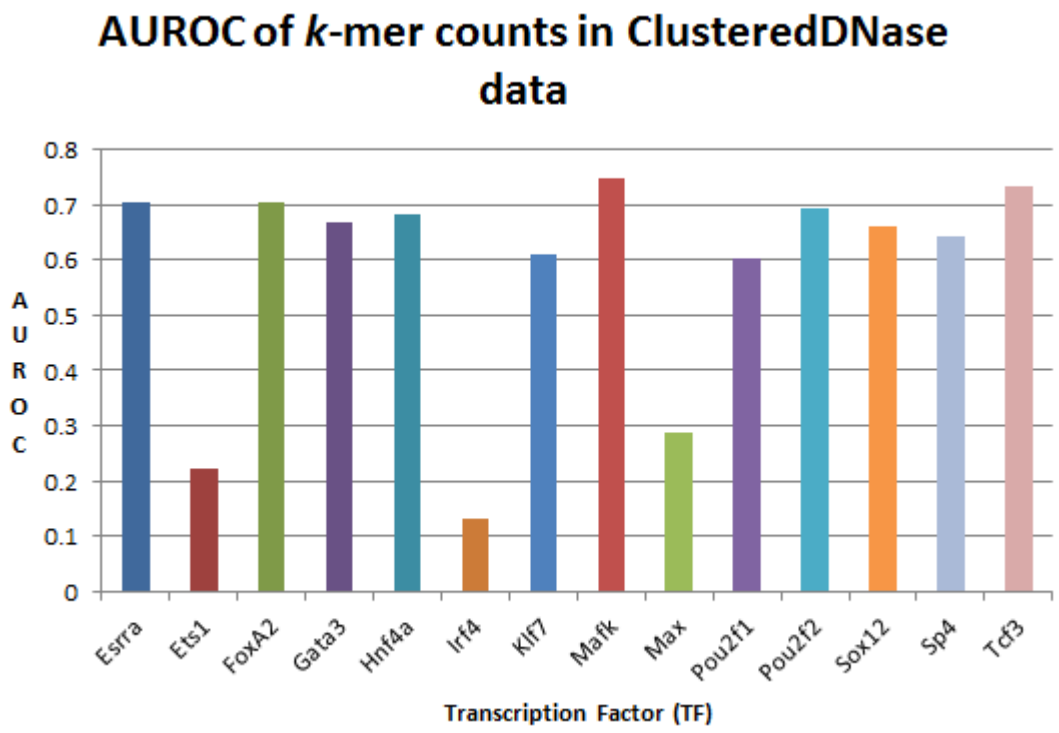

Figure 4-2. AUROC values for ClusteredDNase data (DNase I hypersensitive sites).

4.7 Poor discrimination for Hg19 data

We expected AUROC values around 0.5 since the whole genome is a combination of binding and nonbinding sites, both the high scoring and low scoring 8-mer patterns are expected to have an equal distribution thus a failure to discriminate data by the approach. As expected, quite a number of TFs have an AUROC less than 0.5 and these are Irf 4, Klf7, Ets1, Max and Sp4. The remaining have AUROC values within the range 0.5 - 0.7 (Figure 4-3). The AUROC values are also different from the ones for ChIP-seq and DNase I hypersensitive sites data. The POU TFs have a higher AUROC value (about 0.7-0.75) as was the case for their ChIP-seq data. We suspect this high correlation to the whole genome is because the TF binds to the genome in the presence of other transcription factors or the UniPROBE 8-mers can be ubiquitous binding patterns.

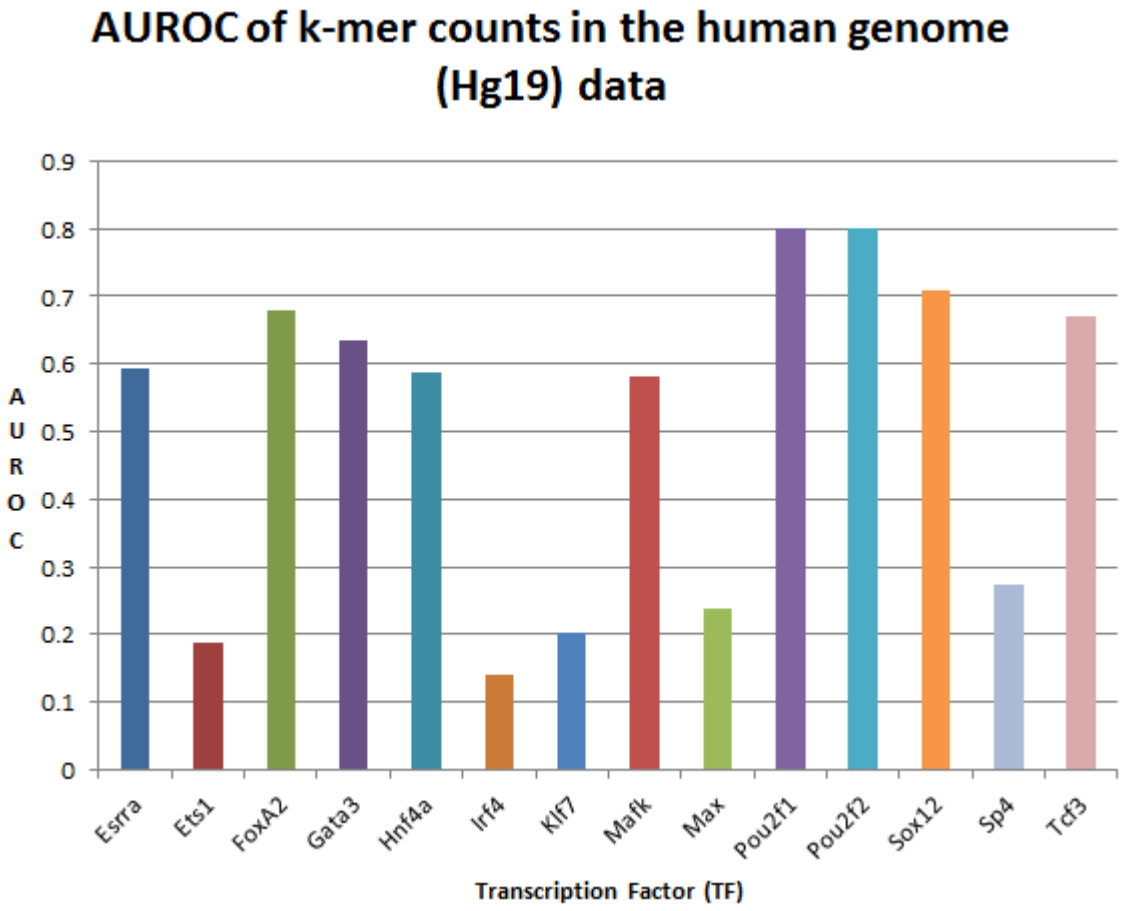

Figure 4-3. AUROC values for Hg19 data.

4.8 Combining our ChIP-seq correlations with CentriMo results

CentriMo results fall into four categories based on the Spearman's rank correlation coefficient for ChIP-seq data (Table 4-1). The categories are: (i) motifs with a correlation coefficient greater than or equal to 0.7 (Table 4-4), (ii) motifs with a correlation coefficient between 0.5 and 0.7 (Table 4-5), (iii) motifs with a correlation coefficient between 0 and 0.5 (Table 4-6), and (iv) motifs with a negative correlation coefficient (Table 4-7). We recall that with ChIP-seq data, we know the bound TF unlike DNase I hypersensitive sites which are likely binding regions but the bound protein is not known. Our interpretation is that if the UniPROBE 8-mer patterns match *in vivo* specificity, then a significant positive correlation for ChIP-seq data should be accompanied by more central enrichment of UniPROBE motifs with highly significant p-values and shorter region widths as compared to JASPAR motifs in the TF's respective ChIP-seq peak region. A significant negative correlation therefore means the opposite is true. We may also refer to correlations for other data sets and AUROC values as well in this discussion.

4.8.1 Motifs with a correlation coefficient greater than or equal to 0.7

Table 4-4. An extract of Table 4-1 showing motifs with a correlation coefficient greater than or equal to 0.7

TF	Correlation Coefficient	P-value
FoxA2	0.709	$2.2e^{-16}$
Klf7	0.752	$2.2e^{-16}$
Mafk	0.763	$2.2e^{-16}$
Sp4	0.757	$2.2e^{-16}$

FoxA2: Results show that the UniPROBE 8-mer patterns for FoxA2 match its *in vivo* binding specificity. From Table 4-4, FoxA2 has a highly significant correlation coefficient (0.709). Although, in our CentriMo output, the site probability curve for the FoxA2 JASPAR motif is slightly higher than that for the UniPROBE primary motif, the UniPROBE primary motif has a narrower region of central enrichment (w = 58) than the JASPAR motif (w = 63 bp) (Figure 4-4a). Both motifs central enrichment is highly significant since their p-values are smaller (p = 4.3e-12715 and p = 6.2e-12274). Their peaks are also unimodal and located at the center which can be a result of direct binding. From FoxA2 UniPROBE 8-mer patterns, GTAAACA has the highest E-score (0.497). The reverse complement of this pattern is the one observed on the sequence logo for the JASPAR motif. High degree of similarity between these patterns, large positive correlation coefficient and the little difference between the central enrichment of the UniPROBE and JASPAR motif suggest that the PBM method was able to predict *in vivo* binding specificity of FoxA2.

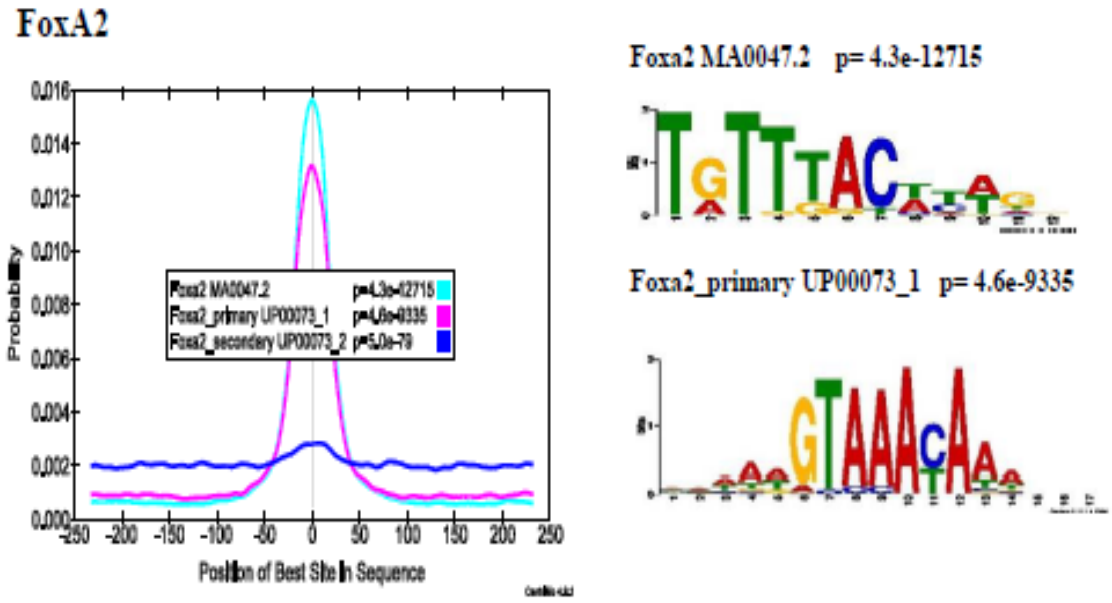

Figure 4-4a. CentriMo results for UniPROBE and JASPAR motifs in FoxA2 ChIP-seq peak region. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

Klf7 (Klf4): The highly significant positive correlation suggests that the UniPROBE 8-mer patterns match the *in vivo* binding specificity of Klf7. However, this is inconsistent with our CentriMo output (Figure 4-4b). Klf4 JASPAR and Klf7 UniPROBE primary motif are among the top enriched motifs. Both have unimodal peaks of almost the same height (~0.008). Although the UniPROBE motif has much narrower region of central enrichment (w = 113), its p-value is less significant (p = 8.4e-915) as compared to the JASPAR motif (p = 4.1e-1098). Both Klf7 and Klf4 are members of the Kruppel-like Factors and these bind to "GT-box" or "CACCC" sites on DNA (Bieker, 2001). At the moment, there is no UniPROBE motif for Klf4 but the top scored UniPROBE 8-mer patterns for Klf7 are highly rich in GC nucleotides, for example CCACGCCC (E-score = 0.497). The CGCCC pattern can be observed on the sequence logo for the Klf7 UniPROBE primary motif and on the reverse complement of Klf4 JASPAR motif. In this case, it looks like the PBM approach also allocates a high score to some flanking regions since the Klf7 secondary motif has a poor CentriMo distribution with a less significant p-value (p = 7.3e-56).

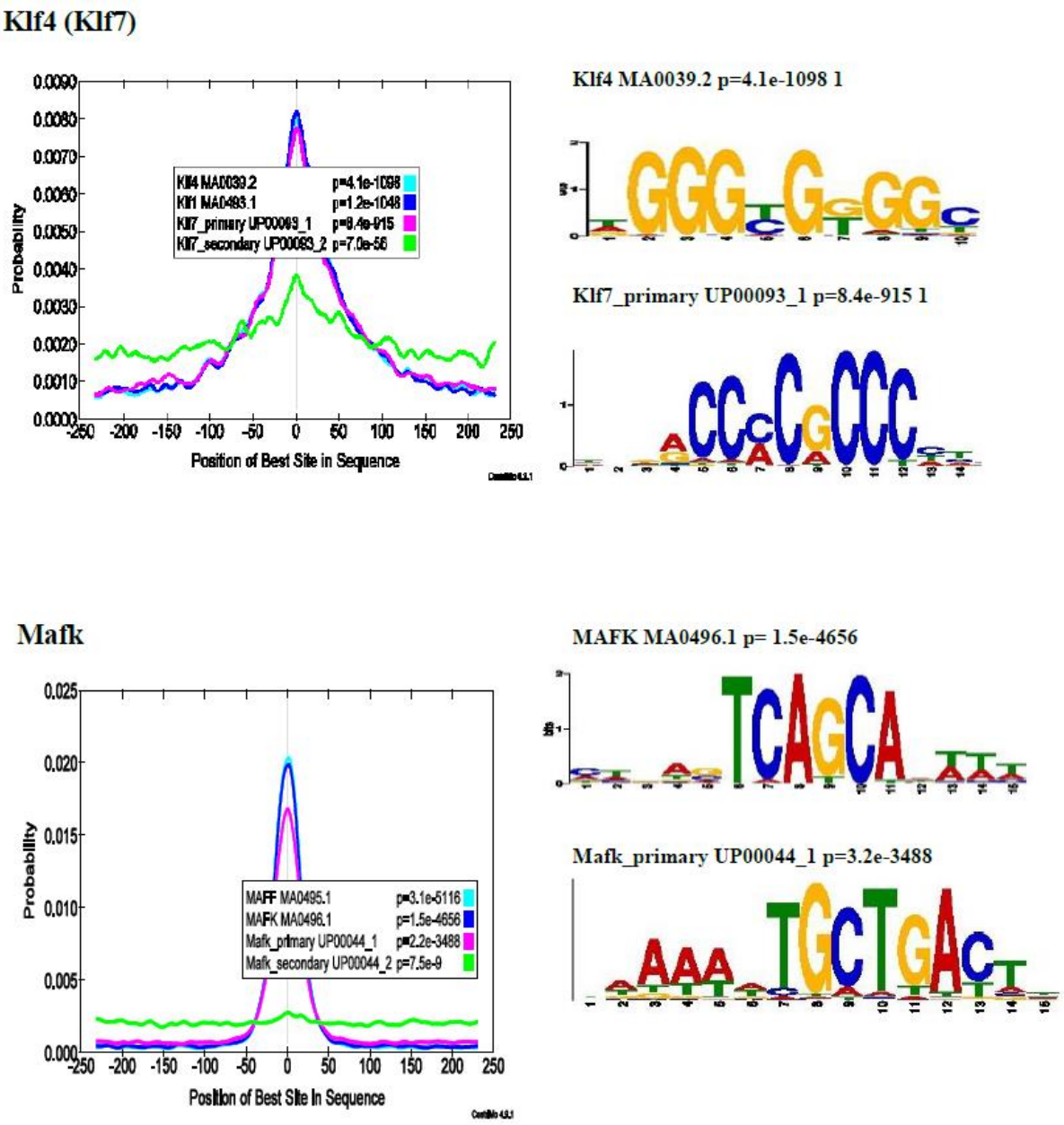

Figure 4-4b. CentriMo results for UniPROBE and JASPAR motifs in Klf4 (Klf7) and Mafk ChIP-seq peak regions. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

Mafk: CentriMo analysis of Mafk shows that the Mafk JASPAR motif predicts *in vivo* binding better than the UniPROBE primary motif although there is no huge difference between the two in terms of their site probability curves, region of central enrichment and p-values (Figure 4-4b). The two have unimodal peaks clearly indicating direct binding. The maximum site probability curve for the JASPAR motif (~0.020) is slightly higher than the UniPROBE one (~0.017) but both are at the center of the peak regions. Both have a very

narrow region of central enrichment (w = 56) although the JASPAR motif is slightly more significant (p = 1.5e-4656) than the UniPROBE one (p = 2.0e-4181). However, both p-values are nearly zero. The JASPAR MAFF motif also appears to be the most centrally enriched motif with a more significant p-value (p = 3.1e-5116) and much narrower region of central enrichment (w = 55). MafF, MafK and MafG are closely related small Maf family TFs (Kannan *et al*., 2012) therefore we suspect that MafF and MafK have similar binding patterns *in vivo*. The top scored UniPROBE Mafk 8-mer pattern is AA.TGCTGA and the TGCTGA pattern can be observed on the logos for MafF and MAfK JASPAR motifs. These results are also consistent with the large significant correlation for Mafk (Table 4-4). We assume that the *in vitro* binding site prediction matched *in vivo* specificity and the little difference in the motifs enrichment can be attributed to the similar *in vivo* binding specificities for MafF and MafK.

Sp4: Our CentriMo results for Sp4 are inconsistent with its highly significant correlation (Figure 4-4c). The peaks are less sharply defined, not well centered and the region of maximum central enrichment is broad (w > 150 bp). Only the Sp4 UniPROBE motifs are among the enriched motifs and the motifs also show less significant enrichment p-values (p = 5.7e-39 and (p = 5.0e-12). There are many reasons for such a distribution. These include failure or poor resolution of the ChIP-seq experiment, indirect binding and the presence of cofactors.

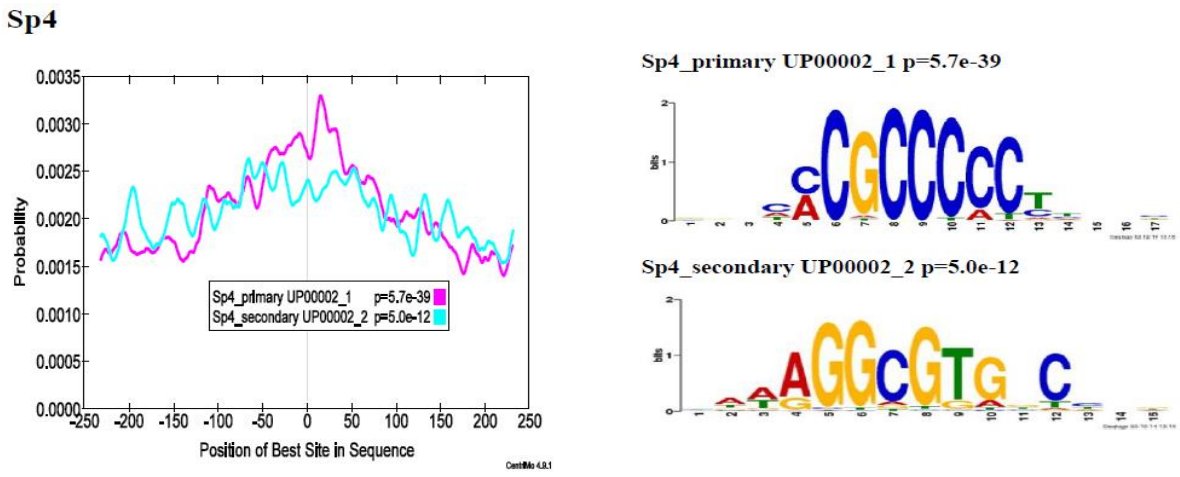

Figure 4-4c. CentriMo results for UniPROBE and JASPAR motifs in Sp4 ChIP-seq peak region. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

4.8.2 Motifs with a correlation coefficient between 0.5 and 0.7

Table 4-5. An extract of Table 4-1 showing motifs with a correlation coefficient between 0.5 and 0.7.

TF	Correlation Coefficient	P-Value
Ets1	0.626	$3.213e^{-12}$
Gata3	0.626	$4.161e^{-12}$
Max	0.628	$2.749e^{-12}$
Sox12	0.614	$1.122e^{-11}$

Ets1: Ets1 is one good example where we suggest the PBM method might have failed to predict the binding specificity of Ets1 *in vivo*. With its average significant correlation, we expected UniPROBE motifs to be among the centrally enriched motifs. However, this is not the case. None of the five centrally enriched motifs is from the Ets family (Figure 4-5b) and none of them has a unimodal peak (all have big dips) suggesting binding Ets1 binds in the presence of cofactors. Although they have higher maximum site probability curves, their regions of central enrichment are broad (w > 100 bp). Only JASPAR Ets1 motif appears to be centrally enriched but it has a much less significant p-value (p=1.8e-143) and a much lower maximum site probability curve than the top five enriched motif. Apparently, ELK4 and GABPA (both from ChIP-seq) are among the top enriched motifs. From the UniPROBE top scored 8-mer patterns, the pattern ACCGGAAG has the highest E-score of 0.497. The pattern CCGGAA can be observed on the motif logos for ELK4 and GABPA. Both GABPA and ELK4 are members of the ETS transcription factor family and they all have a very high binding affinity for the sequence CCGGAAGT (Hollenhorst *et al*., 2011). This shows that the PBM technique can fail differentiate TFs with the same binding specificity *in vivo* such as the case with Ets1, GABPA and ELK4.

Gata3: Gata3 is an example where we assume the PBM method does not tell whether the binding is indirect or cooperative although it can match *in vivo* binding specificity. Its correlation coefficient is similar to that of Ets1. CentriMo output (Figure 4-5a) shows that Gata3 binding can be indirect or cooperative. Gata3 UniPROBE primary motif appears to be more enriched than the JASPAR motif although it is not among the top four enriched motifs. However, motifs have broader regions of maximum central enrichment (w > 100 bp); their

site probability curves achieve highest values off the center of their ChIP-seq regions with some dips at the top and their central enrichment is much less significant (p = 9.1e-166 and p = 1.3e-136). Although the JASPAR motif has a higher site probability curve, the UniPROBE primary motif has a narrower region of enrichment (w = 103 bp) than its region (w = 118 bp). GATA2 and Gata1 are also among the top enriched motifs which is not surprising since they are all members of the same family. Gata3 and Gata2 are known close paralogs and all the three have a very high affinity for the consensus sequence AGATAA (Ko & Engel, 1993). From the UniPROBE Gata3 top scored 8-mers, the sequence AGATAAGA has the highest E-score (0.497). The sequence is an exact match of the one mentioned in literature except for the last two nucleotides (GA). However, the sequence is specific for some GATA family TFs yet our UniPROBE data set is for Gata3. This is an indication that the PBM technique sometimes might fail to predict *in vivo* binding specificities for TFs from the same family

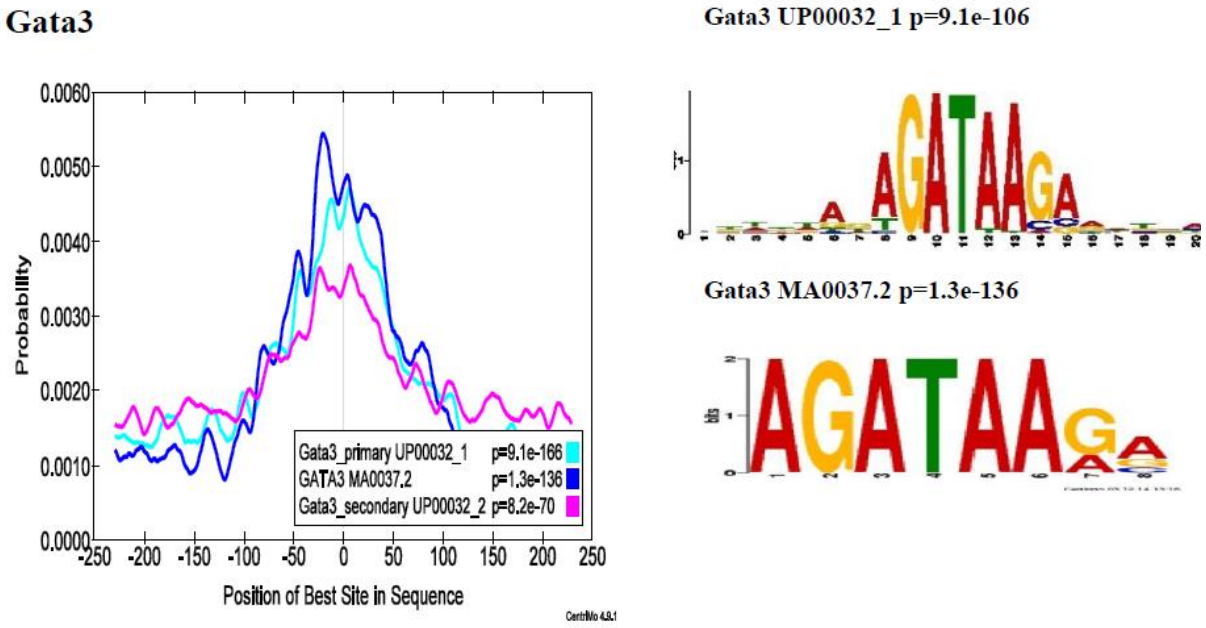

Figure 4-5a. CentriMo results for UniPROBE and JASPAR motifs in Gata3 ChIP-seq peak region. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

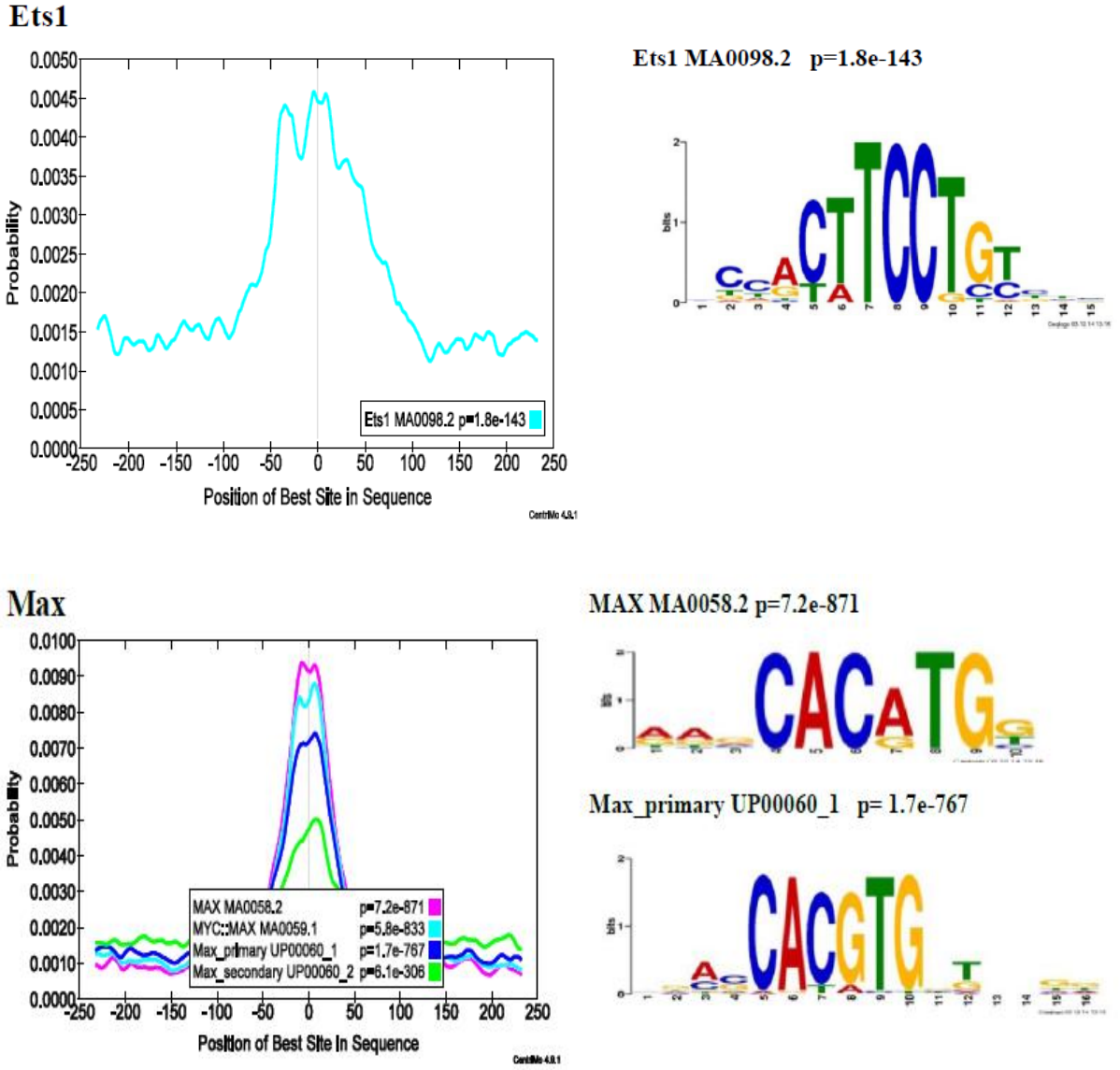

Figure 4-5b. CentriMo results for UniPROBE and JASPAR motifs in Ets1 and Max ChIP-seq peak regions. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

Max: An average correlation could be a result of cooperative *in vivo* binding of Max as suggested by our CentriMo results (Figure 4-5b). All the peaks have a slight dip at the top suggesting some cooperative or indirect binding. Cooperative binding is a possibility since the MYC::MAX and Mycn JASPAR motifs were among the top enriched motifs. The maximum site probability curve for the JASPAR motif (~0.009) is higher than the UniPROBE one (~0.007) but both are at the center of the peak regions. The UniPROBE

motif has a narrower region of central enrichment (w = 79 bp) than the JASPAR motif (w = 81 bp) but its p-value is slightly less significant (p = 1.7e-767) compared to that of the JASPAR motif (p = 7.2e-871). Literature evidence (James & Eisenman, 2002) shows that Myc:Max is an active repressor of many regulatory genes. Max binds specifically to the DNA sequence GGGCAC(G/A)TGCC either alone or with c-Myc (Luscher & Larsson, 1999) but with a higher affinity for CACGTG. CACGTG is the sequence observed on the logos of all the enriched motifs and the entire top scored Max UniPROBE 8-mer patterns contain this sequence. The PBM approach shows a good prediction of *in vivo* binding specificity but more central enrichment of the JASPAR motif can be attributed to cooperative binding hence the binding sites tend to be widespread.

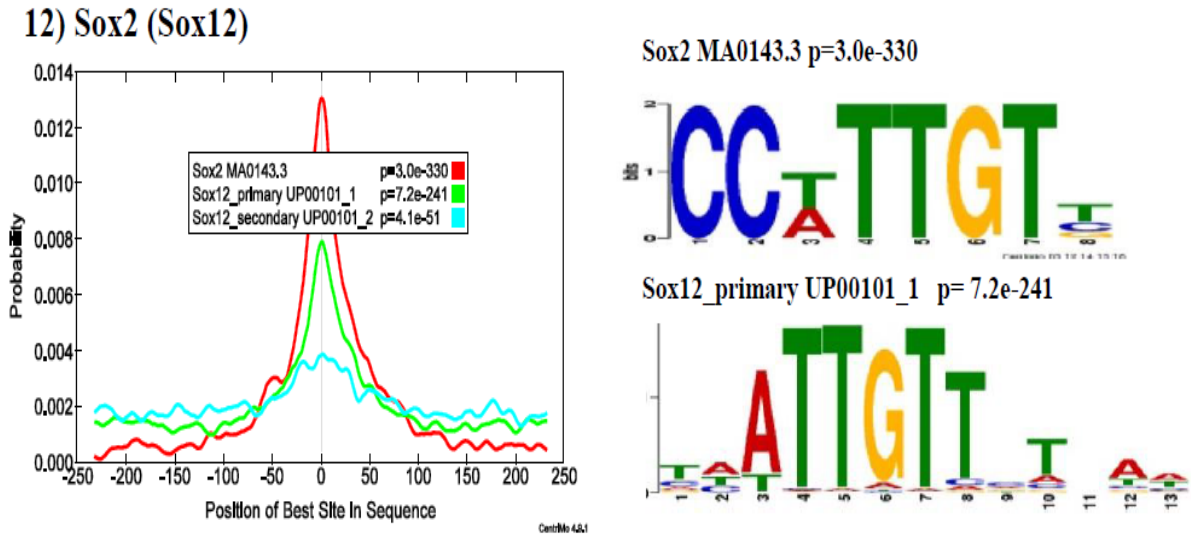

Figure 4-5c CentriMo results for UniPROBE and JASPAR motifs in Sox2 ChIP-seq peak region. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

Sox 2 (Sox12): Sox2 is another example where the PBM approach fails to differentiate the *in vivo* binding specificities of TFs from the same family. None of the Sox2 UniPROBE motifs is among the centrally enriched motifs (Figure 4-5c). Instead, Sox12 UniPROBE motifs are among the centrally enriched motifs together with Sox2 JASPAR motif. The JASPAR Sox2 motif is more centrally enriched but with a less significant p-value (p = 3.0e-330). Sox 12 UniPROBE primary motif has a much less significant p-value (p = 7.2e-241) and a lower site probability curve (~0.08). The region of maximum central enrichment is much narrower (w = 75 bp) than the one for the JASPAR motif (w = 113 bp). The sequence logos for both motifs

show some similar pattern ATTGT. Both Sox2 and Sox12 are members of the Sox family of TFs. Looking at the Sox12 UniPROBE 8-mer patterns, the entire top enriched 8-mer patterns contain the pattern ATTGT. The PBM approach matches the *in vivo* binding specificity but the average correlation can be a result of ubiquitous binding patterns.

4.8.3. Motifs with a correlation coefficient between 0 and 0.5

Table 4-6. An extract of Table 4-1 showing motifs with a correlation coefficient between 0 and 0.5.

TF	Correlation Coefficient	P-value
Hnf4a	0.395	$4.713e^{-05}$
Pou2f1	0.293	0.003
Pou2f2	0.421	$1.299e^{-05}$
Tcf3	0.284	0.004

Hnf4a: Low correlation coefficient plus more central enrichment of the UniPROBE secondary motif than the primary motif (Figure 4-6a) suggests the PBM approach may have failed to match *in vivo* binding specificity of Hnf4a. Compared with the two UniPROBE motifs, the JASPAR motif has the highest maximum site probability (~0.012) and a highly significant p-value (1.3e-5482). Hnf4a binds DNA as a homodimer and in form of repeats (AGGTCA x AGGTCA) (Bolotin *et al*., 2011). The top scored UniPROBE 8-mer for Hnf4a is GGGGTCAA with an E-score of 0.496. The pattern is completely different from the one observed pattern on the sequence logo for the JASPAR motif which is the most centrally enriched motif indicating a possible failure of the PBM technique.

Pou2f1 (Oct4): None of the Pou2f1 motifs is among the centrally enriched motifs (Figure 4-6a). This observation is consistent with the low correlation (0.293) between the UniPROBE 8-mer enrichment scores and the 8-mer occurrences in likely binding regions. We also suspect failure of the PBM approach. Members of the POU TF family appear to be the most centrally enriched. Pou5f1::Sox2 JASPAR motif is among the centrally enriched motifs. Pou2f2, Pou2f3 and Pou3f3 UniPROBE primary motifs are more centrally enriched than Pou2f2 JASPAR motif. Oct4 is a member of the POU domain TFs which is known to be specific to the DNA octamer 5'-ATGCAAAT-3' (Kang *et al*., 2009). The top scored 8-mer patterns for Pou2f1 have completely different sequences from the ATGCAAT pattern. Lack

of centrally enriched Pou2f1 UniPROBE motifs leaves a question on the accuracy of the PBM approach on prediction of Pou2f1 *in vivo* binding specificity.

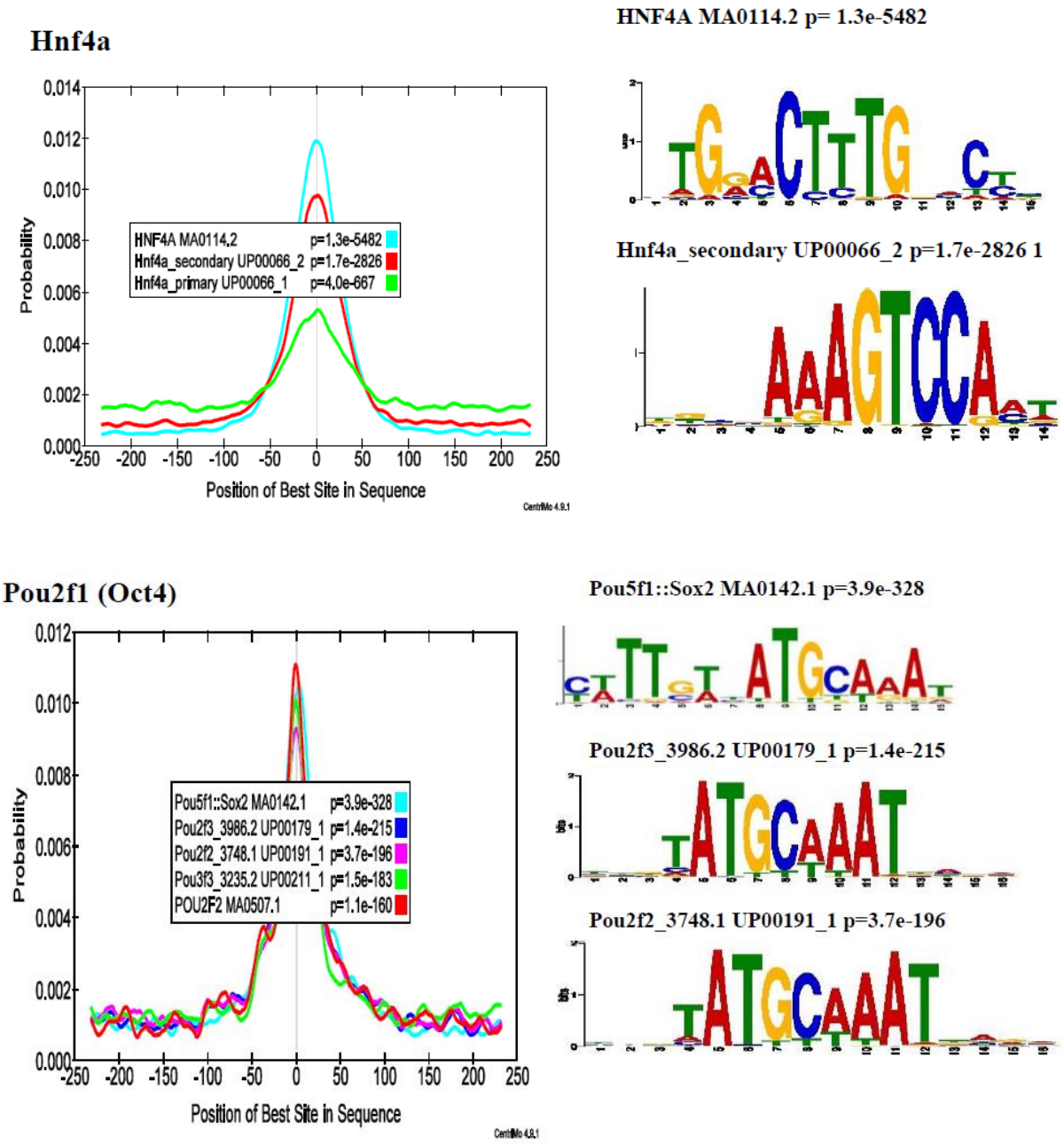

Figure 4-6a: CentriMo results for UniPROBE and JASPAR motifs in Hnf4a and Pou2f1 ChIP-seq peak regions. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

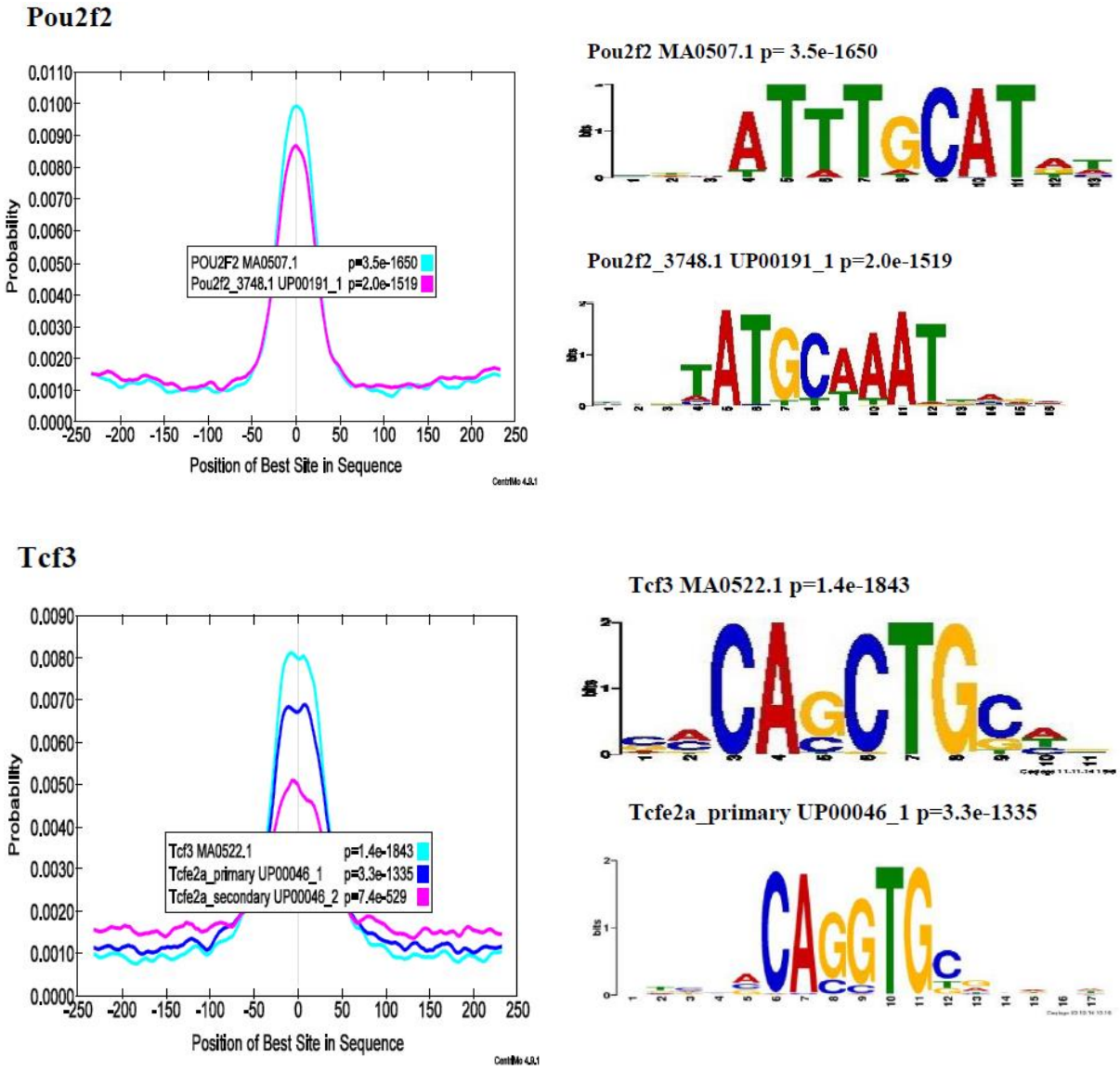

Figure 4-6b. CentriMo results for UniPROBE and JASPAR motifs in Pou2f2 and Tcf3 ChIP-seq peak regions. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

Pou2f2: Although it has a low correlation coefficient, CentriMo analysis of Pou2f2 shows good prediction of its *in vivo* binding specificity by the PBM technique. CentriMo output shows that both Pou2f2 motifs (JASPAR and UniPROBE primary) are centrally enriched (Figure 4-6b). Both have unimodal peaks of almost equal height. The JASPAR motif has a higher maximum site probability curve (~0.01), narrower region of central enrichment (w = 60 bp) and a more significant p-value (p = 3.5e-1650) than the UniPROBE motif (~0.009, w = 63 bp and p = 2.0e-1519). Pou2f2 (also known as Oct2) is another member of the POU domain TFs which also binds to the octamer ATGCAAAT. The enriched UniPROBE motif

logos show this sequence with high information content at all position. From Pou2f2 UniPROBE 8-mer patterns, the same sequence has the largest enrichment score (0.497). Although it was outperformed by the JASPAR motif, our results show that the Pou2f2 UniPROBE motif is also a good description of *in vivo* binding specificity for Pou2f2.

Tcf3: Results for Tcf3 also suggest failure of the PBM approach in predicting its *in vivo* binding specificity. Tcf3 has a low significant correlation which is also consistent with the lack Tcf3 UniPROBE motifs among the centrally enriched motifs (Figure 4-6b). Furthermore, the top scored 8-mer patterns from UniPROBE do not match any of the sequence logos for the enriched motifs. We can say that the Tcf3 UniPROBE motif might not represent Tcf3 *in vivo* binding specificity.

4.8.4. Motifs with a negative correlation coefficient

Table 4-7. An extract of Table 4-1 showing motifs with a negative correlation coefficient

TF	Correlation Coefficient	P-value
Esrra	-0.736	$2.2e^{-16}$
Irf4	-0.239	$2.2e^{-16}$

Esrra (Esrrb): The statistically significant correlation (Table 4-7) for Esrra requires further investigations. We suspect this can be a result of cooperative binding since Esrra binding is also dependent on the presence of other proteins (cofactors) (Deblois & Giguere, 2010) but our CentriMo peaks are sharply defined. CentriMo output shows two JASPAR (Esrrb and Esrra) and one UniPROBE motif (Esrra primary) as the top enriched motifs (Figure 4-7). The site probability curves and regions of central enrichment of the JASPAR motifs and UniPROBE primary motif are almost the same but the JASPAR ones have significantly higher p-values (p=3.0e-3347 and p=2.6e-3295) than UniPROBE one (p=6.9e-3043). From Esrra UniPROBE 8-mer patterns, the top scored pattern is CAAGGTCA with an E-score of 0.499. This sequence is the one observed on the sequence logos of the top two enriched motifs from our CentriMo results and the motifs were derived from ChIP-seq data. The UniPROBE primary motif also contains the same sequence. In such a scenario, we expected a significant positive correlation coefficient since the *in vitro* technique matches *in vivo* binding specificity of the motif in consideration. From this scenario, we suspect the PBM technique also scores some sequences highly where they may not be true transcription factor binding

sites since a negative correlation means that most of the low-scoring 8-mer patterns had high occurrences in likely binding regions.

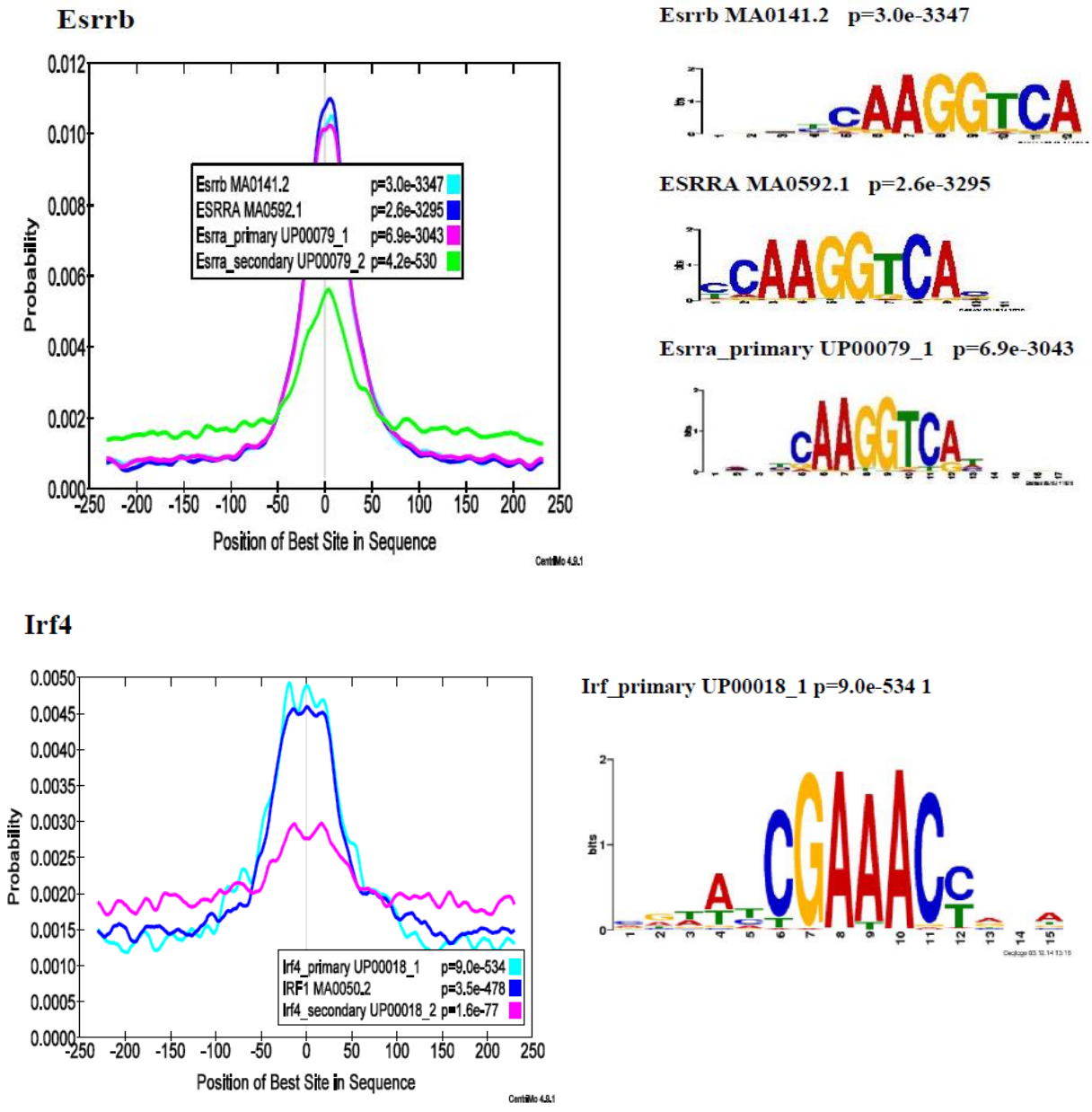

Figure 4-7. CentriMo results for UniPROBE and JASPAR motifs in Esrrb and Irf4 ChIP-seq peak regions. The curve shows the density of the best strong site for the named motif at each position in the peak region. The legend shows the motif and its central enrichment p-value. The enriched motifs are shown as sequence logos on the right side of the curves.

Irf4: The negative correlation for Irf4 plus lack of both JASPAR and UniPROBE Irf4 motifs among the most centrally enriched motifs suggests that Irf4 has a very narrow binding specificity *in vivo*. None of the Irf4 motifs was among the top nine centrally enriched motifs (Figure 4-7). Irf4 UniPROBE primary motif is ranked number ten followed by the secondary

motif located somewhere at the bottom of the rankings. Shaffer *et al*., (2009) reported that Irf4 alone weakly interacts with DNA due to its regulatory domain and it is capable of interacting with multiple TFs like ETS-family TFs so as to allow stronger DNA binding. Literature evidence (Furuita *et al*., 2006) shows that Irf4 prefers the sequence CCGAAA although IRF-family TFs prefer NNGAAA and Irf4-DNA binding is indirect. The negative correlation can also be attributed to variability in binding specificity of Irf4, the presence of cofactors (broad peaks from CentriMo output) or more background sequences than true binding sites.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Our sample was too small to come out with conclusive results. Although binding of TFs in a living tissue is controlled by other factors such as chromatin accessibility and the presence of cofactors which the PBM technique does not take into consideration, our study did not come out with a particular trend showing failure of the PBM approach to predict *in vivo* binding specificity. One or two scenarios, such as Ets1, Hnf4a and Tcf3 show prediction failure by the PBM technique in terms of our Spearman's rank correlation for ChIP-seq data and central motif enrichment analysis. Failure of the PBM approach was found to be a result of variability in the TF' binding specificity, the presence of cofactors, narrow binding specificity and the presence of a large number of background sequences. However, the PBM technique also matched the *in vivo* binding specificities of FoxA2, Pou2f2 and Mafk. There are some cases where we doubt the quality of our data set according to our CentriMo results such as Gata3 and Sp4. According to our Spearman's correlations, there is a possibility for a relationship between the enrichment scores of UniPROBE 8-mer patterns (PBM-derived 8-mers) and the occurrences of these patterns in both likely and unlikely binding regions. However, the nature of the relationship is not clear.

Our results are based on ChIP-seq data which also has artifacts. For example, there is a number of reasons for less sharply defined peaks from CentriMo output which include indirect binding, failure or poor resolution of the ChIP-seq experiment. Therefore further investigations are required for conclusive results from this study. For example, although it is tempting to say that the most frequent 8-mer in likely binding regions is the true TF binding site, this is not always the case. It can be some sequence necessary for TF-DNA binding (background). Further studies should take into consideration how to differentiate the true transcription factor binding sites from these background sequences. Instead of checking for the enrichment of motifs at the center of ChIP-seq regions, we also need to find a way of checking for the enrichment of motifs on the flanking regions of the ChIP-seq peak regions.

REFERENCES

Acquaviva, J. Chen, X. Ren, R. 2008. IRF-4 functions as a tumor suppressor in early B-cell development. *Blood*, **112**(9):3798-806.

Andersen, B. Rosenfeld, M.G. 1994. Pit-1 determines cell types during development of the anterior pituitary gland. *J Biol Chem*, **269**:29335-29338.

Ayala, R. Martinez-Lopez, J. Gilsanz, F. 2012. Acute myeloid leukemia and transcription factors: role of erythroid Kruppel-like factor (ELKF). *Cancer Cell International*, **12**:25.

Badis, G. Berger, M.F. Philippakis, A.A. Talukder, S. Gehrke, A.R. Jaeger, S.A. Chan, E.T. *et al*. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**:1720-1723.

Bailey, T. Krajewski, P. Ladunga, I. Lefebvre, C. Li, Q. Liu, T. Madrigal, P. Taslim, C. Zhang, J. 2013. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput Biol*, **9**(11):e1003326.

Bailey, T.L. 2011. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**(12):1653-1659.

Bailey, T.L. Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28-36. AAAI Press, Menlo Park, California.

Bailey, T.L. Machanick, P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, **40**:e128.

Barberis, A. Petrascheck, M. 2003. Transcription Activation in Eukaryotic Cells. *eLS*, **58**:138-189.

Berger, M.F. Badis, G. Gehrke, A.R. Talukder, S. Philippakis, A.A. et al. 2008. Variation in homeodomain DNA binding revealed by high resolution analysis of sequence preferences. *Cell*, **133**:1266-1276.

Berger, M.F. Bulyk, M.L. 2009. Universal protein binding microarrays for the comprehensive characterization of the DNA binding specificities of transcription factors. *Nat Protoc*, **4**(3):393-411.

Bieker, J.J. 2001. Kruppel-like Factors: Three Fingers in Many. *The Journal of Biological Chemistry*, **276**:34355-34358.

Bolotin, E. Chellappa, K. Hwang-Versules, W. Schnabl, J.M. Yang, C. Sladek, F.M. 2011. Nuclear Receptor HNF4α Binding Sequences are Widespread in Alu Repeats. *BMC Genomics*, **12**:560.

Bonett, D.G. Wright, T.A. 2000. Sample size requirements for Pearson, Kendall, and Spearman correlations. *Psychometrika*, **65**:23-28.

Boyle, A. P. Birney, E. Crawford, G.E. Iyer, V. R. Keefe, D. Lee, B. London, D. Song, L. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*, **21**:456-464.

Boyle, A.P. Davi S. Shulha, H.P., Meltzer, P. Margulies, E.H. Weng, Z. et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**:311-322.

Brooker, R.J. Widmaier, E.P. Graham, L.E. Stiling, P.D. 2010. *Biology*. 2nd *edition*. McGraw-Hill. New York.USA.

Bryne, J.C. Valen, E. Tang, M.E. Marstrand, T. Winther, O. Piedade, I. Krogh, A. Lenhard, B. Sandelin, A. 2008. JASPAR, the open access database of transcription factor –binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, **36**:D102-D106.

Buck, M.J. Lieb, J.D. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**(3):349-60.

Bulyk, M. L. Newburger, D. E. 2009. UniProbe: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, **37**:D77-D82.

Bulyk, M. L. Robasky, K. 2011. UniProbe, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, **39**:D124-D128.

Bulyk, M.L. 2006. DNA Microarray Technologies for Measuring Protein-DNA Interactions. *Curr Opin Biotechnol*, **17**(4): 422-430.

Bulyk, M.L. 2007. Protein Binding Microarrays for the Characterization of Protein-DNA Interactions. *Adv Biochem Eng Biotechnol*, **104**:65-85.

Buske, F.A. Boden, M. Bauer, D.C. Bailey, T.L. 2010. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, **26**(7):860-886.

Calkhoven, C.F. Geert, A.B. 1996. Multiple steps in the regulation of transcription-factor level and activity. *Biochem.J*, **317**:329-342.

Carey, M. Smale, S.T. 2000. *Transcriptional Regulation in Eukaryotes*. Cold Spring Harbor Laboratory Press. New York. USA.

Che, D. Jensen, S. Cai, L. Liu, J.S. 2005. BEST: Binding-site Estimation Suite of Tools. *Bioinformatics*, **21**(12):2909-2911.

Chen, X. Guo, L. Fan, Z. Jiang, T. 2008b. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics*, **24**(9):1121-1128.

Chen, X. Xu, H. Yuan, P. Fang, F. Huss, M. Vega, V.B. Wong, E. Orlov, Y.L. Zhang, W. *et al*. 2008a. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, **133**:1106-1117.

Chi, Y. Huddleston, M.J. Zhang, X et al. 2001. Negative regulation of Gcn4 and Msn2 transcription factors by Scrb10 cyclic-dependent kinase. *Genes and Development*, **15**:1078-1092.

Cooper, G.M. Hausman, R.E. 2007. *THE CELL, A molecular approach, 4th edition*. ASM Press. Washington, D.C. USA.

Deblois, G. Giguere, V. 2010. Functional and physiological genomics of estrogen-related receptors (ERRs) in health and diseases. *Biochim Biophys Acta*, **1812**(8):1032-1040.

ENCODE Consortium. 2004. The ENCODE (Encyclopedia of DNA Elements) project. *Science*, **306**: 636-640.

ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**: 57-74.

ENCODE Consortium. 2014. A Comparative Encyclopedia of DNA elements in the Mouse Genome. *Nature*, **515**:355-364.

Engelkamp, D. Van Heyningen, V. 1996. Transcription factors in diseases. *Current Opinion in Genetics & Development*, **6**:334-342.

Fauteux, F. Blanchette, M. Strömvik, M.V. 2008. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, **24**(20):2303-2307.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, **27**:861-874.

Frith, M.C. Hansen, U. Spounge, J.L. Weng, Z. 2004. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, **32**:189-200.

Fu, Y. Weng, Z. 2005. Improvement of TRANSFAC Matrices Using Multiple Local Alignment of Transcription Factor Binding Site Sequences. *Genome Informatics*, **16**(1):68-72.

Fukue, Y. Sumida, N. Tanase, J. Ohyama, T. 2005. A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucleic Acids Res*, **33**(12):3821-3827.

Furuita, K. Ishizaki, I. Fukada, H. Yamamoto, K. Matsuyama, T. Nomura, M. Mishima, M. Kojima, C. 2006. Studies of DNA recognition mechanism of transcription factor IRF-4. *Nucleic Acids Symp Ser*, **50**(1):259-260.

Gaston, K. Jayaraman, P.S. 2003. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cell Mol Life Sci*, **60**(4):721-41.

Gilfillan, G.D. Hughes, T. Sheng, Y. Hjorthaug, H.S. Straub, T. Gervin, K. Harris, J.R. Undlien, D.E. Lyle, R. 2012. Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*, **13**:645.

Gupta, S. Stamatoyannopolous, J.A. Bailey, T. Noble, W.S. 2007. Quantifying similarity between motifs. *Genome Biology*, **8**(2):R24.

Hanczar, B. Hua, J. Sima, C. Weinstein, J. Bitter, M and Dougherty, E.R. 2010. Small-sample precision of ROC-related estimates. *Bioinformatics*, **26**(6):822-830.

Hanna-Rose, W. Hansen, U. 1996. Active repression mechanisms of eukaryotic transcription repressors. *TIG*, **12**(6): 229-234.

He, H.H. Meyer, C.A. Hu, S.S. Chen, M.W. Zang, C. Liu, Y. Rao, P.K. Fei, T. Xu, H. Long, H. Liu, S. Brown, M. 2014. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods*, **11**:73-78.

Hertz, G.Z. Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**:563-577.

Hollenhorst, P.C. Ferris, M.W. Hull, M.A. Chae, H. Kim, S, Graves, B.J. 2011. Oncogenic ETS proteins mimic activated RAS/MAPK signaling in prostate cells. *Genes & Development*, **25**:2147-2157.

Hon, L.S. Jain, A.N. 2006. A deterministic motif finding algorithm with application to the human genome. *Bioinformatics*, **22**(9):1047-54.

Hughes, J.D. Estep, P.W. Tavazoie, S. Church, G.M. 2000. Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, **296**(5):1205-1214.

James, L. Eisenman, R.N. 2002. Myc and Mad bHLHZ domains possess identical DNA-binding specificities but only partially overlapping functions *in vivo. JSTOR*, **99**(16):10429-10434.

Jiang, B. Liu, J.S. Bulyk, M.L. 2013. Bayesian hierarchical model of protein binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, **29**(11):1390-1398.

Kannan, M.B. Solovieva, V. Blank, V. 2012. The small MAF transcription factors MAFF, MAFG and MAFK: Current knowledge and perspectives. *Biochim Biophys Acta*, **1823**(10):1841-1846.

Ko, L.J. Engel, J.D. 1993. DNA-binding specificities of the GATA transcription factor family. *Molecular Cell Biology*, **13**(7):4011-22.

Krajewski, P. Madrigal, P. 2012. Current bioinformatics approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front. Gene*, **3**:230.

Kranthi, B.V. Kumar, R. Kumar, N.V. Rao, D.N. Rangarajan, P.N. 2009. Identification of key elements involved in promoter recognition by Mxr1p, a master regulator of methanol utilization pathway in *Pichia pastoris*. *Biochim Biophys Acta*, **1789**(6-8):460-8.

Kurata, T. Katayama, A. Hiramatsu, M. Kiguchi, Y. Takeuchi, M. Watanabe, T. Ogasawara, H. Ishihama, A. Yamamoto, K. 2013. Identification of the Set of Genes, Including Nonannotated morA, under the Direct Control of ModE in *Escherichia coli*. *J.Bacteriol,* **195**(19):4496-4505.

Latchman, D.S. 2004. *Eukaryotic transcription factors. 4th edition*. Elsevier Academic Press. London. United Kingdom.

Lesluyes, T. Johnson, J. Machanick, P. Bailey, T.L. 2014. Differential motif enrichment analysis of paired ChIP-seq experiments. *BMC Genomics*, **15**:752.

Lieb, J.D. Liu, X. Botstein, D. Brown, P.O. 2001. Promoter specific binding of Rap1 revealed by genome-wide maps of protein DNA association. *Nat. Genet*, **28**:327–334.

Lodish, H. Berk, A. Zipursky, S.L. Matsudaira, P. Baltimore, D. Darnell, J. 2000. *Molecular cell biology, 4th edition*. W.H. Freeman and company. New York. USA.

Luscher, B. Larsson, L. 1999. The basic region/helix-loop-helix/leucine zipper domain of Myc proto-oncoproteins: Function and regulation. *Oncogene*, **18**(19):2955-2966.

Matys, V. Kel-Margoulis, O.V. Fricke, E. Liebich, I. Land, S. Barre-Dirrie, A. Reuter, I. Chekmenev, D. Krull, M. Hornischer, K. Voss, N. Stegmaier, P. Lewicki-Potapov, B. Saxel, H. Kel, A.E. Wingender, E. 2003. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, **31**(1):374-378.

Nagarajan, N. Jones, N. Keich, U. 2005. Computing the P-value of the information content from an alignment of multiple sequences. *Bioinformatics*, **21**:i311-i318.

Neph, S. Vierstra, J. Stergachis, A.B. Reynolds, A.P. Haugen, E. Vernot, B. Thurman, R.E. John, S. Sandstrom, R. Johnson, A.K. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**(7414):83-90.

Orenstein, Y. Linhart, C. Shamir, R. 2012. Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PLos ONE*, **7**:e46165.

Orenstein, Y. Shamir, R. 2014. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, 1-10.

Park, P.J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**(10):669-680.

Piantados, J. Howlett, P. Boland, J. 2007. Matching the grade correlation coefficient using a copula with maximum disorder. *Journal of Industrial and Management Optimization*, **3**(2):305-312.

Primer, A. 2005. Receiver Operating Characteristic Analysis. *Acad Radiol*, **12**:909-916.

Ptashne, M. Gann, A. 2002. *Genes and Signals*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Ralston, A. Brown, W. 2008. Chromatin remodeling and DNase 1 sensitivity. *Nature Education*, **1**(1):15.

Rhee, H.S. Pugh, F.B. 2012. ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. *Curr Protoc Mol Biol*, **021**:24.

Robertson, G. Hirst, M. Bainbridge, M. Bilenky, M. Zhao, Y. Zeng, T. Euskirchen, G. Bernier, B. Varhol, R. Delaney, A. *et al*. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massive parallel sequencing. *Nat Methods*, **4**:651-657.

Rosenbloom, K.R. Dreszer, T.R. Pheasant, M. Barber, G.P. Meyer, L.R. Pohl, A. Raney, B.J. Wang, T. Hinrichs, A.S. Zweig, A.S. et al. 2010. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res*, **38**:D620-5.

Roth, F.P. Roth, Hughes, J.D. Estep, P.W. Church, G.M. 1998. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole genome mRNA quantitation. *Nature Biotechnol*, **16**:939-945.

Roulet, E. Busso, S. Camargo, A.A. Simpson, A.J. Mermod, N. Bucher, P. 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription factor binding sites. *Nature Biotechnol*, **20**:831-835.

Serandour, A.A. Brown, G.D. Cohen, J.D. 2013. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol*, **14**: R147.

Shaffer, A.L. Emre, N.C. Romesser, P.B. Staudt, L.M. 2009. IRF4: Immunity. Malignancy! Therapy? *Clin Cancer Res*, **15**(9):2954-2961.

Siggers, T. Gordan, R. 2014. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res*, **42**(4):2099-2111.

Slattery, M. Zhou, T. Yang, L. Machado, A.C. Gordan, R. Rohs, R. 2014. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, **39**(9):381-394.

Spitz, F. Furlong, E.M. 2012. Transcription Factors: from enhancer binding to developmental control. *Nat Rev Genet*, **13**(9):613-626.

Sterner, D.E. Berger, S.L. 2000. Acetylation of Histones and Transcription-Related Factors. *Microbial.Mol.Biol*, **64**(2):435-459.

Stormo, G.D. Zhao, Y. 2010. Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, **11**:751-760.

Sun, H. Yuan, Y. Wu, Y. Liu, H. Liu, J.S. Xie, H. 2010. Tmod: Toolbox of Motif Discovery. *Bioinformatics*, **26**(3):405-407.

Tanaka, E. Bailey, T.L. Keich, U. 2014. Improving MEME via a two-tiered significance analysis. *Bioinformatics*, **30**(14):1965-1973.

Thijs, G. Lescot, M. Marchal, K. *et al*., 2001. A high-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**(12):1113-1122.

Titz, B. Thomas, S. Rajagopala, S.V. Chiba, T. Ito, T. Uetz, P. 2006. Transcriptional activators in yeast. *Nucleic Acids Res*, **34**(3):955-967.

Tompa, M. Li, N. Bailey, T.L. *et al*. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnol*, **23**(1):137-144.

Villard, J. 2004. Transcription regulation and human diseases. *Swiss Med wkly*, **134**: 571-579.

Waldiminghaus, T. Skarstad, K. 2010. ChIP on chip: surprising results are often artifacts. *BMC Genomics*, **11**:414.

Wei, G-H. Badis, G. Berger, M.F. Kivioja, T. Palin, K. Enge, M. Bonke, M. Jolma, A. et al. 2010. Genome-wide analysis of ETS family DNA-binding in vitro and in vivo. *The EMBO Journal*, **29**(13): 2147-2160.

Weirauch, M.T. Cote, A. Norel, R. Annala, M. Zhao, Y. Riley, T.R. Julio, S. et al. 2013. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnol*, **31**:126-134.

Workman, C.T. Stormo, G.D. 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput*: 467-478.

Yamamoto, K. Ishihama, A. Busby, S.J. Grainger, D.C. 2011. The *Escherichia coli* K-12 MntR Miniregulon Include dps, Which Encodes the Major Stationary-Phase DNA-Binding Protein. *J.Bacteriol*, **193**(6):1477-1480.

Yanagisawa, S. 2001. The transcriptional activation domain of the plant-specific Dof1 Factor functions in plant, animal and yeast cells. *Plant Cell Physiol*, **42**(8):813-822.

Zeitlinger, J. Zinzen, R.P. Stark, A. Kellis, M. Zhang, H. Young, R.A. Levine, M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*, **21**(4):385-90.

Zheng, R. Blobel, G.A. 2010. GATA Transcription Factors and Cancer. *Genes and Cancer*, **1**(12):1178-1188.

Zhong, S. He, X. Bar-Joseph, Z. 2013. Predicting tissue specific transcription factor binding sites. *BMC Genomics*, **14**:796.

APPENDIX

UniPROBE 8-mer patterns for each TF and their counts in the TF's respective ChIP-seq peak region.

Key: 8-mer | reverse compliment count

Esrra least scoring 8-mers

```
AG.TT.ACCG | CGGT.AA.CT  66
AT.TGTTT.A | T.AAACA.AT  301
CGA.GACA.A | T.TGTC.TCG  98
GGA.AT.TCC | GGA.AT.TCC  124
CATT.AA.AA | TT.TT.AATG  352
AAAA.GACC | GGTC.TTTT    261
CT.CC.TTCA | TGAA.GG.AG  635
C.CCC.AGGA | TCCT.GGG.G  594
C.GAC.TTTG | CAAA.GTC.G  345
GCCCT.AT.A | T.AT.AGGGC  165
ATAG..CGGG | CCCG..CTAT  55
GA.CTT.GGA | TCC.AAG.TC  575
ATT..GGACA | TGTCC..AAT  182
AACTG.CCC | GGG.CAGTT    368
A.GT.GTTTA | TAAAC.AC.T  112
A.GGAA.GTA | TAC.TTCC.T  201
AGAT.TG.CC | GG.CA.ATCT  309
AT.AAT.ACA | TGT.ATT.AT  243
GACC.A.ATA | TAT.T.GGTC  89
CCTCG..TCA | TGA..CGAGG  133
GTA.AA.GAC | GTC.TT.TAC  127
AATG..ACAA | TTGT..CATT  274
CGC.CCACC | GGTGG.GCG    359
AG.TCACC.G | C.GGTGA.CT  698
ACA..GGTGA | TCACC..TGT  401
G.TAAAG.AA | TT.CTTTA.C  321
AG.A.TGTCA | TGACA.T.CT  353
A.GGGGAG.G | C.CTCCCC.T  1059
A.AGGG.TTA | TAA.CCCT.T  197
CTT.G.TCTA | TAGA.C.AAG  228
C.T.TTTGAC | GTCAAA.A.G  268
GCG.CCTAC | GTAGG.CGC    73
CT.TAGGAC | GTCCTA.AG    255
AG.T.ATTAT | ATAAT.A.CT  159
T.C.CTTTCA | TGAAAG.G.A  410
AT.ACACAC | GTGTGT.AT   273
A..CTTGTCA | TGACAAG..T  292
AT.GG.TCAA | TTGA.CC.AT  151
ACAC.TT.CC | GG.AA.GTGT  346
A.C.CATTGA | TCAATG.G.T  146
A.CTGC.CTT | AAG.GCAG.T  546
T..TGTGACA | TGTCACA..A  366
AA.GTGATT | AATCAC.TT    222
AAAGA.A.GA | TC.T.TCTTT  710
AT.TCCT.TC | GA.AGGA.AT  399
GTCA..ATAA | TTAT..TGAC  142
G.CCTCTC.C | G.GAGAGG.C  714
CTGG.TCAA | TTGA.CCAG   363
C.GACTTT.A | T.AAAGTC.G  273
GG.GGGTAA | TTACCC.CC  181
```

Esrra top scoring 8-mers

```
AAGGTCAT | ATGACCTT  2108
A.TGACCTT | AAGGTCA.T  1743
T.AAGGTCA | TGACCTT.A  3917
A..TGACCTT | AAGGTCA..T  2302
AAGGTCAC | GTGACCTT  3248
A..AAGGTCA | TGACCTT..T  1784
C.AAGGTCA | TGACCTT.G  3151
AAGGTCA.G | C.TGACCTT  3121
CCAAGGT.A | T.ACCTTGG  2957
GAAGGTCA | TGACCTTC  2308
AAGGTCA..C | G..TGACCTT  2596
AAGGTCA.C | G.TGACCTT  2680
T..AAGGTCA | TGACCTT..A  2571
AAGGTCA..G | C..TGACCTT  3516
AAGGTCAA | TTGACCTT  2416
CAAGG.CA.C | G.TG.CCTTG  2606
G.AAGGTCA | TGACCTT.C  1919
G..AAGGTCA | TGACCTT..C  3068
TCAAGG.CA | TG.CCTTGA  3634
AAGGTCA..A | T..TGACCTT  1971
AATGACCT | AGGTCATT  881
AAAGGTCA | TGACCTTT  1340
T.ACCTTGA | TCAAGGT.A  3099
C..AAGGTCA | TGACCTT..G  3004
CAAGGTCG | CGACCTTG  542
ATGACCT..A | T..AGGTCAT  1265
CAAGGT.A.C | G.T.ACCTTG  2324
A.AAGGTCA | TGACCTT.T  1450
AT.ACCTTG | CAAGGT.AT  1598
AAGGTCA.A | T.TGACCTT  2879
GACCTTGA | TCAAGGTC  2942
ATG.CCTTG | CAAGG.CAT  2001
G.CAAGGT.A | T.ACCTTG.C  2414
G.CAAGG.CA | TG.CCTTG.C  2994
CCAAGG.CA | TG.CCTTGG  3597
C.CAAGGT.A | T.ACCTTG.G  2492
CAAGG.CAA | TTG.CCTTG  2381
CAAGGT.AA | TT.ACCTTG  1934
T.GAGGTCA | TGACCTC.A  1209
CAAGG.CAC | GTG.CCTTG  2905
C.CAAGG.CA | TG.CCTTG.G  2979
CAAGGTC.C | G.GACCTTG  2515
TG.AGGTCA | TGACCT.CA  1979
AGAGGTCA | TGACCTCT  1908
A..CAAGGTC | GACCTTG..T  1695
CAAGGT.AC | GT.ACCTTG  2630
CCAAGGTC | GACCTTGG  2905
CAAGGTC..C | G..GACCTTG  2140
A.CAAGG.CA | TG.CCTTG.T  1589
CAAGGTCA | TGACCTTG  6073
CCTT.CC.TC | GA.GG.AAGG  171
```

Ets1 least scoring 8-mers

```
CC.T.ATGTA | TACAT.A.GG  17
GA.A.GGAAC | GTTCC.T.TC  72
CCG.AG.TTC | GAA.CT.CGG  91
AAATGG.A.A | T.T.CCATTT  49
AACAT.A.AA | TT.T.ATGTT  26
C..CCGGAGA | TCTCCGG..G  131
CA.TGC.GGA | TCC.GCA.TG  68
T.CCTGT.TA | TA.ACAGG.A  39
CAT.C.TTTA | TAAA.G.ATG  26
CGGATC.AG | CT.GATCCG  40
A.CTGCT.CC | GG.AGCAG.T  156
CC.GAAT.AA | TT.ATTC.GG  23
A.AA.AAGAG | CTCTT.TT.T  86
CTCCGGC..A | T..GCCGGAG  94
CGTA.ATCC | GGAT.TACG  9
CGGAATA..A | T..TATTCCG  12
GC.T.AAAAA | TTTTT.A.GC  50
T.G.GCAGGA | TCCTGC.C.A  108
ATAT.T.CTG | CAG.A.ATAT  22
AAG.T.AAAT | ATTT.A.CTT  50
TCT..GGAAA | TTTCC..AGA  93
AT.CGGCT.G | C.AGCCG.AT  37
```

```
A.A.GATGTT | AACATC.T.T  37
A.ATA.CCGC | GCGG.TAT.T  17
CCACGTCC | GGACGTGG  133
GATA.ATC.A | T.GAT.TATC  9
CACT.CC.GG | CC.GG.AGTG  253
ATA.GAA.CG | CG.TTC.TAT  15
GGGAGA.GA | TC.TCTCCC  188
GTGT.CGGA | TCCG.ACAC  42
CATC..GCTA | TAGC..GATG  6
ACTTC..ACC | GGT..GAAGT  47
AA..AAATAG | CTATTT..TT  48
A.GAG.AAGC | GCTT.CTC.T  122
ACAAAAA..C | G..TTTTTGT  66
AA.ATTCC.G | C.GGAAT.TT  31
ATCC..TTCA | TGAA..GGAT  38
TCCT.TAAA | TTTA.AGGA  53
AGGT.GGA.A | T.TCC.ACCT  69
CT.CCT.ATC | GAT.AGG.AG  55
A.CGG.AGGT | ACCT.CCG.T  55
AGG.AA.GGG | CCC.TT.CCT  199
CTAAG.GG.A | T.CC.CTTAG  46
CTT.GTTT.C | G.AAAC.AAG  73
CC..TACAAA | TTTGTA..GG  25
CCCG.AA.AG | CT.TT.CGGG  60
CG.ATGG.TA | TA.CCAT.CG  8
A.CATTAA.A | T.TTAATG.T  19
GG.TT.GAGA | TCTC.AA.CC  66

ACCGGAAG | CTTCCGGT  356
```

Ets1 top scoring 8-mers

```
ACTTCCGG | CCGGAAGT  593
ATTTCCGG | CCGGAAAT  115
AC.TCCGGT | ACCGGA.GT  200
A.TTCCGGT | ACCGGAA.T  221
ACCGGAAA | TTTCCGGT  120
A.CGGAAGT | ACTTCCG.T  299
ACATCCGG | CCGGATGT  110
AC.GGAAGT | ACTTCC.GT  310
CTTCCGG.A | T.CCGGAAG  259
A.TTCCGG.A | T.CCGGAA.T  177
T..ACCGGAA | TTCCGGT..A  60
TTCCGGAA | TTCCGGAA  79
G.ACCGGAA | TTCCGGT.C  215
AAC.TCCGG | CCGGA.GTT  154
A..CCGGAAG | CTTCCGG..T  299
CCGGAAAC | GTTTCCGG  194
C.A.TTCCGG | CCGGAA.T.G  301
A.CGGAAAT | ATTTCCG.T  86
AA.TTCCGG | CCGGAA.TT  143
T.CCGGAAA | TTTCCGG.A  105
CCGGAAGC | GCTTCCGG  450
ACCGGA.AT | AT.TCCGGT  44
C..TTTCCGG | CCGGAAA..G  179
ACTTCCG.C | G.CGGAAGT  422
A.CCGGAAA | TTTCCGG.T  109
ACCGGAA.C | G.TTCCGGT  183
A.CTTCCGG | CCGGAAG.T  203
A..CTTCCGG | CCGGAAG..T  201
CTTCCGG..A | T..CCGGAAG  221
C.TCCGGT.C | G.ACCGGA.G  207
ACCGGAA..A | T..TTCCGGT  105
CCGGAAG.A | T.CTTCCGG  252
C.CTTCCGG | CCGGAAG.G  643
CATCCGGC | GCCGGATG  89
ACCGGA.GC | GC.TCCGGT  170
CCGGAAG..C | G..CTTCCGG  388
CCGGAAA.C | G.TTTCCGG  136
C..CTTCCGG | CCGGAAG..G  574
TACCGGAA | TTCCGGTA  63
CCGGATGC | GCATCCGG  84
A.TTCCGGC | GCCGGAA.T  321
C..CATCCGG | CCGGATG..G  139
GCCGGAAA | TTTCCGGC  144
ATTTCCG.C | G.CGGAAAT  86
A.CCGGA.GT | AC.TCCGG.T  160
C.CCGGAAG | CTTCCGG.G  453
CATCCGG.C | G.CCGGATG  97
CCGGAA.TC | GA.TTCCGG  184
CTTCCGGC | GCCGGAAG  523

CCC..TACGC | GCGTA..GGG  52
```

FoxA2 least scoring 8-mers

```
ACAACAC.C | G.GTGTTGT  386
CTGTATT..A | T..AATACAG  698
AAACA.C.AA | TT.G.TGTTT  1317
A.AAGTA.AT | AT.TACTT.T  1042
AAC.ACACC | GGTGT.GTT  322
GTGT.GTTA | TAAC.ACAC  342
AT.TGAAT.A | T.ATTCA.AT  762
TATA.TATA | TATA.TATA  213
CAT.TTATA | TATAA.ATG  587
GAA.ATAAA | TTTAT.TTC  1619
TCCA.T.CCA | TGG.A.TGGA  757
AG.A.GTGTA | TACAC.T.CT  467
AT..ACATTA | TAATGT..AT  659
GC.C.CCCCC | GGGGG.G.GC  537
ATAC.GT.TG | CA.AC.GTAT  323
C.CCG.CAAA | TTTG.CGG.G  107
T.T.CCAATA | TATTGG.A.A  528
CAA.ACAAA | TTTGT.TTG  1848
AA.TGG.AAA | TTT.CCA.TT  1468
TA.AGA.GGA | TCC.TCT.TA  580
TT.GG.AAAA | TTTT.CC.AA  1130
CCAC.CCA.A | T.TGG.GTGG  642
ACA.TAT.TG | CA.ATA.TGT  683
AC.GTGT.TA | TA.ACAC.GT  599
C..TTCTGTA | TACAGAA..G  696
GG..GGCGAA | TTCGCC..CC  118
T.CT.ATCAA | TTGAT.AG.A  549
T.AA.ATTTA | TAAAT.TT.A  1871
C.CCCT.ATC | GAT.AGGG.G  371
A.GTTAAAC | GTTTAAC.T  582
AT..TTATTC | GAATAA..AT  772
AGT.TTATA | TATAA.ACT  449
A.TAATG.AT | AT.CATTA.T  611
TA.AAAT.AA | TT.ATTT.TA  1955
ACCAA.TAT | ATA.TTGGT  479
C.AAG.AAAC | GTTT.CTT.G  1562
AGGG..ATGG | CCAT..CCCT  776
ATA.A.TAGA | TCTA.T.TAT  492
AGTA.A.ACG | CGT.T.TACT  197
GT.GA.TGGA | TCCA.TC.AC  375
AT.CACAA.A | T.TTGTG.AT  711
AA.TAA.ACC | GGT.TTA.TT  614
TAA..GAAAA | TTTTC..TTA  1560
ACAC..AATA | TATT..GTGT  551
AA.AATA.GT | AC.TATT.TT  727
CAAAGTA.A | T.TACTTTG  1455
CAAACA.C.A | T.G.TGTTTG  1079
ACAT.A.AGA | TCT.T.ATGT  864
ATTGA.AAA | TTT.TCAAT  1001

GTAAACAA | TTGTTTAC  2436
```

FoxA2 top scoring 8-mers

```
A.GTAAACA | TGTTTAC.T  3558
A..GTAAACA | TGTTTAC..T  3261
TAAACAAA | TTTGTTTA  2671
G..TGTTTAC | GTAAACA..C  1924
GGTAAACA | TGTTTACC  1863
GTAAATAA | TTATTTAC  1799
AGTAAACA | TGTTTACT  3222
GTAAACA.A | T.TGTTTAC  2771
AGTAAATA | TATTTACT  2278
T..GTAAACA | TGTTTAC..A  3336
A.TGTTTAC | GTAAACA.T  1997
A..TAAACAA | TTGTTTA..T  2502
A..GTAAATA | TATTTAC..T  2165
GTAAACA..A | T..TGTTTAC  2425
```

TATTTACA | TGTAAATA 2327
G.GTAAATA | TATTTAC.C 1465
TGTAAACA | TGTTTACA 3386
GTAAATA.A | T.TATTTAC 1753
A.GTAAATA | TATTTAC.T 2331
T..TTGTTTA | TAAACAA..A 2246
ATAAACAA | TTGTTTAT 1601
T..GTAAATA | TATTTAC..A 2440
GT.AATAAA | TTTATT.AC 1471
G.TT.TTTAC | GTAAA.AA.C 1156
C..TGTTTAC | GTAAACA..G 2588
AT.TAAACA | TGTTTA.AT 2133
G..TAAACAA | TTGTTTA..C 1450
ATGTTTAC | GTAAACAT 2050
T..ATAAACA | TGTTTAT..A 1796
TGTAAA.AA | TT.TTTACA 2166
GTAAATA..A | T..TATTTAC 1540
CAAACAAA | TTTGTTTG 2658
T.GTAAACA | TGTTTAC.A 1443
A.TAAACAA | TTGTTTA.T 2260
TATTGACA | TGTCAATA 1119
ATGTAAA.A | T.TTTACAT 2414
A.GTAAA.AA | TT.TTTAC.T 2355
TAAATAAA | TTTATTTA 2485
G.TGTTTAC | GTAAACA.C 1720
A.ATAAA.AA | TT.TTTAT.T 2424
A..TTGTTTA | TAAACAA..T 1441
T..TAAACAA | TTGTTTA..A 1359
G.TAAACAA | TTGTTTA.C 1249
A.TTGTTTA | TAAACAA.T 1523
TA..TAAACA | TGTTTA..TA 2190
T.T.TAAACA | TGTTTA.A.A 2817
A.TATTTAC | GTAAATA.T 2054
GTAAA.AAA | TTT.TTTAC 2201
T.TTGTTTA | TAAACAA.A 1749

AAT.TC.GAG | CTC.GA.ATT 105

Gata3 least scoring 8-mers

CAATC..CAA | TTG..GATTG 79
AA.AGA.TTA | TAA.TCT.TT 196
TCA..AATCA | TGATT..TGA 137
ATCTTC..CA | TG..GAAGAT 117
AG.TTA.AAG | CTT.TAA.CT 122
C.AA.CTCAA | TTGAG.TT.G 92
CA.T.ATCGG | CCGAT.A.TG 5
A.CGATA.CC | GG.TATCG.T 4
CATC.AT.AG | CT.AT.GATG 60
AATC..TCGC | GCGA..GATT 3
ATCA.CAG.C | G.CTG.TGAT 85
AAGA.A.CTG | CAG.T.TCTT 163
ATA.CTCT.C | G.AGAG.TAT 68
GATT.AC.GC | GC.GT.AATC 29
GTGA.A.AAA | TTT.T.TCAC 198
A.AATC.AGT | ACT.GATT.T 118
GTCT.GA.AA | TT.TC.AGAC 73
ATA.TCAAA | TTTGA.TAT 154
AGA.CCT.AT | AT.AGG.TCT 86
A.CGGAT.AC | GT.ATCCG.T 3
AT.TACA.TA | TA.TGTA.AT 144
CATCAT.A.A | T.T.ATGATG 139
G.GAT.ATCC | GGAT.ATC.C 22
AATA.TC.AT | AT.GA.TATT 122
AT..ATATCC | GGATAT..AT 61
CA.AAGA.TA | TA.TCTT.TG 115
ATA.A.GCTG | CAGC.T.TAT 97
C.ATCT.CGG | CCG.AGAT.G 7
AGA.CA.ACA | TGT.TG.TCT 184
AATAT.T.TA | TA.A.ATATT 386
A.TCTTTT.A | T.AAAAGA.T 257
CA..ATACAA | TTGTAT..TG 118
GG.AA.ATCA | TGAT.TT.CC 115
C.AT.ACGTA | TACGT.AT.G 11
CGTATC.T.A | T.A.GATACG 15
AATT.A.CAT | ATG.T.AATT 179
AT.G.CAAAT | ATTTG.C.AT 122
AAAT.G.ACG | CGT.C.ATTT 20
CTAT.AG.AA | TT.CT.ATAG 94
C.GT.TAAGA | TCTTA.AC.G 45
CGAT.TCC.C | G.GGA.ATCG 9
G.AATCATA | TATGATT.C 82
A.A.TAGTAA | TTACTA.T.T 123
G.TATCG.CC | GG.CGATA.C 7
GGT.T.CAAA | TTTG.A.ACC 112
T.TCT.TCTA | TAGA.AGA.A 202
ATCTA..GGG | CCC..TAGAT 44
AA.AG.AATA | TATT.CT.TT 321
AAT.AGT.AT | AT.ACT.ATT 185

AGATAAGA | TCTTATCT 364

Gata3 top scoring 8-mers

CTTATCT..A | T..AGATAAG 254
A.AGATAA.A | T.TTATCT.T 441
CTTATCTC | GAGATAAG 203
A.AGATAAG | CTTATCT.T 286
A..AGATAAG | CTTATCT..T 258
AGATAA.A.A | T.T.TTATCT 382
AGATAAG..A | T..CTTATCT 285
AGATAAGC | GCTTATCT 171
AGATAACA | TGTTATCT 244
GAGATAA.A | T.TTATCTC 284
CTTATCT.A | T.AGATAAG 234
AAGATAA.A | T.TTATCTT 426
AGAGATAA | TTATCTCT 282
AGATAAG.A | T.CTTATCT 356
G.A.AGATAA | TTATCT.T.C 239
GA..AGATAA | TTATCT..TC 232
AGATAA.A.C | G.T.TTATCT 360
AAGATAAG | CTTATCTT 265
CTTATCTA | TAGATAAG 146
AGATAA.AG | CT.TTATCT 227
A..CTTATCT | AGATAAG..T 231
A.AAGATAA | TTATCTT.T 338
T.TTATCTA | TAGATAA.A 273
AGAT..TATC | GATA..ATCT 482
AT..AGATAA | TTATCT..AT 265
A.A.AGATAA | TTATCT.T.T 395
AA.AGATAA | TTATCT.TT 342
C.AGATAA.A | T.TTATCT.G 218
AGATAA.AT | AT.TTATCT 487
AGATAA..GA | TC..TTATCT 187
GAGATAAC | GTTATCTC 92
AGATAACG | CGTTATCT 24
T..GAGATAA | TTATCTC..A 189
AGATAA.AA | TT.TTATCT 446
T.GAGATAA | TTATCTC.A 199
T..GATAAGA | TCTTATC..A 188
C.AGATAAG | CTTATCT.G 148
ATAGATAA | TTATCTAT 164
G..TAGATAA | TTATCTA..C 130
A.GATAAGA | TCTTATC.T 183
TA.AGATAA | TTATCT.TA 261
T.AGATAA.A | T.TTATCT.A 362
CGAGATAA | TTATCTCG 24
C.TTATCTC | GAGATAA.G 181
GAT.TTATC | GATAA.ATC 453
TCTTATCA | TGATAAGA 179
A.T.TTATCT | AGATAA.A.T 362
AGATAAG..C | G..CTTATCT 190

Hnf4a least scoring 8-mers

C.CTCTGC.C | G.GCAGAG.G 825
AC.TTAGG.C | G.CCTAA.GT 120
A..CCCGGAC | GTCCGGG..T 96
CGAA.TCCG | CGGA.TTCG 39
GTC.AAGTA | TACTT.GAC 143
CTCTT.G.CC | GG.C.AAGAG 389
GGTCGG..GA | TC..CCGACC 72
GGC.AGGT.A | T.ACCT.GCC 247
CGAGGT.GA | TC.ACCTCG 61

AC.ATGAC.C | G.GTCAT.GT 162
AT.ACCT.CG | CG.AGGT.AT 40
A..GGTTCGC | GCGAACC..T 21
ATTG.GG.TC | GA.CC.CAAT 99
C..AAGGTAC | GTACCTT..G 115
AATGGT.CA | TG.ACCATT 204
CAAT.GG.CA | TG.CC.ATTG 203
A.GG.CCACC | GGTGG.CC.T 284
G.ACCA.GCC | GGC.TGGT.C 249
ATGG.C.TCG | CGA.G.CCAT 51
G.ACTCT.TA | TA.AGAGT.C 228
GGTT..CCTA | TAGG..AACC 100
CT.TCG.CCC | GGG.CGA.AG 105
G.AC.TGAAC | GTTCA.GT.C 192
GGT..CCTTA | TAAGG..ACC 143
ATGG.TG.CC | GG.CA.CCAT 222
A.CGAAC.TT | AA.GTTCG.T 48
CAC.TCT.AA | TT.AGA.GTG 257
CTCA.A.TCC | GGA.T.TGAG 568
AT.GAAC.TA | TA.GTTC.AT 101
C.G.TCCACA | TGTGGA.C.G 299
AATT..GACT | AGTC..AATT 226
CCAT.GCAA | TTGC.ATGG 212
A..GTCAGAG | CTCTGAC..T 512
G.TGGG.GTC | GAC.CCCA.C 311
GAT..ACCTA | TAGGT..ATC 70
A.GTT.TACC | GGTA.AAC.T 111
AAC.GTGAC | GTCAC.GTT 239
CG.ACG.ACC | GGT.CGT.CG 31
ATGTG..CTC | GAG..CACAT 354
A..GTACGAA | TTCGTAC..T 25
ACCCA.TAC | GTA.TGGGT 104
CCCGCCC..C | G..GGGCGGG 827
C.GGTCG.GG | CC.CGACC.G 128
G.A.CGTACC | GGTACG.T.C 13
AAG.GCGG.A | T.CCGC.CTT 82
T.AAAA.GGA | TCC.TTTT.A 488
AGAGGG.CC | GG.CCCTCT 444
AAAGTA.A.C | G.T.TACTTT 354
ACCTT..GCT | AGC..AAGGT 345
CCTG.TACA | TGTA.CAGG 233

Hnf4a top scoring 8-mers

GGGGTCAA | TTGACCCC 322
CGGGGTCA | TGACCCCG 180
A..TGACCCC | GGGGTCA..T 307
AAAGGTCA | TGACCTTT 1233
GGGTCCAA | TTGGACCC 390
AGGGTCCA | TGGACCCT 573
C.GGGGTCA | TGACCCC.G 572
A.TGACCCC | GGGGTCA.T 271
AAAGTCCA | TGGACTTT 2014
C..GGGGTCA | TGACCCC..G 412
A.GGGTCCA | TGGACCC.T 253
G..GGGGTCA | TGACCCC..C 446
TGACCCCA | TGGGGTCA 564
G.TGACCCC | GGGGTCA.C 301
A..TGGACCC | GGGTCCA..T 312
G..TGACCCC | GGGGTCA..C 300
AGGGGTCA | TGACCCCT 486
A.GGGGTCA | TGACCCC.T 288
G.GGGGTCA | TGACCCC.C 339
A..GGGGTCA | TGACCCC..T 279
G..GGGTCCA | TGGACCC..C 382
T.GGGTCCA | TGGACCC.A 319
G.TGGACCC | GGGTCCA.C 300
A.TGGACCC | GGGTCCA.T 228
GGGGTCAC | GTGACCCC 308
T..GGGTCCA | TGGACCC..A 361
GGGTCCAC | GTGGACCC 298
GGGGTCA..A | T..TGACCCC 300
GGGTCCA..A | T..TGGACCC 249
CGGGGTC.A | T.GACCCCG 101
CTGACCCC | GGGGTCAG 670
T.GGGGTCA | TGACCCC.A 320
C..TGGACCC | GGGTCCA..G 576
GGGTCCA.A | T.TGGACCC 471
AT.GACCCC | GGGGTC.AT 96
A..GGGTCCA | TGGACCC..T 225
GGGGTC.AA | TT.GACCCC 256
C.GGGTCCA | TGGACCC.G 633
T..GGGGTCA | TGACCCC..A 375
ATGGACCC | GGGTCCAT 184
G.T.GACCCC | GGGGTC.A.C 133
GGGGTCA.A | T.TGACCCC 438
AA..TGACCC | GGGTCA..TT 299
G.GTCCAAA | TTTGGAC.C 500
AAAGGTC.A | T.GACCTTT 499
A.AGTCCAA | TTGGACT.T 1009
AAGGGTC.A | T.GACCCTT 264
TGGACCCA | TGGGTCCA 464
C..GGGTCCA | TGGACCC..G 487
C.TGACCCC | GGGGTCA.G 496

Irf4 least scoring 8-mers

A.ATTCG.GA | TC.CGAAT.T 41
CG..ACCAAC | GTTGGT..CG 54
GA.T.AGTAC | GTACT.A.TC 104
GT.TCTGG.A | T.CCAGA.AC 315
A.TC.GTATC | GATAC.GA.T 115
CTTCT.AGA | TCT.AGAAG 499
CCCG.TA.TC | GA.TA.CGGG 37
ATACTT.T.G | C.A.AAGTAT 128
A.T.TTATTT | AAATAA.A.T 724
A.TCTCAC.A | T.GTGAGA.T 230
C.C.GTTCCG | CGGAAC.G.G 95
CAG.TTC.AC | GT.GAA.CTG 247
GT.CCAAGA | TCTTGG.AC 209
CCAC.TT.GA | TC.AA.GTGG 174
GG.ATCA.AC | GT.TGAT.CC 119
GG.TTCTCA | TGAGAA.CC 376
ACTAC..CTA | TAG..GTAGT 91
A.CG.AATCA | TGATT.CG.T 43
TCG.AA.GAA | TTC.TT.CGA 68
AGT.A.TACT | AGTA.T.ACT 224
A.G.AACTGC | GCAGTT.C.T 292
CGAA.GTAG | CTAC.TTCG 27
A.GAGTTT.A | T.AAACTC.T 341
TCATA.C.AA | TT.G.TATGA 200
ATA.TAG.AC | GT.CTA.TAT 101
CGGT..CAAA | TTTG..ACCG 53
A.CCGAA.AC | GT.TTCGG.T 51
CAA.T.CCGG | CCGG.A.TTG 96
T.GTGTT.CA | TG.AACAC.A 235
AA.ACGA.AG | CT.TCGT.TT 106
AAT.TT.GTA | TAC.AA.ATT 204
CAATTCT..A | T..AGAATTG 273
GTT.CCAT.A | T.ATGG.AAC 226
ACCTTT.CG | CG.AAAGGT 50
AACCGGA.G | C.TCCGGTT 107
GGGTTC..GA | TC..GAACCC 162
AA..ATCGAT | ATCGAT..TT 28
G.ACT.CCGA | TCGG.AGT.C 43
AC.GG.ATGG | CCAT.CC.GT 156
AC.AATAGT | ACTATT.GT 108
AACTG..AGT | ACT..CAGTT 470
C.AGAATC.A | T.GATTCT.G 161
C.GAT.CCTA | TAGG.ATC.G 93
ACCTG.GAG | CTC.CAGGT 330
A.TGATA.CT | AG.TATCA.T 180
T.A.AGAAAA | TTTTCT.T.A 944
CAAT..AGAA | TTCT..ATTG 328
TAA.AC.AAA | TTT.GT.TTA 399
CGGAAC..CA | TG..GTTCCG 83
A.CG.GAGTG | CACTC.CG.T 88

Irf4 top scoring 8-mers

A.CGAAACC | GGTTTCG.T 149
GGTTTCG..A | T..CGAAACC 64
ACCGAAAC | GTTTCGGT 161
G.A.CGAAAC | GTTTCG.T.C 148

AGTTTCGG | CCGAAACT 118
GGTTTCGA | TCGAAACC 61
AA.CGAAAC | GTTTCG.TT 237
G..CCGAAAC | GTTTCGG..C 106
CGAAACCA | TGGTTTCG 120
ATCGAAAC | GTTTCGAT 69
A..CGAAACT | AGTTTCG..T 145
CGAAACTA | TAGTTTCG 49
AGTTTCG..A | T..CGAAACT 85
A.CGAAAC.A | T.GTTTCG.T 113
A.CGAAACT | AGTTTCG.T 165
CGAAACT..C | G..AGTTTCG 86
AGTTTCG.A | T.CGAAACT 78
AGTTTCGA | TCGAAACT 89
CGAAACCG | CGGTTTCG 49
GGTTTCG.A | T.CGAAACC 68
CGAAACC..A | T..GGTTTCG 102
CGAAACT.C | G.AGTTTCG 96
G..CGAAACC | GGTTTCG..C 70
CCGAAACC | GGTTTCGG 142
GTTTCG.AA | TT.CGAAAC 84
GTTTCG.TA | TA.CGAAAC 56
G.T.CGAAAC | GTTTCG.A.C 56
C..GGTTTCG | CGAAACC..G 92
A.CCGAAAC | GTTTCGG.T 147
A.CGAAAC.G | C.GTTTCG.T 141
GA..CGAAAC | GTTTCG..TC 156
A..AGTTTCG | CGAAACT..T 90
GTTTCGG.A | T.CCGAAAC 69
T.CGAAAC.A | T.GTTTCG.A 50
A..CGAAACC | GGTTTCG..T 132
AGTTTCG.C | G.CGAAACT 55
CGAAACC.A | T.GGTTTCG 103
A.TTTCGGT | ACCGAAA.T 122
A.GGTTTCG | CGAAACC.T 77
GT..CGAAAC | GTTTCG..AC 47
AAGTTTCG | CGAAACTT 76
CGAAACTC | GAGTTTCG 92
AGTTTCG..G | C..CGAAACT 72
GTTTCGGA | TCCGAAAC 72
CGAAACT.A | T.AGTTTCG 101
C.A.CGAAAC | GTTTCG.T.G 67
CCGAAAC.A | T.GTTTCGG 92
AC.GAAACT | AGTTTC.GT 664
CGAAAC..GA | TC..GTTTCG 77
CGAAAC.A.A | T.T.GTTTCG 97

Klf7 least scoring 8-mers

AGATG.AT.A | T.AT.CATCT 74
T.TAATATA | TATATTA.A 43
A.TATATTA | TAATATA.T 46
TAAACT.CA | TG.AGTTTA 111
A..AATATCA | TGATATT..T 40
GGTTG..ATA | TAT..CAACC 46
CACCC.TG.C | G.CA.GGGTG 340
ACGATA..TA | TA..TATCGT 10
CGA..TTATC | GATAA..TCG 16
T.G.ACAGTA | TACTGT.C.A 65
TCAT.ATCA | TGAT.ATGA 72
A.ATATAG.T | A.CTATAT.T 50
AAGAAG.AA | TT.CTTCTT 408
GA..TAATAA | TTATTA..TC 58
CCAAA..AAA | TTT..TTTGG 325
T.TGGA.CTA | TAG.TCCA.A 96
ATTCA.G.TG | CA.C.TGAAT 100
A.ATAAA.TT | AA.TTTAT.T 110
CGTCCGC.C | G.GCGGACG 194
C.AT.CAACA | TGTTG.AT.G 57
AT..ACTAAC | GTTAGT..AT 36
C.AG.TATTA | TAATA.CT.G 37
TA.TATTGA | TCAATA.TA 27
ATTA.AAAT | ATTT.TAAT 93
TAGT.A.TGA | TCA.T.ACTA 88
T.TCAC.CTA | TAG.GTGA.A 71
CC..GCCCTA | TAGGGC..GG 316
TC.ATC.CGA | TCG.GAT.GA 33
GGGC.TGTA | TACA.GCCC 92
AT..ACTTTA | TAAAGT..AT 74
AATAG.TAA | TTA.CTATT 62
AT.G.CAGAA | TTCTG.C.AT 100
T.CGAA.TTA | TAA.TTCG.A 22
T.TTT.ATCA | TGAT.AAA.A 103
C.TT.TGATA | TATCA.AA.G 66
CTTAAT.AA | TT.ATTAAG 91
ATAGT.TAA | TTA.ACTAT 42
AAA.A.CATA | TATG.T.TTT 69
A.ATATTG.A | T.CAATAT.T 42
TA.AAGTGA | TCACTT.TA 107
A.CCA.TAAT | ATTA.TGG.T 66
TAG..TATAA | TTATA..CTA 55
TT.GATA.AA | TT.TATC.AA 68
GGAG.T.AAC | GTT.A.CTCC 153
C.GGGGCG.C | G.CGCCCC.G 404
CAATTAT..A | T..ATAATTG 48
A.GGTGT.AT | AT.ACACC.T 73
AACATAA.A | T.TTATGTT 84
A.TTGG.ACA | TGT.CCAA.T 100
T.C.TTGTTA | TAACAA.G.A 121

Klf7 top scoring 8-mers

G.C.ACGCCC | GGGCGT.G.C 1008
G.CACGCCC | GGGCGTG.C 1432
CCACGCCC | GGGCGTGG 1706
CCCCGCCC | GGGCGGGG 2654
GCC.CGCCC | GGGCG.GGC 2764
GCCCC.CCC | GGG.GGGGC 3261
G.CC.CGCCC | GGGCG.GG.C 2073
ACC.CGCCC | GGGCG.GGT 1088
GC.ACGCCC | GGGCGT.GC 1172
CC.CGCCCC | GGGGCG.GG 2178
GCCAC.CCC | GGG.GTGGC 2111
GA..ACGCCC | GGGCGT..TC 415
G..CACGCCC | GGGCGTG..C 1174
C.ACGCCCC | GGGGCGT.G 819
GGGGCGTA | TACGCCCC 105
G.CCCGCCC | GGGCGGG.C 2309
A.GGGCG.GG | CC.CGCCC.T 1396
CCCCACCC | GGGTGGGG 2647
CACGCCC.C | G.GGGCGTG 1047
AC.ACGCCC | GGGCGT.GT 544
CCACACCC | GGGTGTGG 2009
AA..GGGCGG | CCGCCC..TT 431
A..GGGCGGG | CCCGCCC..T 1133
G..ACGCCCC | GGGGCGT..C 729
CCCGCCC.A | T.GGGCGGG 759
A.GGGCGT.G | C.ACGCCC.T 620
CCCC.CCCA | TGGG.GGGG 2049
CACGCCC..A | T..GGGCGTG 513
CACGCCCA | TGGGCGTG 858
A.A.GGGCGT | ACGCCC.T.T 243
AACC.C.CCC | GGG.G.GGTT 706
A.CC.CGCCC | GGGCG.GG.T 1324
CC.CGCCC.C | G.GGGCG.GG 2164
ACGCCCCT | AGGGGCGT 512
ACGCCCCA | TGGGGCGT 300
AGGGCG.GG | CC.CGCCCT 1520
G.C.CCGCCC | GGGCGG.G.C 2093
G.CCAC.CCC | GGG.GTGG.C 1787
GG..ACGCCC | GGGCGT..CC 1024
G..ACGCCCA | TGGGCGT..C 685
GCC.CACCC | GGGTG.GGC 2286
AGGGCGT.G | C.ACGCCCT 641
A.GGGCGTG | CACGCCC.T 764
G.C.CGCCCA | TGGGCG.G.C 1255
ACCAC.CCC | GGG.GTGGT 1297
CCGCCCCA | TGGGGCGG 736
ACGCCCC.C | G.GGGGCGT 546
CCGCCCC..C | G..GGGGCGG 1633
AGGG.GTGG | CCAC.CCCT 1505
G.A.ACGCCC | GGGCGT.T.C 363

Mafk least scoring 8-mers

C.GCCA.AAA | TTT.TGGC.G 173
ACT.AA.CAA | TTG.TT.AGT 208
ATTCTA.TG | CA.TAGAAT 163
T.TT.CGCTA | TAGCG.AA.A 22
TA.G.AGTAA | TTACT.C.TA 125
TGAA.TA.AA | TT.TA.TTCA 370
CTTGCTG.A | T.CAGCAAG 188
GAAA.AACA | TGTT.TTTC 363
CGTGA.AA.A | T.TT.TCACG 43
GTAA..GTTA | TAAC..TTAC 99
A..TCCATTT | AAATGGA..T 328
TATCGTAA | TTACGATA 11
A.CTTGA.AT | AT.TCAAG.T 195
AATTTA..GC | GC..TAAATT 174
TAAG.A.CAA | TTG.T.CTTA 198
CATT.TGC.C | G.GCA.AATG 157
ATA.ATAA.G | C.TTAT.TAT 190
AT.TT.GGTG | CACC.AA.AT 134
ACCAT..AAA | TTT..ATGGT 232
AA.CT.ATTT | AAAT.AG.TT 422
CA.TCA.AAG | CTT.TGA.TG 200
TC.ACGAAA | TTTCGT.GA 27
T.ATGAG.GA | TC.CTCAT.A 152
GTA.TGT.AA | TT.ACA.TAC 115
TATAAT..AA | TT..ATTATA 329
G.ATA.TATA | TATA.TAT.C 147
AAT.ACAT.C | G.ATGT.ATT 179
A.TG.GTTTT | AAAAC.CA.T 293
AT.T.AGAAA | TTTCT.A.AT 507
CA.AGTGCA | TGCACT.TG 148
AG.GTGTTA | TAACAC.CT 101
ATGTACAA | TTGTACAT 146
CAA..ATAAG | CTTAT..TTG 152
AAT.C.CCTC | GAGG.G.ATT 121
G.A.GCAGAA | TTCTGC.T.C 314
TGA..AGAAA | TTTCT..TCA 455
A.ATAAG.TA | TA.CTTAT.T 190
AAA.TC.CTA | TAG.GA.TTT 139
CA.AA.GCAA | TTGC.TT.TG 278
AATGG.A.TA | TA.T.CCATT 214
CAT.AGCA.A | T.TGCT.ATG 267
CA.CAAT.AA | TT.ATTG.TG 251
A.TT.TTAGG | CCTAA.AA.T 184
AATG.TAAG | CTTA.CATT 185
T..GCTTACA | TGTAAGC..A 116
ATT.TCC.AC | GT.GGA.AAT 122
T.TC.TGGAA | TTCCA.GA.A 243
TA.TGGTTA | TAACCA.TA 88
AT..TATTCC | GGAATA..AT 167
CTT..TTTCA | TGAAA..AAG 439

Mafk top scoring 8-mers

AA.TGCTGA | TCAGCA.TT 1670
A..TGCTGAC | GTCAGCA..T 1301
GTCAGCA..A | T..TGCTGAC 905
AA..TGCTGA | TCAGCA..TT 1628
A.TTGCTGA | TCAGCAA.T 898
TCAGCA.AA | TT.TGCTGA 1193
GTCAGCAA | TTGCTGAC 900
TA..TGCTGA | TCAGCA..TA 601
GTCAGCA.A | T.TGCTGAC 1055
TCAGCAAA | TTTGCTGA 1176
A.TCAGCA.T | A.TGCTGA.T 1633
A.TCAGCA.A | T.TGCTGA.T 1564
A.A.TGCTGA | TCAGCA.T.T 1300
AAAA.TGCA | TGCA.TTTT 473
A.TGCTGAC | GTCAGCA.T 1123
AT.TGCTGA | TCAGCA.AT 937
AATGCTGA | TCAGCATT 1151
AAA.TGCT.A | T.AGCA.TTT 1235
A.AAAA.TGC | GCA.TTTT.T 548
AAAA.TGC.G | C.GCA.TTTT 613
AAAAA.TGC | GCA.TTTTT 528
A.T.TGCTGA | TCAGCA.A.T 1028
ATCAGCA..A | T..TGCTGAT 386
A.TGCTGAT | ATCAGCA.T 419
AAAATTGC | GCAATTTT 332
CGTCAGCA | TGCTGACG 128
T..GTCAGCA | TGCTGAC..A 1392
GT.AGCA.AA | TT.TGCT.AC 641
GTCAGC..AA | TT..GCTGAC 624
T.GTCAGCA | TGCTGAC.A 483
A.GTCAGCA | TGCTGAC.T 747
T.TTGCTGA | TCAGCAA.A 922
AAAAATGC | GCATTTTT 634
ATGCTGAC | GTCAGCAT 867
T.GCAAAAA | TTTTTGC.A 283
AATTTTGC | GCAAAATT 352
A.GCA.TTTT | AAAA.TGC.T 461
AAA.TGCAA | TTGCA.TTT 437
T.A.TGCTGA | TCAGCA.T.A 652
AAA..GCTGA | TCAGC..TTT 1262
AAAA.TGCT | AGCA.TTTT 827
T.T.TGCTGA | TCAGCA.A.A 923
AA.TTGCT.A | T.AGCAA.TT 582
AAAA.TGC.A | T.GCA.TTTT 346
C..GTCAGCA | TGCTGAC..G 545
ATTTTGCA | TGCAAAAT 396
GTCAGC.A.A | T.T.GCTGAC 495
ATTTTTGC | GCAAAAAT 399
AAT.TGCT.A | T.AGCA.ATT 691
TCAGCA..AA | TT..TGCTGA 914

Max least scoring 8-mers

C.TGGA.AAA | TTT.TCCA.G 190
C.TG.TTCTA | TAGAA.CA.G 112
GAG.TAA.GA | TC.TTA.CTC 129
AG.AAC.ACG | CGT.GTT.CT 100
CAAAA.CAC | GTG.TTTTG 268
CATGGCC.C | G.GGCCATG 419
ATTAC..ATG | CAT..GTAAT 41
G..CCACGCC | GGCGTGG..C 514
CAT.TGGA.C | G.TCCA.ATG 135
CA..CTCCCC | GGGGAG..TG 625
CTTC.CCC.A | T.GGG.GAAG 340
CCACGA.TC | GA.TCGTGG 112
GAGTGG.A.A | T.T.CCACTC 196
AAGCG..CAC | GTG..CGCTT 138
CAC.C.CACA | TGTG.G.GTG 634
ACC.CACGT | ACGTG.GGT 136
GC.CGAGCA | TGCTCG.GC 232
CCCCA..AAA | TTT..TGGGG 251
GAT.AAG.CA | TG.CTT.ATC 118
AC.ATGCTG | CAGCAT.GT 127
ACGTGCG..A | T..CGCACGT 112
ATTCTCT.A | T.AGAGAAT 111
C.C.TGCGGC | GCCGCA.G.G 452
GA.AA.ATTA | TAAT.TT.TC 85
TACCCA.CA | TG.TGGGTA 103
AATAA..CAC | GTG..TTATT 103
GCGCAT.CC | GG.ATGCGC 192
TGTT.TGAA | TTCA.AACA 149
C.AGATCAA | TTGATCT.G 68
A.GTGGCAC | GTGCCAC.T 195
CC.ATTA.GC | GC.TAAT.GG 30
ACAA.ATGG | CCAT.TTGT 206
GCCA..TGAC | GTCA..TGGC 294
CGTGTGA..C | G..TCACACG 136
T.TGC.CCTA | TAGG.GCA.A 85
AA.ATGTGT | ACACAT.TT 145
AAAGG.G.GG | CC.C.CCTTT 563
C.A.ATTATC | GATAAT.T.G 31
TGGGG.A.AA | TT.T.CCCCA 314
T.AGA.CTAA | TTAG.TCT.A 80
ACT.AAGA.A | T.TCTT.AGT 136
AA..GAGCAT | ATGCTC..TT 105
C.AT.AAGCA | TGCTT.AT.G 81
A..ACGTTGG | CCAACGT..T 51
A.G.ATGGGG | CCCCAT.C.T 258
C.TGTC.ATA | TAT.GACA.G 31

```
AAAAG..CGT | ACG..CTTTT  89
C.GAAGCA.G | C.TGCTTC.G  308
GCAC.AG.CA | TG.CT.GTGC  196
CATG..ACAA | TTGT..CATG  82
```

Max top scoring 8-mers

```
A..CACGTGG | CCACGTG..T  533
CA..CACGTG | CACGTG..TG  508
A.CCACGTG | CACGTGG.T  528
GCACGTG..A | T..CACGTGC  264
CACGTGCG | CGCACGTG  381
CACGTGGA | TCCACGTG  419
AG.CACGTG | CACGTG.CT  581
G.CACGTG.A | T.CACGTG.C  335
A.CACGTG.G | C.CACGTG.T  412
A..CACGTGT | ACACGTG..T  258
CACGTGG..A | T..CCACGTG  440
A.CACGTGG | CCACGTG.T  583
A..CCACGTG | CACGTGG..T  485
CCACGTG.C | G.CACGTGG  877
AC.CACGTG | CACGTG.GT  355
CCACGTGG | CCACGTGG  513
CCACGTGC | GCACGTGG  698
A.GCACGTG | CACGTGC.T  330
CACGTGAC | GTCACGTG  799
CACGTG..TC | GA..CACGTG  447
ACACGTG.A | T.CACGTGT  174
A.CACGTG.T | A.CACGTG.T  189
ACACGTGT | ACACGTGT  123
G.CACGTG.C | G.CACGTG.C  482
CACGTGTC | GACACGTG  300
G.CACGTGC | GCACGTG.C  523
CACGTG..AA | TT..CACGTG  214
ACACGTGG | CCACGTGT  381
CACGTGG.A | T.CCACGTG  420
AACACGTG | CACGTGTT  286
ACCACGTG | CACGTGGT  503
CACGTG.C.C | G.G.CACGTG  792
CACGTGT.A | T.ACACGTG  161
AA.CACGTG | CACGTG.TT  358
CACGTG.CA | TG.CACGTG  387
A.ACACGTG | CACGTGT.T  298
CACGTG.TC | GA.CACGTG  516
C.CACGTGG | CCACGTG.G  659
A.CACGTGC | GCACGTG.T  326
CACGTGT..C | G..ACACGTG  318
AA..CACGTG | CACGTG..TT  303
C..ACACGTG | CACGTGT..G  349
C.G.CACGTG | CACGTG.C.G  772
CACGTG..AC | GT..CACGTG  265
TCACGTGA | TCACGTGA  337
CCACGTG..A | T..CACGTGG  368
CACGTG.T.A | T.A.CACGTG  266
C.A.CACGTG | CACGTG.T.G  400
A..CACGTGC | GCACGTG..T  381
ACACGTG.C | G.CACGTGT  304
```

Pou2f1 least scoring 8-mers

```
CTTGCA.AA | TT.TGCAAG  62
ATCG.AT.CA | TG.AT.CGAT  11
ATA.TCGAA | TTCGA.TAT  5
ACA.T.TCAA | TTGA.A.TGT  51
T.AGAA.AAA | TTT.TTCT.A  121
GCA..CGTAA | TTACG..TGC  7
T.AA.GCGAA | TTCGC.TT.A  9
GG.CTAA.CA | TG.TTAG.CC  26
C.GGTATTA | TAATACC.G  11
GGGGTAT..A | T..ATACCCC  17
ACA.GCAAA | TTTGC.TGT  97
GTA.ATGAA | TTCAT.TAC  44
G.G.GTATAA | TTATAC.C.C  11
TATC.T.CAA | TTG.A.GATA  40
CA.A.AATAG | CTATT.T.TG  65
AGAG..AATA | TATT..CTCT  52
ATCA.AT.CA | TG.AT.TGAT  47
CCT.GATT.A | T.AATC.AGG  32
CCT.TAAT.C | G.ATTA.AGG  35
CTCTT..ATA | TAT..AAGAG  50
AA.GAG.ATA | TAT.CTC.TT  36
G.T.ATATAC | GTATAT.A.C  9
T.CCA.CTTA | TAAG.TGG.A  39
C.AGTGA.TA | TA.TCACT.G  33
C.TT.AGTAA | TTACT.AA.G  42
GC.A.TTTAA | TTAAA.T.GC  31
GATTAC.C.A | T.G.GTAATC  11
CTA.TT.GGC | GCC.AA.TAG  19
GC.TCT.TTA | TAA.AGA.GC  54
A.GGTA.TTT | AAA.TACC.T  44
CTAC.TA.GC | GC.TA.GTAG  18
AA.A.GCCTA | TAGGC.T.TT  29
ATAT..GGAG | CTCC..ATAT  22
CAT.AGA.AA | TT.TCT.ATG  56
C.GAATATC | GATATTC.G  17
A.GAAT.ACT | AGT.ATTC.T  35
C..ACATTAG | CTAATGT..G  26
GA.TT.CTTA | TAAG.AA.TC  53
GAAT.TTAC | GTAA.ATTC  29
AA..CCTTTA | TAAAGG..TT  67
TCGT.ATCA | TGAT.ACGA  10
AGACA.TA.A | T.TA.TGTCT  51
AATA..ACGG | CCGT..TATT  9
AATGTG.A.A | T.T.CACATT  69
TGCTT.T.AA | TT.A.AAGCA  95
C.CATTTGC | GCAAATG.G  114
ACCT.C.AAA | TTT.G.AGGT  48
AA.C.GGATT | AATCC.G.TT  44
A..AGTATGT | ACATACT..T  24
AT.TCATGC | GCATGA.AT  27
```

Pou2f1 top scoring 8-mers

```
A.A.TAATTA | TAATTA.T.T  53
AATTAATT | AATTAATT  35
ATTA.CATA | TATG.TAAT  151
A..ATAATTA | TAATTAT..T  52
A.TTAATTA | TAATTAA.T  67
G.TAATTA.C | G.TAATTA.C  26
AAATTAAT | ATTAATTT  56
A.TAATTAA | TTAATTA.T  50
TA.TAATTA | TAATTA.TA  25
AT.ATAAT.A | T.ATTAT.AT  55
ATTAATTA | TAATTAAT  48
ATAATGAG | CTCATTAT  41
A.TAATTAG | CTAATTA.T  43
CATAATTA | TAATTATG  47
A..TAATTAG | CTAATTA..T  49
T..TAATTAA | TTAATTA..A  62
ATTAATT..G | C..AATTAAT  35
A.CTAATTA | TAATTAG.T  31
CCTCATTA | TAATGAGG  55
AT..TAATTA | TAATTA..AT  70
TTAATTAA | TTAATTAA  34
ATAATTAG | CTAATTAT  35
TATGCAAA | TTTGCATA  312
AA..TAATTA | TAATTA..TT  60
G.ATAATTA | TAATTAT.C  49
A.TAATTA.G | C.TAATTA.T  38
TATGCATA | TATGCATA  65
ATGCAAAT | ATTTGCAT  404
A..TTAATTA | TAATTAA..T  65
A..CTCATTA | TAATGAG..T  52
C.TAATTAA | TTAATTA.G  51
A..TAATTAT | ATAATTA..T  40
ATT.GCATA | TATGC.AAT  265
ATA.TAAT.A | T.ATTA.TAT  41
CTAATTAA | TTAATTAG  60
AATTAAT.A | T.ATTAATT  67
A.ATTAAT.A | T.ATTAAT.T  49
ATTATGCA | TGCATAAT  106
CTAATTA.C | G.TAATTAG  44
AT.TGCATA | TATGCA.AT  298
AG.TAATTA | TAATTA.CT  59
```

C..TAATTAA | TTAATTA..G 46
CTAATTAG | CTAATTAG 20
GCATAAT.A | T.ATTATGC 67
A.CT.ATTAT | ATAAT.AG.T 32
T.TATGCA.A | T.TGCATA.A 102
AATAATTA | TAATTATT 49
AT.TAATTA | TAATTA.AT 42
G.TAATTAA | TTAATTA.C 56
GTTAATTA | TAATTAAC 44

Pou2f2 least scoring 8-mers

AT.ATCC.TG | CA.GGAT.AT 150
AACG.T.ATT | AAT.A.CGTT 69
AGT.TG.AAA | TTT.CA.ACT 569
A.GCG.TTTA | TAAA.CGC.T 60
ATTG..TATC | GATA..CAAT 87
AATTTCCC | GGGAAATT 441
ACATTT..TA | TA..AAATGT 405
ATATAA.C.C | G.G.TTATAT 130
ATT.CAC.AA | TT.GTG.AAT 242
AATTGC.AA | TT.GCAATT 239
A.TC.ATATA | TATAT.GA.T 89
CATGC..AAC | GTT..GCATG 209
A.TTGC.TGG | CCA.GCAA.T 340
T.ACTAGA.A | T.TCTAGT.A 215
ATTT.G.ATG | CAT.C.AAAT 242
ATA.TC.CGA | TCG.GA.TAT 39
A.T.GCGACA | TGTCGC.A.T 58
CGTTA..ATA | TAT..TAACG 23
G..CTCGTTA | TAACGAG..C 57
CG.AGATTA | TAATCT.CG 43
GATT.GT.AA | TT.AC.AATC 173
GGCA.G.ATA | TAT.C.TGCC 134
T..CTACATA | TATGTAG..A 159
A.TCCTA.TT | AA.TAGGA.T 234
CAG.ATA.AA | TT.TAT.CTG 278
ATA.A.TCAC | GTGA.T.TAT 183
AATACGGA | TCCGTATT 67
T.TATCC.GA | TC.GGATA.A 133
TT.G.GATAA | TTATC.C.AA 165
ATCATC.TA | TA.GATGAT 96
T.GAAAATA | TATTTTC.A 467
TATT.AC.GA | TC.GT.AATA 172
AACG.AT.TA | TA.AT.CGTT 63
CCT.ATA.TC | GA.TAT.AGG 124
AGTA..ATAT | ATAT..TACT 160
GG.CTG.TAA | TTA.CAG.CC 187
G.TATA.TAC | GTA.TATA.C 54
GATAAA.GA | TC.TTTATC 220
A.TACCC.TC | GA.GGGTA.T 122
G.TA.TTGGA | TCCAA.TA.C 184
AAATTCTG | CAGAATTT 488
A.AATCGG.T | A.CCGATT.T 53
CAC.TAAG.A | T.CTTA.GTG 190
AA.TAG.ACG | CGT.CTA.TT 59
ATA..CGGTA | TACCG..TAT 31
GG.CGCATA | TATGCG.CC 46
CCGC.A.TAA | TTA.T.GCGG 47
G.GG.AATCA | TGATT.CC.C 251
TG.ATC.TAA | TTA.GAT.CA 145
G.CTGTTTA | TAAACAG.C 236

Pou2f2 top scoring 8-mers

ATGCAAAT | ATTTGCAT 2448
A..ATAATTA | TAATTAT..T 271
A.TTAATTA | TAATTAA.T 250
AG.TAATTA | TAATTA.CT 234
G.ATAATTA | TAATTAT.C 192
G.TAATTA.C | G.TAATTA.C 77
A..TAATTAG | CTAATTA..T 207
ATA.T.ATTA | TAAT.A.TAT 246
ATTTAAAT | ATTTAAAT 323
GTTAATTA | TAATTAAC 191
ATAATTAG | CTAATTAT 209
AATTAATT | AATTAATT 179
ATTA.CATA | TATG.TAAT 266
A.TAATTA.G | C.TAATTA.T 182
CTAATTAA | TTAATTAG 244
ATAATTAA | TTAATTAT 300
A..CTCATTA | TAATGAG..T 259
CCTCATTA | TAATGAGG 245
A.A.TAATTA | TAATTA.T.T 279
AT..TAATTA | TAATTA..AT 277
G.TAATTAA | TTAATTA.C 227
ATTTG.ATA | TAT.CAAAT 1389
A.G.TAATTA | TAATTA.C.T 209
A.TAATTAA | TTAATTA.T 282
C.C.TAATTA | TAATTA.G.G 133
C..TTAATTA | TAATTAA..G 198
C.TAATTAA | TTAATTA.G 196
GCTAATTA | TAATTAGC 174
CA..TAATTA | TAATTA..TG 227
ATT.GCATA | TATGC.AAT 1231
AT.TGCATA | TATGCA.AT 1348
ATAATTA..A | T..TAATTAT 325
T.CTAATTA | TAATTAG.A 250
GC.TAATTA | TAATTA.GC 146
C.TAATTAG | CTAATTA.G 150
C..CTAATTA | TAATTAG..G 148
ATAATTA.C | G.TAATTAT 190
A.TTAAT.AG | CT.ATTAA.T 262
TATGCAAA | TTTGCATA 1856
A.TAATTAG | CTAATTA.T 171
CATAATTA | TAATTATG 244
CTAATTA..C | G..TAATTAG 150
A.CTAATTA | TAATTAG.T 180
A..TAATTAA | TTAATTA..T 306
T..TAATTAA | TTAATTA..A 285
T..TTAATTA | TAATTAA..A 318
CTAATTAG | CTAATTAG 79
CTAATTA.C | G.TAATTAG 160
ATTT.CATA | TATG.AAAT 1552
ATTAATTA | TAATTAAT 225

Sox12 least scoring 8-mers

CAG.TCA.AC | GT.TGA.CTG 66
A.TCAAGT.C | G.ACTTGA.T 63
ATTAA.T.GC | GC.A.TTAAT 38
A.TAAAT.TT | AA.ATTTA.T 99
ATG.ATAG.G | C.CTAT.CAT 33
T.AAAAG.CA | TG.CTTTT.A 186
CGAATAT.A | T.ATATTCG 3
A.AG.TATCG | CGATA.CT.T 5
AAA..AACTC | GAGTT..TTT 115
CTACA..ATA | TAT..TGTAG 35
ATCAAT..TC | GA..ATTGAT 34
AAGT.AC.CA | TG.GT.ACTT 59
T.GCTTATA | TATAAGC.A 16
CAAATTTA | TAAATTTG 64
ATGTTATG | CATAACAT 22
T..TATTCTA | TAGAATA..A 68
ACAA.CTAA | TTAG.TTGT 44
ACAT.T.TAA | TTA.A.ATGT 93
AATAAG.TA | TA.CTTATT 29
T.CACA.TGA | TCA.TGTG.A 102
ATA.AC.TAG | CTA.GT.TAT 21
AAA..GTACA | TGTAC..TTT 69
A.ATTCGC.T | A.GCGAAT.T 6
C.AACTACA | TGTAGTT.G 33
AAC.C.TTAG | CTAA.G.GTT 36
GTTT.TACA | TGTA.AAAC 65
TAAGCA..AA | TT..TGCTTA 74
TACTCG..AA | TT..CGAGTA 10
A.AC.AAATT | AATTT.GT.T 54
TT.AGGAAA | TTTCCT.AA 147
AAAAAC..AT | AT..GTTTTT 80
A.ACGAA.AA | TT.TTCGT.T 20
A..AATCGAG | CTCGATT..T 9
ACATAA.GC | GC.TTATGT 31
T.AC.AACGA | TCGTT.GT.A 1
T.CAAAG.CA | TG.CTTTG.A 173

A.GATA.TAC | GTA.TATC.T 17
CA.TT.AAAC | GTTT.AA.TG 80
ATCA.AT.AA | TT.AT.TGAT 54
C.TACTT.AC | GT.AAGTA.G 33
GTT.TAAG.A | T.CTTA.AAC 60
ACTT.T.CAA | TTG.A.AAGT 55
AGTACAA.A | T.TTGTACT 56
CG.AAC.AAA | TTT.GTT.CG 21
C.ACAT.TTG | CAA.ATGT.G 37
ACAA.ATCC | GGAT.TTGT 44
AAAC.GATA | TATC.GTTT 45
CA.C.TTGTC | GACAA.G.TG 111
A.GAG.TAAC | GTTA.CTC.T 28
TA.AGTT.CA | TG.AACT.TA 62

Sox12 top scoring 8-mers

GAACAATA | TATTGTTC 160
ATTGTT.T.A | T.A.AACAAT 160
AACAATAG | CTATTGTT 240
T.AACAATA | TATTGTT.A 87
T..AACAATA | TATTGTT..A 98
A.AACAATT | AATTGTT.T 127
AT..AACAAT | ATTGTT..AT 76
AATTGTTC | GAACAATT 75
ATTGTT.TA | TA.AACAAT 114
ATTGTTC.A | T.GAACAAT 95
A..AATTGTT | AACAATT..T 72
A.AACAAT.A | T.ATTGTT.T 225
AACAATTA | TAATTGTT 83
A.AACAATG | CATTGTT.T 391
C.ATTGTT.A | T.AACAAT.G 110
C.ATTGTTC | GAACAAT.G 278
AGAACAAT | ATTGTTCT 310
GAACAAT.A | T.ATTGTTC 145
A.T.AACAAT | ATTGTT.A.T 91
T.AACAAT.A | T.ATTGTT.A 107
ATTGTT.TC | GA.AACAAT 215
AACAATA.A | T.TATTGTT 174
AACAAT..TA | TA..ATTGTT 69
G.AACAATA | TATTGTT.C 127
A..TAACAAT | ATTGTTA..T 97
ATTGTTC..A | T..GAACAAT 141
ATTGTTC..G | C..GAACAAT 128
A.G.AACAAT | ATTGTT.C.T 125
TA.ACAATA | TATTGT.TA 86
C..AACAATA | TATTGTT..G 104
AATTGTT..A | T..AACAATT 68
A..GAACAAT | ATTGTTC..T 187
ATTGTTCC | GGAACAAT 181
A..AACAATA | TATTGTT..T 166
GAACAAT..A | T..ATTGTTC 195
AATTGTT..G | C..AACAATT 71
AACAATT.A | T.AATTGTT 97
AACAATAA | TTATTGTT 104
TAACAATA | TATTGTTA 103
A.AACAAT.G | C.ATTGTT.T 308
ATAACAAT | ATTGTTAT 197
AACAAT.AA | TT.ATTGTT 186
A..AAACAAT | ATTGTTT..T 189
ATTGTT.TG | CA.AACAAT 257
A.AATTGTT | AACAATT.T 62
ATTGTTA.C | G.TAACAAT 68
AATTGT..TG | CA..ACAATT 105
A.AACAATA | TATTGTT.T 203
A.A.AACAAT | ATTGTT.T.T 270
CTATTGT..A | T..ACAATAG 116

Sp4 least scoring 8-mers

CA.TA.CATA | TATG.TA.TG 12
CT.CTTC.CC | GG.GAAG.AG 329
AAGTA.ATG | CAT.TACTT 28
GTT..TTGTA | TACAA..AAC 22
G.GTTCAAA | TTTGAAC.C 50
T.TCTCAA.A | T.TTGAGA.A 48
AAT.TC.TAA | TTA.GA.ATT 23
TTAT..AAAA | TTTT..ATAA 50
CTTA.ATCC | GGAT.TAAG 40
A.CTCC.TCA | TGA.GGAG.T 90
CTCTTA..CC | GG..TAAGAG 76
AA.GTGGGC | GCCCAC.TT 128
GAT.GT.AAC | GTT.AC.ATC 21
AC.TAG.ACA | TGT.CTA.GT 25
GCA.AGTAA | TTACT.TGC 32
CAC.CA.TTA | TAA.TG.GTG 37
AC.CCC.TAC | GTA.GGG.GT 52
GAC..CCTCC | GGAGG..GTC 179
ATGT.C.TAA | TTA.G.ACAT 17
AC..CCAACC | GGTTGG..GT 91
A.TA.GGGGG | CCCCC.TA.T 40
CGCC.ATT.C | G.AAT.GGCG 79
ACT.TT.CAT | ATG.AA.AGT 34
C.C.CCTAGA | TCTAGG.G.G 98
GAGGAGAA | TTCTCCTC 251
AAA.GG.GCA | TGC.CC.TTT 82
AATGA.TTG | CAA.TCATT 27
ACGG.GGA.G | C.TCC.CCGT 159
TATAT.T.CA | TG.A.ATATA 33
ATGAACG..A | T..CGTTCAT 22
C..CTATTAG | CTAATAG..G 7
GTTG.CCA.A | T.TGG.CAAC 38
CAATTAAA | TTTAATTG 49
CTTTG.TCA | TGA.CAAAG 76
ACAAATA.C | G.TATTTGT 30
ATAAATG.A | T.CATTTAT 45
C.TA.AGGCG | CGCCT.TA.G 71
G.GGGA.GAC | GTC.TCCC.C 172
CCAGA.ACA | TGT.TCTGG 124
ATAC.GAAA | TTTC.GTAT 26
C.TTT.TAGA | TCTA.AAA.G 52
ACAG.AT.AA | TT.AT.CTGT 36
CT.TCT.ACA | TGT.AGA.AG 42
CACA.AGAA | TTCT.TGTG 108
T..TGTTTTA | TAAAACA..A 49
TACGC.C.CA | TG.G.GCGTA 34
AACAATAT | ATATTGTT 24
AAT.A.AAAA | TTTT.T.ATT 108
GGGGG.GCA | TGC.CCCCC 205
A.A.ATATCA | TGATAT.T.T 23

Sp4 top scoring 8-mers

CACGCCC.C | G.GGGCGTG 516
CCGCCCCC | GGGGGCGG 1466
ACGCCCCC | GGGGGCGT 377
AGGGGGCG | CGCCCCCT 1037
G.CACGCCC | GGGCGTG.C 578
G..CCGCCCC | GGGGCGG..C 1693
CCCGCCCC | GGGGCGGG 2024
ACGCCC.CT | AG.GGGCGT 292
A.GGGGCGG | CCGCCCC.T 1040
CCCGCCC.C | G.GGGCGGG 1748
C.CGCCCCC | GGGGGCG.G 1384
CCCCGCCC | GGGCGGGG 2092
CCGCCCC.C | G.GGGGCGG 1284
GGGGGCG..A | T..CGCCCCC 474
CCACGCCC | GGGCGTGG 657
G..ACGCCCC | GGGGCGT..C 524
GGGGGCG.A | T.CGCCCCC 412
C.CC.CCCCC | GGGGG.GG.G 1625
A..CGCCCCC | GGGGGCG..T 265
G.CCC.CCCC | GGGG.GGG.C 1617
A.GGGGGCG | CGCCCCC.T 493
CCGCCCC..C | G..GGGGCGG 1524
AGGGGG.GG | CC.CCCCCT 938
CCC.CCCCC | GGGGG.GGG 1607
C.CCGCCCC | GGGGCGG.G 1998
CCGCCCAC | GTGGGCGG 720
CACGCCCA | TGGGCGTG 248
G.C.CGCCCC | GGGGCG.G.C 1615
CGCCCCC..C | G..GGGGGCG 1114
ACGCCC.C.C | G.G.GGGCGT 311

ACGCCCAC | GTGGGCGT 179
CCCC.CCCC | GGGG.GGGG 2098
A.GGGGCGT | ACGCCCC.T 300
CCAC.CCCC | GGGG.GTGG 681
CCGCCC.CC | GG.GGGCGG 1065
ACGCCC.CC | GG.GGGCGT 285
CACGCCCC | GGGGCGTG 613
CCCC.CCC.C | G.GGG.GGGG 1852
CAC.CCCCC | GGGGG.GTG 428
A..GGGCGGG | CCCGCCC..T 1123
CGCCCCC..A | T..GGGGGCG 442
CC.CCCCCC | GGGGGG.GG 543
GCC.C.CCCC | GGGG.G.GGC 1634
CG..ACGCCC | GGGCGT..CG 189
CC.CGCCC.C | G.GGGCG.GG 1644
A.GGGGG.GG | CC.CCCCC.T 526
ACGCCCC..G | C..GGGGCGT 286
A.CGCCCCC | GGGGGCG.T 410
CGCCCCC.C | G.GGGGGCG 918
CGCCCCCC | GGGGGGCG 429
AGTAC.AAG | CTT.GTACT 91

Tcf3 least scoring 8-mers

AC.TC.TATA | TATA.GA.GT 45
AT..GAATAT | ATATTC..AT 101
A.CTTAT.TA | TA.ATAAG.T 87
AA.CAAAA.G | C.TTTTG.TT 332
TAAAC.AGA | TCT.GTTTA 118
ATA..ACTAA | TTAGT..TAT 98
G.AAT.GATC | GATC.ATT.C 60
GCT.T.ATAA | TTAT.A.AGC 106
ATT.A.TATG | CATA.T.AAT 84
CGCAT.AAA | TTT.ATGCG 41
C..TAGTTTA | TAAACTA..G 56
GTTGTTAA | TTAACAAC 95
A.GCTTGAA | TTCAAGC.T 157
AGAAGA.AT | AT.TCTTCT 265
GCC.TGAA.A | T.TTCA.GGC 212
A..AATGAAC | GTTCATT..T 193
ATGGATG..A | T..CATCCAT 121
G.GTAAGAA | TTCTTAC.C 158
GT.T.CTAAA | TTTAG.A.AC 130
AAACGT.TG | CA.ACGTTT 81
GAAGATC.A | T.GATCTTC 92
ATAAG.TTG | CAA.CTTAT 65
GTC.ACGA.A | T.TCGT.GAC 40
GATT..TTTA | TAAA..AATC 155
CTTA.CAAA | TTTG.TAAG 156
T.GA.TAGAA | TTCTA.TC.A 120
TGCAT.T.AA | TT.A.ATGCA 230
TAAC..CATA | TATG..GTTA 62
ACAAAC.A.A | T.T.GTTTGT 280
AGGACA..AA | TT..TGTCCT 277
T.TGTT.GTA | TAC.AACA.A 102
ATTAA.TT.C | G.AA.TTAAT 110
T.ATGCAA.A | T.TTGCAT.A 173
G.AT.ACAAC | GTTGT.AT.C 91
A.TA.AAGAC | GTCTT.TA.T 102
ATTCCA..AC | GT..TGGAAT 137
AGT.GATAA | TTATC.ACT 61
AA.ATT.ATC | GAT.AAT.TT 94
A.AGAC.GAA | TTC.GTCT.T 247
A..TTGTGTA | TACACAA..T 102
AC.CTA.AAG | CTT.TAG.GT 84
AAAA.GAT.A | T.ATC.TTTT 247
A.CC.ATAAA | TTTAT.GG.T 105
ACC.TT.ATA | TAT.AA.GGT 57
AGTA.T.GAT | ATC.A.TACT 46
T.TACT.TCA | TGA.AGTA.A 169
G.TAAGAG.A | T.CTCTTA.C 152
AG..CAAACA | TGTTTG..CT 240
CA.AACTAA | TTAGTT.TG 150

A.ATCAAAG | CTTTGAT.T 200

Tcf3 top scoring 8-mers

CTTTGATC | GATCAAAG 137
CATCAAAG | CTTTGATG 176
AGATCAAA | TTTGATCT 143
ATCAAAG..A | T..CTTTGAT 175
ACATCAAA | TTTGATGT 155
ACAT.AAAG | CTTT.ATGT 199
ATCAAAG.A | T.CTTTGAT 173
ATCAAAG.G | C.CTTTGAT 144
A..ATCAAAG | CTTTGAT..T 152
CTTTGAT..A | T..ATCAAAG 145
CTTTGAT.A | T.ATCAAAG 144
ATCAAAGG | CCTTTGAT 167
AC.TCAAAG | CTTTGA.GT 190
AGAT.AAAG | CTTT.ATCT 235
A.ATCAAA.A | T.TTTGAT.T 189
ATCAAAG..C | G..CTTTGAT 153
A.ATCAAA.G | C.TTTGAT.T 165
ATCAAAGA | TCTTTGAT 146
AA.ATCAAA | TTTGAT.TT 229
CTTTGAT..C | G..ATCAAAG 169
CTTCAAAG | CTTTGAAG 264
ACTTTGAT | ATCAAAGT 122
A..GATCAAA | TTTGATC..T 100
ACTTCAAA | TTTGAAGT 232
GATCAAA.A | T.TTTGATC 107
A.CTTT.ATC | GAT.AAAG.T 116
C.TTTGATC | GATCAAA.G 88
C.TTTGATG | CATCAAA.G 141
AG.TCAAAG | CTTTGA.CT 230
ATCAAAGC | GCTTTGAT 164
ACTT.AAAG | CTTT.AAGT 235
TTT.AT.AAA | TTT.AT.AAA 166
C..GATCAAA | TTTGATC..G 108
A.TTCAAAG | CTTTGAA.T 289
CTTT.ATC.C | G.GAT.AAAG 170
C.A.ATCAAA | TTTGAT.T.G 150
A.A.ATCAAA | TTTGAT.T.T 204
CTTTGATA | TATCAAAG 96
TA.ATCAAA | TTTGAT.TA 99
A.GATCAAA | TTTGATC.T 120
ATCAAA.GA | TC.TTTGAT 146
CTTT.AT.TC | GA.AT.AAAG 305
CTTT.AT.TA | TA.AT.AAAG 177
A..CTTTGAT | ATCAAAG..T 119
ATCAAA.A.A | T.T.TTTGAT 174
AA..TCAAAG | CTTTGA..TT 263
A.CTTTGAT | ATCAAAG.T 118
ATCAAAG..G | C..CTTTGAT 146
GAT.AAAG.A | T.CTTT.ATC 186