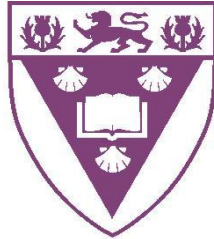


A CENTRAL ENRICHMENT-BASED COMPARISON OF TWO ALTERNATIVE
METHODS OF GENERATING TRANSCRIPTION FACTOR BINDING MOTIFS
FROM PROTEIN BINDING MICROARRAY DATA



RHODES UNIVERSITY
Grahamstown • 6140 • South Africa

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

Master of Science

in

Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

By

Ntombikayise Mahaye

Department of Biochemistry, Microbiology and Biotechnology

December 2012

Supervisor: Prof Philip Machanick

Department of Computer Science, Rhodes University



Abstract

Characterising transcription factor binding sites (TFBS) is an important problem in bioinformatics, since predicting binding sites has many applications such as predicting gene regulation. ChIP-seq is a powerful in vivo method for generating genome-wide putative binding regions for transcription factors (TFs). CentriMo is an algorithm that measures central enrichment of a motif and has previously been used as motif enrichment analysis (MEA) tool. CentriMo uses the fact that ChIP-seq peak calling methods are likely to be biased towards the centre of the putative binding region, at least in cases where there is direct binding. CentriMo calculates a binomial p-value representing central enrichment, based on the central bias of the binding site with the highest likelihood ratio. In cases where binding is indirect or involves cofactors, a more complex distribution of preferred binding sites may occur but, in many cases, a low CentriMo p-value and low width of maximum enrichment (about 100bp) are strong evidence that the motif in question is the true binding motif. Several other MEA tools have been developed, but they do not consider motif central enrichment.

The study investigates the claim made by Zhao and Stormo (2011) that they have identified a simpler method than that used to derive the UniPROBE motif database for creating motifs from protein binding microarray (PBM) data, which they call BEEML-PBM (Binding Energy Estimation by Maximum Likelihood-PBM). To accomplish this, CentriMo is employed on 13 motifs from both motif databases. The results indicate that there is no conclusive difference in the quality of motifs from the original PBM and BEEML-PBM approaches. CentriMo provides an understanding of the mechanisms by which TFs bind to DNA. Out of 13 TFs for which ChIP-seq data is used, BEEML-PBM reports five better motifs and twice it has not had any central enrichment when the best PBM motif does. PBM approach finds seven motifs with better central enrichment. On the other hand, across all variations, the number of examples where PBM is better is not high enough to conclude that it is overall the better approach. Some TFs bind directly to DNA, some indirect or in combination with other TFs. Some of the predicted mechanisms are supported by literature evidence. This study further revealed that the binding specificity of a TF is different in different cell types and development stages. A TF is up-regulated in a cell line where it performs its biological function. The discovery of cell line differences, which has not been done before in any CentriMo study, is interesting and provides reasons to study this further.

Declaration

I, Ntombikayise Mahaye declare that this thesis is my own work except AME and CentriMo scripts provided by Prof Philip Machanick. It is submitted for the degree of Master of Science in Bioinformatics and Computational Molecular Biology in the Faculty of Science at Rhodes University. It has not been submitted before for any degree or examination. I also state that all the sources that I have used have been acknowledged.

Signature:



Date: 12 December 2012

I confirm that this thesis has been submitted with my approval as the university supervisor.

Supervisor: Prof Philip Machanick

Date: 12 December 2012

Signature:



Dedication

This thesis is dedicated to:

My late father, Mr N.J Mahaye,

my mother, Mrs S.J Mahaye

and

my late brother Mr S.G.C Mahaye.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof Philip Machanick. All the work in this thesis is the result of our discussions and the accomplishment of this work would not have been successful without his guidance and support. It is my great pleasure to work with him.

I also thank my course co-ordinator, Dr Özlem Taştan Bishop, for accepting me in the programme and her hard work in getting the Rhodes University Sandisa Imbewu scholarship for me. I also thank all the MSc in Bioinformatics and Computational Molecular Biology lecturers for sharing their information with us.

Special thanks also goes to all Rhodes University Bioinformatics (RUBi) and Department of Computer Science postgraduate students for being my brothers and sisters, making Rhodes University a home away from home and providing me with a good atmosphere for doing my research.

I would like to thank the funding bodies: Rhodes University Sandisa Imbewu and National Research Foundation (NRF) of South Africa for their financial assistance. Without their support, I would not have been able to do this degree.

I would like to thank my family especially my mother, Mrs S.J Mahaye, for her endless love, prayers, encouragement and spiritual support throughout my life. My deepest gratitude also goes to my late brother, Mr Sanele G.C Mahaye, who until his last day had a kind concern regarding my academic requirements. Without his support I would not have gone that far to the MSc level. He was always encouraging me with his best wishes and he never got tired of encouraging me to study hard.

Lastly, I thank Almighty God for blessing me with the opportunity to do this MSc degree.

Contents

Abstract.....	i
Declaration.....	II
Dedication.....	III
Acknowledgements.....	IV
List of figures.....	VII
List of Tables	VIII
List of Abbreviations	IX
CHAPTER 1: INTRODUCTION.....	1
1.1 LITERATURE REVIEW	4
1.1.1 TRANSCRIPTION	4
1.1.1.1 TRANSCRIPTIONAL CONTROL.....	5
1.1.1.2 POST-TRANSCRIPTIONAL CONTROL.....	5
1.1.2 MOTIF DISCOVERY TOOLS.....	8
1.1.3 MOTIF ENRICHMENT ANALYSIS (MEA) TOOLS.....	12
1.1.4 CLASSIFICATION OF TRANSCRIPTION FACTORS.....	16
1.1.5 OVERVIEW OF TFs AND THEIR BINDING PARTNERS	18
1.2 PROBLEM STATEMENT.....	21
1.3 AIMS AND OBJECTIVES.....	23
CHAPTER 2: METHODS.....	24
2.1 CENTRIMO.....	24
2.2 SPAMO	25
2.3 AME.....	26
2.4 DATA	26
CHAPTER 3: RESULTS.....	30
3.1 Comparing the BEEML-PBM motifs with the PBM motifs using CentriMo.....	30
3.1.1 Motifs with clear central enrichment from both motif databases	30
3.1.2 Central enrichment with less sharply defined peaks.....	32

3.1.3	Central enrichment from only one database.....	34
3.1.4	No motif from either database with central enrichment	35
3.2	CentriMo results visualization using logos.....	37
3.3	Independent measure: AME	39
3.4	Comparing central motif enrichment from different cell lines.....	43
3.5	CentriMo run on combined JASPAR/UniPROBE database.....	49
3.6	Further investigation of CentriMo results using SpaMo.....	56
CHAPTER 4: DISCUSSION.....		64
CHAPTER 5: CONCLUSION.....		68
	FUTURE PROSPECTS.....	69
	BIBLIOPGRAPHY	70
	WEBSITE REFERENCES.....	80
	APPENDIX.....	81

List of figures

Figure 3-1: AME analysis of Egr1 motifs.....	39
Figure 3-2: AME analysis of Irf4 motifs.	40
Figure 3-3: AME analysis of Pou2f2 motifs.....	40
Figure 3-4: AME analysis of Tcf3 motifs.....	41
Figure 3-5: AME analysis of Ets1 motifs.	42
Figure 3-6: Evidence of Irf4 binding partners during DNA binding.	57
Figure 3-7: Searching for evidence of Sp4 complex formation during DNA binding.	58
Figure 3-8: Searching for evidence of Foxa2 complex formation during DNA binding.....	59
Figure 3-9: Searching for evidence of Srf complex formation during DNA binding.....	60
Figure 3-10: Searching for evidence of Egr1 complex formation during DNA binding.....	61
Figure 3-11: Searching for evidence of Ets1 complex formation during DNA binding.	62
Figure 3-12: SpaMo analysis of Pou2f2 motifs.	63

List of Tables

Table 1-1: IUPAC DNA codes.	2
Table 2-1: Data used to run CentriMo.	27
Table 2-2: Data used to investigate variations across cell lines.	28
Table 2-3: ChIP-seq data from ENCODE/Stanford/Yale/USC/Harvard Lab.	28
Table 2-4: Explanation of cell lines.	29
Table 3-1: BEEML-PBM and UniPROBE successes.	31
Table 3-2: BEEML-PBM and UniPROBE CentriMo.	33
Table 3-3: BEEML-PBM fails.	34
Table 3-4: CentriMo fail.	36
Table 3-5: TFs variability in different cell lines.	44
Table 3-6: Foxa2 CentriMo distribution.	45
Table 3-7: Further investigation of the BEEML-PBM failure.	46
Table 3-8: Further investigation of CentriMo failure.	47
Table 3-9: ChIP-seq data from the ENCODE/Stanford/Yale/USC/Harvard Lab.	48
Table 3-10: CentriMo run on combined JASPAR/UniPROBE database.	50
Table 3-11: Investigating the mechanism of binding.	52
Table 3-12: Further investigation of BEEML-PBM approach failure.	53
Table 3-13: Investigation on enriched motifs with off-centred peaks.	55
Table 4-1: Summary of each TF.	67

List of Abbreviations

A	Adenine
AME	Analysis of Motif Enrichment
AREs	AU-Rich Elements
B-ZIP	Basic Region Leucine Zipper
BEEML	Binding Energy Estimation by Maximum Likelihood
C	Cytosine
CentriMo	Centrality of Motifs
CLOVER	Cis-eLement OVERrepresentation
CMEA	Central Motif Enrichment Analysis
ChIP	Chromatin ImmunoPrecipitation
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing
DBD	DNA Binding Domain
DNA	Deoxyribonucleic Acid
Egr	Early Growth Response
EM	Expectation Maximization
ENCODE	Encyclopedia of DNA Elements
EST	Expressed Sequence Tags
G	Guanine
hESC	Human Embryonic Stem Cell
HLH	Helix-loop-helix
HRE	Human Response Element
Irf4	Interferon regulatory factor 4
MEME	Multiple EM for Motif Elicitation

mESC	Mouse Embryonic Stem Cell	
NR2F	Nuclear receptor subfamily 2	
PASTAA	Predicting ASsociated Transcription factors from Affinities	Annotated
PBM	Protein Binding Microarray	
PFMs	Position Specific Frequency Matrix	
PWM	Positional Weight Matrix	
QuEST	Quantitative Enrichment of Short Tags	
RBPs	RNA-binding proteins	
RAR	Retinoic acid receptor	
RNA	Ribonucleic Acid	
RXR	Retinoid X receptor	
tRNA	Transfer RNA	
mRNA	Messenger RNA	
SpaMo	Spaced Motif Analysis	
TF	Transcription factor	
TFBS	Transcription Factor Binding Sites	
T	Thymine	
TSS	Transcription Start Site	
Tcf	T-cell factor	
UCSC	University of California, Santa Cruz	
UTRs	Untranslated Regions	

CHAPTER 1: INTRODUCTION

Transcription factors (TFs) are proteins that bind to specific sequences in DNA and activate or repress the expression of their target genes. One of the greatest challenges facing bioinformatics is the understanding of the complex mechanisms regulating gene expression. Predicting binding sites has many applications such as predicting gene regulation and also predicting how genetic variation alters the normal expression of the gene. This study involves the use of ChIP-seq data to characterise transcription factor binding sites (TFBS). ChIP-seq has proven to be a powerful method for mapping *in vivo* locations of TFs (Hu et al., 2010). Knowledge of how DNA-protein interactions regulates gene expression is crucial for understanding biological processes (Valouev et al., 2008). Chromatin immunoprecipitation is used to study the interactions between DNA and proteins.

In the ChIP-seq process, DNA is cross-linked to a specific protein resulting in DNA-protein complexes. The cross-linked DNA is broken into short pieces (usually 0.2 to 1kb) suitable for sequencing (Jothi et al., 2008). An antibody against the protein of interest is added to the complex to isolate DNA that has been bound to the protein. This is followed by high throughput DNA sequencing. Sequencing output is mapped against a reference genome to infer TF binding location. A peak calling algorithm is then used to identify centered ChIP-seq regions. Computational analysis is applied to get biological information from a TF's ChIP-seq data. When binding sites have been validated the next step is to find how TFs bind at genomic regions (Whittington et al., 2011). Transcription factors may bind directly to the corresponding DNA sequences making it possible to declare a sharp peak. There are cases where TFs may show indirect binding (TFs may bind to their target DNA, then other TFs bind to these) or cooperative binding (when TFs bind in combination with cofactors).

A TFBS is generally represented using a motif. Motif are patterns that appear more often in a DNA sequence and they have a biological function (Patrik, 2006). Motifs are represented using a position weight matrix (PWM) (Wasserman & Sandelin, 2004). In a PWM, the rows have a numeric score for each of the four DNA bases at each position (columns) (Nishida et al., 2009). A motif is sometimes named as its consensus sequence with ambiguous positions with the appropriate code (see Table 1-1) for ambiguous positions (www.bioinformatics.org/sms2/iupac.html).

Nucleotide Code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

Table 1-1: IUPAC DNA codes.

This study aims to investigate the mechanisms by which TFs bind to DNA and compare two methods (BEEML-PBM and PBM) for creating motifs from protein binding microarray (PBM) data. PBM is a technique used to measure in vitro TF-DNA (double stranded) binding affinities. The detailed information on TF binding specificity is obtained from the microarrays representing all possible 10bp long binding sites. A fluorophore-antibody complex specific to the epitope is used to tag the protein bound array. Successful TF-DNA interactions are those showing enrichment on the brightest probes on the array. TF binding specificities are represented using position weight matrices (PWMs); the assumption that all motifs exert independent binding effect is not true for all TFs. There are some cases where

nucleotide in one position depends on the nucleotide(s) on the other position (Berger & Bulyk, 2009). Different methods have been developed to estimate the quantitative specificity of TF-DNA binding. Some of these methods consider statistical analysis of TF's binding sites, while others are based on probabilistic model of binding. Probabilistic models have a disadvantage of not considering the non-linear association among binding energy and probability which may contain crucial functional sites (Zhao et al., 2009). BEEML-PBM algorithm is an extension of the BEEML (an approach that takes the TF's concentration into consideration when predicting the binding sites) algorithm for estimation of TFs model specificity. It is used to estimate the maximum likelihood of binding energies between protein-DNA associations based on the model of specificity (Zhao & Stormo, 2011).

The limitations that the currently used motif enrichment analysis (MEA) tools are facing are the selection of background sequences, poor ChIP-seq data and highly enriched co-factor binding sites. To overcome these limitations, this study uses CentriMo, an algorithm that measures central enrichment of a motif as a MEA tool (for more details on CentriMo see section 2.1). CentriMo is superior to other MEA tools because it does not require selection of background sequences. In CentriMo, the flanking regions serve as a control. In cases of poor ChIP-seq data quality CentriMo can point to a problem. As an example, if a TF is expected to have a clear binding specificity but CentriMo has a high p-value or other indications of less specific binding, the possibility that the ChIP-seq experiment has failed is considered. CentriMo is also used against a wider motif database to determine whether there may be cofactors, cooperative binding or indirect binding. This overcomes the problem with highly enriched co-factor binding sites (Bailey and Machanick, 2012).

1.1 LITERATURE REVIEW

1.1.1 TRANSCRIPTION

To understand the mechanism of binding between DNA and TFs it is important to have an idea of what DNA is and also how TFs are involved in gene regulation. DNA is a unit of heredity found in almost all living organisms (except RNA viruses). DNA is an essential molecule in life. The information about proteins is encoded in it. The process of protein synthesis requires DNA as a starting material. DNA replicates itself and is transcribed to mRNA. mRNA is then translated into a protein product (Crick, 1970). Structurally, DNA is made up of a deoxyribose sugar, a base and a phosphate. DNA is made up of four bases adenine (A), cytosine (C), guanine (G) and thymine (T). Adenine and guanine are long double ring purines and thymine and cytosine are short single ring pyrimidines. The DNA bases need to be complementary to each other so as to form a stable double helix structure. For base pairing a purine is paired with a pyrimidine and the opposite is true. The two polynucleotide bases in DNA coil around each other and run in opposite directions. The two DNA bases are joined together by hydrogen bonds and the pairing is specific. Adenine forms a pair with thymine and guanine base pairs with cytosine (Watson and Crick, 1953).

As mentioned earlier that the greatest problem is to understand the complex mechanisms regulating gene expression, so an overview of the TF's mechanism of action in transcription is important. Transcription refers to the synthesis of RNA from a DNA template. Transcription is divided into three main steps including initiation, mRNA chain elongation and chain termination. Initiation involves the correct binding of RNA polymerase to the core promoter of the transcription start site and control of transcription rate. During elongation, nucleotides are covalently added to the 3' end of the growing chain producing the mRNA transcript. Termination occurs when the stop codon is recognized and RNA polymerase is released (Roeder, 1996; von Hippel, 1998; Maston et al., 2006).

The most important concept to understand is the regulation of transcription since the predicted binding sites play a major role in transcription. The regulation of transcription is controlled by the interaction between *cis* regulatory elements and *trans* factors. The interaction of these molecules forms transcriptional regulatory system. The regulatory region consists of the promoter and TFBS where RNA polymerase and TFs bind (Balleza et al., 2009).

1.1.1.1 TRANSCRIPTIONAL CONTROL

Transcriptional control is required to ensure that the genes are transcribed only when they are needed and are switched off when not needed. Each cell can control when and how a particular gene is expressed. Usually, transcription is controlled at the promoter region during initiation. The promoter region has the initiation site, where transcription starts. Nearly all genes contain long or short regulatory DNA sequences. These sequences need to be recognised by TFs to perform their function. It is the DNA sequence together with the protein molecules that control transcription. This DNA-protein interaction is strong and highly specific (Kadonaga, 2004). TFs interpret genetic information and direct the appropriate response to the transcriptional machinery. The specificity of protein coding genes is controlled by sequence specific DNA binding proteins. TFs interpret and transmit information from DNA sequence to the factors and cofactors that mediate RNA synthesis from the DNA template. Transcription is regulated by the collective action of TFs together with RNA polymerase, an enzyme responsible for the synthesis of RNA from DNA template. It also relies on various enzymes involved in modification of histones and other proteins. Three classes of RNA polymerase have been identified; these are RNA polymerase I, II and III. RNA polymerase II was found to play an important role in transcription of protein coding regions (Lee et al., 2004).

1.1.1.2 POST-TRANSCRIPTIONAL CONTROL

Gene expression is controlled at various stages and to cluster the information present in the genome, cells need to interact and coordinate with different layers of regulation. Post transcriptional levels of gene regulation include transcript turnover and translational control and these are the main parts of gene expression. The regulation of gene expression is an essential process in biology and is important for cell proliferation and differentiation during development. To assign a correct biological function, cells gather intrinsic and extrinsic information and coordinate several regulatory mechanisms of gene expression. At any stage, errors in regulation of gene expression can cause disease. Transcriptional control is an important step in gene regulation and is not as complex as post-transcriptional regulation. The complexity is due to the interconnection between the steps of different pathways of DNA and proteins. The control is accomplished by a combination of different RNA-binding proteins (RBPs). RBPs regulate a specific set of mRNAs. Small interfering RNAs, microRNAs and

protein effector complexes bind each other and control the degradation and translation of a particular transcript as a complex (Mata et al., 2005).

TRANSCRIPT TURNOVER

The degradation rate is regulated by control elements located in the 3'-untranslated regions (UTRs) of mRNA and these regions are recognised by many RBPs. Degradation occurs in the cytoplasm of both human and yeast cells (Wilusz & Wilusz, 2004). Even though most expression profiling methods are based on control of transcription, the target measured is the mRNA steady state levels which show the production and the stability of the transcript. mRNA turnover has been determined in various organisms by measuring mRNA levels at various time intervals after RNA polymerase II has been inactivated and this indicates the importance of decay in mRNA levels control (Khodursky & Bernstein, 2003). Like transcription rate, decay rate is also controlled. The half-life of mRNA depends on the composition of the molecular complex. As an example, transcripts that encode core metabolic proteins show a longer half-life compared to those encoding transcription factors (Wang et al., 2002). Transcripts with shorter half-lives allow dramatic changes in mRNA levels as a response to various conditions; this might be good for transcripts encoding regulatory proteins (Yang et al., 2003).

REGULATION OF TRANSLATION

Post-transcriptional regulation of gene expression also occurs during the translation stage and it involves global and transcript-specific actions to control protein synthesis. Proteins are regulated by post-transcriptional modifications and protein degradation (Lackner & Bahler, 2008). Transcript-specific regulation is specific to a particular group of mRNAs and is modulated by various mechanisms. It involves the interaction between RBPs and control element found in the UTRs of the target transcript. Translational regulation is crucial in cases where a sudden change in protein level is required. This includes cellular response to particular conditions and apoptosis, control of cell growth and division and cell differentiation and development. Conditions that lead to inhibition of translation show increase in the synthesis of proteins needed for cell survival. There is evidence that de-

regulation of control of translation can cause cancer (Watkins & Norbury, 2002; Pandolfi, 2004). The level of translation is measured by separating transcripts according to the number of associated ribosomes, the resulting segments are then examined with microarrays to get information about the control of translation (Beilharz & Preiss, 2004).

1.1.2 MOTIF DISCOVERY TOOLS

The study reviews motif discovery tools because it is important to know the limits on the quality of found motifs, when performing a quality measure. MEA is one form of quality measure. To put the research examples in context it is a good idea to know a bit about motifs and how TFs differ.

Motif discovery tools can discover motifs corresponding to the TF's DNA-binding domain. Ab initio methods can discover the DNA-binding motif of the ChIP-ed transcription factor from ChIP-seq peak regions. MEME is one of the most widely used motif discovery tools and DREME is a more recent example. Given a set of DNA sequences in FASTA format containing likely binding sites these methods discover motifs in DNA sequences. The methods search the input sequence and report the statistically significant sites they then represent as motifs. There are web based tools available in MEME suite. The MEME motif discovery algorithm performs both DNA and protein sequence motif discovery. It discovers TF binding sites and protein-protein interaction domains. To improve accuracy repeats may be masked from the input sequence (<http://www.repeatmasker.org>). By default, MEME searches for up to 3 motifs and considers the motif width between six and fifty, but users can change these default settings. The output is a coloured graphical alignment and the E-value, which is a measure of statistical significance of the motif. A low E-value indicates that the motif is statistical significant (Sharma et al., 2011). MEME output can be sent to other web services such as MAST to find if are there any genes containing the motif found by MEME. The user has the option of choosing the database to search (Bailey et al., 2006).

The performance of these tools had been tested. A study on *Saccharomyces cerevisiae* microarray data which include Rox1p and YAP1 was conducted. The aim was to identify the TF binding motifs of these genes based on their over-expression in the upstream region of the gene using MEME (input sequence sizes of 10, 25, 50 and 100). The MEME approach failed to find a motif matching known ROX1 consensus. For input size of ten sequences the correct YAP1 motif was ranked first by MEME and third for both 25 and 50 sequences. For sequence size 100, MEME did not find the motif (Conlon et al., 2003). The fact that MEME fails to find a known ROX1 binding site shows that we cannot only rely on its output. It is therefore important to use more than one method to be certain about the results.

DREME is a motif discovery tool used to find short (up to 8 bases) core DNA-binding motifs based on regular expressions. This algorithm is very fast, can analyse a large ChIP-seq datasets and can complete the motif search in minutes. Most algorithms fail to find a motif from ChIP-seq regions if the input consists of thousands of sequences. The algorithms use a fraction of the data, thus decreasing the specificity of motif discovery. DREME can discover primary and co-factor motifs and indicates the mechanism of binding of the TF. It is simpler, but shares some similarities with other algorithms like Trawler and Amadeus (Linhart et al., 2008) . In the interest of speed DREME search is restricted to regular expressions. The search is both exhaustive and heuristic. The Fisher exact test is used to locate statistically significant and discriminative motifs. To avoid self-overlapping motifs, DREME does not count the number of times the motif has occurred, but the number of sequences containing the motif in both data sets (Bailey, 2011). The algorithm takes two DNA sequences as input and a threshold is set, and then searches for regular expressions using a heuristic approach. The search is iterated until no new motif with an E-value less than a given threshold is found.

The performance of DREME was evaluated and compared with other motif discovery methods using ChIP-seq datasets. The evaluation considered the speed and ability to identify primary and secondary motifs. The study was conducted from mouse erythrocytes, mouse embryonic stem cells and human lymphoblastic cell lines. The output was compared to a database of known motifs using TOMTOM (Gupta et al., 2007). The author ran DREME, MEME (Bailey & Elkan, 1995) and nestedMICA (Down & Hubbard, 2005) using 13 mouse embryonic stem cell (mESC) ChIP-seq datasets. The two methods show poorer performance on more than 500 sequences than DREME. Other algorithms, WEEDER (Pavesi et al., 2004), Trawler and Amadeus were performed on the same datasets (Bailey, 2011). After running for 134 hours on E2f1 dataset nestedMICA did not finish. Bailey reports that DREME found about 33 significant motifs in mESC ChIP-seq data and these motifs correlate with the input. DREME found ChIP-ed FT motifs in 10 out of 13 mESC ChIP-seq data sets. The hypothesis that the known Nanog motif cannot be found was rejected when DREME found it. This suggests that DREME is superior to other methods. Chen et al., 2009 used WEEDER and nestedMICA motif discovery algorithms and reported that there is no motif for E2f1. Using DREME, Bailey found the motif matching known E2f1 in the E2f1 ChIP-seq dataset. A previous study of E2f1 ChIP-chip data also failed to find the E2f1 motif. DREME found

more cofactor motifs compared to other algorithms, however, it was slower than Amedeus and Trawler.

The MEME-ChIP tool set is used to analyse ChIP-seq peaks of large DNA data sets. It can be used as a both motif discovery and a motif enrichment analysis tool. It identifies and visualizes motifs. MEME-ChIP (Machanick & Bailey, 2011) employs both MEME and DREME algorithms in the discovery of motifs. The discovered motifs are visualized and their binding affinity is inferred. MEME is complemented by DREME, which uses a non-probabilistic model and restricted in finding short motifs. MEME-ChIP compares the motifs found by the two algorithms to a database of known motifs using TOMTOM algorithm. The input to MEME-ChIP is a set of FASTA formatted sequences each not less than 100bp long centered on the ChIP-seq peak. The sequence is centered and trimmed to 100bp. The resulting fragments are input to DREME and MEME algorithms to produce motifs. Machanick & Bailey. (2011) used MEME-ChIP and analysed ChIP-seq regions for T-cell acute lymphocytic leukemia 1 (Tal1) reported by (Kassouf et al., 2010) and report that both MEME and DREME found a known Tal1 motif. MEME found evidence that Tal1 forms a complex with GATA-1, but DREME did not find this complex. DREME reported Tal1 to be less significant than GATA-1 motif, indicating that this TF binds together with GATA-1. DREME found nine significant motifs while MEME found three (Machanick & Bailey, 2011). All results from previous studies indicate that no method is superior to another, therefore revealing the importance of using more than one tool so as to be certain about the results. Also it is important to include relevant previous knowledge.

In contrast to the above mentioned motif discovery algorithms, the PBM approach differs from them in that it does not start with DNA containing likely binding sites. The PBM approach aims to examine the TF binding specificity in an unbiased manner. PBM experiment measures the double stranded DNA-TF binding affinity. In a PBM experiment, a double stranded DNA containing 10 nucleotides long binding sites is bound by an epitope-tagged TF (Berger et al., 2006). This is followed by a wash step to remove all unbound TFs. The bound TFs are then labelled with a fluorophore tagged antibody specific for the protein. The microarray is now scanned in triplicate for the spots that are specifically bound by the TF. A p-value for specific binding is calculated for each spot based on its log likelihood ratio relative to the standard deviation of the Gaussian distribution. The binding site motif corresponds to the highly bound spot (Mukherjee et al., 2004).

A previous PBM study of mouse TFs suggested that TF-DNA interaction energies are complex (Badis et al., 2009). They report that 41 of 104 TFs studied had secondary binding preferences not captured by the primary PWM. They also report that 89 TFs were better represented by a linear combination of multiple PWMs than single PWM. They used 3 different methods to find PWMs and observed that no one method is superior to the others. Using 10 nucleotides long PWM compared to 8 nucleotides long sequence reduces the number of required model parameters to thirty. Simple model specificity allows easy analysis of TFs binding specificity to DNA (Zhao & Stormo, 2011).

BEEML method infers binding site sequences. Unlike other methods used to infer sequence motifs, BEEML considers the effect of a TF concentration and non-specific factors that contribute to the binding affinity (Zhao et al., 2009). BEEML-PBM algorithm is an extension of the BEEML algorithm for estimation of TFs model specificity. It is used to estimate the maximum likelihood of binding energies between protein-DNA associations based on the model of specificity. The advantage of using this method is that it minimises the number of parameters used to model the specificity of the TF. In cases where a large number of parameters is required, it becomes difficult to make predictions. Considering the binding energy solves this problem as each base pair contributes independently to the total binding energy (Zhao & Stormo, 2011).

In previous studies using BEEML-PBM, Zhao and Stormo (2011) found that their simple PWM model of specificity performs very well for most TFs. The authors then evaluated the ability of the PWM to predict TF binding preferences on a different PBM data. They used two arrays with different probes for PBM experiments, each representing all possible 10 nucleotides long binding sites. However, since the arrays have different probes, there was no clear explanation of the differences in probe intensities. The authors then conducted the same study now using 8 nucleotides long sequences instead of 10 nucleotides long sequences. 8-mer median intensities provide information about the variance of TF that can be explained by the PWM model. Zhao and Stormo found that a single BEEML-PBM PWM is sufficient to give enough information about PBM data. This supports the study conducted on the mouse factor *Plagl1*, where the PWM estimated from replicate one performs very well on replicate two data (Mukherjee et al., 2004). Zhao and Stormo's findings disagree with Badis et al.'s (2009) conclusion that *Plagl1* requires multiple PWMs. The BEEML-PBM PWM is

qualitatively not the same as the primary PWM identified by Badis et al., 2009. Zhao and Stormo suggested that a single BEEML-PBM PWM performs better than a combination of primary and secondary PWMs in the UniPROBE database (Zhao & Stormo, 2011). This study aims to investigate if the BEEML-PBM approach is better than the original PBM approach using CentriMo as a central motif enrichment tool.

1.1.3 MOTIF ENRICHMENT ANALYSIS (MEA) TOOLS

MEA tools are used to find under-represented and over-represented known motifs in a set of genes. Once the TFBS has been found it is important to extract biologically relevant information from it. As mentioned that the problem that we are currently facing is to determine the mechanism of gene regulation. MEA aims to determine which DNA-binding motifs are involved in regulation of transcription of a set of genes. The algorithm works by discovering enrichment of known binding motifs in the gene regulatory regions. As input, MEA takes in a set of genomic DNA sequences believed to be co-regulated and a set of known DNA-binding motifs discovered by motif discovery tools for a TF of interest. It then determines which (if any) of the discovered motifs may be direct regulators of the genes. MEA is becoming more and more useful as the number of known DNA-TF binding motif databases increases (McLeay & Bailey, 2010).

MEME-ChIP uses the AME algorithm to perform motif enrichment analysis. AME inputs a set of DNA sequences in FASTA format and motif database file(s) in MEME format. AME finds enriched motifs from a given database of known motifs. AME ranks the motifs by computing their Fisher exact test p-value based on the number of sites for each motif above a threshold in the original sequences versus a shuffled version of the original sequences (McLeay & Bailey, 2010). Compared with other tools AME has an advantage of allowing a very large set of data and is capable of detecting very low levels of enrichment of TF-DNA binding motifs that *ab initio* discovery algorithms fail to detect. This algorithm achieves higher sensitivity by restricting the motif search to DNA binding motifs that have been already identified. The input sequences are trimmed by MEME-ChIP to 100bp and AME is performed using these sequences. AME as used in MEME-ChIP compares the sequences with databases of known motifs and outputs statistical enrichment. A previous study indicates that AME was successful in finding a known Tal1 binding motif. AME reports that Tal1 binds in combination with GATA1 and these TFs function in erythropoiesis. AME algorithm found Tal1 ChIP-seq regions to be enriched for fifteen known vertebrate motifs. Tal1 was

found to be less enriched than Tal1-GATA-1 complex. The authors applied the AME algorithm to NFIC ChIP-seq regions to determine a motif matching a known consensus sequence, which other algorithms fail to find. AME found the motif and ranked it 85th out of 532 motifs in JASPAR and UniPROBE databases (Machanick & Bailey, 2011).

Clover is a MEA algorithm that identifies functional sites in DNA sequences. The input to clover is a set of DNA sequences having similar functions. It then compares the input to TF binding patterns and identifies which motifs are statistically over-represented. Firstly the method calculates the raw score, to measure the quality of the motif found in the input sequence. Secondly, it calculates the p-value corresponding to the raw score. A very low p-value (<0.01) shows that the motif is over-represented in the input sequence, suggesting that the motif has a biological function. A higher p-value (>0.99) means that the motif is over-represented in the background sequence rather than in the input sequence (Frith et al., 2004). McLeay & Bailey, (2010) used yeast TF binding data available from ChIP-chip experiments, considering only the data with known DNA-binding motifs. The aim was to investigate the ability of MEA methods to determine and rank the known motif of the ChIP-ed TF from 237 yeast ChIP-chip data set. Four MEA methods (Fisher, mHG, Ranksum and Clover) were tested for the ability to identify a known yeast TF motif. They used ChIP-chip fluorescence p-value as the biological signal. The authors observed that Clover performed very well compared to other methods. Clover identified ChIP-ed TF in all yeast datasets in consideration. The correct TF was ranked in the 84th percentile by the superior method, clover, and the 80th percentile by mHG-YFP. Ranksum-YFP had poor performance compared with other three methods. They found that no method identified a ChIP-ed TF very well on all datasets. Clover was found to be a better method especially if the data has low signal to noise.

PASTAA is also a MEA method that predicts the binding affinity of a transcription factor to a promoter. This method is primarily designed to determine TFs involved in regulation of tissue specific transcription. Genes expressed in response to a particular signal are believed to have the same binding location for the same TF. The method scans the promoter regions of the genes expressed together and identifies over-represented sequence patterns. The method is sensitive to repeat-containing sequences. Wasserman & Fickett, (1998) were the first to use PASTAA and successfully predict the TFs involved in the regulation of muscle specific genes. Other studies by (Frith et al., 2004; Qian et al., 2005; Yu et al., 2006) found more TF-

tissue associations. The authors were interested in proximal promoters of tissue specific genes obtained from EST data. A method requires a correct definition of a set of genes co-regulated by a particular factor. The information about these groups of genes can be extracted from Gene Ontology (Hill et al., 2002) and KEGG databases. PASTAA ranks the genes based on the predicted TF-promoter binding affinity. The idea is that the target genes of the TF are expected to rank high. The degree of gene-tissue association from the ranked list is expected to match another ranking of the same gene (McLeay & Bailey, 2010).

The performance of PASTAA method has been tested on other examples. Roeder et al., (2009) used the yeast genome-wide datasets on in vitro TF–DNA interactions for the three TFs (Rap1, Mig1 and Abf1) from (Mukherjee et al., 2004), and in vivo ChIP–chip data from Harbison *et al.* (2004) for more than 200 TFs in different cellular conditions. They then analysed the data sets corresponding to 25 TFs for which position specific frequency matrix (PFMs) is available in TRANSFAC. Odom et al., (2004) conducted a study where they measured the binding of HNF1, HNF4 and HNF6 to human promoters. PASTAA uses the available data (p -values) to rank the promoters for a given TF. PASTAA reports a significant overlap between predicted targets and experimentally bound sequences. The ChIP-chip data for the 3 TFs in question was analysed to test the ability of PASTAA to determine enrichment of TF targets from a set of vertebrate sequences. As input, they ranked promoter sequences based on their binding affinity to HNF in consideration. PASTAA reports the highest association for position frequency matrices (PFMs) corresponding to HNF1 and HNF4. The HNF6 motif was ranked fifth. Their findings agree with Smith et al.'s (2005) results who found the same factors as highly associated with HNF6 dataset.

CentriMo is a new algorithm, which measures central enrichment of a motif and has previously been used as a MEA tool. CentriMo is based on the idea that ChIP-seq peak calling methods are likely to be biased towards the centre of the putative binding region, at least in cases where there is direct binding. CentriMo calculates a binomial p -value representing central enrichment, based on the central bias of the binding site with the highest likelihood ratio. In cases where binding is indirect or involves cofactors, a more complex distribution of preferred binding sites may occur. CentriMo works by identifying the best binding site in each sequence, based on the log likelihood ratio, for a given motif. The algorithm does not score a sequence if no site has a score above a set log likelihood ratio

threshold (default: 5). It calculates a binomial p -value for each central bin width, counting the number of “best” sites inside and outside the bin, and selects the lowest p -value, as well as recording the bin width of the lowest p -value. In cases where the CentriMo distribution has a clear central peak and a very low p -value (10×10^{-1000} or less) and low p -value in a very narrow window, then we are confident that there is a direct binding by a single factor.

The algorithm has been used as a CMEA tool and its performance has been compared to other MEA tools. A previous study has been conducted on an NFIC dataset. The aim was to find a motif matching the known *in vitro* motif from NFIC dataset. Different methods such as MEME, DREME, WEEDER, Amadeus and Trawler were employed and failed to find this motif. The consensus sequence of a motif found by DREME only matches half of the NFIC motif consensus sequence. This motif is less significant because it ranks nineteenth out of 24 motifs. The AME algorithm also failed to rank this motif. Using CentriMo it was found that the known NFIC motif is the most centrally enriched among 532 motifs in the combined JASPAR and UniPROBE database, although the width was too large and the number of ChIP-seq regions containing a known motif was small. This indicates how CentriMo can be useful even if the ChIP-seq data used is of low quality and when the motif discovery algorithms have failed. Of the five motifs that CentriMo reported three matches with a known consensus (Bailey and Machanick, 2012).

In previous studies DREME and SELEX were used to identify motifs from the Nanog ChIP-seq data. However, the two did not agree. Bailey and Machanick. (2012) then used CentriMo to decide which motif is correct. To decide between the Nanog motifs found by DREME and SELEX, CentriMo confirmed that a motif found by DREME represents *in vivo* binding of Nanog in mouse embryonic stem cell (mESC). The authors used CentriMo and changed the last position of the CVATYA (found by DREME) motif to cytosine or thymine. A drop in central enrichment was observed, indicating the importance of adenine rather than cytosine and thymine for Nanog binding. CentriMo was also used to analyse E2f1 dataset. CentriMo reports that E2f1 does not bind directly to DNA, but in combination with others, since all the motifs in the combined JASPAR and UniPROBE database found by CentriMo were not centrally enriched. The findings suggested that GABPA binds DNA in close proximity to E2f1. In previous studies by (Bieda et al., 2006) it was suggested that YY1 interacts with E2f family (E2F2 and E2F3) to stimulate transcription. The finding, using CentriMo that

suggested cooperative binding between YY1 and E2f1 supports this study (Bailey & Machanick, 2012).

These results show that including central motif enrichment can be useful. CentriMo is superior to other algorithms as a previous study by (Bailey & Machanick, 2012) indicated that CentriMo can be used in ChIP-seq data as a MEA tool and to determine direct or cooperative TF-DNA binding in ChIP-seq data. CentriMo can be used or is useful in cases where motif discovery algorithms have failed to discover a motif.

1.1.4 CLASSIFICATION OF TRANSCRIPTION FACTORS

The mechanism of gene regulation can be better understood if the classes where TFs belong are known. TFs are classified into different groups based on their DNA binding domains (DBD). This classification aids when a novel protein is identified. To understand the biological function of the protein, it has to be classified into a correct class. Proteins are grouped based on their biological functions. To date, computational methods have been used to classify TF and non-TF proteins (Qian et al., 2006). Also, TFs are classified based on TFBS clustering using information content. Therefore, any TF belonging to a specific TF class category has a chance of binding to any TFBS of the same group (Reddy et al., 2006). This classification information can be useful when CentriMo gives ambiguous results and also to validate if we have found a correct motif. The DBD mediates interactions between TFs and their target DNA. Various TFs can bind to the same regulatory sequence (Wittkopp, 2010). TFs are classified based on DBD structure. Many TFs share a similar DBD and hence the same binding site and specificity. As an example, there are 17 Kruppel-like factors (Klf) family members in mammals. The DBD of these factors is characterized by the presence of conserved Cys2His2 zinc fingers that bind to a CACC-box motif (Pearson et al., 2008). Therefore, knowledge of the relevant TF is important.

The basic helix-loop-helix (HLH) structure consists of two amphipathic alpha helices on the amino-terminal and hydrophobic residues at the carboxyl-terminal end. The helices are connected by amino acid residues of different length that form reverse turns and loops (Luscombe et al., 2000). TFs that belong to this family play a vital role in the transcriptional network of most developmental stages. They play a role in cell proliferation and

differentiation, cell lineage determination, sex determination, and other essential processes in organisms ranging from yeast to mammals. HLH proteins are distinguished by the presence of conserved bipartite domains for DNA binding and protein–protein interaction. A basic residue motif permits HLH proteins to bind to an E-box (CANNTG). A second motif of hydrophobic residues (HLH domain) allows these proteins to interact and to form either as both a homodimer or heterodimer (Atchley & Fitch, 1997).

The zinc finger motif characterizes the zinc finger-type TFs and is one of the common motifs in the eukaryotic cell found from enzymes to TFs (Suzuki et al., 2005). The zinc finger motifs bind to G-C rich DNA sequences (Yang et al., 2009). The DBD of this class of TFs is characterized by the presence of nine repeats of a 30 amino acid sequence of the form Tyr/Phe-Xaa-Cys-Xaa-Cys-Xaa₂, 4-Cys-Xaa₃-Phe-Xaa- Leu-Xaa₂- His-Xaa₃ 4- His-Xaa₅, where Xaa is a variable amino acid (Miller et al., 1985). These repeats have two consistent pairs of histidine and cysteine residues, which co-ordinate a single zinc atom. This results in a finger-like structure with conserved phenylalanine and leucine residues and some basic residues in the finger project from the surface of the protein. The contact is between the tips of the fingers and the major groove of the DNA and other fingers bind on the opposite sides of the helix (Pabo & Sauer, 1992). The largest group of TFs in eukaryotic genomes is made up of zinc coordinating proteins. The DNA binding motif is distinguished by the tetrahedral coordination of zinc ions and conserves cysteine and histidine residues. The zinc coordinating motifs are not restricted to DNA binding. They are also common in domains involved in protein-protein interactions (Luscombe et al., 2000).

The leucine zipper is a domain of the basic region leucine zipper (B-ZIP) transcription factors. This domain plays an important role in dimerization and leucines appear every 7 residues. The structure of this class of TFs was proposed by Crick in 1953. It consists of lysine and arginine in the N-terminal region that bound to the major groove of DNA and also forms a dimer GCN4 that binds to DNA. The C-terminal region is made up of an amphipathic alpha helix that forms the leucine zipper through dimerization. In previous studies the interaction of leucine zipper proteins with DNA was examined. The DNA binding specificity of the leucine zipper proteins is not yet well understood. The homodimers bind to a palindromic DNA sequence, while heterodimers bind to any half site. For DNA binding, the leucine zippers have to come into contact with each other in a way that basic regions are in

the major groove of DNA. To determine biological function (which is the important problem in bioinformatics) for this class of proteins, it is important to use proteins that act against the functioning of B-ZIP genes and repress DNA binding. This repression shows whether a family of B-ZIP TFs in question activates or inhibits transcription. A study of the leucine zipper region amino acids shows that all seven residues that form B-ZIP proteins are conserved (Krylov et al., 2001).

1.1.5 OVERVIEW OF TFs AND THEIR BINDING PARTNERS

Literature evidence for the sample TFs used for this study (see section 2.4) is needed to validate if the motifs identified using CentriMo are true binding motifs, rather than concluding based only on CentriMo output. Prior evidence is also useful in cases where the CentriMo distributions are not well centered and it becomes less clear what has happened, as in cooperative or indirect binding cases.

Max is a member of the basic helix-loop-helix protein family. Max has been shown to form a heterodimer with Myc and Mad proteins and regulate gene expression by binding to the E-box DNA sequence (Larsson et al., 1997; Hurlin et al., 2006). The authors reported that Myc and Max interact with each other via helix-loop-helix/Leucine zipper domain and specifically bind an ACGTG DNA sequence. C-Myc always requires interaction with Max to perform its functions as a TF. Mad family TFs also interact with Max and recognize the same binding sequence as Myc. Their findings lead to a hypothesis that Max, Myc and Mad TFs form a network to perform their functions as transcriptional regulators. Egr1 is a member of the immediate early gene family. It is essential in the control of cell growth, differentiation and apoptosis. A previous study reported that Egr1 performs its function by directly interacting and activating the promoter. Wagner et al. (2008) reported that Egr1 share a similar binding site with Sp family members. Egr1 also known as krox-24, ZIF268, and TIS8, is an early response protein and is a member of zinc finger family of TFs. The Egr1 DBD contains 3 zinc fingers positioned between amino acids 332-416 in the C-terminal region of the protein. Egr1 binds to G-C rich regions of DNA sequences (Yan et al., 2000).

The Ets family is characterized by the presence of a unique Ets domain that binds to purine-rich DNA sequence. Members of the Ets family have different binding specificities; a single amino acid change in the Ets domain can lead to a change in DNA binding specificity. A change in moloney sarcoma virus enhancer sequence reduces Elf-1 binding but does not

affect Ets-1 binding (Oikawa et al., 2003). Tcf is the family of TFs consisting of four members, that is, Tcf-1, Lef-1, Tcf-3, and Tcf-4 and they bind in the minor groove of a DNA sequence. Gustavson et al. (2004) found that the binding sites for Tcf have a G/C base pair at position 8 of the binding site (TTCAAAGG) show an increase in Lef-1 HMG box binding affinity. TCF protein family (Elk1, SAP1 and SAP2) and the myocardin protein family (Mkl1 and Mkl) are known Srf cofactors (Halene et al., 2010).

A previous work on analysis of the Srf data set has shown that Srf motifs are highly similar to Ets1 and GABP (Wallerman & Motallebipour, 2009). A prior publication using QuEST reports that Srf and GABP peaks are in close proximity to another, suggesting that the two TFs interact with each other. Srf and GABP are thought to function as transcription activators. Srf is a member of the Mad family of TFs expressed in hematopoietic cells. It has been shown to play an important role in muscle differentiation. Its role in hematopoiesis is not yet understood. Mkl1 has been found to bind in combination with Srf and to function in megakaryopoiesis. Srf has been found to have binding specificity to CCTTATATGG consensus sequence. Srf is activated by interaction with its co-factors to enhance transcription (Valouev et al., 2008).

A study on analysis of Irf4 data sets suggests that Irf4 forms a complex with Jundm2 secondary and binds DNA cooperatively with it. The authors used SpaMo to predict the interactions between Irf4 and Jundm2 TFs (Whittington et al., 2011). A prior study on computational analysis of Irf4 from ChIP-chip experiments reports that both Irf4 and Irf5 recognize the same consensus sequence (CGAAAC) (Badis et al., 2009). Irf4 is an interferon regulatory factor essential for lymphocyte activation. Irf4 associates with NFATc2 to facilitate NFATc2-driven transcriptional activation of the IL-4 promoter. Irf4 physically interacts with NFATc2 to perform this. Irf4 combines with NFATc2 and the IL-4-inducing TF, c-maf, to augment IL-4 promoter activity. The identification of Irf4 as a partner for NFATc2 in IL-4 gene control gives an essential molecular function for Irf4 in T helper cell differentiation (Rengarajan et al., 2002).

GATA3 is expressed in T cells and erythroid cells and it does encode a polypeptide containing two zinc fingers almost similar (92% identity) to GATA-1. A previous study indicates that GATA-3 binds to an AGATAG sequence in human T cell receptor alpha

enhancer (Ho et al., 1991). The GATA family of TFs recognizes the WGATAR consensus sequence (Ko & Engel, 1993).

POU2f2 is a member of the POU class 2 family of TFs characterized by the presence of POU domain. POU domain is divided into two subdomains, POU specific (POUs) subdomain and POU homeodomain (POUh). These domains consist of helix turn helix motifs. A previous study reports that POU interacts with oct-1 and oct-2, which also belong to POU family. This suggests that POU family TFs interact with other TFs containing a POU domain and cooperatively bind DNA (Veenstra et al., 1997). Most Pou domain binding sites identified are A/T rich. The co-activator domains have been identified from DNA-binding sites and examined. The Pou homeodomain binds to the TAAT region of the binding region while the specific domain is flexible can bind to either GARAT or TAATGARAT regions. The expressed Pou domain protein in the developing nervous system has been suggested to play an important role in embryogenesis (Ryan & Rosenfeld, 1997).

A previous study suggests interactions between Foxa2, GABP and Hnf4a (Wallerman & Motallebipour, 2009). Fang et al. (2012) conducted a study and examined the DNA binding specificity of three members of the nuclear receptor family, that is, Hnf4a, Nr2f2 and Rxra by PBM and reported that Rxra does not bind to DNA in the absence of Nr2f2. Fang et al. (2012) used PBM data and reported that Rxra binding sites are almost identical to those of Nr2f2. The authors also suggested competitive binding between these two TFs, Rxra was found to activate transcription, while Nr2f2 inhibits transcription.

1.2 PROBLEM STATEMENT

To date, a lot of work has been done in predicting transcription factor binding sites (TFBS). However, the underlying mechanism of how transcription factors bind to DNA is not yet fully understood (Ahmad et al., 2004). Knowledge of how DNA-protein interactions regulate gene expression is crucial for understanding biological processes. To understand these processes it is important that the motifs in the sequence and their binding factors are known. Generally, the shared motifs can be found using motif discovery tools, but the problem with *ab initio* methods is that they do not give information about factors thought to bind to the predicted motifs. MEA tools include this information by showing which motifs bind to which factors. A TFBS is generally represented using a motif, a representation of the probability that a given base is present at a given position in a DNA sequence where binding occurs. If a motif is a good representation of a TFBS for a particular TF, we expect to find that motif over-represented in locations in DNA that are likely binding sites. There are various ways this over-representation can be measured, including measuring the match of a motif to every possible binding site, and comparing the probability of the level of matching found to a background probability.

Several MEA tools that detect statistical over-representation of a motif in a set of DNA sequenced have been developed and published (Liu et al., 2003; Zheng et al., 2003; Elkon et al., 2003; Sharan et al., 2003; Haverty et al., 2004). Most previously used methods tend to discard important information by collapsing matrix scores at each position to a binary quantity that is above or below the threshold. The methods are also unclear whether to count one match per sequence, a few matches per sequence or all matches in each sequence. It has been found that the regulatory regions of higher eukaryotes usually contain multiple binding sites for the same TF (Papatsenko et al., 2002). Therefore taking multiple sequence matches into consideration is useful in the discovery of motifs by statistical representation. MEA tools are used to find under-represented and over-represented known motifs in a set of genes. The latest-generation technology for identifying TFBS, ChIP-seq, produces a large set of short DNA sequences that a particular TF should have bound to, if the experiment was error-free. The mechanism of binding is not always straightforward because many TFs bind only in combination with others (cofactors), and some only bind indirectly (other TFs bind to the DNA, and they bind to these). MEA tools are becoming more and more useful as the number of known DNA-TF binding motif databases increases (McLeay & Bailey, 2010).

To investigate which method for creating motif databases is better than the other, CentriMo is used as a central motif enrichment analysis (CMEA) tool to compare motifs for a given TF from ChIP-seq experiment corresponding to TFs for which there are UniPROBE and BEEML-PBM motifs. It is expected that, should the BEEML-PBM approach be comparable to PBM, it should perform competitively as measured by the number of examples for which it achieves a lower CentriMo p -value than PBM for a like motif. The claim made by Zhao and Stormo (2011) that their approach is simpler and on the whole produces acceptable results is validated using CentriMo. CentriMo is a statistical method that ranks motifs based on the central enrichment p -value. This algorithm solves the problem of choosing background sequences as the flanking sequences perform this function. CentriMo can be used as a tool for motif enrichment analysis and for comparing variants on a motif. CentriMo can differentiate if the central enriched motif is a result of direct DNA binding of the ChIP-ed TF or is a result of co-factor motifs. Also, the algorithm is faster compared to other MEA tools. It compares known motifs with the input sequence and returns the site-probability curve in minutes. CentriMo is useful in cases where motif discovery algorithms have failed to discover a motif (Bailey & Machanick, 2012). Other MEA tools are appropriate in cases where central enrichment is inadequate, for instance, in detecting cofactors.

1.3 AIMS AND OBJECTIVES

This study aims to investigate the claim made by Zhao & Stormo. (2011) that they have identified a simpler method than that used to derive the UniPROBE motif database for creating motifs from protein binding microarray (PBM) data (Berger et al., 2006), which they call the BEEML-PBM (Binding Energy Estimation by Maximum Likelihood for PBMs). A specific claim made of the Zhao-Stormo is that the BEEML-PBM method is not only simpler, but provides comparable specificity to the UniPROBE PBM method.

The approach is to use CentriMo as a central motif enrichment analysis tool to compare motifs for a given TF. Should BEEML-PBM approach be comparable to PBM, it should perform competitively as measured by the number of TFs for which it achieves a lower CentriMo central enrichment p-value than PBM for a like motif. CentriMo measures central enrichment of the TFBS represented by a motif in a set of DNA sequences. To measure the quality of motifs from BEEML-PBM and the UniPROBE motif databases, the degree of central enrichment of motifs from each database is tested.

A secondary goal of this study is to identify cases where a motif is of poor quality, or there are problems with ChIP-seq data set. The purpose of MEA is to find which motifs from a database of known motifs are most "enriched" by a particular measure and see if the motif or motifs associated with binding are the most enriched, as further evidence that de novo motif finders have found the right motifs.

The approach analyses a set of candidate TFs and determines where a particular TF binds in a DNA sequence, and which cluster of similar TFs are involved in a particular co-operative binding. Since this is a statistical method it is expected to rank motifs directly involved in DNA binding near the top.

The study also aims to identify alternative motifs in the literature and use CentriMo on ChIP-seq data that correspond to those motifs and gives comparisons of the alternative motifs.

RESEARCH QUESTION

Does BEEML-PBM method gives better or worse motifs compared to the PBM method used to create the UniPROBE motif database? This work is extended by evaluation of the quality of motifs in different cell lines using CentriMo.

CHAPTER 2: METHODS

This study uses CentriMo as a central motif enrichment analysis (CMEA) tool on BEEML-PBM and UniPROBE databases to investigate the claim made by Zhao and Stormo (2011) that they have identified a simpler method than that used to derive the UniPROBE motif database for creating motifs from PBM. CentriMo is also used against a wider motif database to determine whether there may be cofactors or indirect binding. CentriMo is useful in cases of poor ChIP-seq data quality, where it can point to a problem. If there is a TF expected to have a clear binding specificity but CentriMo has a high p -value or other indications of less specific binding, the possibility that the ChIP-seq experiment has failed is considered. Because there is no method that is perfect, AME and SpaMo are used for further investigation. SpaMo is used to validate the CentriMo output especially in cases where CentriMo distribution is not well centered and it is less clear what has happened. SpaMo infers physical interactions between a given TF and TFs that bind at a neighbouring sites at the DNA interface. AME is employed as an independent measure to see if CentriMo has found the correct motifs. AME can give an indication of an over-represented TF or TFs that are not well centred, and that is useful for picking up cofactors or if CentriMo fails to find a well centred TF for a reason like indirect binding, AME may find enrichment of TFs to which the TF of interest binds.

2.1 CENTRIMO

CentriMo is an algorithm that measures central enrichment of a motif and has previously been used as MEA tool. CentriMo takes a set of equal length (500bp) genomic sequences in FASTA format and TF binding motifs as input. These sequences are centered on the ChIP-seq peaks. CentriMo is based on the idea that ChIP-seq peak calling methods are likely to be biased towards the center of the putative binding region, at least in cases where there is direct binding. CentriMo works by identifying the best binding site in each sequence, based on the log likelihood ratio, for a given motif. The algorithm does not score a sequence if no site has a score at or above a set threshold (default: 5). It calculates a binomial p -value for each central bin width, counting the number of “best” sites inside and outside the bin, and selects the lowest p -value, as well as recording the bin width of the lowest p -value. In cases where binding is indirect or involves cofactors, a more complex distribution of preferred binding sites may occur. A low CentriMo p -value and low width of maximum enrichment (about 100bp) are strong evidence that the motif in consideration is the true binding motif. When the

CentriMo distribution has a clear central peak, a very low p -value (sometimes on the order of 10^{-1000} or less) and a minimal p -value in a very narrow window, this indicates binding by a single factor. CentriMo also plots the site-probability curve with the center of the curve as position zero. CentriMo is run with the database described in section 2.4 against the CHIP-seq binding regions. For all CentriMo results reported here, the default threshold score of ≥ 5 bits is used. A previous study has shown that using a threshold between 3 to 8 bits generally gives the same results. (Bailey and Machanick, 2012).

As a further evidence that the lack of distinct peaks in CentriMo distributions is a result of indirect or cooperative binding, CentriMo is run using a compendium of motifs containing all vertebrate motifs in the JASPAR CORE 2009 database (Sandelin et al., 2004) and the UniPROBE mouse database. The procedure is the same as above except that now the UniPROBE and JASPAR motif databases are used. A script that trims the sequences to 500bp and converts the file into a FASTA file is created and CentriMo is run on the command line. For the script used, see the appendix.

2.2 SPAMO

The spaced motif analysis (SpaMo) algorithm determines enriched spacings and identifies the interactions between a given motif and TFs that form complexes with the given TF. SpaMo takes in a set of DNA sequences in FASTA format for a given TF, a primary motif represented as a PWM (Stormo, 2000) and a database of secondary motifs. The algorithm tries to find enriched motif spacing patterns by searching for the strongest primary motif binding site for the given TF then searches the neighbouring sites for secondary motif binding sites and then calculates the probability of the spacing relative to the primary TF. It applies a binomial test to assess significance of the enriched spacings. SpaMo outputs a graph of secondary-primary motif site displacement. The graph is divided into four categories, indicating the position where primary and secondary motifs appear on the same strand and when they appear on opposite strands. The output also shows the position (downstream or upstream) of the secondary motif site relative to the primary motif site. Analysing these resulting displacements is useful as they indicate physical interactions between the TFs. Some TFs share the same DNA binding domain and therefore bind with same specificity. Because multiple TFs can bind to the same consensus sequence, it is important that the prior knowledge of a given TF is applied so as to identify the TF corresponding to the reported secondary motif spacing enrichment (Whittington et al., 2011).

2.3 AME

AME is used as an independent measure for this research. Like CentriMo, AME is a MEA tool that given an input sequence, searches for the enriched DNA binding motifs in one or more databases of known motifs. Unlike CentriMo, AME is *not* based on central enrichment. Using AME, the motif enrichment from databases of known motifs (BEEML-PBM and UniPROBE) is compared to motifs reported by CentriMo. The enrichment p-values are compared to explore if the two algorithms find the same motif. AME inputs a set of DNA sequences in FASTA format and motif database files in MEME format. Given a set of equal size DNA sequences (500bp) used for CentriMo, AME finds enriched motifs from a given database of known motifs. AME ranks the motifs by computing their Fisher exact test p-value based on the number of sites for each motif above a threshold in the original sequences versus a shuffled version of the original sequences (McLeay & Bailey, 2010). AME outputs two files, HTML and a text file. For the script used, see the appendix.

2.4 DATA

The ENCODE project (The ENCODE Project Consortium, 2011) contains ChIP-seq data for many TFs. A total of 13 TFs is used in this study. The ChIP-seq data used is from the ENCODE project, corresponding to TFs for which there are UniPROBE (Newburger et al., 2009) and hence BEEML-PBM (Zhao & Stormo, 2011) motifs. Because ENCODE data are from diverse sources, the variations are minimized as far as possible. Since GM12878 is the most studied and common cell line in the ENCODE project, for example, where that cell line is available, ChIP-seq sequences from that cell line are used. The data used for the study are from the HudsonAlpha Institute for Biotechnology (HAIB) (tables 2-1 and 2-2) and Stanford/Yale/USC/Harvard (table 2-3) laboratories. However, Stanford/Yale/USC/Harvard laboratory only contains the ChIP-seq data for only 3 out of 13 TFs used in this study. Table 2-4 summarises the useful information about the cell lines used. The ENCODE peak file is converted into a BED file, which is then submitted to the UCSC genome browser (Karolchik et al., 2004) to obtain a FASTA sequence. The coordinates of ChIP-seq sequences are widened to 500 bp centered on the peaks, and repeat-masked sequence data is used (hg19). For each TF, a custom motif database containing all BEEML-PBM and UniPROBE motifs for that TF is created.

TF	Cell line	Protocol	Treatment
Egr1	GM12878	Biorupter, PCR 1-round	None
Hnf4a	HepG2	Biorupter, PCR 1-round	None
Ets1	GM12878	ChIP AMPure XP	EtoH, 0.02% 1h
Foxa2	HepG2	Biorupter,PCR 1-round	None
Gata3	T-47D	Sonicator,PCR 1-round	DMSO_002pct
Irf4	GM12878	PCR 1-round	None
Max	K562	Sonicator,PCR 1-round	None
Nr2f2	K562	ChIP AMPure XP	None
Pou2f2	GM12878	PCR 1-round	None
Rxra	GM12878	PCR 1-round	None
Sp4	H1-hESC	ChIP AMPure XP	None
Srf	GM12878	PCR 1-round	None
Tcf3	GM12878	PCR 1-round	None

Table 2-1: Data used to run CentriMo.

The table columns show the TF name, cell line (representing the tissue in which the ChIP-seq experiment was carried out), protocol and treatment (if any) used for ChIP-seq experiment. The TFs used are available from ENCODE Project.

TF	Cell line	Protocol	Treatment
Egr1	H1-hESC	Sonicator,PCR 1-round	None
Rxra	H1-hESC	Sonicator,PCR 1-round	None
Srf	H1-hESC	PCR 1-round	None
Ets1	A549	ChIP AMPure XP	EtoH, 0.02% 1h
Foxa2	A549	Sonicator,PCR 1-round	EtoH, 0.02% 1h
Gata3	A549	ChIP AMPure XP	None
Max	A549	ChIP AMPure XP	None
Nr2f2	HepG2	ChIP AMPure XP	None
Pou2f2	GM12891	PCR 1-round	None

Table 2-2:Data used to investigate variations across cell lines.

The TF ChIP-seq data obtained from different cell lines than those in table 2-1. The results are compared with those found using the cell lines in table 2-1. Columns as in table 2-1.

TF	Cell line	Control	Treatment
Max	A549	IgG-rab	None
Gata3	MCF-7	UCDavis input control	None
Hnf4a	HepG2	Standard control	forskolin

Table 2-3: ChIP-seq data from ENCODE/Stanford/Yale/USC/Harvard Lab.

Stanford/Yale/USC/Harvard Lab contains only 3 TFs that have motifs in the BEEML-PBM and hence UniPROBE motif database. This data is used to validate variation across cell lines using ChIP-seq data from different laboratories and also to validate the consistency with the previous cell lines used .

Cell line	Description	Tissue	karyotype
GM12878	lymphoblastoid cell line	blood	normal
GM12891	lymphoblastoid cell line	blood	normal
H1-hESC	embryonic stem cell	embryonic stem cell	normal
HepG2	derived from a male patient with hepatocellular carcinoma	liver	cancer
K562	produced from a female patient with chronic myelogenous leukemia	blood	cancer
A549	derived from a lung carcinoma	epithelium	cancer
MCF-7	derived from mammary glands	breast	cancer
T-47D	epithelial cell line derived from a mammary ductal carcinoma	breast	cancer

Table 2-4: Explanation of cell lines.

Table 2-4 is adapted from the ENCODE sites (<http://encodeproject.org/ENCODE/cellTypes.html>) and ENCODE Project common cell types, last updated 9 March 2012, online: <http://www.genome.gov/26524238>.

CHAPTER 3: RESULTS

CentriMo is used as a CMEA tool for a given TF to investigate which method gives better motifs between the BEEML-PBM and the original PBM method used to create UniPROBE motif database. We expect that should the BEEML-PBM approach be comparable to PBM it should perform competitively as measured by the number of TFs for which it achieves a lower p-value than PBM for a like motif. AME and SpaMo are used as tools for further investigation. SpaMo infers physical interactions between a given TF and TFs bound at a neighbouring sites at the DNA interface. AME can give an indication of an over-represented TF or TFs that are not well centred, and that is useful for picking up cofactors or if CentriMo fails to find a well centred TF for a reason like indirect binding. The results also show the binding specificity of the TFs in different cell lines.

3.1 Comparing the BEEML-PBM motifs with the PBM motifs using CentriMo

A total of thirteen TF ChIP-seq data from ENCODE project is analysed using CentriMo on combined BEEML-PBM/UniPROBE motif database. For all CentriMo results reported here, the threshold score of ≥ 5 bits is used. CentriMo results fall into four different categories based on the motif's central enrichment. The categories involve motifs with: (i) clear central enrichment with well defined peaks from both motif databases, (ii) central enrichment with less sharply defined peaks, (iii) central enrichment from only one database (UniPROBE), and (iv) no motif from either database with central enrichment.

3.1.1 Motifs with clear central enrichment from both motif databases

Of the 13 TFs analysed five (Egr1, Max, Nr2f2, Gata3 and Hnf4a) show a clear central enrichment with well-defined unimodal peaks and bin widths of about 100bp from both motif databases. The TFs also have statistically significant central enrichment p-values. A low p-value and narrow bin width indicate direct DNA binding. BEEML-PBM has the lowest p-value for 3 out of 5 cases, and in one of the two cases where the UniPROBE motif has the lowest p-value, it is the secondary motif (see Table 3-1). The slight shift away from the center on the Hnf4a motifs may be due to the peak calling algorithm failing to correctly locate the binding site. The binding motifs for Max from both databases show that Max is an E-box TF; E-box TFs bind to the CANNTG sequence.

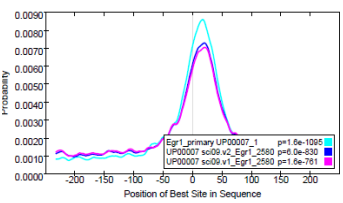


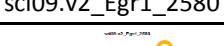
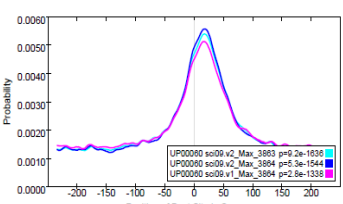

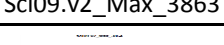
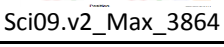
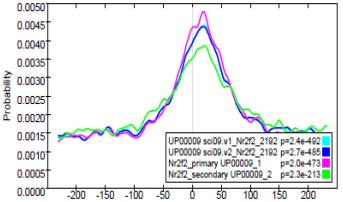

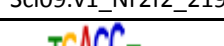
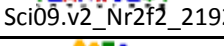
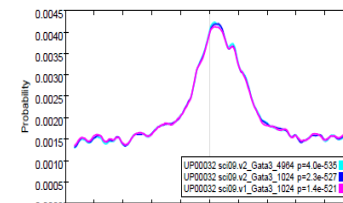


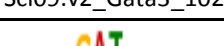
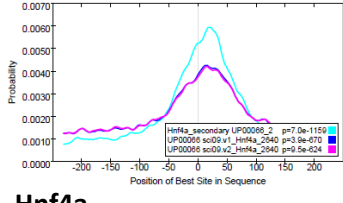



CentriMo Distribution	Logos	p-values	Sequences Best DB	Matches	Width
 <p>Egr1</p>	 Egr1_primary	1.6e-1095	10295 UniPROBE	8926	107
	 sci09.v2_Egr1_2580	6.0e-830		4996	107
	 sci09.v1_Egr1_2580	1.6e-761		4565	95
 <p>Max</p>	 Sci09.v2_Max_3863	9.2e-1636	43838 BEEML-PBM	39273	121
	 Sci09.v2_Max_3864	5.3e-1544		34165	115
	 Sci09.v1_Max_3864	2.8e-1338		38834	115
 <p>Nr2f2</p>	 Sci09.v1_Nr2f2_2192	2.4e-492	24408 BEEML-PBM	19010	127
	 Sci09.v2_Nr2f2_2192	2.7e-485		19039	121
	 Nr2f2_primary	2.0e-473		14747	129
 <p>Gata3</p>	 Sci09.v2_Gata3_4964	4.0e-535	26299 BEEML-PBM	23083	141
	 Sci09.v2_Gata3_1024	2.3e-527		23380	125
	 Sci09.v1_Gata3_1024	1.4e-521		23851	125
 <p>Hnf4a</p>	 Hnf4_secondary	7.0e-1159	24405 UniPROBE	15072	129
	 Sci09.v1_Hnf4a_2640	3.9e-670		22073	127
	 Sci09.v2_Hnf4a_2640	9.5e-624		22216	127

Table 3-1: BEEML-PBM and UniPROBE successes.

The 5 TFs all have reasonably well-centred peaks and low p-values. Columns from left to right: TF name and CentriMo distribution, logos for the first three highly enriched motifs, central enrichment p-value, database of the lowest p-value motif and number of ChIP-seq sequences, and for each motif, number of sequences containing a motif and bin width. BEEML-PBM motifs names all start with “sci”.

3.1.2 Central enrichment with less sharply defined peaks

The second set of results includes three TFs (Foxa2, Irf4, and Sp4) that have central enrichment, but the peaks are not well centered and the region of maximal central enrichment is less narrowly centered (about 200bp). But the motifs show statistically significant central enrichment *p*-values. The broad bin width and a lack of well centered distribution could be due to various reasons including poor resolution of the ChIP-seq experiment, indirect binding, cooperative binding (more than one motif may have a central tendency) or ChIP-seq failure. Because there are so many reasons for a CentriMo distribution that differs from a strongly unimodal distribution, further investigation and literature evidence are required. Of the 3 TFs, UniPROBE has the lowest central enrichment *p*-value motif for two motifs (Foxa2 and Irf4) (see table 3-2). CMEA against a wider motif database to determine whether there may be cofactors, cooperative binding or indirect binding can provide evidence that these TFs are involved in cooperative binding. Cofactors can affect the CentriMo distribution in that their motifs can be more centrally enriched than the motif of the ChIP-ed factor. Also indirect binding can affect the distribution because the motifs that bind directly to DNA may appear more enriched than the motif for the ChIP-ed TF (Bailey & Machanick, 2012). Therefore it is important to take the presence of cofactors into consideration because they might lead to false conclusions if their effect is not considered.

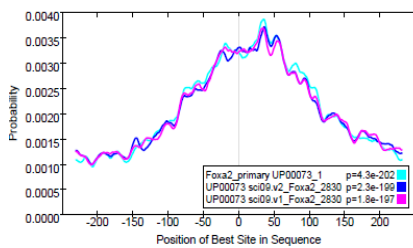



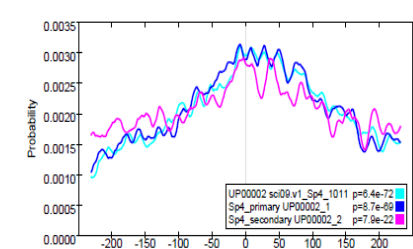



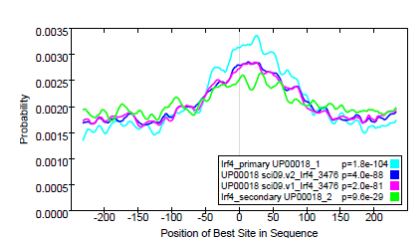



CentriMo distribution	Logos	p-values	Sequences Best DB	Matches	Width
 <p>Foxa2</p>	 <p>Foxa2_primary</p>	4.3e-202	8472 UniPROBE	7099	224
	 <p>Sci09.v2_Foxa2_2380</p>	2.3e-199		7944	199
	 <p>Sci09.v1_Foxa2_2380</p>	1.8e-197		7950	199
 <p>Sp4</p>	 <p>Sci09.v1_Sp4_1011</p>	6.4e-72	6063 BEEML-PBM	5971	237
	 <p>Sp4_primary</p>	8.7e-69		5348	206
	 <p>Sp4_secondary</p>	7.9e-22		5276	176
 <p>Irf4</p>	 <p>Irf4-primary</p>	1.8e-104	19205 UniPROBE	9159	180
	 <p>Sci09.v2_Irf4_3476</p>	4.0e-88		16712	183
	 <p>Sci09.v1_Irf4_3476</p>	2.0e-81		16650	183

Table 3-2: BEEML-PBM and UniPROBE CentriMo.

The distributions show central enrichment with less sharply-defined peaks – for Foxa2 and Irf4 UniPROBE has the lowest p-value motifs (4.3×10^{-202} and 1.8×10^{-104} , respectively). Columns as in Table 3-1.

3.1.3 Central enrichment from only one database

In the third category, only UniPROBE motifs have central enrichment for Pou2f2 and Srf with a narrow region of maximal enrichment 73bp and 51bp, respectively. POU TFs have two different binding domains, octamer (ATGCAAAT) and POU. A POU binding domain can be octamer or homeodomain containing (ATTA) binding sequence (Sharif et al., 2001; Lennard et al., 2006). The Pou2f2 BEEML-PBM motif has insignificant p-value (1.0), implying it is junk, but contains a correct homeodomain binding site ATTA (Table 3-3). The UniPROBE motif is centered and has a significant central enrichment p-value, suggesting the motif represents the true binding site. For Srf, CentriMo finds a p-value and distribution of best sites only for UniPROBE motifs (only Srf primary and secondary) and does not find any motif from the BEEML-PBM database with sites above the score threshold. Further investigation is needed to confirm if the BEEML-PBM database does not contain motifs that represent the true binding motifs of Srf.

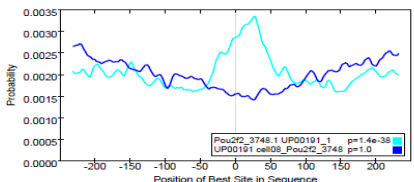


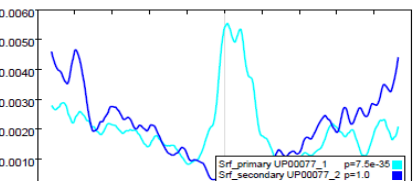


CentriMo distribution	Logos	p-values	Sequences Best DB	Matches	Width
 <p>Pou2f2</p>	 <p>Pou2f2_3748.1</p>	1.4e-38	20589 UniPROBE	8761	73
	 <p>Cell08_Pou2f2</p>	1.0		16100	1
 <p>Srf</p>	 <p>Srf_primary</p>	7.5e-35	43838 UniPROBE	1514	51
	 <p>Srf_secondary</p>	1.0		2106	2

Table 3-3: BEEML-PBM fails.

In each example, only 1 motif (both from UniPROBE) has significant central enrichment. BEEML-PBM motifs all have names starting with “cell” or “sci”. BEEML-PBM Srf motifs are too low in information for a meaningful a logo (column 2). Table columns as in Table 3-1.

3.1.4 No motif from either database with central enrichment

For three TFs (Ets1, Rxra and Tcf3) there is no motif from either database with central enrichment. CentriMo reports the BEEML-PBM and UniPROBE motifs with statistically insignificant central enrichment p-values (Table 3-4). This could be due to poor antibody performance or ChIP-seq experiment failure. Further investigation is needed to be sure. In previous studies it has shown that Ets1 interacts with bHLH-zip proteins especially TFE3 and USF and activate transcription. The authors suggested that Ets1 and TFE3 interacts directly with each other (Tian et al., 1999). That Centrimo fails to find centrally enriched motifs on both databases suggests that these TFs bind cooperatively with other TFs or may be that there are no motifs in the databases that accurately represent the true binding motif.

The Tcf3 distribution is the opposite shape to that expected for all variations, suggesting that binding is more likely away from the center. Also the central enrichment p-value for all the motifs is insignificant (1.0). However, the Tcf3 UniPROBE primary motif (rank third) shows the Tcf binding consensus sequence (TTCAAAGG) suggested by prior studies (Gustavson et al., 2004). It is possible that the inverse distribution is due to the peak calling algorithms failing to localise the binding site. The distribution of Rxra is flat. Rxra has been found to function as a helping protein for retinoic acid receptors (RARs). Therefore it forms a complex with other receptors (Huan et al., 1992). The lack of central enrichment could be due to complex mode of binding as suggested by the literature or CentriMo failure. Further investigation is needed to get a valid reason.

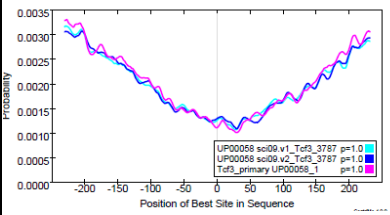



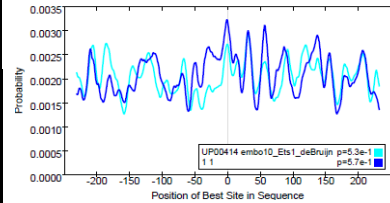
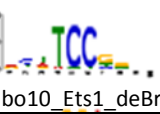

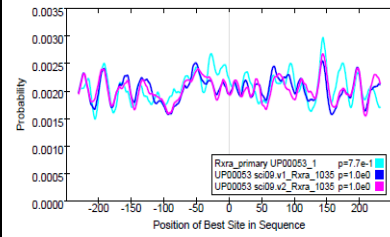



CentriMo distribution	Logos	p-values	Sequences Best DB	Matches	Width
 <p>Tcf3</p>	 Sci09.v1_Tcf3_3787	1.0	32044 None	23991	1
	 Sci09.v2_Tcf3_3787	1.0		22437	1
	 Tcf3_primary	1.0		18318	2
 <p>Ets1</p>	 Embo10_Ets1_deBruijn	0.53	1086 None	966	3
	 Ets1_primary	0.57		568	8
 <p>Rxra</p>	 Rxra_primary	0.77	4411 None	2291	64
	 Sci09.v1_Rxra_1035	1.0		3517	141
	 Sci09.v2_Rxra_1035	1.0		3731	141

Table 3-4: CentriMo fail.

Neither the PBM nor BEEML-PBM motifs are centered. Columns as in Table 3-1.

3.2 CentriMo results visualization using logos

Logos are used to visualise the motifs for the most highly enriched motifs reported by CentriMo, and are shown in column 2 of all tables in chapter 3. The motifs are extracted from UniPROBE and BEEML-PBM databases. A logo combines different important information into a single graphical output. It shows the consensus sequence for the input and the amount of information it contains. It represents the predominance and the relative frequency of each residue at each position and outputs graphical represented bases at each position. The height of each base at each position is proportional to the information contained in that base. Bases represented by bigger symbols are more conserved than those with smaller symbols (Koch et al., 2006). The conserved positions have 2 bits of information. In cases where two of the four bases have an equal chance (50%) of occurring, they contain 1 bit of information. If all four bases appear equally often that position contains no information. If the sample is small it is possible for the information content to be overestimated, therefore a small sample correction has to be employed. The total information content of a motif is directly proportional to its expected frequency of occurring within a random DNA sequence (Patrik, 2006).

A sequence logo provides more information about the binding site compared to the consensus sequence. Useful binding information can be extracted from logos. Each position of the logo output contains a stack of bases (A/T/G/C) indicating how often each residue appears at a particular position. The conserved bases are represented by bigger symbols and less conserved bases by smaller symbols. Sometimes it is not clear which base is preferred at a particular position. In this case the graphical output contains a mixture of bases in one position. The base that appears more often is on the top of the stack. The overall height of the stack in a position is consistent with the information contained in the binding site in that position. The lower height of the base indicates the possibility of alternative bases at that position (Schneider et al., 1990).

Logos indicate that more biological information can be extracted from the well centered motifs (table 3-1, column 2). Interestingly, the most highly enriched motifs in Gata3 ChIP-seq dataset are represented by the WGATAR consensus sequence. This supports the two independent previous studies (Ko et al., 1993; Tallack et al., 2010). Using logos it is clear that Egr1 binds to the G-C rich region DNA sequences as suggested in previous studies (Yan et al., 2000; Kubosaki et al., 2009). Both the BEEML-PBM and PBM approaches report CANNTG motif binding site for Max, and this is consistent with Larsson et al.'s (1997)

findings. The finding that a centrally enriched UniPROBE motif for Irf4 (table 3-2) has a CGAAAC binding site is consistent with a prior publication (Badis et al., 2009). From these results it is observed that centrally enriched motifs appear to be more conserved (bigger letters) compared to the ones that are not centrally enriched as in Tcf3 BEEML-PBM motif examples.

3.3 Independent measure: AME

AME is used as an independent measure for comparison with CentriMo results. AME ranks enriched motifs based on the *enrichment* p-value, compared to CentriMo which ranks motifs based on *central enrichment*. Compared with other tools, AME has an advantage of allowing a very large set of data and is capable of detecting very low levels of enrichment of TF-DNA binding motifs that *ab initio* discovery algorithms fail to detect. Both algorithms achieve higher sensitivity by restricting the motif search to DNA binding motifs that have already been identified. The algorithm calculates the enrichment p-value by Fisher exact test using a shuffled version of the original sequences as a background. As mentioned earlier that CentriMo results are divided into four categories. For each category of results one TF is chosen to run AME. The aim is to observe if CentriMo results are consistent with AME results so as to be certain that CentriMo has found correct motifs.

For all AME results reported here a threshold p-value of 0.001 and a custom database (BEEML-PBM and UniPROBE) for each motif are used. The UniPROBE motif for Egr1 that ranked first by CentriMo now ranks second by AME (fig 3-1). This indicates the importance of considering central motif enrichment. For Irf4, the centrally enriched motifs reported by CentriMo are less enriched using AME (figure 3-2). For Pou2f2, motifs reported by AME have lower p-values compared to those reported by CentriMo (figure 3-3). Both methods rank the octamer first followed by a homeodomain motif.

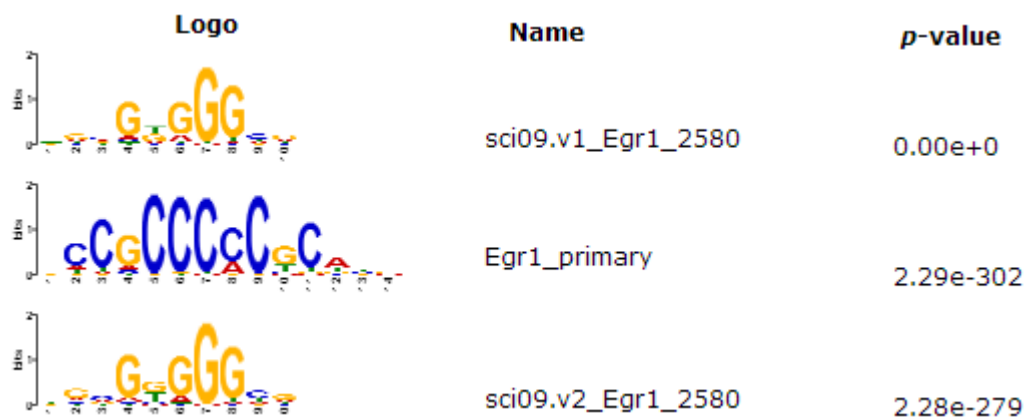


Figure 3-1: AME analysis of Egr1 motifs.

From left, figure shows the logos for the highly enriched motif according to the enrichment p-values, TF name and p-value for each enriched motif.

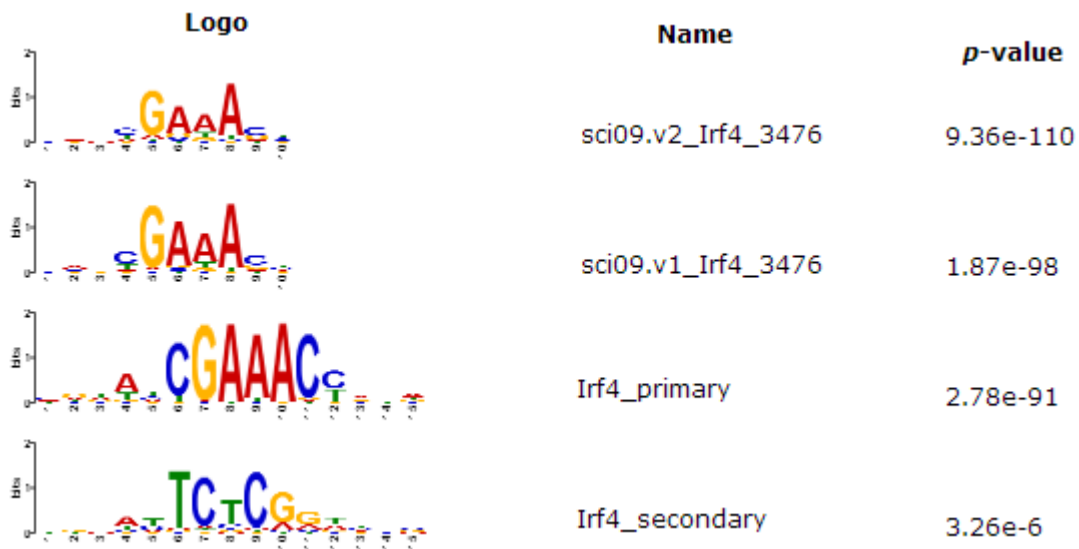


Figure 3-2: AME analysis of Irf4 motifs.

Irf4 analysed using ChIP-seq data from the ENCODE project. Columns as in figure 3- 1.

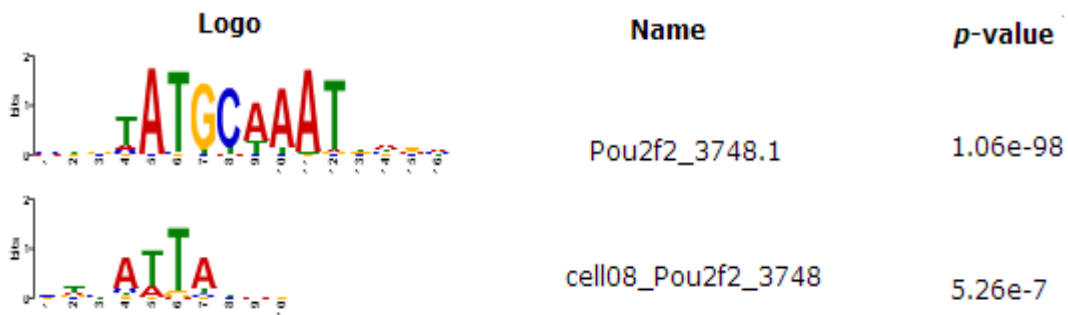
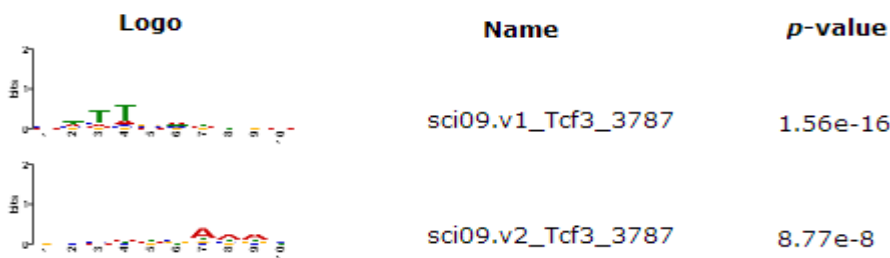


Figure 3-3: AME analysis of Pou2f2 motifs.

Tcf3 motifs are too low in information for a meaningful logo (figure 3-4 (a)). The Tcf3 example is repeated using AME on the entire database. AME ranks E-box motifs as the most enriched motifs in Tcf3 ChIP-seq peak regions (fig 3-4 (b)). The results suggest cooperative or indirect binding of Tcf3. From the literature it is known that Tcf3 binds in cooperation with TAL1, which has an E-box motif (CANNTG) (Kassouf et al., 2010). Interestingly, the highly enriched E-box TFs have statistical significant p-values (0.0). These results show that E-box is a very common motif, therefore, the TF cannot be identified purely based on binding specificity.



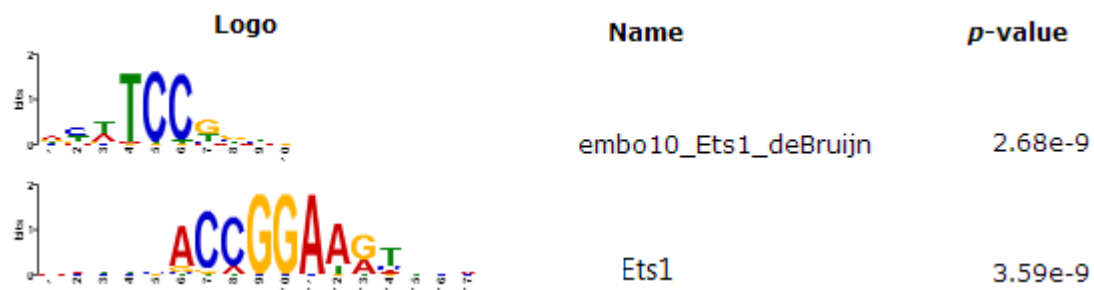
(a) Tcf3 motif on the custom database.



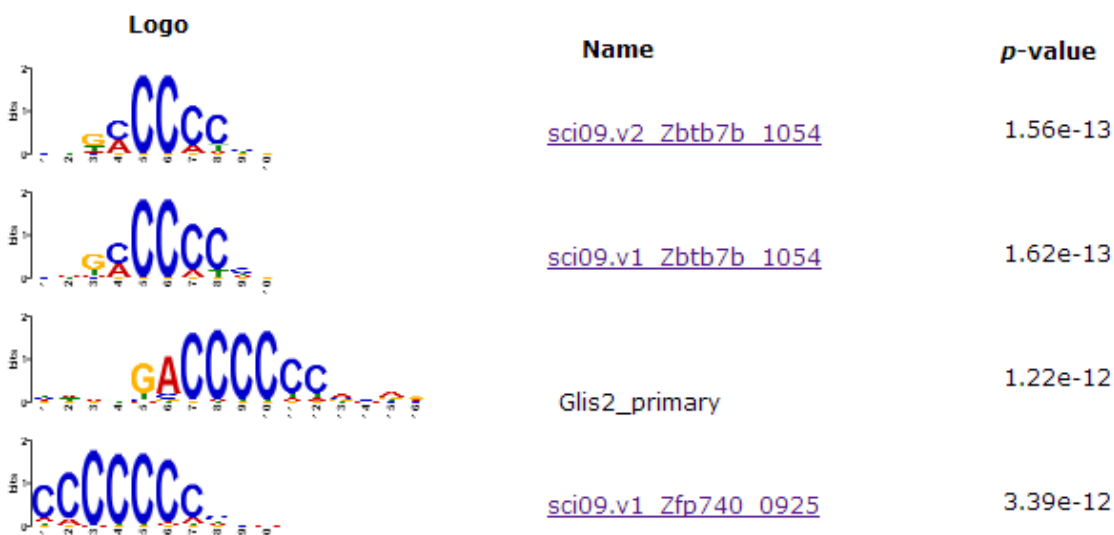
(b) Further investigation of Tcf3 motif on the entire database.

Figure 3-4: AME analysis of Tcf3 motifs.

The p-values for Ets1 motifs are still not that low. The UniPROBE motif contains a known binding site. Ets1 has been found to be associated with different types of human cancers and has been shown to bind to GGA(A/T) motif (Wei et al., 2010). That Ets1 is not over-represented suggest binding in combination with other TFs. Further investigation using AME on the entire database shows that the most enriched motifs are the BEEML-PBM motifs for Zbtb7b. Ets1 motif does not appear near the top of the list of enriched motifs. Zbtb is also involved in oncogenesis (Stogios et al., 2007). The enriched motifs contain mostly C's indicating that there are a lot of similar binding sites in that region. The results suggest that Ets1 bind in combination with other TFs involved in cancer.



(a) Ets1 motif on the custom database.



(b) AME analysis of Ets1 motifs on the entire database

Figure 3-5: AME analysis of Ets1 motifs.

3.4 Comparing central motif enrichment from different cell lines

Transcription is controlled by the interactions between TFs and the genome. However, it is difficult to associate variation in genome sequence with that in TF binding because of the differences in individuals and cell types (Reddy et al., 2012). GM12878, H1-hESC and K562 are the most studied cell lines in the ENCODE project. The K562 cell line is derived from plural effusion in patients with chronic myelogenous leukemia (CML) (Lozzio et al., 1975). GM12878, K562 and H1-hESC are tier 1 cell lines derived from lymphoblastoid cells, CML and embryonic stem cells, respectively (Raney et al., 2011). Tier 1 ENCODE TFs may be higher quality experiments. More information on cell lines is shown in table 2-4. In general, the differences in cell lines could be due to different antibodies used create that particular cell line and also the peak calling algorithm used to declare the TF ChIP-seq regions. And also differences in development stages and disease.

In a total of 13 TFs used in the study, only 9 TFs are available from more than one cell line. The cell lines are shown in table 2-2. CentriMo is used on these 9 TFs and the results are compared with those of the first case study. To confirm if the five clearly centrally enriched motifs (table 3-1) are involved in direct DNA binding, CentriMo is employed to the same TFs using ChIP-seq data from different cell lines, except for Hnf4a which is available only in one cell line. It is observed that the the distribution is now less clearly centered and has a broader bin width compared to the first cell lines used (table 3-5). This suggests that TFs have different binding specificity in the different cell lines. Gata3 results show that it is involved in direct binding when treated with DMSO_002pct compared to the cell line without treatment (table 3-5). Gata3 is up-regulated in T-47D cell line than in A549 cell line. T-47D is a breast cancer cell line (Strom et al., 2004) and Gata3 has been shown to be highly expressed in breast cancer cells (Voduc et al., 2008). The Nr2f2 motifs have the lowest p-values in K562 cell line (table 3-1) compared to HepG2 cell line (table 3-5). Overall, the study shows that the binding specificity of the TF depends on its biological function and the tissue type (normal versus cancer). As an example, cancer associated TFs have higher binding specificity to cancer cell lines compared with normal cell lines.

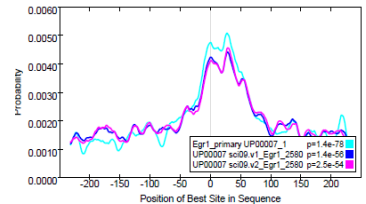



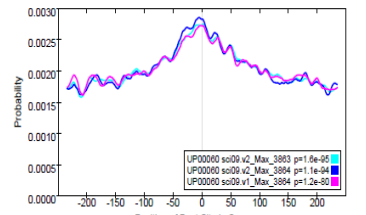


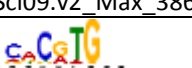
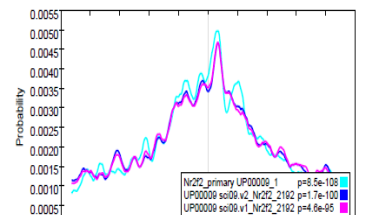


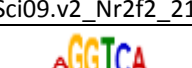
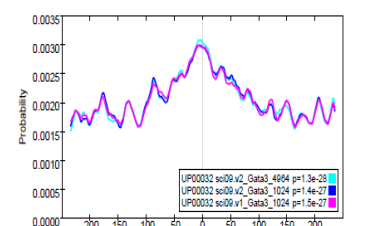



CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Egr1</p>	 <p>Egr1_primary</p>	1.4e-78	2712	2114	113
	 <p>sci09.v1_Egr1_2580</p>	1.4e-56		2576	115
	 <p>sci09.v2_Egr1_2580</p>	2.5e-54		2565	113
 <p>Max</p>	 <p>ci09.v2_Max_3863</p>	1.6e-95	31878	28485	145
	 <p>Sci09.v2_Max_3864</p>	1.1e-94		24321	107
	 <p>Sci09.v1_Max_3864</p>	1.2e-90		28425	145
 <p>Nr2f2</p>	 <p>Nr2f2_primary</p>	8.5e-108	3512	2686	131
	 <p>Sci09.v2_Nr2f2_2192</p>	1.7e-100		3085	133
	 <p>Sci.v1_Nr2f2_2192</p>	4.6e-95		3086	131
 <p>Gata3</p>	 <p>Sci09.v2_Gata3_4964</p>	1.3e-28	5842	5233	125
	 <p>Sci09.v2_Gata3_1024</p>	1.4e-27		5252	179
	 <p>Sci09.v1_Gata3_1024</p>	1.5e-27		5389	179

Table 3-5: TFs variability in different cell lines.

Variability check using CentriMo on the same TFs, but from different cell lines than those in table 2-1. For the cell lines used see table 2-2. Columns from left to right: TF name and CentriMo distribution, logos for the graphed motifs, central enrichment p-values, number of ChIP-seq sequences, and for each motif, number of sequences containing a motif and bin width.

Foxa2, Irf4, and Sp4 TFs have central enrichment, but the peaks are not well centered. Of these 3 TFs, only Foxa2 is available in different cell lines. CentriMo is run on Foxa2 ChIP-seq data from A549 cell line (table 3-6). The distribution is now better centered compared to the first case using the HepG2 cell line. The BEEML-PBM motif that ranked third in the first case is now the most highly centrally enriched motif and the p-value is now low ($4.3e-202$ versus $4.6e-580$). The UniPROBE motif that ranks first now ranks second. The over-expression of Foxa2 in A549 cell line (derived from lung carcinoma) versus HepG2 is interesting and consistent with the previous studies suggested that Foxa2 is over-expressed in human lung carcinoma cell lines (Tang et al., 2011). That the distribution is not well centered in different cell lines could be an artifact of peak calling algorithm or that Foxa2 binds DNA indirectly or in combination with other TFs. Previous studies have shown that Foxa2 binds DNA in combination with other TFs (Wederell et al., 2008). Using TFBS models from TRANSFAC on Foxa2 peaks a prior study's findings show enrichment of HNF4, NHF1, CoupTF, GATA4 and Pax6, which are known to interact with Foxa2 (Odom et al., 2006; Gauthier et al., 2002). The possibility of Foxa2 cooperative binding is validated using CentriMo against a wider motif database in section 3.5.

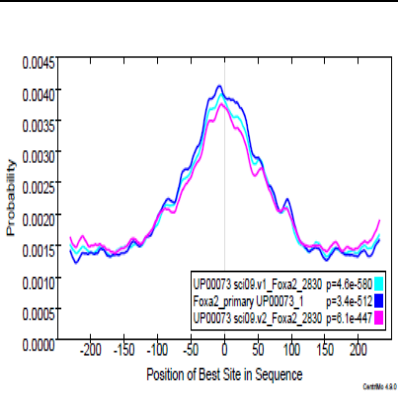
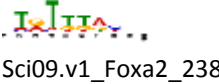


CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Legend: UPO0073 sci09.v1_Foxa2_2380 p=4.6e-580 Foxa2_primary UPO0073_1 p=3.4e-512 UPO0073 sci09.v2_Foxa2_2380 p=6.1e-447</p>	 Sci09.v1_Foxa2_2380	4.6e-580	27361	23900	143
	 Foxa2_primary	3.4e-512		18083	142
	 Sci09.v2_Foxa2_2380	6.1e-447		23908	141

Table 3-6: Foxa2 CentriMo distribution.

Central enrichment of Foxa2 motif from A549 cell line. Table columns as in table 3-5.

In Srf and Pou2f2 results there is no change in the p-values for the Pou2f2 BEEML-PBM and Srf secondary UniPROBE motifs, supporting CentriMo evidence that there is no motif in the BEEML-PBM database for Pou2f2 and Srf with central enrichment. The UniPROBE motifs have the lowest p-values here compared to GM12878 cell line. Overall, the results suggest that the BEEML-PBM binding motif for Pou2f2 is more likely away from the center and there is no motif in the BEEML-PBM database that represents a true binding motif for Srf.

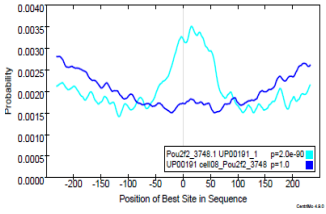


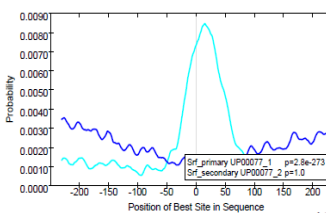


CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Pou2f2</p>	 Pou2f2_3748.1	2.0e-90	22696	10125	119
	 Cell08_Pou2f2_3748	1.0		18588	1
 <p>Srf</p>	 Srf_primary	2.8e-273	3485	2136	97
	 Srf_secondary	1.0		2630	2

Table 3-7: Further investigation of the BEEML-PBM failure.

Validation of CentriMo results using Srf and Pou2f2 ChIP-seq from H1-hESC and GM12891 cell lines, respectively. Table columns as in table 3-5.

Using GM12878 cell line, Ets1 has no centrally enriched motif from either database. Now using A549 cell line the motifs have an improved statistically significant p-value and the distribution is now centered. Ets1 motif from GM12878 cell line has a different binding specificity compared to A549 cell line. The results suggest that Ets1 is up-regulated in cancer cells compared with normal cells. That different cell lines produce very different results illustrates that we cannot just download arbitrary data and process it without thinking about what it represents; maybe that cell line is one where that particular TF is down-regulated. The H1-hESC cell line is more enriched for Rxra compared with GM12878 cell line. Retinoic acids have been found to play a role in the development of vertebrate embryos. This is consistent with CentriMo results that Rxra is up-regulated in embryonic stem cells. Although the p-value is now statistically significant, the window of central enrichment not very narrow (about 160bp) (table 3-8). This could be due to poor ChIP-seq data resolution.

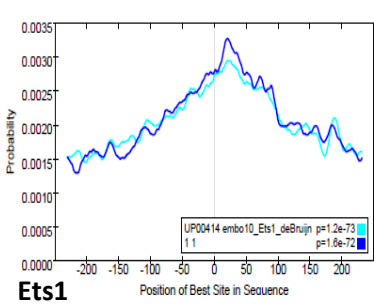


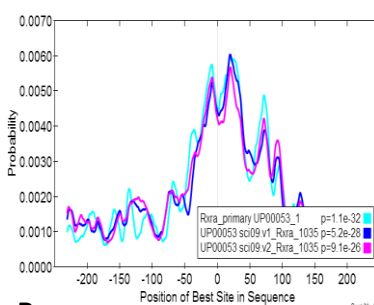



CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Ets1</p>	 <p>Embo10_Ets1_deBruijn</p>	1.2e-73	11401	11102	201
	 <p>Ets1-primary</p>	1.6e-72		7938	184
 <p>Rxra</p>	 <p>Rxra_primary</p>	1.1e-32	596	430	158
	 <p>Sci09.v1_Rxra_1035</p>	5.2e-28		507	159
	 <p>Sci09.v2_Rxra_1035</p>	9.1e-26		515	153

Table 3-8: Further investigation of CentriMo failure.

Validation of CentriMo results using the same TFs, but from different cell lines. Ets1 is from A549 cell line and Rxra is from the H1-hESC cell line. Columns as in Table 3-5.

Comparing motifs found using ChIP-seq data from HAIB lab with that from SYDH lab.

To date, ENCODE has a small sample of TFs that are in both the BEEML-PBM and UniPROBE databases. The SYDH lab contains only 3 TFs that have motifs in both databases. The CentriMo results are compared with those from the HAIB lab. The results using ChIP-seq data obtained from different laboratories agree with each other providing evidence that Max, Hnf4a and Gata3 are centrally enriched (table 3-9). Like in the ChIP-seq data from HAIB lab, these results show that Hnf4a secondary is the most highly enriched motif in Hnf4a peaks. Overall, CentriMo results using ChIP-seq data from HAIB lab and SYDH lab are similar.

CentriMo distribution	Logos	p-values	Sequences	Matches	Width
<p>Max</p>	<p>Sci09.v2_Max_3863</p>	1.2e-294	21281	19397	133
	<p>Sci09.v2_Max_3864</p>	2.5e-285		17153	125
	<p>Max_primary</p>	3.4e-257		11302	133
<p>Hnf4a</p>	<p>Hnf4a_secondary</p>	1.6e-163	6373	4301	117
	<p>Sci09.v1_Hnf4a_2640</p>	1.2e-129		6037	113
	<p>Sci09.v2_Hnf4a_2640</p>	1.1e-122		6032	119
<p>Gata3</p>	<p>Sci09.v2_Gata3_4964</p>	3.0e-289	10287	9033	125
	<p>_Sci09.v2_Gata3_1024</p>	4.3e-281		9094	125
	<p>Sci09.v1_Gata3_1024</p>	6.6e-271		9232	125

Table 3-9: ChIP-seq data from the ENCODE/Stanford/Yale/USC/Harvard Lab.

The cell lines used are shown in table 2-3.

3.5 CentriMo run on combined JASPAR/UniPROBE database

Because the different cell lines do not always agree, further investigation is performed using CentriMo on combined JASPAR/UniPROBE database to see whether the highly centrally enriched motifs correspond to the ChIP-ed TF or are results of co-factors. Also JASPAR provides alternatives, making it possible to check if a previous failure arose because both BEEML-PBM and PBM failed to find a good motif. The output is compared to see whether CentriMo on combined JASPAR/UniPROBE gives similar results to those obtained using CentriMo on combined BEEML-PBM/UniPROBE database, using the same TFs and cell lines. For all CentriMo results reported here, the TF cell lines in table 2-1 are used.

The CentriMo distribution for Egr1, Max, Nr2f2, Gata3 and Hnf4a TFs is unimodal and well centered agreeing with our previous study using BEEML-PBM and UniPROBE databases (see table 3-10). CentriMo reports UniPROBE Gata6 motif and JASPAR GATA1 motif to be centrally enriched (ranks first and fourth, respectively) in GATA3 peaks. This is consistent with a prior study which reports that Gata3 and Gata1 TFs share 92% identity (Ho et al., 1991). GATA3 ranks tenth and thirteenth by JASPAR and UniPROBE, respectively. The sharp peaks in the Egr1 site probability plot also suggest direct DNA binding. Interestingly, the UniPROBE motif for Sp4 ranks third. A previous study by (Wagner et al., 2008) reports that Egr1 shares similar binding sites with Sp family members, but they only report interaction with Sp1.

The peak of the site-probability curve for Max is very well centered indicating direct DNA binding. The JASPAR motifs for Mycn and Myc are most centrally enriched in Max ChIP-seq data and these motifs are very similar to Max. The findings show that using CentriMo and a specified motif database(s) a motif from the same family as the ChIP-ed factor is most likely to be centrally enriched. The central enrichment of Myc peaks in Max ChIP-seq data reported by CentriMo is in agreement with Larsson et al.'s (1997); Orian et al.'s (2003); Hurlin et al.'s (2006) findings. The JASPAR motif of Hnf4a ranks first and the peak is well centered (table 3-10) indicating direct DNA binding. This finding is consistent with the prior studies where the authors used chromatin immunoprecipitation together with promoter microarrays to find the genes occupied by Hnf1a, Hnf4a, and Hnf6 in human liver cells. They concluded that Hnf4a takes part in the regulation of the liver transcriptomes by directly interacting with actively transcribed genes (Odom et al., 2004). From these results it can be concluded that these five TFs have direct binding.

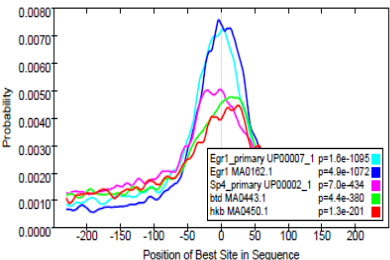



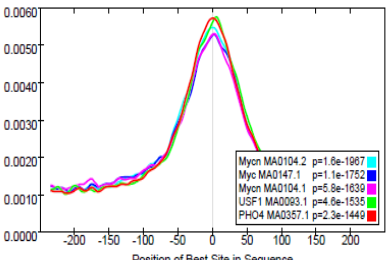



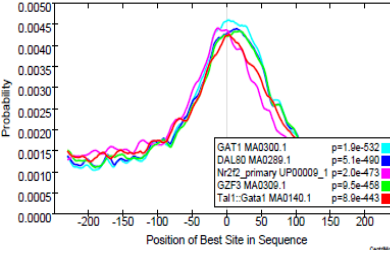



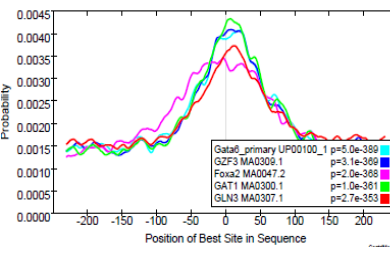



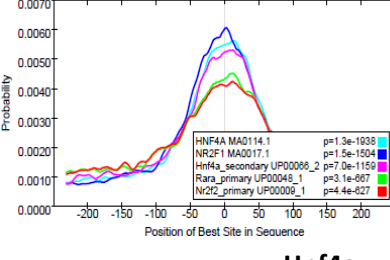



CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Egr1</p>	 Egr1_primary	1.6e-1095	10295	8926	107
	 Egr1MA0162.1	4.9e-1072		7585	106
	 Sp4_primary	7.0e-434		8529	112
 <p>Max</p>	 Mycn_MA0104.2	1.6e-1967	43838	30219	121
	 Myc MA0147.1	1.1e-1752		30766	121
	 Mycn MA0104.2	5.8e-1639		29786	117
 <p>Nr2f2</p>	 GAT1 MA0300.1	1.9e-532	24408	10592	157
	 DAL80 MA0289.1	5.1e-490		11672	156
	 Nr2f2_primary	2.0e-473		14747	129
 <p>Gata3</p>	 Gata6_primary	5.0e-389	26299	16236	106
	 GZF3 MA0309.1	3.1e-369		13944	107
	 Foxa2_MS0047.2	2.0e-368		17271	157
 <p>Hnf4a</p>	 HNF4A_MA0114.1	1.3e-1938	24405	21389	136
	 NR2F1 MA0017.1	1.5e-1504		14731	139
	 Hnf4a_secondary	7.0e-1159		15072	129

Table 3-10: CentriMo run on combined JASPAR/UniPROBE database.

Table shows: the ChIP-ed TF name and the distribution of the first five highly enriched motifs ranked according to their central enrichment p-values. Table columns as in table 3-5.

The JASPAR motif for Foxa2 ranks first (table 3-11), indicating that Foxa2 is highly enriched. But the peak is still not well centered as in the first case using BEEML-PBM and UniPROBE databases. In prior studies Foxa2 has been found to bind together with other TFs to control gene expression. Previous studies have suggested cooperative binding between Foxa2 and HNF4a, HNF6, CREB1, USF1, HNF1a or GATA4 (Odom et al., 2006; Bossard & Zaret, 1998). Using SAGE their findings revealed new interactions, including HNF4, HNF1 and GATA4, as well as novel TFs such as Maz, Mazr and Smad3, all of which are expressed in the adult liver. Also Foxa2 interacts with proteins in close proximity and binds DNA through those TFs. CentriMo ranks motifs of the fox family first and the motifs suggested from the literature are not enriched. The most highly enriched JASPAR motif for Foxa2 has a TTT consensus sequence appearing more than once. This shows the variation in DNA binding specificity. It is possible that the motif matches with other TFs with low information content and might bind to the sequences that are not true binding motifs. Therefore, further investigation is still needed to be certain whether Foxa2 binds cooperatively or not.

For the Irf4 dataset, the highly enriched motifs are the JASPAR motif for GCN4 and the UniPROBE secondary motif for Jundm2 (table 3-11). The observed off-center peaks in Irf4 distribution and the literature evidence confirm that Irf4 binds DNA in combination with Jundm2 secondary, which shows a statistically significant central enrichment p-value ($7.7e-150$) and ranks second in Irf4 ChIP-seq peaks. This supports Whittington et al.,'s finding (2011). The authors used SpaMo and reports that Irf4 binds in combination with Jundm2 secondary. CentriMo ranks UniPROBE motifs for Irf4 and Irf5 seventh and twelfth, respectively. Irf4 is less enriched suggesting cooperative binding with the highly enriched motifs. The GABPA motif (ranks third) is similar to Sfp1 motif (ranks first) reported by AME.

For Sp4, the results suggest that Sp4 binds indirect or in combination with other TFs especially Klf family. The most centrally enriched motifs in Sp4 ChIP-seq peak are Klf7 and Klf4 ranked by UniPROBE and JASPAR, respectively (table 3-11). Literature evidence, which indicated that Klf binding domains are similar to that of Sp family (Waby et al., 2008) and CentriMo results using combined JASPAR/UniPROBE database and on different cell lines suggest that the observed off-center peak distributions are results of cooperative binding.

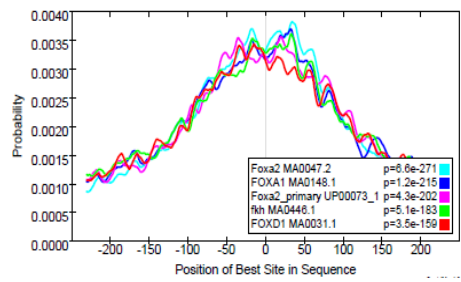



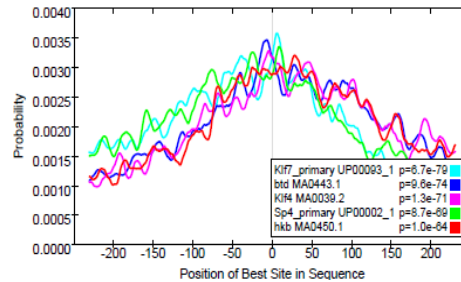



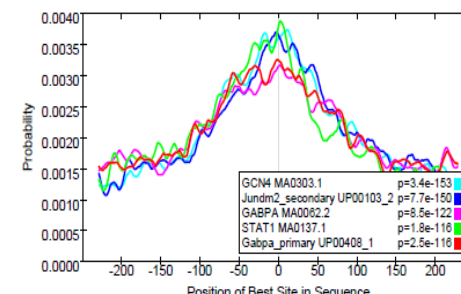



CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Foxa2</p>	 Foxa2_MA0047.2	6.6e-271	8472	7124	203
	 FOXA1_MA0148.1	1.2e-215		7395	220
	 Foxa2_primary	4.3e-202		7099	224
 <p>Sp4</p>	 Klf7_primary	6.7e-79	6063	5494	219
	 btd_MA0443.1	9.6e-74		5367	243
	 Klf4_MA0039.1	1.3e-71		5619	217
 <p>Irf4</p>	 GCN4_MA0303.1	3.4e-153	19205	6928	152
	 Jundm2_secondary	7.7e-150		7592	153
	 GABPA_MA0062.2	8.5e-122		9723	170

Table 3-11: Investigating the mechanism of binding.

Investigating the mechanism of binding (indirect or cooperative) using CentriMo on combined JASPAR/UniPROBE motif database. Columns as in Table 3-5.

For both Pou2f2 and Srf, GABPA motifs from the JASPAR database are highly centrally enriched (table 3-12). None of the enriched motifs look like a POU domain in Pou2f2 data set. CTCF had been suggested to interact with Oct4, which is a POU TF (Lee et al., 2012). The enrichment of CTCF suggests that it interacts with POU TFs. Previous studies have shown that GABP is recruited together with Oct1 (POU TF) at the enhancer region in GABP mediated transcription (Phillips et al., 2000). The enrichment of GABP motifs in Srf data set supports the two independent prior studies, where there authors report that the two TFs bind in close proximity to each other (Wallerman et al., 2009; Valouev et al., 2008). In previous studies, it has been shown that Srf binds in combination with Tcf protein family, which includes Elk1, SAP1 and SAP2 and myocardin proteins (Mk11 and Mk12) (Miano, 2003; Miralles et al., 2003). Also interesting, the JASPAR motif for Elk4 (Tcf family member) ranks third in the Srf ChIP-seq regions indicating that Srf interacts with most members of the Tcf family. That the JASPAR motif for Srf is highly centrally enriched in its dataset shows that the BEEML-PBM method failed or there is no motif in the BEEML-PBM database that accurately represent the true binding motif. The JASPAR motif for Srf is similar to PBM motif, suggesting JASPAR has a better version.

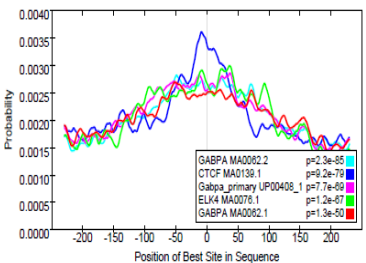



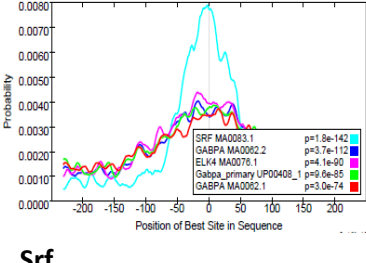



CentriMo distribution	Logos	p-values	Sequences	Matches	Width
 <p>Pou2f2</p>	 GABPA_MA0062.2	2.3e85	20589	11483	206
	 CTCF_MA0139.1	9.2e-79		7610	122
	 Gabpa_primary	7.7e-69		9236	232
 <p>Srf</p>	 SRF_MA0083.1	1.8e-142	5068	869	125
	 GABPA_MA0062.2	3.7e-112		3155	164
	 ELK4_MA0076.1	4.1e-90		1870	172

Table 3-12: Further investigation of BEEML-PBM approach failure.

Table shows the CentriMo output on combined JASPAR/UniPROBE motif database as a further investigation for BEEML-PBM approach failure. Columns as in Table 3-5.

Again, the Ets1 motif does not appear in the set of reported enriched motifs (table 3-13) indicating that there is no motif in the databases that represents the true binding motif. It is possible that the enriched motifs in Ets1 ChIP-seq regions are binding partners of Ets1. In previous studies it has been shown that Ets1 functionally and physically interacts with different TFs and other proteins. Most proteins that directly contact Ets1 bind to the Ets domain (Dittmer, 2003). Other proteins interact with the exon VII domain, the N-terminal part of Ets1 or the activation domain. Some proteins, such as Sp100 or TFE3, interact with several domains of the Ets1 protein (Wasylyk et al., 2002; Tian et al., 1999). A number of TFs had been found to control the transcriptional activity of Ets1 by regulating Ets1 DNA binding affinity. Ets1 has also been shown to increase the DNA binding activity of its partners (Kim et al., 1999).

Both methods fail to find an Rxra motif with central enrichment. The distribution is almost flat. This could be due to the low quality of the ChIP-seq data as results of errors or contamination incorporated during sample preparation or the antibody used is not specific to the TF. The UniPROBE motif for Hnf4a ranks first, suggesting that Rxra binds DNA in combination with Hnf4a. Hnf4a has been suggested as the homodimer receptor for RXRs (retinoid X receptors). The central enrichment of the JASPAR motif for Nr2f1 (table 3-13) is interesting knowing that Rxra does not bind to DNA in the absence of Nr2f2 (Fang et al., 2012). The previous study only suggested binding with Nr2f2 and this study shows that Rxra binds with most members of the nuclear receptor family of TFs. RXRs has been found to form heterodimers with RARs (retinoic acid receptors) and stimulate their binding (Mangelsdorf et al., 1995). RXRs bind to the human response element. Interestingly, the UniPROBE motif for Rara is enriched supporting the previous evidence of RXR/RAR heterodimer formation. The enrichment of RORA and other motifs shows that Rxra binds in combination with any nuclear receptor family of TFs. Therefore lack distinct peak could be due to cooperative binding.

For Tcf3, CentriMo finds that E-box motifs are generally centrally enriched. The first ten motifs in order of CentriMo *p*-value are E-boxes (the first three motifs are shown in table 3-13). This supports our previous finding using AME which reported enrichment of Ascl2 and Tcf2a which are also E-box TFs. Tcf3 example demonstrates how CentriMo “failures” can be informative. Had we not known that Tcf3 binds in cooperation with a TF with an E-box motif, CentriMo results would have provided a reason to investigate this possibility. The results show that the motif for a ChIP-ed TF or a member of its family appears to have most statistical significant central enrichment *p*-value.

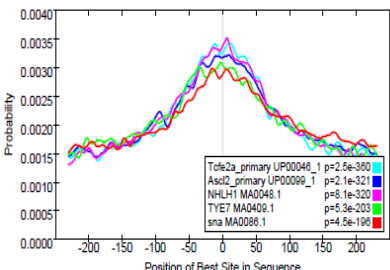



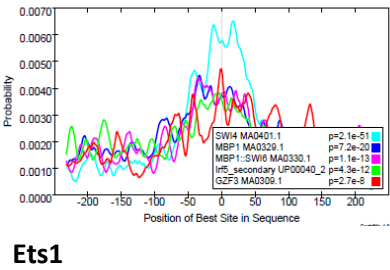


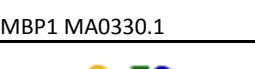
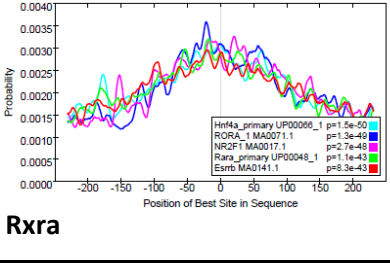

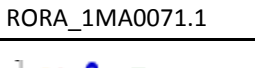
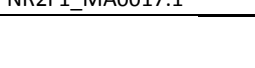
CentriMo distribution	Logos	p-values	Sequences	Width
 <p>Tcf3</p>	 Tcf2a_primary	2.5e-360	32044	128
	 Ascl2_primary	2.1e-321		150
	 NHLH1_MA0048.1	8.1e-320		165
 <p>Ets1</p>	 SWI4_MA0401.1	2.10E-51	1086	129
	 MBP1_MA0329.1	7.20E-20		194
	 MBP1_MA0330.1	1.10E-13		84
 <p>Rxra</p>	 Hnf4a-primary	1.50E-50	8357	214
	 RORA_1MA0071.1	1.30E-49		151
	 NR2F1_MA0017.1	2.70E-48		215

Table 3-13: Investigation on enriched motifs with off-centred peaks.

Further investigation on CentriMo results where there is no motif from BEEML-PBM and UniPROBE databases with central enrichment. Columns as in table 3-5.

3.6 Further investigation of CentriMo results using SpaMo

To explore if the lack of a distinct peak in the CentriMo distribution for Foxa2, Sp4 and Irf4 results from complex binding or binding in close proximity to another TF, spaced motif analysis (SpaMo) is used. SpaMo detects enriched spacings and identifies the interactions between a given motif and TFs that form complexes with the given TF. The algorithm tries to find enriched motif spacing patterns by searching for the strongest primary motif binding site for the given TF then searches the neighbouring sites for secondary motif binding sites and then calculates the probability of each spacing from the primary TFs (Whittington et al., 2011). SpaMo takes in a set of DNA sequences in FASTA format for a given TF, a primary motif represented as a PWM (Stormo, 2000) and a database of secondary motifs. For primary motif input, we use the most highly enriched motif ranked by CentriMo and use BEEML-PBM and UniPROBE databases as secondary motif databases. SpaMo is also used to validate if the centrally enriched motifs reported by CentriMo represent direct binding, we expect *not* to find a single conserved spacing if these TFs bind directly to DNA.

The CentriMo distribution for Foxa2, Irf4 and Sp4 suggests that these TFs bind in combination or in close proximity with other TFs. Investigating this possibility using SpaMo for Irf4, the results show one conserved spacing downstream on the same strand with a p-value of $9.1e-6$ (fig. 3-6). The larger gap of 18bp between primary and secondary motifs suggests the formation of multi-protein-DNA interactions. The results suggest that Irf4 is likely to bind to DNA as a complex with Gabpa. This is in agreement with CentriMo results using JASPAR/UniPROBE database in which Gabpa motifs are statistically enriched in Irf4 ChIP-seq regions. The JASPAR motif for Gabpa ranks third and UniPROBE motif ranks fifth. SpaMo also reports Jundm2 from BEEML-PBM database as a secondary motif with similar spacing and this is also supported by literature evidence, which suggested complex binding between Irf4 and Jundm2 (Whittington et al., 2011). SpaMo found a lot of E-box TFs as secondary motifs in Irf4 ChIP-seq data suggesting cooperative binding or binding in close proximity to E-box motifs, see appendix B.

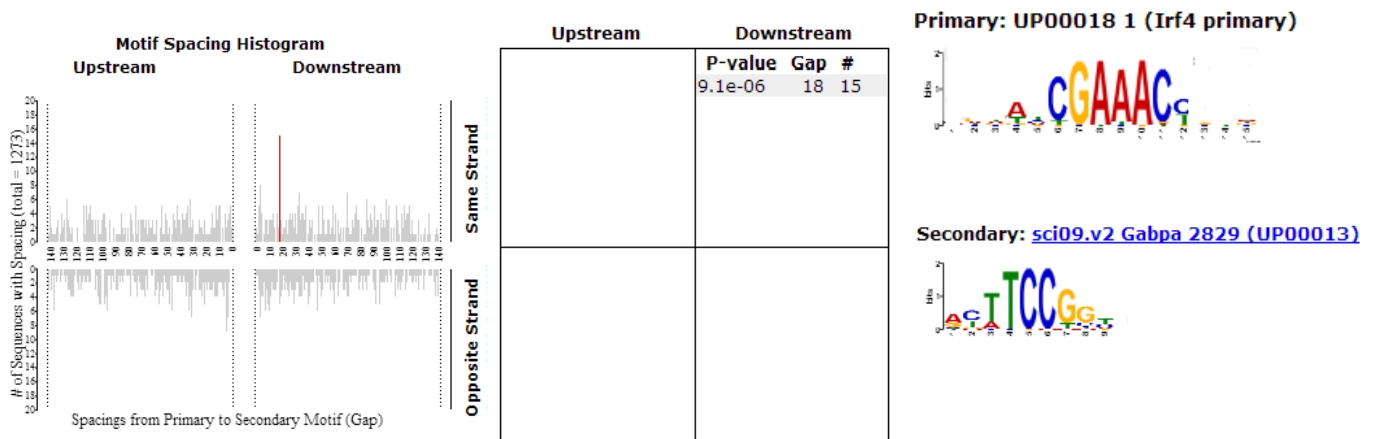
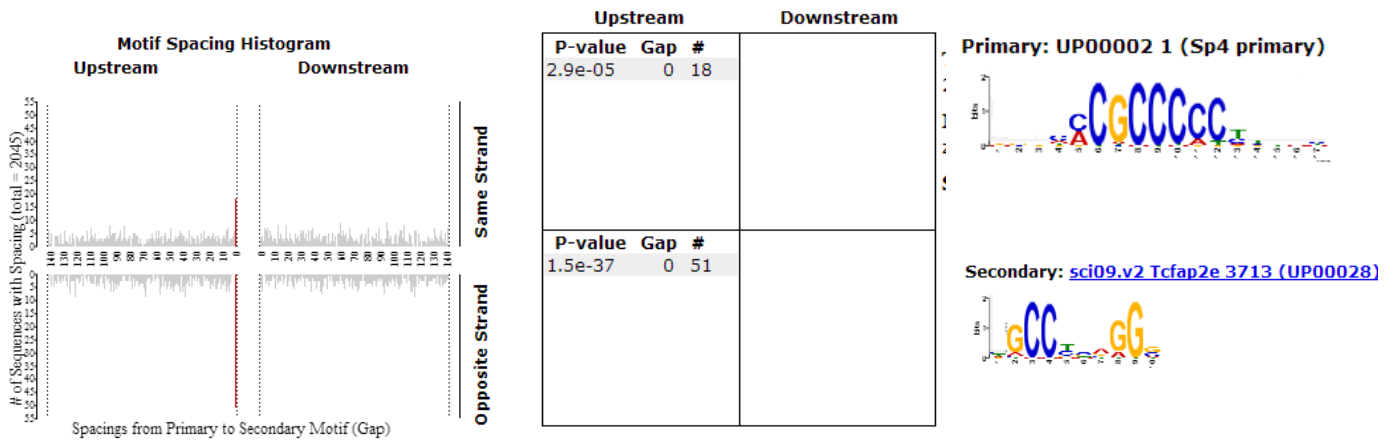


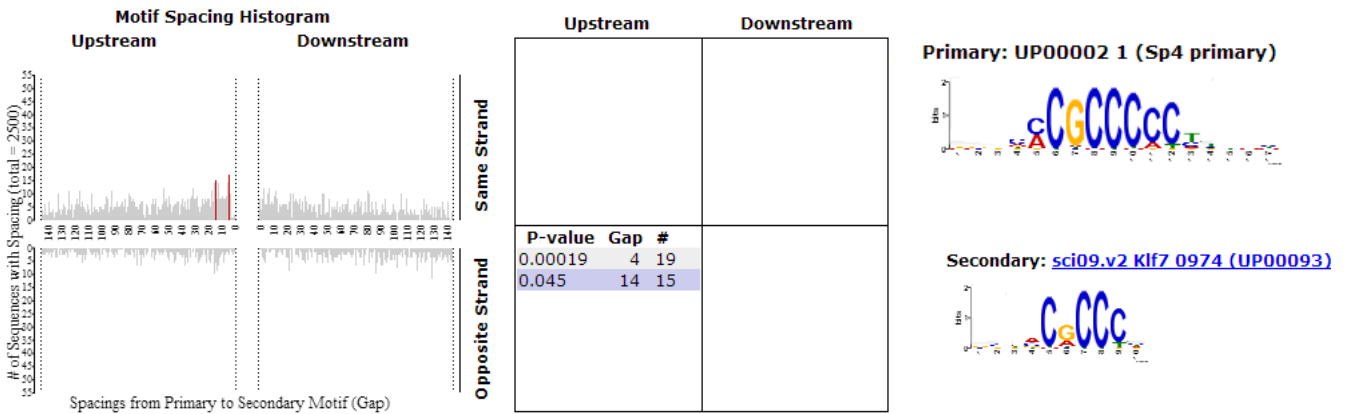
Figure 3-6: Evidence of Irf4 binding partners during DNA binding.

Shown from the left is SpaMo histogram, p-values histogram and on the right are sequence logos for the primary motif (top) and secondary motif (bottom). The red line on the histogram represents statistically enriched spacing.

Figure 3-7a shows that two spacings are enriched for Sp4 on different strands. The positions for the enriched spacings are upstream on the same strand (p-value = 2.9e-05) and upstream on the opposite strand (p-value = 1.5e-37). The spacings enrichment observed on different quadrants suggests that the two TFs are more likely to bind in close proximity to each other than as a complex. The Sp4 motif logo shows that it binds in a G-C rich region and the secondary motifs appear as truncated versions of the primary, indicating the "secondary" (Klf7) is not really a binding site for another motif but just an indication that there are a lot of similar sites in that region. The results also show two enriched spacing very close to each other in one quadrant (upstream on the same strand) with p-values of 0.0019 and 0.045 (fig. 3-7b). Also the p-values are not that low indicating that the instances of that spacing is not very high.



(a) Sp4 motif with two enriched spacings on *different* strands.



(b) Sp4 motif with two enriched spacings on the *same* strand.

Figure 3-7: Searching for evidence of Sp4 complex formation during DNA binding.

Figure 3-8 shows a complex binding formation between Foxa2 and Obox1. The histogram shows one significant motif spacing in one quadrant with a p-value of 7.6e-05. The small gap of 3bp between Foxa2 primary and Obox1 secondary motifs may indicate dimer formation. A complex binding with Hoxa1 and Rara is also suggested. The detected enriched spacing logo shows that primary motif is different from the secondary motif, suggesting complex formation with one or more distinct TFs. These results indicate that the lack of distinct CentriMo distribution peak may be a result of complex binding or binding in close proximity. Foxa2 has been suggested to bind DNA in combination with other TFs, but combination with Obox1 reported by SpaMo had not been suggested.

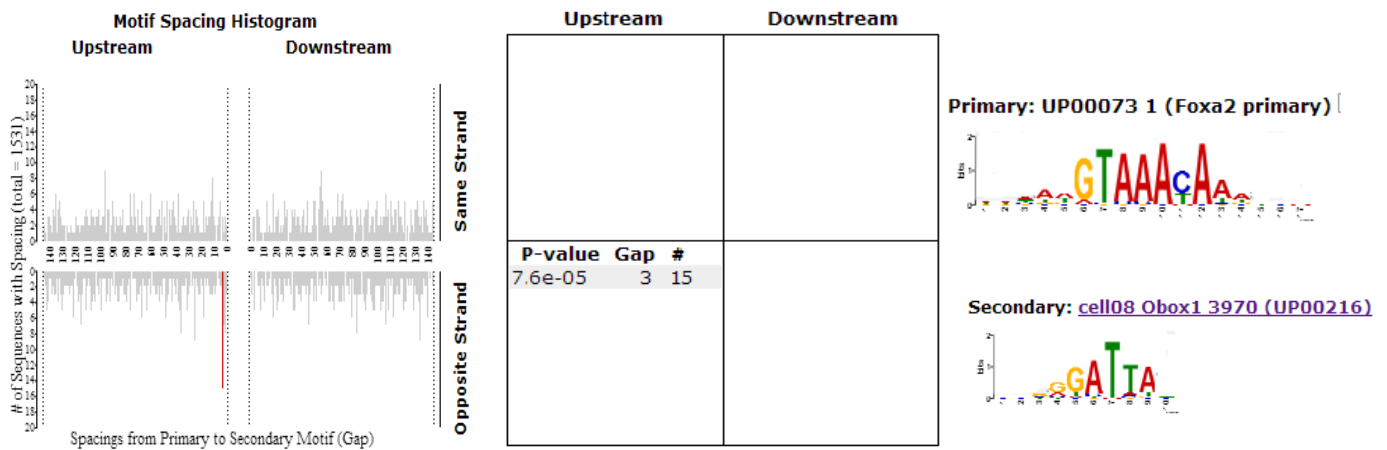
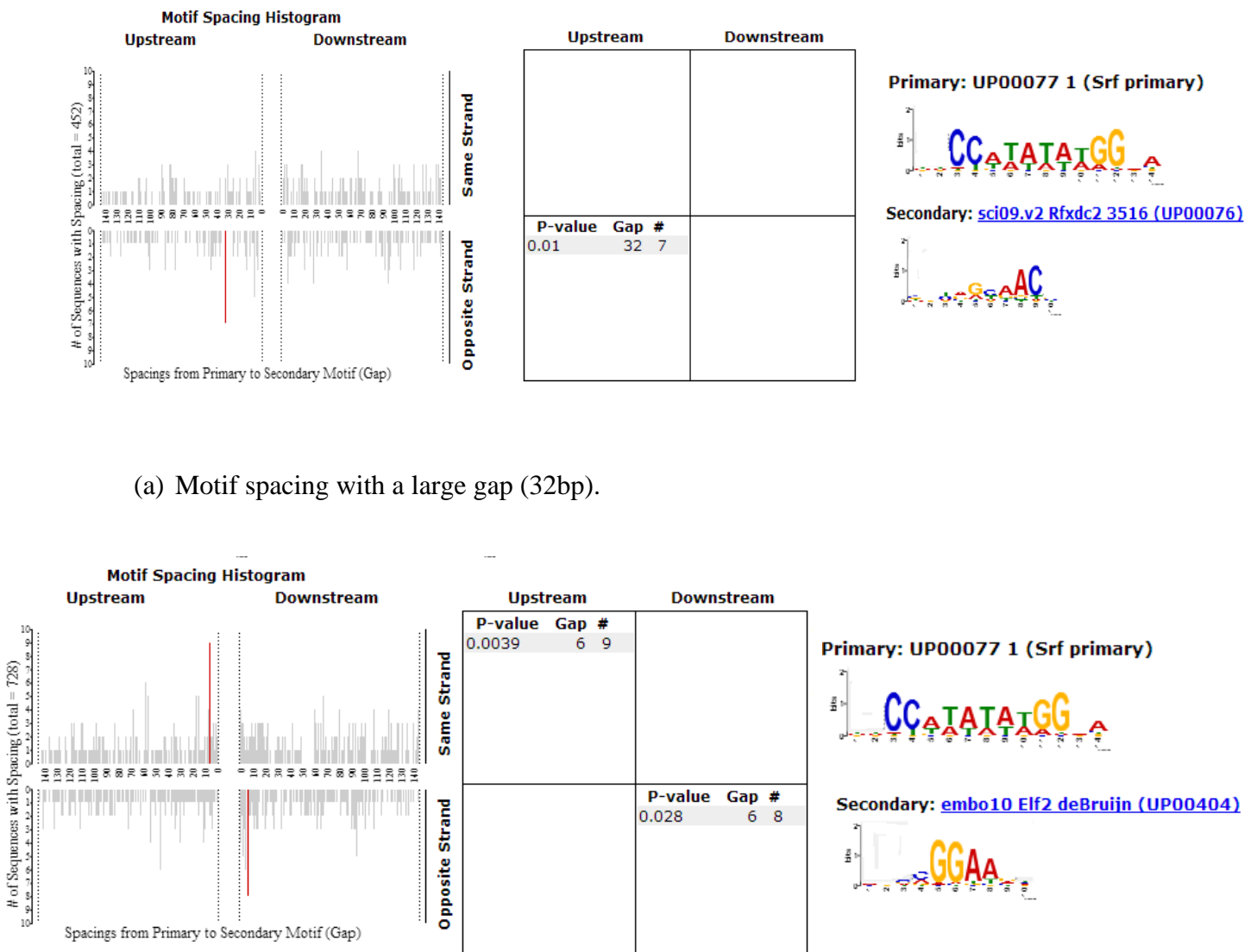


Figure 3-8: Searching for evidence of Foxa2 complex formation during DNA binding.

For Srf, the predicted binding partner is Rfxdc2 BEEML-PBM motif. A secondary motif has very low information for a meaningful logo (fig. 3-9a). The results suggest that there is no motif in the BEEML-PBM database that represents the true binding motif for Srf. A gap between primary and secondary motif is too large (32bp) suggesting that Srf is involved in multi-protein/DNA complex formation. This is in agreement with prior studies which suggested that Srf is involved in SAP-1/SRF/c-fos SRE DNA complex binding. Srf is known to form a complex with Ets proteins like SAP-1 and Elk1 (Mo et al., 2001). Interestingly, using CentriMo on combined JASPAR/UniPROBE Elk4 motif is centrally enriched on Srf ChIP-seq regions suggesting indirect or complex binding. SpaMo also detected enriched

spacing with Erg and Rax. Because Srf is almost a palindrome the two enriched spacings on opposite strands when Elf2 is a secondary motif can be treated as one spacing (fig. 3-9b) since the second spacing could be the result of the reverse strand binding.



(b) Two enriched spacings on different strands and position.

Figure 3-9: Searching for evidence of Srf complex formation during DNA binding.

SpaMo reports several enriched spacings for Egr1 on different orientations suggesting it does not bind to DNA in combination with other TFs. Egr1 binds in a G-C rich region and the Sp4 secondary (fig. 3-10) indicates that there are a lot of similar binding sites in that region. A prior study has reported that Egr1 binds to similar binding sites to the Sp binding site and physically interact with Sp1 to facilitate maximal IL-2R β promoter activity. They also report that Sp1, Sp3 and Egr1 bind to the -170 to -139 enhancer region of the human IL-2R β promoter (Lin et al., 1997). SpaMo results do not support the physical interaction between Egr1 and Sp4. Wagner et al., (2008) also suggested that Egr1 and Sp1 share a similar binding site. Also there are several spacings with low p-value for that "secondary" motif indicating that the TFs do not bind DNA cooperatively. This supports the CentriMo evidence that Egr1 binds directly to DNA.

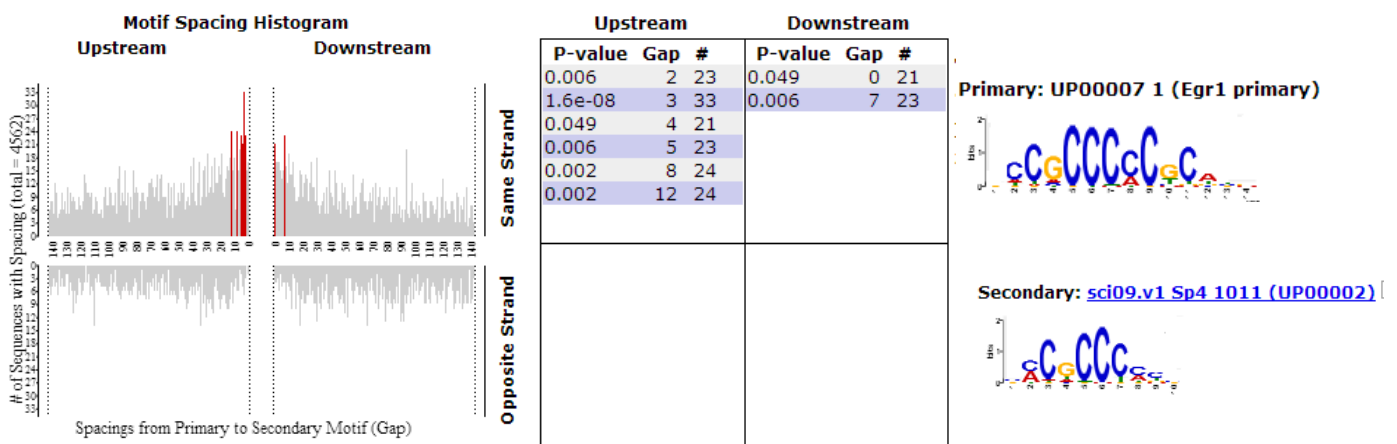


Figure 3-10: Searching for evidence of Egr1 complex formation during DNA binding.

From the category of the CentriMo results where there is no motif from either database with central enrichment Ets1 is chosen and used to run SpaMo. The aim is to investigate whether CentriMo has failed to find a correct motif or the ChIP-seq data used is of low quality therefore the correct motif is not centrally enriched. A single observed spacing in one quadrant (fig. 3-11) suggests complex binding with Atf1 TF. A recent study shows that Ets1 DNA binding depends on partner proteins. These proteins bind to adjacent sequences of Ets binding site (EBS). The authors solved the crystal structure of Ets1 and found a homodimer. The authors also performed DNA binding modelling and report that the Ets1 dimer can bind to two antiparallel pieces of DNA. The Ets1 DNA binding dimer resulted in the formation of additional intermolecular protein-DNA interactions, indicating the complex formation is cooperative (Babayeva et al., 2012). SpaMo results suggest cooperative binding with Atf1. CentriMo results suggest that the mechanism of Ets1 binding is indirect or cooperative. Also AME suggests cooperative binding with TFs involved in cancer.

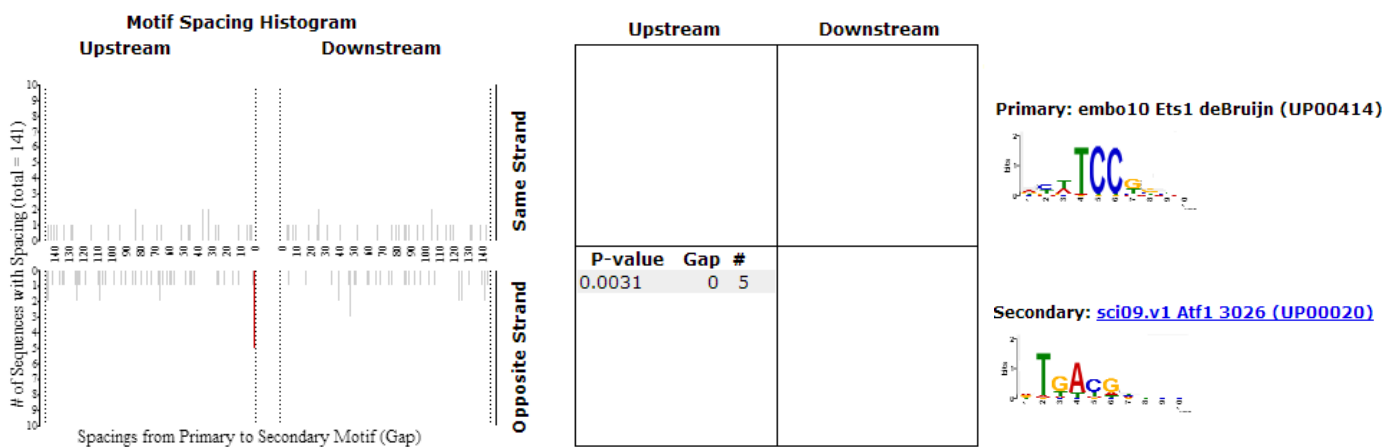
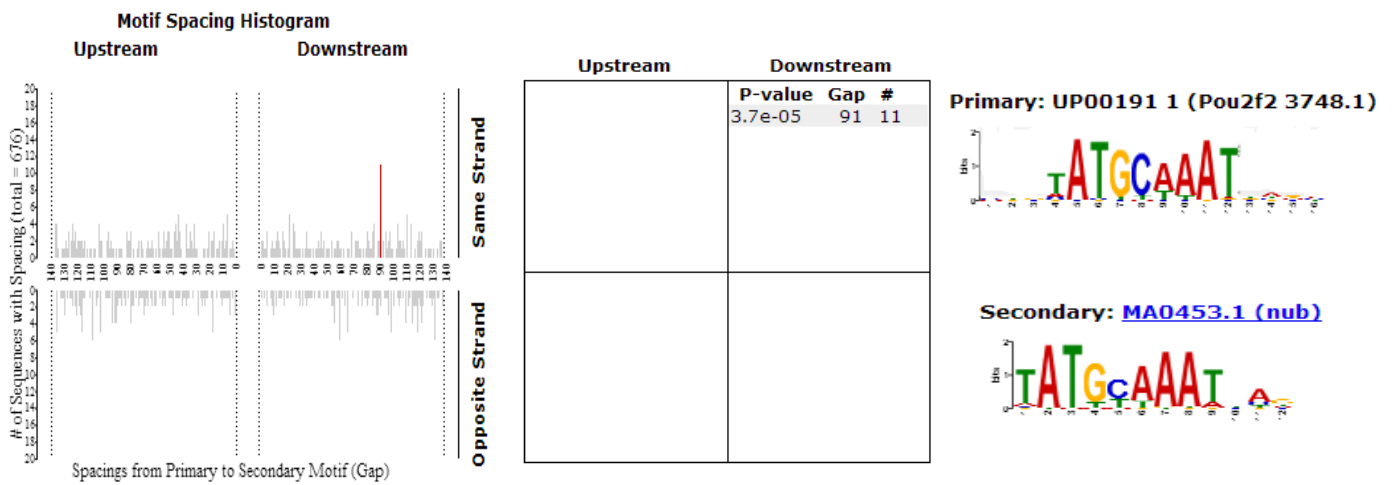
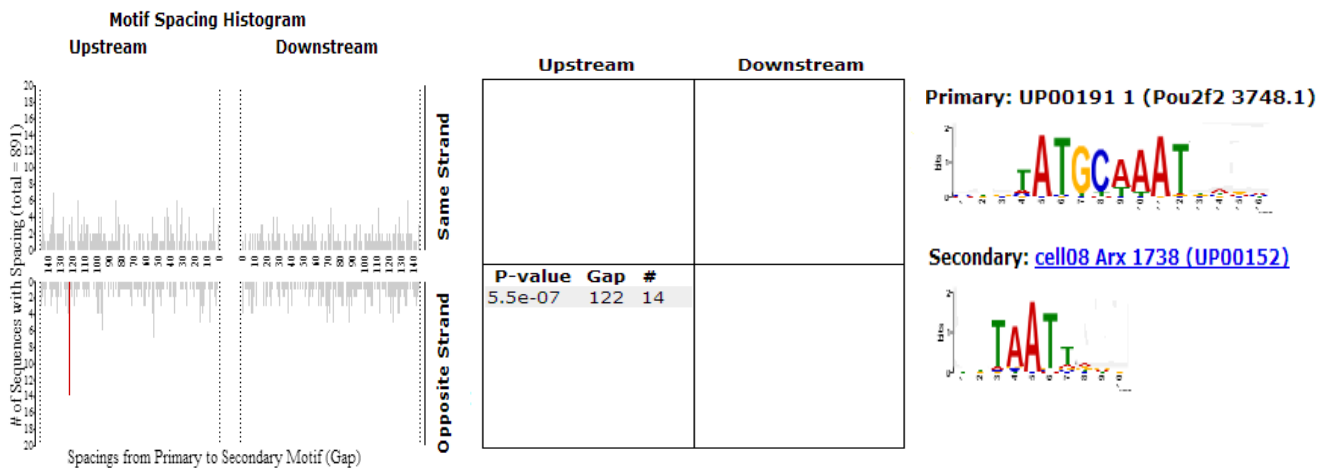


Figure 3-11: Searching for evidence of Ets1 complex formation during DNA binding.

Using AME and CentriMo on combined JASPAR/UniPROBE none of the enriched motifs look like a POU motif. Further investigation using SpaMo shows that Pou2f2 binds in combination with TFs containing ATTTGCAT (octamer) motif (fig. 3-12a) or homeodomain TFs containing ATTA motif (fig. 3-12b). SpaMo also found Cxr, Hoxd8, NK7.1, Phox2b, Pou2f1, eve and Otx2 to have significant spacing in one or different quadrants indicating complex binding or binding nearby these TFs. These TFs are members of the homeo family.



(a) POU-octamer interaction.



(b) Pou-homeodomain interaction.

Figure 3-12: SpaMo analysis of Pou2f2 motifs.

CHAPTER 4: DISCUSSION

CentriMo can be used as a CMEA tool to compare motifs for a given TF and can identify a better method for creating motifs from PBM. A low CentriMo p -value and low width of maximum enrichment (about 100bp) are strong evidence that the motif in consideration is the true binding motif. When the CentriMo distribution has a clear central peak, a very low p -value (sometimes of the order of 10^{-1000} or less) and a minimal p -value in a very narrow window, there is a clear case of binding by a single factor. Where the peak of the site-probability plot is not well centered, it becomes less clear what has happened. Possibilities include, (i) problems with the ChIP-seq experiment that has failed to localise the binding sites, (ii) problems with peak calling failing to localise the binding site, (iii) the specific tissue type or cell line may have unusual binding specificity, (iv) no motif in the database(s) that accurately represents the true binding motif, (v) more complex instances of binding where identifying a single site related to a motif may be problematic (Wang et al., 2011), resulting in peaks that are not centered. This could be due to indirect binding; where there is no motif corresponding to the TF in question is well-centered on the binding regions or cooperative binding, where the TF binds in combination with other TFs and their motifs become highly enriched. Also it could be that the CentriMo algorithm may have failed.

Using CentriMo, it is possible to compare central motif enrichment from two different databases directly based on the above mentioned features. CMEA is applied to ChIP-seq data together with databases of known motifs. In this study the quality of motifs from the BEEML-PBM and UniPROBE databases is compared. CentriMo provides the evidence that Egr1, Max, Nr2f2, Gata3 and Hnf4a bind directly to DNA. Using CentriMo on combined BEEML-PBM/UniPROBE database (containing 980 motifs) and also on combined JASPAR/UniPROBE (containing 862 motifs) the distribution for these TFs is centrally enriched with low p -values. CentriMo provides evidence that Irf4, Sp4 and Foxa2 bind DNA in combination with other TFs. The CMEA findings are validated using SpaMo, which infers physical interactions between a given TF and nearby TFs and literature evidence. The agreement in the SpaMo, AME and CMEA analysis shows that the binding of these TFs is cooperative. CentriMo shares some similarity with SpaMo. SpaMo predicts the enriched spacings between the DNA binding site of the ChIP-ed TF and the co-factor. CentriMo and SpaMo use a binomial test to detect enriched spacings.

CentriMo also shows that TFs from different cell lines have different binding specificity depending on the tissue type and the stage of development. As an example, Ets1 from GM12878 and A549 cell lines shows a huge difference, from a flat distribution to a centered peak (p-value of 0.53 versus $1.2e-73$). The GM12878 cell line is derived from the human lymphoblastoid cells in the blood tissue infected with Epstein-Barr virus while A549 (first developed by D.J Giard in 1972) is the epithelial cell line derived from a lung carcinoma tissue. Ets1 has been found to play a role in cellular differentiation in hematopoietic cells and to facilitate invasive behavior of epithelial cancer cells (Dittmer, 2003). The previous studies have shown that Ets1 is over-expressed in ovarian cancer cells indicating it up-regulates key enzymes involved in antioxidant defense. Their overall finding was that, Ets1 plays an important role in response to oxidative stress in ovarian cancer cells (Verschoor et al., 2010). Therefore, Ets1 has a higher binding affinity in A549 cell line compared with GM12878 cell line, indicating it is where the TF is up-regulated.

Rxra also shows a big difference in central enrichment, a p-value of 0.77 versus $1.1e-32$ in GM12878 and H1-hESC cell lines, respectively. The H1-hESC cell line is derived from the human embryonic stem cells developed from the in vitro fertilised egg. This cell line was developed due to lack of hepatic tissue useful for the treatment of liver disease and drug discovery (Zamule et al., 2011). The retinoids (Rxra) are needed for normal growth and has been found to be involved in various physiological processes such as vision, reproduction and cell differentiation (Abdel-Bakky et al., 2011). Rxra has a strong binding affinity in H1-hESC cell line and is involved in reproduction.

The T-47D cell line is more enriched for Gata3 compared to A549 cell line ($1.3e-28$ versus $4.0e-535$). T-47D is the epithelial cell line derived from a mammary ductal carcinoma in breast tissue. The results suggest that Gata3 is associated with breast cancer than lung cancer. The enrichment of Gata3 in T-47D is interesting knowing that it plays a crucial role in regulation of epithelial cell phenotype in various stages of development of mammary glands (Naylor et al., 2007). It has been shown that Gata3 is expressed by epithelium in mammary glands (Visvader et al., 2003). CentriMo findings agree with the prior studies that expression of Gata3 is associated with breast cancer. It has been suggested that Gata3 maintains a differentiated state in breast epithelial cells. Also the abnormal expression has been found in breast, pancreatic and cervical cancers (Usary et al., 2004; Gulbinas et al., 2006; Steenbergen et al., 2002), respectively.

For Egr1, the distribution is well centered when the ChIP-seq data from GM12878 is used, but when using H1-hESC cell line the p-value becomes less significant ($1.6e-1095$ versus $1.4e-78$). Because Egr1 plays a role in control of homeostasis of hematopoietic cells (Min et al., 2008) and that GM12878 cell line is derived from blood tissue is consistent with CentriMo findings that Egr1 is highly enriched or up-regulated in GM12878 than in H1-hESC cell line.

Overall, the results show that a particular TF may be down-regulated or up-regulated in a cell line depending on the tissue type used to create that cell line and also on the biological function of the TF. CentriMo is developed recently and not much prior work has been done. The discovery of cell line differences, which has not been done before in any CentriMo study, is interesting and provides reasons to study this further. This is treated as a case for future investigation because currently, the ENCODE project has a small number of TFs that are in the UniPROBE and hence BEEML-PBM database. Also CentriMo results show that the motif for a ChIP-ed TF or a member of its family may appear to be highly enriched. As in Max example, the most highly enriched motifs are Myc and Mycn motifs, which are very similar to Max.

Various MEA tools have been developed, but they do not consider central enrichment of a motif. MEA determines whether motifs appear more often than expected by chance in a given TF ChIP-seq peak regions. CentriMo is able to infer direct, indirect and cooperative binding with some of the predictions supported by evidence from prior publications. CentriMo is useful in cases of poor ChIP-seq data quality, it can point to a problem. If there is a TF expected to have a clear binding specificity but CentriMo has a high p-value or other indications of less specific binding, the possibility that the ChIP-seq experiment has failed is considered. The study shows that there is no motif in the BEEML-PBM database that accurately represents the true binding motif of Pou2f2 and Srf while UniPROBE has the binding motifs corresponding to these TFs. In cases where CentriMo does not give clear central enrichment, using SpaMo on combined BEEML-PBM/UniPROBE database shows that the off-center peaks or the lack of distinct peak in CentriMo distribution result from cooperative binding or binding in close proximity to other TFs. SpaMo results show that Pou2f2 binds in combination or nearby TFs containing octamer or homeodomain binding sites. It is also found that both BEEML-PBM and UniPROBE databases have no binding

motif for Tcf3 with central enrichment. The lack of central enrichment is possible if the binding is most likely away from the center an artifact of the peak calling algorithm used.

Motifs involved in direct, indirect or cooperative binding are not clearly visible if central enrichment is not taken into consideration. AME ranks the motifs differently from CentriMo. As an example, the UniPROBE motif for Egr1 ranks first by CMEA and has significant p-value. However, using AME on the same TF used for CentriMo the same motif is now less enriched with less significant p-value compared to CentriMo p-value. Centrally enriched motifs reported by CentriMo are less enriched in AME. Using AME there is no improvement in the quality of motifs that show poor central enrichment indicating CMEA is a better method than standard MEA tools. However, AME can give an indication of an over-represented TF or TFs that are not well centred, and that is useful for picking up cofactors or if CentriMo fails to find a well centred TF for a reason like indirect binding, AME may find enrichment of TFs to which the TF of interest binds. Table 4-1 below summarises the important findings and some known properties for each TF used in this study.

TF	Database	Cofactors	Specificity in cancer	Comments
Egr1	UniPROBE	no	no	Up-regulated in normal cells
Hnf4a	UniPROBE	no	yes	Associated with liver cancer
Max	BEEML-PBM	no	yes	Associated with leukemia
Nr2f2	BEEML-PBM	Not clear	yes	Centrally enriched in lung cancer cells
Gata3	BEEML-PBM	no	yes	Highly associated with breast cancer
Foxa2	UniPROBE	yes	yes	Enriched in lung cancer cells
Irf4	UniPROBE	yes	No evidence	Binds mostly with E-box TFs
Sp4	BEEML-PBM	yes	No evidence	Binds mostly with Klf family members
Pou2f2	UniPROBE	yes	No evidence	Binds with octamer or homeodomain
Srf	UniPROBE	yes	No evidence	Binds in combination with Ets proteins
Ets1	BEEML-PBM	yes	yes	Up-regulated in lung cancer cell line
Rxra	UniPROBE	yes	no	Up-regulated in normal cells
Tcf3	None	yes	No evidence	E-box is the common motif

Table 4-1: Summary of each TF.

Table columns from left to right show: The TF name, database of the lowest p-value, TF involved in cooperative binding, up-regulation of TF in normal versus cancer cells and comments.

CHAPTER 5: CONCLUSION

The sample (13 TFs) is too small to draw general conclusions. Out of the total of 13 TFs, BEEML-PBM twice has not had any central enrichment when the best PBM motif does. In total the BEEML-PBM approach reports five motifs and UniPROBE approach finds seven motifs with better central enrichment. On the other hand, across all variations, the number of examples where PBM is better is not high enough to convince us that it is overall the better approach. There are some cases where multiple motifs found by the BEEML-PBM are of highly variable quality. A difficulty with a comprehensive study is that the match between available ChIP-seq data sets and motifs in the UniPROBE data set is poor. ENCODE may be one of the biggest collections of ChIP-seq data, but currently it contains only a small sample of TF peak regions that are in the UniPROBE and hence BEEML-PBM database.

There is no motif in the BEEML-PBM database that accurately represents the true binding motif of Pou2f2 and Srf TFs. Irf4, Sp4, Tcf3 and Foxa2 bind DNA cooperatively with other TFs. Ets1 is highly associated with cancer. CentriMo results suggest that a TF involved in cancer is more enriched in cancer cell lines than in normal cell lines. It is also found that a cell line can make a big difference. This study revealed that the binding specificity of a TF is different in different cell types and development stages. A TF is up-regulated in a cell line where it performs its biological function. The main contribution of this study is the discovery of cell line differences, which has not been done before in any CentriMo study. This finding is interesting and provides reasons to study the relationship between TFs and tissue types further.

At this point there is no conclusive difference in the quality of motifs from the original PBM and BEEML-PBM approaches. Because there are so many reasons for a CentriMo distribution that differs from a strongly unimodal distribution, we treat such scenarios as cases for further investigation, rather than drawing unequivocal conclusions.

FUTURE PROSPECTS

That both approaches fail on several examples supports the case that evaluation of any found motif against in vivo data is essential. Further evaluation should consider the cause of failure.

This should address the following questions:

- (i) Are cases that appear to fail a result of a poor motif?
- (ii) Errors in the ChIP-seq experiment?
- (iii) Unusual or novel binding specificity?
- (iv) A failure of CentriMo? or
- (v) A more complex mode of binding?

These questions for “failed” cases can be evaluated by:

- (i) Comparing the “failed” motif to one found by *ab initio* motif discovery tools,
- (ii) comparing the CentriMo results presented here with that for another ChIP-seq data set for the same TF, and
- (iii) determining whether there is a case where CentriMo’s statistical approach does not work.

Any reasonable study of the quality of an in vitro or in silico derivation of motifs should be compared against the ground reality of in vivo data. Seeking out ChIP-seq data sets against which to evaluate both the PBM and BEEML-PBM approaches is important to enhance these analysis techniques further.

BIBLIOGRAPHY

- Abdel-Bakky, M. S., Hammad, M. A., Walker, L. A., & Ashfaq, M. K. (2011). Tissue factor dependent liver injury causes release of retinoid receptors (RXR- α and RAR- α) as lipid droplets. *Biochemical and biophysical research communications*, 410(1), 146-51. Elsevier Inc. doi:10.1016/j.bbrc.2011.05.127
- Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4), 477-86. doi:10.1093/bioinformatics/btg432
- Atchley, W. R., & Fitch, W. M. (1997). A natural classification of the basic helix – loop – helix class of transcription factors. *Proc. Natl. Acad. Sci. USA*, 94, 5172-5176.
- Babayeva, N. D., Baranovskaya, O. I., & Tahirov, T. H. (2012). Structural basis of Ets1 cooperative binding to widely separated sites on promoter DNA. *PLoS one*, 7(3), e33698. doi:10.1371/journal.pone.0033698
- 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(5935), 1720-3. doi:10.1126/science.1162327
- Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12), 1653-1659. doi:10.1093/bioinformatics/btr261
- Bailey, T. L., & Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol*, 3, 21-29.
- Bailey, T. L., & Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17) e128. doi:10.1093/nar/gks433
- Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research*, 34(Web Server issue), W369-73. doi:10.1093/nar/gkl198
- Balleza, E., Lopez-Bojorquez, L. N., Martinez-Antonio, A., Resendis-Antonio, O., Lozada-Chavez, I., Balderas-Martinez, Encarnacion, S., et al. (2009). Regulation by transcription factors in bacteria: beyond description. *FEMS microbiology*, 33, 133-151. doi:10.1111/j.1574-6976.2008.00145.x
- Beilharz, T. H., & Preiss, T. (2004). Translational profiling: the genome-wide measure of the nascent proteome. *Briefings in functional genomics & proteomics*, 3(2), 103-111. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15355593>
- 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. doi:10.1038/nprot.2008.195.

- Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., & Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, *24*(11), 1429-35. doi:10.1038/nbt1246
- Bieda, M., Xu, X., Singer, M. A., Green, R., & Farnham, P. J. (2006). Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome research*, *16*(5), 595-605. doi:10.1101/gr.4887606
- Bossard, P., & Zaret, K. S. (1998). GATA transcription factors as potentiators of gut endoderm differentiation. *Development*, *125*, 4909-4917.
- Conlon, E. M., Liu, X. S., Lieb, J. D., & Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(6), 3339-44. doi:10.1073/pnas.0630591100
- Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(6), 561-563.
- Dittmer, J. (2003). The Biology of the Ets1 Proto-Oncogene. *Molecular Cancer*, *2*, 1-21.
- Down, T. A., & Hubbard, T. J. P. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic acids research*, *33*(5), 1445-53. doi:10.1093/nar/gki282
- Elkon, R., Linhart, C., Sharan, R., Shamir, R., & Shiloh, Y. (2003). Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome research*, *13*(5), 773-80. doi:10.1101/gr.947203
- Fang, B., Mane-padros, D., Bolotin, E., Jiang, T., & Sladek, F. M. (2012). Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors. *Nucleic Acids Research*, 1-14. doi:10.1093/nar/gks190
- Frith, M. C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., & Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. *Nucleic acids research*, *32*(4), 1372-81. doi:10.1093/nar/gkh299
- Gauthier, B. R., Schwitzgebel, V. M., Zaiko, M., Mamin, A., Ritz-laser, B., & Philippe, J. (2002). Hepatic Nuclear Factor-3 (HNF-3 or Foxa2) Regulates Glucagon Gene Transcription by Binding to the G1 and G2 Promoter Elements. *Molecular Endocrinology*, *16*(1), 170-183.
- Gulbinas, A., Berberat, P. O., Dambrauskas, Z., Giese, T., Giese, N., Autschbach, F., Kleeff, J., et al. (2006). Aberrant gata-3 expression in human pancreatic cancer. *The journal of histochemistry and cytochemistry*, *54*(2), 161-9. doi:10.1369/jhc.5A6626.2005
- Gupta, S., Stamatoyannopoulos, J. a, Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, *8*(2), R24. doi:10.1186/gb-2007-8-2-r24

- Gustavson, M. D., Crawford, H. C., Fingleton, B., & Matrisian, L. M. (2004). Tcf binding sequence and position determines beta-catenin and Lef-1 responsiveness of MMP-7 promoters. *Molecular carcinogenesis*, *41*(3), 125-39. doi:10.1002/mc.20049
- Halene, S., Gao, Y., Hahn, K., Massaro, S., Italiano, J. E., Lin, S., Kupfer, G. M., et al. (2010). Serum response factor is an essential transcription factor in megakaryocytic maturation. *Blood*, 1-26. doi:10.1182/blood-2010-01-261743
- Haverty, P. M., Hansen, U., & Weng, Z. (2004). Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic acids research*, *32*(1), 179-188. doi:10.1093/nar/gkh183
- Hill, D. P., Blake, J. A., Richardson, J. E., & Ringwald, M. (2002). Extension and Integration of the Gene Ontology (GO): Combining GO Vocabularies With External Vocabularies. *Genome Research*, *12*, 1982-1991. doi:10.1101/gr.580102.11
- Ho, I., Vorhees, P., Marin, N., Oakley, B. K., Tsai, S., Orkin, S. H., & Leiden, J. M. (1991). Human GATA-3 : a lineage-restricted transcription factor that regulates the expression of the T cell receptor OI gene. *The EMBO journal*, *10*(5), 1187-1192.
- Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M., & Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research*, *38*(7), 2154-67. doi:10.1093/nar/gkp1180
- Huan, B., & Siddiqui, A. (1992). Retinoid X receptor RXR alpha binds to and trans-activates the hepatitis B virus enhancer. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(19), 9059-63. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=50064&tool=pmcentrez&rendertype=abstract>
- Hurlin, P. J., & Huang, J. (2006). The MAX-interacting transcription factor network. *Seminars in cancer biology*, *16*(4), 265-74. doi:10.1016/j.semcancer.2006.07.009
- Jothi, R., Cuddapah, S., Barski, A., Cui, K., & Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucleic acids research*, *36*(16), 5221-31. doi:10.1093/nar/gkn488
- Kadonaga, J. T. (2004). Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell*, *116*, 247-257.
- Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic acids research*, *32*(Database issue), D493-6. doi:10.1093/nar/gkh103
- Kassouf, M. T., Hughes, J. R., Taylor, S., Cheng, Y., King, D. C., & Dore, L. C. (2010). Genome-wide identification of TAL1's functional targets : Insights into its mechanisms of action in primary erythroid cells. *Genome research*, 1064-1083. doi:10.1101/gr.104935.110

- Khodursky, A. B., & Bernstein, J. A. (2003). Life after transcription--revisiting the fate of messenger RNA. *Trends in genetics*, *19*(3), 113-5. doi:10.1016/S0168-9525(02)00047-1
- Kim, W., Sieweke, M., Ogawa, E., Wee, H., Englmeier, U., Graf, T., & Ito, Y. (1999). Mutual activation of Ets-1 and AML1 DNA binding by direct interaction of their autoinhibitory domains. *The EMBO Journal*, *18*(6), 1609-1620.
- Ko, L. J., & Engel, J. D. (1993). DNA-Binding Specificities of the GATA Transcription Factor Family. *Molecular and cellular biology*, *13*(7), 4011-4022.
- Koch, J., & Bu, T. R. (2006). Sequence analysis LogoBar : bar graph visualization of protein logos with gaps. *Bioinformatics*, *22*(1), 112-114. doi:10.1093/bioinformatics/bti761
- Krylov, D., & Vinson, C. R. (2001). Leucine Zipper. *Encyclopedia of life sciences, Nature*, 1-7.
- Kubosaki, A., Tomaru, Y., Tagami, M., Arner, E., Miura, H., Suzuki, T., Suzuki, M., et al. (2009). Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation. *Genome biology*, *10*(4), R41. doi:10.1186/gb-2009-10-4-r41
- Lackner, D. H., & Bahler, J. (2008). Translational Control of Gene Expression : From Transcripts to Transcriptomes. *International Review of Cell and Molecular Biology*, *271*(08), 199-251. doi:10.1016/S1937-6448(08)01205-7
- Larsson, L., Bahram, F., Burkhardt, H., & Lüscher, B. (1997). Analysis of the DNA-binding activities of Myc/Max/Mad network complexes during induced differentiation of U-937 monoblasts and F9 teratocarcinoma cells. *Oncogene*, *15*, 737-748. Retrieved from <http://ukpmc.ac.uk/abstract/MED/9264414>
- Lee, B., & Iyer, V. R. (2012). Genome-wide Studies of CCCTC-binding Factor (CTCF) and Cohesin Provide Insight into Chromatin Structure and Regulation. *The Journal of biological chemistry*, *287*(37), 30906-13. doi:10.1074/jbc.R111.324962
- Lee, Y., Kim, M., Han, J., Yeom, K., Lee, S., Baek, S. H., & Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, *23*, 4051-60. doi:10.1038/sj.emboj.7600385
- Lennard, M. L., Wilson, M. R., Miller, N. W., Clem, L. W., Warr, G. W., & Hikima, J.-ichi. (2006). Oct2 transcription factors in fish--a comparative genomic analysis. *Fish & shellfish immunology*, *20*(2), 227-38. doi:10.1016/j.fsi.2005.01.011
- Lin, J. X., & Leonard, W. J. (1997). The immediate-early gene product Egr-1 regulates the human interleukin-2 receptor beta-chain promoter through noncanonical Egr and Sp1 binding sites . *Mol. Cell. Biol.*, *17*(7), 3714-3722.
- Linhart, C., Halperin, Y., & Shamir, R. (2008). Transcription factor and microRNA motif discovery : The Amadeus platform and a compendium of metazoan target sets. *Genome Research*, *18*, 1180-1189. doi:10.1101/gr.076117.108.2

- Liu, R., McEachin, R. C., & States, D. J. (2003). Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. *Genome research*, *13*(4), 654-61. doi:10.1101/gr.911803
- Lozzio, C. B., & Lozzio, B. B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood*, *45*(3), 321-334.
- Luscombe, N., Austin, S., Berman, H., & Thornton, J. (2000). An overview of the structures of protein-DNA complexes. *Genome biology*, *1*(1), 1-10. Retrieved from <http://genomebiology.com/2000/1/1/rEVIIEWS/001>
- Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, *27*(12), 1696-1697. doi:10.1093/bioinformatics/btr189
- Mangelsdorf, D. J., & Evanst, R. M. (1995). The RXR Heterodimers and Orphan Receptors. *Cell*, *83*, 841-850.
- Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics*, *7*, 29-59. doi:10.1146/annurev.genom.7.080505.115623
- Mata, J., Marguerat, S., & Bähler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. *Trends in biochemical sciences*, *30*(9), 506-14. doi:10.1016/j.tibs.2005.07.005
- McLeay, R. C., & Bailey, T. L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC bioinformatics*, *11*, 1-11. doi:10.1186/1471-2105-11-165
- Miano, J. (2003). Serum response factor: toggling between disparate programs of gene expression. *Journal of Molecular and Cellular Cardiology*, *35*(6), 577-593. doi:10.1016/S0022-2828(03)00110-X
- Miller, J., McLachlan, A., & Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus oocytes*. *The EMBO journal*, *4*(6), 1609-1614. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC554390/>
- Min, I. M., Pietramaggiore, G., Kim, F. S., Passegué, E., Stevenson, K. E., & Wagers, A. J. (2008). The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell*, *2*(4), 380-91. doi:10.1016/j.stem.2008.01.015
- Miralles, F., Posern, G., Zaromytidou, A., & Treisman, R. (2003). Actin Dynamics Control SRF Activity by Regulation of Its Coactivator MAL. *Cell*, *113*, 329-342.
- Mo, Y., Ho, W., Johnston, K., & Marmorstein, R. (2001). Crystal Structure of a Ternary SAP-1 / SRF / c-fos SRE DNA Complex. *Journal of Molecular Biology*, *314*, 495-506. doi:10.1006/jmbi.2001.5138

- Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., et al. (2004). Rapid analysis of the DNA Binding Specificities of Transcription Factors with DNA Microarrays. *Nature genetics*, *36*(12), 1331-1339. doi:10.1038/ng1473.Rapid
- Naylor, M. J., & Ormandy, C. J. (2007). Gata-3 and mammary cell fate. *Breast cancer research : BCR*, *9*(2), 302-303. doi:10.1186/bcr1661
- Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research*, *37*(Database issue), D77-82. doi:10.1093/nar/gkn660
- Nishida, K., Frith, M. C., & Nakai, K. (2009). Pseudocounts for transcription factor binding sites. *Nucleic acids research*, *37*(3), 939-44. doi:10.1093/nar/gkn1019
- Odom, D. T., Dowell, R. D., Jacobsen, E. S., Nekludova, L., Rolfe, P. A., Danford, T. W., Gifford, D. K., et al. (2006). Core transcriptional regulatory circuitry in human hepatocytes. *Molecular systems biology*, *2*, 1-5. doi:10.1038/msb4100059
- Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Nicola, J., Murray, H. L., Volkert, T. L., et al. (2004). Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. *Science*, *303*(5662), 1378-1381. doi:10.1126/science.1089769.Control
- Oikawa, T., & Yamada, T. (2003). Molecular biology of the Ets family of transcription factors. *Gene*, *303*, 11-34. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0378111902011563>
- Orian, A., van Steensel, B., Delrow, J., Bussemaker, H. J., Li, L., Sawado, T., Williams, E., et al. (2003). Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. *Genes & development*, *17*(9), 1101-14. doi:10.1101/gad.1066903
- Pabo, C., & Sauer, R. (1992). Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry*, *61*, 1053-95. Retrieved from <http://www.annualreviews.org/doi/pdf/10.1146/annurev.bi.61.070192.005201>
- Pandolfi, P. P. (2004). Aberrant mRNA translation in cancer pathogenesis: an old concept revisited comes finally of age. *Oncogene*, *23*(18), 3134-7. doi:10.1038/sj.onc.1207618
- Papatsenko, D. A., Makeev, V. J., Lifanov, A. P., Regnier, M., Nazina, A. G., & Desplan, C. (2002). Extraction of Functional Binding Sites from Unique Regulatory Regions: The Drosophila Early Developmental Enhancers. *Genome Research*, *12*(3), 470-481. doi:10.1101/gr.212502
- Patrik, D. (2006). What are DNA sequence motifs ? *Nature Biotechnology*, *24*(4), 423-425.
- Pavesi, G., Mereghetti, P., Mauri, G., & Pesole, G. (2004). Weeder Web : discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acid Research*, *32*, 199-203. doi:10.1093/nar/gkh465

- Pearson, R., Fleetwood, J., Eaton, S., Crossley, M., & Bao, S. (2008). Krüppel-like transcription factors: a functional family. *The international journal of biochemistry & cell biology*, 40(10), 1996-2001. doi:10.1016/j.biocel.2007.07.018
- Phillips, K., & Luisi, B. (2000). The virtuoso of versatility: POU proteins that flex to fit. *Journal of molecular biology*, 302(5), 1023-39. doi:10.1006/jmbi.2000.4107
- Qian, J., Esumi, N., Chen, Y., Wang, Q., Chowers, I., & Zack, D. J. (2005). Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic acids research*, 33(11), 3479-91. doi:10.1093/nar/gki658
- Qian, Z., Cai, Y., & Li, Y. (2006). Automatic transcription factor classifier based on functional domain composition. *Biochemical and biophysical research*, 347, 141-144. doi:10.1016/j.bbrc.2006.06.060
- Raney, B. J., Cline, M. S., Rosenbloom, K. R., Dreszer, T. R., Learned, K., Barber, G. P., Meyer, L. R., et al. (2011). ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic acids research*, 39(Database issue), D871-5. doi:10.1093/nar/gkq1017
- Reddy, D. A., Prasad, B. V. L. S., & Mitra, C. K. (2006). Functional classification of transcription factor binding sites : Information content as a metric. *Journal of Integrative Bioinformatics*, 1-13.
- Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., Newberry, K. M., Marinov, G. K., et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research*, 22, 860-9. doi:10.1101/gr.131201.111
- Rengarajan, J., Mowen, K. A., McBride, K. D., Smith, E. D., Singh, H., & Glimcher, L. H. (2002). Interferon Regulatory Factor 4 (IRF4) Interacts with NFATc2 to Modulate Interleukin 4 Gene Expression. *Journal of Experimental Medicine*, 195(8), 1003-1012. doi:10.1084/jem.20011128
- Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, 21(9), 327-335.
- Roider, H. G., Manke, T., O'Keefe, S., Vingron, M., & Haas, S. a. (2009). PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4), 435-42. doi:10.1093/bioinformatics/btn627
- Ryan, A. K., & Rosenfeld, M. G. (1997). POU domain family values: flexibility, partnerships, and developmental codes. *Genes & Development*, 11(10), 1207-1225. doi:10.1101/gad.11.10.1207
- Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(Database issue), D91-4. doi:10.1093/nar/gkh012

- Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097-6100.
- Sharan, R., Ovcharenko, I., Ben-Hur, A., & Karp, R. M. (2003). CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, *19*(Suppl 1), i283-i291. doi:10.1093/bioinformatics/btg1039
- Sharif, N. M., Radomska, H. S., Miller, D. M., & Eckhardt, L. A. (2001). Unique Function for Carboxyl-Terminal Domain of Oct-2 in Ig-Secreting Cells. *Journal of Immunology*, *167*, 4421-29.
- Sharma, S., Singh, R., Rana, S., & Uniyal, R. (2011). IDENTIFICATION OF MOTIFS IN BIOACTIVE PEPTIDES PRECURSORS. *International Journal of Bioinformatics and Biosciences*, *1*(1), 13-26. Retrieved from <http://www.ss-journals.com/index.php/ijbar/article/view/221>
- Smith, A. D., Sumazin, P., Das, D., & Zhang, M. Q. (2005). Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, *21*, i403-12. doi:10.1093/bioinformatics/bti1043
- Steenbergen, R. D. M., Oudeengberink, V. E., Kramer, D., Schrijnemakers, H. F. J., Verheijen, R. H. M., Meijer, C. J. L. M., & Snijders, P. J. F. (2002). Down-Regulation of GATA-3 Expression during Human Papillomavirus-Mediated Immortalization and Cervical Carcinogenesis. *American Journal of Pathology*, *160*(6), 1945-1951.
- Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, *16*(1), 16-23.
- Strom, A., Hartman, J., Foster, J. S., Kietz, S., Wimalasena, J., & Gustafsson, J.-åke. (2004). Estrogen receptor beta inhibits 17beta-estradiol-stimulated proliferation of the breast cancer cell line T47D. *PNAS*, *101*(6), 1566-1571.
- Suzuki, T., Aizawa, K., Matsumura, T., & Nagai, R. (2005). Vascular implications of the Krüppel-like family of transcription factors. *Arteriosclerosis, thrombosis, and vascular biology*, *25*(6), 1135-41. doi:10.1161/01.ATV.0000165656.65359.23
- Tallack, M. R., Whittington, T., Yuen, W. S., Wainwright, E. N., Keys, J. R., Gardiner, B. B., Nourbakhsh, E., et al. (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome research*, *20*(8), 1052-63. doi:10.1101/gr.106575.110
- Tang, Y., Shu, G., Yuan, X., Jing, N., & Song, J. (2011). FOXA2 functions as a suppressor of tumor metastasis by inhibition of epithelial-to-mesenchymal transition in human lung cancers. *Cell research*, *21*, 316-26. Nature Publishing Group. doi:10.1038/cr.2010.126
- The ENCODE Project Consortium. (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS biology*, *9*(4), 1-21. doi:10.1371/journal.pbio.1001046

- Tian, G., Erman, B., Ishii, H., Samudra, S., Tian, G., Erman, B., Ishii, H., et al. (1999). Transcriptional Activation by ETS and Leucine Zipper-Containing Basic Helix-Loop-Helix Proteins. *Molecular and Cellular Biology*, 19(4), 2946–2957.
- Usary, J., Llaca, V., Karaca, G., Presswala, S., Karaca, M., He, X., Langerød, A., et al. (2004). Mutation of GATA3 in human breast tumors. *Oncogene*, 23(46), 7669-78. doi:10.1038/sj.onc.1207966
- Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Batzoglou, S., Myers, R. M., & Sidow, A. (2008). Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-seq Data. *Nature Methods*, 5(9), 829-834. doi:10.1038/nmeth.1246.Genome-Wide
- Veenstra, G., Vliet, P. van der, & Destrée, O. (1997). POU domain transcription factors in embryonic development. *Molecular biology reports*, 24, 139-155. Retrieved from <http://www.springerlink.com/index/L85022G003802284.pdf>
- Verschuur, M. L., Wilson, L. A., Verschuur, C. P., & Singh, G. (2010). Ets-1 regulates energy metabolism in cancer cells. *PloS one*, 5(10), e13565. doi:10.1371/journal.pone.0013565
- Visvader, J. E., & Lindeman, G. J. (2003). Transcriptional regulators in mammary gland development and cancer. *The International Journal of Biochemistry & Cell Biology*, 35(7), 1034-1051. doi:10.1016/S1357-2725(03)00030-X
- Voduc, D., Cheang, M., & Nielsen, T. (2008). GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer epidemiology, biomarkers & prevention*, 17(2), 365-73. doi:10.1158/1055-9965.EPI-06-1090
- Von Hippel, P. H. (1998). An Integrated Model of the Transcription Complex in Elongation, Termination, and Editing. *Science*, 281, 660-665. doi:10.1126/science.281.5377.660
- Waby, J. S., Bingle, C. D., & Corfe, B. M. (2008). Post-Translational Control of Sp-Family Transcription Factors. *Current genomics*, 9, 301-311.
- Wagner, M., Schmelz, K., Dörken, B., & Tamm, I. (2008). Transcriptional regulation of human survivin by early growth response (Egr)-1 transcription factor. *International journal of cancer. Journal international du cancer*, 122(6), 1278-87. doi:10.1002/ijc.23183
- Wallerman, O., & Motallebipour, M. (2009). Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Research*, 37(22), 7498-7508. doi:10.1093/nar/gkp823
- Wang, X., & Zhang, X. (2011). Pinpointing transcription factor binding sites from ChIP-seq data with SeqSite. *BMC systems biology*, 5(2), 1-14. BioMed Central Ltd. doi:10.1186/1752-0509-5-S2-S3

- Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., & Brown, P. O. (2002). Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9), 5860-5. doi:10.1073/pnas.092538799
- Wasserman, W W, & Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *Journal of molecular biology*, 278(1), 167-81. doi:10.1006/jmbi.1998.1700
- Wasserman, Wyeth W, & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature reviews Genetics*, 5(4), 276-87. doi:10.1038/nrg1315
- Wasylyk, C., Schlumberger, S. E., Wasylyk, C., Schlumberger, S. E., Criqui-filipe, P., & Wasylyk, B. (2002). Sp100 Interacts with ETS-1 and Stimulates Its Transcriptional Activity. *Molecular and Cellular Biology*, 22(8), 2687-2702. doi:10.1128/MCB.22.8.2687
- Watkins, S. J., & Norbury, C. J. (2002). Translation initiation and its deregulation during tumorigenesis. *British journal of cancer*, 86, 1023-1027. doi:10.1038/sj/bjc/6600222
- Watson, J., & Crick, F. (1953). The structure of DNA, 532-539. Retrieved from <http://symposium.cshlp.org/content/18/123.short>
- Wederell, E. D., Bilenky, M., Cullum, R., Thiessen, N., Dagpinar, M., Delaney, A., Varhol, R., et al. (2008). Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic acids research*, 36(14), 4549-64. doi:10.1093/nar/gkn382
- Whittington, T., Frith, M. C., Johnson, J., & Bailey, Timothy, L. (2011). Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Research*, 39(15), 1-11. doi:10.1093/nar/gkr341
- Wilusz, C. J., & Wilusz, J. (2004). Bringing the role of mRNA decay in the control of gene expression into focus. *Trends in genetics*, 20(10), 491-7. doi:10.1016/j.tig.2004.07.011
- Wittkopp, P. (2010). Variable transcription factor binding: a mechanism of evolutionary change. *PLoS biology*, 8(3), 2-4. doi:10.1371/journal.pbio.1000342
- Yan, S., Pinsky, D., Mackman, N., & Stern, D. (2000). Egr-1: is it always immediate and early? *The Journal of Clinical Investigation*, 105(5), 553-554. Retrieved from <http://www.jci.org/cgi/content/abstract/105/5/553>
- Yang, E., Nimwegen, E. V., & Zavolan, M. (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome research*, 13, 1863-1872. doi:10.1101/gr.1272403.7
- Yang, Z., Wen, H., Minhas, V., & Wood, C. (2009). The zinc finger DNA-binding domain of K-RBP plays an important role in regulating Kaposi's sarcoma-associated herpesvirus

RTA-mediated gene expression. *Virology*, 391(2), 221-31.
doi:10.1016/j.virol.2009.06.014

Yu, X., Lin, J., Zack, D. J., & Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic acids research*, 34(17), 4925-36. doi:10.1093/nar/gkl595

Zamule, S. M., Coslo, D. M., Chen, F., & Omiecinski, C. J. (2011). Differentiation of human embryonic stem cells along a hepatic lineage. *Chemico-biological interactions*, 190(1), 62-72. Elsevier Ireland Ltd. doi:10.1016/j.cbi.2011.01.009

Zhao, Y., Granas, D., & Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5(12), 1-8.
doi:10.1371/journal.pcbi.1000590

Zhao, Y., & Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, 29(6), 480-483.

Zheng, J., Wu, J., & Sun, Z. (2003). An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Research*, 31(7), 1995-2005. doi:10.1093/nar/gkg287

WEBSITE REFERENCES

<http://genome.ucsc.edu/ENCODE/> (used for CentriMo, SpaMo and AME sequence retrieval)

<http://meme.nbcr.net> (used to run SpaMo)

APPENDIX

Lines typed on the command line have a "\$" in front to distinguish from script files.

APPENDIX: A

Motif extraction

```
$ export tf=Irf4
$ export dbdir=$HOME/meme/db/motif_databases
$ export db=zhao2011 #can be any database
$ grep $tf $dbdir/$db.meme
#MOTIF sci09.v1_Irf4_3476 UP00018#output
#MOTIF sci09.v2_Irf4_3476 UP00018
$ meme-get-motif -id sci09.v1_Irf4_3476 -id sci09.v2_Irf4_3476
< $dbdir/$db.meme > $dbdir/$tf-$db.meme
```

Script

#The script obtains a FASTA file from hg 19, given a BED file: get-fasta.bash

```
#!/bin/bash
ddir=$HOME/Downloads
width=500
genome=hg19

gdir=$HOME/Downloads
gfile=all-repeat-N.fa
for tf in Irf4
do
  echo $tf

  if [ -e /tmp/${pre}${tf}.fa ]

  then
    rm -f /tmp/${pre}${tf}.fa
  fi

  bed-widen -width ${width} < ${ddir}/${pre}${tf}.peaks |
  fastaFromBed -fi ${gdir}/${gfile} -bed - -fo
  /tmp/${pre}${tf}.fa
  fasta-remove-all-n -excl N < /tmp/${pre}${tf}.fa >
  ${ddir}/${pre}${tf}.fa
done
script provided by Prof Philip Machanick
```

CentriMo on combined BEEML-PBM/UniPROBE database (command line)

```
$/get-fasta.bash
```

```
$ centriMo --verbosity 1 -oc Irf4_centrimo Downloads  
/Irf4/Irf4.fa ~/meme/db/motif_databases/Irf4.meme
```

#The Irf4.meme file contains the motifs from BEEML-PBM and UniPROBE databases.

CentriMo on combined JASPAR/UniPROBE database (command line)

```
$centrimo --verbosity 1 -oc Irf4_centrimo  
Project/JASPAR_UniPROBE_CentriMo/Irf4/Irf4.fa/home/ntombi/meme  
/db/motif_databases/uniprobe_mouse.meme  
$HOME/meme/db/motif_databases/JASPAR_CORE_2009.meme
```

CEQLOGO COMMAND LINE

```
$ceqlogo -i Irf4-uniprobe_mouse.meme -t Irf4 > Irf4.eps
```

#can use motif from any database.

#ceqlogo is one of many ways used to generate motif logos.

AME SCRIPT

```
#!/bin/bash

if [ $# -lt 2 ]
then
    echo "USAGE: $0 fastafile DB+"
    echo "must supply fasta file plus at least one DB"
    exit 1
fi

export tf=`echo $1|cut -d '-' -f 1`

echo "TF="$tf

export seq=$1

shift

export DBs=$*

fasta-dinucleotide-shuffle -f $seq -t -dinuc > $tf-shuffled.fa

cat $seq $tf-shuffled.fa > $tf-w_bg.fa

ame --verbose 1 --oc ${tf}_ame_out --fix-partition `getsize
$seq | cut -f 1 -d " "` --bgformat 0 $tf-w_bg.fa $DBs

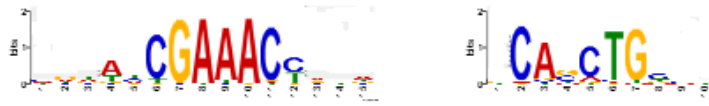
#script provided by Philip Machanick

#For script explanation, see AME electronic data.
```

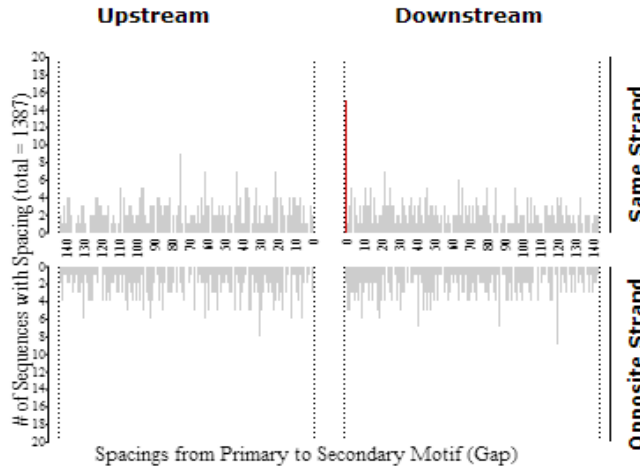
APPENDIX: B

Primary: UP00018 1 (Irf4 primary)

Secondary: [sci09.v1 Ascl2 2654 \(UP00099\)](#)



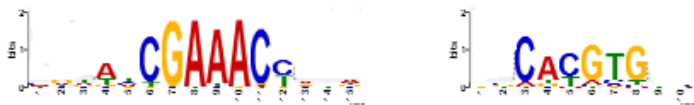
Motif Spacing Histogram



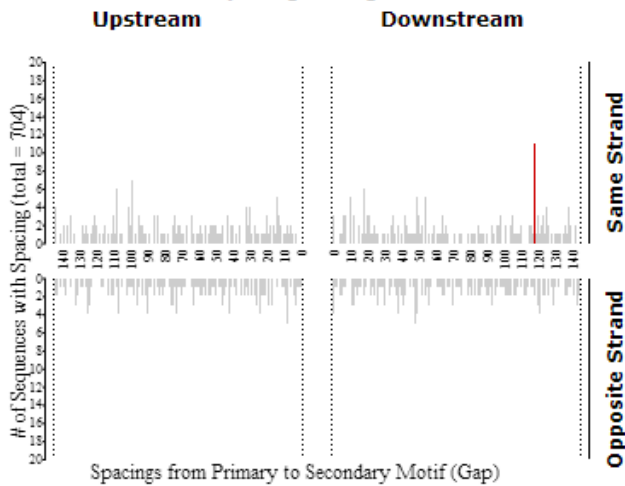
	Upstream	Downstream
Same Strand		P-value Gap # 2.3e-05 0 15
Opposite Strand		

Primary: UP00018 1 (Irf4 primary)

Secondary: [sci09.v1 Max 3863 \(UP00060\)](#)



Motif Spacing Histogram



	Upstream	Downstream
Same Strand		P-value Gap # 3.8e-05 118 11
Opposite Strand		

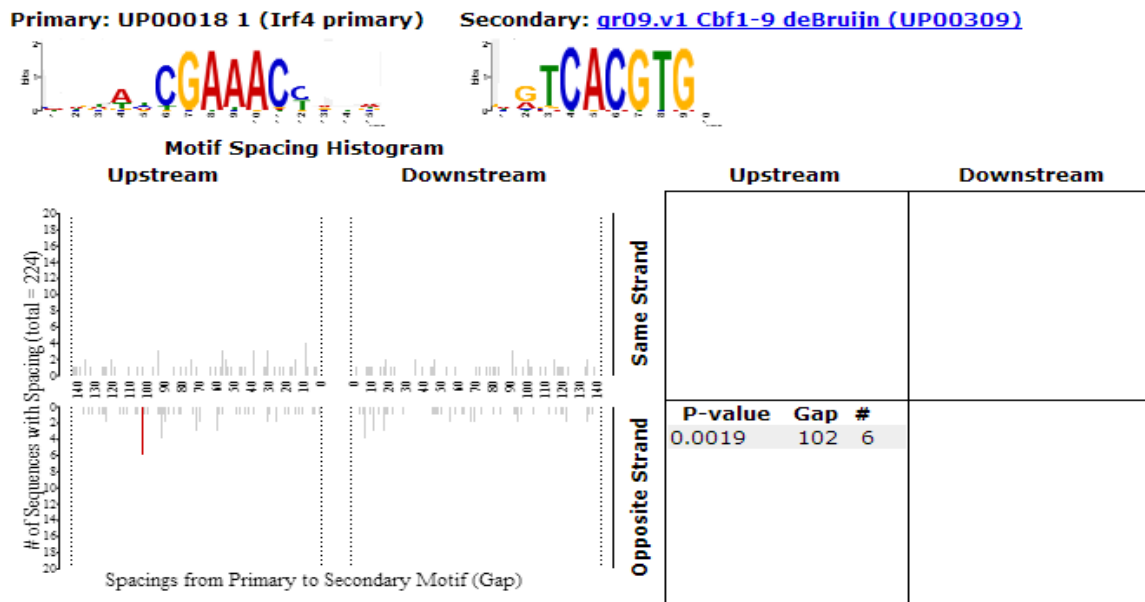
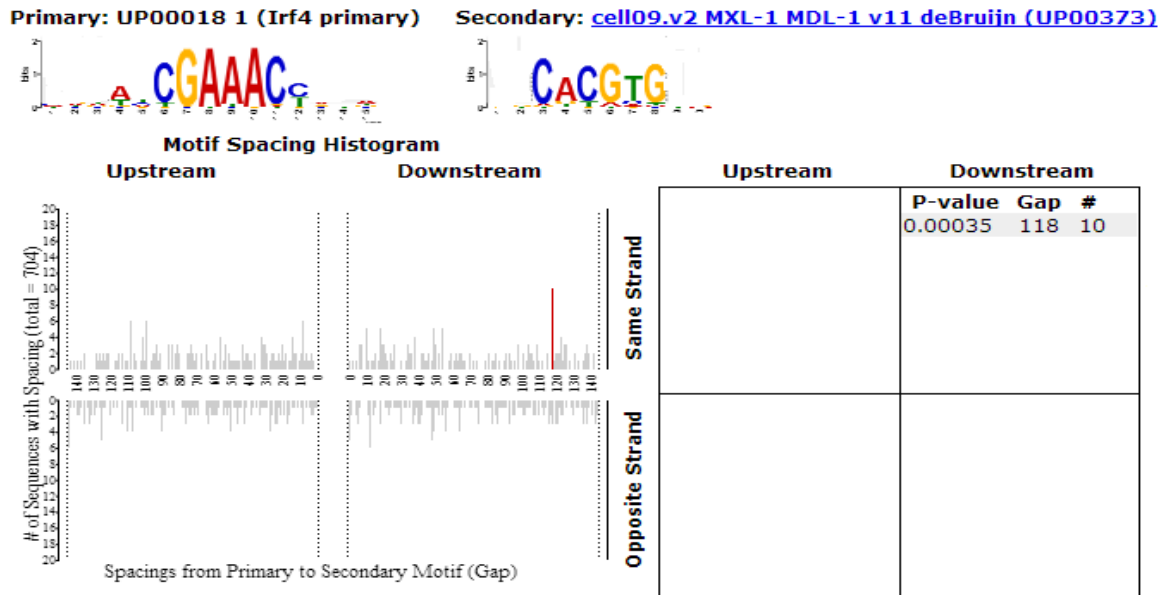


Figure 1: Evidence that Irf4 binds in combination with E-box motifs.