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

Master of Science in Bioinformatics and Computational Molecular Biology by coursework and thesis

Comparison of Protein Binding Microarray Derived and ChIP-seq Derived Transcription Factor Binding DNA motifs

Submitted by: Nkosikhona Hlatshwayo

Supervisor: Professor Philip Machanick

Department of Biochemistry, Microbiology and Biotechnology
and
Department of Computer Science

Rhodes University
Grahamstown
South Africa

January 2015

Plagiarism Declaration

I, *Nkosikhona Rejoyce Hlatshwayo* (g14H7829), declare that this thesis is my own work except AME, CentriMo and bedwiden scripts provided by Professor Philip Machanick. I acknowledge and understand that plagiarism is wrong, and that it constitutes academic theft. This thesis 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. All the sources used have been acknowledged.

Signature: _____ Date: _____

Acknowledgements

I would like to thank my supervisor Professor P. Machanick for first and foremost helping me get into the programme, helping me find funding, patiently teaching me and for all the assistance he provided through the course of 2014.

I am thankful to our lecturers who taught us during the first part of the year, for their patiently teaching us what seemed foreign to us, and doing so best they could.

I would also like to thank Professor O. Tastan Bishop for accepting me and allowing the opportunity to be part of this fantastic MSc. Programme. I am grateful to other students in the programme as well, more especially Caroline Ross for her selflessness and invaluable contribution towards my learning.

I am especially thankful to Caleb Kibet for the continued assistance and moral support in the course of the last part of the year.

Thanks to the former MSc students of RuBi, more especially Ngoni Faya and Thommas Musyoka for being awesome and supportive.

I am also grateful to Dr. B. Kullin for inspiring me and helping me get started.

I acknowledge and express gratitude for the financial assistance from the National Research Foundation (NRF).

Abstract

Transcription factors (TFs) are biologically important proteins that interact with transcription machinery and bind DNA regulatory sequences to regulate gene expression by modulating the synthesis of the messenger RNA. The regulatory sequences comprise of short conserved regions of a specific length called *motifs*. TFs have very diverse roles in different cells and play a very significant role in development. TFs have been associated with carcinogenesis in various tissue types, as well as developmental and hormone response disorders. They may be responsible for the regulation of oncogenes and can be oncogenic. Consequently, understanding TF binding and knowing the motifs to which they bind is worthy of attention and research focus.

Various projects have made the study of TF binding their main focus; nevertheless, much about TF binding remains confounding. Chromatin immunoprecipitation in conjunction with deep sequencing (ChIP-seq) techniques are a popular method used to investigate DNA-TF interactions *in vivo.* This procedure is followed by motif discovery and motif enrichment analysis using relevant tools. Protein Binding Microarrays (PBMs) are an *in vitro* method for investigating DNA-TF interactions. We use a motif enrichment analysis tools (CentriMo and AME) and an empirical quality assessment tool (Area under the ROC curve) to investigate which method yields motifs that are a true representation of *in vivo* binding.

Motif enrichment analysis: On average, ChIP-seq derived motifs from the JASPAR Core database outperformed PBM derived ones from the UniPROBE mouse database. However, the performance of motifs derived using these two methods is not much different from each other when using CentriMo and AME. The E-values from Motif enrichment analysis were not too different from each other or 0. CentriMo showed that in 35 cases JASPAR Core ChIP-seq derived motifs outperformed UniPROBE mouse PBM derived motifs, while it was only in 11 cases that PBM derived motifs outperformed ChIP-seq derived motifs. AME showed that in 18 cases JASPAR Core ChIP-seq derived motifs did better, while only it was only in 3 cases that UniPROBE motifs outperformed ChIP-seq derived motifs. We could not distinguish the performance in 25 cases.

Empirical quality assessment: Area under the ROC curve values computations followed by a two-sided t-test showed that there is no significant difference in the average performances of the motifs from the two databases (with 95% confidence, mean of differences=0.0088125 p-value= 0.4874, DF=47).

Table of Contents

1. Introduction

1.1. Background

1.1.1. DNA: The Basic Unit of Inheritance

All cellular function and development of living organisms is dictated by the genetic code. This code is contained in a macromolecule called deoxyribonucleic acid (DNA). DNA has four different basic units or monomers; namely adenine, thymine, cytosine and guanine (A, T, C and G). These four monomers are joined to form a very long DNA molecule. The specific sequence that these monomers form is the code that dictates the entire being and physical appearance (phenotype) of the living organism.

For a long time biologists believed that the genetic information in the DNA is *transcribed* to messenger Ribonucleic Acid (mRNA) and subsequently the mRNA is *translated* into protein. This is referred to as the "central dogma". Biologists then discovered that in some viruses transcription can occur from RNA to DNA (Crick, 1970; Gerstein *et al.*, 2007). The identity and function of the protein produced is mainly dependant on the DNA sequence from which the mRNA which the protein was translated from was transcribed. Proteins then perform functions in a cell that are essential for life.

1.1.2. From DNA to Genes

As mentioned, DNA contains regions that are coding sequences or genes, that is to say DNA comprises of regions that are transcribed and code for proteins. Pearson (2006) defines a "gene" as a "locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions". This is the current definition of the term "gene" although a gene is much more complicated than this. There are several problems with the current definition of the term "gene" and these are discussed by Gerstein *et al.* (2007).

In humans, sequences that are translated (coding regions or *exons*) make up only approximately 1.2% of the genome. The word genome refers to the total DNA in all cells. The rest of the genome is composed of non-coding sequences which include transcription start sites, intergenic regions, *introns* and repetitive sequences. Intergenic regions comprise of regulatory and functional sequences (Venter *et al.*, 2001).

In humans and other eukaryotes, DNA is not always accessible to transcription machinery for transcription. This is because it is important that the genetic material is stored compactly and safely and that the expression of genes is regulated and that only the needed genes are transcribed. Exactly how this regulation occurs is dependant on the cell type, environmental stimuli, current developmental stage of the cell and other conditions such as the presence of stress. Different cells need different genes expressed, and some other genes are only needed under special conditions. TFs play an important role in ensuring correct gene expression according to the needs of each living organism (Field *et al.*, 2011; Reynolds *et al.*, 2013).

One way this happens is through how DNA is packaged in the cell. There are several hierarchical orders of packaging DNA. It is packaged into *nucleosomes* by a special class of proteins called *histones*. This is the basic unit of packaging which is dynamic and can be transiently unwrapped. There is free, unwrapped DNA between the nucleosomes called the linker DNA. At this stage DNA is accessible for transcription but with decreasing accessibility in the middle (Field *et al.*, 2011; Reynolds *et al.*, 2013).

DNA is further packaged by another type of histone called a linker histone. This stabilises the inaccessibility of DNA. In the next order of packaging, the nucleosomes coil into a condensed chromatin structure. This reduces the accessibility of the linker DNA. The accessibility of DNA depends on chemical modification of histone proteins by other proteins. Activity of these proteins called *chromatin remodellers* or *chromatin modifiers* could lead to the chromatin being open (DNA is accessible) or closed (DNA is not accessible). For instance, it is believed that acetylation (a chemical modification whereby an acetyl group is added to chromatin proteins) of chromatin leads to accessibility of DNA and has been associated with active transcription, while deacetylation is associated with silencing of genes. Chromatin conformation plays an important role in transcription regulation by either allowing or restricting access to DNA and TFs have been shown to play a central role in recruiting chromatin remodellers (Field *et al.*, 2011; Reynolds *et al.*, 2013)

For a long time biologists believed that these non-coding sequences had no functions and referred to them as "junk DNA" (Ohno, 1972; Gerstein *et al.*, 2007). However, it is now understood that these non-coding sequences have very important functions. These non-coding sequences contain specific sequence patterns such as promoters, silencers and other regulatory elements such as enhancers. These regulatory elements are recognised and bound by transcription machinery (such as the RNA polymerase II and transcription

factors) to initiate the transcription of DNA to RNA as needed by the cell. Although other factors (such as chromatin conformation) play a role in gene expression regulation, regulatory elements and the transcription factors (TFs) which bind them play a key role (Gerstein *et al.*, 2007; Wingeder *et al.*, 2013).

1.1.3. Transcription Factors (TFs)

TFs are a class of proteins that bind DNA along with others, such as the RNA polymerase II. Their activity could either repress or activate transcription of DNA to RNA. They have very diverse roles in cells and perform functions that are very important for life. Identifying the sequence patterns to which TFs bind is essential for decoding TF binding location in the genome, identifying genes associated with the TF and assigning it function. Furthermore, this could open a way for us to understand transcriptional regulation under various conditions and gene regulatory networks (Park, 2009; Zhong *et al.*, 2013; Baumgart *et al.*, 2013; Orenstein & Shamir, 2014).

1.1.4. Study Overview

In this study we examine two methods used to decipher TF binding location and identify the sequence patterns (or motifs) recognised by TFs. We also compared the ability of the two methods to identify sequence patterns accurately to see if one method performs significantly better than the other using three different measures.

2. Literature Review

2.1. Transcription Factors

Transcription factors (TFs) are proteins that interact with transcription machinery and bind DNA regulatory sequences to regulate gene expression. This is done through modulating the synthesis of the messenger RNA (Santolini *et al.*, 2014). Although transcription factors play a critical role in regulating gene expression, transcription factor activity is not solely responsible for the regulations of gene expression. Other factors such as histone modification also play a critical role (Chen *et al.*, 2014, Qu & Fang, 2013).

2.2. Transcription Factor Motifs

Transcription factors can be classified into two types namely, general and sequence specific. General transcription factors ubiquitously bind to DNA along with the RNA polymerase II in transcription of many genes. Sequence specific transcription factors, on the other hand, bind to DNA sequences associated with a specific set of genes, all having conserved sequence patterns (Chen *et al.*, 2014). We are interested in the sequence specific TFs. Sequence specific transcription factors mainly bind regulatory regions such as the promoter and enhancer regions in the genome (From here onwards we will refer to "sequence specific TFs" as "TFs" for simplicity). These regulatory regions comprise of shorter, conserved DNA sequences or patterns of a specific length. These short conserved DNA sequences or patterns of a specific length are called **motifs**. TFs have varying binding affinities for each motif (Lesluyes *et al.*, 2014, Orenstein & Shamir, 2014).

2.3. Transcription Factor-DNA Binding

TFs can bind in different ways to a DNA site. Some TFs bind DNA directly as homodimers or heterodimers while some bind indirectly through other

cofactors (Gordân *et* al., 2009; Ridinger-Saison *et al.,* 2012; Hunt *et al.*, 2014). Ridinger-Saison *et al.* (2012) found that Spi-1 can selectively bind DNA directly, preferably at the transcription start site. They also found that Spi-1 can bind DNA indirectly through another TF; GATA1 and TAL1 heterodimer. Furthermore, Spi-1 could both repress and activate genes through direct DNA binding, while there was evidence of Spi-1 repressing transcription through indirect DNA binding.

2.4. Biological Significance of Transcription Factors

TFs have very diverse roles that differ from cell to cell and play a very significant role in development. Through their regulation action, TFs are responsible for regulating many different crucial cellular functions including cell-proliferation, apoptosis, metabolism and differentiation (Dang, 1999; Tremblay *et al.*, 2010; Wang *et al.*, 2010; Baumgart *et al.*, 2013). In co-operation with other factors, transcription factors have been associated with carcinogenesis in various tissue types, developmental and hormone response disorders. An understanding of genome-wide TF-DNA binding is needed in order to find therapy to these disorders and cancers. (Huang *et al.*, 2006; Wang *et al.*, 2010; Baumgart *et al.*, 2013; Zhong *et al.*, 2013).

Deciphering TF binding locations and sequence patterns to which they bind is important. It is key to determining the link between TF binding and disease, the cause of TF associated diseases and disorders and unravelling transcriptional regulatory networks, the correlation between motifs and gene regulation and possibly identifying probable therapeutic targets (Clarke & Granek, 2003; Zhong *et al.*, 2013; Baumgart *et al.*, 2013; Orenstein & Shamir, 2014). Consequently, understanding TF binding and identifying motifs involved in transcription is worthy of attention and research focus.

Although determination of TF binding sites has been the the focus in many studies in computational biology, much remains to be understood about TF binding sites. For instance, not all binding locations for well characterised TFs have been identified. This is true even for well-studied organisms (Tompa *et al.*, 2005; Santolini *et al.*, 2014)

2.5. Efforts to Investigate Transcription Factor Binding

In efforts to understand TF binding, studies have used many different approaches. In this study we focus on two popular techniques, namely the Chromatin Immunoprecipitation coupled with deep sequencing (ChIP-seq) as well as the Protein Binding Microarray (PBM) techniques. These two techniques are used to construct the JASPAR Core and the UniPROBE databases. The JASPAR Core database comprises of motifs derived using *in vivo* ChIP-seq, *in vitro* PBM and *in vitro* SELEX. The UniPROBE database comprises of motifs derived using the PBM technique (Mukherjee *et al.*, 2004 ; Portales-Casamar et *al.,* 2010; Mathelier et *al.,* 2014). In this study we focus on comparing results obtained from only two methods, ChIP-seq and PBM.

2.5.1. Chromatin Immunoprecipitation and Deep Sequencing (ChIP-seq)

Chromatin immunoprecipitation (ChIP) assay is a popular technique used to assay DNA-protein interactions *in vivo*. In this procedure, cell contents are cross-linked with formaldehyde to ensure that DNA-TF complexes do not dissociate. Then cell nuclei are isolated and exposed to sonication to shear chromatin. Immunoprecipitation follows; where primary antibodies against the TF of interest and secondary antibodies against the primary antibodies are used to precipitate the DNA-TF complex out, the cross-links are reversed and DNA is purified and amplified using gene specific primers. This is then followed by high-throughput sequencing; the DNA fragments with the putative binding sites and the motif of the TF of interest are sequenced. The combination of the two techniques is referred to as ChIP-seq (Weinmann and

Farnham, 2002; Schmidt *et al.*, 2009; Santolini *et al.*, 2014).

2.5.1.1. Peak Calling

Following ChIP-seq techniques, the sequences are mapped onto the genome and "peak-calling" software is used to detect putative TF binding sites involved in DNA-TF binding (either direct or indirect). When the DNA fragments are mapped onto the genome they form clusters, and statistically significant clusters or peaks are more likely to be TF binding sites. These binding sites contain motifs. The peak files are deposited onto the *Enc*yclopedia *o*f *D*NA *E*lements (ENCODE) database (Schmidt *et al.*, 2009; Machanick and Bailey, 2011).

2.5.1.2. Motif Discovery

The next step is to find motifs from these putative binding sites. This is called *ab initio* motif discovery, and there are several algorithms employed to discover motifs. The motif discovery algorithm that is of interest to us is Multiple Expectation Maximization for Motif Elicitation or MEME (Schmidt *et al.*, 2009; Machanick and Bailey, 2011; Bailey and Machanick, 2012).

2.5.1.2.1. MEME (Multiple Expectation Maximization for Motif Elicitation)

The MEME algorithm searches for sequence patterns (or motifs) that are enriched in a ChIP-seq dataset. MEME accepts a dataset of DNA sequences and searches for the motifs using a statistical model to automatically choose optimum sequence width, frequency and description of each motif. MEME then gives an output of user-specified number of motifs. Frequently occurring or highly enriched sequence patterns are discovered MEME represents the motif profile in a matrix. The MEME algorithm only searches for gapless motifs. Gapped motifs are split into two motifs or more*.

*http://meme.nbcr.net/meme

2.5.1.3. Motif Representation

The conserved sequence within the TF binding sites have the same length yet some nucleotides in the same position may vary. Thus a motif is not represented by a single sequence but are expressed in Position Weight Matrices (PWMs) which are derived from a given nucleotide frequency at each position (Stormo, 2000).

PWMs can be represented by sequence logos that give a clear description of the consensus sequence. The sequence logo is made up of a stack of all nucleotides that occur in a given position. The total height of the all the nucleotides in a given position depicts how conserved the position in the sequence is (the taller the more conserved) and the amount of information that we can infer about the position or the information content, while the height (for of each nucleotide in the nucleotide stack tells us how frequently that nucleotide appears in that particular position (Xiong, 2006; Bailey and Machanick, 2012; Figure 1).

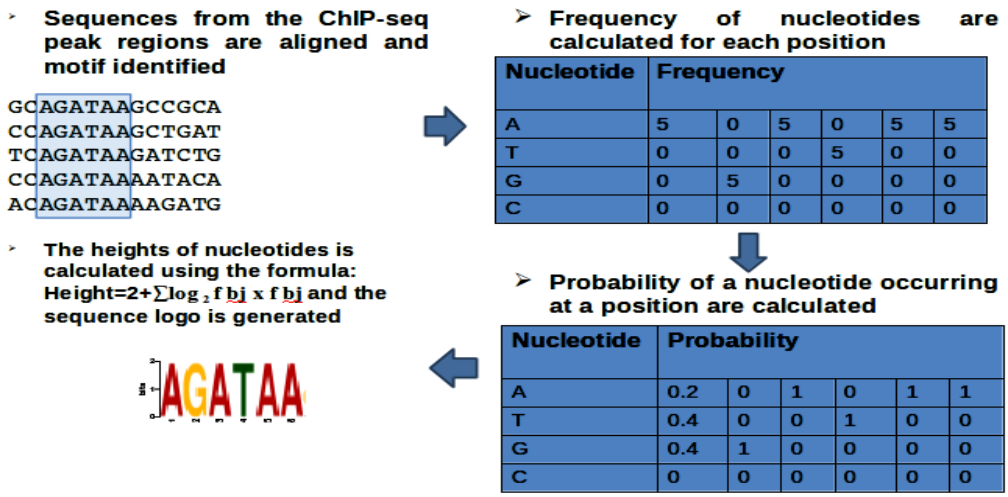


Figure 1: Showing the steps involved in generating Position Weight Matrices (PWMs) and representing motifs as sequence logos.

2.5.1.4. Data Repository for ChIP-seq Data and ChIP-seq Derived Motifs

A large amount of ChIP-seq data is deposited on the ENCODE database. Motifs discovered from the ChIP-seq data are discovered through *de novo* motif discovery using methods like MEME and are deposited into the JASPAR database along with other *in vitro* (PBM) derived motifs (Portales-Casamar *et al.*, 2010).

2.5.1.5. Advantages and Disadvantages of ChIP-seq

ChIP-seq is very popular and is considered a reliable method to investigate transcription factor binding sites in addition to investigating other DNA-protein interactions because the assay takes place *in vivo*. The ChIP-seq method has several advantages compared to the PBM method. First off, the experiments take place in actual cells, thus ensuring that the binding observed does actually occur in living cells and the binding site is accessible for TF binding. Additionally, known cells are used, so to a certain extent, we know the context in which the binding occurs. ChIP-seq is not limited to sequences on a microarray, since no microarrays are used and the required concentration of starting DNA is lower than in PBM (Park, 2009; Schmidt *et al.*, 2009).

Noise from mismatched cross-hybridization on arrays does not affect ChIP-seq thus there is overall less noise when using this method. Furthermore, in ChIP-seq it is the fragments from the experiment that are sequenced as opposed to hybridisation to a microarray of known fragments. This ensures a much more accurate binding prediction. Lastly, arrays have been observed to mask repeats that naturally occur in the genome. This does not happen in ChIP-seq (Park, 2009; Schmidt *et al.*, 2009; Field *et al*,. 2011).

In spite of all these advantages, this combination of techniques has its disadvantages. First off, ChIP-seq is laborious, time consuming and costly. Because there are multiple steps involved, this technique is also prone to human error, as there could be errors along the execution of the experiments. Because of the dynamic chromatin structure, some sites may not be available for binding, and thus we may not have access to all possible *k*-mers. Additionally, in fragment selection, a bias is observed towards GC-rich regions (Park *et* al., 2009; Orenstein & Shamir, 2014).

The accuracy of the data relies on an accuracy of the reference genome sequence used to align results. Furthermore, the success of this technique is also dependant on available resources, for example, there may be TFs which do not have specific antibodies, meaning that these TFs' motifs cannot be determined. Because these experiments are context dependant, information we glean is specific to the cell line or specific condition, consequently, we cannot get a general idea of TF binding. Using this technique we can only assay TF binding in available cell lines (Park *et* al., 2009; Schmidt *et al.*, 2009; Orenstein & Shamir, 2014)

2.5.2. Protein Binding Microarrays (PBM)

Protein Binding Microarrays (PBMs) is an *in vitro* technique also used to investigate TF binding on a genome-wide scale. This technique is much faster and DNA-TF interactions can be determined in one day. In this technique, microarrays containing known sequence fragments are used. This microarray is constructed using a de Bruijn sequence. The de Bruijn sequence is a sequence that is artificially constructed in such a way that all possible *k*-mers are represented within an approximately 40,000 sequences of 60-bases length. A *k*-mer is a DNA fragment and *k* is a variable for the length of the sequence. The *k*-mers appear for more than instance on a microarray. In each instance a specific *k*-mer is flanked by variable sequences to ensure that observed TF

binding is not binding the flanking regions but the *k*-mer itself (Badis *et al.*, 2009; Badis *et al.*, 2009 Supporting Online Material; Matthew *et al.*, 2013, Jiang *et al.*, 2013).

The TF of interest is expressed, purified and tagged with an epitope, then hybridised to the microarray. The unbound TF is washed off the microarray. A fluorophore tagged antibody against the epitope is hybridised to the microarray to bind the epitope-TF complex. If there are TF molecules binding onto the epitopes the fluorophore tagged antibody will find the epitope-TF compex and bind. The unbound antibodies are washed off. A scanner that detects the fluorophore is used to detect DNA-TF complexes on the microarray. The data obtained is normalised and the highest scoring *k*-mers are used to construct PWMs and sequence logos of motifs (Mukherjee *et al.*, 2004).

The motifs are derived using the Seed-and-Wobble algorithm. The Seed-and-Wobble algorithm identifies the most enriched gapped or ungapped *k*-mer. Afterwards the relative nucleotide preference in each position is tested within and outside that specific k-mer. Then finally, a PWM is constructed. Recently other motif finding algorithms were incorporated (Badis *et al.*, 2009).

2.5.2.1. Data Repository for PBM Derived Motifs.

PBM data is deposited into the *Uni*versal *P*BM *R*esource for *O*ligonucleotide-*B*inding *E*valuation (UniPROBE) database where it can be obtained for studies (Robasky and Bulyk, 2010; Chen *et al.*, 2014)

2.5.3. Advantages and Disadvantages of the PBM Technique

PBMs have an advantage in that they are cheaper, faster, highly scalable and results in enriched TF binding sites. PBM also do not have any biases. Using this method we have all possible *k*-mers. In addition, all binding on the array

is purely DNA-TF interaction (Mukherjee *et al.*, 2004; Orenstein & Shamir, 2014).

However, it is good to keep in mind that the genetic code is not random, and not all permutations of a given *k*-mer are present and functional in genomes of living organisms. This method is also performed *in vitro* and may not characterise true binding *in vivo*. Because of the way DNA is wound around histones to form a chromatin complex in eukarytotes, even DNA sequences that are existent in the genome may not be available for binding of TFs and transcription machinery (Field *et al.*, 2011; Mukherjee *et al.*, 2004).

Because PBM is microarray based, it cannot be used to assay TF binding in regions of the genome that contain palindromic sequences or repetitive sequences such as microsatellites and heterochromatin as repeats are masked on arrays. From a PBM experiment we may know the sequence to which the TF binds in the genome, and the sequence of that region but with PBM experiments we do not know the context in which binding occurs (Mukherjee *et al.*, 2004; Orenstein & Shamir, 2014). Another weakness this technique has is a consistent over-representation of some *k*-mers. This phenomenon is called the "sticky" *k*-mers. These "sticky" *k*-mers are background noise but are highly enriched for an unknown reason (Jiang *et al.*, 2013; Orenstein & Shamir, 2014).

3. Comparison of Protein Binding Microarray Derived and ChIP-seq Derived Transcription Factor Binding DNA Motifs

3.1. Problem Statement

It is clear that both techniques have remarkable advantages. Although ChIP-seq is laborious and expensive, it is an *in vivo* technique and is thus expected to yield more accurate models. PBM on the other side is not laborious and is much cheaper. It would be beneficial to investigate whether the convenience of using PBM compromises the accuracy and reliability of yielded TF binding profiles. It is of interest to compare how well the motifs derived using these different methods model actual *in vivo* binding.

3.2. Objectives and Motivation

Our objective in this study was to investigate how well ChIP-seq derived TF binding profiles in the JASPAR Core and PBM derived TF binding profiles model TF binding in living cells. We also investigated how they perform in comparison to each other in living cells.

3.3. Methods

3.3.1. Retrieval of Data and Software Resources

3.3.1.1. ChIP-seq and Human Genome Data

In this study we identified TFs that have ChIP-seq data peak files available for motif profiles that occur both in the JASPAR Core (specifically ones that were ChIP-seq derived) database and the UniPROBE mouse database from the ENCODE database. The human genome was downloaded from an open repository (Refer to Appendix A). We used version 19 of the human genome in this study.

3.3.1.2. Motif Databases

The UniPROBE mouse database were downloaded from the UniPROBE web site. The JASPAR Core database was downloaded from their website. From the JASPAR Core database, only ChIP-seq derived motifs were selected to feature in the database we constructed (Refer to Appendix A for resource urls). This database was named ChIP-seq.meme.

3.3.2. Software Resources

3.3.2.1. Motif Discovery and Motif Enrichment Tools

Motif enrichment tools (CentriMo and AME) were obtained from the MEME website (Refer to Appendix A).

3.3.3. Retrieval of Data Preparation Tools

BEDTools (version 2.17.0) for preparing ChIP-seq data for motif analysis were downloaded from an open source repository (Refer to Appendix A). A bed-widen Perl script written by P. Machanick was downloaded from an open repository (Refer to Appendix A). Fasta-dinucleotide-shuffle was obtained along with MEME suite upon MEME installation.

3.3.4. Retrieval and Preparation of Data for Analysis

3.3.4.1. Extraction of ChIP-seq Peak Regions Sequences from the Human Genome

ChIP-seq peak files were downloaded from ENCODE (Refer to Appendix A). The files were converted from peak files into BED files and subsequently into FASTA files using bed-widen and fastaFromBed. The BED files contain co-ordinates of the ChIP-seq peak regions in the human genome. fastaFromBed used co-ordinates in the bed file to extract the peak regions from the human genome. The bed-widen Perl script trimmed the DNA sequences to 500 bases

wide with the putative binding regions being in the centre of the sequence fragments. The repeats were N-masked, using the Linux *"sed"* command.

Fasta-dinucleotide-shuffle was used to compose control datasets for AME motif enrichment analysis. Fasta-dinucleotide-shuffle generates random sequences from a given ChIP-seq dataset.

3.3.5. Motif Enrichment Analysis

Motif enrichment analysis uses curated database(s) and a ChIP-seq dataset. MEA tools check for enrichment of each motif in the database(s) of interest from a given dataset. Motif enrichment refers to how frequent a motif appears in the dataset compared to background or random sequences. Highly enriched motifs appear most frequently. The search is restricted to the database(s) being queried thus increasing statistical power and sensitivity in detecting under-represented motifs. MEA returns a p-value or E-value for each motif to show the probability of the high enrichment being by chance. It is common practice to accept motifs with an E-value of lower than 0.05 as being truly significantly enriched (Bailey, 2008; Schmidt *et al.*, 2009; Machanick & Bailey, 2011; Bailey and Machanick, 2012; Lesluyes *et al.*, 2014).

3.3.5.1. *Centr*al*i*ty of *Mo*tifs (CentriMo)

Bailey and Machanick (2012) proposed central motif enrichment analysis (CMEA), to overcome the problem with standard motif enrichment analysis and *ab initio* motif discovery approaches that at times failed to correctly identify the motif of the TF of interest. Available *ab initio* motif discovery algorithms assume that the most enriched motif is the site of binding of the TF of interest, which may be incorrect in some cases. For instance, if the immunoprecipitation step in the experiment was unsuccessful the correct motif may not be enriched in the discovered motifs. Additionally, if the TF of interest binds co-operatively with other factors, it is possible that their motifs

appear more represented in the discovered motifs. It is also still possible that the TF of interest does not bind directly to DNA, but binds with factor(s) that are bound to DNA.

CMEA addresses all these issues. CMEA is based on the general observation that binding sites of any given TF in a successful ChIP-seq experiment cluster near the centre of the ChIP-seq peaks, meaning that the best sites for true motifs of any assayed TF have the tendency to lie in the centre of the ChIP-seq peaks. This approached was termed *Centr*al*i*ty of *Mo*tifs (CentriMo).

CentriMo takes sequences of equal length obtained from a ChIP-seq experiment, and at least one motif expressed as a PWM. The length of the ChIP-seq region provided needs to be long enough to contain the motifs and flanking regions as the flanking regions are used as the control in the binomial statistical test. CentriMo tests how enriched the motifs are at the centre in comparison to the flanking regions.

CentriMo returns a graphical representation of the probability that a motif occurs at each position in the length of the ChIP-seq region for each motif, the position of enrichment, an E-value (an adjusted p-value of the significance of the enrichment), total number of matches, region matches and few other informative results. Additionally, it returns the width of the most enriched central region in the ChIP-seq binding sites according to a one-tailed binomial statistical test for each motif and the p-value adjusted for multiple tests. This p-value is referred to as the 'central enrichment p-value' of the motif. CentriMo serves as a visualisation and statistical tool for assessing central enrichment of motifs (Bailey and Machanick, 2012).

In this study we used both MEA and CMEA to investigate how well the motifs in our two databases of interest represent true binding *in vivo*. In ChIP-seq

data that was ChIPped for a particular TF, we expect that the ChIP-seq data should be enriched for sequence patterns which the TF of interest binds. In turn we expect that the enriched sequences match the best TF binding profile in our two databases of interest.

3.3.5.2. *A*nalysis of *M*otif *E*nrichment (AME)

This MEA programme takes in two datasets; the experimental dataset (the ChIP-seq dataset) and the background dataset (a randomly shuffled version of the ChIP-seq dataset) and one or more databases. It then looks for each motif in a database in a DNA dataset. AME treats each position of the sequences in the dataset as a possible TF binding site. It then scores motif enrichment for each motif in the database using a wide variety of methods of testing the scored motif enrichment for significance. Motif enrichment refers to how frequent a motif appears the experimental dataset in comparison to the control sequences. Highly enriched motifs appear most frequently.

The default settings are set such that AME counts the number of significant hits. Significant hits are the number of times a motif in the database matches a sequence pattern in the ChIP-seq dataset with the probability of it happening due to chance (p-value) being lower than the given threshold. Generally a p-value of 0.05 is an acceptable significance threshold, meaning we only accept p-values below 0.05 as significant (Bailey, 2008). A Fisher exact test is performed where the number of binding events in the dataset ChIPped for the TF of interest is compared to the number of binding events in the control sequence set and a p-value is returned for each motif to determine the *p*-value of the count for each motif.

Both CentriMo and AME were executed using the command line in Linux (Ubuntu, 14.04.1 LTS).

3.3.6. Logo Drawing

Logos were drawn using ceqlogo (available along with MEME suite on installation) using the command line. The output pictures were saved in "EPS" format.

3.3.7. Empirical Quality Assessment: Area Under the ROC Curve

We used area under a Receiver Operator Characteristic (ROC) curve as a means to test the quality of the motifs. This method shows the ratio of false positive rate and true positive rate. This is based on rank-ordering the sequences. At first the sequences are ranked according to enrichment, with the most enriched being ranked first. Sequences that are equally enriched are assigned the same rank in such a way that they all take the rank of the sequences that originally ranked lowest. This is done so that the rank for each motif is equal to the number of motifs that have the same or higher enrichment. The list is then set in a descending order. At each rank the number of sequences in that rank or higher that match the motif and those that do not match the motif is counted. These values are then plotted against each other to form the ROC curve (Clarke & Granek, 2003; Xiao *et al.*, 2005).

The area under the ROC curve was computed. An area under the ROC curve value of 0.5, corresponds to the diagonal line, which means there are equal number of false positives as false negatives. This implies that the motif is not a good predictor of TF binding, implying that the motif is by chance. A good quality motif should have a value as close to 1 as possible (Clarke & Granek, 2003; Xiao *et al.*, 2005).

In this study we used resources (Python script) made available by Clarke & Granek, (2003). These resources compute the area under the ROC curve for each TF binding motif model and give the only the values as output. A bash script written by Caleb Kibet was used to prepare ChIP-seq datasets for

empirical quality assessments.

3.3.8. Statistical Analysis

First, the data was tested for normality Shapiro-Wilk normality test because t-test assumes normality. To determine whether there was a significant difference between the area under the ROC curve, a paired t-test was executed and a boxplot presentation of the data was visualised using R. A parametric test could be used on the data since each dataset was continuous. Statistical analysis were performed using R on Rkward console.

3.4. Results and Discussion

There was one case where the motifs for the TF of interest from both databases were not found to be significantly enriched in ChIP-seq data that was supposedly ChIPped for the same TF. Both CentriMo and AME could not find significant enrichment of RXRA motifs (IDs: MA0512.1, UP00053_1 or even UP00053_2) from the ChIP-seq dataset extracted from Gm12878 cell line (Refer to Appendix C and Appendix D, number 4 for PEAK file name). On performing the empirical quality assessment, it became apparent that the motifs may be of poor quality (All results are shown in Appendix D: Table C). However, a more logical explanation would be that the ChIP-seq experiment for this dataset had failed. On further investigation we found that Gm12878 is a B-lymphocyte cell line and RXRA is a retinoic acid receptor that expressed in the the liver, kidneys, epidermis and intestines (Refer to Appendix B: Table A). It is an interesting that this cell line was ChiPped for a TF that is not usually expressed in that particular cell line.

3.4.1. CentriMo

Most of the results showed that the performances of motifs from JASPAR Core and UniPROBE databases do not differ from each other in a noteworthy way. For the first two motifs of the TF of interest, the E-values CentriMo gave were not too different from each other or from 0.00. Nevertheless, of the cases we investigated, the 46 that yielded results, CentriMo showed that in 35 cases JASPAR Core ChIP-seq derived motifs did better, while it was in only 11 cases that PBM derived UniPROBE motifs outperformed ChIP-seq derived JASPAR Core motifs. Evidently, ChIP-seq derived JASPAR Core motifs outperform their UniPROBE counterparts on average (Refer to Table 1).

There is one noteworthy case, whereby CentriMo found that only the JASPAR Core ChIP-seq derived TCF3 motif to be significantly enriched, and did not find any of the UniPROBE TCF3 motifs to be significantly enriched in the

dataset.

The number of total matches returned by CentriMo helps to resolve the differences in the E-values. We expect a higher E-value where there is a higher proportion of centralised matches. There are cases where this was not observed however; EGR1 (for all 3 cell lines), SRF (for 3 cell lines), RXRA (for cell 1 line), MAX (for 8 cell lines), GATA3 (for 3 cell lines), MAFK (all 6 cell lines) and TCF7L2 (for 4 cell lines). In such cases, the lack of correlation in number of matches and E-values can be attributed to low information content or the lack of centrality of the motif in question. Low information content motifs are more likely to find matches due to lack of conservation in each position thus those motifs will always find a specific match. The E-value could be low in a case where the number of matches are very high when the matches were not sharply clustered towards the centre.

3.4.2. AME

Most results returned by AME could not be distinguished from each other because of the inability of the programme to distinguish numbers with an exponent below -300 from each other and 0.00.

The results with E-values above 1e-300, there was no noteworthy difference in the motif performance. However, ChIP-seq derived JASPAR Core motifs seemed to consistently perform better than their UniPROBE counterparts. Of the cases we investigated, the 46 that yielded results, AME showed that in 14 cases JASPAR Core ChIP-seq derived motifs did better, while only it was only in 3 cases that UniPROBE motifs outperformed ChIP-seq derived motifs. We could not distinguish the performance in 29 cases (Refer to Table 1).

3.4.3. Area Under the ROC curve

Empirical quality assessment results showed that ChIP-seq derived JASPAR Core motifs outperformed UniPROBE motifs in 25 cases while UniPROBE motifs outperformed JASPAR Core motifs in 22 cases (Refer to Appendix D: Table C).

3.4.4. Statistical Analysis

The Shapiro-Wilk normality test showed that both the JASPAR Core and UniPROBE motif area under the ROC curve values for each datasets were both normally distributed. The p-values were 2.653e-4 and 4.615e-05 for JASPAR Core and UniPROBE resectively. Thus a parametric paired t-test was carried out on the data.

A two-sided paired t-test showed that there was no significant difference between the average performances of these two groups of motifs with 95% confidence (p-value= 0.4874, DF=47). The mean area under the ROC curve values for ChIP-seq derived motifs in the JASPAR Core and PBM derived motifs in the UniPROBE databases were 0.8357917 and 0.8269792 respectively. We used a boxplot to visualise the distribution of area under the ROC curve values. A boxplot summary revealed that on the distribution of the area under the ROC curve values were similar and that each had an outlier due to the RXRA ChIP-seq dataset whose ChIP-seq experiment, we suspect, had failed. The UniPROBE had an additional outlier, where TCF3 motif was poor quality and had a low information content (See Figure 2 and Table 2).

We then executed further one-sided paired t-tests to determine which models performed better. The null hypothesis is: Their average performances do not differ significantly from each other. The alternative hypothesis was: JASPAR Core motifs perform significantly better than UniPROBE using empirical quality assessment. The results showed that JASPAR Core motifs did not in fact

perform significantly better than UniPROBE motifs (mean of differences=0.0088125, p-value=0.2437, DF=47).

3.4.5. Comparison of CentriMo and AME Results

The GABPA, A549 results shown in Table 2 are representative of cases where CentriMo could give a definite answer of which model was better and AME could not distinguish the performances. The CentriMo distribution of enrichment shows that the performances do not differ much, such that the graphs overlap. Furthermore, these results are representative of cases where CentriMo seemingly failed to find a third motif when AME was seemingly sensitive enough to. However, the third motif AME found to be significantly enriched has a very low information content and is likely to find a match due to having low information content. This is also the case with TCF3 (Refer to Table 2).

There were other cases where CentriMo proved more sensitive than AME, (Represented by EGR1, K562 cell line in Table 2). Although relatively less enriched, the UniPROBE motif was significantly more centrally enriched. A similar case is found in MAFK, H1-hESC cell line (Table 2), where the JASPAR Core motif was relatively less enriched but reletively significantly more enriched.

There were cases where there was a disagreement between CentriMo and AME. For instance, results for GABPA, H1-hESC cell line; using CentriMo we found that ChIP-seq derived JASPAR Core motif did better while AME showed that PBM derived UniPROBE motifs did better. The distribution from CentriMo helps to resolve this disparity. CentriMo finds enrichment of the centre of the dataset, so the PBM motif scored lower because the distribution is not as centred as the ChIP-seq motif. Therefore, although the PBM motif is more enriched in the dataset, the enrichment is not centred. A similar case to this

is GATA3, T-47D results except here CentriMo found PBM motif to perform better. This shows that both MEA methods are not biased towards either PBM motifs or ChIP-seq motifs (Refer to Table 2).

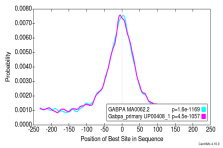
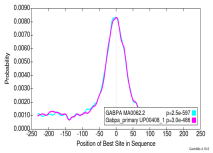
Cases represented by RXRA H1-hESC show how useful the CentriMo distribution graphical representation is in clarifying relatively high (although significant) E-values. The distribution shows the enrichment is not sharply centred for the RXRA motifs in the datasets.

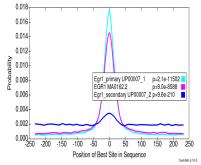
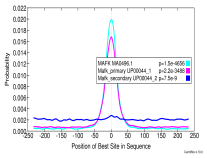
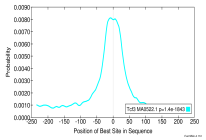
It is noteworthy that in most cases, empirical quality analysis agreed with CentriMo, except for in 14 cases. GABPA, H1-hESC cell line of cases where AME and CentriMo results were not in agreement and the empirical quality assessment confirmed CentriMo results. There were 2 cases, however, where empirical quality analysis agreed with AME when CentriMo and AME gave different answers (See MAX, Gm12878 cell line and ESSRA, HepG2 cell line in Appendix D, Table C).

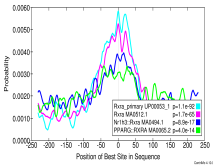
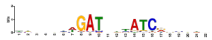
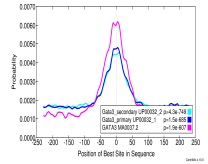
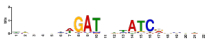
Table 1: Showing the summary of results. Refer to Appendix D, Table C for all the results.

Measure	Better ChIP-seq motif performance	Better PBM motif performance
CentriMo	35	11
AME	14	3
Area under ROC curve	25	22

Table 2: Showing results for MEA analysis and empirical quality assessment. Column 1 shows the names of the TFs of interest. Column 2 shows the cell line from which the ChIP-seq data was extracted and in brackets, the number of sequences in the ChIP-seq dataset. Columns 3-7 shows CentriMo results - 3: Motif ID and Logo representation; 4: CentriMo distribution (the turquoise curve represents the best scoring motif, the purple curve represents the second best and the blue curve the least scoring motif. In other cases there are other colours that represent motifs that do worse than the first three); 5: The rank of the motif in relation to other motifs in the two databases combined; 6: The adjusted p-value (E-value) and in brackets, the number of significant matches; 7: Shows which method modelled binding better according to CentriMo analysis. Columns 8-10 shows AME results – 8: Motif ID and Logo representation; 9: The adjusted p-value (E-value); 10: Shows which method modelled binding better according to AME analysis. Column 11: Shows area under the ROC curve values for the two best scoring motifs. Motifs from JASPAR Core database can be spotted by an ID starting with MA, while UniPROBE can be spotted by an ID starting with UP.

		CentriMo					AME			ROC curve
Motif Name	**Cell line (No. of seqs)**	**Motif ID (Logo)**	**Distribution**	**Rank**	**E-value (Matches)**	**Better *in vivo* modelling**	**Motif ID (Logo)**	**E-value**	**Better *in vivo* modelling**	**Area under ROC curve**
1. GABPA	**A549 (12348)**	**MA0062.2**  **UP00408_1** 		**1** **3**	**7.9e-1167 (10070)** **2.3e-1054 (9287)**	**ChIP-seq**	**MA0062.2**  **UP00408_1**  **UP00408_2** 	**0.00e+0** **0.00e+0** **4.35e-98**	**Cannot be distinguished**	**MA0062.2 0.866** **UP00408_1 0.844**
	H1-hESC (5653)	**MA0062.2**  **UP00408_1** 		**1** **3**	**1.3e-594 (4517)** **1.5e-483 (3771)**	**ChIP-seq**	**UP00408_1**  **MA0062.2**  **UP00408_2** 	**5.759e-251** **1.505e-222** **3.775e-32**	**PBM**	**MA0062.2 0.977** **UP00408_1 0.972**

2. EGR1	**K562 (36997)**	**UP00007_1**  **MA0162.2**  **UP00007_2** 		**1** **3** **23**	**1.0e-11499 (34617)** **4.5e-8585 (35310)** **4.8e-207 (35569)**	**PBM**	**MA0162.2**  **UP00007_1** 	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**UP00007_1 0.951** **MA0162.2 0.925**
3. MAFK	**H1-hESC (11425)**	**MA0496.1**  **UP00044_1**  **UP00044_2** 		**2** **5** **69**	**7.8e-4654 (10435)** **1.1e-3485 (10851)** **3.8e-6 (10643)**	**ChIP-seq**	**MA0496.1**  **UP00044_1** 	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**MA0496.1 0.936** **UP00044_1 0.904**
4. TCF3	**Gm12878 (16021)**	**MA0522.1** 		**1**	**7.2e-1841 (12810)**	**ChIP-seq**	**MA0522.1**  **UP00058_2** 	**0.00e+0** **6.188e-101**	**ChIP-seq**	**MA0522.1 0.849** **UP00058_2 0.356**
5. RXRA	**Gm12878 (1704)**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**MA0512.1 0.482** **UP00053_1 0.472**

	H1-hESC (1306)	**UP00053_1**  **MA0512.1**  **MA0494.1**  **MA0065.2** 		**1** **7** **22** **25**	**5.7e-90 (1135)** **8.7e-63 (1110)** **4.5e-14 (970)** **2.0e-11 (1270)**	**PBM**	**UP00053_1**  **MA0512.1**  **MA0494.1**  **MA0065.2** 	**6.39e-27** **3.39e-22** **1.342e-08** **3.366e-06**	**PBM**	**UP00053_1** **0.757** **MA0512.1** **0.717**
6. GATA3	**T-47D (37199)**	**UP00032_2**  **UP00032_1**  **MA0037.2** 		**8** **9** **10**	**2.2e-746 (30947)** **7.5e-683 (25340)** **9.5e-605 (8399)**	**PBM**	**MA0037.2**  **UP00032_1**  **UP00032_2** 	**1.324e-278** **8.837e-206** **1.336e-172**	**ChIP-seq**	**UP00032_2** **0.638** **UP00032_1** **0.595** **MA0037.2** **0.555**

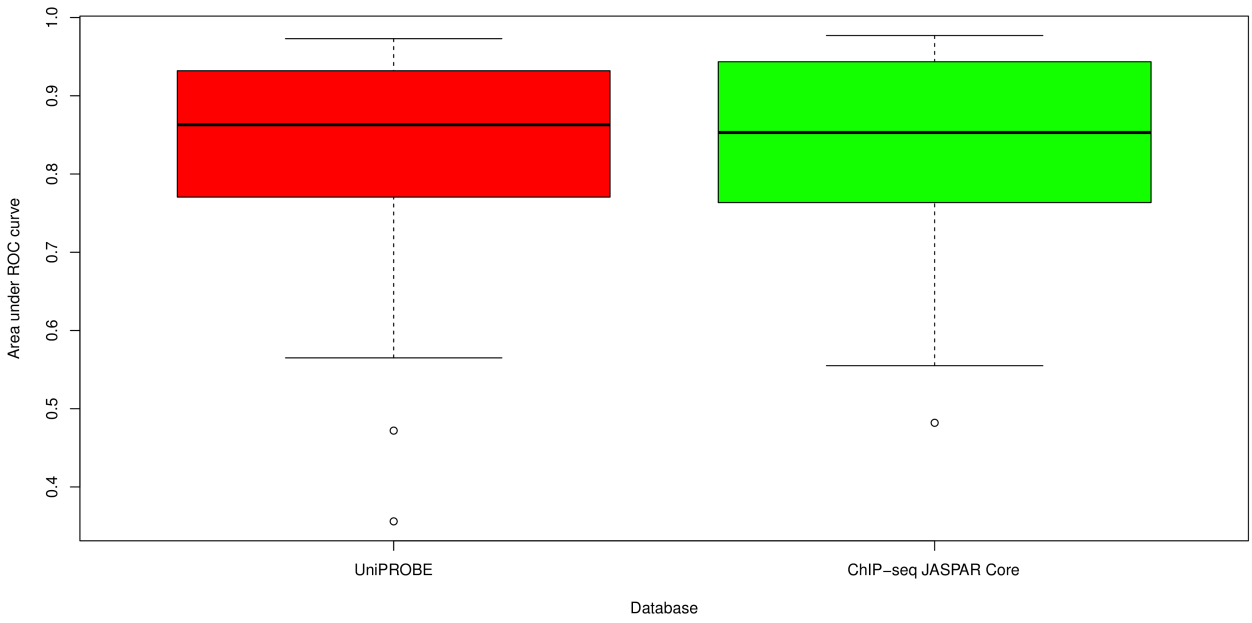


Figure 2: Showing a box plot of the values of the area under the ROC curve for UniPROBE and ChIP-seq derived JASPAR Core motifs.

3.5. Limitations of the Study

The sample size was limited by the number of motifs that were available in both the JASPAR Core and UniPROBE mouse databases. Usually cancerous cell lines are studied and have ChIP-seq data available more readily than normal cell lines. The sample size was further reduced by the lack of availability of ChIP-seq data ChIPped for the TF with binding profiles available in both our databases of interest.

AME was used as one of our means to test the quality of the TF motif models. AME does not report the number of matches that matched the TF motif model significantly in the ChIP-seq dataset. Furthermore, AME cannot distinguish numbers that are smaller than 1e-300 from each other and from 0.00, while CentriMo is able to distinguish this excellently. In a setting where the adjusted p-values (E-values) were smaller than 1e-300, we could not tell which model performed better. This brought confusion because there would be cases where

more than 10 TF binding profiles from both databases would have an E-value of 0.00. In cases like this we cannot be too confident of the binding site, since all motif models seem to match so well.

3.6. Conclusions

ChIP-seq derived JASPAR Core motifs often outranked UniPROBE motifs most of the time, the performance of motifs in JASPAR Core and UniPROBE databases are not too different from each other using motif enrichment analysis. Their E-values do not differ much from 0 and thus from each other.

Although there were cases where each measure would give a different answer with regards to which method gave a better model; it is evident that ChIP-seq derived JASPAR Core motifs and PBM derived UniPROBE motifs model *in vivo* binding with a way that is not significantly different from each other.

Executing statistical analysis on empirical quality assessment results showed that the differences in average performance are not significant. The performance of UniPROBE PBM models was remarkable. However, caution should be exercised when the PBM method being used to construct TF binding profiles, as in some cases the motifs yielded do not model true *in vivo* binding. A more cautious way to achieve accurate models would be to use another method to validate the models derived from PBM.

3.7. Future Work

As the ENCODE database grows, there should be more ChIP-seq data for analysis for a larger sample size study. A larger sample size study would be free from errors that come with analysing a small sample, give more statistical power and would give off more accurate results.

References

Badis G *et al*. (2009) Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*. **324**:1720-1723 and Supporting Online Material page 5

Bailey T. (2008) Discovering Sequence Motifs. (Chapter 12) *Bioinformatics*, **452** 243

Bailey TL & Machanick P. (2012) Inferring Direct DNA Binding from ChIP-seq. *Nucleic Acids Research.* **40**(17):1-10

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW and Noble WS. (2009) MEME SUITE: Tools for Motif Discovery and Searching. *Nucleic Acids Res*. **37**: W2O2

Baumgart S, Ellenrieder V, and Fernandez-Zapico ME. (2013) Oncogenic Transcription Factors: Cornerstones of Inflammation-Linked Pancreatic Carcinogenesis. *Gut*. **62**(2): 310-316

Chen T *et al*. (2014) ChIPseek, A Web-based Analysis Tool for ChIP Data. *BMC Genomics*. **15**(539): 1-13

Clarke ND & Granek JA. (2003) Rank Order Metrics for Quantifying the Association of Sequence Features with Gene Regulation. *Bioinformatics*. **19**(2): 212-218

Cooper SJ, Nathan D. Trinklein ND, Nguyen L & Myers MR. (2007) Serum Response Factor Binding Sites Differ in Three Human Cell Yypes. Genome Research. **17**:136–144

Crick F. (1970). Central Dogma of Molecular Biology. *Nature*. **227**(6): 561-563

Dang CV. (1999) c-Myc Target Genes Involved in Cell Growth, Apoptosis, and Metabolism. *Molecular and Cell Biology*. **19**(1): 1-11

Field Y, Sharon E & Segal E. (2011) How Transcription Factors Identify Regulatory Sites in Genomic Sequence in A Handbook of Transcription Factors (Ed. Hughes TR). *Springer*. **52**: 193-204

Frietze S *et al*. (2012) Cell Type-specific Binding PatternsRreveal that TCF7L2 can be Tethered to the Genome by Association with GATA3. Genome Biology **13**(R52):1-18

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Korbel JO, Emanelson O, Zhang ZD, Weissman S & Snyder M. (2007). What is a gene, post-ENCODE?History & updated definition. *Genome Research*. **17**:669–681

Gordân R, Hartemink AJ & Bulyk ML. (2009) Distinguishing Direct Versus Indirect Transcription Factor-DNA Interactions. *Genome Research*. **19**(11):2090-100

Huang M, Hhengb Y, Liuc C, Shufan Linb & Liub E. (2006) A Small-molecule c-Myc Inhibitor,10058-F4, Induces Cell-cycle arrest, Apoptosis, and Myeloid

Differentiation of Human Acute Myeloid Leukemia. *Experimental Hematology* **34**: 1480–1489

Hunt RW, Mathelier A, del Peso L & Wasserman WW. (2014) Improving Analysis of Transcription Factor Binding Sites Within ChIP-Seq Data Based on Topological Motif Enrichment. *BMC Genomics* **15**(472): 1-19

Hwang J *et al*. (2013) MafK Positively Regulates NF-κB Activity by Enhancing CBP-Mediated p65 Acetylation. *Scientific Reports.* **3**:3242

Jiang B, Liu JS, Bulyk ML. (2013) Bayesian Hierarchical Model of Protein-binding Microarray k-mer Data Reduces Noise and Identifies Transcription Factor Subclasses and Preferred k-mers. *Bioinformatics*. **29**(11):1390–1398

Kitamura T *et al*. (2002) The Forkhead Transcription Factor Foxo1 Links Insulin Signaling to Pdx1 Regulation of Pancreatic β-cell Growth. *J Clin Invest.* **110**(12):1839-1847

Lesluyes T, Johnson J, Machanick P and Bailey TL. (2014) CentriMo: Differential Motif Enrichment Analysis of Paired ChIP-seq experiments. *BMG Genomics*. **15**(752):1-7

Machanick P and Bailey TL. (2011) MEME-ChIP: Motif Analysis of Large DNA Datasets. *Bioinformatics*. **27**(2): 1696–1697

Marconette CN *et al*. (2013) Integrated Transcriptomic and Epigenomic Analysis of Primary Human Lung Epithelial Cell Differentiation. *PLOS Genetics*. **9**(6): 1-15

Mathelier *et al*. (2014) Jaspar 2014: the Greatly Expanded Open-access Database of Transcription Factor Binding Profiles. *Nucleic Acids Res.* **42**(Database issue):D142-147

Matthew *et al*. (2013) Evaluation of Methods for Modeling Transcription Factor Sequence Specificity. *Nature Biotechnology.* **31**: 126–134

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, & Bulyk ML. (2004) Rapid Analysis of the DNA Binding Specificities of Factors with DNA Microarrays. *Nature Genetics.* **36**(12): 1331–1339

Nichol JN, Assouline S, & Miller WH. (2013) The Etiology of Acute Leukemia in: Neoplastic Diseases of the Blood (5th Edition, Ed: Wiernik PH, Goldman JM, Dutcher JP & Kyle RA) *Springer*. 177-198

Ohno S. (1972). So Much "junk" DNA in: Our Genome in Evolution of Genetic Systems. (Ed. HH Smith) *Brookhaven* Symposium. **23**:366-370

Orenstein Y & Shamir R. (2014) A Comparative Analysis of Transcription Factor Binding Models Learned from PBM, HT-SELEX and ChIP data. *Nucleic Acid Research*. **42**(8):e63

Park JP. (2009) ChIP–seq: Advantages & Challenges of a Maturing Technology. *Nature Reviews Genetics*. **10**: 669-679

Pearson H. (2006) Genetics: What is a gene? *Nature*. **441**:398–401

Plutzky J. (2011) The PPAR-RXR Transcriptional Complex in the Vasculature: Energy in the Balance. *Circulation Research*. **108**(8)1002-1016

Portales-Casamar E *et al*. (2010). Jaspar 2010: the Greatly Expanded Open-access Database of Transcription Factor Binding Profiles. *Nucleic Acids Res.* **38**(Database issue):D105-110

Qu H & Fang X. (2013) A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. *Genomics Proteomics Bioinformatics*. **11**:135–141

Reynolds N, O'Shaughnessy A & Hendrich B. (2013) Transcriptional Repressors: Multifaceted Regulators of Gene Expression. *Development*. **140**: 505-512

Ridinger-Saison M *et al*. (2012) Spi-1/PU.1 Activates Transcription Through Clustered DNA Occupancy in Erythroleukemia. *Nucleic Acids Research* **40**(18): 8927–8941

Robasky K & Bulyk ML. (2010) UniPROBE, Update 2011: Expanded Content and Search Tools in the Online Database of Protein-Binding Microarray Data on Protein–DNA Interactions. *Nucleic Acids Research.* **3**: D124-128

Rosmarin AG, Resendes KK, Yang Z, McMillan JN, & Flemin SL. (2004) GA-binding Protein Transcription Factor: A Review of GABP as an Integrator of Intracellular Signaling and Protein–Protein Interactions . Blood Cells, Molecules, and Diseases **32**:143–154

Santolini M, Mora T, Hakim V. (2014) A General Pairwise Interaction Model Provides an Accurate Description of In Vivo Transcription Factor Binding Sites. *PLoS ONE*. **9**(6): 99015

Schmidt, Wilson MD, Spyrou C , Brown GD, Hadfield J, Odom DT. (2009) ChIP-seq: Using High- throughput Sequencing to Discover Protein–DNA Interactions. *Methods.* **48**: 240–248

Stormo GD. (2000) DNA Binding Sites: Representation and Discovery. *Bioinformatics*. **16**(1), 16-23

Thiel G & Cibelli G. (2002) Regulation of Life and Death by the Zinc Finger Transcription Factor Egr-1. *Journal of Cellular Physiology*. **193**:287–292

Tompa M *et al*. (2005) Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites. *Nature Biotechnology*, **23**(1): 137–145

Tremblay M, Tremblay CS, Herblot S, Aplan PD, Hébert J, Perreault C, and Hoang T. (2012) Modeling T-cell Acute Lymphoblastic Leukemia Induced by the SCL and LMO1 Oncogenes. *Genes & development* **24**: 1093–1105

van Waveren C and Morae CT. (2008) Transcriptional Co-expression and Co-

Regulation of Genes Coding for Components of the Oxidative Phosphorylation System. *BMC Genomics.* **9**(18):1-15

Venter JC *et al*. (2001) The Sequence of the Human Genome. *Nature*. **291**(5507): 1304-1351

Walhout AJM, Gubbels JM, Bernards R, van der Vliet PC & Timmers HTM. (1997) c-Myc/Max Heterodimers Bind Cooperatively to the E-Box Sequences Located in the First Intron of the Rat Ornithine Decarboxylase (ODC) Gene. *Nucleic Acids Research*. **25**(8):1493-1501

Wang D, Qiu C, Zhang H, Wang J, Cui Q, Yin Y. (2010) Human MicroRNA Oncogenes and Tumor Suppressors Show Significantly Different Biological Patterns: From Functions to Targets. *PLoS ONE* **5**(9): 1-7

Weinmann AS, Farnham PJ. (2002) Identification of Unknown Target Genes of Human Transcription Factors Using Chromatin Immunoprecipitation. *Methods.* **26:** 37-47

Wingeder E, Schoeps T & Dönitz J. (2013) TFClass: an Expandable Hierarchical Classification of Human Transcription Factors. *Nucleic Acids Research*. **41**: D165-D170

Xiao L, Noll DM, Lieb JD & Clarke ND. (2005) DIP-chip: Rapid and Accurate Determination of DNA-binding Specificity. *Genome Research*. **15**:421-427

Xiong J. (2006) Essential Bioinformatics. *Cambridge University Press*. 114

Zhong S, He X, Bar-Joseph Z. (2013) Predicting Tissue Specific Transcription Factor Binding Sites. *BMC Genomics*. **14**(796):1-14

Appendices

Appendix A

Resources used in this study were downloaded from the url states below:

BED Tools

https://code.google.com/p/bedtools/downloads/detail?name=BEDTools.v2.17.0.tar.gz

ChIP-seq Data

http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform

Human Genome

http://homes.cs.ru.ac.za/philip/data/hg19/

UniPROBE Database

http://thebrain.bwh.harvard.edu/uniprobe/

JASPAR Core Databases

http://jaspar.genereg.net

MEME Website

http://meme.nbcr.net/meme

Perl Script (fastaFromBed)

http://homes.cs.ru.ac.za/philip/data/scripts/

Resources for Calculating the Area Under the ROC Curve

ftp://ftp.bs.jhmi.edu/users/nclarke/MNCP/

Appendix B

Supplementary Information

Table A: Showing description of cell lines used in this study (As described on ENCODE)

Cell Line ID	Cell Line Description
1. A549	Epithelial cell line derived from a lung carcinoma tissue
2. GM12878	B-lymphocyte
3. H1-hESC	Embryonic stem cells inner cell mass
4. HEK293	Embryonic kidney, cells contain Adenovirus 5 DNA
5. HeLa-S3	Cervical carcinoma, ectoderm
6. HepG2	Hepatocellular carcinoma, endoderm
7. HCT-116	Colorectal carcinoma, colon cancer, endoderm
8. HUVEC	Umbilical vein endothelial cells, mesoderm
9. IMR90	Fetal lung fibroblasts
10. K562	Established from a patient with chronic myelogenous leukemia
11. MCF-7	Mammary gland, adenocarcinoma, ectoderm
12. NB4	Acute promyelocytic leukemia cell line.
13. PANC-1	Pancreatic carcinoma
14. SH-SY5Y	Neuroblastoma clonal subline of the neuroepithelioma cell line
15. T-47D	Epithelial cell line derived from a mammary ductal carcinoma

Table B: Showing transcription factor details.

TF name	Full name **Alternative name	Function	Reference
1. GABPA	GA Binding Protein	-Activates genes that control cell cycle, apoptosis, differentiation, cell cycle progression and embryogenesis -Found in myeloid and muscle cells	Rosmarin *et al.* 2004
2. EGR1	Early Growth Response protein 1 **zif268, Krox24, TIS8	-Activates the expression of genes needed for mitogenesis and differentiation. -An early gene in fibroblasts, neuronal cells, lymphoid cells	Thiel and Cibelli 2002
3. SRF	Serum Response Factor	-SRF regulates a number of genes necessary for early development, cell cycle regulation, apoptosis, cell growth, and differentiation. -Found in muscles and neurons	Cooper *et al.* 2007
4. HNF4A	Hepatocyte Nuclear Factor 4 alpha **NR2A1 (nuclear	-May play a role in development of the liver, kidney and intestines -Involved in lung regeneration	Marconette *et al.*, 2013

	receptor subfamily 2, group A, member 1)		
RXRA	Retinoid X Receptor alpha	-Involved in the regulation of energy balance, glucose homeostasis as well as fatty acid handling and storage -Expressed in the liver, kidneys, epidermis, and intestines	Pluzky, 2011
6. TCF3	Transcription Factor 3	-The TCF3 gene encodes two alternatively splices basic helix-loop-helix (bHLH) TFs E12 and E47 -These TFs play a role in early B-cell lineage development	Nichol *et al.* 2013
7. FoxA2	Forkhead box protein A2 **Hepatocyte Nuclear Factor 3-beta (HNF-3B)	-Involved in embryonic development, development of liver, pancreas, pancreatic beta-cells and lungs -Regulates fat metabolism -Maintains glucose homeostasis	Kitamura *et al.*, 2002
8. MAX	MYC-associated factor X	-Involved in cell proliferation, inhibition of differentiation and apoptosis along with c-MYC	Walhout *et al.*, 1997
9. ESRRA	Estrogen related receptor alpha **ERRA	-Involved in mitochondrial gene regulation alongside other co-fctors	van Waveren and Morae, 2008
10. GATA3	GATA binding protein 3	-Regulates of T-cell development and in endothelial cell biology	Frietze *et al.* 2012
11. MAFK	Musculoaponeurotic fibrosarcoma oncogene	-Modulates NF-κB activity	Hwang *et al.* 2013
12. TCF7L2	Transcription factor 7-like 2 **TCF4	-Highly up-regulated in several types of human cancer, (colon, liver, breast, and pancreatic cancer)	Frietze *et al.* 2012

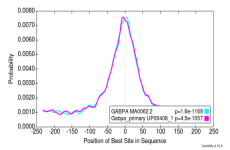
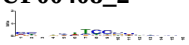
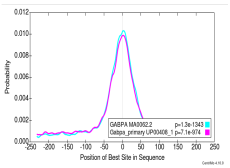
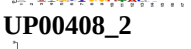
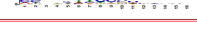
Appendix C

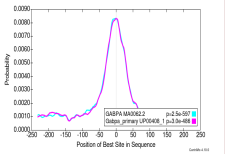
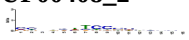
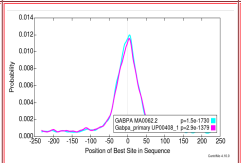
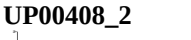
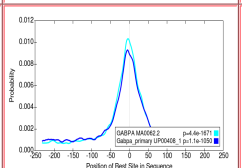
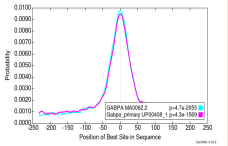
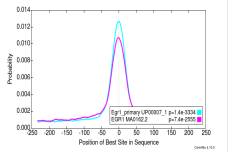
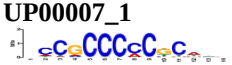
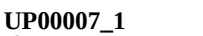
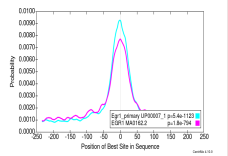
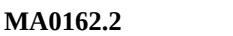
Names of used ChIP-seq peak files

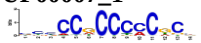
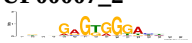
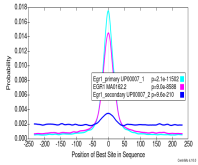
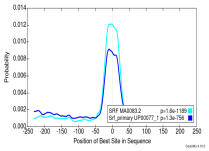
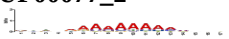
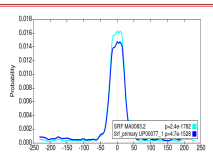
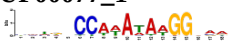
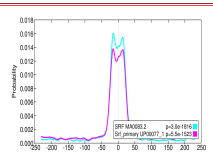
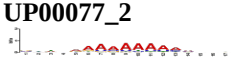
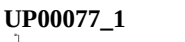
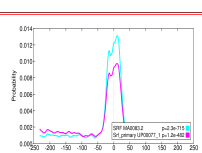
1. wgEncodeAwgTfbsHaibA549GabpV0422111Etoh02UniPk.narrowPeak
2. wgEncodeAwgTfbsHaibGm12878Egr1Pcr2xUniPk.narrowPeak
3. wgEncodeAwgTfbsHaibGm12878GabpPcr2xUniPk.narrowPeak
4. wgEncodeAwgTfbsHaibGm12878RxraPcr1xUniPk.narrowPeak
5. wgEncodeAwgTfbsHaibGm12878SrfPcr2xUniPk.narrowPeak
6. wgEncodeAwgTfbsHaibGm12878Tcf3Pcr1xUniPk.narrowPeak
7. wgEncodeAwgTfbsHaibH1hescEgr1V0416102UniPk.narrowPeak
8. wgEncodeAwgTfbsHaibH1hescGabpPcr1xUniPk.narrowPeak
9. wgEncodeAwgTfbsHaibH1hescRxraV0416102UniPk.narrowPeak
10. wgEncodeAwgTfbsHaibH1hescSrfPcr1xUniPk.narrowPeak
11. wgEncodeAwgTfbsHaibHelas3GabpPcr1xUniPk.narrowPeak
12. wgEncodeAwgTfbsHaibHepg2Foxa2sc6554V0416101UniPk.narrowPeak
13. wgEncodeAwgTfbsHaibHepg2GabpPcr2xUniPk.narrowPeak
14. wgEncodeAwgTfbsHaibHepg2Hnf4asc8987V0416101UniPk.narrowPeak
15. wgEncodeAwgTfbsHaibHepg2RxraPcr1xUniPk.narrowPeak
16. wgEncodeAwgTfbsHaibHepg2SrfV0416101UniPk.narrowPeak
17. wgEncodeAwgTfbsHaibK562Egr1V0416101UniPk.narrowPeak
18. wgEncodeAwgTfbsHaibK562GabpV0416101UniPk.narrowPeak
19. wgEncodeAwgTfbsHaibK562MaxV0416102UniPk.narrowPeak
20. wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak
21. wgEncodeAwgTfbsHaibT47dGata3sc268V0416102Dm002p1hUniPk.narrowPeak
22. wgEncodeAwgTfbsSydhA549MaxIggrabUniPk.narrowPeak
23. wgEncodeAwgTfbsSydhGm12878MaxIggmusUniPk.narrowPeak
24. wgEncodeAwgTfbsSydhH1hescMafkIggrabUniPk.narrowPeak
25. wgEncodeAwgTfbsSydhH1hescMaxUcdUniPk.narrowPeak
26. wgEncodeAwgTfbsSydhHct116Tcf7l2UcdUniPk.narrowPeak
27. wgEncodeAwgTfbsSydhHek293Tcf7l2UcdUniPk.narrowPeak
28. wgEncodeAwgTfbsSydhHelas3MafkIggrabUniPk.narrowPeak
29. wgEncodeAwgTfbsSydhHelas3MaxIggrabUniPk.narrowPeak
30. wgEncodeAwgTfbsSydhHelas3Tcf7l2c9b92565UcdUniPk.narrowPeak
31. wgEncodeAwgTfbsSydhHelas3Tcf7l2UcdUniPk.narrowPeak
32. wgEncodeAwgTfbsSydhHepg2ErraForsklnUniPk.narrowPeak
33. wgEncodeAwgTfbsSydhHepg2Hnf4aForsklnUniPk.narrowPeak
34. wgEncodeAwgTfbsSydhHepg2Mafkab50322IggrabUniPk.narrowPeak
35. wgEncodeAwgTfbsSydhHepg2Mafksc477IggrabUniPk.narrowPeak
36. wgEncodeAwgTfbsSydhHepg2MaxIggrabUniPk.narrowPeak
37. wgEncodeAwgTfbsSydhHepg2Tcf7l2UcdUniPk.narrowPeak
38. wgEncodeAwgTfbsSydhHuvecMaxUniPk.narrowPeak
39. wgEncodeAwgTfbsSydhImr90MafkIggrabUniPk.narrowPeak
40. wgEncodeAwgTfbsSydhK562Mafkab50322IggrabUniPk.narrowPeak
41. wgEncodeAwgTfbsSydhK562MaxIggrabUniPk.narrowPeak
42. wgEncodeAwgTfbsSydhMcf7Gata3sc269UcdUniPk.narrowPeak
43. wgEncodeAwgTfbsSydhMcf7Gata3UcdUniPk.narrowPeak
44. wgEncodeAwgTfbsSydhMcf7Tcf7l2UcdUniPk.narrowPeak
45. wgEncodeAwgTfbsSydhNb4MaxUniPk.narrowPeak
46. wgEncodeAwgTfbsSydhPanc1Tcf7l2UcdUniPk.narrowPeak
47. wgEncodeAwgTfbsSydhShsy5yGata3sc269sc269UcdUniPk.narrowPeak

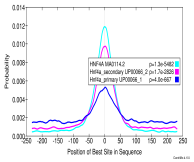
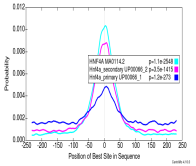
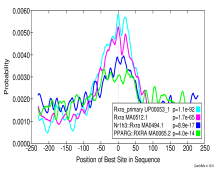
Appendix D

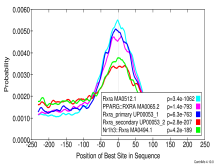
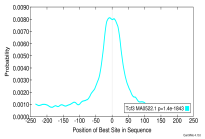
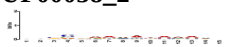
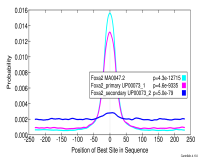
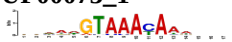
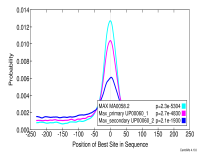
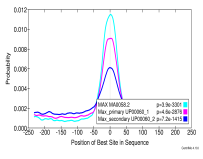
Table C: Showing results for MEA analysis and empirical quality assessment. Column 1 shows the names of the TFs of interest. Column 2 shows the cell line from which the ChIP-seq data was extracted and in brackets, the number of sequences in the ChIP-seq dataset. Columns 3-7 shows CentriMo results - 3: Motif ID and Logo representation; 4: CentriMo distribution (the turquoise curve represents the best scoring motif, the purple curve represents the second best and the blue curve the least scoring motif. In other cases there are other colours that represent motifs that do worse than the first three); 5: The rank of the motif in relation to other motifs in the two databases combined; 6: The adjusted p-value (E-value) and in brackets, the number of significant matches; 7: Shows which method modelled binding better according to CentriMo analysis. Columns 8-10 shows AME results – 8: Motif ID and Logo representation; 9: The adjusted p-value (E-value); 10: Shows which method modelled binding better according to AME analysis. Column 11: Shows area under the ROC curve values for the two best scoring motifs. Motifs from JASPAR Core database can be spotted by an ID starting with MA, while UniPROBE can be spotted by an ID starting with UP.

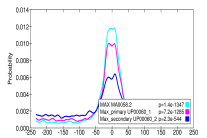
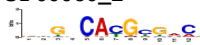
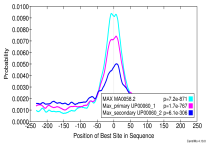
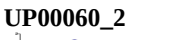
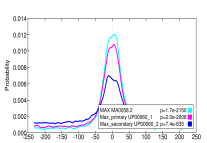
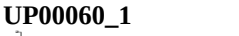
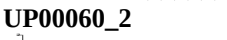
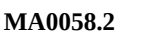
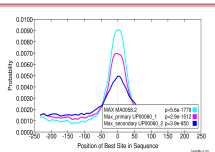
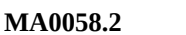
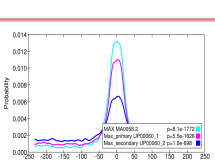
		CentriMo					AME			ROC curve
Motif Name	**Cell line (No. of seqs)**	**Motif ID (Logo)**	**Distribution**	**Rank**	**E-value (Matches)**	**Better *in vivo* modelling**	**Motif ID (Logo)**	**E-value**	**Better *in vivo* modelling**	**Area under ROC curve**
1. GABPA	**A549 (12348)**	**MA0062.2**  **UP00408_1** 		**1** **3**	**7.9e-1167 (10070)** **2.3e-1054 (9287)**	**ChIP-seq**	**MA0062.2**  **UP00408_1**  **UP00408_2** 	**0.00e+0** **0.00e+0** **4.35e-98**	**Cannot be distinguished**	**MA0062.2 0.866** **UP00408_1 0.844**
	Gm12878 (6566)	**MA0062.2**  **UP00408_1** 		**1** **3**	**6.3e-1341 (6087)** **3.6e-971 (5260)**	**ChIP-seq**	**MA0062.2**  **UP00408_1**  **UP00408_2** 	**0.00e+0** **0.00e+0** **4.15e-73**	**Cannot be distinguished**	**MA0062.2 0.951** **UP00408_1 0.922**

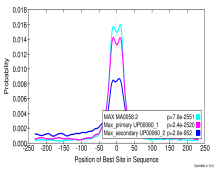
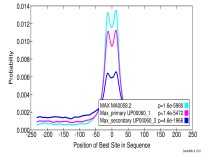
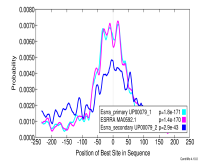
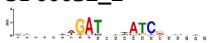
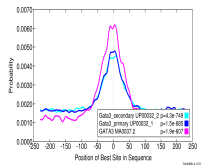
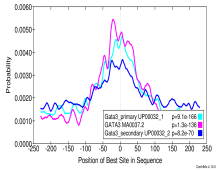
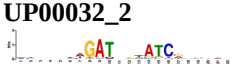
	H1-hESC (5653)	**MA0062.2**  **UP00408_1** 		**1** **3**	**1.3e-594 (4517)** **1.5e-483 (3771)**	**ChIP-seq**	**UP00408_1**  **MA0062.2**  **UP00408_2** 	**5.759e-251** **1.505e-222** **3.775e-32**	**PBM**	**MA0062.2 0.977** **UP00408_1 0.972**
	HeLa-S3 (6761)	**MA0062.**  **UP00408_1** 		**1** **3**	**7.4e-1728 (6339)** **1.5e-1376 (5713)**	**ChIP-seq**	**MA0062.2**  **UP00408_1**  **UP00408_2** 	**0.00e+0** **0.00e+0** **1.676e-60**	**Cannot be distinguished**	**MA0062.2 0.974** **UP00408_1 0.965**
	HepG2 (10109)	**MA0062.2**  **UP00408_1** 		**1** **4**	**2.2e-1668 (9043)** **5.6e-1048 (7387)**	**ChIP-seq**	**MA0062.2**  **UP00408_1** **UP00408_2** 	**0.00e+0** **0.00e+0** **5.118e-92**	**Cannot be distinguished**	**MA0062.2 0.955** **UP00408_1 0.935**
	K562 (14393)	**MA0062.2**  **UP00408_1** 		**1** **3**	**2.3e-2052 (11815)** **2.2e-1566 (10221)**	**ChIP-seq**	**MA0062.2**  **UP00408_1**  **UP00408_2** 	**0.00e+0** **0.00e+0** **5.369e-104**	**Cannot be distinguished**	**MA0062.2 0.966** **UP00408_1 0.952**
2. EGR1	**Gm12878 (16331)**	**UP00007_1**  **MA0162.2** 		**1** **3**	**6.8e-3332 (14751)** **3.7e-2552 (15521)**	**PBM**	**MA0162.2**  **UP00007_1** 	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**UP00007_1 0.942** **MA0162.2 0.913**
	H1-hESC (8743)	**UP00007_1**  **MA0162.2** 		**1** **3**	**2.7e-1120 (7715)** **9.0e-792 (8222)**	**PBM**	**MA0162.2**  **UP00007_1** 	**0.00e+0** **2.971e-172**	**ChIP-seq**	**UP00007_1 0.888** **MA0162.2 0.861**

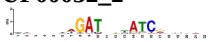
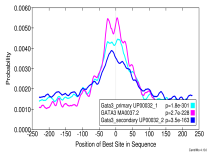
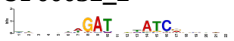
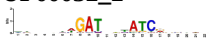
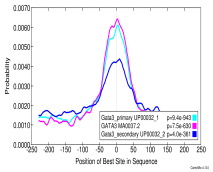
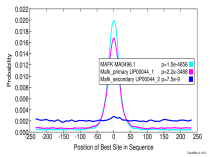
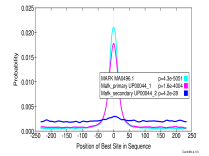
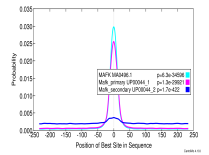
	K562 (36997)	**UP00007_1**  **MA0162.2**  **UP00007_2** 		**1** **3** **23**	**1.0e-11499 (34617)** **4.5e-8585 (35310)** **4.8e-207 (35569)**	**PBM**	**MA0162.2**  **UP00007_1** 	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**UP00007_1 0.951** **MA0162.2 0.925**
3. SRF	**Gm12878 (8544)**	**MA0083.2**  **UP00077_1** 		**1** **2**	**8.0e-1187 (4616)** **6.6e-754 (5802)**	**ChIP-seq**	**MA0083.2**  **UP00077_1**  **UP00077_2** 	**8.068e-244** **2.157e-227** **1.946e-105**	**ChIP-seq**	**MA0083.2 0.724** **UP00077_1 0.681**
	H1-hESC (5105)	**MA0083.2**  **UP00077_1** 		**1** **2**	**1.2e-1759 (3621)** **2.3e-1525 (3950)**	**ChIP-seq**	**MA0083.2**  **UP00077_1**  **UP00077_2** 	**0.00e+0** **0.00e+0** **1.82e-48**	**Cannot be distinguished**	**MA0083.2 0.734** **UP00077_1 0.700**
	HepG2 (5314)	**MA0083.2**  **UP00077_1** 		**1** **2**	**1.5e-1813 (3910)** **2.7e-1520 (4337)**	**ChIP-seq**	**MA0083.2**  **UP00077_1**  **UP00077_2** 	**0.00e+0** **0.00e+0** **5.173e-58**	**Cannot be distinguished**	**MA0083.2 0.828** **UP00077_1 0.812**
	K562 (4717)	**MA0083.2**  **UP00077_1** 		**1** **2**	**1.1e-712 (1497)** **6.1e-480 (1418)**	**ChIP-seq**	**MA0083.2**  **UP00077_1**  **UP00077_2** 	**2.823e-175** **5.029e-163** **6.814e-49**	**ChIP-seq**	**MA0083.2 0.608** **UP00077_1 0.565**

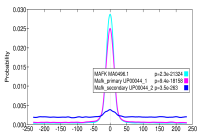
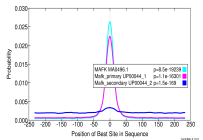
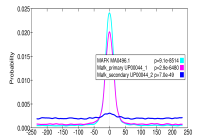
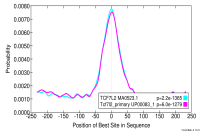
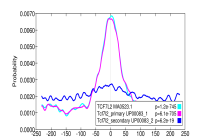
4. HNF4A	**HepG2 (20805)**	**MA0114.2**  **UP00066_2**  **UP00066_1** 		**1** **3** **12**	**6.6e-5480 (20072)** **8.7e-2824 (16522)** **2.0e-664 (17031)**	**ChIP-seq**	**MA0114.2**  **UP00066_2**  **UP00066_1** 	**0.00e+0** **0.00e+0** **8.696e-33**	**Cannot be distinguished**	**MA0114.2 0.958** **UP00066_2 0.842**
	Same cell line (11130)	**MA0114.2**  **UP00066_2**  **UP00066_1** 		**1** **3** **12**	**5.5e-2546 (10732)** **1.8e-1412 (8837)** **2.0e-664 (9121)**	**ChIP-seq**	**MA0114.2**  **UP00066_2**  **UP00066_1** 	**0.00e+0** **0.00e+0** **1.156e-12**	**Cannot be distinguished**	**MA0114.2 0.941** **UP00066_2 0.818**
5. RXRA	**Gm12878 (1704)**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**MA0512.1 0.482** **UP00053_1 0.472**
	H1-hESC (1306)	**UP00053_1**  **MA0512.1**  **MA0494.1**  **MA0065.2** 		**1** **7** **22** **25**	**5.7e-90 (1135)** **8.7e-63 (1110)** **4.5e-14 (970)** **2.0e-11 (1270)**	**PBM**	**UP00053_1**  **MA0512.1**  **MA0494.1**  **MA0065.2** 	**6.39e-27** **3.39e-22** **1.342e-08** **3.366e-06**	**PBM**	**UP00053_1 0.757** **MA0512.1 0.717**

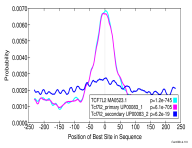
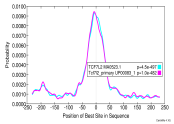
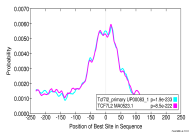
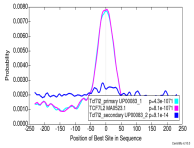
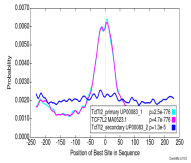
	HepG2 (17063)	**MA0512.1**  **MA0065.2**  **UP00053_1**  **UP00053_2**  **MA0494.1** 		**5** **7** **8** **22** **28**	**1.7e-1059 (15075)** **7.0e-791 (15736)** **3.2e-760 (14604)** **1.4e-204 (16605)** **2.1e-186 (12526)**	**ChIP-seq**	**MA0065.2**  **MA0512.1**  **UP00053_1**  **MA0494.1**  **UP00053_2** 	**0.00e+0** **0.00e+0** **4.741e-195** **1.776e-50** **0.04493**	**Cannot be distinguished**	**UP00053_1 0.784** **MA0512.1 0.775** **MA0065.2 0.731**
6. TCF3	**Gm12878 (16021)**	**MA0522.1** 		**1**	**7.2e-1841 (12810)**	**ChIP-seq**	**MA0522.1**  **UP00058_2** 	**0.00e+0** **6.188e-101**	**ChIP-seq**	**MA0522.1 0.849** **UP00058_2 0.356**
7. FoxA2	**HepG2 (40989)**	**MA0047.2**  **UP00073_1**  **UP00073_2** 		**1** **3** **71**	**2.2e-12712 (38629)** **2.3e-9332 (38488)** **2.5e-76 (40201)**	**ChIP-seq**	**MA0047.2**  **UP00073_1**  **UP00073_2** 	**0.00e+0** **0.00e+0** **3.374e-196**	**Cannot be distinguished**	**MA0047.2 0.917** **UP00073_1 0.866**
8. MAX	**K562 (46171)**	**MA0058.2**  **UP00060_1**  **UP00060_2** 		**3** **4** **9**	**1.1e-5301 (22891)** **1.3e-4827 (32345)** **1.1e-1927 (40722)**	**ChIP-seq**	**MA0058.2**  **UP00060_1**  **UP00060_2** 	**0.00e+0** **0.00e+0** **2.004e-166**	**Cannot be distinguished**	**UP00060_1 0.792** **MA0058.2 0.724**
	Same cell line (31436)	**MA0058.2**  **UP00060_1**  **UP00060_2** 		**1** **4** **8**	**2.0e-3298 (14097)** **2.3e-2873 (20915)** **3.6e-1412 (27582)**	**ChIP-seq**	**MA0058.2**  **UP00060_1**  **UP00060_2** 	**0.00e+0** **0.00e+0** **1.516e-108**	**Cannot be distinguished**	**UP00060_1 0.918** **MA0058.2 0.851**

	A549 (9881)	MA0058.2  UP00060_1  UP00060_2 		2 3 8	7.2e-1345 (5084) 3.6e-1282 (7169) 1.1e-541 (8855)	ChIP-seq	MA0058.2  UP00060_1  UP00060_2 	0.00e+0 0.00e+0 2.755e-50	Cannot be distinguished	UP00060_1 0.918 MA0058.2 0.854
	Gm12878 (12542)	MA0058.2  UP00060_1  UP00060_2 		2 4 8	3.6e-868 (5330) 8.7e-765 (8275) 3.0e-303 (11015)	ChIP-seq	UP00060_1  MA0058.2  UP00060_2 	1.93e-300 1.958e-265 7.691e-30	PBM	UP00060_1 0.869 MA0058.2 0.794
	H1-hESC (11129)	MA0058.2  UP00060_1  UP00060_2 		1 3 9	8.8e-2148 (7261) 1.0e-2005 (9150) 3.7e-832 (10163)	ChIP-seq	MA0058.2  UP00060_1  UP00060_2 	0.00e+0 0.00e+0 4.066e-123	Cannot be distinguished	UP00060_1 0.910 MA0058.2 0.855
	HeLa-S3 (29647)	MA0058.2  UP00060_1  UP00060_2 		1 4 11	2.8e-1776 (11355) 1.4e-1509 (18190) 2.0e-847 (25864)	ChIP-seq	MA0058.2  UP00060_1  UP00060_2 	0.00e+0 0.00e+0 2.004e-166	Cannot be distinguished	UP00060_1 0.958 MA0058.2 0.917
	HepG2 (11854)	MA0058.2 UP00060_1  UP00060_2 		1 3 8	4.1e-1769 (6060) 2.8e-1623 (8418) 5.2e-696 (10541)	ChIP-seq	MA0058.2  UP00060_1  UP00060_2 	0.00e+0 0.00e+0 3.23e-62	Cannot be distinguished	UP00060_1 0.964 MA0058.2 0.927

	HUVEC (9122)	**MA0058.2**  **UP00060_1**  **UP00060_2** 		**1** **2** **9**	**3.5e-2548 (6079)** **1.2e-2517 (7531)** **1.0e-949 (8429)**	**ChIP-seq**	**MA0058.2**  **UP00060_1**  **UP00060_2** 	**0.00e+0** **0.00e+0** **9.684e-96**	**Cannot be distinguished**	**UP00060_1 0.910** **MA0058.2 0.847**
	NB4 (34659)	**MA0058.2**  **UP00060_1**  **UP00060_2** 		**1** **3** **9**	**7.9e-5978 (18975)** **7.1e-5470 (25676)** **2.3e-1963 (30755)**	**ChIP-seq**	**MA0058.2**  **UP00060_1**  **UP00060_2** 	**0.00e+0** **0.00e+0** **4.063e-161**	**Cannot be distinguished**	**UP00060_1 0.973** **MA0058.2 0.953**
9. ESRRA	**HepG2 (1177)**	**UP00079_1**  **MA0592.1**  **UP00079_2** 		**2** **3** **12**	**8.8e-169 (1040)** **6.9e-168 (1037)** **1.4e-40 (971)**	**PBM**	**MA0592.1**  **UP00079_1**  **UP00079_2** 	**1.30e-97** **4.615e-96** **3.659e-07**	**ChIP-seq**	**MA0592.1 0.835** **UP00079_1 0.798**
10. GATA3	**T-47D (37199)**	**UP00032_2**  **UP00032_1**  **MA0037.2** 		**8** **9** **10**	**2.2e-746 (30947)** **7.5e-683 (25340)** **9.5e-605 (8399)**	**PBM**	**MA0037.2**  **UP00032_1**  **UP00032_2** 	**1.324e-278** **8.837e-206** **1.336e-172**	**ChIP-seq**	**UP00032_2 0.638** **UP00032_1 0.595** **MA0037.2 0.555**
	MCF-7 (6081) **Same cell line**	**UP00032_1**  **MA0037.2**  **UP00032_2**  **UP00032_1**		**5** **7** **11**	**4.5e-163 (9183)** **6.8e-134 (3718)** **4.1e-67 (10402)**	**PBM**	**MA0037.2**  **UP00032_1**  **UP00032_2** 	**0.00e+0** **1.135e-273** **1.217e-143**	**ChIP-seq**	**UP00032_1 0.607** **MA0037.2 0.604**

	(12077)	**MA0037.2**  **UP00032_2** 		**6** **7** **9**	**8.9e-299 (4664)** **1.4e-225 (2040)** **1.8e-160 (5297)**	**PBM**	**MA0037.2**  **UP00032_1**  **UP00032_2** 	**2.038e-120** **1.144e-68** **2.567e-41**	**ChIP-seq**	**UP00032_1 0.641** **MA0037.2 0.575**
	SH-SY5Y (15879)	**UP00032_1**  **MA0037.2**  **UP00032_2** 		**5** **7** **9**	**4.7e-940 (13495)** **3.8e-627 (6709)** **2.0e-358 (14047)**	**PBM**	**MA0037.2**  **UP00032_1**  **UP00032_2** 	**0.00e+0** **0.00e+0** **9.531e-115**	**Cannot be distinguished**	**UP00032_1 0.730** **MA0037.2 0.644**
11. MAFK	**H1-hESC (11425)**	**MA0496.1**  **UP00044_1**  **UP00044_2** 		**2** **5** **69**	**7.8e-4654 (10435)** **1.1e-3485 (10851)** **3.8e-6 (10643)**	**ChIP-seq**	**MA0496.1**  **UP00044_1** 	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**MA0496.1 0.936** **UP00044_1 0.904**
	HeLa-S3 (14185)	**MA0496.1**  **UP00044_1**  **UP00044_2** 		**2** **4** **31**	**1.3e-5460 (12490)** **8.2e-4002 (13172)** **2.1e-25 (13197)**	**ChIP-seq**	**MA0496.1**  **UP00044_1** 	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**MA0496.1 0.946** **UP00044_1 0.929**
	HepG2 (61944)	**MA0496.1**  **UP00044_1**  **UP00044_2**  **MA0496.1**		**2** **3** **22** **3**	**3.1e-34593 (58130)** **6.4e-29919 (60873)** **8.5e-420 (58837)** **1.1e-21321**	**ChIP-seq**	**MA0496.1**  **UP00044_1**  **MA0496.1**	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**MA0496.1 0.970** **UP00044_1 0.964** **MA0496.1**

	Same cell line (37628)	**UP00044_1** **UP00044_2**		**4** **22**	**(35395) 4.7e-18155 (36815) 1.7e-260 (35490)**		**UP00044_1**	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**0.951** **UP00044_1 0.923**
	IMR90 (40788)	**MA0496.1** **UP00044_1** **UP00044_2**		**2** **4** **24**	**4.2e-19236 (38075) 5.8e-16299 (39806) 7.7e-167 (38596)**	**ChIP-seq**	**MA0496.1** **UP00044_1**	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**MA0496.1 0.973** **UP00044_1 0.968**
	K562 (19317)	**MA0496.1** **UP00044_1** **UP00044_2**		**2** **5** **29**	**4.6e-8511 (17466) 1.4e-6477 (18302) 3.5e-46 (18037)**	**ChIP-seq**	**MA0496.1** **UP00044_1**	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**MA0496.1 0.951** **UP00044_1 0.944**
12. TCF7L2	**HCT-116 (19463)**	**MA0523.1** **UP00083_1**		**2** **3**	**1.1e-1362 (14205) 3.0e-1276 (14671)**	**ChIP-seq**	**MA0523.1** **UP00083_1**	**0.00e+0** **0.00e+0**	**Cannot be distinguished**	**UP00083_1 0.837** **MA0523.1 0.832**
	HEK293 (8961)	**MA0523.1** **UP00083_1** **UP00083_2**		**2** **3** **61**	**5.9e-743 (7985) 3.1e-702 (8200) 3.1e-16 (7373)**	**ChIP-seq**	**MA0523.1** **UP00083_1**	**3.552e-309** **1.711e-236**	**ChIP-seq**	**UP00083_1 0.825** **MA0523.1 0.801**

	HeLa-S3 (19242) **Same cell line (3198)**	**MA0523.1** **UP00083_1** **UP00083_2** **MA0523.1** **UP00083_1**	 	**2** **3** **139** **1** **3**	**2.8e-1002 (15056)** **6.0e-969 (15620)** **1.2e-10 (14651)** **2.2e-494 (2909)** **5.2e-480 (2948)**	**ChIP-seq** **ChIP-seq**	**MA0523.1** **UP00083_1** **MA0523.1** **UP00083_1**	**4.83e-224** **7.708e-185** **9.92e-211** **2.95e-179**	**ChIP-seq** **ChIP-seq**	**MA0523.1 0.833** **UP00083_1 0.820** **MA0523.1 0.886** **UP00083_1 0.876**
	HepG2 (2742)	**UP00083_1** **MA0523.1**		**1** **3**	**9.3e-231 (2678)** **4.3e-219 (2642)**	**PBM**	**MA0523.1** **UP00083_1**	**7.725e-259** **6.827e-230**	**ChIP-seq**	**MA0523.1 0.852** **UP00083_1 0.844**
	MCF-7 (10293)	**UP00083_1** **MA0523.1** **UP00083_2**		**2** **3** **61**	**2.3e-1070 (9000)** **4.1e-1068 (9230)** **4.6e-11 (8205)**	**PBM**	**MA0523.1** **UP00083_1**	**1.322e-274** **1.164e-226**	**ChIP-seq**	**UP00083_1 0.860** **MA0523.1 0.848**
	PANC-1 (13366)	**UP00083_1** **M A0523.1** **UP00083_2**		**2** **83** **3**	**1.3e-773 (11177)** **2.4e-773 (10869)** **6.4e-3 (9896)**	**PBM**	**MA0523.1** **UP00083_1**	**3.789e-305** **7.994e-260**	**ChIP-seq**	**UP00083_1 0.756** **MA0523.1 0.752**