

**COFACTOR PREDICTION USING  
CHIP-SEQUENCING DATA**

**CHANG CHENG WEI**

**B.Comp. (*Hons*), Computer Science, NUS**

**B.Sc. (*Hons*), Applied Mathematics, NUS**

**A THESIS SUBMITTED  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY  
GENOME INSTITUTE OF SINGAPORE  
NUS GRADUATE SCHOOL FOR INTEGRATIVE  
SCIENCES AND ENGINEERING**

**2014**

## **DECLARATION**

**I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.**

**This thesis has also not been submitted for any degree in any university previously.**

A handwritten signature in cursive script, appearing to read 'C. Cheng Wei', is positioned above a horizontal line.

**Chang Cheng Wei**

**25 July 2014**

## **ACKNOWLEDGEMENTS**

First and foremost I would like to thank my beloved mother for her unconditional love and care during the entire period of my PhD. This journey had been long and strenuous for me and much more so for her. I can truly empathise with the torment and anguish of her seeing me suffer over the period. Same goes for everyone in my family.

Secondly, I would like to thank my supervisors Dr Ken Sung Wing Kin and Dr Edwin Cheung Chong Wing for their unceasing support and guidance over the entire duration of my PhD, tolerating my various shortcomings, procrastinations and lacklustre attitudes. Thanks for continuing to believe in me and giving me chance and again to eventually complete my thesis. I would also like to thank Dr Tara Leah Huber, Director of Student's Affairs in GIS for her support and for helping to organise the various sub-deadlines that are critical for making things move for me. In addition, I would also like to thank various A\*STAR staffs and NGS staffs that have helped in one way or another for the various administrative tasks pertaining to me. For these I would like to offer my deepest and most sincere gratitude to the above people who have made the completion of this thesis possible.

Next, I would like to thank Chng Kern Rei who is my childhood friend since secondary school, Jae Tan Peck Yean and Zhang Zhizhuo for their great support, valuable comments on my thesis and giving me various advices and encouragements during this final stretch. I would also like to thank my lab mates Goh Wan Ling for performing the various wet lab experiments to support the biological findings for CENTDIST pertaining

to the validation of AP4, and Tan Si Kee for her extensive works on AP2 gamma. I would also like to thank my other lab mates which include Meihui, Shin Chet, Ying Ying, Shuzhen, Noel, Yixiang, Michelle, Jiayu, Jasmine, Yang Chong, Zikai and Simon for bringing life, fun and laughter in my PhD journey.

Last but not least I would like to thank my doctor, psychologist and case manager in IMH for their support and treatment while I was suffering psychologically and also my religious organisation Soka Gakkai International for bringing me hope through the guidances and encouragements from my fellow comrades in faith and books written by my life mentor Daisaku Ikeda. Particularly happy to have successfully overcome and recover from my anxiety disorder and depression. This journey had been very eventful and I managed to learn many important and precious life lessons which enabled me to become much more courageous and confident than I was ever before, equipping me with the ability to face any challenges ahead in the future.

## LIST OF TABLES

Table 1 IUPAC codes for nucleic acids .....	22
Table 2 List of curated TF PWM databases.....	29
Table 3 List of existing motif enrichment tools.....	37
Table 4 The ranking of the known co-TFs of AR for each motif enrichment tool.....	55
Table 5 Novel co-TF candidates of AR predicted by CENTDIST.....	56
Table 6 Table describing the four sets of simulated SP1 ChIP-seq sequences.....	78
Table 7 The counts and respective enrichment P-value of SP1 in the four sets of sequences.   81	
Table 8 The counts and normalised count of SP1 in the four sets of sequences. ....	83
Table 9 Performance of MOTIFDIFF in predicting CTCF using CHIPSCORE as gold standard       98	
Table 10 Performance of MOTIFDIFF in predicting all TF by FAMILY using CHIPSCORE as gold standard.....	99
Table 11 Accuracy of MOTIFDIFF with respect to the bidirectionality of gold standard enrichment. The number of points are shown in bracket.....	103

## LIST OF FIGURES

Figure 1.1 Motif model. ....	21
Figure 1.2 Log probability of PWM with pseudocount added. ....	24
Figure 1.3 Illustration of Gibbs sampling algorithm. ....	27
Figure 1.4 Steps in ChIP-seq assay. ....	30
Figure 1.5 Example differential motif comparison illustration using heatmap based on enrichment pvalue score. ....	40
Figure 2.1 AR motif distribution around AR ChIP-seq peaks. ....	44
Figure 2.2 FOXA1 ChIP-seq and motif distributions around AR ChIP-seq peaks. ....	45
Figure 2.3 Determining the frequency score of AR motif around AR ChIP-seq peaks. ....	48
Figure 2.4 Analysis of motifs around RNA PolII ChIP-seq peaks. ....	51
Figure 2.5 Demonstration of CENTDIST Capability ....	53
Figure 2.6 Frequency and velocity graphs of AR and its co-TFs including the newly discovered AP4. ....	57
Figure 2.7 Uniform distribution of AP4 motifs around the AR only ChIP-seq peaks. ....	59
Figure 2.8 AP4 is a novel co-TF of AR. ....	60
Figure 2.9 Top 10 motif hits reported by CENTDIST showing the score and distribution of best motif within each TF family ....	62
Figure 2.10 AP-2 $\gamma$ is identified as a potential collaborative factor of ER $\alpha$ . ....	63
Figure 2.11 AP-2 $\gamma$ is required for the efficient binding of ER $\alpha$ . ....	65
Figure 2.12 CENTDIST web interface and program procedure. ....	66
Figure 2.13 Sample Output page of CENTDIST. ....	68
Figure 3.1 Types of comparison. ....	73
Figure 3.2 Histograms of FOXA1 (V\$HNF3ALPHA_Q6) motif around AR ChIP-seq peaks in LNCaP with different GC content plotted using the same scale shows the additive nature of background motifs. ....	76
Figure 3.3 SP1 motif logo is GC rich. ....	77
Figure 3.4 Enrichment score of CENTDIST fails to provide necessary information to determine differential enrichment among simulated datasets. ....	80
Figure 3.5 Normalisation by subtracting the background enables the accurate comparison of enrichment. ....	83
Figure 3.6 ENCODE TF ChIP-seq Experimental Matrix across 91 Cell Lines and 161 TF. ....	85
Figure 3.7 Different peak distribution profiles observed for pairs of TFs ChIP-seq performed in K562. The graphs have been normalised by dividing by the number of peaks in the ChIP-seq data to be centered upon. ....	86
Figure 3.8 Motif distribution profile of MAZ motif (V\$MAZR_01) around CHD1 ChIP-seq peaks in K562. ....	87

Figure 3.9 The calculation of HISTSCORE from histogram counts. The histogram counts have been scaled by dividing by the number of peaks and multiplying by 10. ....	90
Figure 3.10 HISTSCORE of motif correlates better with actual ChIP-seq peak distribution signal than other score. ....	95
Figure 3.11 CEBPB is preferentially enriched in different regions depending on the Cell Line in which the ChIP is performed. ....	100
Figure 3.12 Scatterplot of CHIPSCOREREV against CHIPSCORE to show the accuracy in relation to the bidirectionality of gold standard enrichment. ....	102
Figure 3.13 Differentially enriched motifs found in ERE2unique over EREGFunique.....	105
Figure 3.14 Differentially enriched motifs found in EREGFunique over ERE2unique.....	106
Figure 3.15 Differentially enriched motifs found in siCTRL unique over SiFoxA1 unique.....	108
Figure 3.16 Differentially enriched motifs found in SiFoxA1 unique over siCTRL unique.....	109
Figure 3.17 Differentially enriched motifs found in ER unique sets over ERAP2.....	111
Figure 3.18 Differentially enriched motifs found in ERAP2 over ERunique.....	112
Figure 3.19 Verification that E2F is more enriched in ER AP2 overlapping sites as compared to ER unique sites.....	113

# TABLE OF CONTENTS

DECLARATION .....	2
ACKNOWLEDGEMENTS .....	3
LIST OF TABLES .....	5
LIST OF FIGURES .....	6
TABLE OF CONTENTS .....	8
SUMMARY .....	10
ABBREVIATIONS .....	13
CHAPTER 1 Introduction .....	14
1.1 Background .....	14
1.1.1 Central Dogma of Molecular Biology .....	14
1.1.2 Gene Regulation by Transcription Factors .....	15
1.1.3 Transcription Factor .....	17
1.1.4 ChIP-Sequencing .....	29
1.2 Research Problems .....	31
1.2.1 Identifying co-TF from ChIP-seq datasets .....	31
1.2.2 Identify Differential Motif Enrichment between two sets of ChIP-seq peaks .....	38
CHAPTER 2 CENTDIST – Web-based tool for Motif Enrichment .....	41
2.1 Introduction .....	41
2.2 Results .....	43
2.2.1 Development of CENTDIST .....	43
2.2.2 CENTDIST Web Server .....	65
2.3 Discussion .....	69
CHAPTER 3 MOTIFDIFF – Web-based tool for Differential Motif Enrichment .....	71
3.1 Introduction .....	71
3.2 Additive nature of motif background .....	75
3.3 Difficulties in identifying differential motifs ....	<b>Error! Bookmark not defined.</b>
3.4 HISTSCORE: An Alternative Enrichment Statistic .....	82
3.4.1 Overview .....	82
3.4.2 Insight from ENCODE datasets .....	84

3.5	MOTIFDIFF Algorithm .....	91
3.6	Results .....	93
3.6.1	Motif HISTSCORE have good correlation with ChIP-seq HISTSCORE .....	93
3.6.2	Large scale validation of MOTIFDIFF using ENCODE.....	95
3.6.3	Application.....	103
3.7	Discussion .....	113
CHAPTER 4	Conclusions .....	116
4.1	Summary of Contributions .....	116
4.2	Future Directions.....	118
4.2.1	CENTDIST.....	119
4.2.2	MOTIFDIFF.....	120
PUBLICATIONS	.....	122
BIBLIOGRAPHY	.....	123

## SUMMARY

Nuclear receptors (NRs) are a special class of transcription factors (TF) whose primary function is to allow cells to react to chemical changes in the environment or to respond to hormones produced by other parts of the body. The response is mediated by the interaction with its cofactors and co-regulators. Sufficiently disrupted transcriptional network involving NRs could jeopardise a cell's functionality and lead to disastrous outcome such as cancer.

Our lab's focus is in understanding the development and progression of prostate cancer and breast cancer for which the nuclear receptors androgen receptor (AR) and estrogen receptor (ER) play central roles. ChIP-seq of AR in prostate cancer cell lines LNCaP and VCaP and that of ER in MCF-7 cell line was performed to obtain a genome-wide view of the binding sites of these NRs in their respective cancer cell-lines. These data allows us to further study the transcriptional network mediated by these NRs in their respective cancer cell-lines.

Genes in a transcriptional network are regulated by regions with TF binding motif clusters referred to as cis-regulatory modules (CRMs), with CRMs near the transcription start site (TSS) of a gene being known as promoters while distal ones being known as enhancers or silencers. The activity of a CRM is often being likened to a binary switch that depends on the binding presence of its members as inputs.

This implies that TFs that are involved in co-regulating a set of genes tend to occur within close vicinity. Having obtained the location of the binding sites of an NR in the genome, we may now predict what are the cofactors of the NRs, i.e other TFs that works closely with the NR in the same CRM to co-regulate target genes, and thus expand our knowledge of the transcriptional network. Henceforth, in-silico screening of probable cofactors of the NR can be performed by looking for over-represented known TF motifs in the vicinity of the NR binding sites using databases of known TF motifs actively deposited in JASPAR and TRANSFAC. Novel motifs could also be discovered using *de novo* motif discovery tools such as MEME, Weeder and Bioprosector.

Motifs have long been used to predict the presence of TF binding sites. However, to date, analysis had been mostly restricted to promoter regions of regulated genes. The main reason is that prediction by motif alone on the entire stretch of genome rarely gives satisfactory results, yielding unacceptable rates of false positives, i.e. most regions that contain motif along the DNA are not bound by the TF, possibly due to closed chromatin or some other biological reasons.

ChIP-seq is a revolutionary assay used to detect the binding sites of TF in the entire genome, made possible through the rapid advancement in high-throughput sequencing technology. The precision of the binding location of TF can be detected within positional accuracy of  $\pm 200$  bp. Motif predictions of TF binding sites tend to yield a much greater accuracy when restricted to regions near ChIP-seq peaks. Moreover, observation of the motif distribution profile of known cofactor around ChIP-seq peaks suggests that true

cofactors tend to exhibit certain imbalanced distribution around ChIP-seq peaks. Utilizing this property, we developed CENTDIST that ranks TF candidates based on the imbalanced distribution around ChIP-seq peaks. CENTDIST performs positively in comparison with existing tools of similar nature such as CEAS and CORE\_TF using mouse ES cells. By applying CENTDIST to our inhouse generated AR and ER ChIP-seq in LNCaP and MCF7 respectively, we managed to uncover novel co-factors that potentially play important roles in prostate and breast cancer.

Though CENTDIST is well suited for identifying enriched motif, it is not suitable for comparing enrichment among sets of peaks. As such, we developed MOTIFDIFF with the aim of comparing motif enrichment between peak sets. As with CENTDIST, MOTIFDIFF makes use of the motif to accurately quantify the motif in the sets which can then be used for comparison. Validating using large number of ChIP-seq datasets from ENCODE, we showed that MOTIFDIFF is able to correctly predict the relative abundance of TFs using motifs. Applying MOTIFDIFF to our in-house ER and AP2 ChIP-Seq, we identify several potential candidates that play specific roles in ER in the presence of AP2 and those in the absence of AP2.

## ABBREVIATIONS

AR	Androgen receptor
bp	Base Pairs
ChIP	Chromatin immuno-precipitation
ChIP-chip	Chromatin immuno-precipitation coupled with chip hybridization
ChIP-seq	Chromatin immuno-precipitation coupled with high throughput sequencing
CRMs	Cis-regulatory modules
DBD	DNA binding domain
DNA	Deoxyribonucleic Acid
DWM	Dinucleotide weight matrix
EM	Expectation Maximisation
ER	Estrogen receptor
HMM	Hidden Markov Model
IUPAC	International Union of Pure and Applied Chemistry
LBD	Ligand-binding domain
mRNA	Messenger RNA
NRs	Nuclear Receptors
PCM	Positional Count Matrix
PWM	Positional Weight Matrix
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
TF	Transcription Factor
tRNA	Transfer RNA

## **CHAPTER 1    Introduction**

### **1.1 Background**

#### **1.1.1 Central Dogma of Molecular Biology**

Following the completion of the Human Genome Project (Venter et al. 2001), the understanding of human biology had grown by leaps and bounds. It has enabled us to finally study and understand human genetic diseases such as cancer and eventually develop treatments for them. This is particularly accelerated by the development of computational technology which is well suited to tackle the combinatorial complexities that have been encoded in the depths of human genome. One of the earliest concepts that we learnt through genetics is that each individual has a unique genetic makeup which translates to specific phenotypic characteristics and this information is contained in the genes which represent a mere 3% of the entire genome.

The genome comprises of deoxyribonucleic acid (DNA) molecules made up of nucleotide bases: A, C, G, T (viz. Adenine, Cytosine, Guanine, and Thymine) stringed together along a phosphate backbone. DNA molecules occur in paired strands with complementary pairing of A to T, G to C, forming a double helix.

The idea of how DNA and protein are related has been ingeniously represented by the central dogma of molecular biology proposed by Crick. The central dogma describes major transitions among the three key entities in the cell: DNA, ribonucleic acid (RNA) and protein. Within each cell, the DNA is organized into chromosomes residing within

the nucleus. A cell divides to produce two daughter cells. Prior to the cellular division (also known as mitosis), the genomic DNA replicates so that each daughter cell receives exactly 23 pairs of chromosomes. The chromosomal DNA ranges from 50Mb to 250Mb in size and, to date, is punctuated with more than 30, 000 genes. In the event of transcription, the DNA sequence within each gene is converted to pre-messenger RNA (pre-mRNA) through complementary pairing. This RNA molecule typically consists of coding regions, known as exons, interspersed with non-coding intronic fragments. In order to produce mature RNA, spliceosome removes the introns to form a continuous string of exons which is then exported out of the nucleus to the cytoplasm. Ribosomes then “decode” the genetic information stored in mature RNA by systematically recruiting amino acids which eventually leads to formation of polypeptide and folded protein. However, multiple non-coding RNAs which do not get translated to proteins also exist. Some of the well-known examples include ribosomal RNA (rRNA) which is a core component of the ribosome, and transfer RNA (tRNA) which is responsible for bringing appropriate amino acids to mature messenger RNA (mRNA) during peptide synthesis.

### **1.1.2 Gene Regulation by Transcription Factors**

Despite the fact that the entire human genome carries 30,000 genes, only a fraction of these genes are actually actively transcribed in a particular cell. This is essential considering the fact that protein determines the cells functionality and the functionality of cells are tissue-specific. Special mechanisms are in place to control for the amount of proteins to be produced in the cell, depending on the environment it is in. These mechanisms generally involve various proteins such as RNA polymerases, histones,

histone modifiers, transcription factors, and co-factors. Among these, transcription factors (TFs) play a central role. TFs are proteins containing DNA binding domain (DBD) that recognise specific sequences in the genome which are commonly referred to as the TFs' motifs. TFs typically bind closely together at motif clusters known as cis-regulatory modules (CRMs). CRMs are typically 100-1000bp in length and can be found at gene promoters within 3000 bp from the gene TSS or at distal regions in general which could be located anywhere on the genome. The distal regions can in turn be classified into one of the following a) enhancers that promote the expression of the target genes, b) insulators that indirectly affect gene regulation by interacting with nearby CRMs and c) silencers which repress the expression of the target genes. Several CRMs consisting of various TFs are typically involved in the regulation of a particular gene. Knowing how and when genes are activated is of particular importance; particularly for those disease-causing ones as we can then study the regulatory network and design a remedy that selectively target their aberrant expression. We are still a long way from knowing exactly how and when genes are activated, but following the sequencing of an organism's genome, we can predict regions that are potentially bound by the regulating TFs. The existence of motifs however, is insufficient to determine whether a particular TF will bind at a particular genomic location. It also depends on several other epigenetic factors such as nucleosome positioning and histone marks involving the methylation and acetylation of any of the four cores of the histones.

### **1.1.3 Transcription Factor**

Because of its relation to gene transcription, much emphasis had been placed towards studying them. The primary questions are where they bind and what the regulated genes are. To answer the first question, we could use the motifs they recognized once we have identified their motifs. Special assays such as SELEX (Selection of aptamers by systematic evolution of ligands by exponential enrichment) and PBM (Protein Binding microarray) have been developed specifically for this purpose. SELEX involves the progressive selection and amplification of an initial random pool of DNA towards a final pool of DNA with optimized ability to bind a specific TF (Ellington et al. 1990) whereas PBMs are special microarrays consisting of probes covering all 8-mer motifs and give out fluorescent signal in response to protein binding events (Mukherjee et al. 2004).

Other than from experimental assays described above, another strategy is to look at the promoters of homologous genes across multiple species or genes with correlated expressions in a microarray experiment. The reason why this works is because promoters are typically functional cis-regulatory element where TFs bind. However such approaches typically work well only for simpler organisms such as bacteria and yeast because in these organisms promoters are primarily where TFs bind. For complex eukaryotes such as human and mouse, gene regulation involves interplay of enhancers and silencers which can be located anywhere in the genome. Certain classes of TFs such as nuclear receptors are known not to bind preferentially at promoters.

In recent years, the development of an enabling technology ChIP (Chromatin-Immunoprecipitation) drives the study of transcription factor to a whole new level. Using ChIP, coupled with high throughput sequencing assay (ChIP-seq) (Barski et al. 2007) or with high throughput assays such as chip hybridization (ChIP-chip) (Aparicio et al. 2004), we can now retrieve the genome-wide binding of a target transcription factor in-vivo. ChIP-seq has now become the de-facto standard for identifying genome-wide binding of a particular TF *in vivo* due to its higher resolution, quality and cost effectiveness. Analysing the data churned from ChIP-seq, we could obtain primary information such as the binding location in the genome, the binding affinity, as well as secondary information such as the motif and their likely target genes which are often assumed to be within 50kb from their binding.

The task of determining the motif of a transcription factor from a set of putative binding sequence is a non-trivial problem and is about one of the oldest and most actively studied bioinformatics problems known as *de novo* motif finding.

#### 1.1.3.1 Motif Model

Before proceeding, we need to first decide how the motif is to be represented. There are various methods to represent the motif, ranging from simple consensus, to positional weight matrix (PWM) , to dinucleotide weight matrix (DWM) (Siddharthan 2010), to Hidden Markov Model (HMM) (Gelfond et al. 2009). These methods are increasingly complex motif representation trading off between simplicity and ability to accurately capture the motif's degeneracy and further capturing positional dependencies. Overly

complex models however may not be beneficial, as there is a risk of over-fitting while not necessarily improving the general prediction accuracy. To date, such advanced method has yet to show much benefit. In this thesis, we will mainly focus on using PWM as the preferred way of representing motifs as it is the most popular way and there are many usable motifs deposited in this format in the literature. Moreover it is observed that in reality, PWM and consensus sequence usually provides a good enough approximation (Benos et al. 2002).

We use Figure 1.1 to illustrate how a motif is represented as a PWM. Figure 1.1A shows the binding sequences of a TF recognising a 7bp sequence coloured in blue. Once the sequences are being aligned, we obtain the positional count matrix (PCM) as in Figure 1.1B. The simplest way to represent the motif is by using consensus, i.e. the nucleotide with the highest occurrence. In the example, the consensus sequence is AGCTCAC. However, we note that the TF seems to prefer C and G equally at the second position. In general, within the region that the TF recognises, there could be certain positions that are more tolerant, allowing for degeneracies, ranging from having no preference to any nucleotide at all to strictly preferring only one of the nucleotide. To rectify this issue, instead of the single consensus using the most frequently occurring nucleotide, we allow the base at a particular position to be within a subset of all possible nucleotides [A, C, G, T] (i.e. [A], [C], [G], [T], [A,C], [A,G], [A,T], [G,T], [A,C,G], [A,C,T], [A,G,T], [C,G,T], [A,C,G,T]). For succinctness, there is a single letter representation for all possible subsets, represented by the IUPAC coding (see Table 1). IUPAC consensus gives a compact representation of the motif but suffers from being unable to fully capture the degeneracy,

however it is computationally more efficient to work with and is well suited for a number of data structures that allow for optimisation of algorithms such as suffix array and hash tables. The discrete solution space also allows the possibility of exhaustive enumeration for shorter motifs. It is bulky and less informative to display a motif as a matrix. Therefore, a graphical way of representing the motif, weblogo (Crooks et al. 2004) is developed, where the height of each base corresponds to the information content of that base (see Figure 1.1C). For a base that has no preferences for any nucleotides (i.e. the probability of each nucleotide is 0.25), the information content is 0. If the base is biased to only one particular nucleotide, then the information content achieves its maximum at 2. Information content is a good measure of degeneracy as it intuitively gives us the level of importance and tolerance for mismatch at a particular base. In the example we see that position two has the greatest degeneracy as its information content is the lowest. Precisely, the information content at base  $i$  is calculated as  $-\sum_{j \in [A,C,G,T]} p_{i,j} \log(p_{i,j})$ , where  $p_{i,j}$  represent the empirical probability of the nucleotide  $j$  occurring at base  $i$ .

A

GACTATGCGCGAGCAGTAGTACAGAGCTCACAACTCGTAAAGCCTAATGG  
 AAACCTCACGACGAACAGCAAGACCATGCGTTACCCCTCTTATTGGAACC  
 TGTCGGCGGGAATTAGTGATCCGAGTCGGGGGAAGCTCTCATTATCCTTT  
 CGCATGAGAGAGGTCACATGCATACGAGCTGTTGGGCATCGTCATTCTAA  
 TTAGACCTATGCGTACGAGGGGGCGCAGTAGCGCCGATCTACCTCACATG  
 CTGCATCCAGGGCGAAGCTCAGATGGTGGGAACTCTGGCCGGGCGGTCTC  
 AAGCTGACCGGCAACGACCGACATTAAAACCTCACGATCGCCATTGTCAC  
 TTCTCCCCATCTCACATGCCTGCGGCAGTAGATTGGGCCTGGGAGCTGGC  
 TGAGCAGAGCAATCTTCGCCAAGCAGCCATTCAACGTACACCAAAACAG  
 AGCCACCTGTAGCTCACACTGACCTTGCCATACAAGCTGACACTTTATC

B

AGCTCAC  
 ACCTCAC  
 AGCTCTC  
 AGGTCAC  
 ACCTCAC  
 AGCTCAG  
 ACCTCAC  
 ATCTCAC  
 ACGTCAC



Position	A	C	G	T	IUPAC
1	10	0	0	0	A
2	0	4	5	1	S
3	0	8	2	0	C
4	0	0	0	10	T
5	0	10	0	0	C
6	9	0	0	1	A
7	0	9	1	0	C

C



**Figure 1.1 Motif model.**

A) Sequences colored in blue are where TF binds. B) Alignment of sequence bound by TF are counted to produce the position weight matrix (PWM) of aligned sequences and the IUPAC representation in the last column consisting of the allowed base in the position consensus. Consensus of the sequence is ASCTCAC. C) The sequence logo representation of the aligned sequence using weblogo tool (<http://weblogo.berkeley.edu>)

**Table 1 IUPAC codes for nucleic acids**

code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

### **1.1.3.2 Predicting TF binding using IUPAC consensus and PWM**

To predict TF binding at an arbitrary sequence using our model, first we need to define a measure of goodness for matching. For consensus sequence, the measure of goodness is simply defined by the number of mismatches, the lower the better. The problem with this is granularity. A single increase in the number of mismatches raises the admissible predictions by many folds. For example a length 7bp motif with no degenerate base, admits 1 sequence with no mismatch, 22 sequences with at most one mismatch and 211 sequences with at most two mismatches.

The measure of goodness of matching using PWM is computed using the log likelihood ratio. In other words, we are comparing how good the fit is using our model as compared with a background model with no preference for nucleotides at every position.

The formula for the score is:

$$\sum_i \log_2 \left( \frac{p_{i,S[i]}}{0.25} \right) = \sum_i \log_2(p_{i,S[i]}) - L * \log_2 0.25 = \sum_i \log_2(p_{i,S[i]}) + 2L$$

where  $S[i]$  is the  $i^{\text{th}}$  position of the sequence we would like to measure,  $p_{i,S[i]}$  is the probability that the  $i^{\text{th}}$  position is  $S[i]$ , and  $L$  is the length of the motif. To compute this score, first we need to convert our PWM into a table of log probability.

Figure 1.2 shows the log probability of our PWM in Figure 1.1B after adding a pseudocount of 0.25 to each cell in the table. This is a standard method to avoid overfitting since the observed zero count is possibly due to a lack of data. Note that the contribution of pseudocount decreases as the sample sequences aligned increases.

Using this table with row  $i$  and column  $j$  being represented by  $M[i,j]$ , the score of a sequence AGCTCAC is  $M[1,A] + M[2,G] + M[3,C] + M[4,T] + M[5,C] + M[6,A] + M[7,C] + 2*7 = -0.03067 + -0.3123 + -.12494 + -0.03067 + -0.03067 + -0.07525 + -0.07525 + 14 = 13.32$  which is the maximum score attainable by any sequence.

The corresponding score of a change of the relative weak base at position 2 from G to A (AACTCAC) is 12 and of a change of the strong base as position 5 from G to A (AGCTAAC) is 11.7, fitting our intuition of a higher penalty for mismatches at stronger bases. Analogous to the number of mismatch cut-off for consensus sequence is the PWM cut-off score. All in all, PWM offers a much higher granularity and better handling of mismatches at degenerate bases than consensus.

i	A	C	G	T
1	-0.03067	-1.64345	-1.64345	-1.64345
2	-1.64345	-0.413	-0.32123	-0.94448
3	-1.64345	-0.12494	-0.68921	-1.64345
4	-1.64345	-1.64345	-1.64345	-0.03067
5	-1.64345	-0.03067	-1.64345	-1.64345
6	-0.07525	-1.64345	-1.64345	-0.94448
7	-1.64345	-0.07525	-0.94448	-1.64345

**Figure 1.2 Log probability of PWM with pseudocount added.**

Lookup entry for computing PWM score of AGCTCAC

### 1.1.3.3 Precision and Recall of Prediction

An important measure of the goodness of prediction is the precision and recall. Precision measures the chance that a sequence being predicted is truly being bound, in other words a precise prediction is one that do not report many spurious predictions; whereas recall or sensitivity measures how much of the true bindings we can recover. By relaxing our

criteria for the prediction, (i.e. increasing the number of mismatch or lowering our PWM cut-off score) we decrease the precision as we will likely be reporting more false positives while we increase the recall as we can recover more instances. This trade-off between precision and recall is a constant decision to be made. Other than the cut-off criteria, a weaker motif model (one with more degenerated bases) will have both lower precision and recall than a stronger motif model.

#### 1.1.3.4 ***De novo* motif finding**

*De novo* motif finding is the process of identifying over-represented patterns in a set of sequence. Suppose we are given a set of sequences that we know are bound (or partially) by some DNA-binding protein which binds onto certain specific motif, but we do not know what the motif is. The problem then is to try to describe the motif that is being bound by the protein (based on one of the models described in Section 1.1.3.1) based on the set of sequences given. The following general methodologies are commonly used to solve *de novo* motif finding problem.

##### ***Word-based motif***

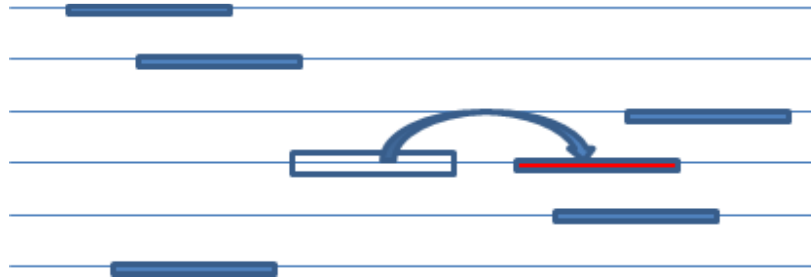
We can efficiently enumerate frequently occurring words in a set of sequences using special data structures such as suffix tree, suffix arrays or hash table. These words with high frequencies are likely instances of certain motif occurrences. By combining and clustering words that differ by only a few bases, the PWM of the motif could be inferred. Examples of existing *de novo* motif finding tools that utilize this methods include RSAT (van Helden et al. 1998), YMF (Sinha et al. 2003) and CisFinder (Sharov et al. 2009).

One of the problems with this approach is that when the length of the motif is long, the word instances of the motif may not occur frequently enough to be identified. And hence the *de novo* motif finding could fail. To rectify this problem, motif finders directly find PWM models instead of indirectly combining from separate word instances. There are two general methods to go about doing this: Gibbs sampling and expectation maximisation (EM).

### ***Gibbs Sampling***

We use Figure 1.3 to illustrate the Gibbs sampling algorithm. Gibbs sampling assumes that each sequence has a motif occurrence. Initially the position of the motif occurrence in each sequence is randomly initialised. Assuming that each sequence has a motif occurrence coming from an assumed PWM model, we can iteratively refine the model to one that is slightly better by repeatedly applying the following step. Take turn leaving out each sequence, find the segment that has the highest score using the PWM aligned using the occurrences in the other sequences and then update the motif position to the highest scoring segment scored using the PWM. After repeating this process several times, at some point, the motif position of all sequences the motif score will remain unchanged. At this point, the PWM aligned using all the sequences will be reported. As the algorithm requires an initial randomized initialization, this PWM reported may not be optimum as it is possible that the algorithm may get stuck at a local optimum. Several modifications to the algorithm were employed by different variants to address and relax several assumptions such as the restriction of having exactly one motif occurrence per sequence. Examples of *de novo* motif finders utilising the Gibbs sampling approach are:

GibbsDNA(Lawrence et al. 1993), AlignACE(Roth et al. 1998), MotifSampler(Thijs et al. 2001), BioProspector(Liu et al. 2002) and ANN-spec(Workman et al. 2000).



**Figure 1.3 Illustration of Gibbs sampling algorithm.**

In the current iteration, it is the fourth sequence's turn to be updated. The new position of the motif is being updated for the fourth sequence to the red motif which is the best scoring position using the PWM aligned using the other blue motifs. In next iteration, the fifth sequence is being considered.

### ***Expectation Maximisation***

Expectation maximisation is a standard statistical tool used to estimate the prior's unknown parameters that give the maximum likelihood i.e. the best explaining model that gives the highest probability of observing the given data. In the context of *de novo* motif finding problem, the unknown prior we are interested in is the PWM motif model. The statistical formulation is as follows: Given  $N$  sequences  $s_1, s_2, s_3, \dots, s_N$ , each length  $L$  substring of these  $N$  sequences are presumed to come from either the motif model  $\Theta$  or the background model  $\Theta_0$ . The EM algorithm will try to iteratively update the motif model  $\Theta$  to obtain one in which using the model, the probability of observing the input sequences is maximised. In the EM algorithm there are two phases. In the E-phase (Expectation phase) we use the current motif model to compute for each position the

probability that it comes from the motif model. In the M-phase (Maximisation phase) we use differentiation to obtain the stationary point of the expectation function. These parameters are then evaluated to obtain the new motif model. The MEME suite consists of a large arsenal of *de novo* motif finders catered towards different purposes (Bailey et al. 2009).

### ***Graph-based***

In graph-based *de novo* finder, each sequence is represented by a node and pairs of sequences sharing certain substrings that are close in terms of Hamming distance from one another will be represented as an edge in a graph. The problem then translates into a problem of cliques finding (Liang et al. 2004; Zhang et al. 2011) or dense subgraph (Zhang et al. 2011).

#### **1.1.3.5 Existing curated Protein-DNA binding motif databases**

Motifs of TFs have been actively published and curated by the scientific community. TRANSFAC and JASPAR are two of the most commonly used TF motif databases of mammals.

Table 2 shows a list of available curated TF PWM databases for various organisms. The analysis in this thesis is based mostly on TRANSFAC for comparisons with other methods.

**Table 2 List of curated TF PWM databases**

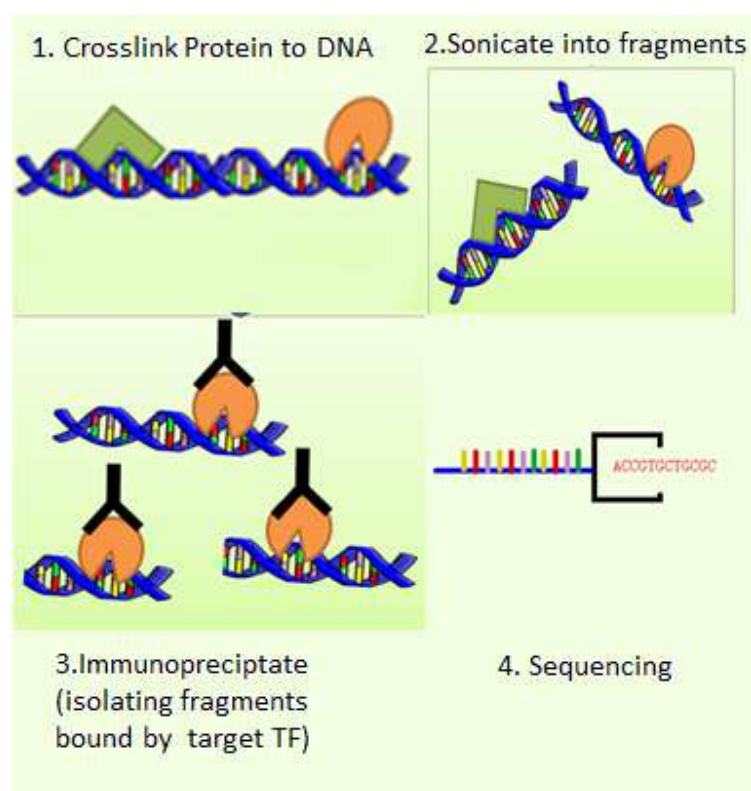
<b>Name</b>	<b>Organisms</b>
RegPrecise(Qi et al. 2014)	Prokaryotes
RegTransBase(He et al. 2013)	Prokaryotes
RegulonDB(Xu et al. 2013)	Escherichia coli
PRODORIC(Machiels et al. 2013)	Prokaryotes
TRANSFAC(Dai et al. 2012)	Mammals, Bacteria, Plant
JASPAR(Sandelin et al. 2004)	Mammals, Bacteria, Plant
TRED(Akutsu et al. 2012)	Human, Mouse, Rat
HOCOMOCO(Iwagami et al. 2012)	Human

### **1.1.4 ChIP-Sequencing**

#### **1.1.4.1 The Experiment**

When the project started, chromatin immuno-precipitation followed by sequencing (ChIP-seq) (Johnson et al. 2007) is the de facto standard for the identification of TF binding in cells *in vivo*. It is the combination of two techniques, namely chromatin immuno-precipitation (ChIP) and high throughput sequencing. Figure 1.4 consists of the steps for ChIP-seq. First, formaldehyde is added to crosslink the protein to the DNA

which would enable the proteins to stick to the DNA while the fragments are being sonicated in step two. Subsequently, antibody for the target TF is being added to separate the DNA fragments bound by target TF from other fragments. After that the extracted fragments are decrosslinked to remove the protein from the DNA and sequenced by a sequencing machine.



**Figure 1.4 Steps in ChIP-seq assay.**

#### 1.1.4.2 Mapping

Subsequently reads have to first be mapped onto the genome. The reads are typically of the order of tens of millions of length 30-80 bp. Mapping is a process of converting a sequence into a location in the genome. The main idea behind mapping is that reads that

are long enough eventually will be uniquely represented on the genome. The longer the sequenced read, the higher the chance that the read will be uniquely identifiable in the genome. Typically, sequence mappers will allow for a number of mismatches due to sequencing error and mutations in DNA relative to the reference genome. Some of the popular mappers are BWA (Li et al. 2009), Bowtie (Langmead et al. 2012) and BatMis (Tennakoon et al. 2012).

### **1.1.4.3 Peak Calling**

The objective of ChIP-seq is to identify regions in the genome that have been bound by target TF. Peak callers do so by identifying regions in the genome with mapped read counts that are statistically over-represented relative to a background run in which ChIP is not performed. The local maximum of a statistically over-represented region is called a peak. Examples of popular peak calling programs includes SISSRs (Narlikar et al. 2012), MACS (Zhang et al. 2008) and CCAT (Xu et al. 2010).

## **1.2 Research Problems**

### **1.2.1 Identifying co-TF from ChIP-seq datasets**

#### **1.2.1.1 Description**

Once the ChIP-seq of a particular TF is generated, we could try to answer the question of what are the potential cofactors which frequently cooperate with the TF without further performing experiments. The reason is because cofactors reside in the same cis-regulatory modules and hence the motifs occur closely around one another. Therefore, we expect to be able to find comotifs at close vicinity ( $\pm 500\text{bp}$ ) around the peaks. There

are three approaches to this problem: 1) perform *de novo* motif finding algorithms to see what motifs are enriched and then match the motif to motif database, 2) scan the sequences using all motifs from the motif database and look for those that are enriched, 3) scan the sequences using all motifs from the motif database and look for sets of motifs that that are combinatorially enriched.

### **1.2.1.2 Literature Review of Existing Methods**

#### **1.2.1.3 Large scale *de novo* motif finding**

By performing *de novo* motif finding on the set of ChIP-seq peaks, we would identify ChIPed TF motif as well as the motifs of its co-factors. However because of the large size of ChIP-seq data sets, most of the older generation *de novo* motif finders such as MEME and YMF that were designed for promoter analysis could not handle such big datasets and typically only restricted to just the top binding sites and therefore unable to comprehensively obtain all the co-TFs that the primary TF potentially works with. Recent *de novo* motif finding algorithms that are tailored toward large datasets include MDScan (Liu et al. 2002), Trawler (Ettwiller et al. 2007), Amadeus (Linhart et al. 2008), DREME (Bailey 2011) and CisFinder (Sharov et al. 2009). These tools typically tapped on data structures such as suffix tree, suffix array and hash tables to speed up searches.

##### **1.2.1.3.1 *Motif Scanning followed by Motif Enrichment Scoring***

Instead of performing *de novo* motif finding, using the various motif databases available such as TRANSFAC and JASPAR, we can perform motif scan to predict cofactors whose motif is available.

This process typically consists of two phases: motif scanning and motif enrichment scoring. We will explore the existing tools that can be used for this purpose. In the subsequent section we will first separately look at the various ways of performing motif scanning and motif enrichment scoring.

#### *1.2.1.3.1.1 Variations in Motif Scanning*

Section 1.1.3.2 describes one of the ways of performing motif scan based on PWM. To recap, to determine whether a particular string of nucleotides belongs to the motif model, first the log-likelihood score is being computed. We can do likewise for each position in a relevant background (e.g. a set of genomic or promoter sequences). The stringency can be set by deciding the number of occurrences we expect to observe in the background, say 1 occurrence per 10000 bp. The score corresponding to this expected occurrence is used as the cutoff. In this case, we say the pvalue cutoff is 0.0001. In the method described in Section 1.1.3.2, we assumed a background probability for the nucleotides to be uniformly 0.25 each. Another approach is to estimate the background nucleotide probability based on the individual counts of the nucleotides in the genome of the studied organism. The rationale for this is that in a genome that is GC rich, it is not surprising to observe a GC rich motif, and hence we would expect a greater penalty in the denominator. Though it is more accurate to account for, it suffers from having to use a separate formula when applying to different contexts and we shall later see that the choice of background, whether promoter or global plays a part as well. The method in Section 1.1.3.2 is preferred because of its simplicity and also background biases will be accounted for by using the p-value as the cutoff.

Other than using p-value as cutoff, other scoring method includes TRANSFAC Match which tries to minimise false positive and negative using the set of positive and negative sequences supplied for training. Firstly, the log-likelihood score are rescaled linearly such that the minimum and maximum achievable score corresponds to 0 and 1. To determine if there is a match, TRANSFAC also tracks the score for the most informative 5bp segment in the entire motif called the core of the motif. For each of the positive and negative sequences supplied during training, the highest score is being computed and corresponding core score of the best alignment recorded. Using these scores TRANSFAC Match computes separate cutoffs that minimise false positive rate, or false negative rate, or the sum of these rates.

#### *1.2.1.3.1.2 Common Motif Enrichment Scoring*

To determine whether a particular motif is interesting within a set of sequences, we make use of p-value. To explain what p-value is, we make use of a coin-flipping example. Suppose we flipped 10 coins and 9 of them turned out to be heads. Had each of the coins been a fair coin, the chance of observing at least 9 heads would have been  $P(9 \text{ heads}) + P(10 \text{ heads}) = \binom{10}{9}0.5^90.5 + \binom{10}{10}0.5^90.5 \approx 0.0107$ , which is rather unlikely. This therefore gives us evidence to suspect that not all the coins are fair coins. In this case, 0.0107 is our p-value and the smaller this value, the more “surprising” our observation and the more evidence we have to suspect that at least some of the coins are biased towards heads. P-value is computed with respect to a null hypothesis which is a reasonable model to explain the observation when there is nothing special. In our coin

flipping example, our null hypothesis is the background model that all the 10 coins are fair, with a 50% probability of coming up head.

Motif enrichment analysis in general consists of three steps: 1) perform motif scan to determine motif hits, 2) compute enrichment score and 3) report enriched motifs.

The enrichment score used by motif enrichment analysis tools is usually a p-value.

### ***Binom1***

Total number of motifs in input sequences follows the distribution  $Binomial(N, p)$

where

$N$  = Total length of input sequences

and

$$p = \frac{\text{Number of motifs in background sequences}}{\text{Total length of background sequences}}$$

### ***Binom2***

Total number of input sequences with motifs follows the distribution

$Binomial(N, p)$

where

$N$  = Number of input sequences

and

$$p = \frac{\text{Number of background sequence with motif}}{\text{Number of background sequences}}$$

### ***Hypergeom***

Total number of input sequences with motif follows the distribution

$$\text{Hypergeometric}(N, K, n, k)$$

where

$N$  = Total number of sequences which includes input and background sequences

$n$  = Number of input sequences

$k$  = Number of input sequences with motifs

#### ***1.2.1.3.1.3 Existing tools***

The existing tools differ in the way motif scan and motif enrichment scoring are being performed. Moreover, before high throughput experiments such as ChIP-seq and ChIP-chip became popular, genomic analysis is restricted to the promoter of genes, typically guided by gene co-expression obtained by microarray experiments. Promoter of groups of genes that are co-expressed will be subjected to motif enrichment analysis to predict candidate TFs that explain the co-expression. Up till recently, most motif enrichment tools had been restricted to just the analysis of promoters.

Table 3 shows a list of motif enrichment tools comprising of ConTra (Hooghe et al. 2008), CORE\_TF (Hestand et al. 2008), oPOSSUM (Ho Sui et al. 2005), PASTAA (Roeder et al. 2009), GATHER (Chang et al. 2006) and CEAS (Ji et al. 2006).

**Table 3** List of existing motif enrichment tools

WebTools	Promoter/ Genomewide	Motif Scanning	Motif Enrichment	Background
ConTra	Promoter	TRANSFAC MATCH	Binom1	Selected Promoters
oPOSSUM	Promoter	TRANSFAC MATCH	Binom1	Selected Promoters
PASTAA	Promoter	TRANSFAC MATCH	Binom1	Selected Promoters
GATHER	Promoter	TRANSFAC MATCH	Hypergeom	All Promoters
CEAS	Genomewide	EVALUE	Binom1	Whole Genome
CORE_TF	Promoter/ Custom Sequence	TRANSFAC MATCH	Binom1,Binom2	Selected Promoters/ Custom Sequence

Though these tools could be used for motif enrichment analysis, most of them will provide interface for further analysis. For example, ConTra, oPOSSUM and PASTAA allow the user to zoom into the promoter regions containing motifs. ConTra and CORE\_TF will use the TF candidates reported for further analysis.

#### 1.2.1.4 Summary

*De novo* motif finding in general is time consuming especially for large number of datasets and has difficulty in detecting subtle signals. As for motif enrichment, there are no hard and fast rules of how to perform and there are lots of parameters that average biologists have difficulty understanding and fiddling with. Some of these parameters are 1. the background (which models the non-binding sites), 2. the enrichment window size (which models the distance between the co-TF and the peak), and 3. the PWM score (Stormo 2000) cut-off (which determines if a site can be bound by the co-TF or not).

To provide biologists with an easy to use interface to predict candidate TF binding in target set, we develop CENTDIST in CHAPTER 2.

## **1.2.2 Identify Differential Motif Enrichment between two sets of ChIP-seq peaks**

### **1.2.2.1 Description**

In certain situations, we are interested in comparing two sets of genomic regions for differences in motif that could explain the underlying differential properties distinguishing the two sets. For example if we had two sets of ChIP-seq peaks that were performed under two different conditions, such as under two types of drug treatments, we could perform differential motif analysis to help predict the transcription factors that are active in each scenario. This can be done in several ways: 1) Discriminative *de novo* motif finding, 2) Using motif enrichment tools to compare the list that are filtered based on certain cut-off criteria, 3) Perform motif scanning and derive a differential motif enrichment score based on the counts.

### **1.2.2.2 Literature Review of Existing Methods**

#### ***1.2.2.2.1 Discriminative De Novo Motif Discovery***

Unlike usual *de novo* motif discovery algorithms as illustrated in Section 1.1.3.4, discriminative motif finding algorithms take into consideration of an additional negative set of sequences to be discriminated against. For a motif to be enriched, not only must it occur frequently in the positive set, it should also not occur frequently in the negative set.

Discriminative motif finders are typically designed to compare bound set with an unbound control set (Barash et al. 2001; Smith et al. 2005; Elemento et al. 2007; Redhead et al. 2007; Bailey 2011). Recently contrast motif finder (CMF) (Mason et al. 2010) has been developed to compare two sets of arbitrary binding peak sets and identify context-dependent motifs.

#### ***1.2.2.2.2 Motif Scanning Based***

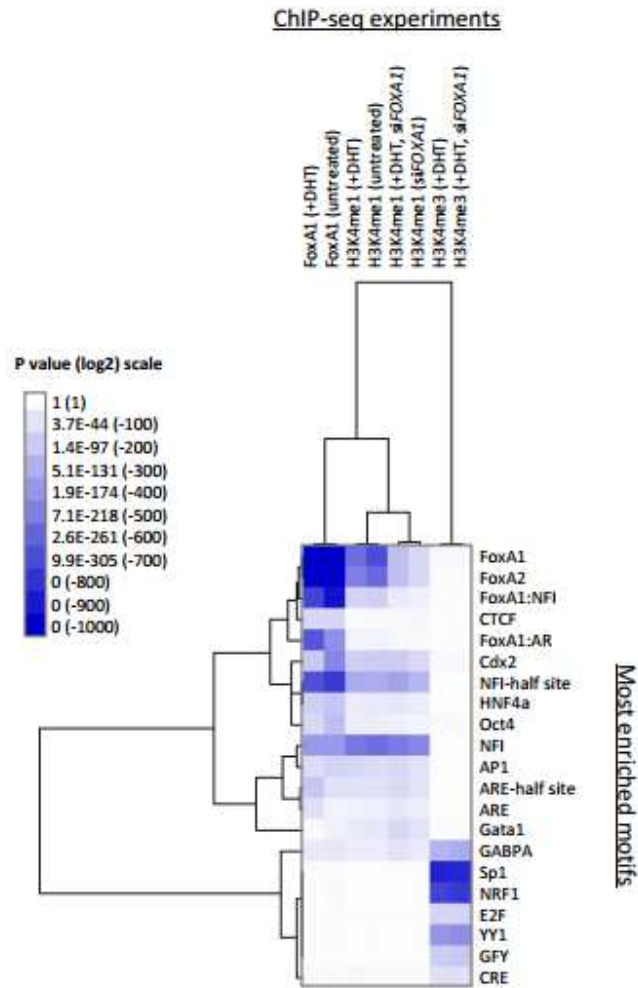
Suppose we are given two sets of ChIP-seq datasets and a database of PWMs, as with for single set, we would first perform motif scan for each set of ChIP-seq dataset for each PWM. General ways of identifying differentially enriched motifs: 1) Applying Motif Enrichment on each set and compare reported results 2) Applying Motif Enrichment on each set and comparing the scores reported.

##### *1.2.2.2.2.1 Comparing Motif Enrichment Result List*

We perform motif enrichment as per section 1.2.1 for each data set. Using the default cutoff for the motif enrichment analysis to determine enriched motifs in each set, we then decide whether a motif is differentially enriched by checking whether they appear in the other set.

##### *1.2.2.2.2.2 Comparing Motif Enrichment Scores*

After performing motif enrichment as per section 1.2.1 for each data set. Instead of cutoff, we compare the enrichment score reported. This approach is most popular and has been adopted by Wang et al. (2011) and represented by a heat map (see Figure 1.5).



**Figure 1.5 Example differential motif comparison illustration using heatmap based on enrichment pvalue score.**

### 1.2.2.3 Summary

Though several *de novo* based algorithms exist to identify differential motifs in two sets, no tool exists that is based on PWM database scanning. As such, we proceed to develop MOTIFDIFF in CHAPTER 3.

## CHAPTER 2     **CENTDIST – Web-based tool for Motif Enrichment**

This chapter describes CENTDIST, a user friendly tool we developed to identify colocalized motifs. This is a joint work between Zhang Zhizhuo and me. Parts of the material covered in this chapter were originally published in (Zhang et al. 2011).

### **2.1 Introduction**

With the revolutionary improvements of high throughput sequencing technologies, ChIP-seq has become increasingly affordable and effective to the extent of becoming the de facto standard for identifying the genome-wide binding profile of a particular TF in-vivo under a specific cell condition. Large amounts of high quality cistromic data rapidly produced by the biology research community calls forth effective, efficient and easy-to-use computational tools so that biologists can easily perform useful computational analyses without requiring much computational knowledge. One such important analysis is the identification of co-TFs, the cooperative partners in gene transcriptional regulation. Recent advances in ChIP-seq and the wide adoption of the technology in mapping TF binding sites have allowed researchers to identify novel co-TFs (Johnson et al. 2007).

Currently, co-TFs of a selected TF are identified in the following manner. First, a peak calling program such as MACS (Zhang et al. 2008) or CCAT (Xu et al. 2010) is used to determine which peaks in the ChIP-seq data are binding sites. Next, candidate co-TFs are predicted by examining if their motifs (position weight matrix, PWM) are enriched near the ChIP-seq peaks after normalizing against a chosen background model. TFs with

enriched motifs are classified as potential co-TF candidates and subsequently validated experimentally. This approach, known as the enrichment based method, has been widely used to identify novel co-TFs in web-based programs such as CEAS (Shin et al. 2009), CORE\_TF(Hestand et al. 2008), ConTra (Hooghe et al. 2008), and oPOSSUM(Ho Sui et al. 2007). However, there are occasions when this approach fails to find co-TFs. This is because the accuracy of enrichment-based methods is highly dependent on several user-specific parameters including: 1. the background (which models the non-binding sites), 2. the enrichment window size (which models the distance between the co-TF and the peak), and 3. the PWM score (Stormo 2000) cut-off (which determines if a site can be bound by the co-TF or not). Since different co-TFs require different parameters, existing methods can only identify co-TFs that satisfy the parameters specified by the user. This restriction thus limits the accuracy of existing methods. To avoid this problem, it would be ideal to have a method that does not require the user to specify a background while the method automatically estimates the enrichment window size as well as the PWM score cut-off for every co-TF.

Accurately predicting the co-TFs of a particular TF from a ChIP-seq experiment is computationally challenging because some co-TFs may occur infrequently while the location of others are less certain than that of the ChIPed TF (Chromatin Immunoprecipitated TF in ChIP-seq experiment). Previous reports suggested that motifs of co-TFs are enriched around ChIP-seq peaks (Wederell et al. 2008; Sharov et al. 2009). Several studies also showed that if two TFs are co-associated, their ChIP-seq peaks (or their binding sites) are not only in close proximity with each other, but the relative

distance of each TF with respect to the other exhibits a peak-like distribution (Chen et al. 2008; Cheung et al. 2010; He et al. 2010) We call this property the center distribution. Herein, we examine whether center distribution can be utilized for co-TF discovery.

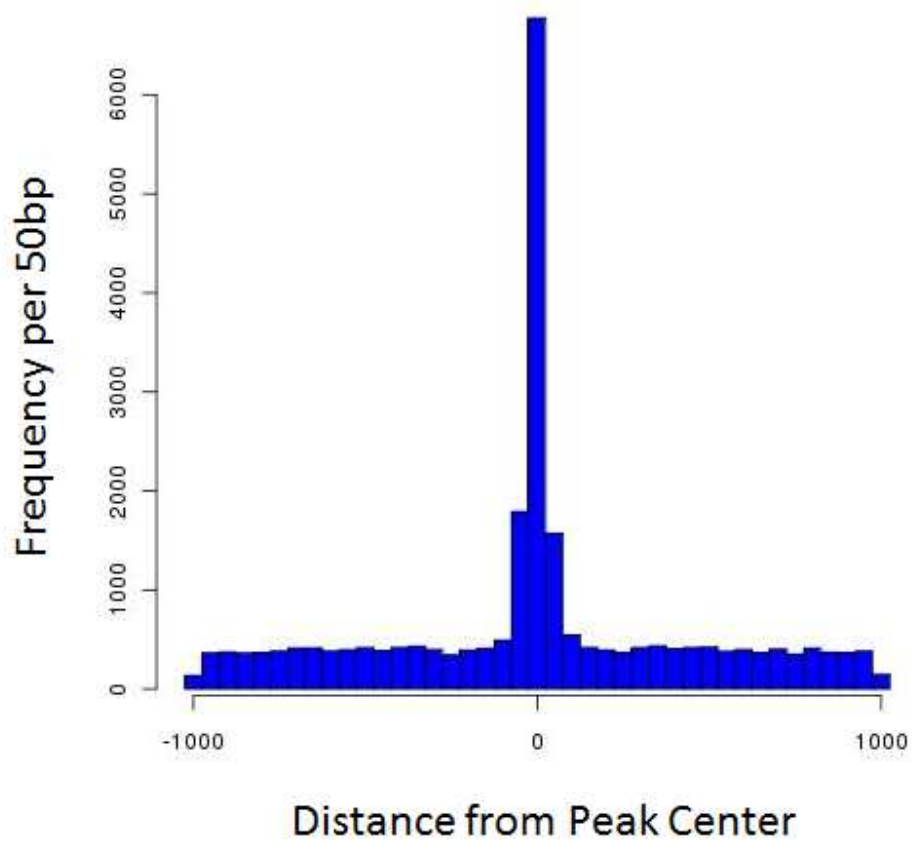
## **2.2 Results**

### **2.2.1 Development of CENTDIST**

We utilised our inhouse generated AR ChIP-seq and its cofactor FOXA1 ChIP-seq to gain some insights on true co-motif distribution. We first look at the AR motif distribution around ChIP-Seq peaks. Figure 2.1 shows the distribution of AR motif (V\$AR\_02) around the AR ChIP-seq peak. Strong motif enrichment can be seen within 100bp away from the peak center, peaking at the ChIP-seq peak center and practically flat elsewhere.

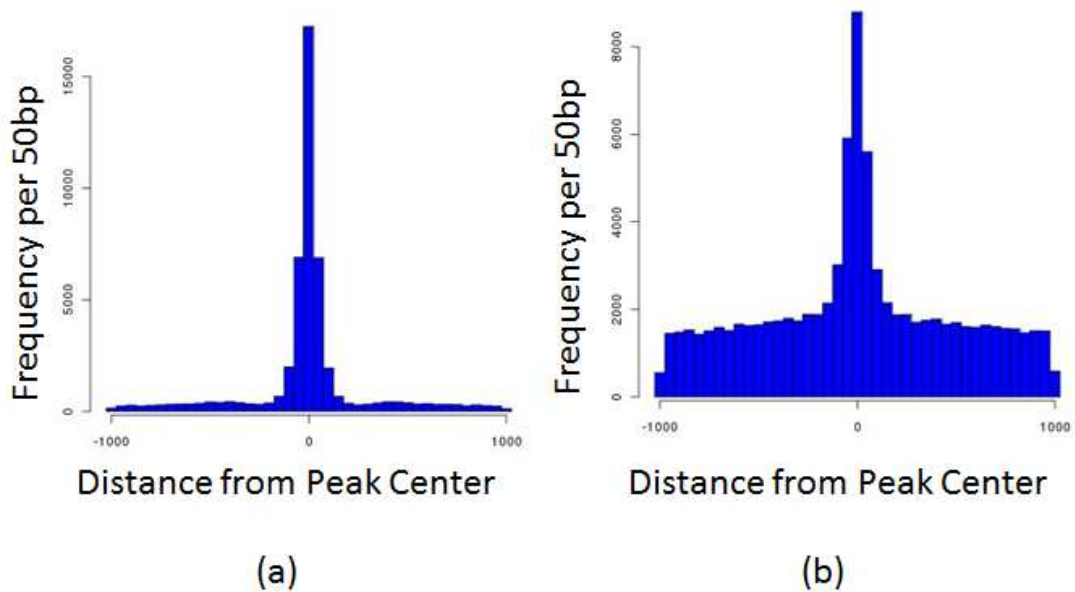
We plotted the histogram for both FOXA1 ChIP-seq peak (Figure 2.2a) and FOXA1 motif (Figure 2.2b) around AR ChIP-seq peaks. As compared to AR motif, similar enrichment can be observed for FOXA1 motif peaking at the AR ChIP-seq peak center though the width of enrichment is wider, extending to 200bp from AR ChIP-seq peak center. In addition, we observe that the distribution of FOXA1 motif closely resemble the distribution of FOXA1 ChIP-seq peak, albeit with higher background, showing that motif is a good predictor for the distribution of actual TF binding.

This observation of co-motif distribution is what we called center distribution and we proceed to develop an algorithm called CENTDIST around this idea.



**Figure 2.1 AR motif distribution around AR ChIP-seq peaks.**

The histogram of AR motif matches (V\$AR\_02) around AR ChIP-seq peaks, binned at 50bp intervals. Motif enrichment can be observed within 100bp from peak center.



**Figure 2.2 FOXA1 ChIP-seq and motif distributions around AR ChIP-seq peaks.**

a) Histogram of FOXA1 ChIP-seq peak around AR ChIP-seq peak. b) Histogram of FOXA1 motif matches (V\$HNF3ALPHA\_Q6) around AR ChIP-seq peaks, binned at 50bp intervals.

### 2.2.1.1 Algorithm behind CENTDIST

#### 2.2.1.1.1 *Strategy for Removing the Need for Secondary*

##### ***Parameters***

As previously mentioned, earlier methods require users to decide on the background model, the PWM cut-off and the proper enrichment window size for scanning. One of the aims of CENTDIST is to seek to take these responsibilities off users and simplify the analysis process by helping to select the optimal recommended parameters automatically. As such, users should only be required to input a set of genomic locations representing

ChIP-seq peaks (chromosome-peak summit position) and a list of candidate PWM motifs (provided by users or obtained from either the TRANSFAC (Matys et al. 2003) or JASPAR (Sandelin et al. 2004) databases representing co-TF binding sites. Utilising a fast motif scanning approach employed in CISFINDER (Sharov et al. 2009), we can scan for good matches of large number of motifs in a fraction of time compared with the naïve approach. This is possible because for the purpose of motif enrichment in high throughput data, we believe we should not go beyond an E-value (see Section 1.2.1.3.1.1 for description of E-value) of 0.001 which is roughly 3000000 hits in the entire human genome.

This allowed us to be able to scan over a larger region and focus on analysing the motif histogram over a larger window instead of just looking at the counts in a specific small window. We employed a score maximisation strategy to automatically determine the best parameters for the enrichment window and the motif score threshold. The score that CENTDIST tries to maximise, the frequency score, will be described in the next section.

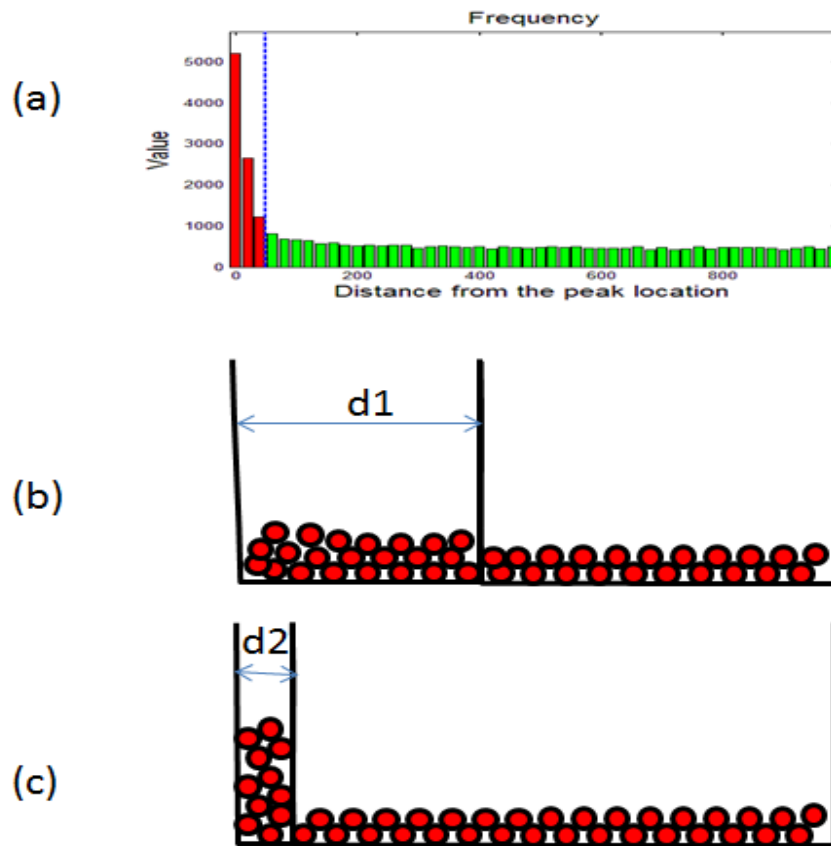
#### **2.2.1.1.2 *Frequency Score***

The core of the CENTDIST algorithm is the analysis of the frequency graph of TF motif matches with respect to the distance to the ChIP-seq peak center. As seen in Figure 2.1 and Figure 2.2(a) and (b), we observe symmetry about the ChIP-seq peak center indicating there is no directional bias of the motif with respect to the ChIP-seq peaks, equally likely to occur on either side. For the purpose of score computation, we consider only the magnitude of the distance from the peak center, ignoring the direction of

occurrence. We use a bin size of 20bp to obtain the histogram which can subsequently be represented by a vector of frequencies  $f_i$  for  $i=1,2,\dots,50$  corresponding to the count of motifs within the range of distance from ChIP-seq peak center  $20i-20$  to  $20i$  respectively, covering distances up to 1000bp from the ChIP-seq peak center.

After obtaining the frequency graph, we compute the frequency score,  $Z_{\text{frequency}}$  by the formula,  $Z_{\text{frequency}} = Z\left(m_i, \frac{m_0}{m_i+m_0}, m_i + m_0\right) = (m_i - m_0) \sqrt{\frac{m_i+m_0}{m_0 m_i}}$ , where  $m_i$  is the number of motif occurrences within a distance  $d$  (which will be varied to obtain the best score) from the ChIP-seq peak centers and  $m_0$  is the number of motif occurrences that are not within distance  $d$  from the peak center and the function  $Z(x,p,n) = (x - np) / \sqrt{np(1-p)}$  measures the number of standard deviations count  $x$  is from the expected value. This statistic is frequently used as the normal approximation to the probability of observing at least  $x$  successes in a binomial distribution with parameters  $n$  and  $p$ .

As an example, we look at the frequency graph of AR motif distribution about AR ChIP-seq peaks. (see Figure 2.3(a)) We represent the motif occurrences by balls and illustrate for two different partition sizes,  $d_1$ (see Figure 2.3(b)) and  $d_2$ (see Figure 2.3(c)), where the number of balls in each partition is derived from the frequency graph in Figure 2.3(a). The event represented by Figure 2.3(c) is harder to observe than that in Figure 2.3(b) (i.e. lower  $p$ -value). Hence, enrichment window  $d_2$  is chosen over  $d_1$ . The best such enrichment window will be determined by trying all possible distance  $d$ . Figure 2.3(a) shows the best enrichment window obtained, where the enriched region is colored red.



**Figure 2.3 Determining the frequency score of AR motif around AR ChIP-seq peaks.**

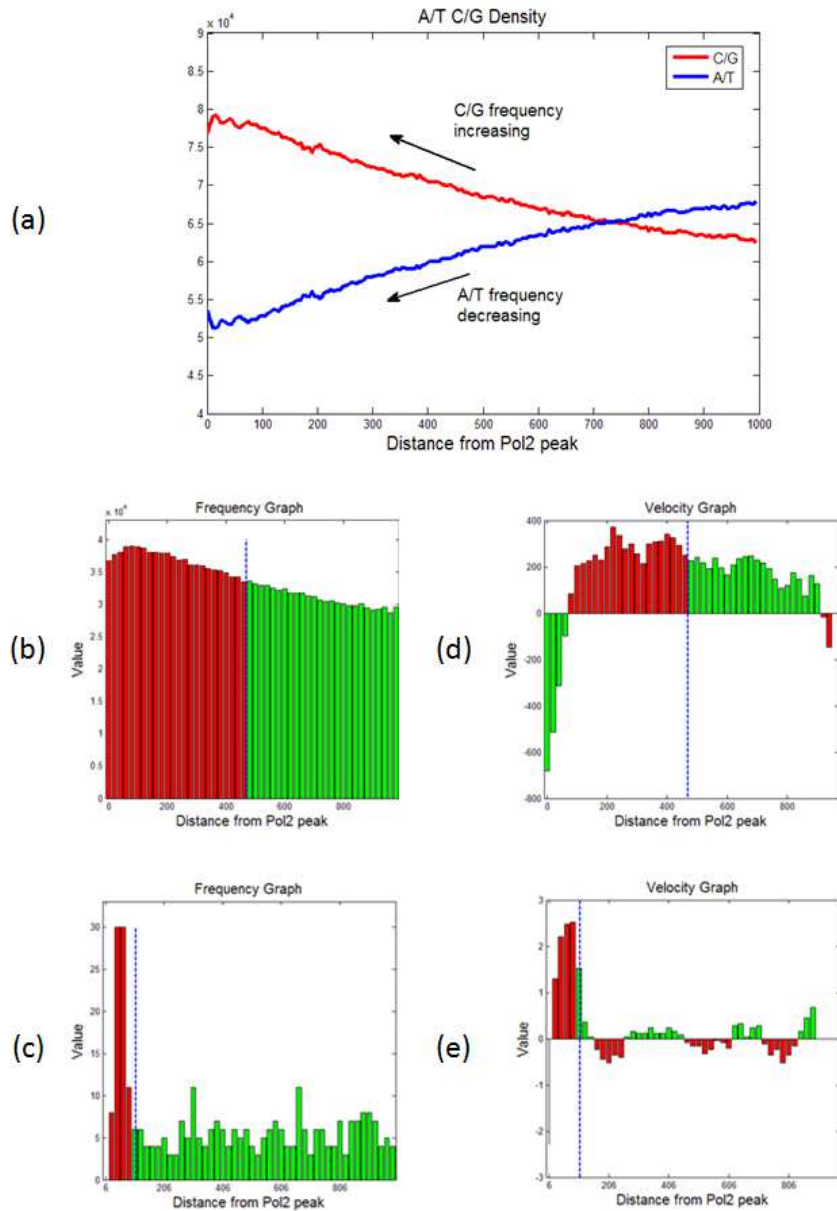
a) Frequency graph of AR motif around AR ChIP-seq peak (best scoring partition colored red). b) Treating motif occurrences as balls that are randomly dropped into two partitions  $d1$  and  $1000-d1$  with the probability of being dropped in the partition being proportional to the width of the partitions. c) Using a smaller enrichment window  $d2$ , we obtain better p-value (more unexpected), hence enrichment window  $d2$  is preferred over  $d1$ . The best such enrichment window will be determined by trying all possible windows.

### **2.2.1.1.3 Velocity Score**

For both the motif distribution of the primary TF, AR seen in Figure 2.1 and the co-TF, FOXA1 seen in Figure 2.2, not only do we observe enrichment of frequency near the peak center, we could also observe an increasing gradient towards the center. This is another feature that could be used to discern the differences among real motif and biological artifacts primarily brought about by GC or AT bias. We illustrate using RNA PolII ChIP-seq in K562 cells (Raha et al. 2010) (GEO accession numbers: GSM487431) as an example for which this feature is particularly beneficial. PolII, being the main protein responsible for transcription of genes are expected to bind near gene TSS. As such, it is situated within a promoter CpG island with much higher density of C/G then A/T (see Figure 2.4(a)) and hence is likely to be enriched by GC rich motifs, while the enrichment of TATA-box motif, a well-known binding anchor for PolII which is AT-rich is likely to be suppressed. To demonstrate, we examined the occurrences of the string pattern “CC” (not a real TF motif) and the TATA-box motif (Transfac ID: V\$TATA\_01) within the vicinity of the RNA PolII ChIP-seq peaks. Comparing the frequency distribution of “CC” (Figure 2.4(b)) and TATA-box (Figure 2.4(c)), using frequency score, we would have obtained a higher score for “CC”. However, we see that for the TATA-box motif, we observe increasing gradient towards the center, just like our model TF and co-TF distribution we looked at earlier, whereas the observation is absent for the false motif “CC”. To utilize this feature of increasing gradient towards the center, we first compute the velocity which is a smoothed gradient for the underlying frequency graph. Given the frequency graph represented by a vector of frequencies  $f_i$  for  $i=1,2,\dots,50$ . The velocity graph is defined to be the vector of differences  $v_i=(f_i-f_{i+5})/5$  for  $i=1,2,\dots,45$ .

Figure 2.4(d) and Figure 2.4(e) shows the velocity graph of “CC” di-nucleotide and TATA-box motif respectively. We see that the velocity graph is color-coded with red and green. Roughly speaking, the bars colored red are “good” velocities and the bars colored green are “bad” velocities. With respect to the best partition obtained for the frequency score, we define “good” velocities to contain the positive velocities within the enrichment partition and the negative velocities within the non-enrichment partition, and vice versa. Velocity score is highest when there is a decrease towards the center outside the enrichment partition and increase towards the center within the enrichment partition.

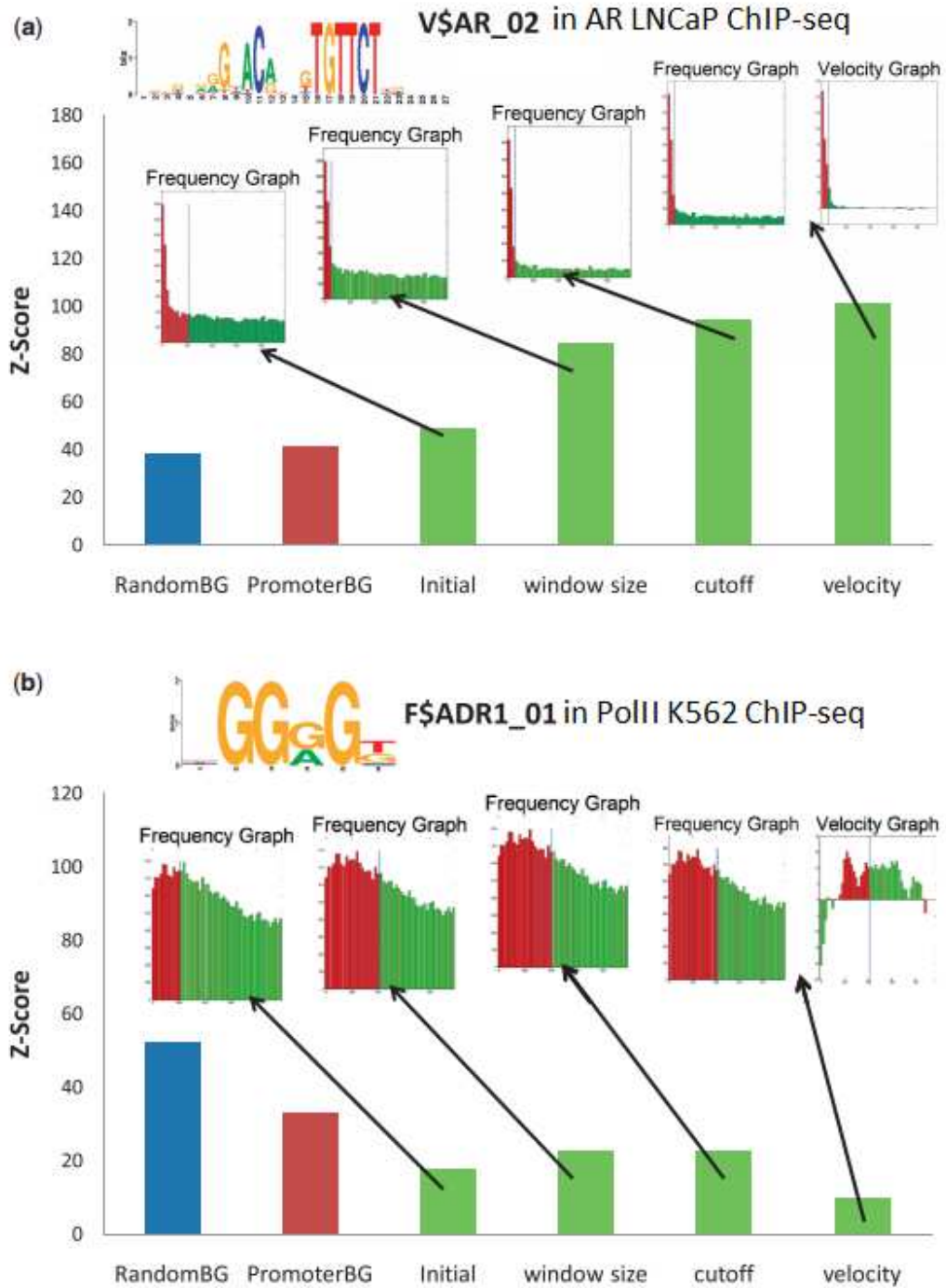
Denoting the sum of good velocity as  $G$  and sum of bad velocity as  $B$ , the velocity score is then calculated by:  $Z_{velocity} = Z(G, \frac{B}{B+G}, B + G) = (G - B) \sqrt{\frac{B+G}{BG}}$  where the function  $Z(x, p, n)$  is as described previously.



**Figure 2.4 Analysis of motifs around RNA PolII ChIP-seq peaks.**

(a) A/T or C/G density changes when approaching the PolII peak. (b) Frequency graph of the “CC” di-nucleotide distribution around PolII peak. (c) Frequency graph of TATA-box motif distribution around PolII peak. (d) Velocity graph of the “CC” di-nucleotide distribution around PolII peak. (Red are “good” velocities and green are “bad” velocities”) (e) Velocity graph of TATA-box motif distribution around PolII peak. (Red are “good” velocities and green are “bad” velocities”)

The velocity score serves to correct the frequency score biases due to CG (or AT) variation in the regions around the ChIP-seq peaks. The overall scoring function used by CENTDIST to assess motif distribution which is basically the sum of  $Z_{\text{frequency}}$  and  $Z_{\text{velocity}}$ , is hereby called the center distribution score. As a summary, Figure 2.5 demonstrates the capability of CENTDIST to promote true positive and repress false positive. To demonstrate the former, we consider the motif occurrence of V\$AR\_02 around AR ChIP-seq peaks. As shown in Figure 2.5(a), the Z-score progressively increases as we use flanking region as background (instead of promoter or random region), select the optimal window, the optimal PWM cut-off and finally considering the velocity. To demonstrate the latter, we study the CG-rich yeast TF motif, F\$ADR1\_01, which would have been determined incorrectly to be enriched around the PolII (RNA polymerase II) ChIP-seq peaks in human K562 cells using traditional approach. We know this motif is not actually enriched because PolII-binding sites are enriched for CpG islands, which are regions known to contain many CG repeats. As shown in Figure Figure 2.5(b), this motif has a modest center distribution score based on only the frequency score, but the final center distribution score was significantly lower after taking the velocity score into consideration.



**Figure 2.5 Demonstration of CENTDIST Capability**

(a) CENTDIST enhances the Z-score of the AR motif in the AR ChIP-seq data set (LNCaP cell line). The blue bar and red bar show the Z-scores of the AR motif computed using the traditional

enrichment method under the window size of 200 bp and the default PWM cut-off (1.32), respectively. The green bars show the Z-score of the AR motif computed by CENTDIST after it optimized different parameters. In the initial stage, the frequency Z-score was calculated using flanking regions at 200 bp as background and default PWM cut-off. In the second stage (window size), CENTDIST finds the best window size to maximize the Z-Score, in which the enrichment window size of AR is changed from 200 to 60 bp. In the third stage (cut-off), CENTDIST finds the best PWM cut-off to maximize the Z-Score, which leads to the flanking region noise level dropping significantly. In the fourth stage, CENTDIST combines the Z-scores of both the frequency graph and the velocity graph, thus further increasing the Z-Score. (b) CENTDIST can repress the Z-score of the false CG-rich motif in the PolII ChIP-seq data set compared to the traditional overrepresentation methods. All Z-scores are computed exactly as in (a). Since CENTDIST considers the velocity graph of the false CG-rich motif, the combined Z-score of CENTDIST finally drops and is significantly lower than that computed by the traditional enrichment based method. As a side note, this figure also showed that random background can produce quite different results compared to promoter background, which highlights the difficulty of choosing a correct background in existing enrichment based methods.

### 2.2.1.2 Comparison with Existing Tools

To assess CENTDIST's performance with respect to CEAS and CORE\_TF, and also to discover potential new co-TFs of AR, we compared the performance of CENTDIST against two enrichment-based programs, CEAS and CORE\_TF, on our AR ChIP-seq dataset. To ensure version compatibility among TRANSFAC databases used, we restrict the matrices used in the comparison to only those in TRANSFAC 11.2 which was used by CORE\_TF while CEAS used an older version. CENTDIST was sensitive enough to discover AR and all seven known co-TFs within top 20 hits (first two columns in Table 4). Well-characterised AR co-factors such as FOXA1, Oct1 and Ets1, were among some of the highly ranked factors in our analysis (Cheung et al. 2010; He et al. 2010). This result was significantly better than CEAS, which failed to find 5 of the known AR co-TFs. To make sure that the failure to identify the TFs is not due to using an older version of

TRANSFAC, we checked that the database used by CEAS contain matrices for these TFs by generating genomic locations corresponding to perfect match of the individual matrices as input to CEAS for verification. CORE\_TF, optimized with a random background setting and 400 bp extracted window size, identified all known AR co-TFs, however, this was within the top 37 hits. AUC analysis also indicated that CENTDIST outperformed the other motif enrichment tools even under their best configurations (see Table 4).

**Table 4 The ranking of the known co-TFs of AR for each motif enrichment tool.**

	CENTDIST*	CENTDIST	CORE_TF prombg 200	CORE_TF prombg 400	CORE_TF prombg 1000	CORE_TF randbg 200	CORE_TF randbg 400	CORE_TF randbg 1000	CEAS 200	CEAS 400	CEAS 1000
AR	1	1	2	2	6	1	1	1	2	2	1
CEBP	14	14	12	16	25	20	15	7			
ETS	10	9	64	61	66	37	37	47			
FOX	2	2	1	1	1	2	2	2	1	1	2
GATA	12	10	10	13	12	16	12	14			
NF1	9	11	40	60	70	10	21	31	3	3	
NKX	7	8	11	5	2	12	4	3			
OCT	19	19	4	8	5	15	19	26			
AP4	25	21				65	70				
AUC†	0.9683	0.9683	0.91	0.8917	0.8742	0.9358	0.9375	0.9208	0.6854	0.6875	0.625

\* The output result of CENTDIST\* is ranked by the Z-score of frequency graph only.

† The AUC score computation excludes AP4.

The columns 4th-6th are the results for CORE\_TF using promoter background (default background for CORE\_TF) with window size 200-1000 respectively, and the column 7th-9th are the result of CORE\_TF using random genome

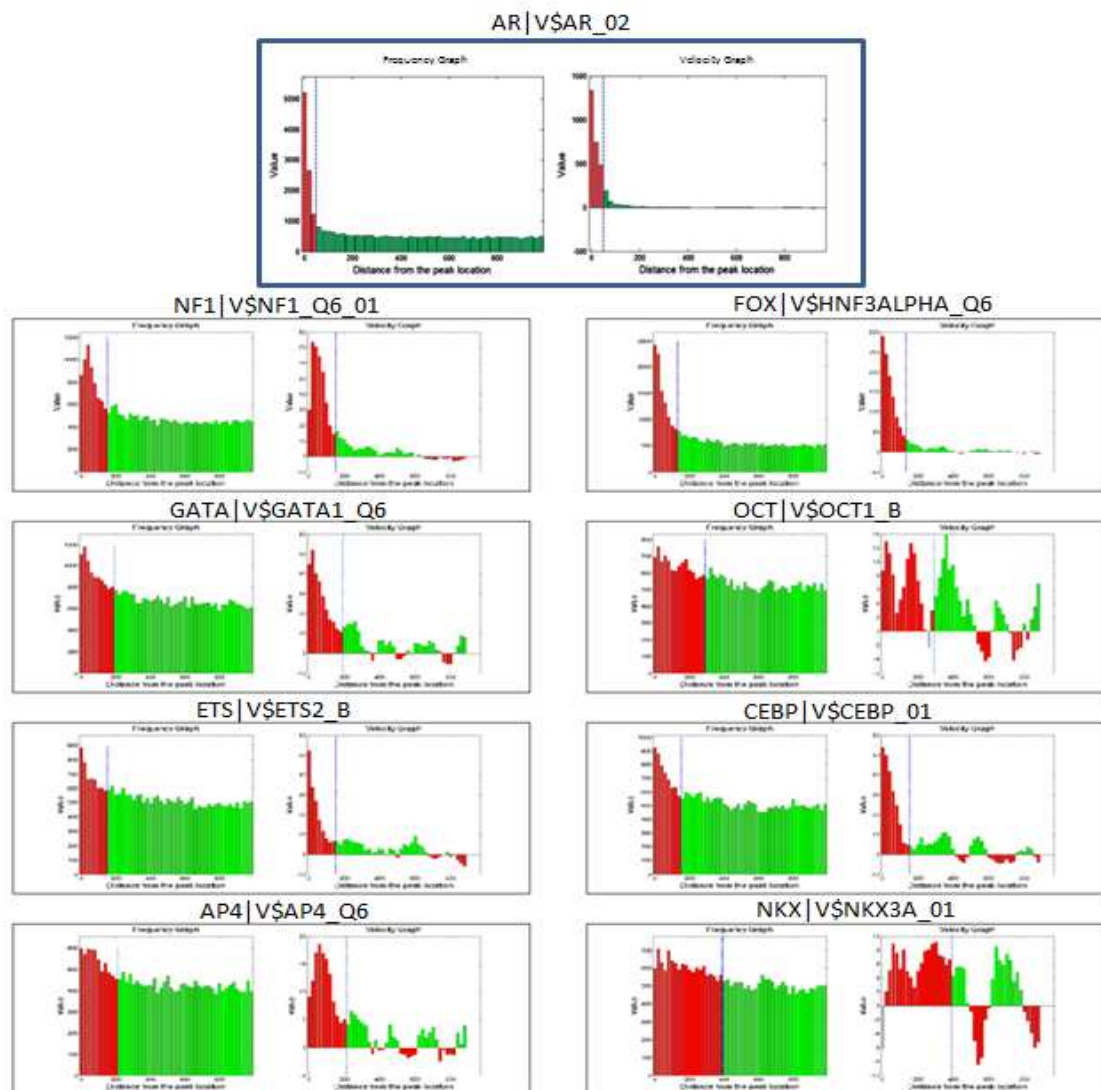
background with window size 200-1000 respectively. The last three columns are the results of CEAS with window size 200-1000 respectively.

CENTDIST discovered 10 co-TFs that were unique to the program. For five of these co-TFs, evidence from literature suggests that they play a functional role in prostate cancer development (Table 5). Among the other five for which link with prostate cancer is not well established, we focused on AP4 to validate it as a potential co-TF of AR. From Table 4 we could see that AP4 motif is ranked 21 in CENTDIST but is not reported within top 50 by the other enrichment tools.

**Table 5 Novel co-TF candidates of AR predicted by CENTDIST.**

Family	Best Motif	RANK	Function
CACCT	V\$AREB6_04	29	AREB6 also known as ZEB1 has a role in prostate cancer, which enhances transendothelial migration and represses the epithelial phenotype; ZEB1 and AR regulate each other to promote cell migration in triple negative breast cancer cells. (Park et al. 2000; Ergen et al. 2007)
BRCA	V\$BRCA_01	34	BRCA1/BRCA2 is BReast CAncer genes, and mutation in these genes increase risk of prostate cancer; BRCA1 is coactivator of the androgen receptor in both transfected prostate and breast cancer cell lines.(Chi et al. 1994; Maggiolini et al. 2004)
P53	V\$P53_02	42	p53 in prostate cancer: frequent expressed transition mutations. (Kang et al. 2004)
ERE	V\$T3R_Q6	13	Oestrogen receptor beta (ERbeta) is necessary for androgen-stimulated proliferation of LNCaP prostate cancer cells.(He et al. 2010)
CDX	V\$CDX_Q5	12	CDX can form complex with AR in LNCaP cell lines. (Cheung et al. 2010)

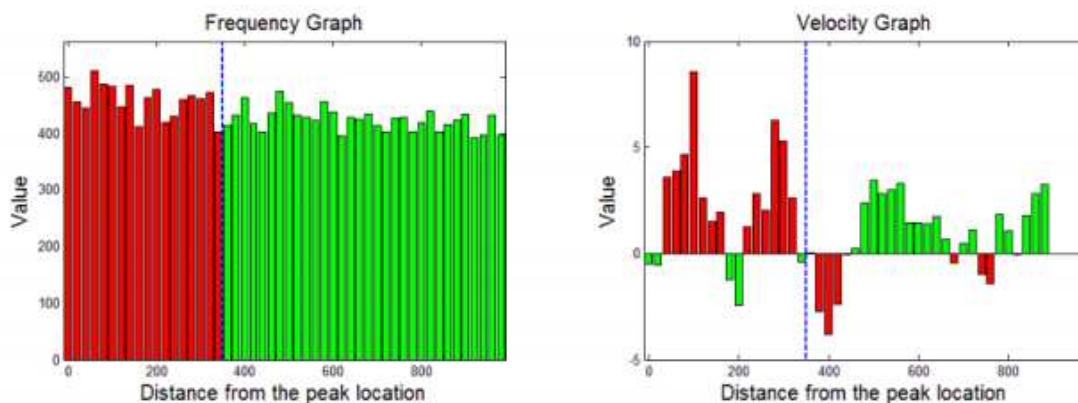
In Figure 2.6, we observed that both the frequency and the velocity of most co-TF motifs with respect to the AR peaks have good shape even though their enrichment were not as significant as that of AR. Taken together, our observation suggests that the frequency and velocity of co-motifs are useful information for determining true motif signals.



**Figure 2.6** Frequency and velocity graphs of AR and its co-TFs including the newly discovered AP4.

### 2.2.1.3 Validation of AP4 as a novel cofactor of AR

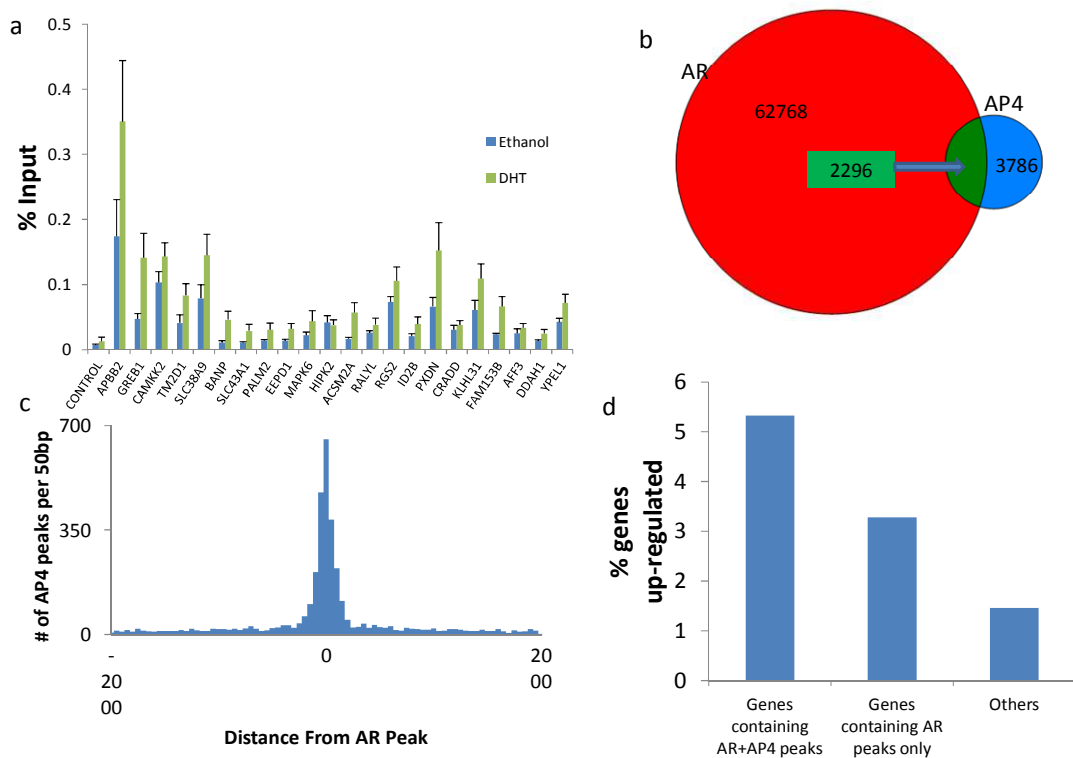
AP4 belongs to the basic helix-loop-helix (bHLH) family of transcription factors. It functions as a homodimer and is known to play important roles in colorectal cancer (Cao et al. 2009), however our understanding of this TF in prostate cancer is limited. To test if AP4 is a co-TF of AR, we randomly selected 22 AR ChIP-seq peaks that contain the AP4 motif and performed ChIP-qPCR in LNCaP cells treated with and without DHT. As shown in Figure 2.8(a), all 22 binding sites showed enrichment compared to the genomic control site, suggesting that AP4 is co-localized at AR binding sites. Furthermore, under DHT treatment (which recruits AR), the binding of AP4 was enhanced compared to vehicle (Ethanol) treatment. To further validate whether AP4 and AR are co-binding, we took an unbiased approach and performed a ChIP-seq of AP4. As shown in Figure 2.8(b), a large number (2,296 out of 6082/38%) of AP4 ChIP-seq peaks overlapped with AR. A distribution analysis of AP4 ChIP-seq peaks around AR binding sites confirmed that AP4 binds in close proximity (within  $\pm 200$  bp) to AR (Figure 2.8(c)). We scanned for the AP4 motif in the ChIP-seq peaks and found that 79.4% of the AR-AP4 overlapping peaks contain AP4 motif. In contrast, although 40.8% of the AR only peaks contain AP4 motif, the center distribution score for the AP4 motif around these peaks was low (see Figure 2.7).



**Figure 2.7 Uniform distribution of AP4 motifs around the AR only ChIP-seq peaks.**

We scanned the AR-only ChIP-seq peaks (excluding the peaks overlapping with AP4 ChIP-seq peaks) with AP4 motif and found the motif distribution look like a uniform distribution. The left panel is the frequency graph, which shows frequency uniformly distributed across the different distance range. The right panel is the velocity graph, which shows the different color velocities distributed equally.

Finally, we examined the fraction of androgen up-regulated genes near AR and AP4 peaks. Genes are defined as up-regulated if there exist at least one probe having a fold change of 1.5 or above at one of the three time points 3, 6 and 12 hours upon DHT treatment compared to vehicle treatment. We divided the genes into three groups: genes with AR+AP4 peaks, genes with AR only peaks, and genes with no AR peaks. We found that the proportion of up-regulated genes in group 1 is 1.6 fold and 3.7 fold more than that in groups 2 and 3, respectively (Figure 2.8(d)), suggesting that AP4 may co-localize with AR to directly up-regulate the transcription of androgen target genes.



**Figure 2.8 AP4 is a novel co-TF of AR.**

(a) ChIP-qPCR of AP4 was performed on 22 randomly selected AR peaks containing AP4 motifs in LNCaP cells before and after 2 h of DHT treatment. (b) Venn Diagram depicting the overlap between the ChIP-seq peaks of AR and AP4. (c) AP4 ChIP-seq peak distribution around AR ChIP-seq peak. (d) Association of up-regulated genes with binding sites containing AR+AP4, AR only, or others.

### 2.2.1.4 Validation of AP2 as a novel cofactor of ER


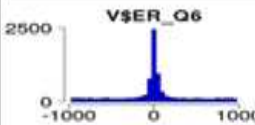

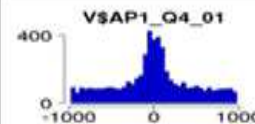

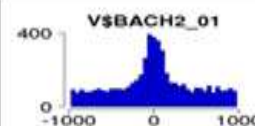

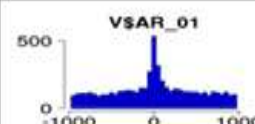



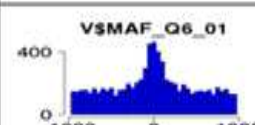



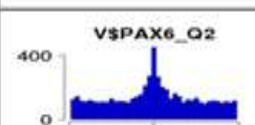

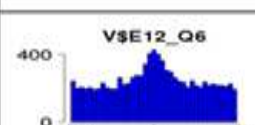

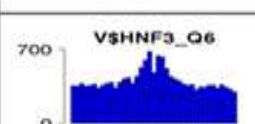
Parts of the results in this section are due to my fellow lab mate Tan Si Kee.

One of the nuclear hormone receptors that our lab studies extensively other than AR is estrogen receptor (ER). It is often highly expressed in breast cancer cells and plays an important role in the progression of breast cancer (Turner et al. 1998). In hoping to find

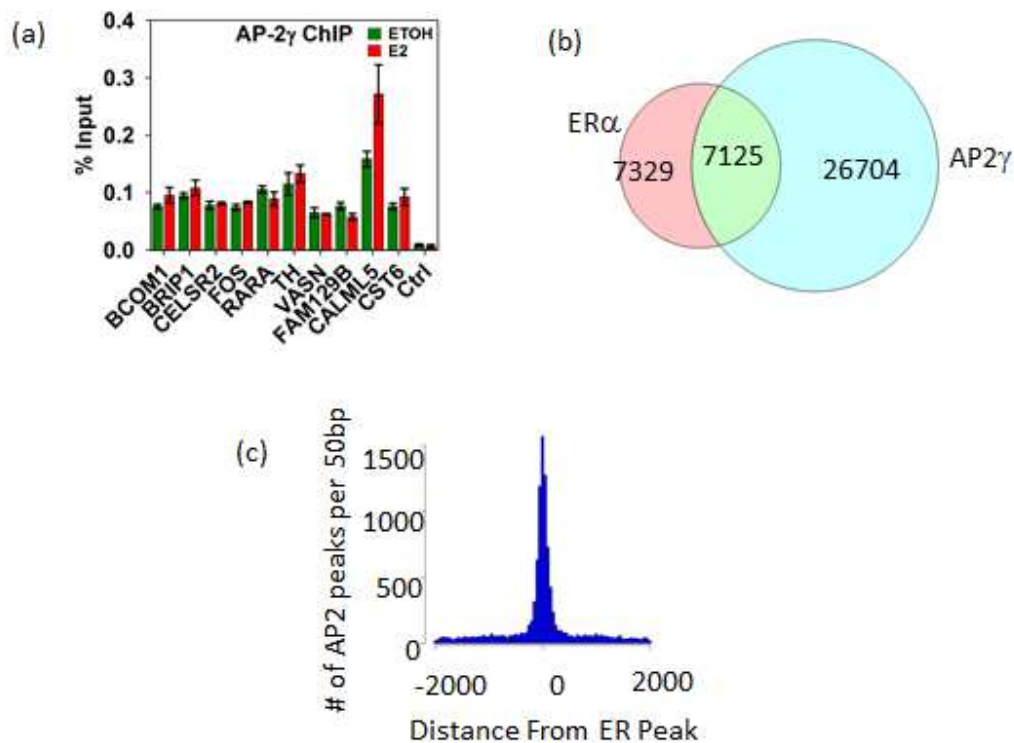
novel cofactors of ER, we performed CENTDIST on ER ChIA-PET (a technique similar to ChIP-seq) peaks (Fullwood et al. 2009).

Figure 2.9 shows the top 10 motif hits as reported by CENTDIST. CENTDIST ranks ERE top and also reported binding motifs of known collaborative factors of ER $\alpha$  such as AP-1, BACH1 and FOXA1. In addition we also found the AP-2 family of transcription factors to be highly enriched within the set of ER binding sites. This preliminary finding led us to investigate the role of AP-2 $\gamma$  in detail, specifically to address its functions in hormone-responsive breast cancers.

Of particular interest is the AP2 family which was ranked 5th. Activating Protein 2 (AP-2) is a family of transcription factors which is known to be mainly involved in the gene expression regulation during early developmental stage. Among the AP-2 family of transcription factors, AP-2 $\gamma$  is the most commonly expressed protein in breast cancer cells and has been known to be the main driver of mammary oncogenesis. Its role as a cofactor of ER, however, has yet to be discovered. We validated AP-2's co-localization with ER by performing ChIP-qPCR at ten ER binding sites which contain AP-2 motifs. Figure 2.10(a) shows that all ten binding sites have increased AP-2 $\gamma$  ChIP enrichment with respect to genomic control site. Moreover, it can be seen that its enrichment is independent of E2 treatment, suggesting that AP-2 $\gamma$  possibly plays some role prior to the binding of ER.

RANK	TF NAME	TF FAMILY	MOTIF LOGO	SCORE	DISTRIBUTION
1	V\$ER_Q6	ERE		164.191	
2	V\$AP1_Q4_01	AP1		53.2943	
3	V\$BACH2_01	BACH		48.8697	
4	V\$AR_01	AR		40.6783	
5	V\$AP2ALPHA_01	AP2		40.2189	
6	V\$MAF_Q6_01	MAF		39.9482	
7	V\$NRF2_Q4	NRF		35.7714	
8	V\$PAX6_Q2	PAX		34.3179	
9	V\$E12_Q6	EBOX		27.7194	
10	V\$HNF3_Q6	FOX		27.5104	

**Figure 2.9** Top 10 motif hits reported by CENTDIST showing the score and distribution of best motif within each TF family



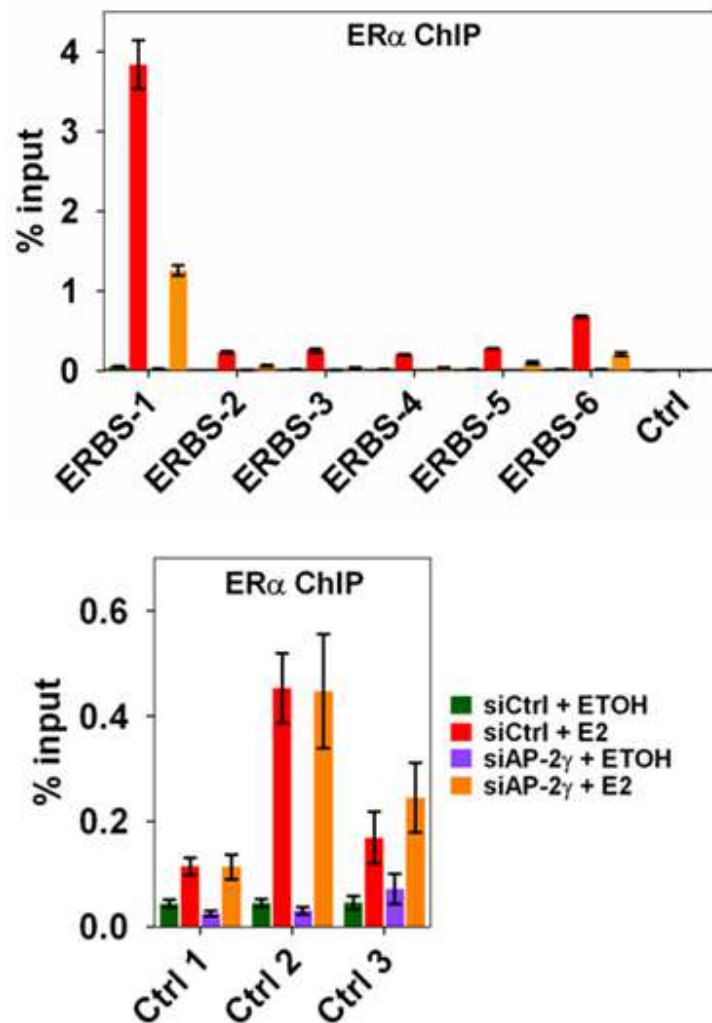
**Figure 2.10 AP-2 $\gamma$  is identified as a potential collaborative factor of ER $\alpha$ .**

(a) ChIP-qPCR of AP-2 $\gamma$  was performed on ten ChIA-PET ERBS with predicted AP-2 motifs in MCF7 cells before and after 45 mins of E2 treatment. (b) Venn Diagram showing the overlap between the ChIP-PET peaks of ER $\alpha$  and ChIP-Seq peaks of AP-2 $\gamma$ . (c) AP-2 $\gamma$  ChIP-seq peak distribution around ER $\alpha$  ChIA-PET peaks.

Following the successful ChIP-qPCR validation of AP-2 $\gamma$ , we performed a ChIP-seq of AP-2 $\gamma$ . Overlapping the AP-2 $\gamma$  ChIP-Seq binding sites with the ER ChIA-PET binding sites revealed that roughly half of all ER binding sites contain AP-2 $\gamma$ , and one-fifth of AP-2 $\gamma$  binding sites contain ER (See Figure 2.10(b)). Distribution analysis of AP-2 $\gamma$  ChIP-seq peaks around ER bindings sites confirmed that AP2 binds in close proximity (within  $\pm 200$  bp) to ER (See Figure 2.10(c)). We performed microarray with AP-2 $\gamma$  knockdown to identify genes that are regulated differently in the absence of AP-2 $\gamma$ . One such gene is REarranged after Transfection (*RET*) proto-oncogene. Our microarray shows

that RET is E2-induced, but became less activated upon the knockdown of AP-2 $\gamma$ . Another interesting thing about RET is that around RET gene locus, there are six ERBS and all six of them harbours AP-2 $\gamma$  motifs. Previous studies showed that *RET* expression was up-regulated by estrogen, which is verified by our microarray and correlates with ER $\alpha$  expression in primary breast tumors and cell lines (Frasor et al. 2003; Tozlu et al. 2006; Boulay et al. 2008), and that mutation of this gene has also been shown to be involved in the progression of thyroid carcinoma (Boulay et al. 2008).

To validate the observation from our microarray, we performed experiment to measure the mRNA expression of RET upon knockdown of AP-2 $\gamma$  and observed a significant drop in its expression when compared to control. Next we performed ChIP-qPCR on the six ERBS mentioned above to assess the effect of the presence of AP-2 $\gamma$  on ER $\alpha$  binding. As seen in Figure 2.11, all six binding sites show sharp decrease in ChIP enrichment upon the knock down of AP-2 $\gamma$ . This shows that AP-2 $\gamma$  is in fact required for the efficient binding of ER. This suggests a model for a mechanism of gene transcription by which AP-2 $\gamma$  is first present at the locus and serves to promote the recruitment of ER and subsequently regulate the transcription of the gene.



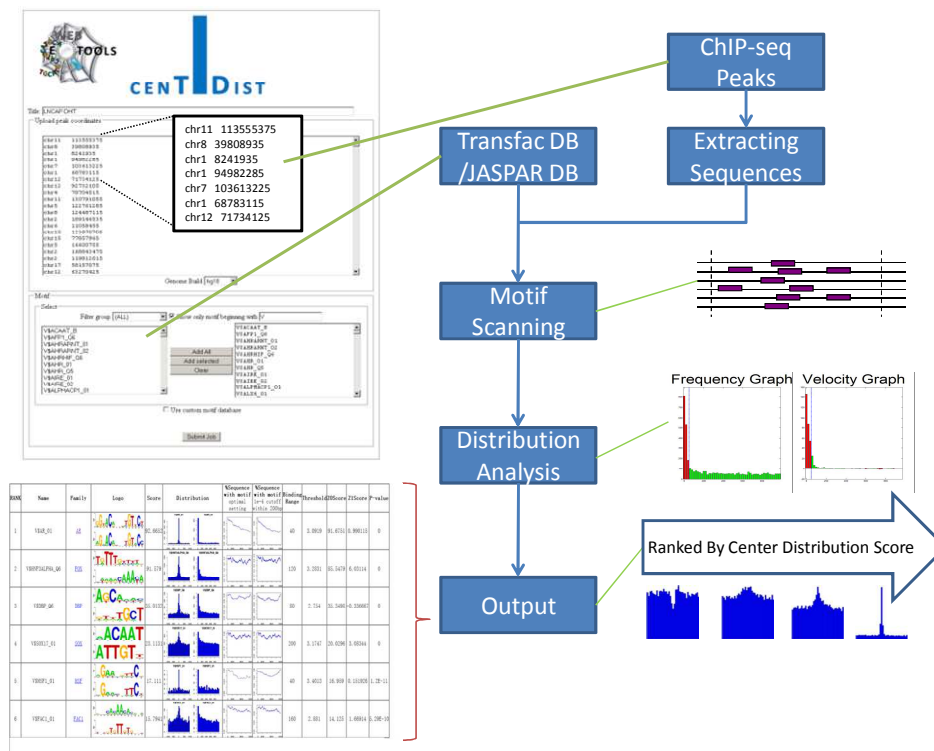
**Figure 2.11 AP-2γ is required for the efficient binding of ERα.**

ERα ChIP was performed on MCF-7 cells transfected with control or AP-2γ siRNA and treated with or without E2 for 45 mins. ERα binding was assessed at the *RET*-associated ERBS (top panel) and at control ERBS that do not coincide with AP-2γ binding (bottom panel)

## 2.2.2 CENTDIST Web Server

We developed a method called CENTDIST which can be accessible at <http://compbio.ddns.comp.nus.edu.sg/~chipseq/centdist/>. CENTDIST is a novel web-application for identifying co-localized transcription factors around ChIP-seq peaks based

on the skewness of their motif distribution around the peaks. The general pipeline of CENTDIST is shown in Figure 2.12.



**Figure 2.12 CENTDIST web interface and program procedure.**

Users can input or upload ChIP-seq peak locations or the bed format peak region data, and select the corresponding reference genome and the motif candidates (TRANSFAC, JASPAR, or custom database).

After submitting the job, the data will automatically be processed according to the CENTDIST analysis pipeline. Specifically, CENTDIST will scan the sequences ( $\pm 1000$  bp around the peaks) and obtain the occurrences of each PWM motif to generate the frequency graph and the velocity graph. Z-score is used to assess the enrichment around peaks for each graph. The center distribution score of each PWM motif is calculated as

the sum of the two Z-scores. Finally, CENTDIST outputs a list of TF families ranked by the center distribution scores.

CENTDIST is designed for analyzing high-throughput ChIP-seq data. Its web user interface contains three main parts: input, job management, and output. For input, CENTDIST accepts a list of ChIP-seq peaks. The ChIP-seq peak information can be formatted in the form of chromosome-position pairs or BED format genomic regions. CENTDIST is capable of supporting more than 1 million peak coordinates. The motifs used for scanning can be entered in the form of PWM or selected from either the TRANSFAC database (version 11.3), which contains 849 matrices or the JASPAR database, which has 459 matrices. CENTDIST also provides options for users to easily filter PWM motif candidates by string pattern, taxonomy, or transcription factor (TF) family. Finally, unlike other motif scanning programs, CENTDIST is totally parameter free. Users are not required to provide the background, the enrichment window size, or even choose the FDR or PWM cut-off for the PWM motifs. All these parameters will be estimated by CENTDIST automatically.

With regards to job management, submitted jobs will be sent to the job queue on the server and processed based on a first come first serve policy. Users can view the status of their submitted jobs, and access or delete the results of previous runs at the 'viewjob' page. The page refreshes automatically and email notifications will be sent to users once the jobs are completed.

The main output page for CENTDIST is a table containing PWM motifs ranked according to center distribution scores. Each row in the table presents the enriched TF family, and user can click on a link associated with each TF family to browse the result of each individual member. The output also contains visualization features like the PWM logo (Schneider and Stephens, 1990) of the motif, the frequency graph (center view and folding view), and other useful numeric features like binding range (the enrichment window size), PWM threshold (the cut-off that maximizes the center distribution score), center distribution score, and p-value. In addition, the output page provides the motif distribution across different peak ranks (column 7 and 8 in Figure 2.13), which is useful when the input peaks are sorted by some quality measure like ChIP-seq intensity.

Results for Run: Incap\_AR sample

VERSION: 2011.03.29

GO TF:

Show top  Families  Download As Text

Rank	Name	Family	Logo	Score	Distribution	%Sequence with motif optimal setting	%Sequence with motif 1e-4 cutoff within 200bp	Binding Range	PWM Score Cutoff	Z0Score	Z1Score	P-value	
1	V\$AR_01	AR		92.6652					40	3.0919	91.6751	0.990115	0
2	V\$HNF3ALPHA_Q6	FOX		91.579					120	3.2831	85.5479	6.03114	0
3	V\$DBP_Q6	DBP		35.0132					80	2.754	35.3498	-0.336667	0
4	V\$SOX17_01	SOX		23.1131					200	3.1747	20.0296	3.08344	0
5	V\$HSF1_01	HSF		17.111					40	3.4013	16.959	0.151926	8.23E-12

Figure 2.13 Sample Output page of CENTDIST.

## 2.3 Discussion

In this chapter, we presented a new computational method called CENTDIST that utilizes frequency information as well as slope information (velocity) to predict whether a motif is real or not. CENTDIST does not require an explicit background model. Using the velocity score, CENTDIST is also insensitive to CG- or AT-biases. Because CENTDIST automatically selects the optimal configuration, minimal expert knowledge is required by the user. From the ChIP-seq of AR in LNCaP cells, CENTDIST discovered AP4 as a novel co-TF of AR, which was missed by existing enrichment based methods. (Validation of AP4 is discussed in section 2.2.1.3). Other than AP4, CENTDIST also predicted 9 additional co-TFs that were missed by the other programs. For 5 of these co-TFs, evidence from literature suggests that they could be potential collaborators of AR.

A reason for the poor performance of existing motif enrichment tools is because of the heavy reliance on the selection of the proper background and other parameter settings. Choosing the correct background, however, is currently considered an art. What's more is that there is no one set of parameters that can satisfy all co-TFs. Stronger evidence of this point can be seen in the comparison of 14 ChIP-seq of mouse ES cells described in Zhang et al. (2011). Finally, the assumption that noise is uniformly distributed may not be true when CG (or AT) content varies in ChIP-enriched regions.

CENTDIST does have certain limitations. For example, CENTDIST may fail to identify co-TFs whose binding site distribution does not follow the proximity assumption (i.e. co-TFs that are not co-localized with the ChIPed TF). However, the latter would not be

found by traditional enrichment based methods either since their binding sites are not enriched.

CENTDIST is a user-friendly web-based application that is capable of analyzing large-scale ChIP-seq datasets. It can scan approximately seven hundred TRANSFAC motifs over a ChIP-seq dataset containing 10,000 peaks in only 10 minutes. With CENTDIST, users do not have to set any parameters except to upload the ChIP-seq peak locations and select the PWM motif library they wish to use for scanning. The output of CENTDIST contains clean and rich information for users. Specifically, it groups the list of enriched motifs into TF families, and provides other information including PWM logo, motif distribution graph, enrichment P-value, and the enriched window size of the enriched motifs.

To the best of our knowledge, CENTDIST is the first motif enrichment tool for ChIP-seq data that utilizes the shape information (velocity) of the motif distribution, and automatically detects the size of the motif enriched region and PWM score cut-off. It compares the enrichment inside/outside of the enriched motif region without the need for additional background information. Although there is still room to improve the methodology, this study opens a new door for utilizing the shape information to extract biologically meaningful co-TFs in a ChIP-seq data set.

## CHAPTER 3    **MOTIFDIFF – Web-based tool for Differential Motif Enrichment**

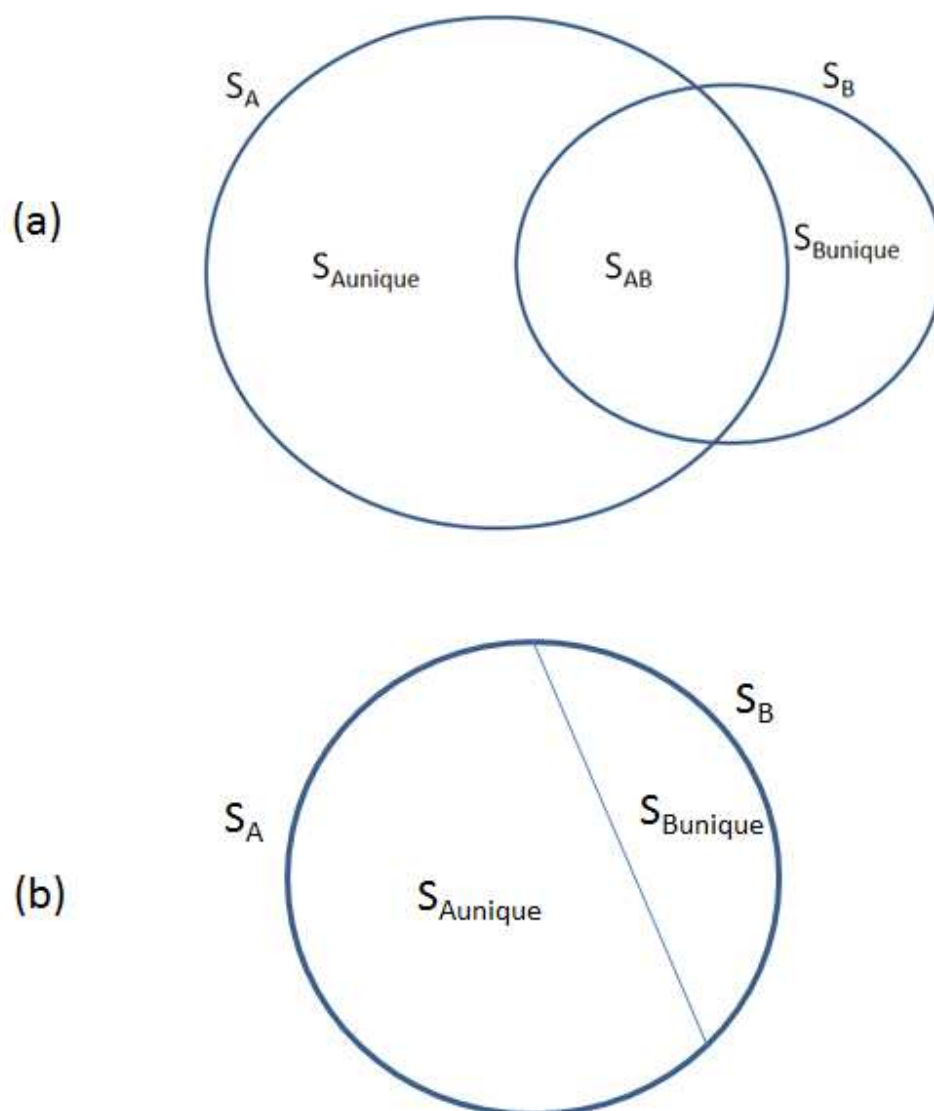
### **3.1 Introduction**

In the previous chapter, we developed a tool, CENTDIST that utilizes the imbalance of co-motif's distribution around ChIP-seq peaks to help identify potential co-TFs with high accuracy. The method is resistant to biases brought about by window size selection and choice of background. CENTDIST had been designed with the aim of identifying enriched motifs that would be difficult to detect by existing methods that relied on enrichment over a predetermined background.

Although CENTDIST is useful for finding enriched motifs, CENTDIST cannot help to compare if two sets of ChIP-seq peaks have different composition of motifs. Moreover, identifying differential motifs between two sets of ChIP-seq peaks is useful since it allows us to predict specific co-TFs that play a more significant role in one set than the other. In general, the target is to compare two sets of ChIP-seq peaks with mostly common properties, differing by some specific property. This analysis can then lead us to correlate the differential property of the ChIP-seq sets with the predicted differential co-TFs reported. For example, it could be ChIP-seq performed on a single TF in different signaling pathways and we want to determine what are the specific co-factors exclusively active in the respective cell lines; or perhaps we can partition a set of ChIP-seq peaks into two sets based on certain property such as whether the peaks are co-bound by a second TF (for which ChIP-seq data is available) and ask what are the specific co-factors that are involved in each partition.

The two examples described above are in fact two types of comparison. For the first type of comparison, we are comparing two sets of ChIP-seq peaks,  $S_A$  and  $S_B$ . Peaks common to both sets correspond to the common TF binding sites. Typically, we group peaks 500 bp apart as same peak, or we say the two peaks overlap. After we perform overlap on the two peak sets, we will in general obtain three sets of mutually exclusive peaks, namely: the set of peaks in  $S_A$  but not in  $S_B$  ( $S_{A\text{unique}}$ ), the set of peaks in both  $S_A$  and  $S_B$  ( $S_{AB}$ ), and the set of peaks in  $S_B$  but not in  $S_A$  ( $S_{B\text{unique}}$ ). (See Figure 3.1(a)) For the second type of comparison, we partition a single set  $S$  into two subsets  $S_A$  and  $S_B$  based on certain property. In this case, there are no overlapping peaks and therefore only two sets of mutually exclusive peak sets:  $S_{A\text{unique}}$  and  $S_{B\text{unique}}$  corresponding to  $S_A$  and  $S_B$  respectively. (See Figure 3.1(b))

For the first scenario, as the overlapping peak sets ( $S_{AB}$ ) are common in both sets and our purpose is to discern the difference among set  $S_A$  and  $S_B$ , these common regions are ignored in our analysis. However, suppose we are interested in for example comparing  $S_{A\text{unique}}$  and  $S_{AB}$ , it can be reformulated as a partition of  $S_A$  using the property of whether the peaks overlap with peaks in  $S_B$ . Upon doing so, both will just be a comparison between two mutually exclusive sets. The problem then is to search for interesting co-TFs that are significantly differentially enriched in  $S_{A\text{unique}}$  and  $S_{B\text{unique}}$ . This can be achieved by analyzing the motifs in  $S_{A\text{unique}}$  and  $S_{B\text{unique}}$ .



**Figure 3.1 Types of comparison.**

- a) The three mutually exclusive sets formed by overlapping  $S_A$  with  $S_B$  are  $S_{Aunique}$ ,  $S_{AB}$  and  $S_{Bunique}$ .
- b) The two mutually exclusive sets formed by partitioning peaks that satisfy a particular property versus peaks that do not.

One of the ways to perform this is by performing *de novo* motif finding that are specially designed to identify discriminative motif using some measure of separation between two sequence set. Such tools include CMF (Mason et al. 2010), DEME (Redhead et al. 2007), DME (Smith et al. 2005), FIRE (Elemento et al. 2007). Background CG/AT content of sequences could bias the occurrence of CG/AT rich motifs, resulting in spurious false positives, posing a large problem for *de novo* finders in general. The more recent tools, CMF attempts to account for such enrichment by normalizing against the respective background model using Markov models of individual sets. However we show using a simulated example in Section 3.3 that such consideration is still insufficient to take into account of biases due to background. Also, as usual, *de novo* motif finding is typically computationally intensive and time consuming and limited to finding motifs at within 100-200bp from the ChIP-seq peaks due to the sheer size when in general the interactions between TF can have much greater range.

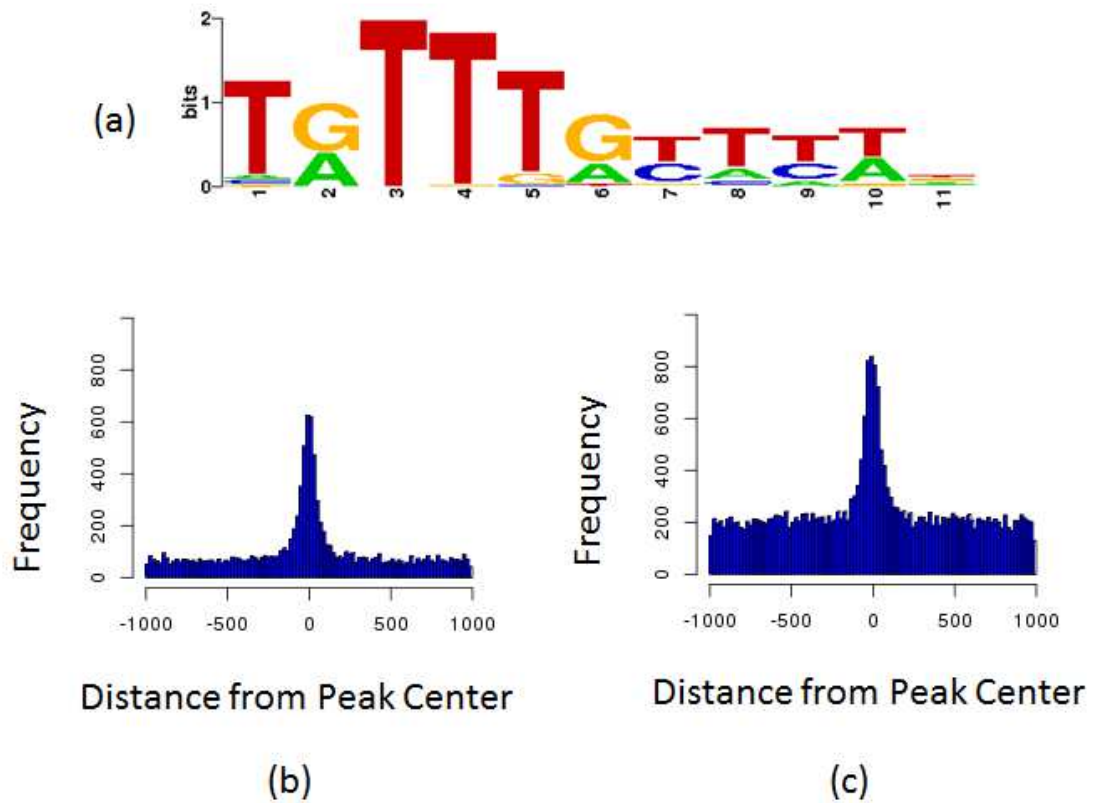
Other than *de novo* motif finding, another option to approach this problem is to predict enrichment of co-TF using known motifs. This can be done in two ways: 1) to perform motif enrichment analysis such as CENTDIST separately on the two sets of ChIP-seq peak list and then compare the resulting lists with certain cutoff in place and 2) to perform motif scan for each known motif and then to compare the enrichment scores among the two sets obtained using the counts in the sets against appropriate background. We will discuss the drawbacks of the various comparison measures using the simulated example in Section 3.3.

### 3.2 Additive nature of motif background

As we have discussed before, the background noise level of a motif is highly dependent on the GC content in the region. To illustrate this, we look at the distribution of FOXA1 motifs around AR ChIP-seq peaks in LNCaP. We first obtain two sets of AR ChIP-seq peaks, one set with high GC content (above 75<sup>th</sup> percentile) within 1000bp from peak center and the other with low GC content (below 25<sup>th</sup> percentile). As FOXA1 motif is AT rich (see Figure 3.2 (a)), it tends to occur less in GC-rich regions. We see the frequency of FOXA1 motif at the flanking region of the low GC set of AR peaks is much higher than that of the high GC ones. As a result, we also notice the frequency close to peak center to be additionally enriched by the same amount (see Figure 3.2 (b) and (c)). This observation suggests the additive nature of background motif to the true FOXA1FOXA1 motifs corresponding to true ChIP-seq binding.

### 3.3 Difficulties in identifying differential motifs

We proceed to generate simulated datasets based on the additive model suggested in Section 3.2. The simulated datasets are designed to have varying amounts of SP1 motifs implanted on two sets of artificially generated ChIP-seq peak regions, one generated from promoter background and the other generated from genomic background using chromosome 1. SP1 (V\$SP1\_Q6) having the sequence GGGGGGCGGGGCC (motif logo shown in Figure 3.3) has a GC rich motif that has a high chance of being reported in a GC-rich background.



**Figure 3.2 Histograms of FOXA1 (V\$HNF3ALPHA\_Q6) motif around AR ChIP-seq peaks in LNCaP with different GC content plotted using the same scale shows the additive nature of background motifs.**

a) FOXA1 motif is AT-rich. b) Histogram about AR ChIP-seq peaks with high GC content within +/- 1000bp from peak center. c) Histogram about AR ChIP-seq peaks with low GC content within +/- 1000bp from peak center.



**Figure 3.3 SP1 motif logo is GC rich.**

We would like to investigate the effects of GC contents in the background on the enrichment scores of GC-rich motifs such as SP1. In the genome, there exist GC-rich regions such as the promoter regions and CpG islands, and the non-GC-rich regions which form the majority. Markov models are often used to model the inherent nucleotide dependencies observed in these regions that are not due to motif enrichments. In general higher order Markov models are better (Thijs et al. 2001), but practically it is limited by technical constraints in terms of computation time and space, and depending on the amount of training data available, high orders also suffer from overtraining. To generate backgrounds corresponding to these two types of regions, we constructed two 7<sup>th</sup> order Markov models, one from the entire of chromosome 1 (calling it genomic background) and the other from the 1000 bp upstream of all non-redundant REFSEQ genes (calling it promoter background), and from these we generate 10000 random promoter background sequences and 10000 random genomic background sequences of length 4000bp using GenRGenS program (Ponty et al. 2006).

Next, for each of the two different backgrounds, we implant two levels of SP1 motifs such that 25% and 50% respectively contains SP1 motifs, simulating corresponding amount of SP1 TF binding in the peak sets. The *rseq* function of ‘cosmo’ package in R (Bembom et al. 2007) is used to generate motifs instance from the PWM, and

subsequently implanted into the background sequences such that the distance from the peak center follows a normal distribution such that 99.9% of all motifs fall within 500bp from the peak center.

Upon doing so, there will now be four sets of 10000 sequences of length 4000bp: promoter background with 50% SP1 motifs (Promoter High=**PH**), promoter background with 25% SP1 motifs (Promoter Low=**PL**), genomic background with 50% SP1 motifs (Genomic High=**GH**) and genomic background with 25% SP1 motifs (Genomic Low=**GL**). Based on our simulation, we would expect **GH** and **PH** to have similar level of SP1 motifs, and that **GL** and **PL** to have similar level of SP1 motif. We also expect SP1 motifs to be more enriched in **GH** and **PH** when compare to **GL** and **PL**.

**Table 6 Table describing the four sets of simulated SP1 ChIP-seq sequences.**

	Genomic Background, Low GC	Promoter Background, High GC
Low Level of SP1 implanted (25%)	<b>GL</b>	<b>PL</b>
High Level of SP1 implanted (50%)	<b>GH</b>	<b>PH</b>

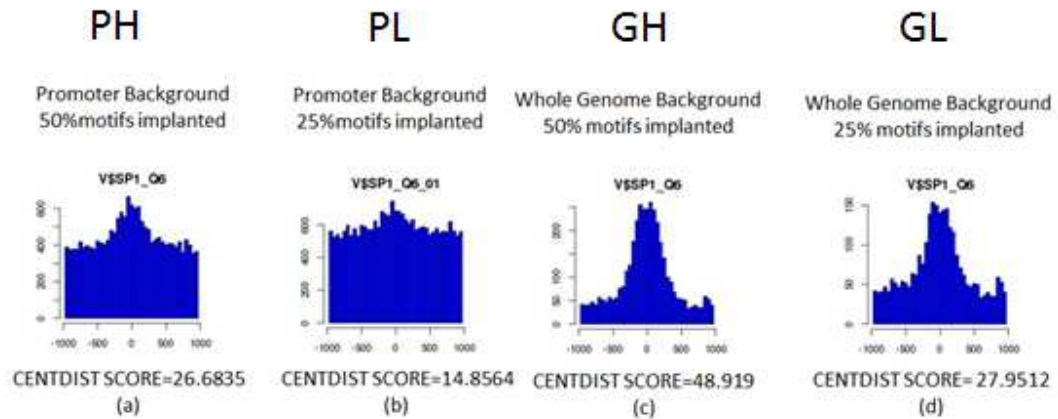
As mentioned in the introduction, there are several ways one may approach the problem of comparing motif enrichments in two sets: 1) Discriminative *de novo* motif finding, 2) Using motif enrichment tools such as CENTDIST and compare the list that are being

reported to be enriched (cut-off using p-value and/or score), 3) Perform motif scanning and derive a differential motif enrichment score based on the counts.

To study the effectiveness of *de novo* motif finders in this situation, we chose CMF out of the various motif finders due to it being one of the most recent discriminative motif finders developed and that it has features designed to account for different backgrounds. Out of all the comparisons, of particular interest is the comparison between **GH** and **PL**. Because we have implanted higher amount of SP1 motifs in **GH**, we would expect the *de novo* motif finder to report the SP1 motif as one of its candidates. However, on the contrary, CMF reported AT rich motifs such AAATAAATAAAA, ATATATA and AAAAATAAATA instead. This shows that under this circumstance, *de novo* motif finders fail to identify the correct motif.

Next we would like to discuss the drawbacks of comparing list of enriched motifs output by motif enrichment tools. We input the four sets **PH**, **PL**, **GH** and **GL** individually into CENTDIST. SP1 showed up as the top motif in all four sets, which is expected as 25% sequence with implanted motifs is definitely considered significant, though not as significant as compared to a set with 50% implanted motifs. This is not favourable for the purpose of comparison as we would not know which of these sets are more enriched. We may then ask what if we compare the scores reported by CENTDIST. The scores reported by CENTDIST are shown in Figure 3.4. Using the scores, we will falsely report that SP1 is more enriched in **GH** (having a score of 48.919) than in **PH** (having a score of 26.6835), similarly **GL** (having a score of 27.9512) as compared to **PL** (having a score of

14.8564), giving rise to false positive when SP1 is equally enriched in both of the comparisons resulting in false negative. This shows that the problem could not be solved by comparing the candidate list from motif enrichment tools.



**Figure 3.4 Enrichment score of CENTDIST fails to provide necessary information to determine differential enrichment among simulated datasets.**

CENTDIST score and optimal histogram for (a)**PH**: promoter background with 50% SP1 motifs implanted, (b)**PL**: promoter background with 25% SP1 motifs implanted, (c) **GH**: genomic background with 50% SP1 motifs implanted and (d) **GL**: genomic background with 25% SP1 motifs implanted.

To illustrate the problem further, we investigate various other enrichment scores based on the motif counts that are frequently used such as hypergeometric and binomial p-value against appropriate backgrounds, and fold overrepresentation, that similarly face the drawbacks experienced by CENTDIST's scoring function. The appropriate backgrounds to use are naturally the Markov model in which the sequences had been generated from.

Table 7 shows the counts of SP1 motifs in the four sets of sequences PL, GL, PH and GH. The counts of the respective promoter and genomic Markov-generated backgrounds

(same number of sequence and length) are also shown in the last two columns Pbg and Gbg respectively. While GH is consistently higher than GL, and PH is consistently higher than PL in all enrichment statistics which includes Binomial P-value using counts within 200bp, Hypergeometric Pvalue within 200bp and Fold Enrichment over background count. We see that there enrichment scores are significantly downplayed for the promoter backgrounds, leading to PH to be less enriched than GL when in fact PH has higher number of implanted motifs. The results using 100bp, 500bp and 1000bp window sizes (not shown) are similar. This shows the inherent problem when comparing motif enrichments among sets having different background.

**Table 7 The counts and respective enrichment P-value of SP1 in the four sets of sequences.**

	PL	GL	PH	GH	Pbg	Gbg
<b>motif count within 200 bp</b>	2340	621	2835	1072	1897	159
<b>numseq with motif within 200 bp</b>	2099	604	2517	1050	1718	158
<b>Binomial Pvalue using 200 bp</b>	5.23E-23	3.24E-169	9.11E-90	0		
<b>Hypergeometric Pvalue using 200 bp</b>	3.90E-12	9.33E-65	6.85E-44	1.85E-171		
<b>Fold count within 200bp over background</b>	1.233527	3.90566	1.494465	6.742138		

## 3.4 HISTSCORE: An Alternative Enrichment Statistic

### 3.4.1 Overview

The previous section showed that the various statistics that are commonly used for enrichment cannot give an accurate representation of the TF abundance in the simulated peak sets for comparison. In this section we attempt to develop an enrichment score that quantify the goodness of a given histogram based on our notion of center distribution.

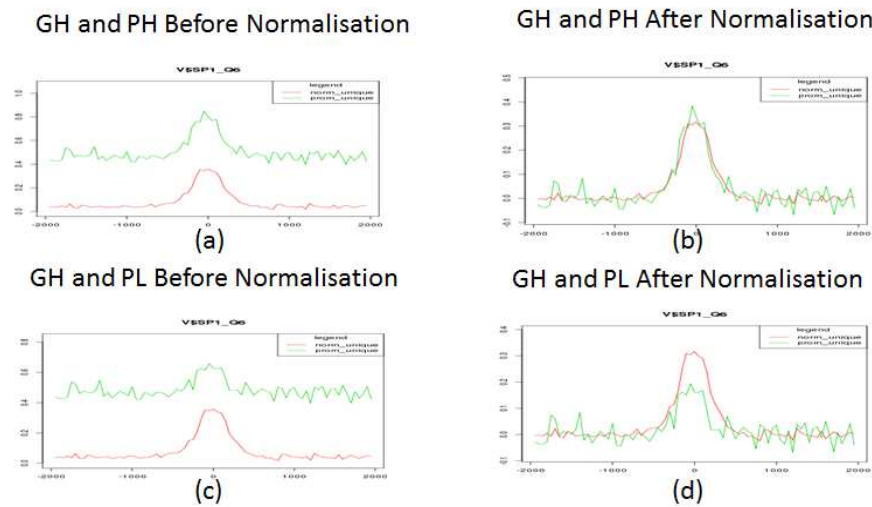
One statistic not often used is the normalised count which is the difference between the counts in the enrichment region and the background count. It is in fact closely related to the Z-score, with Z-score being computed by dividing the standard deviation of the background after obtaining the difference. Z-score is commonly used to provide estimation for Binomial P-value and this also explains why the enrichment score in **PH** and **PL** are greatly penalised (as compared to **GH** and **GL**) as the high motif counts in the promoter backgrounds are associated with much higher standard deviation. Because dividing the difference by the standard deviation makes the statistic too conservative, we chose to forgo the division.

To justify that this statistic is a better choice compared with p-value based statistic, in Table 8 we computed the normalised count within 200bp after subtracting the respective background counts. We see that using this, we managed to recover the actual relative TF abundance with **PH** and **GH** having normalised count of 938 and 913 compared with the normalised counts of 443 and 462 of **PL** and **GL** respectively, being roughly twice as enriched, which is as expected. Figure 3.5 shows the SP1 motif distribution around

respective ChIP-seq peaks before and after subtracting the background when comparing enrichment between **GH** and **PH** (Figure 3.5 (a)), and between **GH** and **PL** (Figure 3.5 (b)). Because of this favourable property, we therefore chose this statistic as the basis to improve upon for developing our enrichment score.

**Table 8 The counts and normalised count of SP1 in the four sets of sequences.**

	PL	GL	PH	GH	Pbg	Gbg
<b>motif count within 200 bp</b>	2340	621	2835	1072	1897	159
<b>normalised count within 200 bp after subtracting background</b>	443	462	938	913		

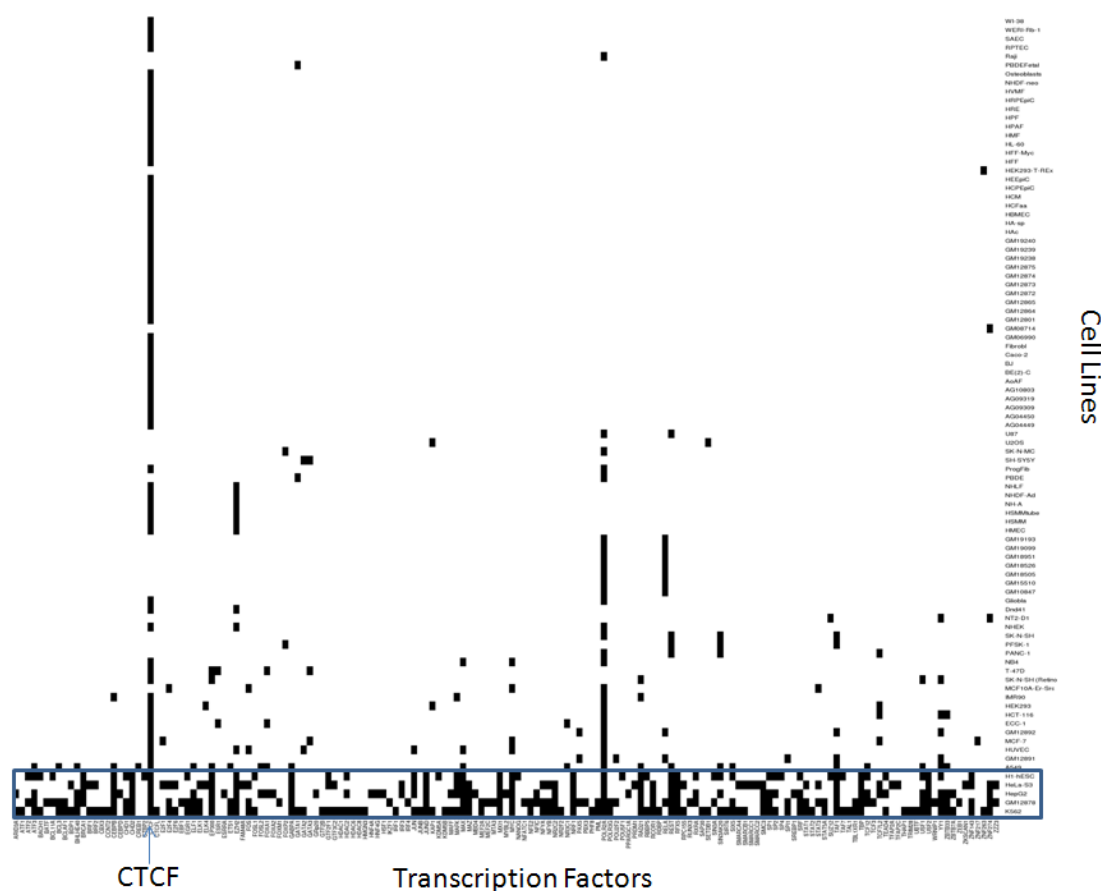


**Figure 3.5 Normalisation by subtracting the background enables the accurate comparison of enrichment.**

a) Motif density of SP1 in GH (red) and PH (green) before normalization b) Motif density overlay of SP1 in GH (red) and PH (green) after normalization c) Motif density overlay of SP1 in GH (red) and PL (green) before normalization d) Motif density overlay of SP1 in GH (red) and PL (green) after normalization

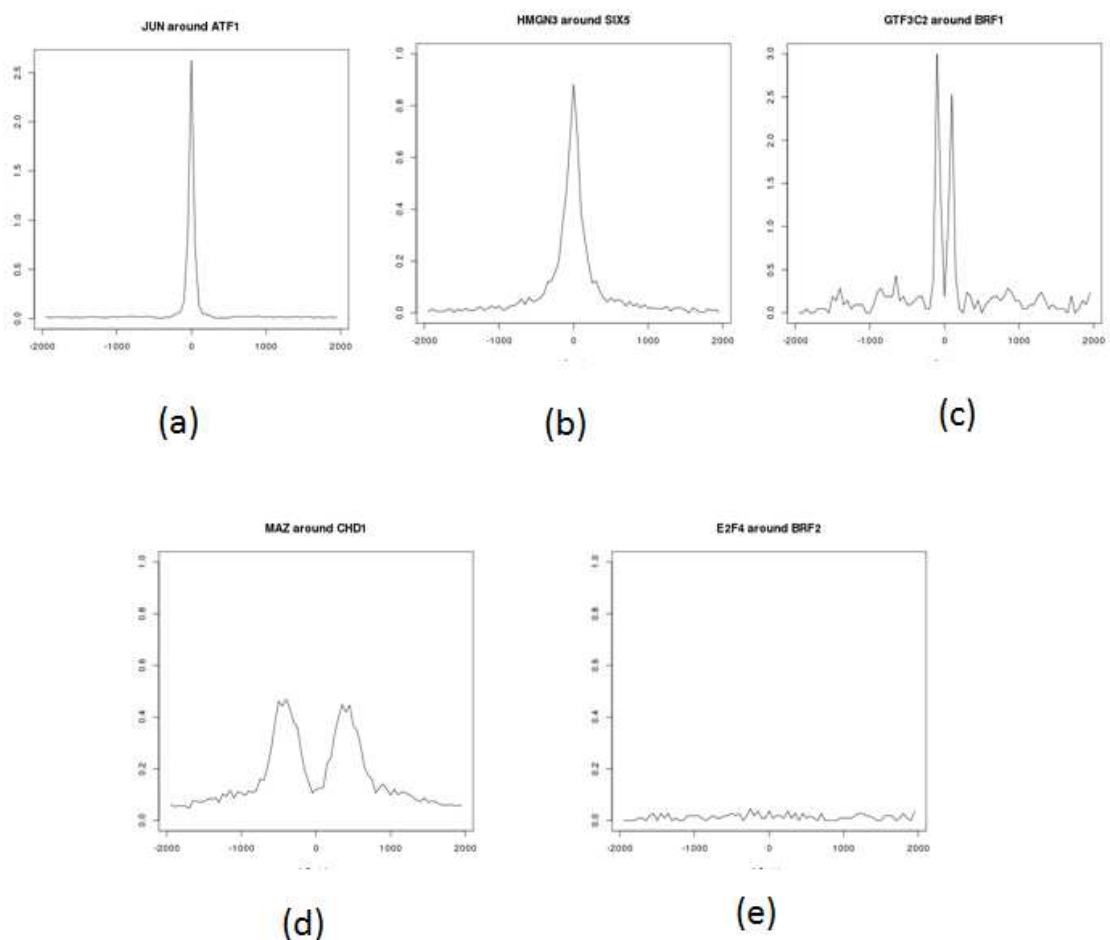
### 3.4.2 Insight from ENCODE datasets

Recently, the ENCODE project (Bernstein et al. 2012) released a large number of ChIP-seq datasets performed on various cell lines and TFs with pilot emphasis on the following five cell lines: K562, GM12878, Hela-S3, H1-Hesc and HepG2, and the TF CTCF (see Figure 3.6). Using the most comprehensive cell line K562, we wish to first look at the different types of ChIP-seq peak profiles we could observe relative to another set of ChIP-seq peaks. We generated the profiles for all ChIP-seq dataset pairs in K562. On the whole, the main types of profiles we observed are single peak profiles of various widths (Figure 3.7 (a) and (b)), double-peak profiles peaking at various distances (Figure 3.7 (c) and (d)) and profiles that are flat without enrichment (Figure 3.7 (e)). In particular, Figure 3.7 (c) shows MAZ around CHD1 having a double peak profile. Plotting MAZ motif (V\$MAZR\_01) in place of MAZ ChIP-seq also yield the same double peak profile (Figure 3.8). This observation also motivates the idea of predicting ChIP-seq distribution using motifs.



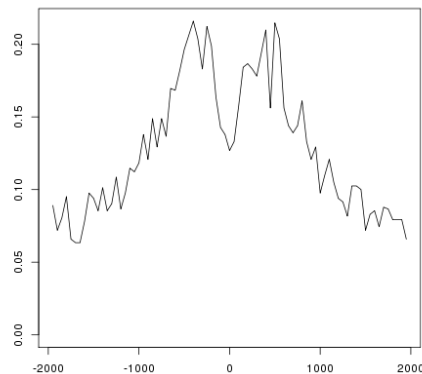
**Figure 3.6 ENCODE TF ChIP-seq Experimental Matrix across 91 Cell Lines and 161 TF.**

ENCODE ChIP-seq data matrix with pilot emphasis on CTCF and five cell lines, namely K562, GM12878, Hela-S3, H1-Hesc and HepG2 (boxed).



**Figure 3.7 Different peak distribution profiles observed for pairs of TFs ChIP-seq performed in K562. The graphs have been normalised by dividing by the number of peaks in the ChIP-seq data to be centered upon.**

(a) JUN around ATF1 showing narrow sharp peak profile. (b) HMGN3 around SIX5 showing wide peak profile. (c) GTF3C2 around BRF1 showing double peak profile with a dip in the center. (d) MAZ around CHD1 showing another double peak profile peaking at a further distance (500bp) from the center. (e) E2F4 around BRF2 showing no enrichment.



**Figure 3.8 Motif distribution profile of MAZ motif (V\$MAZR\_01) around CHD1 ChIP-seq peaks in K562.**

Double peak profile of MAZ motif corresponding to the double peak profile seen in Figure 3.7 (d)

Using the profiles which have been normalized by dividing the number of peaks in the ChIP-seq data to be centered upon, we can essentially infer the percentage of the ChIP-seq peaks containing the secondary ChIP-seq peak (or motif) at a certain distance from the center. This in turns enables us to infer the strength of the interaction among the primary TF (the one being centered) and the secondary TF (being counted). Throughout this chapter, all such distribution graph seen, be it for motif or ChIP-seq peaks are binned at 50bp and scaled by dividing by the number of ChIP-seq peaks centered upon, and then multiplying by 10. A height of 1 is considered very good. The maximum of the Y-axis is being set to 1 or the maximum height of the graph if greater than 1.

Judging by the shapes of distributions we observed, especially those with distal enrichment having double peak profile such as in Figure 3.7 (d), the strength of interaction cannot be naively considered to be based on the enrichment at the center.

Hence, it is necessary to develop a scoring based on the distribution graph that enables double peak profile to be scored appropriately. We shall call this scoring HISTSCORE (short for histogram score).

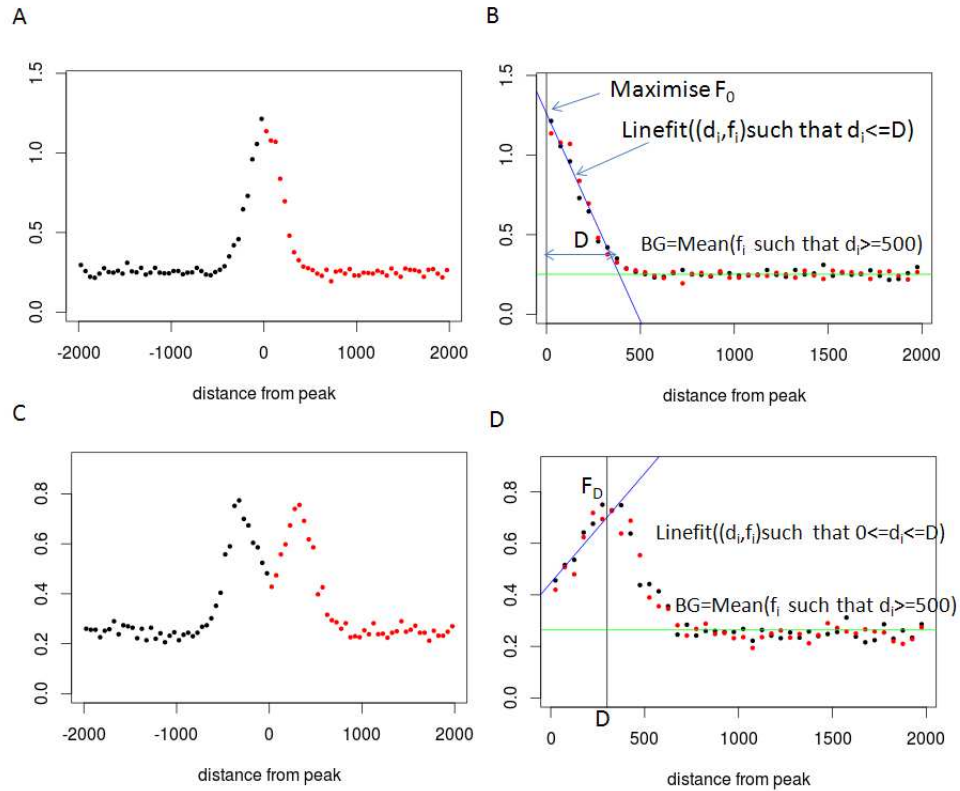
We shall now describe how HISTSCORE is being calculated based on the histogram counts (of motifs or peaks binned at 50bp over +/-2000 bp) around our set of ChIP-seq peaks of interest. Co-TF peak or motif distributions around ChIP-seq peaks can have single peak profile like Figure 3.9(a) or double peak profile like Figure 3.9(d). These two types of profiles will be scored separately and the better of these two scores will be used as the final HISTSCORE for that profile.

To compute the score for the single peak profile, we first fold the graph representing the histogram counts in Figure 3.9(a) such that the points on the left of the y-axis are reflected along the y-axis as in Figure 3.9(b). Fixing a certain distance  $D$ , we can fit a regression line over the points not more than  $D$  from the center. For that  $D$  we take note of the value  $F_0$  where the line cuts the y-axis. This is repeated for various  $D$  up to 500bp. The maximum such  $F_0$  obtained is then subtracted by the background mean comprising of points greater than 1000bp from the center to obtain the single peak profile score.

Likewise, to determine the double peak profile score, the graph representing the histogram counts which typically look like Figure 3.9(c) is folded such that the points on the left of the y-axis are reflected along the y-axis as in Figure 3.9(d). Fixing a certain distance  $D$ , we can fit a regression line over the points not more than  $D$  from the center.

But this time, for that  $D$  we take note of the value  $F_D$  where the line cuts the line  $x=D$  instead. This is repeated for various  $D$  up to 500bp. The maximum such  $F_D$  obtained is then subtracted by the background mean comprising of points greater than 1000bp from the center to obtain the double peak profile score.

The rationale behind HISTSCORE is to tap on the notion of measuring the goodness of the histogram taking the whole histogram into account rather than focusing on counts within a specific window. This HISTSCORE focuses on the highest point of the graph which could be at the center or some distance  $D$  from the center. The reason for not taking just the maximum is because the method has been intentionally designed to be robust and penalize against irregularly shaped graphs that may spike at certain points for some reasons.



**Figure 3.9 The calculation of HISTSCORE from histogram counts. The histogram counts have been scaled by dividing by the number of peaks and multiplying by 10.**

(a) Typical histogram count graph with single peak profile. Each dot represent the count of motif at 50bp bin at the respective distance from the peak denoted by the x-axis.

(b) Folded version of (a) with black dots on the left reflected to the right along the y-axis. Blue line shows the best fitted line which maximises the y-intercept  $F_0$ . Green line is the mean of points greater than 1000 bp from center.

(c) Typical histogram count graph with double peak profile. Each dot represent the count of motif at 50bp bin at the respective distance from the peak denoted by the x-axis.

(d) Folded version of (c) with black dots on the left reflected to the right along the y-axis. Blue line shows the best fitted line which maximises the y-intercept  $F_D$ . Green line is the mean of points greater than 1000 bp from center.

### 3.5 MOTIFDIFF Algorithm

Having defined our appropriate enrichment statistic, HISTSCORE, we can now proceed to describe our algorithm to determine differentially enriched motifs in two contrasting sets for a set of known motifs. The essential input of MOTIFDIFF are two sets of genomic locations representing ChIP-seq peaks (chromosome-peak summit position) denoted by  $S_A$  and  $S_B$  that are to be compared to find the list of differential motifs, and a list of candidate PWM motifs (provided by users or obtained from either the TRANSFAC (Matys et al., 2003) or JASPAR (Sandelin et al., 2004) databases).

Given  $S_A$  and  $S_B$ , we obtain the set  $S_{A\text{unique}}$  by filtering those peaks in  $S_A$  that are not within 500bp from any peaks in  $S_B$  and likewise,  $S_{B\text{unique}}$  by filtering those peaks in  $S_B$  that are not within 500bp from any peaks in  $S_A$ . Next, we extract  $\pm 2000$ bp from each of the two sets  $S_{A\text{unique}}$  and  $S_{B\text{unique}}$  from the reference genome and proceed to scan the sequence using each PWM in the input motif database. The E-value cutoff for the motif scan is set at 0.00025 in the flanking background where we limit the number of motif hits to 2.5 motif hits per 10000 bp in the flanking background region. The rationale for setting Evalue cutoff of 0.00025 is that it gives sufficient hits for analysis of peak size larger than 1000 while enabling our motif scanning algorithm to take advantage of the relative stringent cutoff to speed up the search. Moreover, it is roughly the probability of a perfect 6-mer match, which is a typical baseline for a proper motif.

Using the motif hits of motif  $M$ , we can then generate the histogram graphs (50 bp bins dividing the counts by the number of peaks and multiplying by 10),  $H_{A,M}$  and  $H_{B,M}$  of

motif  $M$  in  $S_{A\text{unique}}$  and  $S_{B\text{unique}}$  respectively and compute MDscore which is a simple ratio test of the HISTSCORE of  $H_{A,M}$  over  $H_{B,M}$ , with a small pseudocount added to numerator and denominator to avoid division by zero, i.e.

$$MDscore_M(A, B) = \frac{HISTSCORE(H_{A,M}) + \rho}{HISTSCORE(H_{B,M}) + \rho}$$

where  $\rho$  is some small pseudocount (chosen to be 0.0025).

We utilize a Z-score-like measure to help us filter noisy results. We define:

$$MDzscore_M(A, B) = \frac{HISTSCORE(H_{A,M}) - HISTSCORE(H_{B,M})}{BGdev(H_{B,M})}$$

where  $BGdev(S_{B\text{unique}})$  is the standard deviation of the flanking background of  $S_{B\text{unique}}$  made up of bins more than 500bp away from the center.

We reject the result if the MDzscore is below 3.719 corresponding to a Z-test pvalue of 0.0001. This is so as to reduce the false positives due to noisy background. In addition to MDzscore, we arbitrarily set our cutoff for MDscore to 1.3.

## 3.6 Results

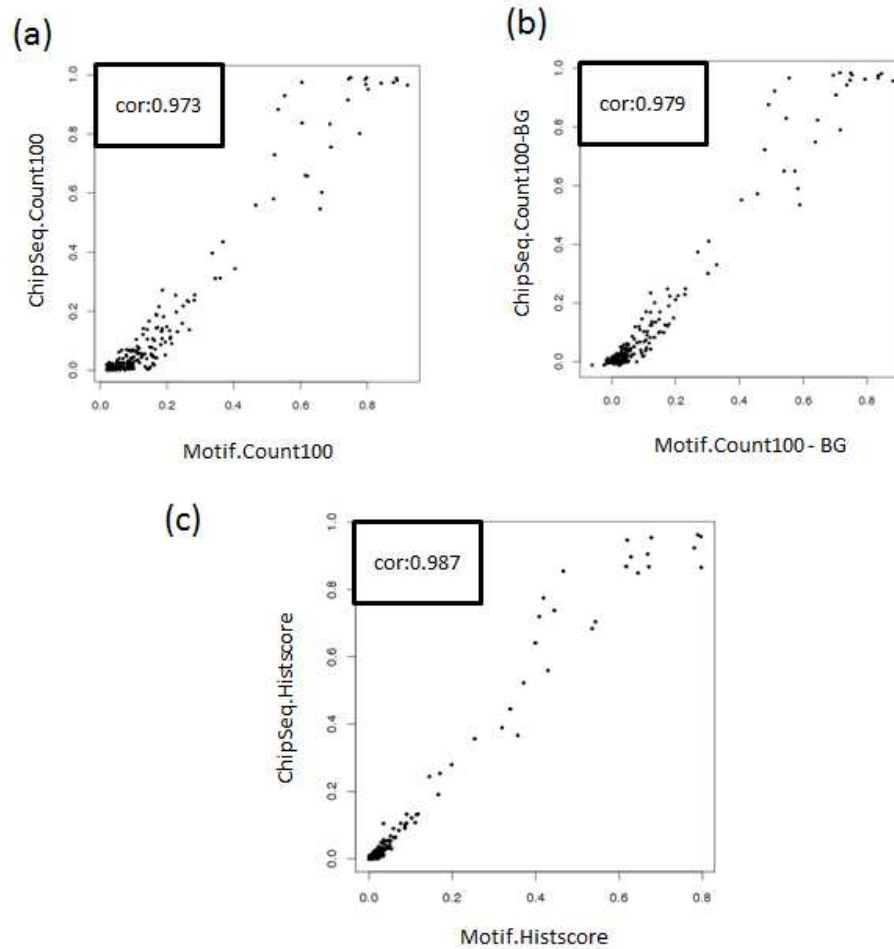
### 3.6.1 Motif HISTSCORE have good correlation with ChIP-seq HISTSCORE

We make use of the large number of CTCF ChIP-seq data in ENCODE to help evaluate HISTSCORE as a good measure of ChIP-seq localization signal by showing that the HISTSCORE using CTCF motif correlates well with the HISTSCORE computing using CTCF ChIP-seq compared with other measures.

Using the binding peaks of all other TFs in ENCODE, we can obtain sets of regions of various backgrounds by performing clustering of all TF ChIP-seq peaks (other than CTCF) in a particular cell line. For example, the cell line NB4 has 4 TF ChIP-seq peak sets, namely: CTCF, POLR2A, MAX and MYC. Excluding CTCF, using the peaks of POLR2A, MAX and MYC, we can cluster sets of peaks together if they are within 500bp from some peak in the cluster. After clustering, we will have 7 cluster groups containing: POLR2A only (cluster<sub>100</sub>), MAX only (cluster<sub>010</sub>), MYC only (cluster<sub>001</sub>), POLR2A and MAX only (cluster<sub>110</sub>), POLR2A and MYC only (cluster<sub>101</sub>), MYC and MAX only (cluster<sub>011</sub>) and all POLR2A, MYC and MAX. (cluster<sub>111</sub>)

We perform clustering for all cell lines and then select only those sets of clusters which have at least 1000 clusters in the set. The mean coordinate of the cluster will then be treated as a ChIP-seq peak. For each of these clusters, we plot the histogram using CTCF motif to obtain  $H_{\text{motif}}$  and using the CTCF ChIP-seq peaks in the cell line to obtain  $H_{\text{chipseq}}$ .

We then try to correlate the three different scores COUNT100, COUNT100-BG and HISTSCORE applied onto  $H_{\text{motif}}$  and  $H_{\text{chipseq}}$ . COUNT100 takes the sum of count within 100bp from center. COUNT100-BG takes the sum of count within 100bp from center and subtract the expected background using bins 1000bp away from the center. The scatter plots for the three scores COUNT100, COUNT100-BG and HISTSCORE are shown in Figure 3.10. We also perform with other window sizes but do not show them as COUNT100 has the best correlation (0.973, see Figure 3.10(a)) among them. COUNT100-BG has slightly better correlation (0.979, see Figure 3.10(b)) but HISTSCORE has the highest correlation (0.987, see Figure 3.10(c)).



**Figure 3.10 HISTSCORE of motif correlates better with actual ChIP-seq peak distribution signal than other score.**

Scatterplot of (a) count of CTCF chipseq peaks within 100bp of cluster center against count of CTCF motif within 100bp of cluster center. (b) count of CTCF chipseq peaks within 100bp of cluster center with flanking background subtracted against count of CTCF motif within 100bp of cluster center with flanking background subtracted. (c) HISTSCORE( $H_{\text{ChIPseq}}$ ) against HISTSCORE( $H_{\text{motif}}$ ).

### 3.6.2 Large scale validation of MOTIFDIFF using ENCODE

One of the applications of MOTIFDIFF is to identify the differential co-TF partners of a particular TF in two different cell lines. For example say we have two cell lines *C1* and *C2* and we have performed ChIP-seq experiment on a particular factor *TF1* for both cell lines, calling the ChIP-seq peak sets corresponding to the above two cell lines *C1TF1* and

*C2TF1*. MOTIFDIFF will try to predict a factor *TF2* which is specifically enriched in one of the two cell lines. Suppose we have the ChIP-seq for *TF2* in both cell lines as well, say *C1TF2* and *C2TF2*, then we can try to define enrichment using these ChIP-seq libraries as follows. The peak sets *C1TF1* and *C2TF1* may not be disjoint. The set of peaks exclusive to *C1TF1* is named *C1TF1unique* and the set of peaks exclusive to *C2TF1* is named *C2TF1unique*. Using the actual ChIP-seq data, we define the gold standard to be:

$$CHIPSCORE_{TF1,C1,C2}(TF2) = \frac{HISTSCORE(H_{C1TF1unique,C1TF2}) + \varepsilon}{HISTSCORE(H_{C2TF1unique,C1TF2}) + \varepsilon}$$

where  $\varepsilon$  is set to be 0.0025.  $H_{A,M}$  is the histogram of *M* around *A* as defined in Section 3.5 (also see ChIP-seq HISTSCORE in Section 3.6.1). Note the asymmetric definition of the function and therefore  $CHIPSCORE_{TF1,C1,C2}(TF2)$  is not the reciprocal of  $CHIPSCORE_{TF1,C2,C1}(TF2)$ . When computing the enrichment of *C1* over *C2*, we make use of the *TF2* ChIP-seq datasets for *C1* and ignore the *TF2* ChIP-seq for *C2*.

Since we require the ChIP-seq of the predicted TF in both cell lines to be able to validate the predictions reported by MOTIFDIFF, we need to use pairs of cell lines with high degree of overlapping ChIP-Seq experiments. As seen in Figure 3.6, the concentration of experimental data in the five cell lines namely: GM12878, Hela-S3, H1-hESC, K562 and HepG2, make the cross product of these cell lines a great choice.

As with CENTDIST, due to the similarities among motifs of the same family such as the various FOX motifs, we are satisfied when the predicted motif family is correct. For those TF families which have multiple encode data, the encode fold score used is the maximum score within the TF family. Motif family assignment of TRANSFAC is as per Zhang et al. (2011). Each TF in ENCODE datasets is assigned to a TRANSFAC family if its motif can be found in TRANSFAC. We get the best score in the FAMILY as the representative score for the FAMILY, i.e.

$$CHIPSCORE_{TF1,C1,C2}(FAMILY) = \max_{TF2 \in FAMILY} CHIPSCORE_{TF1,C1,C2}(TF2)$$

From the output of MOTIFDIFF, we also have

$$MDSCORE_{TF1,C1,C2}(FAMILY) = \max_{TF2 \in FAMILY} MDSCORE_{TF1,C1,C2}(TF2)$$

For this calculation, we only consider those TF2 that have passed our MDzscore cutoff of 3.719.

We deem a FAMILY to be correct by gold standard in the direction of C1 over C2 if

$$CHIPSCORE_{TF1,C1,C2}(FAMILY) \geq 2$$

We deem a FAMILY to be positively predicted by MOTIFDIFF in the direction of C1 over C2 if

$$MDSCORE_{TF1,C1,C2}(FAMILY) \geq 1.3$$

### 3.6.2.1 Prediction of differential CTCF binding in all common TF peak sets in selected cell line pairs

As CTCF has been performed on all cells studied in ENCODE project has a specific conserved motif, we first decide to look at the prediction accuracy of CTCF (considering TF2=CTCF) for all comparisons of C1TF1 versus C2TF1 (TF1 not CTCF). Table 9 shows the prediction's precision and recall is 84.3% and 88.2% respectively, meaning that the enrichment determined by MOTIFDIFF is correct 84.3% of the time, while it is able to accurately identify 88.2% of the total true enrichments as defined by our gold standard.

**Table 9 Performance of MOTIFDIFF in predicting CTCF using CHIPSCORE as gold standard**

		Pass MOTIFDIFF	
		FALSE	TRUE
CHIPSCORE>2	FALSE	336	20
	TRUE	28	150

recall=84.3%

precision=88.2%

### 3.6.2.2 Prediction of differential binding of other TF in all common TF peak sets in selected cell line pairs

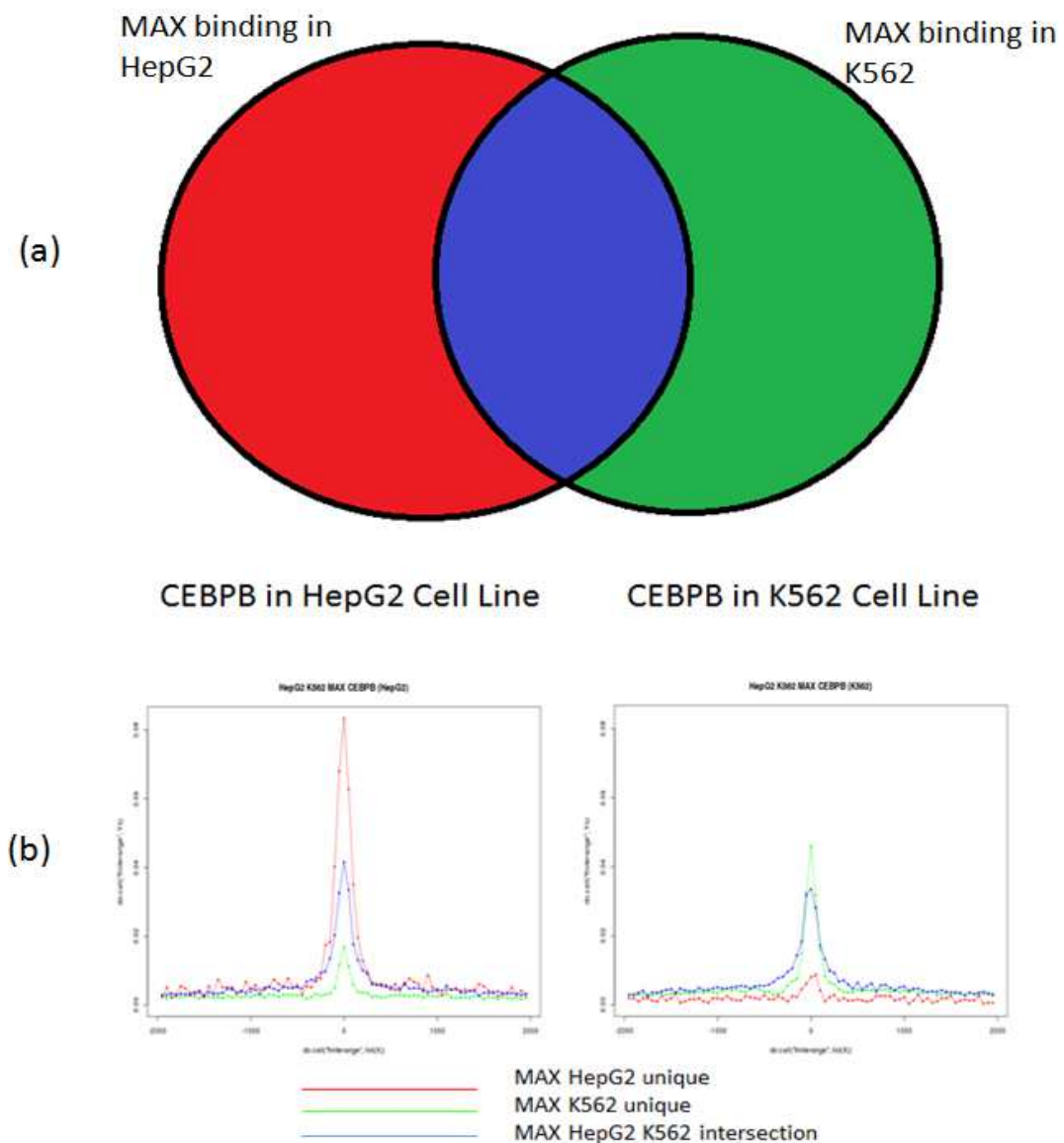
Next, we proceed to determine the prediction accuracy of all TFs that had been assigned a TF family. Table 10 shows the prediction performance of MOTIFDIFF. The result shows the overall prediction's precision and recall are only 76.8% and 51.1% respectively.

Though precision of 76.8% is still roughly acceptable, but we observe that the recall of 51.1% is rather low.

**Table 10 Performance of MOTIFDIFF in predicting all TF by FAMILY using CHIPSCORE as gold standard**

		Pass MOTIFDIFF		
		FALSE	TRUE	
CHIPSCORE>2	FALSE	3150	945	recall=51.1% precision=76.8%
	TRUE	2991	3122	

To investigate why, we look at some example of situation where CHIPSCORE is high but MDSCORE is low. Figure 3.11 shows the histogram of TF CEBPB in respective regions of the overlap between the MAX sites of HepG2 and K562. CEBPB ChIP-seq performed in HepG2 is enriched preferentially in the HepG2 unique regions while CEBPB ChIP-seq performed in K562 is enriched preferentially in the K562 unique regions. This is caused by the inherent problem that existence of motif does not perfectly determine the binding of the corresponding TF at a particular region. Other factors such as epigenetics, chromatin conformation, histone modifications etc play important roles especially when we are looking at completely different cell types.



**Figure 3.11 CEBPB is preferentially enriched in different regions depending on the Cell Line in which the ChIP is performed.**

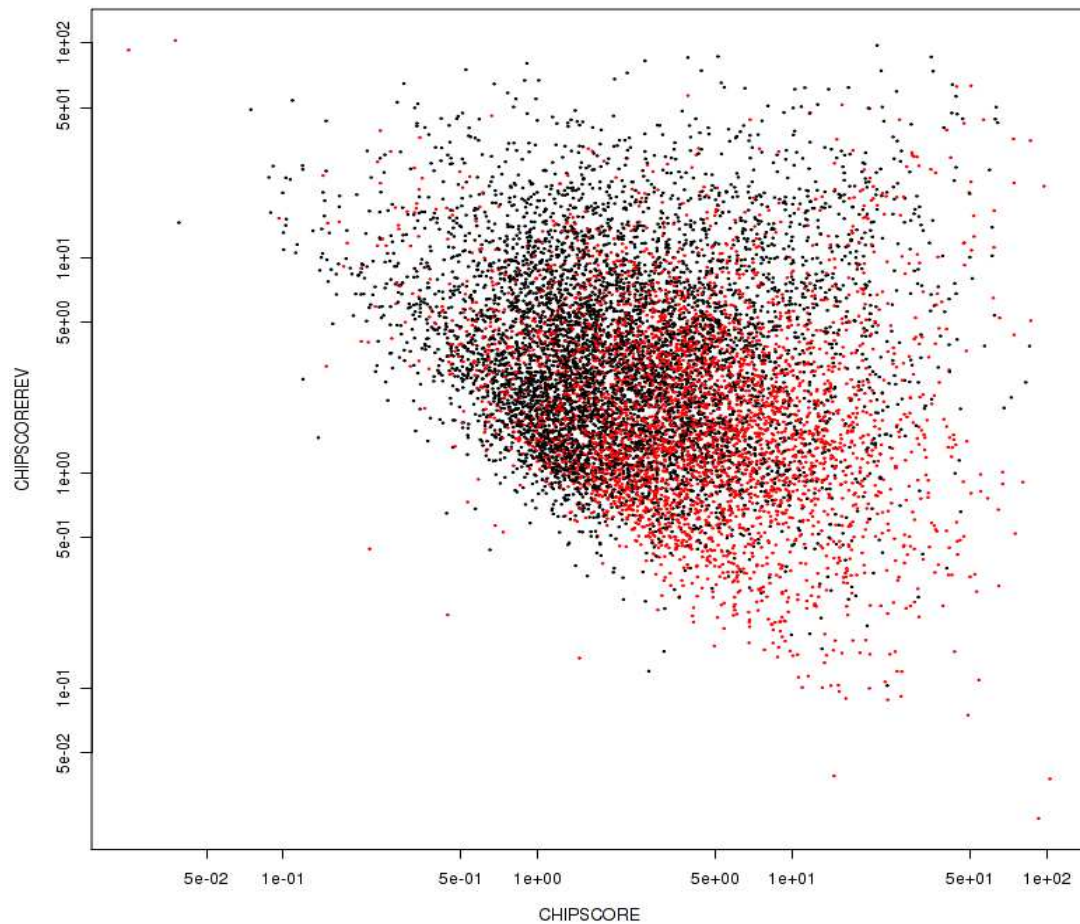
(a) Venn diagram color-coded to represent the partition of MAX binding used for the plot in (b).  
 (b) The left plot shows the distribution of CEBPB (performed in HepG2 cell line) around respective groups of MAX TF binding while the right plot shows the distribution of CEBPB (performed in K562 cell line) around MAX TF binding in K562 cell line. With reference to CHIPSCORE, the C1, C2, TF1, TF2 for the left plot are HepG2, K562, MAX, CEBPB respectively and CHIPSCORE is the enrichment of red graph over the green graph. For the right plot, C1, C2, TF1, TF2 are K562, HepG2, MAX, CEBPB respectively, and the CHIPSCORE is the enrichment of green graph over the red graph. The CHIPSCORE of the right plot is CHIPSCOREREV corresponding to parameters for the left plot.

We define the reverse CHIPSCORE, CHIPSCOREREV as:

$$CHIPSCOREREV_{TF1,C1,C2}(FAMILY) = CHIPSCORE_{TF1,C2,C1}(FAMILY)$$

which is the corresponding highest family score in the reverse direction of the comparison. We therefore plot the scatter plot of *CHIPSCORE* against *CHIPSCOREREV* over all TF1, C1 and C2 and mark the point as red if the enrichment is being reported

by MOTIFDIFF. (See Figure 3.12) From the plot we see that in the lower right quadrant which is where *CHIPSCORE* is high and *CHIPSCOREREV* is low, we have a high proportion of accurate prediction by MOTIFDIFF and the accuracy decreases as *CHIPSCOREREV* increases. Another thing that can be seen on the diagram is that when *CHIPSCORE* is low, MOTIFDIFF seldom predict the enrichment to be positive. Table 11 represents the scatterplot in the form of a table containing the percentage positive prediction in the respective section on the graph. From this table, we see that the TF's *CHIPSCOREREV* in the two sets affects whether MOTIFDIFF can accurately predict its enrichment.



**Figure 3.12 Scatterplot of *CHIPSCOREREV* against *CHIPSCORE* to show the accuracy in relation to the bidirectionality of gold standard enrichment.**

Scatter plot of *CHIPSCORE* against *CHIPSCOREREV* for all the comparisons. Points are marked red if the enrichment is being reported by MOTIFDIFF. Observe that higher proportion of points is marked red at the lower right region of the plot.

**Table 11 Accuracy of MOTIFDIFF with respect to the bidirectionality of gold standard enrichment. The number of points are shown in bracket.**

Prediction reported by MOTIFDIFF when CHIPSCORE>2 is true positives and when CHIPSCORE<2 is false positives. We observe that the true positive is highest when CHIPSCOREREV is low.

		CHIPSCORE		
		(-Inf,1] (FP)	(1,2] (FP)	(2, Inf] (TP)
CHIPSCOREREV	(-Inf,1]	0.269 ( 26 )	0.335 ( 248 )	0.731 ( 1325 )
	(1,2]	0.222 ( 248 )	0.309 ( 754 )	0.579 ( 1494 )
	(2, Inf]	0.171 ( 1325 )	0.228 ( 1494 )	0.391 ( 3294 )

### 3.6.3 Application

Having evaluated the performance using large scale datasets from ENCODE, we proceed to apply our method on some existing comparison study performed by other researchers and also on our in-house generated data.

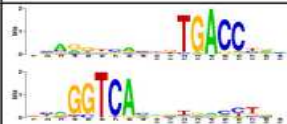
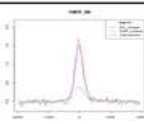
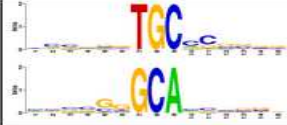
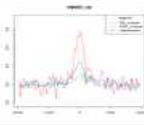

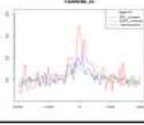
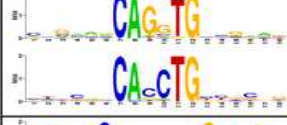
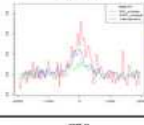
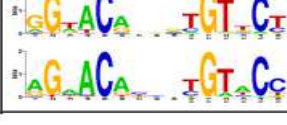
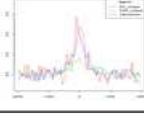
#### 3.6.3.1 Cofactors in different Signaling Pathway: MCF7 ERE2 vs EREGF

Lupien et al. (2009) showed that ER-alpha binds at different locations corresponding to E2 signaling and EGF-signaling in MCF7 cell lines. In the paper, it was found that FOXA1 and AP-1 motifs were enriched in the shared as well as EGF-unique binding sites.

We performed MOTIFDIFF analysis on the ER-alpha binding sites under E2 and EGF treatment. We observed FOXA1 (rank 3) and AP-1 (rank 1) are enriched preferentially in ER-alpha binding in EGF-treated cells (see Figure 3.14). It agrees with the finding of Lupien et al. In addition, we also observed GATA motif (rank 2) to be enriched in the binding sites unique to EGF. This probably suggests that some member of GATA Family may likely interact with ER in the EGF signaling pathway. GATA3 has previously been

shown to be required for the luminal A type of breast cancer. Dydensborg et al. (2009) showed that GATA3 interact with LMO1, ER and FOXA1 whereas Albergaria et al. (2009) showed that the nuclear expression of GATA3 in breast cancer is considered a marker of luminal cancer in ER+ cancer and luminal androgen responsive cancer in ER-/AR+ tumors. It is highly coexpressed with FOXA1 and serves as negative predictor of basal subtype and ERBB2 subtype. Hence, it is likely that GATA3 may mediate gene regulation in some specific way unique to EGF signaling.

## ER E2 unique over ER EGF unique


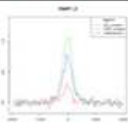

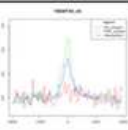
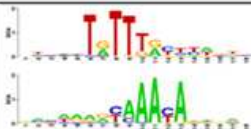
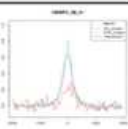

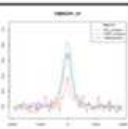

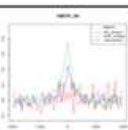
RANK	LOGO	IMAGE	FAMILY	MOTIF	SCORE
1			<a href="#">ERE</a>	V\$ER_Q6	SCORE=2.92 ZSCORE=25.524
2			<a href="#">HIC1</a>	V\$HIC1_02	SCORE=2.41 ZSCORE=8.068
3			<a href="#">CACCT</a>	V\$AREB6_03	SCORE=2.09 ZSCORE=6.531
4			<a href="#">EBOX</a>	V\$MYOD_Q6_01	SCORE=2.05 ZSCORE=5.829
5			<a href="#">AR</a>	V\$AR_01	SCORE=1.98 ZSCORE=7.906

Total reported with SCORE>=1.3 and ZSCORE>=3.719 : 31

**Figure 3.13** Differentially enriched motifs found in ERE2unique over EREGFunique

The top 5 reported TF families are shown. Total number of families that are above cutoff is 31.

## ER EGF unique over ER E2 unique

RANK	LOGO	IMAGE	FAMILY	MOTIF	SCORE
1			<a href="#">AP1</a>	V\$AP1_C	SCORE=2.51 ZSCORE=29.027
2			<a href="#">GATA</a>	V\$GATA3_02	SCORE=2.25 ZSCORE=15.592
3			<a href="#">FOX</a>	V\$HNF3_Q6_01	SCORE=2.14 ZSCORE=24.631
4			<a href="#">BACH</a>	V\$BACH1_01	SCORE=1.79 ZSCORE=22.08
5			<a href="#">GATA_DIMER</a>	V\$EV11_05	SCORE=1.76 ZSCORE=9.56

Total reported with SCORE>=1.3 and ZSCORE>=3.719 : 25

**Figure 3.14 Differentially enriched motifs found in EREGFunique over ERE2unique**

The top 5 reported TF families are shown. Total number of families that are above cutoff is 25.


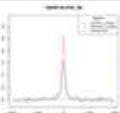

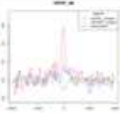

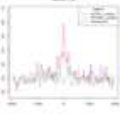

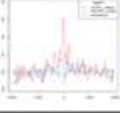

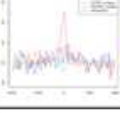

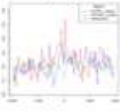
### 3.6.3.2 Knockdown study: LNCaP SiFoxA1 vs SiCtrl

Wang et al. (2011) showed that the presence of FOXA1 in LNCaP cells determines the mode of AR binding. In an attempt to figure out what other potential TFs are involved in such a change, we performed MOTIFDIFF analysis on two ChIP-seq datasets from the paper, i.e. the AR ChIP-seq peak sets before (siCTRL) and after knockdown of FOXA1

(SiFoxA1). Figure 3.15 shows the MOTIFDIFF comparison results reporting the motif families that are enriched in normal over FOXA1 knockdown AR peaks. The result shows the forkhead (FOX) family to be highly differentially enriched, which is not surprising as in the presence of FOXA1, AR binding is highly dependent on FOXA1 as FOXA1 acts as a pioneering factor (Qiao et al. 2011). In the absence of the factor, AR no longer depends on FOXA1 for binding and hence do not occur preferentially near FOXA1 motifs.

As for the enrichment of FOXA1 knockdown AR peaks over normal (see Figure 3.16), in agreement with the paper's result, we found NF1 (MDscore=1.92, Zscore=19.385 or Pvalue=5e-84) and AR (MDscore=1.8, Zscore=101.51 or Pvalue<1e-300) motifs to be preferentially enriched. Interestingly, however NF1 and AR though highly enriched are not the highest ranked factors, at 14 and 18 respectively. We feel that those ranked at the top are potentially important factors too. Using ONCOMINE (Choi et al. 2011), we found Myc belonging to the EBOX family which is reported as the 2<sup>nd</sup> most enriched family, to be up-expressed in five Prostate cancer vs normal studies (Varambally, LaTulippe, Grasso, Lapointe and Tomlins), suggesting its importance in the development of Prostate cancer. Pax3 belonging to the PAX family which is reported as the 3<sup>rd</sup> most enriched family by MOTIFDIFF is found to be up-regulated upon DHT treatment in LNCaP using our in-house generated microarray data, suggesting a feedback mechanism involving Pax3 in the absence of FOXA1.

## SiCTRL unique over SiFoxA1 unique

RANK	LOGO	IMAGE	FAMILY	MOTIF	SCORE
1			<a href="#">FOX</a>	V\$HNF3ALPHA_Q6	SCORE=8.22 ZSCORE=45.467
2			<a href="#">TEF</a>	V\$TEF_Q6	SCORE=3.09 ZSCORE=9.578
3			<a href="#">SOX</a>	V\$SOX17_01	SCORE=2.91 ZSCORE=8.807
4			<a href="#">NKX</a>	V\$NKX62_Q2	SCORE=2.9 ZSCORE=9.087
5			<a href="#">POLYA</a>	V\$LPOLYA_B	SCORE=2.35 ZSCORE=7.434
9			<a href="#">CDX</a>	V\$CDX2_Q5	SCORE=2.14 ZSCORE=5.531

Total reported with SCORE $\geq$ 1.3 and ZSCORE $\geq$ 3.719 : 29

**Figure 3.15 Differentially enriched motifs found in siCTRL unique over SiFoxA1 unique.**

The top 5 reported TF families are shown, together with CDX which has been reported by the author of the study. Total number of families that are above cutoff is 29.

## SiFoxA1 unique over SiCtrl unique

1			AHR	VSAHRHIF_Q6	SCORE=2.5 ZSCORE=16.78
2			EBOX	VSHIF1_Q5	SCORE=2.49 ZSCORE=19.217
3			PAX	VSPAX6_Q2	SCORE=2.32 ZSCORE=17.63
4			HEN	VSLBP1_Q6	SCORE=2.31 ZSCORE=10.924
5			CREB	VSNRSF_Q4	SCORE=2.28 ZSCORE=17.62
14			NF1	VSNF1_Q6	SCORE=1.92 ZSCORE=19.384
18			AR	V\$AR_01	SCORE=1.8 ZSCORE=101.51

Total reported with SCORE $\geq$ 1.3 and ZSCORE $\geq$ 3.719 : 54

**Figure 3.16 Differentially enriched motifs found in SiFoxA1 unique over siCTRL unique**

The top 5 reported TF families are shown, together with AR and NF1 which has been reported by the author of the study. Total number of families that are above cutoff is 54.

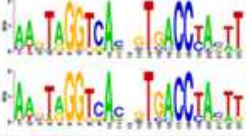
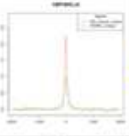
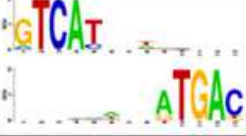
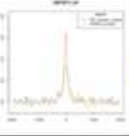

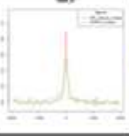

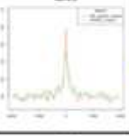

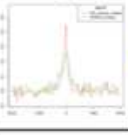
### 3.6.3.3 Cofactor modules: MCF7 ER vs ERAP2

Previously, using CENTDIST, we discovered AP2 gamma as a co-factor of ERa in MCF7. By partitioning the ERa peaks into those that contains AP2 gamma and those without, we can learn more about the role of AP2 gamma and what factors interacts specifically with AP2 gamma. Figure 3.17 shows the MOTIFDIFF comparison results. From the result, we see several enriched factors in the ERa AP2 overlapping set, noticeably E2F1 and SP1. Both of them have previously been found to be important co-factors of ERa. Using MOTIFDIFF, we manage to show that these factors possibly act in tandem with AP2 gamma since in the absence of AP2 gamma there is minimal enrichments of the factors.

Recently, Cao et al. (2011) performed E2F1 ChIP-seq in MCF7 cell line. To verify our findings, we perform overlap using the E2F1 ChIP-seq (see Figure 3.19(b)), and as we would expect, we found a large proportion of the ER AP2 overlapping sites to contain E2F1 binding peaks in comparison with the ER unique sites. Moreover, it also verified our assumption that our MOTIFDIFF raw scores is correlated with the actual percentage of peaks containing the co-TF.

As for SP1, no ChIP-seq has yet been performed in MCF7 cell line. However, interestingly the three transcription factors E2F, AP2 and SP1 are closely involved in the development of squamous cell carcinoma cell line. (Wong et al. 2005)

# ER unique over ERAP2


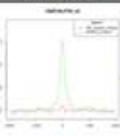

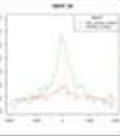
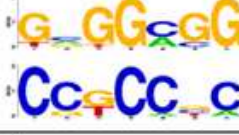
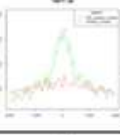
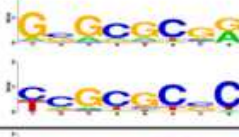
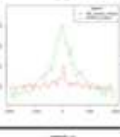
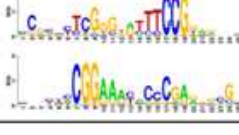

RANK	LOGO	IMAGE	FAMILY	MOTIF	SCORE
1			<a href="#">ERE</a>	VSPARG_02	SCORE=2.01 ZSCORE=49.516
2			<a href="#">AP1</a>	VSTCF11_01	SCORE=1.49 ZSCORE=12.782
3			<a href="#">AR</a>	VSAR_01	SCORE=1.43 ZSCORE=17.104
4			<a href="#">CREB</a>	VSATF6_01	SCORE=1.33 ZSCORE=9.745
5			<a href="#">EXR</a>	VSPXR_Q2	SCORE=1.31 ZSCORE=5.843

Total reported with SCORE $\geq$ 1.3 and ZSCORE $\geq$ 3.719 : 5

**Figure 3.17 Differentially enriched motifs found in ER unique sets over ERAP2.**

The top 5 reported TF families are shown. Total number of families that are above cutoff is 5.

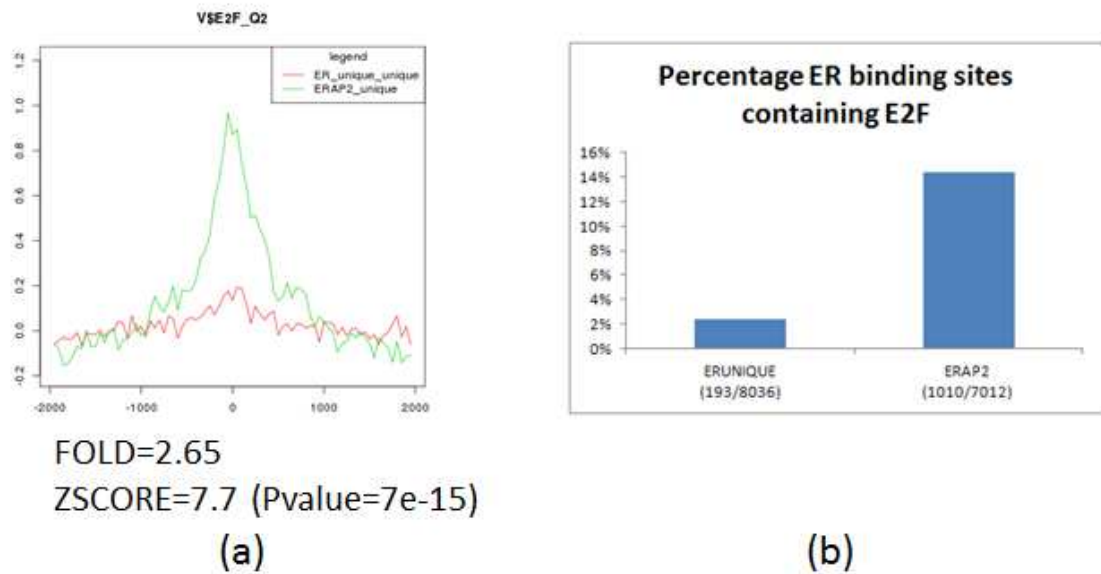
# ERAP2 over ER unique

RANK	LOGO	IMAGE	FAMILY	MOTIF	SCORE
1			<a href="#">AP2</a>	VSAP2ALPHA_02	SCORE=5 ZSCORE=39.845
2			<a href="#">E2F</a>	VSE2F_Q2	SCORE=2.65 ZSCORE=16.282
3			<a href="#">SP1</a>	VSETF_Q6	SCORE=2.19 ZSCORE=12.322
4			<a href="#">ZF5</a>	VSZF5_01	SCORE=1.91 ZSCORE=11.907
5			<a href="#">DEAF1</a>	VSDEAF1_01	SCORE=1.65 ZSCORE=8.322

Total reported with SCORE $\geq$ 1.3 and ZSCORE $\geq$ 3.719 : 27

**Figure 3.18 Differentially enriched motifs found in ERAP2 over ERunique.**

The top 5 reported TF families are shown. Total number of families that are above cutoff is 27.



**Figure 3.19 Verification that E2F is more enriched in ER AP2 overlapping sites as compared to ER unique sites**

a) Normalised motif histogram of E2F in ER containing AP2 binding (green) and those that do not (red). b) Percentage of ER binding containing E2F. We see that a higher percentage of ER colocalising with AP2 are bound with E2F as compared with those that do not.

### 3.7 Discussion

In this chapter, we developed a method that given two sets of ChIP-seq peaks attempt to predict the set of differentially enriched co-TFs in the two sets. Two main approaches are *de novo* motif finding and perform scanning using known motifs. Using a simulated example of varying degree of SP1 motifs implanted on two types of background: promoter and genomic, we show the various drawbacks of *de novo* motif finding and comparison using popular enrichment statistics such as binomial test pvalue and hypergeometric test pvalue and also overrepresentation against background. The main problems with these methods are the overpenalisation when background count is high. We consider the unpopular option of subtracting the background count without

penalisation and showed it is a good statistic for comparing actual abundance of TF in a set. Developing upon this statistic, we introduce HISTSCORE which assigns a score taking into account the shape of the entire histogram in general. Using large number of CTCF ChIP-seq experiments from ENCODE, we proved that HISTSCORE obtained using motif can predict the corresponding abundance using ChIP-seq peaks. Moving on, we set up an experiment using all ENCODE ChIP-seq in the five cell lines: K562, GM12878, HeLa-S3, H1-Hesc and HepG2 to test the overall prediction accuracy of MOTIFDIFF. We showed that under good circumstance where the motif is specific and of high quality such as CTCF, the prediction accuracy is good, having a precision and recall of 88.2% and 84.3% respectively. The ultimate downfall in the method however lies in the fact that not all bindings could be satisfactorily explained by motif. There are other factors in consideration such as chromatin structures, epigenetics, histone modification etc. Because of these factors, the general sensitivity of our method is not impressive. However, we showed that if we condition our predictions on just those “well-behaved” ones (i.e. those TFs for which  $\text{CHIPSCORE}_{\text{REV}} < 1$  and  $\text{CHIPSCORE} > 2$ ) our sensitivity is still quite acceptable at 73.1%. (as seen in Table 11)

Next, we applied MOTIFDIFF on several existing studies and verified their findings. In addition, we showed some potential co-TFs that are not reported by their studies that may have biological significance. Utilising our in-house generated AP2 and ER ChIP-seq in MCF7, we discovered two highly potential co-TF candidates E2F1 and SP1 that are likely to interact closely with AP2 and ER in MCF7. As a strong validation using E2F1 ChIP-seq generated in MCF7 by Cao et al. (2011), we see a significantly larger

percentage of ER in ERAP2 overlapping with E2F1 than those ER binding that do not contain AP2. As for SP1, no ChIP-seq on MCF7 has yet been performed but we show significant evidence using gene association that suggests its biological significance.

Though prediction using motifs alone are inherently weak, they are useful in giving insights and directions for later studies without incurring much experimental costs.

## CHAPTER 4     **Conclusions**

### **4.1 Summary of Contributions**

In this thesis, the topic is mainly to develop tools to aid biologist to identify potentially novel transcription factors making use of the high accuracy and precision of ChIP-seq and similar high throughput technology such as ChIA-pet. Making use of the high resolution of ChIP-seq (ability to predict the location of binding site to within a few hundred base pairs), we expect to be able to improve the sensitivity of co-TFs prediction using motifs. By analyzing the motif distribution around the ChIP-seq peaks, we are able to identify co-TFs with strong confidence. CENTDIST was thus developed to perform this. We showed that existing methods perform poorly as there was a need to specify a background. The selection of background greatly affects whether a motif would be enriched in a set of genomic regions. In addition to background selection, other parameters that need to be considered are the window size of the targeted regions and the motif matching score cutoff. Varying these parameters is not an easy task and also laborious for the average biologist. Henceforth, we develop CENTDIST as a user-friendly tool that is implemented as a web server and utilizes a scoring function that maximizes across all possible parameters. The method had been proven to perform significantly better than existing approaches, being able to identify the cofactors of AR in LNCaP prostate cancer cells with better accuracy than the existing methods. With the accuracy proven using known cofactors, we decide to look at novel co-TFs not yet known. One such candidate is AP4 which was chosen because it was not identified by the other motif enrichment programs. Our web lab validation and gene association suggests AP4 to be a probable co-factor of AR. We also performed motif enrichment analysis using

CENTDIST to identify novel cofactors of ER in MCF7 breast cancer cells. Our result identified AP2 gamma as a novel cofactor, and subsequent validation concludes that AP2 gamma potentially plays a pioneering role in the development of breast cancer, comparable to FOXA1.

Though CENTDIST solves the problem of identifying novel cofactors in a set of genomic regions, it cannot effectively compare motif enrichments among multiple sets of genomic regions. In Section 3.3 we discussed the difficulties in finding differential motifs. A common problem faced by existing motif enrichment methods is that their enrichment scores or pvalues cannot be compared directly. This therefore calls for a need to develop alternative method to perform such task, and hence we develop MOTIFDIFF to address this issue. MOTIFDIFF aims to solve the problem by doing normalizing to account for the background it resides in. The normalized enrichment score of a co-TF in a set of regions is the highest enrichment subtracted by the background noise. This score is designed to be correlated with the percentage of regions containing the actual co-TF binding. Using datasets from existing literature and our own in-house generated datasets, we verified MOTIFDIFF with the results obtained by them, and in addition we suggested several co-TFs that are potentially important specifically to a particular set. In particular, when comparing the cofactors in ER binding peaks for those that contain AP2 binding peaks and those that don't, we identified E2F and SP1 as cofactors that potentially works specifically in the presence of AP2. Using E2F ChIP-seq recently performed in MCF7, we verified the accuracy of our MOTIFDIFF model which asserted that the score reported by MOTIFDIFF is correlated and provide an accurate estimate of the percentage

of binding peaks containing the co-TF. For SP1, though no ChIP-seq has yet been performed. We decide to verify it using microarray. As we only have motif prediction of SP1 binding, usual gene association analysis is not sensitive enough to be able to detect the association of SP1 binding. As such, we devised a new method of performing gene association enrichment by considering the improvements introduced by subsequently adding more information. The approach is being verified by E2F which had been shown to be preferentially enriched in ERAP2 overlapping peaks. Using this approach, we show that the additional information of SP1 leads to improved prediction of the upregulation of genes by E2 at 6h. Precisely how and why they interact specifically with AP2 are yet to be known, but there are studies that showed that the three transcription factors E2F, AP2 and SP1 are closely involved in the development of squamous cell carcinoma cell line. (Wong et al. 2005)

Accompanying each tool is a web interface that enables biologist with no computational background to use with. These tools will greatly help advance biologists to develop understanding in transcription factors related to their area of study.

## **4.2 Future Directions**

Though we managed to show the merits of our two developed tools CENTDIST and MOTIFDIFF, there is still much room left for improvements. In the following sections, we will describe some of the ways that we could improve them.

## **4.2.1 CENTDIST**

### **4.2.1.1 Comotif rank preference**

In our current implementation of CENTDIST, though we provided a histogram displaying comotif occurrence preference with respect to the ranking of ChIP-seq peak intensities as an additional output for users to consider in their assessment, we did not incorporate it into our CENTDIST scoring. Certain co-TFs exhibit a preference to TFs having stronger ChIP-seq intensity, while some exhibit preference towards lower ChIP-seq intensity. An explanation for why some co-TFs preferentially bound to low intensity ChIP-seq peak is that the lower intensity ChIP-seq peaks are indirectly bound that are picked up subtly via the ChIP-seq protocol. At the moment this feature serves as extra information for user's assessments. Given more observations we could possibly devise a scoring function that incorporate it so as to improve the sensitivity and specificity of our method.

### **4.2.1.2 Utilise conservation scores**

Motifs occurring in highly conserved DNA regions are more likely to be functional and that is the reason why many methods had been developed to zoom in on conserved regions. Utilising conservation scores such as PHASTCONS (Siepel et al. 2005) in conjunction with motif scanning may help to improve the specificity of motif predictions which in turn improve the sensitivity and specificity of co-TFs identified.

## **4.2.2 MOTIFDIFF**

### **4.2.2.1 More Validations for MOTIFDIFF**

Due to lack of studies with appropriate validations, we do not yet have very solid evidence to prove the effectiveness of MOTIFDIFF in identifying novel differential motifs. In section 3.6.3.1 and 3.6.3.2 we predicted a number of candidate differential motifs that are not being reported by the respective studies. Given more time and resources, we would hope to perform ChIP-seq on the respective conditions to help validate our findings.

### **4.2.2.2 Differential *de novo* motif finding**

After developing a method to screen database of motifs to determine differential enrichments, the next step would be to extend the idea to identify motifs a priori. While there are already several algorithms performing discriminative motif finding, none utilize the motif shape information around ChIP-seq peak as done by MOTIFDIFF. We believe that utilizing such information could improve the sensitivity and specificity in which novel interesting motifs could be detected.

### **4.2.2.3 Utilising quantitative differential ChIP-seq signal obtained using MAnorm(Shao et al. 2012)**

Currently in MOTIFDIFF, given two sets of ChIP-seq peaks that are called using some peak calling programs, we perform peak overlap to categorize the peak sets as described in Section 3.1 into three mutually exclusive sets, consisting of one overlapping sets and two non-overlapping sets. MOTIFDIFF will then proceed to compare the motif signals in the two non-overlapping sets. The downside of this is that in doing so, we discard a large

portion of the peak sets which belong to the overlapping sets. One way to overcome this problem is to have a way of quantifying the level of differential ChIP-seq signal among the two ChIP-seq libraries. Shao et al. (2012) describe a novel methodology to do that called MAnorm. MAnorm utilizes the raw reads of the two ChIP-seq libraries in addition to their peak coordinates to obtain the differential ChIP-seq signal among the two ChIP-seq libraries that are normalized to account for biases due to sequencing depth and background noise etc. One down side is that to perform MAnorm we require the original reads in addition to the peak coordinates, which may not be readily available and would require a long time to be uploaded to the server. Alternatively we could accept the outputs of MAnorm which had been run separately on the user's computer as inputs for MOTIFDIFF.

#### **4.2.2.4 Compare across multiple peak sets**

As described in Section 3.4, the HISTSCORE derived in MOTIFDIFF is suitable for comparison among multiple libraries. This can then be plotted using a heat map and for clustering. HISTSCORE suffer less from background-related issues and sample-size related issues than p-value-based enrichment statistics and several statistics discussed in the section.

## PUBLICATIONS

### First author papers

Zhang Z\*, Chang CW\*, Goh WL, Sung WK and Cheung E (2011). "CENTDIST: discovery of co-associated factors by motif distribution." Nucleic Acids Res **39**(Web Server issue): W391-399.

### Co-author papers

Chng KR, Chang CW, Tan SK, Yang C, Hong SZ, Sng NY and Cheung E (2012). "A transcriptional repressor co-regulatory network governing androgen response in prostate cancers." EMBO J **31**(12): 2810-2823.

Tan PY, Chang CW, Chng KR, Wansa KD, Sung WK and Cheung E (2012). "Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival." Mol Cell Biol **32**(2): 399-414.

Tan SK, Lin ZH, Chang CW, Varang V, Chng KR, Pan YF, Yong EL, Sung WK and Cheung E (2011). "AP-2gamma regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription." EMBO J **30**(13): 2569-2581.

Zhang Z, Chang CW, Hugo W, Cheung E and Sung WK (2013). "Simultaneously learning DNA motif along with its position and sequence rank preferences through expectation maximization algorithm." J Comput Biol **20**(3): 237-248.

## BIBLIOGRAPHY

- Akutsu Y, Shuto K, Kono T, Uesato M, Hoshino I, Shiratori T, Miyazawa Y, Isozaki Y, Akanuma N and Matsubara H (2012). "A phase 1/11 study of second-line chemotherapy with fractionated docetaxel and nedaplatin for 5-FU/cisplatin-resistant esophageal squamous cell carcinoma." Hepatogastroenterology **59**(119): 2095-2098.
- Albergaria A, Paredes J, Sousa B, Milanezi F, Carneiro V, Bastos J, Costa S, Vieira D, Lopes N, Lam EW, Lunet N and Schmitt F (2009). "Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours." Breast Cancer Res **11**(3): R40.
- Aparicio O, Geisberg JV and Struhl K (2004). "Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo." Curr Protoc Cell Biol **Chapter 17**: Unit 17 17.
- Bailey TL (2011). "DREME: motif discovery in transcription factor ChIP-seq data." Bioinformatics **27**(12): 1653-1659.
- 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**(Web Server issue): W202-208.
- Barash Y, Bejerano G and Friedman N (2001). A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites. Proceedings of the First International Workshop on Algorithms in Bioinformatics, Springer-Verlag: 278-293.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I and Zhao K (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.
- Bembom O, Keles S and van der Laan MJ (2007). "Supervised detection of conserved motifs in DNA sequences with cosmo." Stat Appl Genet Mol Biol **6**: Article8.
- Benos PV, Bulyk ML and Stormo GD (2002). "Additivity in protein-DNA interactions: how good an approximation is it?" Nucleic Acids Res **30**(20): 4442-4451.
- Bernstein BE, Birney E, Dunham I, Green ED, Gunter C and Snyder M (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.
- Boulay A, Breuleux M, Stephan C, Fux C, Briskin C, Fiche M, Wartmann M, Stumm M, Lane HA and Hynes NE (2008). "The Ret receptor tyrosine kinase pathway functionally interacts with the ERalpha pathway in breast cancer." Cancer Res **68**(10): 3743-3751.
- Cao AR, Rabinovich R, Xu M, Xu X, Jin VX and Farnham PJ (2011). "Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome." J Biol Chem **286**(14): 11985-11996.
- Cao J, Tang M, Li WL, Xie J, Du H, Tang WB, Wang H, Chen XW, Xiao H and Li Y (2009). "Upregulation of activator protein-4 in human colorectal cancer with metastasis." International journal of surgical pathology **17**(1): 16-21.

- Chang JT and Nevins JR (2006). "GATHER: a systems approach to interpreting genomic signatures." *Bioinformatics* **22**(23): 2926-2933.
- Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-H, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-K, Clarke ND, Wei C-L and Ng H-H (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." *Cell* **133**(6): 1106-1117.
- Cheung E and Kraus WL (2010). "Genomic analyses of hormone signaling and gene regulation." *Annual review of physiology* **72**: 191-218.
- Chi SG, deVere White RW, Meyers FJ, Siders DB, Lee F and Gumerlock PH (1994). "p53 in prostate cancer: frequent expressed transition mutations." *Journal of the National Cancer Institute* **86**(12): 926-933.
- Choi J, Lee MK, Oh KH, Kim YS, Choi HY, Baek SK, Jung KY, Woo JS, Lee SH and Kwon SY (2011). "Interaction effect between the receptor for advanced glycation end products (RAGE) and high-mobility group box-1 (HMGB-1) for the migration of a squamous cell carcinoma cell line." *Tumori* **97**(2): 196-202.
- Crooks GE, Hon G, Chandonia JM and Brenner SE (2004). "WebLogo: a sequence logo generator." *Genome Res* **14**(6): 1188-1190.
- Dai W, Zhang EJ, Duan WY and Zhou Q (2012). "[Overexpression of Twist1 promotes tumor invasion in human tongue squamous cell carcinoma cell line Tca8113]." *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi* **28**(12): 1246-1249.
- Dydensborg AB, Rose AA, Wilson BJ, Grote D, Paquet M, Giguere V, Siegel PM and Bouchard M (2009). "GATA3 inhibits breast cancer growth and pulmonary breast cancer metastasis." *Oncogene* **28**(29): 2634-2642.
- Elemento O, Slonim N and Tavazoie S (2007). "A universal framework for regulatory element discovery across all genomes and data types." *Mol Cell* **28**(2): 337-350.
- Ellington AD and Szostak JW (1990). "In vitro selection of RNA molecules that bind specific ligands." *Nature* **346**(6287): 818-822.
- Ergen HA, Narter F, Timirci O and Isbir T (2007). "Effects of manganese superoxide dismutase Ala-9Val polymorphism on prostate cancer: a case-control study." *Anticancer research* **27**(2): 1227-1230.
- Ettwiller L, Paten B, Ramialison M, Birney E and Wittbrodt J (2007). "Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation." *Nat Methods* **4**(7): 563-565.
- Frasor J, Danes JM, Komm B, Chang KC, Lyttle CR and Katzenellenbogen BS (2003). "Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype." *Endocrinology* **144**(10): 4562-4574.
- Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E and Ruan Y (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome." *Nature* **462**(7269): 58-64.

- Gelfond JA, Gupta M and Ibrahim JG (2009). "A Bayesian hidden Markov model for motif discovery through joint modeling of genomic sequence and ChIP-chip data." *Biometrics* **65**(4): 1087-1095.
- He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, Mieczkowski P, Lieb JD, Zhao K, Brown M and Liu XS (2010). "Nucleosome dynamics define transcriptional enhancers." *Nature genetics* **42**(4): 343-347.
- He YF, Ji CS, Hu B, Fan PS, Hu CL, Jiang FS, Chen J, Zhu L, Yao YW and Wang W (2013). "A phase II study of paclitaxel and nedaplatin as front-line chemotherapy in Chinese patients with metastatic esophageal squamous cell carcinoma." *World J Gastroenterol* **19**(35): 5910-5916.
- Hestand MS, van Galen M, Villerius MP, van Ommen GJ, den Dunnen JT and t Hoen PA (2008). "CORE\_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes." *BMC Bioinformatics* **9**: 495.
- Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT and Wasserman WW (2007). "oPOSSUM: integrated tools for analysis of regulatory motif over-representation." *Nucleic Acids Res* **35**(Web Server issue): W245-252.
- Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP and Wasserman WW (2005). "oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes." *Nucleic Acids Res* **33**(10): 3154-3164.
- Hooghe B, Hulpiau P, van Roy F and De Bleser P (2008). "ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species." *Nucleic Acids Res* **36**(Web Server issue): W128-132.
- Iwagami S, Baba Y, Watanabe M, Shigaki H, Miyake K, Ida S, Nagai Y, Ishimoto T, Iwatsuki M, Sakamoto Y, Miyamoto Y and Baba H (2012). "Pyrosequencing assay to measure LINE-1 methylation level in esophageal squamous cell carcinoma." *Ann Surg Oncol* **19**(8): 2726-2732.
- Ji X, Li W, Song J, Wei L and Liu XS (2006). "CEAS: cis-regulatory element annotation system." *Nucleic Acids Res* **34**(Web Server issue): W551-554.
- Johnson DS, Mortazavi A, Myers RM and Wold B (2007). "Genome-wide mapping of in vivo protein-DNA interactions." *Science* **316**(5830): 1497-1502.
- Kang Z, Jänne OA and Palvimo JJ (2004). "Coregulator recruitment and histone modifications in transcriptional regulation by the androgen receptor." *Molecular endocrinology (Baltimore, Md.)* **18**(11): 2633-2648.
- Langmead B and Salzberg SL (2012). "Fast gapped-read alignment with Bowtie 2." *Nat Methods* **9**(4): 357-359.
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." *Science* **262**(5131): 208-214.
- Li H and Durbin R (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* **25**(14): 1754-1760.
- Liang S, Samanta MP and Biegel BA (2004). "cWINNOWER algorithm for finding fuzzy dna motifs." *J Bioinform Comput Biol* **2**(1): 47-60.
- Linhart C, Halperin Y and Shamir R (2008). "Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets." *Genome Res* **18**(7): 1180-1189.

- Liu XS, Brutlag DL and Liu JS (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nat Biotechnol **20**(8): 835-839.
- Lupien M, Eeckhoute J, Meyer CA, Krum SA, Rhodes DR, Liu XS and Brown M (2009). "Coactivator function defines the active estrogen receptor alpha cistrome." Mol Cell Biol **29**(12): 3413-3423.
- Machiels JP, Kaminsky MC, Keller U, Brummendorf TH, Goddemeier T, Forssmann U and Delord JP (2013). "Phase Ib trial of the Toll-like receptor 9 agonist IMO-2055 in combination with 5-fluorouracil, cisplatin, and cetuximab as first-line palliative treatment in patients with recurrent/metastatic squamous cell carcinoma of the head and neck." Invest New Drugs **31**(5): 1207-1216.
- Maggiolini M, Recchia AG, Carpino A, Vivacqua A, Fasanella G, Rago V, Pezzi V, Briand P-A, Picard D and Andò S (2004). "Oestrogen receptor beta is required for androgen-stimulated proliferation of LNCaP prostate cancer cells." Journal of molecular endocrinology **32**(3): 777-791.
- Mason MJ, Plath K and Zhou Q (2010). "Identification of context-dependent motifs by contrasting ChIP binding data." Bioinformatics **26**(22): 2826-2832.
- Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S and Wingender E (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-378.
- Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA and Bulyk ML (2004). "Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays." Nat Genet **36**(12): 1331-1339.
- Narlikar L and Jothi R (2012). "ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder." Methods Mol Biol **802**: 305-322.
- Park JJ, Irvine RA, Buchanan G, Koh SS, Park JM, Tilley WD, Stallcup MR, Press MF and Coetzee GA (2000). "Breast cancer susceptibility gene 1 (BRCA1) is a coactivator of the androgen receptor." Cancer research **60**(21): 5946-5949.
- Ponty Y, Termier M and Denise A (2006). "GenRGenS: software for generating random genomic sequences and structures." Bioinformatics **22**(12): 1534-1535.
- Qi L and Zhang Y (2014). "Truncation of inhibitor of growth family protein 5 effectively induces senescence, but not apoptosis in human tongue squamous cell carcinoma cell line." Tumour Biol **35**(4): 3139-3144.
- Qiao B, Johnson NW, Chen X, Li R, Tao Q and Gao J (2011). "Disclosure of a stem cell phenotype in an oral squamous cell carcinoma cell line induced by BMP-4 via an epithelial-mesenchymal transition." Oncol Rep **26**(2): 455-461.
- Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K and Snyder M (2010). "Close association of RNA polymerase II and many transcription factors with Pol III genes." Proceedings of the National Academy of Sciences of the United States of America **107**(8): 3639-3644.
- Redhead E and Bailey TL (2007). "Discriminative motif discovery in DNA and protein sequences using the DEME algorithm." BMC Bioinformatics **8**: 385.

- Roider HG, Manke T, O'Keeffe S, Vingron M and Haas SA (2009). "PASTAA: identifying transcription factors associated with sets of co-regulated genes." Bioinformatics **25**(4): 435-442.
- Roth FP, Hughes JD, Estep PW and Church GM (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol **16**(10): 939-945.
- Sandelin A, Alkema W, Engstrom P, Wasserman WW and Lenhard B (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-94.
- Sandelin A, Alkema W, Engström P, Wasserman WW and Lenhard B (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Research **32**(Database issue): D91-94.
- Shao Z, Zhang Y, Yuan GC, Orkin SH and Waxman DJ (2012). "MANorm: a robust model for quantitative comparison of ChIP-Seq data sets." Genome Biol **13**(3): R16.
- Sharov AA and Ko MS (2009). "Exhaustive search for over-represented DNA sequence motifs with CisFinder." DNA Res **16**(5): 261-273.
- Sharov AA and Ko MSH (2009). "Exhaustive search for over-represented DNA sequence motifs with CisFinder." DNA research : an international journal for rapid publication of reports on genes and genomes **16**(5): 261-273.
- Shin H, Liu T, Manrai AK and Liu XS (2009). "CEAS: cis-regulatory element annotation system." Bioinformatics **25**(19): 2605-2606.
- Siddharthan R (2010). "Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix." PLoS One **5**(3): e9722.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W and Haussler D (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-1050.
- Sinha S and Tompa M (2003). "YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation." Nucleic Acids Res **31**(13): 3586-3588.
- Smith AD, Sumazin P and Zhang MQ (2005). "Identifying tissue-selective transcription factor binding sites in vertebrate promoters." Proc Natl Acad Sci U S A **102**(5): 1560-1565.
- Stormo GD (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.
- Tennakoon C, Purbojati RW and Sung WK (2012). "BatMis: a fast algorithm for k-mismatch mapping." Bioinformatics **28**(16): 2122-2128.
- Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P and Moreau Y (2001). "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling." Bioinformatics **17**(12): 1113-1122.
- Tozlu S, Girault I, Vacher S, Vendrell J, Andrieu C, Spyrtos F, Cohen P, Lidereau R and Bieche I (2006). "Identification of novel genes that co-cluster with estrogen receptor alpha in breast tumor biopsy specimens, using a large-scale real-time reverse transcription-PCR approach." Endocr Relat Cancer **13**(4): 1109-1120.

- Turner BC, Zhang J, Gumbs AA, Maher MG, Kaplan L, Carter D, Glazer PM, Hurst HC, Haffty BG and Williams T (1998). "Expression of AP-2 transcription factors in human breast cancer correlates with the regulation of multiple growth factor signalling pathways." *Cancer research* **58**(23): 5466-5472.
- van Helden J, Andre B and Collado-Vides J (1998). "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." *J Mol Biol* **281**(5): 827-842.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A and Zhu X (2001). "The sequence of the human genome." *Science (New York, N.Y.)* **291**(5507): 1304-1351.

- Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG and Fu XD (2011). "Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA." Nature **474**(7351): 390-394.
- Wederell ED, Bilenky M, Cullum R, Thiessen N, Dagpinar M, Delaney A, Varhol R, Zhao Y, Zeng T, Bernier B, Ingham M, Hirst M, Robertson G, Marra MA, Jones S and Hoodless PA (2008). "Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing." Nucleic Acids Res **36**(14): 4549-4564.
- Wong CF, Barnes LM, Dahler AL, Smith L, Popa C, Serewko-Auret MM and Saunders NA (2005). "E2F suppression and Sp1 overexpression are sufficient to induce the differentiation-specific marker, transglutaminase type 1, in a squamous cell carcinoma cell line." Oncogene **24**(21): 3525-3534.
- Workman CT and Stormo GD (2000). "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity." Pac Symp Biocomput: 467-478.
- Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F and Sung WK (2010). "A signal-noise model for significance analysis of ChIP-seq with negative control." Bioinformatics **26**(9): 1199-1204.
- Xu JY, Yang LL, Ma C, Huang YL, Zhu GX and Chen QL (2013). "MiR-25-3p attenuates the proliferation of tongue squamous cell carcinoma cell line Tca8113." Asian Pac J Trop Med **6**(9): 743-747.
- Zhang S, Li S, Niu M, Pham PT and Su Z (2011). "MotifClick: prediction of cis-regulatory binding sites via merging cliques." BMC Bioinformatics **12**: 238.
- Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W and Liu XS (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.
- Zhang Z, Chang CW, Goh WL, Sung WK and Cheung E (2011). "CENTDIST: discovery of co-associated factors by motif distribution." Nucleic Acids Res **39**(Web Server issue): W391-399.