Computational Identification of Thyroid Response Elements in Genomic DNA

**By
Rémi Gagné**

**A thesis submitted to
The Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Master of Computer Science**

**Ottawa-Carleton Institute for Computer Science
School of Computer Science
Carleton University
Ottawa, Ontario**

April 2010

**© Copyright
2010, Rémi Gagné**

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN: 978-0-494-68634-8
Our file *Notre référence*
ISBN: 978-0-494-68634-8

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Due to the volume and complexity of data arising from high throughput biological assays, computational analysis becomes increasingly important to assist biologists in forming and testing hypotheses. In the current study, bioinformatics is applied to the fields of microbiology and toxicogenomics in analyzing chIP-chip data to study the thyroid hormone receptor conducted by Health Canada. This data analysis requires normalization and signal detection. A survey of contemporary methods was performed in order to find the most appropriate model for each step, given our experimental platform. Proof of concept experiments using high quality benchmark data revealed that normalization for chIP-chip data didn't improve the accuracy of subsequent peak finding algorithms. Splitter was used to detect peaks, which revealed 230 regions for which the thyroid hormone receptor is believed to be bound to DNA. Once signal detection was complete, the identified DNA segments were examined to model the degenerate sequence motif. Motif finding algorithms (MFAs) from a number of underlying statistical models were also applied to find occurrences of novel motifs not previously known to be linked to the thyroid hormone receptor. In total, 105 thyroid hormone receptor binding sites (thyroid response elements) were identified with an expected false discovery rate of 20%. MFAs found motifs which are very similar to known binding sites for proteins which could interact with the thyroid hormone receptor, such as SP-1, PAX and KROX binding sites. A wet laboratory validation of these sites is now needed in order to reveal the functionality of these sites, i.e. whether the identified motifs truly exhibit a gene regulation function.

Acknowledgments

I would sincerely like to thank everyone that has been involved in this project. Particularly, people that are directly involved in the project who have been providing support from the beginning; Dr. Hongyang Dong who generated the data, Andrew Williams the statistician at Health Canada and all my coworkers (Byron Kuo, John Gingerich and many others) that were there for me in the best and worst times.

I would like to also thank my family for their understanding in these sometimes pretty stressful moments, a especially my lovely wife Paula and my little “pitchounette” Catherine.

I am also grateful to the members of my committee, Prof. Dehne and Prof. Famili who generously agreed to spend time examining this document.

I am very grateful to Health Canada and my work supervisor Dr. Paul White, who supported me financially during this project. And last but not least, my two supervisors, Dr. Carole Yauk and Dr. James Green who agreed to take me under their wing to help me produce this document.

Table of Contents

ABSTRACT	II
ACKNOWLEDGMENTS	III
TABLE OF CONTENTS.....	IV
LIST OF TABLES	VII
LIST OF FIGURES	IX
LIST OF ALGORITHMS	XII
LIST OF EQUATIONS.....	XIII
CHAPTER 1. INTRODUCTION	1
1.1. MOTIVATION.....	4
1.2. OBJECTIVES	4
1.3. THESIS OUTLINE	4
CHAPTER 2. BIOLOGICAL AND TECHNOLOGICAL REVIEW.....	6
2.1. BIOLOGICAL BACKGROUND	6
2.1.1. Thyroid hormones.....	6
2.1.2. Thyroid hormone receptor	7
2.1.3. THR partners for gene regulation	9
2.1.4. Thyroid response elements	10
2.1.5. Summary.....	15
2.2. BIOLOGICAL LABORATORY TECHNOLOGY OVERVIEW AND EXPERIMENTAL DESIGN.....	16
2.2.1. Chromatin immunoprecipitation.....	16
2.2.2. Microarrays.....	17
2.2.3. Biological experimental design	19

2.3. SUMMARY	21
CHAPTER 3. CHIP-CHIP DATA ANALYSIS.....	22
3.1. MAPPING PROBES TO GENOME.....	24
3.2. NORMALIZATION OF CHIP-CHIP DATA	24
3.2.1. Normalization methods developed for gene expression microarrays	25
3.2.2. ChIP-chip normalization methods	30
3.2.3. Summary.....	38
3.3. PEAK FINDING ALGORITHM	39
3.3.1. Splitter [20].....	40
3.3.2. Summary.....	44
3.4. EVALUATION OF PERFORMANCE WITH BENCHMARK DATA	44
3.4.1. Evaluation of the precision of binding site cut-off	48
3.5. OPTIMISATION OF PEAK-FINDING ALGORITHM PARAMETERS TO ACTUAL THR STUDY DATA	50
3.6. RESULTS OF PEAK FINDING TO ACTUAL THR STUDY DATA	51
3.7. EXPERIMENTAL VALIDATION	54
3.8. SUMMARY.....	55
CHAPTER 4. MOTIF IDENTIFICATION	56
4.1. SEARCHING FOR THE KNOWN CONSENSUS TRE MOTIF	56
4.1.1. Models for the identification of the TRE hexamer.....	57
4.1.2. Determination of the correct DNA scanning model for TREs	59
4.1.3. Relative abundance of TRE hexamers in DNA sequences.....	63
4.1.4. Analysis of TRE ChIP-chip sequences for the THR and AP-1 binding site ..	68
4.1.5. Summary.....	70
4.2. IDENTIFICATION OF NOVEL MOTIFS	71
4.2.1. Application of motif finding algorithms to the TRE dataset	73
4.2.2. Results of MFAs on the TRE dataset	76
4.2.3. Summary of the utilization of MFAs.....	91

4.3. SUMMARY	94
CHAPTER 5. CONCLUSION	97
5.1. SUMMARY OF RESEARCH	97
5.2. MAJOR CONCLUSIONS.....	98
5.2.1. Normalization.....	98
5.2.2. Peak finding.....	98
5.2.3. TRE consensus motif searching.....	99
5.2.4. Novel TRE motif searching	99
5.3. FUTURE WORK	99
BIBLIOGRAPHY	101
APPENDIX A	109
APPENDIX B	110
APPENDIX C	121
APPENDIX D	124
APPENDIX E	126

List of Tables

Table 2-1: TRE arrangements for hetero and homo dimer TRE configurations	12
Table 2-2: List of TREs in mouse genome compiled from the literature. the gene that is regulated by the TRE, its accession number, its DNA strand(GS), the location with respect to the transcription start site of the gene, the strand of the TRE (TS), the sequence which contains the TRE with binding site in bold, the type of gene regulation that the TRE performs (up or down regulates gene transcription), the literature reference for the TRE and the TRE configuration are shown.	13
Table 3-1: Comparison of the platform used for Spike-in and our chIP-chip data.....	45
Table 3-2: Results from the Whitehead data set with Splitter at 2.5 SD	46
Table 3-3: Location of peaks with respect to mRNA mapped in MM9 by the UCSC genome browser	52
Table 4-1: Scores of halfSites for murine TREs.....	61
Table 4-2: Scores of halfSites for rat, human, chicken TREs.....	61
Table 4-3: Sample of True Positive and False Positive rate vs. Min Score and Max Score	63
Table 4-4: Output of Bipad (left half site motif logo, the distribution of the length of the spacer, and the right half site motif logo)	77
Table 4-5: Results of the MEME analysis using all the sequences in the TRE chIP-chip dataset ranked by <i>E</i> -values.....	80
Table 4-6: Results of the MEME analysis using TOP5-PND4&TOP25-PND15 sequences in the TRE chIP-chip dataset ranked by <i>E</i> -values.....	82
Table 4-7: Motif targets using TOMTOM for a few example cases	84
Table 4-8: Motif found by MEME (Table 4-5 – 4 and Table 4-6 - 3) in the first column and the SP-1 motif in the Transfac database in column 2.....	85
Table 4-9: Top 3 motifs found by Bioprospector and percentage of motif hit in CpG islands.....	87

Table 4-10: Motif targets using TOMTOM for a few example cases	88
Table 4-11: Top 4 motifs found by Weeder and percentage of motifs in CpG islands	89
Table 4-12: Motif targets using TOMTOM for a few example cases	90

List of Figures

Figure 1-1: Schema of gene promoter region	2
Figure 1-2: Flowchart of experimental process	3
Figure 2-1: Illustration of T3 (left) and T4 (right) hormones	6
Figure 2-2: Sequence (1D) and Structure (3D) of Nuclear Receptors.....	8
Figure 2-3: Logo of TRE hexamer.....	11
Figure 2-4: Mechanism of T3 regulated gene (activation and repression) with FOS and JUN nuclear receptor interaction	14
Figure 2-5: AP-1 binding site sequence logo.....	15
Figure 2-6: Each step of chromatin immunoprecipitation	16
Figure 2-7: Example of a microarray with a zoom-in on some wells.....	19
Figure 2-8: Distribution of the resolution for the "A" (black line) and "B" (blue line) microarray (please note that blue and black line overlaps greatly).....	20
Figure 2-9: Experimental Design of Biological Experiment	20
Figure 3-1: Example of raw chip-chip data and confirmed TRE expressed by probes circled in red. Lines indicate the probe location (x-axis) and the intensity of the Red (ChIP DNA) to the Green (total input DNA).	23
Figure 3-2: Partial flowchart of thesis with normalization algorithms highlighted.....	25
Figure 3-3: Lowess normalization performed on a microarray dataset. a) Data	28
Figure 3-4: Lowess regression performed on the combined replicates (median value) of PND15.....	29
Figure 3-5: Histogram of distribution of TI probes, figure to be compared to Fig. 3a of Peng <i>et al.</i>	31
Figure 3-6: Nucleotide effect along the probe position for the TI channel where a quadratic effect is shown regarding the base (A,G,C) position in the probe. 35	
Figure 3-7: Median Log2 intensity of TI probes vs. the number of individual nucleotides in each probe	36

Figure 3-8: Median corrected probe intensity vs. number of each nucleotide in the probe. The intense parabolic effects was been attenuated tremendously showing no correlation with respect the number of individual nucleotides in the probes (part A of Equation 3).	38
Figure 3-9: Median probe value corrected vs. each nucleotide position in each probe. This figure shows the residuals of the nucleotide position effect in the probe. The narrow Y range indicates that the correction for this factor was useful (part B of Equation 3).	38
Figure 3-10: Partial flowchart of thesis with peak finding algorithm highlighted	39
Figure 3-11: Shows the manual cut-off options (from (a) above) for the combined replicates (median value) for PND15B.	41
Figure 3-12: Example for Splitter in Dynamic mode for Splitter with example values. ...	42
Figure 3-13: Illustration of the 4 cases when a predicted binding site is called as a true positive	46
Figure 3-14: ROC-like curve for the Whitehead dataset	48
Figure 3-15: Difference between the true and predicted start values for the true positives in the Whitehead dataset	49
Figure 3-16: Difference between the true and predicted end values for the true positives in the Whitehead dataset	50
Figure 3-17: Histogram of the distance to the nearest mRNA mapped by the UCSC on build MM9	53
Figure 4-1: Partial flowchart of thesis with Identification of sites that are consistent with the consensus TRE highlighted.....	56
Figure 4-2: Histogram of Scores from half sites from the literature.....	60
Figure 4-3: Frequency of TRE half sites per kbps found in relevant DNA regions	65
Figure 4-4: Density of TRE half sites in promoter regions of 4 TH regulated genes. Each graph shows the NCBI Reference sequence ID of the T3 regulated gene, and its chromosomal location. The Y axes indicate the number of half sites and the X axis indicate the nucleotide position with respect to the gene where 0 = -8kps from the transcription start site. The arrows show precisely where the TREs are located for each known TH regulated gene.	67

Figure 4-5: Histogram of the Scores for the AP-1 Binding Site from Jaspar	70
Figure 4-6: Partial flowchart of thesis with identification of novel TRE highlighted	71
Figure 4-7: Histogram of the distribution of GROIs length presented in Appendix B.....	72
Figure 4-8: Locations of binding motifs in a few test cases	93

List of Algorithms

Algorithm 1: Quantile normalization algorithm	26
Algorithm 2: Lowess algorithm	28
Algorithm 3: The splitter algorithm	43

List of Equations

Equation 1: Log transform of chip signal (R: red channel intensity, G: green channel intensity).....	23
Equation 2: GC content based normalization and parameter estimation.....	31
Equation 3: Quadratic model for chIP-chip normalization derived from Potter <i>et al.</i>	33
Equation 4: Equation for the expected probes per peak	39
Equation 5: Information theory model for finding transcription factor binding site	58
Equation 6: Scoring function for the information theory model.....	58
Equation 7: low and high score instantiation	62

Chapter 1. Introduction

Hormones are small chemical molecules used by cells as communicating agents to send messages across an organism. They are produced by endocrine glands which introduce them directly into the blood stream. Hormones target specific proteins called hormone receptors (HRs), to which they bind in order to achieve their metabolic function. For example, a hormone (the ligand) may bind to a transcription factor (the receptor) which then permits the expression of a gene. As an analogy, hormones are the key that binds to the lock (hormone receptor) leading to other biochemical reactions in the cell. For steroid and thyroid hormones, receptor binding to key elements of deoxyribonucleic acid (DNA) can result in downstream production of the associated gene product (called transcription). The presence or absence of the receptor ligand (the hormone) determines the regulation of the downstream gene. Hormone function can be disrupted by external factors such as xenobiotics. Xenobiotics are chemicals that are present in the natural environment that do not normally occur in nature (i.e., pollutants). Xenobiotic chemicals have been demonstrated to interfere with various hormone-receptor mediated processes [56] [9].

The focus of this thesis is on the interaction of the thyroid hormone, its cognate receptor (thyroid receptor beta), and its DNA binding sites. In order to characterize the mechanisms by which xenobiotics disrupt thyroid hormone action, this project focuses on characterizing DNA binding sequences of the thyroid hormone receptor (THR). This hormone was selected among others due to its crucial role in mammalian development. An unusually low level of thyroid hormone can lead to serious consequences such as mental retardation [43]. Thyroid hormones play an important role in growth, metabolism, and physiological functions of many tissues across many stages of development [62]. Figure 1-1 depicts the promoter region of a gene that is regulated by a transcription factor and a hormone. The dashed portion represents the DNA binding site, which is the sequence of DNA recognized by the receptor.

Figure 1-1: Schema of gene promoter region

HRs recognize a pattern of nucleotides on DNA ('Binding Site' in Figure 1-1) to which they bind to regulate expression for a specific gene. The location on DNA to which the hormone-protein complex binds is referred to as the binding site or the response element (RE). The specific DNA sequence that makes up the RE depends on the specific hormone receptor to which it binds. While these RE sequences are fairly well conserved within a genome and across organisms, they still show considerable sequence variation, and are therefore considered to be degenerate. Factors other than the type of receptor can explain the RE degeneracy, such as mutations in DNA. The binding surface of the nuclear receptor allows some flexibility with respect to the sequence it binds. In the current project, response elements for thyroid hormone are referred to as thyroid response elements (TREs). This brief introduction on thyroid hormone gene regulation is covered in greater detail in chapter 2.

Methods to identify and characterize REs require experiments in a biological laboratory as well as computational models to interpret the raw data gathered by biologists. Chromatin immunoprecipitation paired with microarray (chIP-chip) is a wetlab method performed to isolate hormone receptor proteins with their respective REs. Once chIP-chip is performed, isolated DNA segments, i.e. binding sites, are exposed to their matching complement on a microarray. Biologists gather signals from the

microarray with a laser scanner, allowing the identification of genomic regions that may contain REs (i.e., are enriched following immunoprecipitation of the target receptor). Further experiments are required to confirm the presence of the RE, and to determine the exact location of the RE within the genomic region. This brief introduction to chIP-chip will be given in greater detail in Chapter 2.

Data gathered by biologists need to be normalized in order to remove non-biological artefacts which impart biases on the raw data. Statistical models are implemented in order to process the information and correct for biases. Computational models are then employed to identify enriched DNA regions that contain REs. However, due to the nature of chIP-chip technology, harvested DNA segments are much longer than the length of the RE. Motif identification algorithms are then used to identify consensus DNA segments corresponding to REs within the identified regions. Ideally, this analysis is then used to derive a consensus sequence for the binding site of a hormone receptor. The following flow chart (Figure 1-2) shows the infrastructure of the thesis, where the 2 major chapters are chIP-chip data processing and binding site identification and characterization.

Figure 1-2: Flowchart of experimental process

1.1. Motivation

Characterizing the mode of action of thyroid hormone (TH) through its involvement with TR and DNA will allow for a better understanding of the potential health effects associated with disruption of thyroid hormone levels. The technology used in this project requires the integration of many fields: biology, statistics and computer science. Each step of this work requires a good understanding of these specialized fields in order to identify the relevant literature and perform a valid evaluation of each computational model. As required, computational models were modified in order to solve the particular problem presented in this thesis, i.e. identification of thyroid response elements in genomic DNA.

1.2. Objectives

The main objective of this thesis is to develop a well-defined analysis framework to identify TREs from biological data using computational models. Application of this framework will deliver a list of binding sites ranked by probability of occurrence. This will allow biologists to prioritize candidate sites for follow-up wetlab verification experiments to confirm the results of the *in silico* framework.

1.3. Thesis Outline

This thesis is presented in 5 chapters; each bullet point below represents a chapter.

- Chapter 1: An introduction provides the reader with an overview of the project.
- Chapter 2: Provides an overview of the biological and computational basis underlying the research conducted in this thesis.
- Chapter 3: Reports on the analysis of the chIP-chip data including normalization, genomic mapping of probe locations, and peak identification.
- Chapter 4: Reports on the post-analysis of chIP-chip data where motif finding algorithms will be used to find a consensus binding sequence for THR.

Chapter 5: Summarizes the major contributions and conclusions arising from this research and provides ideas for potential future work.

Chapter 3 & 4 contain a review of relevant previous work, critical assessment of the available algorithms, results obtained from the chosen method and a discussion of the outcome of the results. In order to facilitate the reading of this thesis, the traditional structure of {method, results, discussion} will be provided for each method, rather than for each chapter.

Chapter 2. Biological and Technological Review

Since this project requires the integration of many fields, this chapter aims to inform the reader on the underlying concepts necessary to understand the remainder of the thesis. First, this chapter will cover the necessary biological concepts to comprehend the necessity and application of the computational models.

2.1. Biological Background

2.1.1. Thyroid hormones

As mentioned in the thesis introduction, hormones are key molecules governing gene transcription. The current project focuses on thyroid hormones (THs) which interact directly with the thyroid hormone receptor (see section below).

THs are a class of hormones secreted by the thyroid and composed mainly of triiodothyronine (T3) and thyroxine (T4). T3 and T4 are synthesized by the thyroid gland. Production of T3/T4 is stimulated by thyrotropin-releasing hormone (TRH) and thyroid stimulating hormone (TSH).

Figure 2-1: Illustration of T3 (left) and T4 (right) hormones

The production level of T4 in the thyroid gland is much greater than T3. However, the most potent form of thyroid hormone is T3. This hormone interacts directly with TRs and is directly responsible for gene regulation in many organs. The principal function of T4 is to act as a pro-hormone, which is converted into T3 by the action of deiodinase enzymes [32].

Mechanisms of action of TH are categorized in 2 groups: 'hormone receptor-specific nuclear' and 'extra nuclear' actions [2]. Since extra nuclear actions are not of interest to this study, focus is placed on the hormone receptor-specific group. Hormone receptor-specific refers to a hormone binding to its related hormone receptor in order to regulate gene transcription.

2.1.2. Thyroid hormone receptor

THR is the principal molecule that is studied in this project, i.e. the protein that will target specific DNA segments to regulate gene transcription. To refer to the analogy mentioned in the introduction, THR is the lock that regulates gene transcription.

THR is a type II member of a super family of nuclear receptors (NRs) [34]. NRs are located in the cell's nucleus, regardless of the presence or absence of a ligand. In the context of this thesis, the ligand is the thyroid hormone (i.e. the key that fits into the THR lock). Other members of this group include vitamin D receptors, retinoid acid receptors, and the retinoid X receptor just to name a few. Some of these other types of NRs can combine with a THR in a dimer or trimer configuration to control gene expression, which will be reviewed later.

THR is encoded by two genes, c-erbA-1 and c-erbA-2, which are expressed into THR α and THR β respectively. Each of these has alternative splice variants, i.e. 4 isoforms; THR α 1, THR α 2, THR β 1, THR β 2. Although THR has 4 isoforms, the wet laboratory procedures targeted THR β 1 and therefore any mention of THR below will refer to THR β 1 unless mentioned otherwise.

THRs have common functional domains, i.e. sub sequences of amino acid residues which have conserved roles across the isoforms. THRs contain two activation sites, i.e. domains that are responsible for gene transcription. Due to the nature of wet laboratory technology used, the specific mechanism of gene regulation is not a discriminatory factor for the discovery of binding sites. Below is an overview of the functional domains of THR and Figure 2-2 shows the 1D (i.e. sequence) and 3D (i.e. structure) representations of the protein.

N-terminal domain:

This domain contains a weak activation function (AF-1) for which the presence of a ligand is not required for THR to be active. However, this function is usually weak, and is rarely responsible for gene activation.

Structural Organization of Nuclear Receptors

Figure 2-2: Sequence (1D) and Structure (3D) of Nuclear Receptors

(http://en.wikipedia.org/wiki/Nuclear_receptor)

DNA binding domain (DBD):

This domain is of particular interest to this project. As its name implies, this domain is responsible for binding the protein to DNA. The DBD contains two zinc fingers that bind to a DNA sequence called a hormone response element (HRE). Since this project targets the thyroid hormone, here the DBD binds to a TRE. Zinc finger is a local protein structure where the protein chain wraps around a zinc ion, and allows the protein to bind to DNA. A portion of this domain will also interact with another NR to form a partnership, which will be explored in greater details in a following section.

Hinge region:

This domain is the pivoting point connecting the two most important domains: the DBD and LBD.

Ligand binding domain (LBD):

The structure formed by this domain contains a cavity whose function is to receive a ligand; in the case of THR, the ligand is TH. Depending on the presence or absence of the ligand, THR exhibits 2 different conformations (i.e. alternate structural configurations). An aporeceptor refers to a THR where the LBD is free of a ligand, whereas a holoreceptor is a THR where its activating hormone is contained in the LBD. The LBD contains another activation function (AF-2), which is responsible for triggering gene transcription.

2.1.3. THR partners for gene regulation

Gene regulation by THR doesn't necessarily arise due to one specific type of protein-DNA interaction. Rather, THR can bind to DNA in a number of configurations:

Monomer: A single molecule binding to DNA that will regulate transcription

Homodimer: 2 THR molecules bind together through the DBD and LBD. This complex then binds to DNA to regulate gene transcription.

Heterodimer: 1 THR molecule and 1 other nuclear receptor molecule, such as retinoid X receptor or retinoid acid receptor, bind together and regulate gene transcription.

Multimer: Several NRs will bind with THR creating a complex that binds to DNA and regulates gene transcription. The number of NRs can range from 3-5 according to Mengeling *et al.* [37]

This information demonstrates the ability of THR to form clusters. Section 2.1.4 will demonstrate the importance of the previous information as the entire binding sequence can be composed of many repeated short sequences.

2.1.4. Thyroid response elements

Thyroid response elements (TREs) are DNA sequences which are recognised by THRs and their partners. TREs are the site where THR binding occurs leading to regulation of gene expression. TREs are divided into two regulation classes, positive TREs and negative TREs, which respectively promote or inhibit gene transcription. Since the specific nucleotides forming a RE are not identical from one binding site to another, REs are often characterized by sequence motifs.

2.1.4.1. Known motifs

A consensus motif for TREs has been established in the literature for some time [13]. A hexamer, a DNA sequence containing 6 nucleotides, is the core element in the binding sequence for THR. This hexamer has the following canonical sequence: “AGGTCA”. However, this sequence can be interpreted as a somewhat degenerate consensus. Figure 2-3 provides a sequence logo of the TRE hexamer from sites collected in the literature (see Table 2) and illustrates the degeneracy of the hexamer. Information provided by the figure displays for each position of the hexamer the information content (see weight in Equation 5) for each nucleotide. The likeliness of occurrence is shown proportionally

with the size of the nucleotide. TREs can be found on either of the DNA strands, i.e. 5' – 3' or 3' – 5'.

Figure 2-3: Logo of TRE hexamer

(Image generated with

<http://www.bioconductor.org/packages/2.3/bioc/html/seqLogo.html>)

As mentioned in section 2.1.3, THR can form complexes to bind as a monomer, homodimer, heterodimer, homotrimer, heterotrimer, homotetramer, heterotetramer, homoquintamer, heteroquintamer. Each of these require a 1 to 5 hexamers in order to bind to DNA. Many other NR bind to the THR hexamer motif [34] and therefore other differentiation factors are involved. The spacing between hexamers is believed to be the major discriminating factor. For example, the dimer formed by the vitamin D receptor (VDR) and the retinoid X receptor (RXR) binds to 2 hexamers with a spacer of 3 nucleotides. THR dimers require a spacing of 4 nucleotides between the hexamers of dimers in some cases (see other cases in Table 2-2). However, one must consider the fact that the spacer preferences for TREs are to be seen as a degenerate consensus. From previously identified and confirmed TREs from multiple species, Laudet *et al.* [34] relates the predominant TREs identified to date to be dimers. In this case, each of the two hexamers is referred to as a 'half site'. Table 2-1 shows the diverse arrangements for TRE dimers.

Table 2-1: TRE arrangements for hetero and homo dimer TRE configurations

	Hexamer	spacer	Hexamer
Direct Repeats	AGGTCA →	NNNN	AGGTCA →
Inverted Repeats/ Palindrome	AGGTCA →	None	TGACCT ←
Everted Repeats/ Inverted Palindrome	TGACCT ←	NNNNNN	AGGTCA →

TRE trimers, tetramers and pentamers have no well defined spacer and therefore are difficult to assess. Evidence in the literature suggests that THR can bind as a homopentamer or heteropentamer with RXR (see Section 2.1.3).

2.1.4.1.1. Compiled motifs

A thorough review of the literature produced a list of known THR β TREs for mice. This review was made necessary since no exhaustive list in the literature for mouse was found. A list for other species was compiled by Williams *et al.* [59]. One must specify that the TREs published in literature and their position is highly dependent on the genome assembly in which they have been identified. Genome builds, i.e. their sequences and annotations evolve constantly which makes some TRE references obsolete on current builds.

Table 2-2: List of TREs in mouse genome compiled from the literature. the gene that is regulated by the TRE, its accession number, its DNA strand(GS), the location with respect to the transcription start site of the gene, the strand of the TRE (TS), the sequence which contains the TRE with binding site in bold, the type of gene regulation that the TRE performs (up or down regulates gene transcription), the literature reference for the TRE and the TRE configuration are shown.

Gene	Accession #	GS	Location	TS	Sequence	Type	Ref.	TRE config.
Nr4a1	NM_010444	+	-1182 to- 1218	+	actggg atggagatgtgacctgcagggtga	Neg	[15]	Trimer
c-myc	NM_019660	+	Exon1	+	cgacctaagaagcagctct	Neg	[45]	Dimer
SREBP-1c	NM_011480	-		+	gcctgacaggtgaaatcgcc	Neg	[27]	Monomer
KLF9	NM_010638	+	-3830 to-3804	+	agg tgaagtgaagtca	Pos	[12]	DR4
UCP3	NM_009464	+	-59 to -30	+	tcagaatagggttcaggcagctgggtga	Pos.	[51]	DR1
Myogenin	NM_031189	+	-526 to-494	+	gtgtaggtgttttaggggtctca	Pos.	[16]	DR4
Nas1	NM_019481	-	-436 to-425	-	agg ctatagccc	Pos	[11]	IR0
ChREBP	AB074517	+	-2436 to-2389	unknown	cggg tactagagggga	Pos.	[25]	DR4
ChREBP	AB074517	+	-2436 to-2389	unknown	agg caatgagaggtga	Pos.	[25]	DR4
ABCD2	NM_011994	-	-401	-	tgg cctgattgacct	Pos.	[58]	DR4
CYP7A1	NM_007824	-	-3KB	-	agg gca	Pos.	[50]	Monomer
CYP7A1	NM_007824	-	-3KB	-	agg tcagggtca	Pos.	[50]	DR0
Fgfr1	NM_010206	+	-279 to-264	+	ttg ccccattcaacct	Pos.	[39]	DR4
MBP	NM_001025245	+	-184 to-167	+	ac cteggctgaggaca	Pos.	[17]	DR4
TRH	NM_009426	-	-57 to-51	-	tg acct	Neg	[48]	Monomer
(RAT) MBP	NM_001025289	+	-186 to-163	+	ag acctggctgaggaca	Pos	[18]	ER6

2.1.4.1.2. Differences between species

In addition to the literature describing known TREs in the mouse genome, it is also possible to search for TREs in the genomes of other closely related species. Orthologous genes are genes that have been conserved in biological evolution through time across species. Although high TRE sequence conservation can increase the confidence in a putative site, the absence of strong sequence conservation does not necessarily imply the lack of function for this sequence. This may serve as a criterion to establish priority when biologists perform wet-laboratory experiments to validate the novel TREs.

2.1.4.2. Atypical and novel motifs

Evidence in the literature suggests that THR and other NR partners does not bind directly to DNA, but regulate gene expression indirectly by binding to other proteins that are in turn bound to DNA. As shown in Figure 2-4, Lazar [35] suggests that THR can regulate gene expression through association with the Jun and Fos proteins bound to DNA at an AP-1 binding site. The AP-1 binding site is characterized by the sequence logo shown in Figure 2-5.

Figure 2-4: Mechanism of T3 regulated gene (activation and repression) with FOS and JUN nuclear receptor interaction

Figure 2-5: AP-1 binding site sequence logo

(<http://jaspar.cgb.ki.se/>)

Furthermore, Hashimoto *et al.* [26] demonstrated that the presence of T3, THR and RXR alone may be insufficient for the mediation of gene transcription for specific transcripts in vitro. In this example, these 3 players required intermediate binding in order to regulate liver X receptor alpha confirming the requirement of other proteins for gene regulation of some transcripts.

2.1.5. Summary

The 3 principal factors involved in TH regulation of gene expression are listed below.

- **A hormone:** TH is the hormone that binds to the ligand binding domain of THR and regulates gene transcription
- **A hormone receptor:** THR is the nuclear receptor responsible for the initiation of gene transcription. In most cases, it binds with 1 to 4 partners (e.g.RXR, RAR) in cases found in the literature to date.
- **A binding site:** This thesis categorises TREs in 2 classes;
 - Known hexamer: the “AGGTCA” hexamer is well established in the literature for being a binding site for THR and its partners. One discriminating factor for THR and other nuclear receptors is the spacing

between the hexamers. Although the spacing is a factor, it is not always a criterion that is respected (see Table 2-2).

- Atypical and novel motifs: Some evidence in the literature provides evidence that THR may not bind directly to DNA, although it is a main player in TH regulated genes.

2.2. Biological laboratory technology overview and experimental design

Now that the principal biological concepts have been reviewed, the wet laboratory techniques used to gather the raw data for this project are briefly reviewed below. Chromatin immunoprecipitation on microarray (chIP-chip) is the biological laboratory technology that is used in the present study to detect the presence and location of TREs in the mouse genome. This technique consists of two experiments that are performed consecutively. The first experiment (chIP) stands for chromatin immunoprecipitation. A chIP experiment has 4 steps (described in Figure 2-6) and ultimately produces a solution containing an enriched population of DNA sequences containing the thyroid response elements. The second “chip” refers to the tiled microarray used to measure the signal harvested during chIP.

2.2.1. Chromatin immunoprecipitation

Here are the 4 experimental steps performed in a chIP experiment;

Figure 2-6: Each step of chromatin immunoprecipitation
(<http://www.bio.brandeis.edu/haberlab/jehsite/chIP.html>)

Step 1) Crosslinking the TRE-THR complex

Crosslinking means that covalent bonds are formed between the TREs and the THR complexes. Covalent bonds are strong chemical bonds initiated here by formaldehyde. Formaldehyde has the ability to interact with two amino acid groups to effectively hold the TREs and THR complexes together.

Step 2) Sonication of the DNA

Sonication is used to fragment the DNA into the desired length. As shown in the second diagram of Figure 2-6, sonication breaks the DNA containing the TREs into short local sequences that are associated with the THR complexes. The desired length of DNA depends on the protocol used and is generally from 350 to 1200 bp.

Step 3) Complexing proteins with an antibody

An antibody that specifically targets THR is added to the solution containing the DNA, TREs and THR complexes. The antibody-THR-TRE complexes are then isolated from the remainder of the DNA.

Step 4) Reversing the crosslinking

The last step consists of reversing the crosslinking and digesting the proteins (i.e. removing the antibodies and THR complexes). Afterwards, a technique called polymerase chain reaction (PCR) replicates the DNA segments containing the TREs to produce enough DNA for analysis.

Following enrichment of DNA containing the TREs which had been bound to the THR complexes, the chip products are analyzed using DNA microarrays.

2.2.2. Microarrays

A microarray is a glass slide with a rectangular shape of approximately 2.5 x 7.5 cm. Single stranded DNA segments, referred to as oligonucleotides or probes are fixed at high density in grids on the surface of the microscope slide. The probes are designed from DNA sequence elements within the promoter region of genes. Hybridization of the chip

fragments to the complementary DNA fragments (i.e. probes) located on the microarray is used to determine whether DNA from this location is enriched in the chIP. In the present work, a two color technology was used to perform the experiment using the two fluorescent dyes below;

- Cyanine3 (Cy3) (green dye, total input channel);

Cy3 was used to label reference DNA used for non-discriminated DNA fragments. DNA gathered prior to chIP is sonicated in order to reduce its size and labelled with a fluorescent dye. The labelled DNA binds to complimentary probes. The relative amount of target DNA is visualized using a laser to excite the Cy3 dye at 550nm and detected using an emission wavelength of 532nm. Thus all DNA fragments are included in this total input (TI) sample, not just those fragments found to be bound to THR, and they are used as a reference background.

- Cyanine5 (Cy5) (red dye, IP channel);

The Cy5 channel is used with the DNA fragments from the chIP analysis step. DNA fragments (suspected TREs) collected from chIP are labelled with Cy5 which is excited and emits at a different wavelength, 550nm and 635 nm respectively.

The known composition of the probe DNA fragments located on the microarray allows one to determine the DNA sequence that was bound to the probes. As per the example below, one can observe red, green and yellow spots during a two-channel microarray experiment. Green and red spots show probes where only TI or IP show signal respectively. Yellow spots are apparent when there is a mix of Cy3 and Cy5 signal (approximately equal signal for IP and TI samples). The signal intensity measured by the detector is linearly correlated with number of DNA fragments present in a well. A scanner digitizes the surface of the microarray at the respective emission wavelengths of Cy3 and Cy5 and saves these data as an image file.

Figure 2-7: Example of a microarray with a zoom-in on some wells

(http://en.wikipedia.org/wiki/DNA_microarray)

2.2.3. Biological experimental design

This study makes use of two datasets. The first dataset was used to calibrate the methods and select preprocessing algorithms using a spike-in dataset from Johnson *et al.* [29]. The second dataset was produced at Health Canada for the specific goal of identifying novel TREs in the mouse genome. The experiment conducted by Dr. Hongyang Dong was performed on mice. Figure 2-9 illustrates the study design where 5 male mice were sacrificed on post natal day (PND) 4 and on PND 15. Cerebellum was collected during necropsy and its DNA extracted. Cerebellum tissue was selected since it is highly responsive to thyroid hormone at PND 4 and PND15. Dr. Dong selected and designed a microarray mapping ~5000 genes that were suspected to contain TREs according to the literature. For more details about the experimental design, please refer to Dong *et al.* [15].

The Agilent gene promoter 2 color microarray containing 44 000 probes was selected by Dr. Dong. Regions of the genome are tiled from -8000 base pairs (bp) upstream to +2000 bp downstream relative to each gene transcription start site. Probes have a length of 60 nucleotides spaced by an average of 153 bp along the region covered. Figure 2-8 shows the distribution of inter-probe spacing within the “A” and “B” arrays which respectively map chromosome 1 to 9 and 9 to X. As can be seen from the distribution, the probe spacing is not consistent and shows considerable variation. This is due to multiple

reasons such as repeat masking, where mapping is discontinued due to simple short repeating subsequences.

Figure 2-8: Distribution of the resolution for the "A" (black line) and "B" (blue line) microarray
(please note that blue and black line overlaps greatly)

Figure 2-9: Experimental Design of Biological Experiment

(mouse picture: http://en.wikipedia.org/wiki/File:Apodemus_sylvaticus_bosmuis.jpg)

2.3. Summary

Section 2.2 reviewed the laboratory technologies used to conduct the biological experiments that produced the data analyzed in this thesis. The first experiment, ChIP, aims at isolating DNA segments that contain TREs. The second experiment, tiled microarray (chip), measures the difference in DNA sequence concentrations between the total input (TI) and ChIP (IP) samples. Taken together, these experiments aim to identify the locations where TREs are located in the mouse genome.

The following chapters will describe and present the methods by which the collected chIP-chip laboratory data were analysed.

Chapter 3. ChIP-chip Data Analysis

Chapter 2 provides the biological background and description of the experimental procedures used to collect the raw data. The following chapter describes the methods chosen by the author of this thesis to perform chIP-chip analysis. The choice of methodology at each stage of data processing and analysis is supported by the literature and *in silico* experimental validation. This chapter focuses on normalization and peak finding algorithms for chIP-chip data. In the workflow of chIP-chip analysis, the data must be first pre-processed to remove microarrays of poor quality. The raw spot intensity data must then be normalized to correct for various biases (see section 3.2), such as dye preference or probe content. Peak identification algorithms then seek to identify genomic regions that are believed to contain TREs (see section 3.2.2). These genomic regions are then analyzed to extract specific TRE locations and consensus motifs. This latter topic is covered in Chapter 4.

Following microarray hybridization, scanner-specific software is used to derive an image from the microarray and obtain a signal intensity for each spot, i.e. quantitation. Various measurements are taken for each dye (Cy3, Cy5) such as mean and median spot intensities (i.e. the mean and median intensities of all pixels corresponding to each microarray spot), spot morphology and local background signal intensity. In the present study, the median signal intensity was used as the primary source of data for probe fluorescence since the mean is more susceptible to outliers than the median.

Following image acquisition, quality assessment (QA) of microarrays is the first step in chIP-chip analysis. QA aims to eliminate microarrays that are not hybridized optimally (e.g., low signal to noise ratio, high background, low overall signal intensity relative to the other arrays, etc.) or contain overwhelming artefacts that cannot be corrected by subsequent processing methods. Data distortion can be caused by sub-optimal hybridization protocols or technical errors, or by physical problems with the arrays such as scratches, fingerprints, and other marks on the glass slide. The pre-processing and

filtering of the data to eliminate these poor quality spots from the analysis was performed using the scanner software and is not a major contribution in this work.

Figure 3-1 illustrates raw chIP-chip data from chromosome 15, base pairs 74504474-74507608. A positively identified region containing a TRE is circled in red. The ordinate axis is a $\log_2$ ratio that allows for quick visualization of fold changes relative to the background. The $\log_2$ ratio transform is calculated using following formula:

$$M = \text{Log}_2 R - \text{Log}_2 G$$

Equation 1: Log transform of chip signal (R: red channel intensity, G: green channel intensity)

Figure 3-1: Example of raw chIP-chip data and confirmed TRE expressed by probes circled in red. Lines indicate the probe location (x-axis) and the intensity of the Red (ChIP DNA) to the Green (total input DNA).

3.1. Mapping probes to genome

The first step prior to analyzing chIP-chip data is to map each probe to a specific location of the genome. This process needs to be performed prior to peak finding, since probe mapping has a great influence on the results of the peak finding algorithms. This is because adjacent probes are expected to have correlated intensities. The annotation of the genome varies and evolves from one build to another as new sequencing and mapping data are released in the public domain. As such, it is important to update the actual probe locations using the newest build prior to analysis. Nucleotide sequences covering regions between probes are gathered from genome assembly (readily available from genome browsers such as the UCSC genome browser [46], the Ensembl genome browser [19] or the NCBI genome browser [3]). The UCSC genome browser provides a convenient tool to convert coordinates in batch mode. The design of the array used in this study contained build numbers Mus Musculus (MM) 5 and MM6. These were converted to the most recent build, MM9, using the *liftOver* [6] tool.

3.2. Normalization of ChIP-chip data

Once the raw data have been collected and the initial quality assurance screens have been applied, the data must be normalized. The main goal of normalization is to remove experimental artefacts that do not represent biological phenomena. An example artefact for microarrays is elevated signal intensities for probes that have a high sequence content in Gs and Cs. The additional hydrogen bond between pairs of Gs and Cs can explain this bias as the microarray washing protocol will remove fewer probes having a high G and C content which, in turn, will artificially increase the signal of high G-C content probes. Some chIP-chip normalization methods are designed specifically to address this phenomenon, as will be discussed later in this chapter. Another objective of normalization is to remove variability between microarray replicates (for e.g., resulting from overall DNA labelling efficiency, or dye batches). Figure 3-2 (a subset of Figure 1-2) shows the precise step in the workflow that this section addresses.

Figure 3-2: Partial flowchart of thesis with normalization algorithms highlighted

The most widely used application for DNA microarrays is gene expression analysis [49]. It follows that the majority of microarray data normalization methods were developed for gene expression microarray data as opposed to chIP-chip data. Although very similar hardware is used for both gene expression arrays and chIP-chip arrays, the resulting data are significantly different in a number of ways. Therefore, many assumptions underlying gene expression normalization methods are violated when these methods are applied to chIP-chip data. In the subsequent section a number of gene expression and chIP array normalization methods are presented. These methods are evaluated for their suitability for the present chIP-chip data set. In addition, the available literature was thoroughly reviewed to find the most appropriate method for the analysis of chIP-chip data.

3.2.1. Normalization methods developed for gene expression microarrays

Several methods that have been developed for gene expression array normalization are described below. Although these methods have been recommended by developers of chIP-chip data analysis methods [42] [31], the findings reveal that these methods are not appropriate for chIP-chip normalization.

3.2.1.1. Quantile normalization

The quantile normalization method aims to harmonize the signal distribution of each array to a pooled distribution. Many laboratory and biological variations can affect the distribution of the signal for a dye, e.g. dye labelling differences. Recall that in the present study, five mice were analyzed leading to five replicate microarrays. The underlying assumption made when using this normalization method is that microarray

replicates should have the same signal distribution. Therefore, only batches of arrays that are assumed to have the same distribution should be normalized using this method. The outline of the algorithm is shown in Algorithm 1. Spot intensities are listed in columns, with one microarray per column, resulting in a matrix, X . This matrix X has dimension $p \times N$ (number of spots on one microarray) rows by (number of microarrays to be normalized in the same batch) columns. Each column is then sorted to produce X_{sort} . Vector projection is applied to each row of X_{sort} onto d where d is a vector of length N having all its values set to $1/\sqrt{N}$. This is performed to scale all replicate arrays to a pooled distribution. This allows harmonising the distribution of all arrays, therefore making each replicate data set more comparable to the others by scaling up or down individual distributions.

As stated above, quantile normalization assumes a similar probe signal distribution across all arrays for which the normalization is applied. Thus, special care should be taken when choosing the microarrays that will be normalized together. For example, in an exposure study, arrays referring to the exposed animals shouldn't be normalized in the same batch as the control animals since they are not expected to have the same distribution, i.e. their global gene expression is expected to be different than the control animals.

1. Given N microarray datasets of length p , form matrix X of dimension $p \times N$ where each dataset is a column
2. Set $d = (1/(N^{1/2}), \dots, 1/(N^{1/2}))_N$
3. Sort each column X to give X_{sort}
4. Project each row of X_{sort} onto d to get X'_{sort}
5. Get X_{norm} by rearranging each column of X'_{sort} to have the same ordering as original X

Algorithm 1: Quantile normalization algorithm

Unlike gene expression microarray data, the signals of greatest interest for chIP-chip are precisely located on the right hand extreme of the distribution. For the present dataset, quantile normalization was immediately eliminated as a potential candidate normalization method since this method interferes tremendously with signals of high intensity on the

edges (i.e. tails) of the distribution function. Problems associated with this approach were raised by Bolstad [5] who proposed to center and scale the probes on the edges of the distribution. Also, quantile normalization doesn't correct for any biases like dye or nucleotide effects (explained below). The concerns raised by Bolstad were reiterated by Peng *et al.* [42]. These authors found that the method was too stringent on probes at the edges of the distribution. This was due to the small number of probes located in the right tail of the spot intensity distribution.

3.2.1.2. Lowess normalization [61]

In order to discuss Lowess normalization, one must first introduce ration intensity plots often referred to as MA plots (see Algorithm 2 for construction details). An MA plot graphs the signal intensity of each microarray spot on a slide; the y axis is the log ratio of signal over background and the x axis is the average log intensity of both channels. For Lowess normalization, it is assumed that the center of mass along the abscissa is 0 (i.e., the majority of probes show equal (Cy3,Cy5) signal). Therefore, after the MA plot is built, Lowess curve fitting is performed in order for the center of mass of M to be translated to 0, along the entire A axis. Rather than fitting a parametric model through all points, Lowess normalization performs nonparametric curve fitting at each point, based on a weighted local window of points. Then, each spot signal is shifted by its corresponding regression value. Depending on the implementation of the Lowess fitting algorithm, the degree of the polynomial fitting (see Algorithm 2) and the fraction of data used for each point to calculate the weighted least square can be specified by the user.

1. Construct the MA plot such that

$$A = \frac{1}{2} \log_2(\text{signal} \times \text{background})$$

$$M = \log_2\left(\frac{\text{signal}}{\text{background}}\right)$$

and a probe p has a coordinate (A_p, M_p) . Here signal is IP and background is TI.

2. Perform polynomial curve fitting using the weighted least square criterion.

3. For each point in the graph, normalize the value:

$$M_{\text{norm}} = \log_2\left(\frac{\text{signal}}{\text{background}}\right) - c$$

where c = the ordinate value of the lowess fitting for point p

Algorithm 2: Lowess algorithm

Figure 3-3 illustrates the Lowess normalization curve. Subplot a) clearly shows that the centre of mass does not fall on the A axis for larger A values. Subplot b) illustrates the results of the lowess curve fitting. Subplot c) shows the final normalized data where the centre of mass has been corrected to fall on the A axis.

Figure 3-3: Lowess normalization performed on a microarray dataset. a) Data before normalization; b) Lowess curve fitting; c) Normalized data

(http://www.systemsbiology.nl/datagen/transcriptomics/variation/intens_norm.html)

Lowess normalization is not suited for chIP-chip data since none of the probes on a chIP microarray are expected to show down regulation. Using gene expression arrays, it is expected to see both up and down-regulation of probes. In contrast, the signal measured by chIP-chip is an enrichment of IP over TI. Therefore, it is expected for the signal to be positive (i.e. $IP > TI$) for a subset of spots, and zero (i.e. no enrichment) for the majority of spots. Figure 3-4 shows a MA plot representing the combined replicates (median value) for PND15B. Thus, performing Lowess subtraction would violate this assumption since the majority of ratios are greater or equal to 0. As such, Lowess normalization is not suited to our data.

Figure 3-4: Lowess regression performed on the combined replicates (median value) of PND15

3.2.1.3. Summary

Although the hardware for chIP-chip and gene expression experiments is very similar, the resulting data are fundamentally different. Therefore, the blind application of normalization methods developed for gene expression microarray data analysis to chIP-

chip data would violate underlying assumptions made by these normalization methods. Therefore, this leads to a conclusion that normalization methods intended for gene expression studies are not suited for chIP-chip data. This point was illustrated above for two popular gene expression normalization methods.

3.2.2. ChIP-chip normalization methods

This section reviews two methods designed specifically for normalization of chIP microarray experiments. Verifications are made in order to determine which biases are addressed by each method, and assess whether each algorithm can correct for artefacts expected to be present in our chIP-chip data.

3.2.2.1. GC probe based method

Since the probes on a DNA microarray tile the genome, the sequence composition (i.e. the proportion of C, G, A, and T bases) of each probe can vary dramatically. One bias that can be observed in chIP-chip spot intensity data is due to the strength of the bond between base pairs, where a G-C bond has three hydrogen bonds while an A-T bond only has two. The GC probe based method of normalization uses the composition of each probe to perform spot intensity normalization. Song *et al.* [52] analyzed data produced using the Nimblegen platform (Nimblegen Corporation, New York) chIP arrays and found that probe intensity variance was highly correlated to the number of Gs and Cs contained in each probe. GC probe normalization utilizes the number of occurrences of Cs and Gs as a normalizing factor in order to correct for the poor binding affinity of low GC content probes. As a quick description of the algorithm, probes having the identical number of {G,C} are placed in bins where their spot intensities are characterized by the mean, standard deviation and covariance. Equation 2 describes this normalization technique where k represents the bin index, j the channel index (1=TI, 2=IP), and x_{ij} the signal intensity. The final t_i values, where i is the spot reference, provide normalized intensity values that can later be used for peak discovery.

$$t_i = \frac{(x_{i2} - x_{i1}) - (\hat{\mu}_{2k} - \hat{\mu}_{1k})}{\sqrt{\hat{\sigma}_{1k}^2 + \hat{\sigma}_{2k}^2 - 2\hat{\xi}_k}}$$

$$\hat{\mu}_{jk} = \sum_{\{i|GC_i=k\}} \frac{x_{ij}}{n_k}$$

$$\hat{\sigma}_{jk} = \sum_{\{i|GC_i=k\}} \frac{(x_{ij} - \hat{\mu}_{jk})^2}{n_k}$$

$$\hat{\xi}_k = \sum_{\{i|GC_i=k\}} \frac{(x_{i1} - \hat{\mu}_{1k})(x_{i2} - \hat{\mu}_{2k})}{n_k}$$

Equation 2: GC content based normalization and parameter estimation

An investigation was performed in order to determine whether the low GC probe content bias was observed for the Dr. Dong dataset (2.2.3), since the Agilent platform is used instead of Nimblegen. Both platforms share several characteristics such as probe length and the use of 2 color (i.e. channel) technology. Figure 3-5 shows a histogram of one of Dr. Dong's TI array signals to compare it to Figure 3a from Peng *et al.* Histograms of signal intensities for low GC-content probes (less than 20 {G,C} bases in the probe of length 60) and all probes were plotted separately. If there were a significant GC-content bias, a significant difference in mean intensity between the two histograms would be expected.

Figure 3-5: Histogram of distribution of TI probes, figure to be compared to Fig. 3a of Peng *et al.*

The results of the Agilent microarray do not show a strong correlation between signal intensity and probe GC content. The results of Peng *et al.* clearly show a bimodal distribution and a difference of close to 4 fold between the means. Figure 3-5 shows significantly less bias toward probes that have less than 20 GC compared to the rest of the probes where an approximate difference in sample means of 0.2 can be observed. Although a t-test reveals a difference in means for TI signal (all probes, probes with less than 20 GC), I argue that this method has little impact and limits the normalization to only the GC content factor. The next normalization method presented in section 3.2.2.2 is a multi component method that addresses this bias.

This method was not applied since the correction of the bias would have been negligible. One possible explanation for the probe distribution observed by Peng *et al.* may lie in the specific laboratory protocols used in the microarray experiments. The microarray washing protocol post-hybridization is crucial to produce an evenly distributed TI signal across the array. If washes were not stringent enough, the non-specific probes may more readily be retained and contribute to signal intensity. As such, low GC content probes may produce an artifactually high signal intensity. However, this explanation is only a hypothesis that could be verified by wet laboratory experiments and further statistical analyses.

3.2.2.2. Content and nucleotide location method (CNLN)

This method was initially developed for the Affymetrix platform (1 colour array) by Johnson *et al* [30]. Their model, MAT, is composed of 81 parameters to be computed, such as the respective number of {A, C, G, T} bases in a probe, the number of occurrences of the probe sequence in the genome studied, etc. Their method is integrated in a linear model to define and characterize each parameter. This allows for the assessment of a baseline for each probe, which is later subtracted from the raw probe intensity value. Potter *et al.* [44] provided a slight modification to the above mode. These authors reduced the complexity of MAT by eliminating a number of parameters that failed to show statistically significant correlation with probe intensity. One must mention

that the experimental data Potter *et al* used was from DNA methylation arrays. Due to the similar laboratory protocol and nature of the data (i.e. only positive signals are observed), methods for DNA methylation arrays and chIP-chip are interchangeable.

Potter *et al.* used a quadratic model to correct for a number of experiment-induced biases including probe composition (as above), position of nucleotides within the probe, and also dye effect. A linear model calculates a linear regression with parabolic parameters in order to minimize square error. Equation 3 below shows a modified version of the normalization equation presented by Potter *et al.* with its respective parameters described below.

$$P_d = \alpha_0 + \boxed{\text{A}} + \boxed{\text{B}} + \delta I(d = \text{Green})$$

$$P_d = \alpha_0 + \sum_{j \in \{A,C,G,T\}} (\beta_{j0} n'_j + \beta_{j1} n'_{2j}) + \sum_{j \in \{A,C,G\}} (\gamma_{j0} S'_j + \gamma_{j1} S'_{2j}) + \delta I(d = \text{Green})$$

Equation 3: Quadratic model for chIP-chip normalization derived from Potter *et al.*

The parameters to Equation 3 are defined below:

- P_d = is the expected baseline log transformed probe value for $d = \{\text{Red}, \text{Green}\}$
- l = is the number of nucleotides in the probes, i.e. its length.
- α_0 is the mean baseline signal across the array.
- β_j is the coefficient for the contribution of base j .
- n'_j is the abundance of nucleotide j in the probe divided by l .
- n'_{2j} is the abundance of nucleotide j in the probe squared divided by l .
- S'_j is sum of the position of base j within the sequence of the probe divided by l .
- S'_{2j} is the sum of the square of the position of base j with the sequence of the probe divided by l .
- I is an indicator binary function
- δ is the global dye effect

A linear regression model was performed on these factors in order to normalize the data to account for variation due to these effects. Part **A** of the equation normalizes the signal with respect to its nucleotide content (number of A, C, G and T). One intentional omission in the original Equation 3 presented by Potter *et al.* is the effect of the cut-site frequency. DNA is digested (cut) with a restriction enzyme prior to analysis for DNA methylation arrays (as in Peng et al.). However, the present ChIP-chip dataset did not pre-cut the DNA and thus this portion of the equation was removed.

The following paragraphs examines whether any normalization (i.e. β and γ correction) is actually required for each bias that Equation 3 is able to account for.

Quadratic Nucleotide Position Effect

Prior to the application of this normalization method, plots (Figure 3-6 & Figure 3-7) were generated to determine whether the quadratic nucleotide effect of nucleotide positions (shown as Part **B** in Equation 3) within probes was observed in our data. Figure 3-6 shows the median $\log_2$ intensities for one typical microarray of the TI probes for each nucleotide vs. the position of this nucleotide on the probe. Quadratic regression has been performed on the plot which clearly indicates the fitness of the quadratic model. Therefore the quadratic effect term is justified for use in this study.

Figure 3-6: Nucleotide effect along the probe position for the TI channel where a quadratic effect is shown regarding the base (A,G,C) position in the probe.

Quadratic Nucleotide Count Effect

One addition to Potter's model, is the part A in Equation 3 above. This term was added to account for the effect of the probe nucleotide count squared at each position in the probe. This reflects the reality that 'stickier' C and G bases will impart a greater bias to the spot intensity. This increase in intensity is not linear but quadratic, as shown in Figure 3-7. Another relationship demonstrated by Potter *et al.* is the $\log_2$ signal intensity vs. the number of nucleotides {A, G, C, T} shown in Figure 3-7 and Figure 3-8. Thymidine shows little correlation with respect to the parabolic regression and was left out for this reason. This is the same effect seen by Peng *et al.* Plotting the entire probe signal with respect to the content helps to visualize the effect. In order to determine the significance of each coefficient parameter in Equation 3, a t-test is performed in order to evaluate and reject the null hypothesis of having independent coefficients. Probabilities of rejecting the null hypothesis are given from the *summary(lm())* of the R programming language [53].

Figure 3-7: Median Log2 intensity of TI probes vs. the number of individual nucleotides in each probe

Summary of Changes to Potter Model

The model proposed by Potter *et al.* was modified in order to take into account the effects seen in this experiment.

- The quadratic term **B** in Equation 3 (S' and S_2' in Potter *et al.*'s model) of the model, which corrects for the effect of the nucleotide {A, C, G} location along the probe, was applied to the first 45 nucleotides in the probe. The remainder has high variability and therefore is not well fitted by a regression as seen in Figure 3-6. T was not accounted for since the effect was not significant in previously assayed models as can be visualized in Figure 3-6.
- Potter *et al.* proposed using a simple linear fit for the effect of nucleotide {A, C, G, T} probe content. The effect of nucleotide content clearly follows a quadratic model as can be seen in Figure 3-7. This effect was therefore modeled as a quadratic effect by adding the β_{j1} term. This renders Song's method of correcting for probe GC content (described above in section 3.2.2.1) unnecessary since the minor differences in signal

intensities between low and high GC content probes are accounted for and corrected.

- No correlation was found between the T count in the probe and the α_0 term introduced by Potter *et al.* Therefore, the coefficient of α_0 in Potter *et al.*'s equation was dropped and now constitutes the intercept of the model.
- The effect of the cut-site frequency was intentionally omitted to adapt the model from methylation arrays to ChIP arrays.

Assessing the Quality of Normalized Signal

Unless extensive spike-in studies are performed, one has to rely on model quality assessments performed in the literature to determine the potential performance and appropriateness of the normalization algorithms for a given experiment. Spike-in experiments are conducted by adding a known concentration of DNA sequences in a calibration study. Therefore, a recovery assessment of known spike-ins can be conducted in order to evaluate the performance of the normalization technique. In this study, no such calibration experiment was performed.

The effectiveness of the normalization method can be assessed by plotting the factors that are supposed to be corrected by the normalization method against non-normalized data in order to evaluate the correction. Figure 3-8 & Figure 3-9 show plots of normalized probe values vs. the factor corrected. A normalization method will eliminate or at least reduce any variation between a factor (e.g. number of GCs) and the signal. Figure 3-8 shows the residuals for the GC content correction, where a decrease of correlation can be observed from Figure 3-7 to Figure 3-8. Although Figure 3-8 seems to show residual effects, the reader should take into account the range of the Y axis scale that is much narrower than previously shown in Figure 3-7. Figure 3-9 shows the residuals of the correction for the location nucleotides {A, C, G, T} which also decreases the correlation between the signal and the nucleotide position.

Figure 3-8: Median corrected probe intensity vs. number of each nucleotide in the probe. The intense parabolic effects was been attenuated tremendously showing no correlation with respect the number of individual nucleotides in the probes (part A of Equation 3).

Figure 3-9: Median probe value corrected vs. each nucleotide position in each probe. This figure shows the residuals of the nucleotide position effect in the probe. The narrow Y range indicates that the correction for this factor was useful (part B of Equation 3).

3.2.3. Summary

A modified version of the normalization algorithm created by Potter *et al.* was developed. The modified version takes into account the content and nucleotide location portions of the algorithm since it is able to correct for multiple factors in one model. The statistical significance of each parameter estimated in the model shown in Equation 3 was also assessed. Finally, the quality of the normalization was demonstrated for both probe content correction and nucleotide position correction in Figure 3-8 & Figure 3-9.

Results from applying the selected normalization method to data from a spike-in study will be provided in Section 3.4. Even though the normalization corrects biases, it will be possible to evaluate if a correction of the bias has a high impact on our ultimate goal: the recovery of the enriched peaks.

3.3. Peak finding algorithm

The point of a chIP-chip experiment is to determine what regions of DNA are associated with a transcription factor of interest. Thus, following normalization, identification of enrichment (peak finding) is carried out. In other words, which DNA genomic regions show a significant enrichment in the immunoprecipitate (IP) channel compared to the total input (TI) channel. Figure 3-10 shows the precise step in the workflow for the analysis of chIP-data for this process.

Figure 3-10: Partial flowchart of thesis with peak finding algorithm highlighted

A peak is defined as a set of probes located within a defined genomic region that shows high signal ratio intensity (i.e. elevated IP/TI intensity ratios). The definition of a peak changes slightly depending on the algorithm used to find peaks; a precise definition will be given when describing each algorithm. In order to evaluate the number of probes in a peak, the formula developed by Keles *et al.* [31] was used as shown in Equation 4.

$$P_{pb} = \left(\frac{s_{DNA} + p_l}{p_l + D} \right)$$

Equation 4: Equation for the expected probes per peak

The following parameters are used in the equation above;

- P_{pb} = number of probes per binding site
- s_{DNA} = size of DNA fragments following shredding by sonication , in bp
- p_l = probe length, in bp
- D = average distance, in bp, between the end of one probe, and start of the next probe

Inserting parameters $\{s_{DNA} = 600, p_l = 60, D = 153\}$ from our study into Equation 4 results in $P_{pb}=3.09$. In other words, 3.09 probes are expected to be in a peak. However, Keles *et al.* [31] performed an analysis on data published by Cawley *et al.* [8] which revealed that the number of probes expected to show enrichment is much higher in most cases than the number of probes actually observed to be enriched. That is, for each actual transcription factor binding site, only a subset of those peaks expected to show a significantly elevated signal IP/TI ratio actually does.

3.3.1. Splitter [20]

As discussed in Section 3.4 below, Johnson *et al.* [29] conducted a very large scale spike-in study in order to evaluate the performance of multiple criteria for chIP-chip experiments. These authors evaluated multiple normalization and peak finding algorithms using data from a known mixture. More details of this study will be discussed in section 3.4. Based on their recommendations for the Agilent platform, the Splitter tool was selected for evaluation.

Splitter uses the distribution of the signal intensities as input in order to select a cut-off to discriminate true binding signals from noise signals. Here, noise signals are defined as probes that show elevated $\log_2$ intensity ratios that are not caused by a THR binding phenomenon. First, a histogram of probe frequency vs. chIP signal is produced in order to visualize the signal distribution. Second, a range of true binding signal must be selected by the user. The algorithm is then applied in either the manual (a) or dynamic mode (b) described below.

a. Manual Mode: description.

- i. The number of standard deviation (SD) from the center (mean or median) of the distribution.
- ii. A percentile value.
- iii. A user defined value.

The cut-off value in (a) selects probes which are of high interest because they are in the right tail of the distribution of signal intensities. Moreover, these probes satisfy two other criteria known as GAPMAX AND MINRUN, which are defined below. Figure 3-11 shows example values for the manual cut-off selection to select peaks of true binding signal.

Figure 3-11: Shows the manual cut-off options (from (a) above) for the combined replicates (median value) for PND15B.

b. Dynamic mode: Description

A dynamically defined value for finding peaks based on user defined parameters: from, to, splits and break (described below).

Applying the dynamically defined value for cut-off uses the following algorithm (Algorithm 3) to select $\log_2$ values considered as true binding probes. Once “from” and “to” values are selected by the user on the histogram, the user identifies the number of splits for which peaks will be identified. Splitter moves from one split section to the other, if the number of peaks does not increase by the “break” value, then the algorithm stops. Figure 3-12 shows an example of Splitter in dynamic mode, where “from” = 1, “to”=3.3 and “splits” = 3. Splitter moves from box A to box B to box C until the abundance of peaks from box to box has not increased by factor “break”.

Figure 3-12: Example for Splitter in Dynamic mode for Splitter with example values.

- 1- The user selects a “from” and “to” value in the right tail of the histogram.
- 2- The user identifies the number of “splits” that he wants for Splitter to perform.
- 3- The user identifies the “break” value for which the algorithm will converge and stop.
- 4- The “from” to “to” range is broken into “splits” sections.
- 5- Starting from the lower bound, for each section, if the number of peaks has not increased by a factor of “break”, then the algorithm stops and reports the clusters.

Algorithm 3: The splitter algorithm

Two additional parameters are needed by Splitter:

GAPMAX: is the maximum distance in base pairs where clusters will be joined together.

A cluster is a set of $\log_2$ ratios which are consecutively mapped on the genome.

MINRUN: is the minimum number of consecutive probes showing significantly increased signal intensities relative to TI for which a cluster will be considered to be a positive peak.

Splitter combines array replicates by calculating the mean or the median for a given probe on the replicates. The output of Splitter consists of a table that lists putative peaks (i.e. binding sites) ranked by the median $\log_2$ signal intensity ratio of the probes in each predicted binding site. Columns of the table are the following: *Chr*, the chromosome where the peak is located, *Start*, the chromosome location where the peak location begins, *End*, the chromosome location where the peak location ends, and *Median Signal*, the median signal of the probes contained in the peak. Since analyzing chIP-chip data consists of applying both a normalization method and a peak finding algorithm, the user is invited to section 3.5 for Splitter results and discussion.

3.3.2. Summary

The Splitter method was chosen for the analysis of TRE chIP-chip data based on the recommendations of Johnson *et al.*[29] The Splitter algorithm includes various parameters that must be tuned by the user (described in subsequent section). The next section focuses on finding the optimal combination of normalization and peak finding algorithms using benchmark data that most closely represents the characteristics of the microarrays that were used in the present study (produced by Dong *et al.*). A detailed description of the analysis of the data from Dong *et al.* is described in section 3.5.

3.4. Evaluation of performance with benchmark data

A study carried out by Johnson *et al.* [29] evaluated the performance of a number of available platforms for chIP-chip data collection and analysis. Many performance criteria were evaluated including specificity and sensitivity, along with the data processing methods used (normalization and peak finding). Briefly, the authors of the study conducted a large scale spike-in study. The spike-in consisted of random selections of 100 known DNA segments (497 bp long) added at known concentrations to a sample of human DNA preparation (see Johnson *et al.* for more details). They sent out the spike-in sample to collaborating laboratories which had a role in performing the chIP-chip experiment and analysing the data without knowing the genomic location of the 100 probes.

The Johnson *et al* study included 43 chIP-chip datasets. Of these, the Whitehead dataset used amplification of the IP channel and the Agilent platform. The dataset has 2 arrays containing the same sample (i.e. two replicates). This dataset most resembled our own experimental parameters (see Table 3-1 for a comparison of platform characteristics) and was therefore used to evaluate the performance of this technology and analysis. Parameters were identified for the Whitehead dataset from Johnson *et al.* that were the most efficient in terms of sensitivity and precision. Correlation of these parameters to the Dong *et al.* dataset is then made.

	Johnson <i>et al.</i>	Dong <i>et al.</i>
Probe length (nucleotides)	60	60
Median genomic region between probes (nucleotides)	10	153
Number of probes/array	244 000	44 000
Average length of DNA segment hybridized to the array	497	600
Amplification technique	LM	WGA

Table 3-1: Comparison of the platform used for Spike-in and our chIP-chip data

Since the Splitter method was recommended by Johnson *et al.*, efforts were concentrated on re-analyzing these datasets with the supplied documentation in order to determine the influence of the normalization on the sensitivity and precision factors of the analysis. While the supplemental documents provided by the authors provided insight into the parameters used to perform the analysis by the Whitehead institute, two issues were of concern. First, no information was given about the normalization method that was used and second, the value of the “to” parameter of Splitter (see Algorithm 3) was not mentioned specifically in the supplemental material. Here is a quote: “initial guess about histogram separation based on visual information”. This makes any experimental replication impossible, since they have used the dynamic Splitter and no specific value of the “to” parameter is given. Therefore, in order to have consistency in chIP-chip processing, it was decided to evaluate the datasets based on the content and nucleotide normalization method using the Splitter algorithm in manual mode to find the most beneficial set of parameters. The Splitter manual mode was selected since replicable user selections of “from” and “to” from one data set to another cannot be defined. The standard deviation was used in order to have a standard cut-off value which can be calculated for each dataset. When using the standard deviation as cut-off, only Gapmax and Minrun are meaningful, since the portion of the histogram selected is automated. In our analysis, the standard deviation was used however, its corresponding percentile would have given a similar cut-off since our chIP-chip datasets are quasi log normal. For Gapmax and Minrun, respectively 200 and 2 was used, which are identical to the parameters used in the data analysis for the Whitehead dataset [29]. When evaluating the

performance of the normalization and peak finding methods, a hit is identified as a true positive if a predicted binding site has any overlap with the true starting and ending location of a spike-in DNA segment. This definition of true positive is illustrated in Figure 3-13. A predicted binding site that doesn't fall within the 4 cases below is labelled as a false positive.

Figure 3-13: Illustration of the 4 cases when a predicted binding site is called as a true positive

Table 3-2 shows the results for the combinations of methods (normalization and peak finding) for Splitter when using a 2.5 standard deviation cut-off parameter value. Appendix A holds the other tables that include 2.25 and 2.75 standard deviation as cut-off values for Splitter.

Table 3-2: Results from the Whitehead data set with Splitter at 2.5 SD

	WhiteHead, 2.5 SD		
	SP	CNCL + SP	CNCL2 + SP
Number of True Positives	54	54	67
Number of False Positives	2	2	30
Number of False Negatives	46	46	33
Total number of sites found	56	56	99
Sensitivity	0.54	0.54	0.67
Precision	0.96	0.96	0.67

SP = Splitter with no normalization

CNCLN + SP = Content and nucleotide location method followed by Splitter

CNCLN2 = The I binary function in Equation 3 was removed and the model was applied to each dye followed by Splitter.

CNCL incorporates both dye measurements in the same model, the correction applied to the chIP-chip data is essentially the dye effect when the log₂ ratio of the probe IP/TI is calculated. One can tentatively conclude that the dye effect calculated by the CNCL model for the Whitehead dataset is not a major factor or the model doesn't appropriately fit the residual signal value for each dye. This conclusion is based on the fact that the results in Table 3-2 (number of true positive, false positive, false negative, sensitivity and precision) are identical for SP and (CNCL + SP). CNCL2 calculates the model separately on each channel omitting the δ factor. Each channel is normalized separately before the IP/TI ratio is taken.

In both normalization methods (CNCL and CNCL2), each coefficient shows a high level of statistical significance (F-values, identical to description in section 3.2.2.2). However, it appears that the binding effect has a greater consequence on the normalization method that corrects for the nucleotide content and nucleotide location of the probe. Even though the data were normalized with CNCL2, there are few differences in the results produced by CNCL2 and {SP, CNCL} (Table 3-2). However, these are preliminary results; more arrays would be needed to conclude with certainty the reasons why differences exist in the performance of these methods.

Johnson *et al.* generated performance curves (ref [29]), Figure 2) that are similar to receiver operating characteristic (ROC) curves. They plotted TP/(TP+FN) vs FP/(TP+FN) for a variety of analytical methods using the Whitehead dataset. For the purpose of comparison, Figure 3-14 was included to evaluate similarity. Unlike CNCL2, CNCL normalization had no effect on Splitter. Therefore, only the Splitter curves are shown below. In order to have the largest true positive / false positive rate, one must select the ROC-like curve that is the closest to the upper-left corner of the graph

(indicated with an arrow in Figure 3-14. In this case, the best cut-off value for Splitter is 2.4 SD with no normalization applied.

Figure 3-14: ROC-like curve for the Whitehead dataset

(Note: SP and CNCL + SP curves overlap perfectly)

3.4.1. Evaluation of the precision of binding site cut-off

For each predicted binding site, the output from Splitter contains a chromosome number, a starting location, and an ending location for each peak. From the evaluation of performance where 54 binding sites were found, the difference between the true start and end values was calculated to the predicted start and end values. Since a true positive was labelled when any overlap occurred (see Figure 3-13), histograms were generated for the differences in extents; (true start value - predicted start value) and (true end value - predicted end value). The histograms in Figure 3-15 and Figure 3-16 show the positional biases from the start and end prediction values respectively.

The histograms in Figure 3-15 and Figure 3-16 reveal a strong bias for the prediction of the end position. A high density at (true value – predicted value) would mean that the predicted value is often accurate as opposed to high deviations from 0 like shown in Figure 3-16 means that the algorithm is conservative and provides an estimate of the length of the binding site that is too short. The bias indicates that the predicted ending value is very conservative and therefore should be accounted for. The next section discusses this issue in the context of Splitter parameter optimization.

Figure 3-15: Difference between the true and predicted start values for the true positives in the Whitehead dataset

Figure 3-16: Difference between the true and predicted end values for the true positives in the Whitehead dataset

3.5. Optimisation of peak-finding algorithm parameters to actual THR study data

Based on the evaluation of performance conducted on the Whitehead dataset, no normalization method was applied in the analysis. Even though the sensitivity is higher with CNCL2, the benefit comes with a lesser precision. Higher precision enables us to have greater confidence in the TRE motifs that will be found in the following chapter. The analysis summarized in Figure 3-14 revealed that the optimal cut-off value for the Splitter was 2.4 standard deviations. The analysis performed by the Whitehead laboratory on the Whitehead dataset was performed using a GAPMAX of 200 and MINRUN of 2, below is the discussion of the parameters using in the THR study.

We argue that MINRUN must remain at 2 because diminishing the value to 1 would reveal a very high number of hits with no neighbouring probes to increase confidence in peak identification. Further elevated ratios for single probes can result from random

chance. Since the wet-laboratory manipulations are the limiting factor, this issue is technology-based and could not be avoided.

In contrast, setting the MINRUN to 3 could limit tremendously the number of hits obtained. As formulated earlier by Keles *et al.*, the number of probes expected per peak is 3.09. The conclusions made by Keles *et al.* leads toward a much lower number of probes per peak than expected. All analyses are based on maximizing the balance of true positives to false positives. In a biological experiment where only a limited number of hits may be expected, maximizing the identification of peaks is desired. However, the biologists should be cautioned that these peaks warrant careful scrutiny and validation using alternative approaches.

Since a MINRUN of 1 is too low and a MINRUN of 3 is too high, this leaves us with the single possibility of setting this parameter to 2.

If the GAPMAX value utilized by and for Whitehead is considered, they used a GAPMAX to cover the mapping of 3 probes. In order to cover the mapping of 3 probes, a GAPMAX of 488 would be required due to the Dong *et al.* microarray design. However, 488 is very close to the size of shredded DNA length which is 600. Evaluation of the right estimate for this parameter is not trivial since there are many variables to take into account when doing so. The Whitehead dataset was performed with a spike-in which is generally less noisy than a real sample. Results are reported up to a GAPMAX for 488, but priority to the evaluation of the results should be given to peaks having shorter GAPMAX, i.e. validation should be taken to peaks having lower GAPMAX.

3.6. Results of peak finding to actual THR study data

The following section presents the results for the Dong *et al.*/TRE study. As a reminder, the chromatin immunoprecipitation was conducted on cerebellum DNA from 10 mice sacrificed on post natal days 4 and 15 (n=5 per group). Immunoprecipitated DNA was hybridized to microarrays and the analysis was performed on the data acquired from

these microarrays. The TRE dataset of Dong *et al.* was analyzed without normalization using the manual version of Splitter with a cut-off value of 2.4, a MINRUN of 2 and GAPMAX of 488. Peak ending values were also corrected by adding 200 bps. The list of enriched sites found by applying these parameters can be found in Appendix B. Briefly, 44 peaks were identified for mice collected on PND4 and 186 for PND15, for a total of 230 candidate binding sites.

An analysis was performed to reveal the location of peaks with respect to the closest gene.

Table 3-3 and Appendix B indicate the location of peaks with respect to mRNA mapped in the MM9 genome build of the UCSC genome browser. Although the mapping of the Agilent microarrays used for this experiment was biased toward locations associated with curated genes, it is evident that there was a bias towards TRE locations within the limits of transcription of a gene for the identified peaks.

Table 3-3: Location of peaks with respect to mRNA mapped in MM9 by the UCSC genome browser

In genes		Outside genes
144		86
In introns	In exons	
78	76	

Figure 3-17 shows the density of the distance to the nearest mRNA mapped by the UCSC on build MM9. Although the mapping was performed on genes from -8 to +2 Kbps, updating the genome version can lead to sections of the build that have lost their annotations for diverse reasons (e.g. new scientific evidence). However, one can observe that the majority of peaks are mapped within 25 Kbps of an mRNA mapped by the UCSC on build MM9.

Figure 3-17: Histogram of the distance to the nearest mRNA mapped by the UCSC on build MM9

Physical location relative to the start of a gene may seem to be the most logical predictor for control of gene expression by a transcription factor. However, physical proximity in 1 dimension (i.e., primary DNA structure), does not necessarily correlate with physical proximity of transcription factors due to the three dimensional structure of the DNA elements (e.g., folding of chromosomes) [21].

Appendix B reveals the peaks with their relationships to mRNA mapped by the UCSC genome browser. It indicates the mRNA mapped IDs, the location with respect to the closest mRNA and the number of papers for which the concatenated mRNA name and “thyroid hormone” reveals hits on Pubmed [38]. The number of publications for an mRNA name increases our confidence that this mRNA is regulated by the thyroid hormone receptor. However, as described above, further evidence is needed to validate this finding, since the proximity of a gene to a peak does not necessarily involve a link between a response element and a gene. Experimental validation provides this additional evidence as discussed in the next section.

3.7. Experimental validation

After gathering the results of chIP-chip peak-finding, biologists validate a subset of the predicted binding sites through chIP-PCR experiments. However, this method is very time consuming and only a few candidates can be chosen for wetlab validation. ChIP-PCR can be described in these steps; first, a polymerase chain reaction is used to amplify DNA segments. Secondly, relative copy numbers of specific amplified sites can be visualized by agarose gel electrophoresis. Then, those products determined from ChIP to be positive hits are compared to known negative control to confirm that there indeed was enrichment of the site. Dr. Dong, Health Canada scientist, carried out chip-PCR validation for 19 predicted chIP-chip peaks from the analysis performed in section 3.6. Dong's criterion of selection for choosing which peaks were going to be validated was based on information in the literature of the genes located near the location of the peak. In other words, predicted peaks that were located near genes that were proposed to be regulated by thyroid hormone in previous publications were selected for subsequent experimental validation. From the 19 peaks, 12 were confirmed true positive and 7 were confirmed false positive. This reflects a false positive rate (FPR) of 39%; however, this finding should be interpreted with caution. First, peaks selected for chIP-PCR were not selected with respect to the signal strength (they were selected based on evidence in the literature to support the control of the mRNA by TR). Second, the genomic region selected for validation was much smaller than the predicted peak for the 19 tested, which could lead to false negative (i.e. missed positive) results. Moreover, comparing the FDR to the results of other chIP-chip studies should also be interpreted with caution. Many factors other than the normalization and peak finding algorithms can influence the FDR such as the sample type (e.g., some biological tissues may be more strongly controlled by specific nuclear receptors) and technical expertise in the lab. Johnson *et al.* showed that the platform was not the major factor in the efficiency of peak discovery. However, since Johnson *et al.* sent the same sample to each participating lab, the nature of the sample is a factor that they couldn't evaluate. Taking these factors into consideration, the FDR of the analysis implemented on the Dong et al. dataset is similar to the FDRs obtained in other studies.

3.8. Summary

ChIP-chip data analysis can be carried out using various approaches. The work described above introduces the underlying assumptions that must be verified prior to committing to a specific analytical approach. One must consider that most methods are derived from studies of gene expression using microarrays and that ChIP-chip microarrays are quite different. Rather than measuring an increase or decrease of probe signal intensity (expression arrays), ChIP-chip experiments aim to quantify the presence or absence of a binding site. It must be kept in mind that, even when some biases are detected, they might not be sufficient to hide the signal. Also it is possible that the normalization technique may fail to remove the predominant biases in the data. From the analyses performed on the Whitehead dataset in this thesis, both hypotheses mentioned above are possible. However, more data would have to be generated in order to verify either assumption.

Here was determined the accuracy of the analytical approaches applied to the Whitehead dataset analyses, using approaches similar to those described by Johnson *et al.* In addition, in studying the accuracy of the prediction of the starting and ending point was corrected and found that the prediction is often very conservative and the peak region should be enlarged.

The results generated by the analyses performed in this chapter are predicting 44 peaks for the PND4 mouse dataset and 186 peaks for PND15 dataset. These DNA regions will be analysed by motif finding algorithms in Chapter 4 in an effort to characterize the TRE sequence motif and to narrow the predicted binding site from a relatively large DNA region (~600 bps) to a TRE specific site.

Chapter 4. Motif identification

The objective of this chapter is to find TREs in genomic segments believed to show significant enrichment in the IP channel (i.e., segments corresponding to the chIP-chip peaks predicted in Chapter 3). This chapter is divided into 2 major sections. The first section examines the genomic segments for the presence of known TRE consensus sequences (i.e. motifs), and the second presents exploratory work aimed at identifying novel TRE motifs in chIP-chip predicted peaks. In order to simplify the language in this chapter, peaks identified in Appendix B will be referred to as genomic regions of interest (GROI).

4.1. Searching for the known consensus TRE motif

Section 2.1.4 of this document relates an overview of what is known about TREs from the literature. As described, the consensus motif for TREs consists of the sequence “AGGTCA” arranged in dimers following one of three configurations: a direct repeat with a spacer of 4 nucleotides (DR4), a palindrome with no spacer (IR0), or an inverted palindrome with a spacer of 6 nucleotides (ER6). This section will explore the identification of known consensus TREs within GROIs (shown in Figure 4-1).

Figure 4-1: Partial flowchart of thesis with Identification of sites that are consistent with the consensus TRE highlighted

The objective of this section can be broken down into two main components (identification of hexamers and identification of dimers) aimed at identifying the optimal conditions under which to increase accuracy of TRE identification. Within these

components, analyses were conducted to find optimal scores to reduce Type I and Type II error rates, i.e. false positive and false negative rates. Specific questions are identified below.

For the consensus hexamer (AGGTCA)

- Does the number of half sites in a gene promoter region correlate with the presence of TREs? If so, can this be used to differentiate between true and false GROIs?
- How does the distribution and number of consensus hexamers in the GROIs compare to: randomly selected gene promoter regions, nucleotide permutation of these promoter regions, promoter regions of genes known to contain a TRE, and genes that are regulated by a nuclear receptor other than TRE, but contain the same half site?

For the consensus dimer (DR4, IR0, ER6)

- When a DR4 is found, is the gene it regulates on the same strand as the motif?
- Is there a bias in the location of the motif with respect to the transcription start site (TSS) of the nearest gene that would be indicative of a true positive motif?

4.1.1. Models for the identification of the TRE hexamer

Many approaches have been developed to scan for diverse transcription factor binding sites on DNA. These approaches rely on previously characterized binding sites in order to create a model to search for undiscovered binding site locations. Below is a brief overview of scanning models.

A number of approaches (e.g. [57]) rely on a position weight matrix (PWM) (see Equation 5) based on information theory models. A PWM is a matrix containing the frequency (m) of each nucleotide ($b = \{A, C, G, T\}$) in each position (i) in the binding site. In this approach, the learning set of known transcription factor binding sites (number of binding sites is n) must be aligned to one another prior to calculating the PWM.

$$\begin{aligned}
PWM(b,i) &= \sum_{i=1, b \in \{A,C,G,T\}}^n m_{b,i} \\
PRM(b,i) &= \frac{PWM(b,i)}{n} + 0.01 \\
W(b,i) &= \text{Log}_2 \left(\frac{PRM(b,i)}{\text{Pr}(b)} \right)
\end{aligned}$$

Equation 5: Information theory model for finding transcription factor binding sites

The probability matrix (PRM) scales the probability of occurrence of each nucleotide and position with respect to the number of binding sites used in the training set. A residual is added to this in order to eliminate having any probabilities equal to zero in the PRM. Since the number of binding sites in a training set is often very limited, it may be the case that the observed frequency of a particular nucleotide in a particular position is zero. To avoid making the corresponding probability of ever seeing this nucleotide in that position equal to zero, a small residual (0.01 in equation 5) is added to each probability. This is also referred to as adding a ‘pseudocount’ in the context of maximum likelihood estimation of nucleotide frequencies or incorporating a non-zero prior probability in the context of Bayesian parameter estimation. This small addition avoids making the strong assumption that a specific nucleotide would never be present in any one position of the PRM. The next step computes the log-ratio of each probability in the PRM matrix divided by the corresponding probability of the nucleotide (W). The *Score* function (Equation 6) sums up each nucleotide in the sequence according to its position to assign a score to each nucleotide sequence. Scores calculated by Equation 6 are measured in bits.

$$Score(s_{1..L}) = \sum_{i=1}^L W_{b(i),i}$$

Equation 6: Scoring function for the information theory model

Modifications have been developed to increase the selectivity of this method. For example, Tomovic *et al.* [54] proposed a model that accounts for position dependencies

between nucleotide positions (Equation 5 assumes position independence in the binding site motif). Their model uses various statistical approaches, such as Bayesian and randomization statistics, to evaluate the dependencies. The conclusions of their work suggest that the length of the binding site, as well as the number of samples in the training set, are important factors for evaluating the probability of nucleotide dependence on one another in a set of binding sites. However, the authors conclude that there is no ‘universal’ answer.

Other position dependency models have been developed including Bayesian networks and hidden Markov models (HMMs). Tomovic *et al.* indicate that these models usually perform better, particularly by decreasing the false positive rate. However, they require optimizing architectural parameters prior to training (which is often not trivial) and numerous binding sites for training (which are quite limited in the present study). For example, the order for a HMM must often be specified by the user before a model can be trained. In other words, in order to use these models efficiently (i.e. high sensitivity and specificity), one must have a high number of training data (i.e. known binding sites) and must be able to parameterize models efficiently. As an example of the types of parameters that must be determined from the training data, for an HMM the state transition matrix must be estimated along with the emission probabilities for each node, and the overall structure of the model must also be optimized.

Thus, due to the limited number of characterized binding sites available in the published literature, and the fact that we have chIP-chip evidence of a binding site occurrence (restricts considerably the search space since it targets short DNA sequences), the Equation 5 model was selected for identification of TRE binding sites in the GROIs.

4.1.2. Determination of the correct DNA scanning model for TREs

A DNA scanning model scores the similarities between a set of known binding sites (here TREs) and a specific starting position in a DNA sequence. While these methods will

ultimately be applied to GROIs, known TREs are first analyzed in order to optimize the parameters of the scanning model.

TREs gathered from the literature in Table 2-2 were split in two hexamers (where applicable, i.e. non- $\{\text{DR4, IR0, ER6}\}$ TREs were discarded) and aligned to create a list of half sites. The model from Equation 5 was applied to the known half sites using a leave one out (LOO) test protocol. The resulting left skewed score distribution is shown in Figure 4-2. The distribution can be broken into 2 sections, a right region (blue) with high scores and a small left region (red) (analysis to follow). In the sections that follow, a two-threshold rule is developed for classifying putative TREs. The coloring in Figure 4-2 corresponds to the optimum threshold values from Table 4-3 below. The negative scores are driven by half sites containing two C's in positions 2 & 3 instead of G's.

Figure 4-2: Histogram of Scores from half sites from the literature

One hypothesis for the left shifted distribution in Figure 4-2 is that TRE dimers are primarily composed of low score-high score and high score-high score combinations of half sites (defined below). Table 4-1 and Table 4-2 show the properties of the tests that were made in order to examine this hypothesis. Table 4-1 shows the scores of the half sites for mouse TREs (see Table 2-2) where 'Score 1st hs' and 'Score 2nd hs' indicate the results of applying the score function from Equation 5 to the first and second half site

respectively of the dimer. In order to eliminate any bias in the analysis, these scores were calculated using a leave 1 out cross-validation. The same strategy was employed for the results of the other species (see 2.1.4.1.1) analyzed (Table 4-2).

Bootstrapping was used to evaluate the null hypothesis that the min and max score in each pair of half sites was, in fact, drawn from the same distribution (i.e. the mean difference between min and max of each row of Table 4-1 should be zero). Applying bootstrapping with 1000 bootstrap samples (drawn with replacement), we are able to reject the null hypothesis (with a $p\text{-value} < 0.01$) for Table 4-1. We therefore accept the alternative hypothesis that there is, in fact, a significant difference between min and max.

To enhance the confidence in the mouse TREs bootstrapping results, an identical test was performed on TREs from other species located in Table 4-2. Again, with high confidence ($p\text{-value} < 0.01$), we were able to reject the null hypothesis, and concluded that there is a significant difference in scores between the half sites in previously known TREs.

Table 4-1: Scores of halfSites for murine TREs

TRE ID	Score 1 st hs	Score 2 nd hs
1	7.7	6.0
2	-6.4	6.1
3	5.4	2.5
4	9.2	6.9
5	6.8	7.8
6	1.9	6.0
7	-4.1	3.5
8	6.0	3.6
9	7.3	4.9
10	6.5	9.3
11	5.9	9.2

Table 4-2: Scores of halfSites for rat, human, chicken TREs

TRE ID	Score 1 st hs	Score 2 nd hs
1	2.1	8.8
2	8.6	9.5
3	8.5	1.8
4	8.7	3.1
5	8.7	8.9
6	9.5	8.8
7	8.7	8.9
8	8.7	8.9
9	-3.6	4.7
10	8.0	4.7
11	8.7	8.9
12	3.5	9.5
13	8.6	7.0

The results above suggest a rule for identifying potential TREs. For a given putative TRE site, the score of each half site is computed. The lower of the two scores is subjected to one threshold (lowScore below), and the higher of the two scores is subjected to another threshold (highScore below). If the two scores pass the two thresholds, the putative site is classified as a TRE; otherwise, it is rejected.

An experiment was conducted to optimize the two threshold values in order to minimize the false positive rate while maximizing the sensitivity. This must be done for both for the low-score and high-score thresholds. This was accomplished by gathering the genomic regions around the TREs found in the literature and performing cross validation. Since the median length of the chIP-chip sequences in the present study is 520, we extended half of this length on each side of the published TREs to match the median chIP-chip sequence length. There were also false positives generated by the chIP-chip data analysis that could be used to measure the false positive rate; false positives were identified in the wet-lab (section 3.7) and they were used in the cross-validation analysis.

The LOO cross validation was conducted by evaluating every possible pair of threshold values defined by Equation 7 as the two lower bound thresholds to identify TREs in the sequences mentioned immediately above.

Define 2 ensembles:

$$low = \{3, 3.05, 3.1, 3.15, \dots, 5\}$$

$$high = \{5, 5.05, 5.1, 5.15, \dots, 8\}$$

$$(\text{minScore}, \text{maxScore}) = (p, q)$$

$$\text{where } p \in low \wedge q \in high$$

Equation 7: low and high score instantiation

For this analysis, known TREs that did not correspond to consensus motifs (DR4, IR0, ER6) were left out in order to simplify the analysis. Also, due to the various genome builds available, some TRE sequences could not be found within their respective genomes. Thus, although they were used to train the model, the sequences upstream and

downstream (sequence extension) were not available and as such they were not used for testing.

Table 4-3 shows a portion of the results of the cross validation (11 LOO) for scores. The entire table of results is available on request; only relevant rows are shown here. The table columns are: Min and MaxScore are the cut-off scores (see Equation 5), True Positive are the number of TREs found in 10 sequences containing 1 TRE, False Positive are the false positives found in the respective sequence. That is, false positives are identified as the TREs (satisfying a low and high score) found in the segment of DNA being analyzed, that were not in the same location as those validated TREs in the literature for that segment of DNA. False Positive 2 is the number of false positives found in random sequences selected from false positives confirmed by chIP-PCR (section 3.7).

Table 4-3: Sample of True Positive and False Positive rate vs. Min Score and Max Score

MinScore	MaxScore	TruePositive	False Positive	False Positive 2
3.75	6	8	3	0
3.76	6	8	2	0
3.77	6	8	2	0
3.78	6	8	2	0
3.79	6	7	2	0
3.8	6	7	2	0

The optimum cut-off point was selected by subtracting the true positive from the false positive column with the highest difference. From there, all rows with non zero false positive 2 values were eliminated. A small list resulted and the ratio (false positive / true positive) was calculated. The best range found by the LOO cross validation was: minScore (3.76 to 3.78) and maxScore (6.0) with a false positive rate of 0.2.

4.1.3. Relative abundance of TRE hexamers in DNA sequences

This topic will be broken down into 2 sections. In the first part, the density of TREs (i.e. frequency of TRE hexamer appearance) will be evaluated across various conditions. The

second will examine the density of half sites in the promoter regions of known T3 regulated genes.

Prior to examining TRE density, an appropriate score cut-off value needs to be determined for halfsites. We found that a cut-off value of 6.3 was appropriate for half sites, since it is the optimal high score (defined previously). Only one threshold was selected since this section now looks at the characteristics of individual half sites. A higher threshold was selected in order to diminish the number of half sites that have a low probability of being functional.

4.1.3.1. Density of TRE consensus hexamer

Sequences in groups mentioned for each figure below were scanned for half sites with a lower bound cutoff value equal to or greater than 6.3. Estrogen regulated genes were chosen for this comparison since the characterized motif for their nuclear receptor is highly similar with the exception that the orientation of the motif is a palindrome with a 3 nucleotide spacer. Figure 4-3 shows (A) the density of TRE half sites per 1000 bps in permutations of random promoter regions (n = 50), (B) randomly selected promoter regions (n = 50), (C) promoter region for genes with TH regulated promoter regions (n = 28), (D) promoter region for genes with estrogen regulated promoter regions (n = 50), (E) GROIs regions (n = 100), and (F) GROIs validated regions (n = 36). The results shown by Figure 4-3A are random nucleotide permutations of the identical promoter regions shown in Figure 4-3B. The differences in the number of half sites in both figures suggest a nucleotide position importance with respect the sequences composition since permutations of identical sequences show a much lower number of half sites. Figure 4-3 B, C & D show very similar distributions. Thus, no significant differences are evident between random promoter regions, TH regulated promoter regions (Table 2-2) and estrogen regulated promoter regions.

Figure 4-3: Frequency of TRE half sites per kbps found in relevant DNA regions

Therefore, frequency and density of half sites is not a significant discriminating factor for the identification or characterization of TRE containing promoters. Figure 4-3 E & F shows the distribution of half sites in GROIs and GROIs validated regions. The distribution of TRE motifs across these regions is broader and half site frequency can be much higher. Because these regions are much smaller (300-600 bps) compared to the entire promoter regions, dilution of the concentration of half sites may explain the differences in distribution between the two groups ((Figure 4-3 A, B, C & D) & (Figure 4-3 E & F)).

4.1.3.2. TRE hexamer half site relative abundance across genes known to contain functional TRE

The above findings lead to a closer examination of the promoter regions of genes known to be regulated by TH. A sliding window of 250 bps was used to examine the number of half sites within the promoter regions of 4 TH regulated genes (Figure 4-4).

Figure 4-4: Density of TRE half sites in promoter regions of 4 TH regulated genes. Each graph shows the NCBI Reference sequence ID of the T3 regulated gene, and its chromosomal location. The Y axes indicate the number of half sites and the X axis indicate the nucleotide position with respect to the gene where 0 = -8kps from the transcription start site. The arrows show precisely where the TREs are located for each known TH regulated gene.

This analysis revealed that there is no significant increase in the number of half sites in the promoter region where the binding site (TRE) is located. Therefore, TRE frequency

and density are not predictive for TR controlled genes and we are unable to use these features to garner confidence in true positive peaks identified using ChIP-chip.

4.1.3.3. Summary

In summary, the analysis of TRE distribution across promoter regions known to be regulated and not regulated by thyroid hormone, as well as random permutations of those promoters, revealed highly variable and overlapping distributions. Thus, classification of genes with respect to the half site density within their promoter region is not possible.

4.1.4. Analysis of TRE ChIP-chip sequences for the THR and AP-1 binding site

Here we apply the strategy, developed in section 4.1.2, to identify TRE half sites to the ChIP-chip dataset to identify TR binding sites (i.e., TREs). In addition, a SP-1 motif scanning strategy was applied and will be discussed in detail.

4.1.4.1. THR motif identification in GROIs

Here we present a list of potential candidates found in the GROIs. Initially, an optimal cut-off for finding TRE half sites was found, then the GROIs were scanned for pairs of TRE half sites that complied with the known motif rules (DR, IR, ER) and included the high-high and high-low scores approach developed above that led to a high true positive and low false positive discovery rate. This list can be found in Appendix C. We identified 15 potential TREs in 44 GROIs for PND 4 and 80 potential TREs in 186 GROIs for PND15.

Of the potential TREs (DR4, IR0, ER6), three in PND4 and ten in PND15 show multiple TREs within a single GROI. In other words, more than one TRE consensus was found within a single DNA region corresponding to a GROI. This interesting phenomenon may be explained by the tendency for transcription factor binding sites to appear in groups where one or more could be functional in gene regulation [40]. These additional TF binding sites are often referred to as “shadows”.

Analysis of the orientation of TREs in the GROIs revealed that 56 were DR4s, 28 were ER6s and 11 were IR0s. Although prior information on TREs is limited, these numbers follow the same trend as the TREs listed in Table 2-2, as the relative abundance of each conformation (DR4s, ER6 and IR0) is similar.

4.1.4.2. Ap-1 motif identification in GROIs

The TH mode of action proposed by Lazar *et al.* establishes a link between the Jun/Fos nuclear receptors and the AP-1 binding site. Thus, the PWM for AP-1 was downloaded from the JASPAR database [7] in order to scan the GROIs for the AP-1 consensus with the same scoring function utilized for the TREs. The dataset from which binding sites for the AP-1 motif is learnt contains very few instances. Figure 4-5 illustrates the distribution of the scores for the known JASPAR AP-1 sites which have a consensus of “TGAGTCA”. In order to limit the number of false positives, TRE chIP-chip sequences were scanned for a score equal to, or greater than 8.6 bits, which represents the score of 70% of the known AP-1 binding sites, other AP-1 having lower scores.

Using this approach, the model identified 16 potential AP-1 binding sites in 44 PND4 GROIs and 53 potential binding sites for 186 PND15 GROIs. From these, PND15 had 10 GROIs which had more than 1 AP-1 potential binding sites and 4 were found for PND4.

A combination of TRE and AP-1 searches identified 5 GROIs for PND4 for which a TRE had previously been found. Analysis of PND15 revealed 21 potential AP-1 sites where TREs were previously identified Section 4.1.4.1. In total, 26 GROIs contain both a TRE and an AP-1 site. Thus, this suggests that this is biologically meaningful and suggests that genes regulated by these transcription factors are coordinated in their mode of action.

Figure 4-5: Histogram of the Scores for the AP-1 Binding Site from Jaspar

4.1.5. Summary

In this chapter section, the {min, max}Scoring model was established in order to identify putative TREs. Also studied was the effect of the presence of half sites in DNA sequences. This section identified known THR interacting binding protein motifs in GROIs. A total of 164 potential binding sites were found. A more extensive discussion is reserved for the end of this chapter since more complete information on all potential motifs (known and novel) will be available.

4.2. Identification of novel motifs

After identifying the location of motifs that are known to follow the TRE and AP-1 consensus, further analyses were carried out to identify novel motifs, i.e. short nucleotide sequences that show conservation across DNA regions corresponding to GROIs, and that have a low probability of randomly occurring in the genome. Figure 4-6 identifies which step of this project is currently being elaborated.

Figure 4-6: Partial flowchart of thesis with identification of novel TRE highlighted

Many algorithms have been developed to find enriched motifs in a set of sequences and reviews have been written categorizing motif finding algorithms (MFA). Reviews by Das *et al.* [10], Tompa *et al.* [55] and D’haeseleer [14] categorize algorithms into 3 groups: enumerative methods, deterministic optimization, and probabilistic optimization. In particular, the reader is referred to D’haeseleer *et al.* for a quick and concise overview of the groups of MFA. In addition, Das *et al.* generated an extensive list of motif finding algorithms that can be found in the supplemental materials of their publication.

Hu *et al.* [28] conducted an experiment on prokaryotic sequences to rate the performance of MFAs. However, using the approach of that study for the assessment of TRE motif identification from chIP-chip dataset should be undertaken with great caution. The level of complexity of binding sites (degeneracy, multiple half site motifs) is generally much higher in eukaryote cells, which are a more sophisticated cell type than prokaryotes therefore making the mechanisms of gene regulation more complex. Prokaryotes are different from eukaryotes in that their cells do not have a nucleus and their genomes are

packaged in a DNA/protein complex called nucleoid. These organisms have much less DNA in general, and slightly different mechanisms of gene regulation.

The goal of Hu *et al.* was to evaluate the performance of MFA with minimal human intervention, i.e. by setting a minimum number of parameters to the MFAs. According to the findings of Hu *et al.*, the most important factor that limits motif finding algorithms is the length of the sequences across which the motif search is conducted. Their performance indicators (see section *Measures of prediction accuracy* in [28]) show a constant decrease in performance (see Figures 6, 7, 8 in [28]) as the length of sequences submitted to the MFAs increases. GROIs presented in Appendix B have a length distribution presented in Figure 4-7, which shows a significant variation in the length of the GROIs sequences. Hu *et al.* did not explore the performance of MFAs on a dataset in which sequence lengths vary. It is unclear whether inclusion of shorter sequences within a heterogeneous set of sequences may improve motif finding, possibly through concentrating the maxima of the algorithm which could then be found in longer sequences. In the present study on TRE identification, we have compromised by assuming equi-length with a median length at 520. Given these assumptions, a performance indicator (see Hu *et al.* equation 6) of around 0.4 can be expected.

Figure 4-7: Histogram of the distribution of GROIs length presented in Appendix B

Although there are disagreements over certain aspects of MFAs, all the reviews on this topic have the following consensus; several MFAs of complementary underlying models should be applied to a dataset and the top motifs from each method should be pursued with wet-laboratory validation. This strategy is employed below with algorithms presented in the next section. Once motifs are found by the algorithms, they will be compared to pre-existing motifs from other DNA binding proteins.

4.2.1. Application of motif finding algorithms to the TRE dataset

The following section presents the strategy used for employing MFAs to search for novel motifs within the TRE chIP-chip dataset. Ultimately, the analysis will generate a list of motifs, which will be related to the literature to contrast with known THR motifs.

4.2.1.1. Considerations before applying MFAs

Background Estimation

Background estimation is often employed by MFAs in order to estimate the probability of a novel motif occurring by chance in the genome. As in MEME [1], transition probabilities are gathered in tuples so that a hidden Markov model (HMM) can be generated. From there, each novel motif is evaluated by the HMM in order to reveal the probability of this motif being found in the background.

The first step in this work is to gather sequences from the TRE chIP-chip experiments where no binding is expected. This will form our ‘true negative’ dataset. In section 3.6, true binding signal was identified as having an intensity greater than the mean + 2.4 SD of the log-ratio probe intensities. In order to have a conservative estimation of background, the sequences of probes and surrounding areas that yielded signal intensities that were less than 1 SD greater than the mean were collected and labelled as background sequences.

Gathering sequences from the TRE ChIP-chip as the background dataset permits the assessment of whether or not a novel motif is not expected to occur in non-TRE sequences, which increases the confidence in associating the motif with TH binding.

4.2.1.2. Approach used in applying the MFAs

The approach for identifying novel motifs uses prior evidence. Published research demonstrates that THR binds to DNA in combination with itself or other proteins. Therefore, the first step examines the hypothesis that another gapped motif (other than the consensus) can direct THR binding, or the possibility that THR interacts with another DNA binding protein whose motif or THR interaction is unknown. The second step examines motifs without gaps to test the hypothesis that DNA binding co-regulators may also bind to non-gapped motifs. In summary, here is the approach for identifying motifs;

1. Use MFAs that search for “bi-partite”, i.e. gapped motifs will be executed.
2. Use MFAs to find motifs without gaps.

A literature search revealed that Bipad was the only utility that was able to account for the various orientations of dimer binding sites (see Table 2-1). As such, Bipad [4] was selected to perform the novel motif finding for the gapped motif searches. The Bipad algorithm goes through an iterative process in which random starting positions are given for the left part of the motif. The right part of the motif positions are generated knowing the user specified minimum gap distance allowed, the minimum and maximum gap length set by the user. Entropies for each half site are then calculated for each possible half site considering the min and max gap.

With the nucleotide sequence that contains the minimum entropy, alignment is performed on each sequence of the set in order to find the group of sequences given a left half site motif length, gap min, and gap max as well as the right half site motif length that has the minimal entropy. This is performed until the entropy of the group is less than a threshold delta (a preset parameter on the package web interface). These iterations are performed and run to obtain convergence when “cycles” are down to 0.

The second part of the strategy consists of applying MFAs that are not intended for finding bi-partite motifs. These algorithms can find gapped motifs, however, unlike Bipad, they are not designed for maximizing information content with respect to the orientation of half sites (see Table 2-1). The tools selected were chosen in order to evaluate one method from each category defined in D'haeseleer *et al.* [14] and are outlined below.

1. **Expectation maximization (EM):** MEME [1] was chosen in this category since it has the capability of running with a minimum number of user defined parameters and its performance is comparable to the other EM algorithms tested in [28]. Briefly, a PWM is built from random subsequences. The probability (Z) of matching the PWM of the motif is calculated for all positions of the sequences. This is the expectation step, i.e. calculate the expected frequency of this PWM in all positions in a set of sequences. The PWM is then refined; (Z s) are used as weights to recalculate positions of the PWM and are averaged across all Z (s) for this position. This is the maximization step, i.e. increasing the knowledge from what was learned from the expectation step. The EM steps are performed until a change in the PWM is minimal and below a certain defined threshold.
2. **Gibbs sampling:** Bioprospector [36] uses a strategy involving Gibbs sampling, where the motif binding start sites are chosen iteratively and randomly. Following each iteration, the algorithm probabilistically determines whether to add or remove a site. Therefore, with each iteration the model is refined and more information is acquired. As such, the number of iterations has a great influence on the convergence of the Gibbs sampling model.
3. **Enumeration:** The Weeder [41] enumeration algorithm was chosen since it was favourably reviewed by Tompa *et al.* [55]. The data structure implemented in Weeder is a suffix tree for which sequences with a maximum number of mismatches (e) are allowed. Briefly, the Weeder algorithm proceeds as follow: starting from the root of the tree, patterns are found recursively by increasing the length of the pattern by one nucleotide at each iteration and pruning the rest of the

tree when (e) is exceeded. A heuristic approach is used for long patterns where pruning occurs even before (e) is met, diminishing greatly the search space for a pattern.

As explicitly mentioned by Hu *et al.* [28], two predominant factors influence the success rate for the identification of motifs independent of the choice of algorithm. The first factor is the length of the sequences examined by the MFAs. In our case, the wet laboratory technology dictates the length of sequences, the most influential factor being the resolution of the array. Secondly, the number of sequences submitted to the MFAs can also play a role in the success rate of the identification of motifs (see section Effect of the number of input sequences in [28]).

4.2.2. Results of MFAs on the TRE dataset

This section presents the results of the application of the algorithms described above on the TRE dataset. The findings are presented in the same order as the description of the approaches above. The logos for the discovered motifs are shown below, along with comparable motifs available from the Transfac and Jaspar motif databases.

4.2.2.1. Application of Bipad

Bipad was applied to the DNA sequences identified during the GROIs identification stage, in order to find a two block motif with a width of 6 nucleotides each (i.e. hexamers) and a spacer of 0 to 8 nucleotides. Both DNA strands were searched for motifs using Bipad with a maximum of 500 iterations. Only the sequences that corresponded to the 5 top ranked GROIs of PND4 and top 25 GROIS of PND15 (TOP5-PND4&TOP25-PND15) were searched in order to limit the number of sequences provided to the algorithm to address the second factor discussed above. This allows a GROIs representation of both PND4 and PND15 in the highest score range. Table 4-4 shows the left and right half site motifs and the distribution of the gap between the 2 half sites.

Table 4-4: Output of Bipad (left half site motif logo, the distribution of the length of the spacer, and the right half site motif logo)

It should be noted that Bipad does not allow user input for any background comparison dataset (see 4.2.1.1). Moreover, it doesn't perform an evaluation of the motif found compared to what can be generated from a background dataset. It relies entirely on the information content present in the set of sequences submitted to the algorithm to set background probabilities.

Bipad was used in order to examine the possibility of having a dimer motif that is different than the consensus revealed in the literature. The interesting fact that can be observed in the 2 half sites is the predominance of "GG" in the reverse complement of the first half site at position 2 & 3 (appears as "CC" in positions 4 & 5 of forward strand of left half site above) which can also be found in the position 2 & 3 of the second half site. Further validations in the wet laboratory will have to take place in order to reveal whether or not this motif is biologically relevant or not.

4.2.2.2. Identifying novel motifs with MEME

The TRE chip-chip dataset was analysed by MEME using all the sequences and the subset TOP5-PND4&TOP25-PND15. Although Hu *et al.* recommended examining a subset of sequences for motifs; we felt that applying MEME to the entire dataset could reveal motifs that are not present, or are too degenerate, in one subset to be detected by MEME.

The algorithm was applied using various user-defined parameters. The first user parameter is "-dna", which indicates that MEME is scanning DNA sequences. The

second is the model to be used “-mod tcm”, which determines the number of motifs per sequence (0 or many occurrences of a motif per sequence). The third is the number of motifs to be searched “-nmotif 25”, i.e. a maximum of 25 motifs to be searched. The parameters “-minw 6” and “-maxw 25” are the minimum and maximum allowable lengths of the motifs for which MEME should search. The next parameter is “-revcomp”, which instructs MEME to look for motifs in both DNA strands. MEME uses background nucleotide frequencies (described above) in order to estimate the statistical significance of a putative motif as measured by *E*-value (described below).

Table 4-5 and Table 4-6 present the motif logos that were identified by MEME. The TOP 5 scoring motifs are displayed and are labelled motifs 1 through 5. One condition for inclusion of these motifs in this table is a very low *E*-value ($E\text{-value} < 10^{-13}$). The *E*-value (expected value) measures the expected frequency of this motif in random sequences given the background nucleotide frequencies.

One observation of the motifs discovered by MEME in Table 4-5 (motifs 1, 2, 3, 4) and Table 4-6 (motifs 1, 2) is the predominance of {G, C} in the motifs. This raises the question of whether or not these motifs are found predominantly in CpG islands, thus being an artefact. CpG islands are DNA regions associated with promoters where methylation plays an important role in mediating gene expression. A generalized definition of a CpG island (introduced by Gardiner-Garden *et al.* [22]) is a locus that is greater than or equal to 200 bps in length, which has a G + C nucleotide content and (observed CpG /expected CpG) ratio in excess of 0.6. Another condition is that the CpG island must be in a promoter region of a gene. Since the TRE chIP microarrays mapped promoter regions exclusively, the last condition is true for all the data. Emboss [47] provides a tool in their package “newcpgreport” for which one user can submit sequences to determine if they are CpG islands. Using the definition above, 182 GROIs of 230 did not contain a CpG island. GROIs suspected to contain a CpG island are listed in Appendix E.

Although Table 4-5 (motifs 1, 2, 3, 4) motifs show very low E -values, motifs 1 and 3 are respectively located within CpG islands 92% and 84% of the time (GROIs that match the motif). Table 4-5 (motif 2) is located in CpG islands 33% of the time and 26% of the time for motif 4. MEME implements a normalization procedure to account for local nucleotide biases within submitted sequences [1]. This normalization restrains the algorithm from being trapped in repetitive nucleic regions. Caution should be taken with motifs that are found in CpG islands for follow-up wet-lab validation, since MFAs tend to converge to these sites since they have high information content. The motifs that are not highly present in CpG islands should be prioritized for validation since they are less likely to contain repetitive nucleotide elements.

The MEME package suite has an integrated application called TOMTOM [24]. This application aligns any position weight matrix against publicly available databases (e.g. JASPAR [7] or TRANSFAC [60]) of known motifs for DNA binding proteins. TOMTOM provides q and p values to determine the statistical significance of such an alignment occurring in the databases. The q value is the false discovery rate.

Motifs with a high number of nucleotides with low information content (e.g. Table 4-5 (motif 1, 3) and Table 4-6 (motif 2)) show a high number of significant matches. Because the information content of these motifs is low, preference was given to the other motifs. Table 4-7 shows the results of TOMTOM for the motifs identified in Table 4-5 and Table 4-6. Only the hits with a p -value and q -value of < 0.05 were retained in this analysis.

Table 4-5: Results of the MEME analysis using all the sequences in the TRE chip-chip dataset ranked by *E*-values.

Motif number	Number of sites found in dataset	Percentage of sites in CpG island	Sequence Logo of motif found by MEME
1-	50	92%	
2-	33	33%	
3-	52	84%	

Table 4-6: Results of the MEME analysis using TOP5-PND4&TOP25-PND15 sequences in the TRE chIP-chip dataset ranked by *E*-values

Motif number	Number of sites found in dataset	Percentage of sites in CpG island	Sequence Logo for motif found by MEME
1-	50	10%	<p>MEME (no 55C) 01 02 10 14 39</p>
2-	31	0%	<p>MEME (no 55C) 01 02 10 14 40</p>
3-	40	15%	<p>MEME (no 55C) 01 02 10 14 40</p>

4-	9	20%	<p>bits</p> <p>MEME (no GGC) 01 02 10 14 41</p>
5-	30	13%	<p>bits</p> <p>MEME (no GGC) 01 02 10 14 42</p>
6-	43	2%	<p>bits</p> <p>MEME (no GGC) 01 02 10 14 43</p>

Table 4-7: Motif targets using TOMTOM for a few example cases

Motif ID	Motif database	Target name from database	Evidence in literature of relationship with THR	Note
Table 4-5 - 2	Jaspar	RREB1	None found	
Table 4-5 - 4 Table 4-6 - 3	Transfac	SP1	Yes [33]	Motif has been found in both entire set and subset.
Table 4-5 - 5 Table 4-6 - 4	Transfac	PAX6	No, but PAX6 is known to be a major actor for the development of sensory organs.	The motif alignment was performed on the extreme right hand of the logo.
Table 4-5 - 6	Transfac	SMAD3	None found	

The results showed that some of the motifs were similar to motifs of other transcription factors. For example, motifs Table 4-5 – 4 and Table 4-6 - 3 are highly similar to the SP-1 motif. Table 4-8 shows respectively the logos found by MEME (Table 11 – 4 and Table 12 - 3) in the first column and the SP-1 motif in the Transfac database in column 2.

Table 4-8: Motif found by MEME (Table 4-5 – 4 and Table 4-6 - 3) in the first column and the SP-1 motif in the Transfac database in column 2

In this case the SP-1 motif is especially interesting since a link to TREs is supported by the literature. Kim *et al.* [33] found that SP1 binds to a highly similar consensus sequence just shown above. SP-1 is a known THR coregulator that enhances gene expression. Although targeted genes for these binding sites cannot be identified here, genes which show high up regulation in gene expression studies could be potential targets for this co-regulator. Since there are no parallel gene expression data available at this time, such work could certainly be accomplished. Complete information on the location of these motifs is available upon request, but is too much information to be contained in this thesis.

4.2.2.3. Using BioProspector to find novel motifs

The application Bioproprospector was chosen to further mine the TRE data for novel motifs because it performed very well overall in the review by Hu *et al.* [28], and its high performance capabilities on larger datasets. Bioproprospector was initially run with the default parameters, with the exception of a proprietary format that included the

distribution of background sequences (i.e., sequences mentioned at the beginning of section 4.2.1.1). The length of the desired motif was set to 10 and the number of iterations to search for each motif instance was set to 240 as suggested by Liu et al. [36]. As with the MEME results above, these parameters also led to results encompassing motifs exhibiting a very high bias of CpG islands. Although the algorithm permits the discovery of more than one motif, convergence toward the same nucleotide position was found as the algorithm was given the liberty of finding 25 motifs (resulting in some similarities across motifs, and thus multiple hits at the same site). Unlike MEME, Bioproprospector doesn't mask motifs in DNA once they have been found.

Table 4-9 presents the top 3 motifs found by Bioproprospector. Instances are primarily located in CpG islands; the other motifs are not shown since their logos are quite similar and the motif locations were mostly redundant. Table 4-10 shows motifs for which interesting TOMTOM hits have been found. Since the motifs exhibit a high level of degeneracy, many similar existing motifs were found. Therefore the literature was examined and references are presented below.

Table 4-9: Top 3 motifs found by Bioprospector and percentage of motif hit in CpG islands

Motif number	Number of sites found in dataset	Percentage of sites in CpG island	Sequence logo of motif found by Bioprospector
1-	205	70%	<p>Sequence logo for motif 1- showing high information content at positions 3, 4, 5, 6, 8, and 9. The motif is approximately CGGC CCG.</p>
2-	268	64%	<p>Sequence logo for motif 2- showing high information content at positions 1, 2, 3, 4, 5, 7, 8, 9, and 10. The motif is approximately GG GGG GGG.</p>
3-	280	65%	<p>Sequence logo for motif 3- showing high information content at positions 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. The motif is approximately GG GGGG GGG.</p>

Table 4-10: Motif targets using TOMTOM for a few example cases

Motif ID	Database of motif	Target name in database	Evidence in literature of relationship with THR	Note
Table 4-9-1	Transfac	KROX	Yes [23]	Although this is not proof that KROX is a coregulator of THR, they are closely related in the development of mice.
Table 4-9-2	Transfac	SP-1	Yes [33]	Similar motif has been found by MEME.

Thus, Bioprosector has identified a few additional potential motifs. However, due to the nature and implementation of the algorithm, there is a tendency towards finding multiple maxima at a single location, resulting in multiple motifs reported at a single location. One potential improvement would be to modify the code of Bioprosector in order to have the algorithm mask each motif after each motif instance selection. This could potentially lead to finding motifs that are more degenerate and avoid reporting multiple motifs at a single location.

4.2.2.4. Employing Weeder to find novel motifs

Weeder was selected for further TRE motif mining based of the extensive study presented by Tompa *et al.* [55]. These authors suggest that Weeder performs very well according to various indicators in its collective analysis on benchmark data. They hypothesize that Weeder's good results are due to its relatively conservative approach for determining whether sequences contain a motif or not.

Weeder's parameters were set to search for a consensus sequence of length 6 with 1 mutation, length 8 with 2 mutations, length 10 with 3 mutations and length 12 with 4 mutations on both DNA strands. Unlike the 2 previous MFAs, Weeder has its own format of background distribution files and no tools to create one. Therefore, the Weeder pre-calculated *mus musculus* (MM) background model was employed for this experiment.

Table 4-11 shows the top 4 motifs of length 8 and 10 respectively. Table 4-12 displays the motifs found by TOMTOM when searching for known transcription factor binding site consensus.

Table 4-11: Top 4 motifs found by Weeder and percentage of motifs in CpG islands

Motif number	Number of sites found in dataset	Percentage of sites in CpG island (%)	Sequence logo for motif identified by Weeder
1-	120	26	<p>Sequence logo for motif identified by Weeder</p>
2-	119	23	<p>Sequence logo for motif identified by Weeder</p>

Table 4-12: Motif targets using TOMTOM for a few example cases

Motif ID	Database of motif	Target description in database	Evidence in literature of relationship with THR	note
Table 4-11 -1	Transfac	SMAD	none	Very low p-value but E-value @ 1.9, i.e frequent in background
Table 4-11 -2	Transfac	PAX6	No, but PAX6 is known to be a major actor for the development of sensory organs.	With TOMTOM, the motif alignment was performed on the extreme right hand of the logo.
Table 4-11 -3	Transfac	NRSF	none	Motif discovered by Weeder

				aligns with half of the NRSF motif. There is evidence that the neural-restrictive silencing factor is active in the murine cerebellum.
Table 4-11 -4	Transfac	PCF2	None	Very low p-value but E-value @ 3.2, i.e. frequent in background

Although Weeder identified motifs of length 6 (not shown here), the number of instances in the chIP-chip sequences was so high that they are not presented. Hu *et al.* also demonstrated that the true positive discovery rate for motifs of length 6 is very low. Thus, we believe that a very high proportion of these would be false positives.

Weeder identified one common transcription factor with MEME: PAX6. One observation with Weeder is that the q-values for TOMTOM are much higher than with the other MFAs.

4.2.3. Summary of the utilization of MFAs

The MFAs presented here found several transcription factors for which evidence of interaction with THR has previously been demonstrated. As an indication of the possible interactions between the transcription factor binding sites, Figure 4-8 presents the relative location of the transcription factors. This figure is a representation of DNA sequences for peak ID ≤ 5 for PND4 and peakID ≤ 10 for PND15.

There are many noteworthy features that are evident in this figure. First, the combination of MFAs and TOMTOM do not agree with one another; i.e., the SP-1 binding sites have different locations in MEME versus Bioprosector. This could lead to a potential conclusion that once a novel consensus site has been found by an MFA, sequences should be scanned for the motif in order to verify if other motifs with high information content (bits) coexist. MFAs target the motifs based on the information content and therefore very similar motifs could be found in the same set of sequences for which the information content is slightly inferior.

Surprisingly, AP-1 consensus binding sites were also found in sequences containing TRE consensus binding sites, even though AP-1 is supposed to regulate gene transcription independently of THR. One hypothesis for this observation is that the genes that are regulated by these transcription factors play a very important biological role in development and are co-ordinately regulated.

Sample of Motifs found and their locations on TRE chip-chip peaks

Figure 4-8: Locations of binding motifs in a few test cases

(A = TRE, B = AP-1, C = Bipad site, D = Meme SP-1, E = MEME PAX6, F = Bioproductor KROX, G = Bioproductor SP-1, H = Weeder PAX6)

Interestingly, none of the sites identified by Bipad (2 half sites here displayed by 'C' in Figure 4-8) overlap with instances of the TRE consensus motif ('A' in Figure 4-8). The Bipad algorithm is supposed to find half sites with high information content. One could hypothesize that the consensus of the TRE binding site for the half site is very degenerate and therefore a consensus would be difficult to find. Further discussion is presented on this in section 5.3

Some possible interference from applying MFAs may result from CpG islands because MFAs converge on sites with high information content. Investigators should be aware of the bias of certain motifs towards CpG islands before selecting motifs for wet-lab validation.

Hu *et al.* [28] developed an algorithm called “Ensembl” to reveal consensus from many MFAs. These authors observed an increase in specificity when applying this tool on benchmark data. In the analysis of the TRE data, the application of such an algorithm may have revealed a few similar consensus sequences although the binding site locations predicted by each algorithm were quite different. As such, many of the reviews of MFAs described above recommend wet-lab validation of the top few motifs found by MFAs that explore DNA sequences using different underlying statistical models.

The lists of locations of the motifs presented above are available upon request. In order to eliminate unnecessary appendices, they are not presented here.

4.3. Summary

The questions posed at the beginning of this chapter are re-iterated below. The conclusions drawn in answer to these questions are indicated in *italics*.

For the consensus hexamer (AGGTCA)

- Does the number of half sites in a gene promoter region correlate with the presence of TREs? If so, can this be used to differentiate between true and false GROIs?

Although there is a slight difference in the median distribution of the number of half sites between random genes and THR regulated genes, the distributions still show a large overlap. Therefore, it is unlikely that any classifier performs well on this data.

- How is the distribution and number of hexamers in the GROIs different from: randomly selected genes, nucleotide permutation of these promoter regions, promoter regions of genes known to contain a TRE, and genes that are regulated by a nuclear receptor containing the same half site?

Figure 4-3 shows the different distribution of the frequency of half sites per kilo base pairs. Although some minor differences in the distributions exist, the differences do not appear to be significant. Thus, although there are no definitive conclusions possible at this time, the data strongly suggest that there are no differences in half site distribution among true promoters.

For the consensus dimer

- When a DR4 is found, is the gene it regulates on the same strand as the motif?

This answer can be found in Table 2-2 where the strand on which the TRE resides and the strand of the corresponding genes is displayed. High correlation can be found between the 2 factors. However, one must keep in mind that this hypothesis can only be studied for DRs where palindromes can in fact be read in both directions (5'-3' and 3'-5').

- Is there a bias in the location of the motif with respect to the TSS of the nearest gene that would permit a certain probability of a motif to be a true positive?

There are currently insufficient validated data to answer this question. Wet-lab validation is recommended to build confidence conclusions drawn from the data. Classification with respect to the location of the binding motifs and gene transcription start sites should be explored.

Although this is briefly discussed in section 5.3, an interesting potential validation step that should follow this analysis is the evaluation of gene expression changes in the same tissues following perturbation of thyroid hormone levels. The pairing of transcription factor binding with changes in gene expression following exposure to regulators of the transcription factors is a powerful approach to identify promoter regions with true binding sites.

Chapter 5. Conclusion

5.1. Summary of Research

This thesis presents a framework for the identification of TRE candidates *in silico*. In Chapter 2 an overview of the biology underlying TH gene regulation was presented. Wet laboratory techniques used to validate functional TREs were described in order to facilitate and explain the nature and provenance of the data. The wet-lab experimental design for the present dataset (chIP-chip) was explained in detail to provide a foundation for interpreting the subsequent in-silico experiments that comprise the bulk of this thesis.

In Chapter 3 various normalization methods were investigated to examine their performance and suitability to apply to the chIP-chip data. A benchmark spike-in study was used to determine the optimal parameters in order to obtain the best (compromise) approach to yield the highest potential sensitivity and precision. Ultimately, no normalization was applied and Splitter was used to search for peaks. Application of this approach to the chIP-chip dataset resulted in the identification of 230 chromosomal positions exhibiting a peak indicating enrichment in the IP channel.

In Chapter 4, applying the known classical TRE half site, an analysis of the density and presence of TRE half sites in genomic sequences was undertaken to determine whether it was possible to discriminate a region containing a TRE from one without. A PWM model was then established to scan the chIP-chip peak sequences and reveal the precise location of TRE and AP-1 candidates. From the TREs found in the literature, it was apparent that in most cases, one TRE half site score was much lower than the second half site. I hypothesized that this may be the result of co-factor binding alongside the thyroid hormone receptor (e.g., RXR). With the {min, max}Scoring model, a scan of all of the identified chIP-chip peaks revealed 15 potential TREs in 44 peaks for PND 4 and 80 potential TREs in PND 15. Because of demonstrated laboratory evidence for the

involvement of AP-1, this analysis was also undertaken for the AP-1 sites. This analysis resulted in 69 potential binding sites for AP-1.

Besides the classical TRE half site motif, novel motif finding algorithms were also used in a strategy to find transcription factor binding sites. The employed strategy combined highly similar results from more than one MFA in order to identify any consensus that could be observed. Several motifs that did not reveal any match to the Transfac and Jaspar databases were also found, identifying potentially novel sequence motifs associated with TR α binding. All the records for potential binding site locations will be shared with a biologist who will conduct wet-laboratory experiments in order to reveal if instances of these motifs induce gene regulatory function.

5.2. Major Conclusions

5.2.1. Normalization

- Normalization methods for gene expression arrays are not appropriate for chIP-chip data due to violation of underlying assumptions.
- Normalization techniques developed for chIP-chip failed to address all sources of bias in our study.
- Potter *et al.* algorithm [44] was modified to address GC-content probe bias.

5.2.2. Peak finding

- Analysis of gold standard spike-in dataset from Johnson *et al* [29] indicates that Splitter algorithm [20] is most appropriate for current experimental platform (Agilent, 60bp probe length).
- Application of Splitter algorithm to Dr. Dong's dataset resulted in 230 high confidence peaks in PND4 and PND15.

5.2.3. TRE consensus motif searching

- PWM scoring of individual half sites exhibit statistically significant difference between two half sites comprising dimer TRE motif.
- LOO cross-validation indicates a false positive rate of 0.2 when {min, max} scoring model is applied to a test dataset of known positive (i.e. containing TRE) and negative DNA sequences.
- No evidence that density or position distribution of TRE half sites within promoter regions are correlated with reported THR gene regulation.

5.2.4. Novel TRE motif searching

- Application of MEME [1], BioProspector [36], and Weeder [41] resulted in discovery of novel motifs in GROIs (i.e. genomic regions pulled down during chIP-chip experiment).
- Analysis using TOMTOM tool [24] indicated that several novel motifs were associated with known TF binding sites. One of these, SP-1, has been identified in the literature as enhancing expression of known THR regulated genes [33].

5.3. Future work

A number of very interesting ideas emerged from this study that warrant further investigation. In particular, the search for transcription factor binding sites for gene regulation could be improved:

- The Bipad algorithm could be modified to search for binding sites that are composed of more than two half sites (see first row in Table 2-2).
- With Bipad, only the top consensus is returned. Modifications to the algorithm to output more than one logo could potentially be beneficial since true binding site motifs are not always the ones having the highest information content.
- A strategy of masking could also be implemented in Bioprospector in order to have fewer repetitions in consensus motifs.

- Classification of binding sites with respect to gene expression data could also be conducted. In other words, differentiation of up- versus down-regulated genes to identify motifs specific to these changes.

A number of interesting ideas also emerged for biologists with respect to wet-lab experiments:

- ChIP-chip could be performed using a different antibody to target a known THR partner (e.g. RXR), this could improve the motif searching confidence where classical TREs are more likely to be found if chIP-chip revealed positive signals on both chIP-chip TRE and chIP-chip RXR datasets.
- Later, when TREs are further characterized, chIP-chip could be motif driven, where chIP-chip would map specific genomic regions where TRE motif instances were found in order to determine whether they are functional or not.

Bibliography

- [1] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.
- [2] J. H Duncan Bassett, Clare B Harvey, and Graham R Williams. Mechanisms of thyroid hormone receptor-specific nuclear and extra nuclear actions. *Mol Cell Endocrinol*, 213(1):1–11, Dec 2003.
- [3] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic Acids Res*, 37(Database issue):D26–D31, Jan 2009.
- [4] Chengpeng Bi and Peter K Rogan. Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res*, 32(17):4979–4991, 2004.
- [5] Ben Bolstad. Probe level quantile normalization of high density oligonucleotide array data. *unpublished*, N/A:N/A, 2001.
- [6] UCSC Genome Browser. Lift genome annotations. <http://genome.ucsc.edu/cgi-bin/hgLiftOver>.
- [7] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–D106, Jan 2008.
- [8] Simon Cawley, Stefan Bekiranov, Huck H Ng, Philipp Kapranov, Edward A Sekinger, Dione Kampa, Antonio Piccolboni, Victor Sementchenko, Jill Cheng, Alan J Williams, Raymond Wheeler, Brant Wong, Jorg Drenkow, Mark Yamanaka, Sandeep Patel, Shane Brubaker, Hari Tammana, Gregg Helt, Kevin Struhl, and Thomas R Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116(4):499–509, Feb 2004.
- [9] Kevin M Crofton. Thyroid disrupting chemicals: mechanisms and mixtures. *Int J Androl*, 31(2):209–223, Apr 2008.

- [10] Modan K Das and Ho-Kwok Dai. A survey of dna motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007.
- [11] P. A. Dawson and D. Markovich. Regulation of the mouse *nas1* promoter by vitamin d and thyroid hormone. *Pflugers Arch*, 444(3):353–359, Jun 2002.
- [12] Robert J Denver and Keith E Williamson. Identification of a thyroid hormone response element in the mouse *kruppel-like factor 9* gene to explain its postnatal expression in the brain. *Endocrinology*, 150(8):3935–3943, Aug 2009.
- [13] B. Desvergne. How do thyroid hormone receptors bind to structurally diverse response elements? *Mol Cell Endocrinol*, 100(1-2):125–131, Apr 1994.
- [14] Patrik D'haeseleer. How does dna sequence motif discovery work? *Nat Biotechnol*, 24(8):959–961, Aug 2006.
- [15] Hongyan Dong, Carole L Yauk, Andrew Williams, Alice Lee, George R Douglas, and Michael G Wade. Hepatic gene expression changes in hypothyroid juvenile mice: characterization of a novel negative thyroid-responsive element. *Endocrinology*, 148(8):3932–3940, Aug 2007.
- [16] M. Downes, R. Griggs, A. Atkins, E. N. Olson, and G. E. Muscat. Identification of a thyroid hormone response element in the mouse *myogenin* gene: characterization of the thyroid hormone and retinoid x receptor heterodimeric binding site. *Cell Growth Differ*, 4(11):901–909, Nov 1993.
- [17] A. Farsetti, B. Desvergne, P. Hallenbeck, J. Robbins, and V. M. Nikodem. Characterization of myelin basic protein thyroid hormone response element and its function in the context of native and heterologous promoter. *J Biol Chem*, 267(22):15784–15788, Aug 1992.
- [18] A. Farsetti, T. Mitsuhashi, B. Desvergne, J. Robbins, and V. M. Nikodem. Molecular basis of thyroid hormone regulation of myelin basic protein gene expression in rodent brain. *J Biol Chem*, 266(34):23226–23232, Dec 1991.
- [19] Paul Flicek, Bronwen L Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha,

Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, James Smith, and Stephen M J Searle. Ensembl's 10th year. *Nucleic Acids Res*, 38(Database issue):D557–D562, Jan 2010.

[20] Yutao Fu. Oligo-tiling array signal processing by splitter. <http://zlab.bu.edu/yf/anchor/web/splitter.cgi?step=0>.

[21] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K. D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R. Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, Nov 2009.

[22] M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *J Mol Biol*, 196(2):261–282, Jul 1987.

[23] M. T. Ghorbel, I. Seugnet, N. Hadj-Sahraoui, P. Topilko, G. Levi, and B. Demeneix. Thyroid hormone effects on krox-24 transcription in the post-natal mouse brain are developmentally regulated but are not correlated with mitosis. *Oncogene*, 18(4):917–924, Jan 1999.

[24] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biol*, 8(2):R24, 2007.

[25] Koshi Hashimoto, Emi Ishida, Shunichi Matsumoto, Shuichi Okada, Masanobu Yamada, Teturo Satoh, Tsuyoshi Monden, and Masatomo Mori. Carbohydrate response

element binding protein gene expression is positively regulated by thyroid hormone. *Endocrinology*, 150(7):3417–3424, Jul 2009.

[26] Koshi Hashimoto, Shunichi Matsumoto, Masanobu Yamada, Teturo Satoh, and Masatomo Mori. Liver x receptor-alpha gene expression is positively regulated by thyroid hormone. *Endocrinology*, 148(10):4667–4675, Oct 2007.

[27] Koshi Hashimoto, Masanobu Yamada, Shunichi Matsumoto, Tsuyoshi Monden, Teturo Satoh, and Masatomo Mori. Mouse sterol response element binding protein-1c gene expression is negatively regulated by thyroid hormone. *Endocrinology*, 147(9):4292–4302, Sep 2006.

[28] Jianjun Hu, Bin Li, and Daisuke Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*, 33(15):4899–4913, 2005.

[29] David S Johnson, Wei Li, D. Benjamin Gordon, Arindam Bhattacharjee, Bo Curry, Jayati Ghosh, Leonardo Brizuela, Jason S Carroll, Myles Brown, Paul Flicek, Christoph M Koch, Ian Dunham, Mark Bieda, Xiaoqin Xu, Peggy J Farnham, Philipp Kapranov, David A Nix, Thomas R Gingeras, Xinmin Zhang, Heather Holster, Nan Jiang, Roland D Green, Jun S Song, Scott A McCuine, Elizabeth Anton, Loan Nguyen, Nathan D Trinklein, Zhen Ye, Keith Ching, David Hawkins, Bing Ren, Peter C Scacheri, Joel Rozowsky, Alexander Karpikov, Ghia Euskirchen, Sherman Weissman, Mark Gerstein, Michael Snyder, Annie Yang, Zarmik Moqtaderi, Heather Hirsch, Hennady P Shulha, Yutao Fu, Zhiping Weng, Kevin Struhl, Richard M Myers, Jason D Lieb, and X. Shirley Liu. Systematic evaluation of variability in chip-chip experiments using predefined dna targets. *Genome Res*, 18(3):393–403, Mar 2008.

[30] W. Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X. Shirley Liu. Model-based analysis of tiling-arrays for chip-chip. *Proc Natl Acad Sci U S A*, 103(33):12457–12462, Aug 2006.

[31] Sunduz Keles. Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, 63(1):10–21, Mar 2007.

[32] Josef Köhrle. Iodothyronine deiodinases. *Methods Enzymol*, 347:125–167, 2002.

[33] M. K. Kim, J. S. Lee, and J. H. Chung. In vivo transcription factor recruitment during thyroid hormone receptor-mediated activation. *Proc Natl Acad Sci U S A*, 96(18):10092–10097, Aug 1999.

- [34] Vincent Laudet and Hinrich Gronemeyer. *The Nuclear Receptor Factsbook*. Academic Press London, 2002.
- [35] Mitchell A Lazar. Thyroid hormone action: a binding contract. *J Clin Invest*, 112(4):497–499, Aug 2003.
- [36] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–138, 2001.
- [37] Brenda J Mengeling, Fan Pan, and Martin L Privalsky. Novel mode of deoxyribonucleic acid recognition by thyroid hormone receptors: thyroid hormone receptor beta-isoforms can bind as trimers to natural response elements comprised of reiterated half-sites. *Mol Endocrinol*, 19(1):35–51, Jan 2005.
- [38] NIH. www.ncbi.nlm.nih.gov/pubmed.
- [39] Patrick J O'Shea, Celine J Guigon, Graham R Williams, and Sheue yann Cheng. Regulation of fibroblast growth factor receptor-1 (fgfr1) by thyroid hormone: identification of a thyroid hormone response element in the murine fgfr1 promoter. *Endocrinology*, 148(12):5966–5976, Dec 2007.
- [40] Dmitri A Papatsenko, Vsevolod J Makeev, Alex P Lifanov, Mireille Régnier, Anna G Nazina, and Claude Desplan. Extraction of functional binding sites from unique regulatory regions: the drosophila early developmental enhancers. *Genome Res*, 12(3):470–481, Mar 2002.
- [41] Giulio Pavesi, Paolo Mereghetti, Federico Zambelli, Marco Stefani, Giancarlo Mauri, and Graziano Pesole. Mod tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res*, 34(Web Server issue):W566–W570, Jul 2006.
- [42] Shouyong Peng, Artyom A Alekseyenko, Erica Larschan, Mitzi I Kuroda, and Peter J Park. Normalization and experimental design for chip-chip data. *BMC Bioinformatics*, 8:219, 2007.
- [43] S. P. Porterfield and C. E. Hendrich. The role of thyroid hormones in prenatal and neonatal neurological development—current perspectives. *Endocr Rev*, 14(1):94–106, Feb 1993.

- [44] Dustin P Potter, Pearly Yan, Tim H M Huang, and Shili Lin. Probe signal correction for differential methylation hybridization experiments. *BMC Bioinformatics*, 9:453, 2008.
- [45] G. Pérez-Juste, S. García-Silva, and A. Aranda. An element in the region responsible for premature termination of transcription mediates repression of c-myc gene expression by thyroid hormone in neuroblastoma cells. *J Biol Chem*, 275(2):1307–1314, Jan 2000.
- [46] Brooke Rhead, Donna Karolchik, Robert M Kuhn, Angie S Hinrichs, Ann S Zweig, Pauline A Fujita, Mark Diekhans, Kayla E Smith, Kate R Rosenbloom, Brian J Raney, Andy Pohl, Michael Pheasant, Laurence R Meyer, Katrina Learned, Fan Hsu, Jennifer Hillman-Jackson, Rachel A Harte, Belinda Giardine, Timothy R Dreszer, Hiram Clawson, Galt P Barber, David Haussler, and W. James Kent. The ucsc genome browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, Jan 2010.
- [47] P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6):276–277, Jun 2000.
- [48] T. Satoh, T. Monden, T. Ishizuka, T. Mitsuhashi, M. Yamada, and M. Mori. Dna binding and interaction with the nuclear receptor corepressor of thyroid hormone receptor are required for ligand-independent stimulation of the mouse preprothyrotropin-releasing hormone gene. *Mol Cell Endocrinol*, 154(1-2):137–149, Aug 1999.
- [49] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.
- [50] Dong-Ju Shin, Michelina Plateroti, Jacques Samarut, and Timothy F Osborne. Two uniquely arranged thyroid hormone response elements in the far upstream 5' flanking region confer direct thyroid hormone regulation to the murine cholesterol 7alpha hydroxylase gene. *Nucleic Acids Res*, 34(14):3853–3861, 2006.
- [51] Gemma Solanes, Neus Pedraza, Verónica Calvo, Antonio Vidal-Puig, Bradford B Lowell, and Francesc Villarroya. Thyroid hormones directly activate the expression of the human and mouse uncoupling protein-3 genes through a thyroid response element in the proximal promoter region. *Biochem J*, 386(Pt 3):505–513, Mar 2005.

- [52] Jun S Song, W. Evan Johnson, Xiaopeng Zhu, Xinmin Zhang, Wei Li, Arjun K Manrai, Jun S Liu, Runsheng Chen, and X. Shirley Liu. Model-based analysis of two-color arrays (ma2c). *Genome Biol*, 8(8):R178, 2007.
- [53] R Development Core Team. R: A language and environment for statistical computing. 2009. ISBN 3-900051-07-0.
- [54] Andrija Tomovic and Edward J Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, Apr 2007.
- [55] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W. James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, Jan 2005.
- [56] C. R. Tyler, S. Jobling, and J. P. Sumpter. Endocrine disruption in wildlife: a critical review of the evidence. *Crit Rev Toxicol*, 28(4):319–361, Jul 1998.
- [57] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004.
- [58] Isabelle Weinhofer, Markus Kunze, Heidelinde Rampler, Sonja Forss-Petter, Jacques Samarut, Michelina Plateroti, and Johannes Berger. Distinct modulatory roles for thyroid hormone receptors α and β in srebp1-activated abcd2 expression. *Eur J Cell Biol*, 87(12):933–945, Dec 2008.
- [59] Graham R. Williams and Gregory A. Brent. *Thyroid Hormone Response Elements*, chapter 15, pages 217–239. Raven Press, 1995.
- [60] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer. Transfac: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–319, Jan 2000.
- [61] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, Feb 2002.

[62] P. M. Yen. Physiological and molecular basis of thyroid hormone action. *Physiol Rev*, 81(3):1097–1142, Jul 2001.

Appendix A

Tables presented here show the analysis results performed for different chIP-chip cut offs (see Figure 3-11) with their respective standard deviation values.

	WhiteHead, 2.5 SD		
	SP	CNCL + SP	CNCL2 + SP
Number of True Positives	54	54	67
Number of False Positives	2	2	30
Number of False Negatives	46	46	33
Total number of sites found	56	56	99
Sensitivity	0.54	0.54	0.67
Precision	0.96	0.96	0.67

	WhiteHead 2.25 SD		
	SP	CNCL + SP	CNCL2 + SP
Number of True Positives	65	65	70
Number of False Positives	8	8	84
Number of False Negatives	35	35	30
Total number of sites found	73	73	154
Sensitivity	0.65	0.65	0.70
Precision	0.89	0.89	0.45

	WhiteHead 2.75 SD		
	SP	CNCL + SP	CNCL2 + SP
Number of True Positives	49	49	62
Number of False Positives	1	1	8
Total number of False Negatives	51	51	38
Total number of sites found	50	50	71
Sensitivity	0.49	0.49	0.62
Precision	0.98	0.98	0.88

Appendix B

Appendix B shows the peaks found by applying the strategy elaborated in section 3.5. The 2 tables below are from PND4 and PND15 and include the following columns; the chromosome of the peak, the starting location of the peak, the corrected ending location of the peak, the number of probes in a peak, the gap value for the peak (highest distance between probes of a peak), the median signal intensity across probes in the peak and the validation results (NV = not validated, TP = true positive, FP = false positive (see section 3.7)), the location of the peak with respect to the closest mRNA mapped by the UCSC genome browser on MM9 genome build (e = exon, i = intron, o = outside), the ID of the mRNA mapped, the mRNA symbol, the distance from the closest mRNA for outside peaks, the number of references relating the mRNA symbol to the terms “thyroid hormone” on Pubmed.

PND4 Sites

Peak ID	Chr	Start	End Corrected	N probes	Gapmax	Signal	Validation	Location	mRNA ID	mRNA symbol	Distance	No. of refs
1	chr3	126700734	126701109	2	124	2.224	NV	e25 e17 e16 e12	NM_178655 AK036018 AK082352 AK044634	Ank2 Ank2 Ank2 Ank2		0
2	chr2	174105864	174106699	3	337	2.224	NV	o	AK039035	Gnas	3121	18
3	chr8	98221226	98221816	2	340	2.171	NV	ij	AK129305	mKIAA1180		0
4	chr7	135357520	135357969	2	202	2.063	NV	o	NM_009943	Cox6a2	8089	0
5	chr2	131394967	131395499	2	272	1.988	NV	o	NM_013460	Adra1d	7478	0
6	chr1	93313967	93314423	2	208	1.983	TP	e2	NM_011066	Per2		6
7	chr1	173150329	173150817	2	238	1.978	NV	e5 e5	BC058978 NM_001037170	Tommm40 Tommm40		0
8	chr7	150192500	150193024	2	276	1.974	NV	ij ij e4 e3	AK039657 AK041748	AK039657 Tspan32		0

9	chr8	87489815	87490307	2	244	1.957	NV	o	AY039000 NM_020286 AJ279792 AJ279793 AJ279794 AJ279795	Tspan32 Tspan32 Tspan32 Tspan32 Tspan32 Phemx	397	0
10	chr2	93476336	93477043	2	456	1.932	NV	o	NM_007442	Aix4	5547	0
11	chr3	87946878	87947459	2	336	1.905	NV	ij ij ij	NM_133665 S68893 S68895 AK141270	Mef2d Mef2d Mef2d Mef2d		3
12	chr2	127345397	127346114	2	464	1.902	NV	ij ij ij	NM_019789 AF287732 AK016721	Kcnp3 Kcnp3 DREAM/ calsenilin/ KChIP3		1
13	chr2	24896352	24896934	2	324	1.888	NV	e25 e24 e5 e3	NM_146251 BC021312 EU046569 BC027342	Pnp1a7 Pnp1a7 Pnp1a7 Pnp1a7		0
14	chr8	71318206	71318736	2	281	1.879	NV	e3 e3 e3 e3 e3 e3	AK169819 AK083331 AK047093 NM_027590 AK147743 AK034498	Ints10 Ints10 Ints10 Ints10 Ints10 Ints10		0
15	chr7	148134128	148134628	2	249	1.873	NV	o	NM_026820	Ifitm1	18496	0
16	chr7	28964046	28964516	2	213	1.858	NV	ij ij	NM_010092 NM_001037957	Dyrk1b Dyrk1b		0
17	chr6	112340461	112340825	2	113	1.832	NV	o	AK054244	D630042P16Rik	13104	0
18	chr7	138081260	138081620	2	109	1.821	NV	ij ij	NM_019564	Htra1		0
19	chr1	52785706	52786284	2	324	1.819	NV	o	AK171946	2210010L05Rik	2122	0
20	chr1	172569257	172569727	2	222	1.801	NV	o	NM_177068	Olfml2b	5236	0
21	chr7	86042576	86042939	2	118	1.796	NV	ij ij ij	NM_026531 AK089090 AK017644 BC022614	Isg2011 Isg2011 Isg2011 Isg2011		0
22	chr7	80520916	80521288	2	127	1.783	NV	ij ij	NM_177740 BC059072	Rgma Rgma		0
23	chr7	17477279	17478002	2	465	1.773	NV	e2 ij ij	NM_010320	Gng8		0
24	chr15	74509157	74510173	3	476	1.990	TP	o	AK157822	Arc	7173	37
25	chr11	98208012	98208442	2	176	1.921	NV	o	NM_144828	Ppp1r1b	1609	1
26	chr19	7345196	7345900	2	452	1.895	NV	e7 e6 e5	AK052295 BC055719 NM_054048	Rcor2 Rcor2 Rcor2		1
27	chr15	76519605	76519975	2	122	1.835	NV	e3 ij ij	AK193686 NM_033371 AK220248	Ppp1r16a Ppp1r16a Ppp1r16a		1

28	chr17	43706958	43707376	2	168	1.783	NV	ij ij	AK155678 AK005158 NM_013737	Pla2g7 Pla2g7 Pla2g7		0
29	chr15	79852833	79853205	2	126	1.752	NV	o	AK087840	Pdgb	7967	3
30	chr15	89253902	89254550	2	398	1.682	NV	e14 e14	NM_009948 AK146851	Cpt1b Cpt1b		0
31	chr17	28338616	28338982	2	120	1.671	NV	ij	NM_172617	Zfp523		0
32	chr10	57514939	57515347	2	156	1.658	TP	ij	NM_020561	Smpd13a		0
33	chr11	72854005	72854643	2	384	1.633	NV	e17	NM_018883	Camkk1		0
34	chr10	77813542	77813992	2	206	1.632	NV	ij ij	NM_053014 AK008965	Agpat3 Agpat3		0
35	chr10	126718983	126719561	2	330	1.625	NV	e2 ij ij	BC038259 NM_033072 AK158797	Mbd6 Mbd6 Mbd6		0
36	chr9	108405410	108406012	2	356	1.574	NV	o	NM_133794	Qars	4392	0
37	chr14	121309436	121309972	2	292	1.574	NV	o	AK017701	Ranbp5	443	0
38	chr10	60299885	60300475	2	338	1.552	NV	o	BC057560	Unc5b	6146	0
39	chr9	69246049	69246527	2	222	1.529	NV	ij ij ij	AK042916 AK041772 AK163917 NM_145618	Narg2 Narg2 Narg2 Narg2		0
40	chrX	7325659	7326157	2	254	1.499	NV	o	NM_207202	Ccdc120	7707	0
41	chr15	101090429	101091023	2	348	1.489	NV	o	NM_010444	Nr4a1	6269	92
42	chr17	63233959	63234321	2	110	1.471	NV	o	NM_010109	Efna5	3655	0
43	chr19	38338078	38338774	2	450	1.460	NV	o	AK035873	AK035873	528	0
44	chr14	22319784	22320336	3	178	1.448	NV	ile2 e2 e2 e2 ij	AK052307 AK048336 AK135465 AK083123 AK045188 NM_017479	Myst4 Myst4 Myst4 Myst4 Myst4 Myst4		0

PND15 Sites

Peak ID	Chr	Start	End Corrected	Nprobes	Gap	Signal	Validation	Location	mRNAID	mRNA symbol	Distance	No. of refs
1	chr7	127344457	127344897	2	195	2.689	TP	e8 e3	NM_016669 AK163801	Crym Crym		9
2	chr8	97537009	97537499	2	243	2.631	NV	e15 e15	AK087268 AF166382	Gpr56 Gpr56		1

3	chr1	82831717	82832402	2	439	2.601	NV	e15 e14	NM_018882 AK167547	Gpr56 Gpr56	3655	0
4	chr8	73034376	73034742	2	121	2.555	NV	o	NM_010472	Hrb	476	0
5	chr7	52422510	52422983	2	222	2.551	FP	il	NM_019883	Uba52		0
6	chr8	94882072	94882549	2	232	2.518	NV	e2	BC023331	Slc17a7		0
7	chr2	63936773	63937136	2	118	2.478	FP	o	NM_018826	Irx5	1072	0
8	chr8	15057301	15057953	2	404	2.380	NV	e2	BC059266	Fign		0
9	chr4	128783518	128784221	2	457	2.358	NV	il l	NM_008664	Myom2		0
10	chr6	90569034	90570252	4	401	2.327	NV	il e2	NM_027868	Tmem54 Tmem54		0
11	chr8	125224281	125224888	3	221	2.301	NV	o	NM_001037493	Slc41a3 Slc41a3	2011	0
12	chr8	107749142	107749644	2	257	2.280	NV	e2 e2	NM_009824	Cbfa2t3h		0
13	chr2	131394967	131395499	2	272	2.276	NV	o	AK166937 NM_145604	D230025D16Rik	7478	0
14	chr9	21223744	21224614	3	364	2.252	FP	e8	AK051965 BC066826	D230025D16Rik		0
15	chr3	89022839	89023393	2	309	2.252	NV	e7	NM_013691	D230025D16Rik		0
16	chr3	88066324	88066750	2	176	2.239	NV	o	NM_021375	Rhbg	8144	0
17	chr4	125823592	125824114	2	262	2.223	NV	o	NM_172145	2610027C15Rik	1132	0
18	chr4	154301726	154302271	2	300	2.179	NV	o	NM_178931	Tnfrsf14	85	0
19	chr5	23951682	23952258	2	329	2.166	NV	e2	NM_022418	Tmub1		0
20	chr7	138081260	138081620	2	109	2.164	NV	il	NM_019564	Htra1		0
21	chr5	34868596	34868951	2	110	2.122	NV	il	NM_011893	Sh3bp2		0
22	chr2	131012675	131013294	2	374	2.118	NV	e2 e2	NM_023117 BC057568	Cdc25b Cdc25b		0
23	chr1	192923348	192923815	2	219	2.117	NV	il l	AK171970 NM_030060	913021103Rik		0
24	chr4	128965108	128965688	2	335	2.116	NV	il l	AK160756 NM_023485	913021103Rik		0
25	chr5	143670058	143670564	2	246	2.094	NV	o	NM_007393	Sync Sync	2161	0
26	chr7	108543652	108544111	2	214	2.093	NV	il l il l	NM_001040111 A0621558	Actb		0
									NM_001040112	Centd2 Centd2		0

27	chr3	126700734	126701109	2	124	2.076	NV	e25[e17] e16[e12]	NM_027180 NM_198096 NM_178655 AK036018 AK082352 AK044634	Centd2 Ank2 Ank2 Ank2 Ank2	0
28	chr7	17477279	17478002	2	465	2.037	NV	e2	NM_010320	Gng8	0
29	chr6	97161310	97161673	2	108	2.014	NV	il	NM_022992	Ar16p5	0
30	chr2	179776846	179777294	2	203	2.004	FP	il	NM_178750	Ss181	0
31	chr8	13159417	13159933	2	271	2.003	NV	e2[e2	NM_010684 AK004637	Lamp1 Lamp1	0
32	chr2	174105864	174106699	3	337	1.998	NV	o	AK039035	Gnas	3121 18
33	chr2	180067766	180068281	2	255	1.981	NV	il	NM_008093	Gata5	0
34	chr7	148377990	148378880	3	445	1.965	NV	e4	BC061885	Hras1	1
35	chr7	148464995	148465622	2	382	1.949	NV	o	NM_011328	Sct	592 0
36	chr4	128553419	128553898	2	224	1.944	NV	o	NM_178110	Trim62	7485 0
37	chr5	115608783	115609156	2	128	1.940	NV	o	NM_175403	2410014A08Rik	1001 0
38	chr2	27527348	27527784	2	191	1.929	NV	o	NM_011305	Rxra	4936 7
39	chr7	28964046	28964759	3	254	1.924	NV	e2[e2	NM_010092 NM_001037957	Dyrk1b Dyrk1b	0
40	chr1	75231707	75232223	2	271	1.920	NV	o	BC011090	Dnajb10	820 0
41	chr5	34863611	34864238	2	382	1.905	NV	o	BC052083	Trip2	7610 0
42	chr2	127345397	127346114	2	464	1.897	NV	il il	NM_019789 AF287732 AK016721	Kcnip3 Kcnip3 DREAM/ calsenilin KChIP3	1
43	chr4	45843704	45844103	2	148	1.891	NV	o	NM_018741	Igfbp1	4404 0
44	chr5	136721240	136721777	2	291	1.878	FP	o	BC049777	Sh2b2	1004 0
45	chr1	93313967	93314423	2	208	1.872	TP	e2	NM_011066	Per2	6
46	chr5	4756950	4757494	2	299	1.867	NV	e2	NM_021457	Fzd1	0
47	chr7	148134128	148134628	2	249	1.844	NV	o	NM_026820	Iftm1	18496 0
48	chr2	32636879	32637540	2	401	1.841	NV	e9 il	NM_025619 AK132137	1700019L03Rik Pthr1	0
49	chr8	108467119	108467530	2	166	1.834	NV	il il	NM_009195 AK149969	Slc12a4 Slc12a4	0
50	chr3	90066934	90067820	3	370	1.833	NV	e14[e10 e8[e4	NM_028881 BC023091 AK014553 BC019713	Crtc2 Crtc2 Crtc2 Crtc2	0

51	chr1	78197650	78198091	2	196	1.831	FP	o	AK083424	Pax3	4380	1
52	chr8	96910274	96910637	2	118	1.828	NV	ijl	NM_022331 AK081688	Herpud1 Herpud1		0
53	chr5	97427288	97427929	2	381	1.828	NV	ijl	AK046752 NM_080708	Bmp2k Bmp2k		0
54	chr2	32207807	32208580	3	335	1.826	NV	ijl	NM_010065 L31395	Dnm1 ImKIAA4093		0
55	chr2	103810618	103811403	3	318	1.825	TP	e3 e2	AK013416 NM_008505	Lmo2 Lmo2		0
56	chr6	32536627	32537004	2	117	1.811	TP	il	NM_175750	Pixna4		0
57	chr4	124615773	124616208	2	186	1.806	NV	e4 e3	AK037053 NM_177573	9930104L06Rik 9930104L06Rik		0
58	chr2	24903083	24903455	2	127	1.805	NV	e27 e26 e7 e5	NM_146251 BC021312 EU046569 BC027342	Pnpla7 Pnpla7 Pnpla7 Pnpla7		0
59	chr7	108565215	108565780	2	317	1.799	NV	o	NM_001008548	Pde2a	4431	0
60	chr7	108534298	108534671	2	121	1.791	NV	ijl ijl	NM_001040111 AJ621558 NM_001040112 NM_027180	Centd2 Centd2 Centd2 Centd2		0
61	chr7	109590455	109590981	2	277	1.784	NV	e2	NM_009103	Rrm1		0
62	chr7	31701741	31702370	2	384	1.784	NV	o	AK045564	Mag	2519	12
63	chr7	51030485	51031153	2	423	1.763	NV	ijl	AK134755 AB016227	AK134755 Klk11		0
64	chr4	106984710	106985360	2	405	1.760	NV	o	AK077502	Yipf1	1675	0
65	chr2	160471716	160472085	2	124	1.758	NV	e2 e2	AK166844 NM_009408	Top1 Top1		0
66	chr6	92140136	92140502	2	113	1.758	FP	e2	NM_030081	Zfyve20		0
67	chr7	52869657	52870738	3	464	1.748	NV	e2	NM_020013	Fgf21		1
68	chr7	51923278	51923931	2	407	1.741	NV	il	AY363100	Myh14		0
69	chr5	31548676	31549054	2	133	1.741	NV	ile9	NM_147201 AK046142	Nrbp1 Nrbp1		0
70	chr7	148622490	148623119	2	384	1.739	NV	ijl	NM_026646 AK018760	Slc25a22 Slc25a22		0
71	chr2	174104562	174105234	2	427	1.724	NV	o	AK039035	Gnas	4586	18
72	chr1	75499918	75500487	2	324	1.723	NV	e19 e11 e8 e4	AK029863 NM_178884 AK162473 AK031074	Obsl1 Obsl1 Obsl1 Obsl1		0
73	chr7	31700170	31700729	2	299	1.722	TP	o	AK045564	Mag	878	12
74	chr3	52845317	52845973	2	411	1.716	NV	e2 e2	NM_175386 AK172532	Lhfp Lhfp		0
75	chr7	108533357	108533908	3	188	1.711	TP	e5 e3 il	NM_001040111 AJ621558	Centd2 Centd2		0

76	chr4	137416409	137416960	2	305	1.668	NV	o	NM_001040112	Centd2	1251	0
77	chr2	25481417	25481845	2	183	1.663	NV	e5 j	AK047359	Ece1		0
78	chr4	62186073	62186647	2	329	1.656	NV	e2	NM_007995 AK039020	Fcna AK039020		0
79	chr2	28049852	28050346	2	249	1.653	NV	ij j	NM_177607	493343017Rik		0
80	chr8	94880530	94880904	2	129	1.652	NV	o	NM_001038614	Olfm1 Olfm1		0
81	chr2	160191359	160191999	2	395	1.635	NV	e2	NM_001038613	Irx5	790	0
82	chr7	152488550	152489007	2	212	1.633	NV	ij	NM_010658	Marf		0
83	chr8	80959161	80959661	2	255	1.620	NV	e2	NM_145379	Mrgprf		0
84	chr7	150192500	150193024	2	276	1.619	NV	ij le4 e3 e3 e3	NM_138944	Pou4f2		0
85	chr2	180330388	180330790	2	157	1.618	NV	o	AK039657 AK041748	AK039657 Tspan32		0
86	chr4	154327029	154327729	3	263	1.614	TP	o	AY039000 NM_020286	Tspan32 Tspan32		6
87	chr2	173044846	173045234	2	143	1.594	NV	o	AJ279792 AJ279793	Tspan32 Tspan32		0
88	chr7	29546733	29547185	2	206	1.581	FP	e6 e5	AJ279794 AJ279795	Tspan32 Phemx		0
89	chr7	29172656	29173371	2	463	1.570	NV	ij	NM_010908 AY176058	Nfkbb Nfkbb		0
90	chr2	119146509	119146983	2	228	1.566	NV	ij	NM_026042	Med29		0
91	chr6	85286943	85287650	3	317	1.561	NV	e2 e2	AK039958	AK039958		0
92	chr1	36126785	36127215	2	176	1.557	NV	ij	NM_001003955	Rab11fip5		0
93	chr2	93476336	93477043	2	456	1.552	NV	o	NM_177466	Rab11fip5		0
94	chr7	19935602	19936292	2	445	1.545	NV	e5 e5	NM_015818	Hs6st1		0
95	chr3	37247585	37248298	2	468	1.544	NV	e2	NM_007442	Aix4	5547	0
96	chr3	90316286	90316874	2	343	1.544	NV	ij	AK160703 NM_007948	Ercc1 Ercc1		0
97	chr2	92446287	92446768	2	236	1.542	NV	ij	NM_008006	Fgf2		63
98	chr4	133807818	133808429	3	230	1.489	NV	e2	NM_011309	S100a1		0
99	chr11	69905111	69905765	2	404	3.271	NV	o	AK134997	Chst1		1
100	chr16	17506180	17506696	2	264	2.602	NV	e20 e21	NM_00109296	Grp1		0
				2	404	3.271	NV	o	NM_007493	Asgr2	380	0
				2	264	2.602	NV	e20 e21	AK158809 NM_175178	Aifm3 Aifm3		0

101	chr11	117646206	117646796	2	342	2.582	NV	e7 e7 e6	AK156764 AK171772 NM_181856	Tmc8 Tmc8 Tmc8		0
102	chr15	88965876	88966554	2	434	2.493	NV	e3 e9 i e9	AK178955 NM_013871 AK165741 AK011286	Mapk12 Mapk12 Mapk12 Mapk12		0
103	chr15	79852833	79853205	2	126	2.472	NV	o	AK087840	Pdgb	7967	3
104	chr10	57514939	57515347	2	156	2.460	TP	i	NM_020561	Smpd3a		0
105	chr10	77882614	77882976	2	116	2.443	TP	o	NM_007793	Cstb	5438	0
106	chr11	69613817	69614187	2	120	2.399	NV	i	NM_010198	Fgf11		0
107	chr14	56215446	56215876	2	186	2.316	NV	e12	NM_194346	Rnf31		0
108	chr15	81416435	81416819	2	140	2.304	FP	i i	AK042627 NM_177821	Ep300 Ep300		3
109	chr16	41532845	41533217	2	126	2.289	NV	i	AK164171	Lsamp		8
110	chr16	90221400	90221784	2	138	2.263	NV	i	NM_011434	Sod1		0
111	chr11	95704850	95705216	2	118	2.261	NV	e3 e3	NM_023121 NM_001038664	Gngt2 Gngt2		1
112	chr19	10599441	10599925	2	238	2.209	NV	e2 e2 e2 e2	AK044306 AK165330 NM_172302 AK053375	57304531 6Rik 57304531 6Rik 57304531 6Rik 57304531 6Rik		0
113	chr15	96198147	96198767	2	360	2.208	NV	i	NM_175251	Arid2		0
114	chr10	79460968	79461324	2	112	2.198	NV	e5	NM_013850	Abca7		0
115	chr17	28338616	28338982	2	120	2.186	NV	i	NM_172617	Zfp523		0
116	chr15	102236111	102236629	2	264	2.180	NV	o	EF200008	Sp1	1070	66
117	chr14	31741855	31742223	2	122	2.168	NV	o	NM_008407	Ith3	5495	0
118	chr9	103134524	103134942	2	162	2.146	NV	o	NM_133977	Trf	2326	74
119	chr18	38122382	38122772	2	144	2.126	NV	e5	NM_175684	Fchsd1		0
120	chr15	88971788	88972346	2	312	2.111	NV	o	AK011286	Mapk12	1264	0
121	chr11	3993890	3994400	2	266	2.110	NV	e3	NM_026443	1700020C11Rik		0
122	chr14	56203866	56204318	2	206	2.103	NV	o	AK148628	1500005A01Rik	227	0
123	chr19	5879568	5880244	2	430	2.095	NV	e3 e4 e3	AK133879 AK170869 NM_134154	AW491445 Slc25a45 Slc25a45		0
124	chr8	3622027	3622402	2	115	2.089	NV	i	AK011549	AK011549		0

125	chr15	101090429	101091023	2	348	2.074	NV	o	NM_010444	Nr4a1	6269	92
126	chr14	56502084	56502802	2	462	2.071	NV	e3	NM_019820	Cbln3		0
127	chrX	7414165	7414777	2	364	2.066	NV	o	AK155131	Otd5	4179	0
128	chr17	68623443	68623905	2	218	2.065	NV	ij j	NM_177278 AK053068	L3mbt4 L3mbt4		0
129	chrX	153931157	153931541	2	126	2.060	TP	o	NM_009214	Sms	1563	48
130	chr11	4965092	4965686	2	348	2.023	NV	o	BC031785	Gas2l1	383	1
131	chr18	32090738	32091594	3	396	2.022	NV	e2	NM_144862	Lims2		0
132	chr14	31694983	31695397	2	164	2.016	NV	o	NM_018746	Itih4	4301	0
133	chr11	101526580	101527180	2	354	2.014	NV	e2	NM_025404	Arl4d		0
134	chr18	50287962	50288532	2	316	2.013	NV	e2	NM_008292	Hsd17b4		0
135	chr11	119219152	119219618	2	222	2.003	NV	e4	NM_178743	Slc26a11		0
136	chr18	36898370	36899004	2	374	1.997	NV	e8 e8	AK046207 NM_028036	Tmco6 Tmco6		0
137	chrX	141122825	141123207	2	132	1.993	NV	o	NM_001033795	Zcchc16	242	0
138	chr12	36883987	36884341	2	110	1.990	NV	e2 e2	NM_178910 NM_146033	Ankmy2 Ankmy2		0
139	chr12	113882599	113883185	2	342	1.969	NV	ij j	NM_013929 AF033112	Siva1 Siva		0
140	chr17	27072064	27072504	2	194	1.965	NV	e6 e6 e6 e6 e4	NM_009343 AK144793 BC015077 AK133474 AK076365	Phf1 Phf1 Phf1 Tctex-3 Tctex-3		0
141	chr18	61201167	61201731	2	320	1.957	NV	o	NM_008809	Pdgfrb	3363	7
142	chr11	72973049	72973605	2	310	1.955	NV	o	NM_033321	P2rx5	421	0
143	chr15	74509157	74510173	3	476	1.953	TP	o	AK157822	Arc	7173	37
144	chr15	58912324	58912708	2	124	1.932	NV	ij j j	NM_144800 AK129139 AK046628	Mtss1 Mtss1 mKIAA0429		0
145	chr10	126430219	126430629	2	150	1.916	NV	ij	AK045068	Ctdsp2		0
146	chr11	95702514	95703232	2	474	1.912	NV	ij	NM_025659	Abi3		0
147	chrX	138824647	138825079	2	186	1.906	NV	o	NM_019477	Acs14	1	0
148	chrX	7325659	7326157	2	254	1.896	NV	o	NM_207202	Ccdc120	7707	0
149	chr9	66005449	66005859	2	162	1.889	NV	o	AK158082	DAPK2	214	1
150	chr9	111341952	111342606	2	402	1.867	NV	ij	NM_172928	Dclk3		0
151	chr13	83643058	83643522	2	206	1.866	TP	e2	NM_025282	Meif2c		3

152	chrX	50265416	50265890	2	228	1.861	NV	e2 e2 e2 e2	AK157045 AK038320 NM_027642 AK028498	mKIAA1823 Phf6 Phf6 Phf6		1
153	chr17	31271523	31272051	2	274	1.857	NV	o	NM_011575	Tff3	5495	0
154	chr10	126432485	126433303	4	268	1.848	TP	e7 e2	AK045068 NM_146012	Ctdsp2 Ctdsp2		0
155	chr12	114371679	114372069	2	146	1.847	TP	e18 e19 i	NM_054081 AF463504 BC112910	Mta1 Mta1 Mta1		1
156	chr19	5993526	5993930	2	160	1.832	NV	ij ij	AF084459 NM_007600 AK171405	Capn1 Capn1 Capn1		0
157	chr11	72828889	72829327	2	192	1.827	NV	o	NM_018883	Camkk1	3182	0
158	chr10	126602999	126604075	3	468	1.821	NV	e3 e2 e2 e3 e3 e2	BC022180 AK086849 AK040324 NM_008080 BC019969 BC036996	B4galnt1 B4galnt1 B4galnt1 B4galnt1 B4galnt1 B4galnt1		0
159	chr15	102349043	102349765	2	478	1.812	NV	e2 e3 e3 ij	AK181782 NM_009319 AK143188 AK144205	Tarbp2 Tarbp2 Tarbp2 Tarbp2		0
160	chr14	104872120	104872784	2	420	1.805	NV	o	NM_011143	Pou4f1	5568	0
161	chr8	3631021	3631710	2	429	1.804	NV	e2 e2	NM_011503 BC003477	Stxbp2 Stxbp2		0
162	chr11	53907117	53907667	2	290	1.786	NV	o	AK142624	P4ha2	6809	0
163	chr14	121309436	121309972	2	292	1.779	NV	o	AK017701	Ranbp5	443	0
164	chr11	62058965	62059473	2	248	1.771	NV	o	NM_007413	Adora2b	3012	0
165	chr11	59041905	59042433	2	282	1.749	NV	o	AK168081	Arf1	744	0
166	chr12	28027233	28027819	2	336	1.735	NV	o	NM_009234	Sox11	242	0
167	chr14	45608851	45609390	2	279	1.734	NV	e2 ij	NM_008964 AK144143	Ptger2 AK144143		0
168	chr15	84967866	84968224	2	112	1.732	NV	ij ij	AK006801 BC112373	Ribc2 Ribc2		0
169	chr16	20645376	20645958	2	336	1.728	NV	e21 e20 e20 e19 ij e14	NM_177941 NM_177940 NM_139293 NM_177942 AK031951 BC030900	Ece2 Ece2 Ece2 Ece2 Ece2 Ece2		0
170	chr14	35182104	35182518	2	170	1.727	TP	o	NM_153127	Mmrn2	6171	0
171	chr11	52174093	52174571	2	232	1.716	NV	o	NM_011694	Vdac1	45	0
172	chr11	77029527	77030071	2	298	1.712	NV	e2	NM_177710	Ssh2		0
173	chr14	34025780	34026508	2	484	1.695	NV	o	NM_153800	Arhgap22	3500	0
174	chr9	46030948	46031650	3	338	1.692	NV	o	NM_009692	Apoa1	5142	0
175	chr10	87319252	87319622	2	124	1.678	TP	o	NM_184052	Igf1	3152	24

176	chr15	76519175	76519975	3	430	1.662	NV	e3 l l	AK193686 NM_033371 AK220248	Ppp1r16a Ppp1r16a Ppp1r16a		1
177	chr9	108961370	108962006	2	392	1.652	NV	e2 e2 e2 l l	NM_011637 NM_001012236 BC060205 NM_172774 BC078458	Trex1 Trex1 6620401K05Rik 6620401K05Rik 6620401K05Rik		1
178	chr10	74429151	74429509	2	112	1.649	NV	il	AK017008	Rtdr1		0
179	chr9	44127860	44128402	2	296	1.644	NV	o	AK039220	Tmem24	37	0
180	chr13	43461087	43461651	2	318	1.628	NV	il	AK002609	Sirt5		0
181	chr15	76173899	76174731	3	396	1.625	NV	e2 l	NM_025567 AK170272	Cyc1 Cyc1		0
182	chr10	4522078	4522446	2	122	1.619	NV	o	NM_175374	Mtrf1	151	0
183	chr11	62692979	62693493	2	268	1.605	NV	e2	NM_026210	Fam18b		0
184	chr11	76660019	76660555	2	292	1.602	NV	o	AK142583	Cpd	45	2
185	chr9	107577514	107577888	2	130	1.599	NV	e2	NM_008140	Gnat1		0
186	chr11	97560386	97560760	2	128	1.574	NV	ile12	AK142837 NM_009545	Pcgf2 Pcgf2		0

Appendix C

Tables below show the results of the known TRE motif scanning for PND4 and PND15. Columns are described as the peakID which is the corresponding identifier to the peakID identifier for PND4 and PND15 in Appendix B, strand on which the TRE is located, the start position with respect to the location “start” in Appendix B, the end position with respect to the location “start” in Appendix B, first half site, second half sites, the score of the first half site, the score of the second half site and the type of TRE {DR4=Direct repeat with a spacer of 4 nucleotides, IR0= Inverted repeat with no spacer, ER6 = Everted palindrome with a spacer of 6 nucleotides}.

PND4

peakID	strand	start	end	string	string2	score1	score2	type
2	1	71	86	agggca	aggtgg	6.28	4.96	DR4
3	1	378	393	aggtta	agggtta	7.65	5.04	DR4
4	-1	291	306	agggca	aggagt	6.28	4.21	DR4
4	1	341	358	taacct	agggga	7.65	5.22	ER6
7	1	140	157	tctcct	cgggtga	6.15	4.51	ER6
8	1	193	210	ggacct	aggtca	4.7	8.89	ER6
8	1	446	463	agacct	gggaca	6.94	5.72	ER6
10	1	159	176	tctcct	gggata	6.15	4.48	ER6
12	1	36	53	tctcct	agggga	6.15	5.22	ER6
32	-1	217	232	agggca	ggggca	6.28	4.79	DR4
32	-1	212	227	agggca	agggca	6.28	6.28	DR4
32	-1	207	222	aggcca	agggca	6.96	6.28	DR4
37	1	334	349	aggaca	agggtta	7.2	5.04	DR4
40	-1	390	405	gggcta	aggaca	4.23	7.2	DR4
41	1	563	580	tgacct	aagtca	8.89	4.43	ER6

PND15

peakID	strand	start	end	string	string2	score1	score2	type
2	1	187	204	agaccc	gggtta	5.46	6.16	ER6
3	1	204	221	taacct	aggtgt	7.65	5.89	ER6
6	1	12	29	tgtcct	agggct	7.2	4.34	ER6
8	-1	452	467	aggaga	agctga	6.15	4.13	DR4
9	1	637	654	ttacct	aggaca	6	7.2	ER6
10	-1	254	269	tgttca	aggcca	4.76	6.96	DR4
12	-1	347	362	aggcga	gggtga	5.9	6.34	DR4

14	-1	114	129	gggtca	aggaaa	7.4	4.31	DR4
23	-1	216	231	aggtgt	aggaga	5.89	6.15	DR4
26	1	345	360	aggtct	aggcca	6.94	6.96	DR4
30	-1	20	35	agctga	aggtga	4.13	7.83	DR4
32	1	72	87	agggca	aggtgg	6.28	4.96	DR4
40	-1	54	69	gggaca	aggtga	5.72	7.83	DR4
40	1	406	421	aggcta	aggtct	5.72	6.94	DR4
42	1	37	54	tctcct	agggga	6.15	5.22	ER6
49	-1	190	205	aggact	aggcca	5.26	6.96	DR4
55	1	544	559	gggaga	aggaca	4.66	7.2	DR4
56	-1	225	240	aggtct	aagtca	6.94	4.43	DR4
57	-1	19	34	gggcga	aggaca	4.42	7.2	DR4
58	1	303	318	aggtgt	aggtga	5.89	7.83	DR4
59	1	153	170	cgtcct	aggaga	4.34	6.15	ER6
60	-1	73	88	gggaca	aggtga	5.72	7.83	DR4
63	1	378	389	aggcca	agccct	6.96	4.34	IR0
67	1	441	458	tgccct	atgtca	6.28	4.43	ER6
69	-1	201	216	ggggca	aggtta	4.79	7.65	DR4
70	1	579	590	agatca	tgccct	4.38	6.28	IR0
72	1	227	244	tctcct	acgtca	6.15	5.75	ER6
73	1	257	268	agatca	cgacct	4.38	6.02	IR0
76	-1	413	428	gggcta	aggaca	4.23	7.2	DR4
82	1	300	311	aggccg	tgtcct	4.09	7.2	IR0
84	1	194	211	ggacct	aggtca	4.7	8.89	ER6
84	1	447	464	agacct	gggaca	6.94	5.72	ER6
85	-1	26	41	aggaca	agggca	7.2	6.28	DR4
86	-1	227	242	aggtgg	aggaca	4.96	7.2	DR4
87	1	85	96	aggcca	tgatct	6.96	4.38	IR0
87	1	283	298	aggtct	agctta	6.94	3.95	DR4
88	1	241	256	gggtta	gggcca	6.16	5.47	DR4
89	-1	501	516	aggaca	gggtgt	7.2	4.4	DR4
91	1	310	321	agggct	agacct	4.34	6.94	IR0
91	-1	443	458	aggtaa	ggggca	6	4.79	DR4
93	1	160	177	tctcct	gggata	6.15	4.48	ER6
94	-1	85	100	aggcca	gggtta	6.96	4.51	DR4
96	1	253	270	tgagct	gggtca	5.19	7.4	ER6
96	1	255	270	agctga	gggtca	4.13	7.4	DR4
96	-1	423	438	aggcta	aggaca	5.72	7.2	DR4
96	1	433	450	tgtcct	agctca	7.2	5.19	ER6
99	1	21	32	gggtta	tgactt	6.16	4.43	IR0
104	-1	208	223	aggcca	agggca	6.96	6.28	DR4
104	-1	213	228	agggca	agggca	6.28	6.28	DR4
104	-1	218	233	agggca	ggggca	6.28	4.79	DR4
106	1	87	104	tacct	aggtta	5.04	6	ER6
106	1	255	270	agggca	gggcta	6.28	4.23	DR4
106	1	290	307	cggcct	aggcca	4.09	6.96	ER6
111	1	74	89	tggtca	aggaga	4.76	6.15	DR4

122	1	340	351	gggtca	tacct	7.4	5.04	IR0
123	1	36	53	tggcct	aggcca	6.96	6.96	ER6
123	1	85	102	tgacct	aggtat	7.4	4.06	ER6
125	1	564	581	tgacct	aagtca	8.89	4.43	ER6
130	1	565	580	aggtgg	gggtta	4.96	6.16	DR4
133	1	370	385	aggtaa	gggtgt	6	4.4	DR4
138	1	56	73	acacct	aggcca	5.89	6.28	ER6
146	1	245	260	aggcca	gggcta	6.96	4.23	DR4
148	-1	391	406	gggcta	aggaca	4.23	7.2	DR4
152	-1	165	180	aggaca	gggcga	7.2	4.42	DR4
155	-1	39	54	gggaga	aggaga	4.66	6.15	DR4
155	-1	308	323	aggcca	aggcca	6.96	6.96	DR4
157	1	185	202	tcacct	agggtct	7.83	4.34	ER6
158	-1	1020	1035	aggaga	aggagt	6.15	4.21	DR4
159	-1	631	646	aggttg	aggtga	4.78	7.83	DR4
160	-1	384	399	aggtcg	aggccg	6.02	4.09	DR4
161	1	244	261	ggacct	aggtga	4.7	7.83	ER6
161	1	437	448	acgtct	agacct	3.81	6.94	IR0
162	1	175	186	aggtct	ttacct	6.94	6	IR0
163	1	335	350	aggaca	agggta	7.2	5.04	DR4
164	1	393	408	gggtga	agctca	6.34	5.19	DR4
176	1	318	333	gggtca	aggtat	7.4	4.06	DR4
180	1	321	332	aggtga	tcagct	7.83	4.13	IR0
181	1	212	227	cggaca	aggtga	3.89	7.83	DR4
181	-1	516	531	agggta	gggtca	5.04	7.4	DR4
185	-1	157	172	aggact	aggaca	5.26	7.2	DR4

Appendix D

Tables below show the results of the known AP-1 motif scanning for PND4 and PND15. Columns are described as the peakID which is the corresponding identifier to the peakID identifier for PND4 and PND15 in Appendix B, strand on which the AP-1 consensus is located, the start position with respect to the location “start” in Appendix B, the end position with respect to the location “start” in Appendix B and the score (bits) of the motif.

PND4

Peak ID	Strand	Start	Stop	String	Score
2	-1	8	14	tgagtaa	9.2
2	-1	650	656	tgtgtca	9.3
3	1	106	112	tgtgtca	9.3
15	-1	253	259	tgtgtca	9.3
17	-1	198	204	tgtgtca	9.3
24	1	209	215	tgagtaa	9.2
24	-1	594	600	tgcgta	9.3
26	1	476	482	tgtgtca	9.3
28	1	18	24	tgagtca	11.8
33	1	19	25	tgaatca	9.9
36	-1	284	290	tgtgtca	9.3
36	-1	423	429	tgagtaa	9.2
37	1	327	333	tgagtca	11.8
37	1	360	366	tgagtca	11.8
40	1	490	496	tgtgtca	9.3
41	1	244	250	tgagtca	11.8

PND15

Peak ID	Strand	Start	Stop	String	Score
2	-1	220	226	tgtgtca	9.3
8	1	29	35	tgattca	10.5
21	1	321	327	tgattca	10.5
24	1	113	119	tgtgtca	9.3
32	-1	8	14	tgagtaa	9.2
32	-1	650	656	tgtgtca	9.3
33	1	41	47	tgaatca	9.9
44	-1	362	368	tgtgtca	9.3
47	-1	253	259	tgtgtca	9.3
51	1	97	103	tgagtaa	9.2
52	1	318	324	tgagtcg	9.1

53	1	51	57	tgactca	10.5
53	1	394	400	tgtgtca	9.3
54	1	502	508	tgagtaa	9.2
55	1	249	255	tgagtaa	9.2
60	1	100	106	tgtgtca	9.3
62	1	403	409	tgtgtca	9.3
75	1	406	412	tgagtaa	9.2
81	1	599	605	tgagtca	11.8
82	1	311	317	tgagtca	11.8
88	1	394	400	tgaatca	9.9
90	1	458	464	tgagtca	11.8
91	1	231	237	tgagtca	11.8
94	1	580	586	tgagtcg	9.1
96	1	86	92	tgactca	10.5
97	1	144	150	tgactca	10.5
97	1	184	190	tgtgtca	9.3
102	1	452	458	tgtgtca	9.3
107	1	28	34	tgaatca	9.9
110	1	358	364	tgcgta	9.3
118	1	241	247	tgtgtca	9.3
120	1	164	170	tgagtaa	9.2
120	1	270	276	tgaatca	9.9
123	1	582	588	tgagtca	11.8
125	1	244	250	tgagtca	11.8
132	1	174	180	tgtgtca	9.3
134	1	318	324	tgagtaa	9.2
134	1	508	514	tgattca	10.5
143	1	209	215	tgagtaa	9.2
143	-1	594	600	tgcgta	9.3
146	-1	160	166	tgtgtca	9.3
148	1	490	496	tgtgtca	9.3
149	-1	388	394	tgagtaa	9.2
163	1	327	333	tgagtca	11.8
163	1	360	366	tgagtca	11.8
169	1	140	146	tgtgtca	9.3
169	-1	216	222	tgagtaa	9.2
170	-1	138	144	tgagtcg	9.1
173	1	205	211	tgactca	10.5
173	1	445	451	tgaatca	9.9
185	1	129	135	tgagtca	11.8

Appendix E

The following table is the list of GpC islands in the TRE chIP-chip peaks. The first column is the peak identification, the second and third column are the start and end position with respect to the peak location and the fifth and sixth column are the second (if applicable) coordinates CpG island location of the peak with respect to the peak ID.

TRE peak ID	Start position 1	End position 1	Start position 2	End position 2
PND4-09	0	270		
PND4-16	283	470		
PND4-23	159	386		
PND4-37	460	536		
PND4-44	0	552		
PND15-004	0	366		
PND15-006	0	477		
PND15-012	0	203		
PND15-028	159	386		
PND15-030	0	448		
PND15-031	0	220		
PND15-037	0	102		
PND15-039	283	713		
PND15-046	0	178	213	544
PND15-052	0	363		
PND15-054	130	351	517	773
PND15-055	0	459		
PND15-061	0	265		
PND15-065	0	369		
PND15-072	336	569		
PND15-074	0	451		
PND15-078	143	519		
PND15-081	95	640		

PND15-083	164	500		
PND15-095	0	429		
PND15-098	0	52		
PND15-108	0	117	270	384
PND15-110	0	197		
PND15-112	51	484		
PND15-122	0	128		
PND15-133	147	588		
PND15-134	0	256		
PND15-139	14	586		
PND15-147	13	432		
PND15-152	0	474		
PND15-158	0	354		
PND15-160	141	579		
PND15-163	460	536		
PND15-165	0	143		
PND15-166	0	367	491	586
PND15-167	73	287	310	531
PND15-171	177	478		
PND15-172	0	420	427	544
PND15-179	0	542		
PND15-181	0	381		
PND15-183	0	184		
PND15-184	0	536		
PND15-186	0	136	204	374