

# CHAPTER ONE: INTRODUCTION

## 1.1 IMMUNE SYSTEM

The immune system is an intricate defense mechanism comprised of biological elements that protect the organism against foreign invaders (like bacteria, viruses, etc.) by recognizing and destroying them. The immune system is classified based on the time and nature of their actions; the first section of the immune system is composed of a layer of epithelial cells that acts as a physical barrier by providing an immediate defense against the invading pathogens in a non-specific manner and is termed as innate immune system. Innate immunity is triggered by the incoming pathogens that are identified by the pattern recognition receptor (PPR). The PPR identifies the incoming pathogen by recognizing a pattern specific for each group of microorganisms termed as pathogen-associated molecular patterns (PAMPs) invading the host [1]. After recognizing the pathogens, these epithelial cells secrete chemicals (enzymes like lysozyme) to carry out cell wall lysing of the pathogen. The cell lysing is then followed by engulfing and destroying the pathogen by endocytosis and phagocytosis.

The second component of the immune system is called the adaptive immune system. Adaptive immunity is more specific to the invading pathogens. This type of immunity is characterized by immunological memory, through which each pathogen is recalled. What helped to eradicate the infectious pathogens is also recalled from the lymphocytes [2] . The lymphocytes belong to the category of leukocytes that are mainly concentrated in the central lymphoid systems such as the spleen and lymph nodes.

The two components of the immune systems complement each other very effectively. In summary, the non-specific innate immunity is developed immediately after birth and is effective only when challenged whereas the specific adaptive immunity is developed by an organism over a period of time by exposing it to immunization and is effective for a long time period as it utilizes immunological memory.

## **1.2 CELLS INVOLVED IN IMMUNE RESPONSE**

The inflammatory reaction is the result of immune response. It is stimulated when the entering pathogen injures the host tissue, which in turn recruits cells by secreting chemokines. Some of the essential components that are responsible for the pro-inflammatory response are TNF- $\alpha$ , cytokines like IFN- $\gamma$ , white blood cells like antigen presenting dendritic cells, lymphocytes like T cells and B cells, and effector cells or natural killer cells (NK cells) [3]. Figure 1 gives a clear diversification of the blood stem cell into various immune cells that carry out the immune response effectively.

Out of these components, T cells play a significant role in immune response. These are divided into two major classes: T helper cells (CD4<sup>+</sup> cells) and cytotoxic T cells (CD8<sup>+</sup> cells) [4].

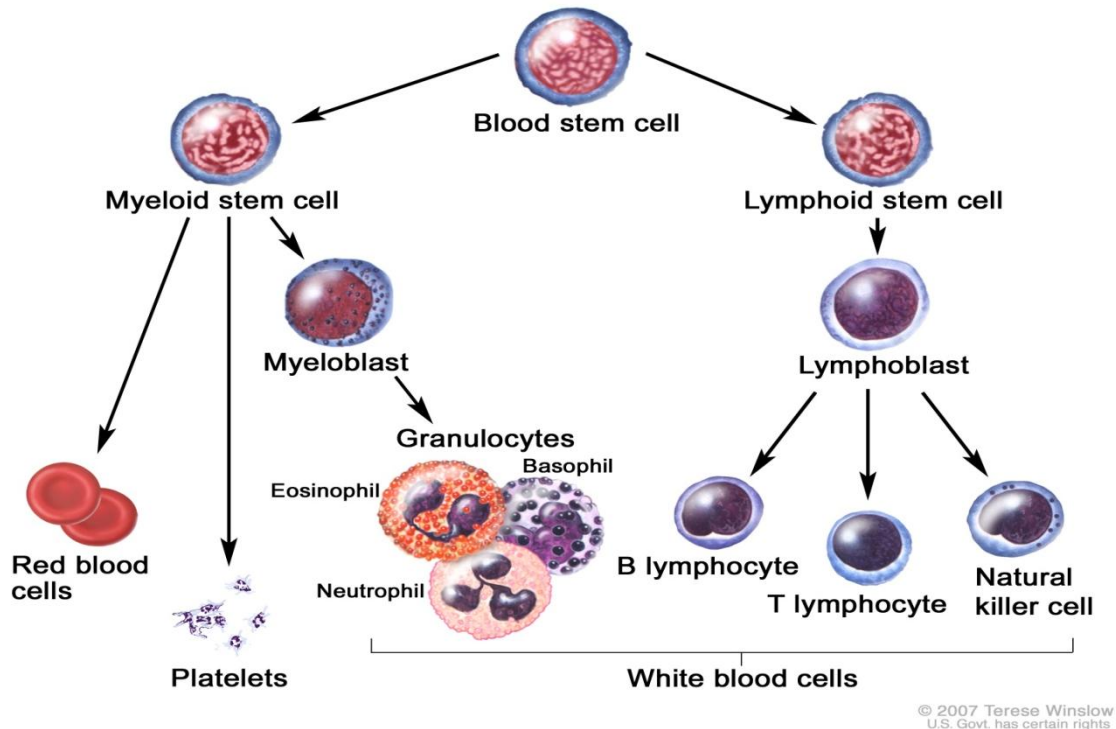


Figure1: Cells involved in the immune response (Terese Winslow, 2007). Out of these cells, the cells that encompassed in the lymphoblast lineage mostly involved in immune response.

### 1.3 T-HELPER CELL DEVELOPMENT AND ACTIVATION

The T helper cells are a subset of lymphocytes that act against the pathogens by stimulating cytokines (especially  $\text{IFN-}\gamma$ ), leading to an inflammatory response. This category of T cells has  $\text{CD4}^+$  proteins on their surface. The activation of this cell lineage is regulated by the antigen peptides binding to the major histocompatibility complex class II (MHC-II) present on the surface of antigen presenting cells (APCs) [5]. Additionally, these T helper cells mature and develop in the thymus region [6].

The T helper cells also differentiate into many sublineages, like Th1, Th2, Treg, Th17, T follicular helper (Tfh) cells, and Th9 (IL-9-expressing) cells. Of these sublineages Th1 and Th2 have a more

established participation in the immune response mechanism (Figure 2) [7].

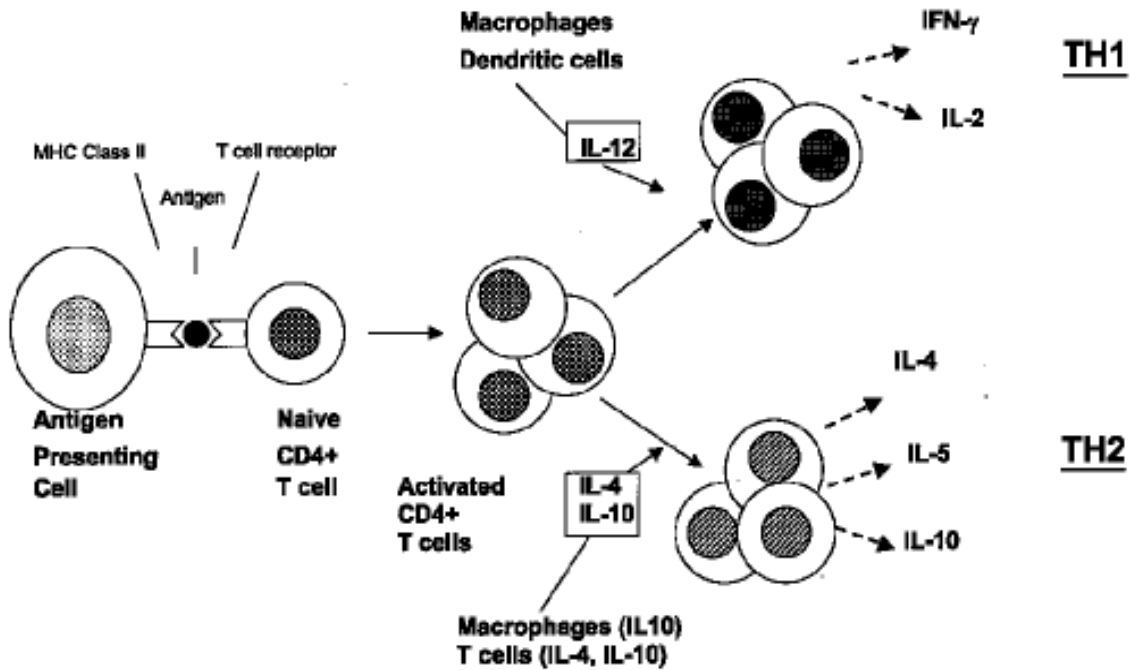


Figure2: T-helper cell lineages [8] . The activated CD4+ T cell differentiates into many T-helper cell lineages which are mediated by different interleukin molecule (Example: The Th1 lineage is mediated by IL-12 whereas Th2 lineage is mediated by IL-4).

The cytokine IFN- $\gamma$  stimulates the macrophages and dendritic cells to produce interleukin-12, which is critical in promoting the Th1 cell lineage development. The IFN- $\gamma$  cytokine stimulation also inhibits the production of other interleukins (like IL-4, IL-10), thus preventing the development of other Th cell lineages such as Th2, Treg, etc.

### 1.3.1 TRANSCRIPTION REGULATION IN Th1 CELL DEVELOPMENT

Transcription regulation is a dynamic process that regulates the expression of genes by employing various transcription regulatory elements such as activators, enhancers, and repressors. One of the complex transcription regulatory networks in higher eukaryotes is the immune cell transcription regulatory network.

The T cell transcription regulatory network is a complex system regulated by various families of factors, such as NF- $\kappa$ B, STAT family, and GATA (FIGURE 3) [9]. The transcription regulatory network plays an essential role in diversification of the T cell lineage into Th1 and Th2 sublineages. The Th1 lineage's transcription network is regulated by T-bet which cooperates with another factor called STAT4 [10].

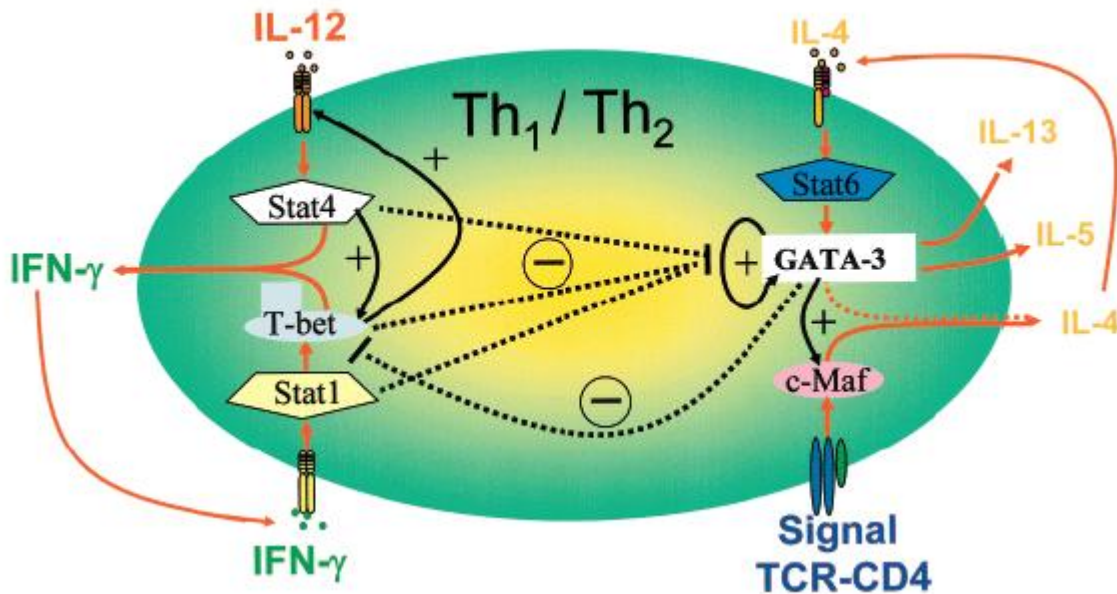


Figure3: Transcription regulation in Th1 cell [11].The T-bet induces IL-12 expression that enables the STAT4 to undergo dimerization process leading the Th1 cell development.

T-bet (also called Tbx-21) is a member of the T-box family of transcription factors (TFs). It plays an essential role in the activation of IFN- $\gamma$  production, which eventually leads to the repression of interleukin-4 and 5 (IL-4 and IL-5); this plays a major role in the development of Th 2 cell lineage [12]. T-bet also induces the expression of IL-12R, enabling STAT4 activation for Th1 cell development [13]. STAT4 is a member of the STAT transcription factor family that is regulated by a cytokine response [14].

The TCR signal transduction cascade activates the master regulators and also the cytokine gene framework to induce the transcription process leading to the development of Th1 cell lineage [15].

#### **1.4 ROLE OF CIS-REGULATORY MODULE IN TRN (TRANSCRIPTION REGULATORY NETWORK)**

A few transcription factors (for example, T-bet and STAT4) work in coordination with each other to regulate the transcription regulatory network. Such factors bind to the specific transcription factor binding sites (TFBS) that are located very close to each other in the genomic region. These co-occurring heterotypic or homotypic clusters of TFBS form the base for the transcription regulatory network and are referred to as Cis-regulatory modules (CRM) [16]. The CRMs are the main components that control the regulation of transcription regulatory networks [17]. The CRM is about 500-1000 bp in length and is mostly positioned in the upstream region to the transcription start site.

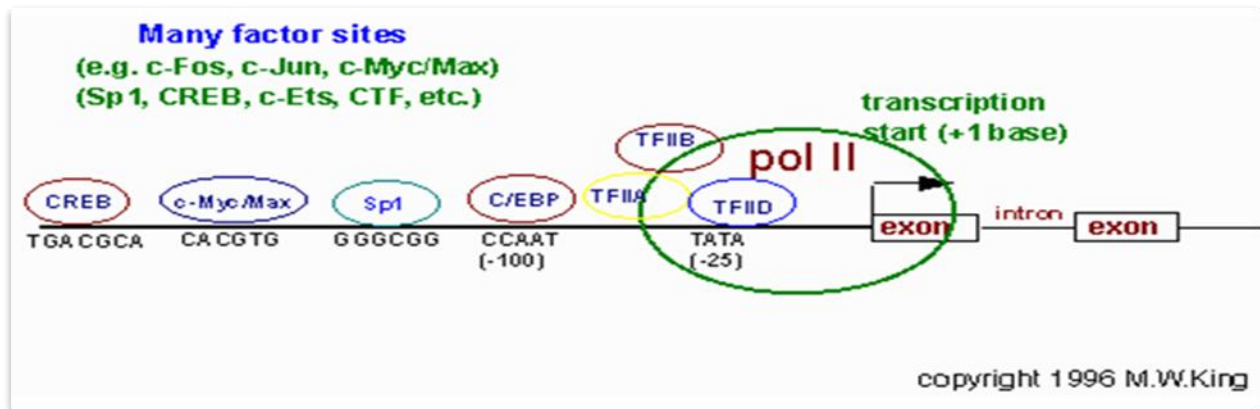


Figure 4: Representation of the cis-regulatory module or Cistrome unit. In this representation, the transcription factors Sp1, c-Myc/Max and others binding to their respective TFBS on the genes forming the *cis*-Regulatory Module or unit. And these sites co-regulated to carry out gene expression.

## **CHAPTER TWO: BACKGROUND**

### **2.1 STAT4 BIOLOGY**

STAT4 transcription factors are categorized under the family of signal transducers and activators of transcription proteins that play a cardinal role in the development of Th1 cell lineage. The activation of STAT4 molecules is an orchestra of interactions with other molecules. The IL-12 molecule binds to the IL-12 receptor on the surface of the naive T helper cell (Th0); this binding in turn mediates the conversion of the inactive STAT4 molecule to an active homodimer STAT4 molecule. The active STAT4 molecule carries the signal from the cell surface to the nucleus, and it then binds to various cis-regulatory modules that control the transcription of a number of genes involved in Th1 cell development.

STAT4 has many similar functional domains that belong to the STAT family of proteins, such as the DNA-binding domain, a conserved SH2 domain that is involved in dimerization processes and a C-terminal transactivation domain [18].



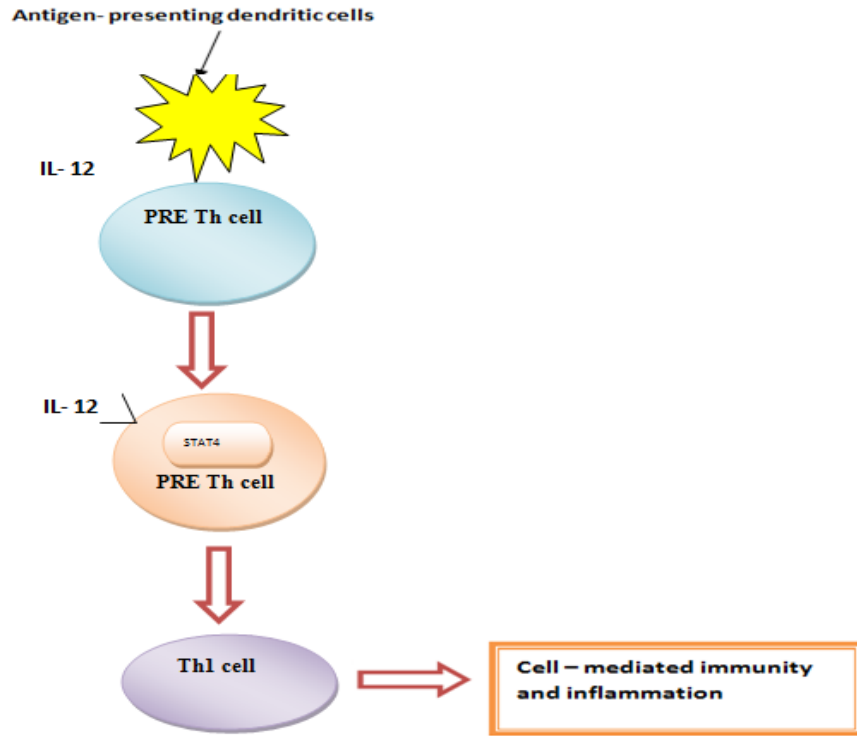
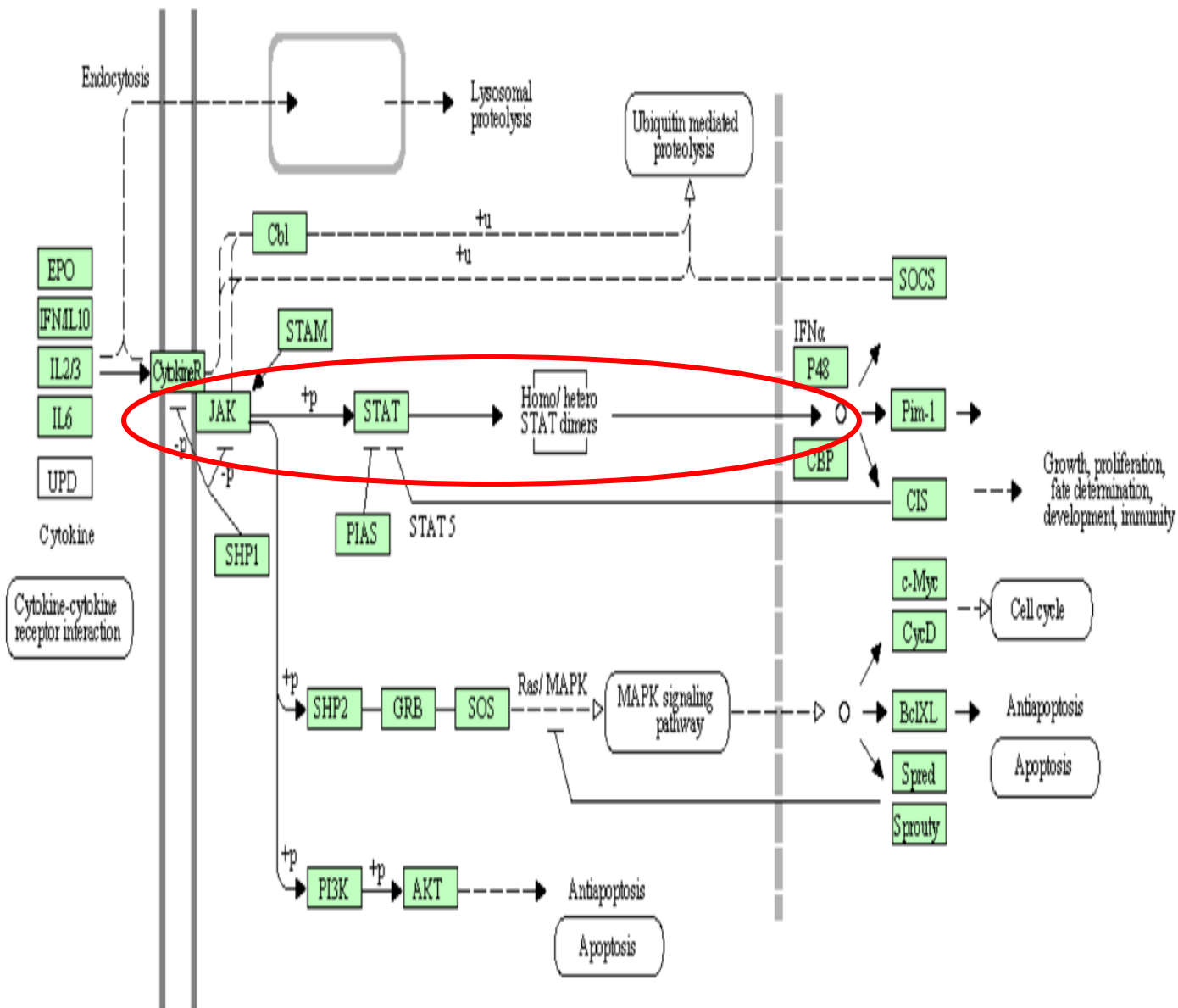


Figure 5: Biology of STAT4 involvement in Th1 cell development.

The activation of STAT4 is mediated by the Jak-STAT (Janus kinase-signal transducers and activators of transcription pathway). In higher eukaryotes, the Jak-STAT pathway initiates the signaling cascades for the stimulation of many cytokines and growth factors. The conversion of homodimer of STAT4 molecules to a heterodimer molecules is also advocated through the Jak-STAT by enhancing the phosphorylation process, which in turn stimulates the conversion of naive Th cells (Th0) to a mature and active state [19].

# JAK-STAT SIGNALING PATHWAY



04630 3/31/09  
 (c) Kanehisa Laboratories

Figure 6: JAK-STAT (Janus kinase/signal transducers and activators of transcription) pathway. The highlighted circle depicts how the STAT dimerization is carried out in the Jak-STAT pathway. (Courtesy: KEGG Database)

## 2.2 CURRENT UNDERSTANDING OF STAT4

Numerous wet lab biological studies have been conducted to decipher the transcription regulation process of the immune system. For example, one study demonstrates the early target genes of IL-12 and STAT4 signaling in Th1 cell differentiation [20], and another study deals with the requirements for the splice forms STAT4 $\alpha$  and STAT4 $\beta$  to mediate responses to IL12 [21]. In addition to these *in vivo* studies, there was also some *in vitro* computational analysis performed to understand the gene transcriptional mechanism of the immune system. Some of the latest computational studies were performed in order to obtain an elaborated knowledge about the significance of STAT4 molecules in Th1 cell development, which will be discussed in this section.

A significant study by Lai et al. deals with the role of STAT4 and STAT6 in epigenetic modification and transcriptional regulation in Th1 cell development. The study analyzes the STAT4 and STAT6 histone methylation ChIP-seq data to investigate the transcriptomics of murine STAT4-mediated Th1 development. The histone methylations focused on in this study were H3k4me3 and H3k27me3. It reports that around 63% STAT4 and 59% STAT6 binding sites were colocalized with H3k4me3 histone methylation patterns in contrary where only 0.3 and 0.5 percentage of STAT4 and STAT6 binding sites, respectively, were colocalized with H3K27me3. From these results, it can be inferred that STAT4 plays a more significant role in promoting the active transcription regulatory network as it promotes H3k4me3 modification patterns more than H3k27me3. On the other hand, STAT6 has an antagonizing, repressive effect, promoting H3k27me3 modification patterns. Thus, the biological significance of the histone methylation process states that transcription is increased by H3k4me3 patterns and negatively regulated in the presence of higher H3k27me3 patterns. This study also identifies a few genes that may bind to either STAT4 or STAT6, which can have negative

regulations in order to provide lineage-specific expression (to Th1 cell lineage or Th2 cell lineage, corresponding to STAT4 or STAT6, respectively) [22].

### **2.3 STAT4 TARGET GENES IN TH1 DEVELOPMENT**

Good et al. identified and characterized STAT4 targets that play a vital role in Th1 cell development and receptor signaling. The few potential STAT4 TF targets that were reported through this study were furin, Ifng, Il12rb2, and Il18r1 [23-25]. Although a number of STAT4 targets are involved in Th1 cell development, the major group of these genes showed a lack of consistency between STAT4 binding and induction of target gene expression. This inconsistency could be attributed to a number of reasons; the strongly binding target gene regulatory sequence need not necessarily induce high levels of target gene expressions, and conversely a weak binder did not always show poor expression. Furthermore, the presence of other TFB motifs near the STAT4 site can either increase or decrease the expression of downstream target genes, which provides confirmatory evidence for this discrepancy. The high throughput of ChIP-on-chip analysis done by Good et al. identified the STAT4 consensus for the binding of STAT4 targets and also potential transcription factor binding sites, such as NF- $\kappa$ B and PPAR $\gamma$ -RXR that play a major role in STAT4-mediated Th1 cell development [23].

The Good et al. study involved a genome-wide analysis of STAT4 ChIP-on-chip data. The ChIP-on-chip experiment was performed on a mouse promoter region of 7.5 kb upstream to 2.5 kb downstream from the transcription start site (TSS) that identified the 3397 STAT4 binding region, or interval sequence that corresponded to 4669 genes. A peak intensity filter cut off at 4 or greater refined this set to 1111 interval sequences corresponding to 1540 genes. The study also identified the

consensus of STAT4 binding region as “TTCNNNGAA”. A *de novo* motif analysis on these 1111 interval sequences identified a couple of potential motifs, such as NF- $\kappa$ B and Ppar $\gamma$ -RXR (apart from STAT4 which may be an essential cis-regulatory module participant in Th1 development). Two of the new potential STAT4 gene targets, Pcgf5 and Mllt3, were also identified from this genome-wide analysis [23].

## 2.4 ROLE OF PPAR $\gamma$ -RXR IN T CELL DEVELOPMENT

Peroxisome proliferator-activated receptor gamma (PPAR- $\gamma$ ) belongs to the family of nuclear receptors that are involved in macrophage development, immune response, and T cell-mediated inflammation [26]. PPAR- $\gamma$  is widely expressed in all major cell types, such as T cells, macrophages, dendritic cells, endothelial cells, and epithelial cells. The nuclear receptor PPAR- $\gamma$ /RXR- $\alpha$  heterodimer is composed of the peroxisome proliferator-activated receptor gamma (PPAR- $\gamma$ ) and the retinoic acid receptor (RXR- $\alpha$ ). The phenotypic effects of PPAR- $\gamma$ -instructed dendritic cells (DC) is to enhance phagocytic activity and lead to the modification of cytokine-production profiles, resulting in elevated natural killer T (NKT) cell activity. PPAR- $\gamma$  are involved at different stages of DC differentiation [27]. The activation of PPAR $\gamma$  at the transcriptional level of DC development impacts the upregulation (TLR4, CD36) and downregulation (IL1R2, IRF4) of the genes involved. Some of these genes contribute immensely to the development of the receptor-specific DC phenotype [28]. T cell PPAR- $\gamma$  is also important in regulating the relative abundance of CD8<sup>+</sup> and CD4<sup>+</sup> T cell subsets. It has a repressive action on NF- $\kappa$ B, thus providing a check over the stimulation of cytokine production. PPAR- $\gamma$ -RXR plays a pivotal role in modulating the JAK/STAT pathway by inhibiting the phosphorylation of specific JAK and STAT molecules, and also the differentiation and activation of Th1 cells [29].

## 2.5 COMPUTATIONAL METHODS INVOLVED IN CRM DISCOVERY

CRM is a complex biological model; it can effectively be deciphered by combining computational algorithms and tools along with the wet lab experiments. For example, the TFBs in each CRM unit are represented by a position weight matrix (PWM) that are obtained by using statistical approaches based on the frequency of the occurrence of the base (A,T,G,C). Databases like JASPAR [30] and TRANSFAC [31] act as repositories that have information pertaining to the PWM of the cis-regulatory DNA sequence and their transacting factors.

Many informatics-based computational approaches have led to the discovery of many tools in order to explore the CRM unit. One such algorithm was developed by Xin et al. that was based on the prediction of CRM, which was built on a probabilistic model of evolution [32]. In the study performed by Sharan et al., a tool called CREME was developed in order to identify and visualize the CRMs in the promoter region for a set of co-regulated genes. Ivan et al. proposed a new algorithm named CSam (CRM sampler) and D2z-set in order to predict cis-regulatory modules without any information about the motifs [33]. Aerts et al. developed a tool suite (called the Toucan) that could predict the CRMs in a set of co-regulated genes. MotifScanner is one of the tools from this suite that detects the pre-defined motifs in a DNA sequence, employing a background model and a probabilistic estimation of the number of hits (or motifs present) [34].

Similarly, there are many computational approaches that are employed to decode the CRM unit that might help biologists to understand the cis-regulatory module involvement in a particular transcriptional regulatory network.

## 2.5 KNOWLEDGE GAP

The various wet lab studies serve as a stepping stone to study the biological significance of STAT4-mediated Th1 cell development. Additionally, the development of technology provides more sophisticated methodologies to study this significance. Some of these methodologies, such as high throughput (e.g. microarray, CHIP-on-chip) and next generation sequencing (like CHIP-seq and RNA-seq), generate a cyclopean amount of data that should be analyzed effectively.

The informatics-based computational approaches are one of the best possible solutions to analyze this huge quantity of data. For instance, scanning of the chip from the CHIP-on-chip experiment may give many false positive signals and background noise that can hinder the specificity of the analysis. The informatics-based algorithms cannot only eliminate the false positive signal but can also normalize the data.

## 2.6 RESEARCH QUESTION

Earlier studies on STAT4 provide a sound insight on the characteristics and the biological importance of the immune response from varied frames of references. Good et al. in his study involving high throughput data identifies crucial and novel target genes and motifs (NF-kB and Ppary-RXR) (Good SR 2009).

The aim of this work is to implement various computational techniques (high throughput approaches) with biological relevance in order to attain the following:

- a. Decoding the STAT4-mediated transcriptional regulatory networks in Th1 cell development.
- b. Aid the identification of potential STAT4 CRM (cis-regulatory modules) involved in Th1 cell development. Furthermore, *in vivo* studies shall serve as an additional validation for the results obtained.



## CHAPTER THREE: MATERIALS AND METHODS

### 3.1 ChIP-on-chip AND MOUSE GENOME DATA

ChIP-on-chip (Chromatin immunoprecipitation–on-microarray) is a genome-wide location analysis technique. This technique identifies and analyzes DNA fragments (or binding sites) that are potentially bound by specific DNA binding proteins called the transcription factors. The wet lab protocol of this technique involves four steps: cross-linking of the DNA fragment with proteins, sonication of the DNA fragment into small pieces, performing immunoprecipitation of DNA-bound proteins with monoclonal antibodies, and purification and hybridization of the DNA fragments (or binding sites) on a micro array chip [35].

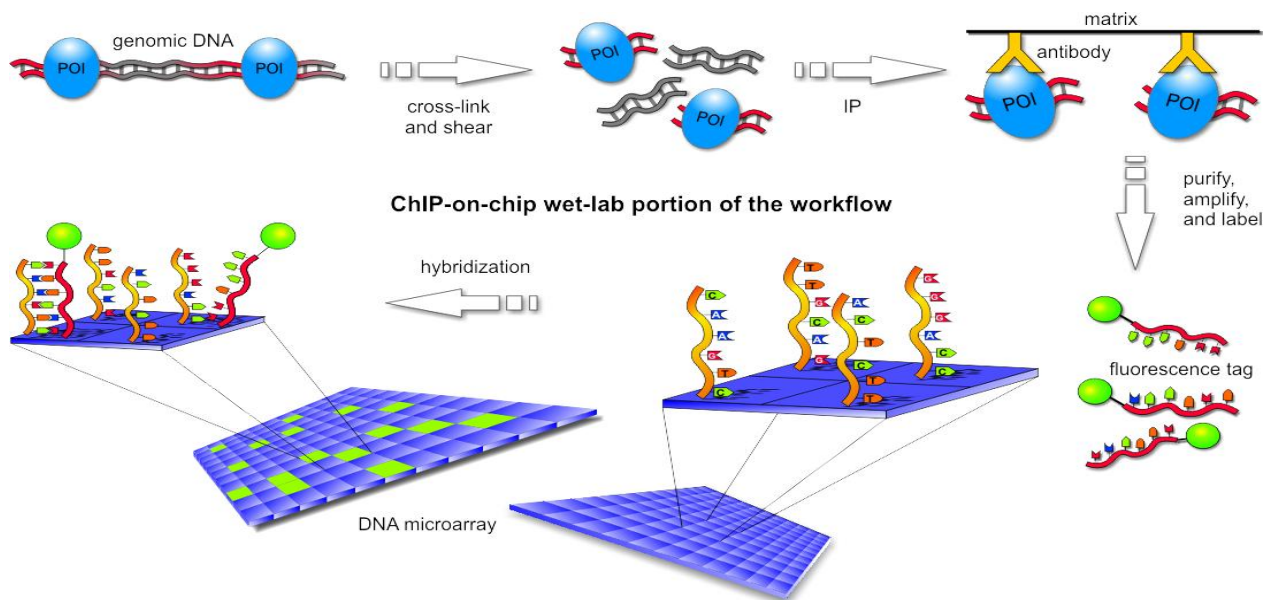


Figure 7: Wet lab ChIP-on-chip work flow [source: [www.Wikipedia.org](http://www.Wikipedia.org)]. The steps observed in this experiment are a) cross-linking and sheering b) immunoprecipitation c) purifying and labeling with

fluorescence tag d) hybridization followed by DNA microarray.

In this study, Th1 cells stimulated by IL-12 are used to generate the STAT4 ChIP-on-chip data. As identified by Good et.al, this data set has 3396 “interval” sequences (with STAT4 binding sites) corresponding to 4669 target genes. A measure used to quantitatively estimate STAT4 binding in their data set was the “peak binding intensity” with a range of values for all the interval sequences from 2.2 to 32.4. In their analysis, they employed a peak binding intensity filter of 4, which reduced the data set to 1111 interval sequences for 1540 genes. We have considered these data sets in this study, viz., the whole set of 3396 interval sequences, and the peak intensity-filtered 1111 sequences [23].

To work with the above data, three sets of sequences were generated based on the 3396 interval sequences using the mm9 build of the mouse genome. Each interval sequence with a STAT4 binding site was extended on either side (or trimmed equally on both sides) to 2000 bases, forming the foreground region (labeled as *Foreground*). Taking the nucleotide at the end of the foreground region, a sequence of a length of 2000 bases was taken to form *Background 1* (3' or downstream of the interval sequence), and from the start nucleotide of the foreground region, an upstream sequence of a length of 2000 bases was taken for *Background 2* (5' or upstream of interval sequence). These three sequence sets for each of the interval sequences were equal in length so as to act as controls for the sequence length in the cis-regulatory modules (CRM) enrichment studies, and the proximity of the background sequences at the 5' and 3' ends of the foreground sequence was to control the GC content of the data.

The STAT4-dependent genes identified from the ChIP-on-chip experiments were then classified into

three categories based on their temporal expression patterns induction. There are minimal to no induction genes set expressed as fold induction patterns of less than or equal to 2 at 4 hours and 18 hours; transient induction genes sets expressed as fold induction patterns of less than or equal to 2 at 4 hours and greater than or equal to 2 at 18 hours and sustained induction genes sets that expressed a fold induction pattern of greater than or equal to 2 at 4 hours and 18 hours [23].

### **3.2 DATABASE FOR ChIP-on-chip DATA**

The need to access both the gene and interval information for establishment of a relationship between a gene and interval would be the first stepping stone for the data analysis. However, the raw data provided by the Genpathway Company has the gene and interval information in separate sheets. To overcome the above difficulty, the data was loaded onto various tables into a MYSQL database on the server (belonging to the School of Informatics, IUPUI). This database depicts in a tabular format the information from Genpathway, like the gene name, chromosome, interval start and stop, interval length, peak value, distance from TSS, and position.

Statistically this database has around 4699 genes and 11680 interval ids corresponding to 6818 interval ids of IL12 (replicate 1 and 2) and 4862 interval ids of IL23 (replicate 1 and 2 ). A frontend in PHP was created in order to retrieve information conveniently from the database.

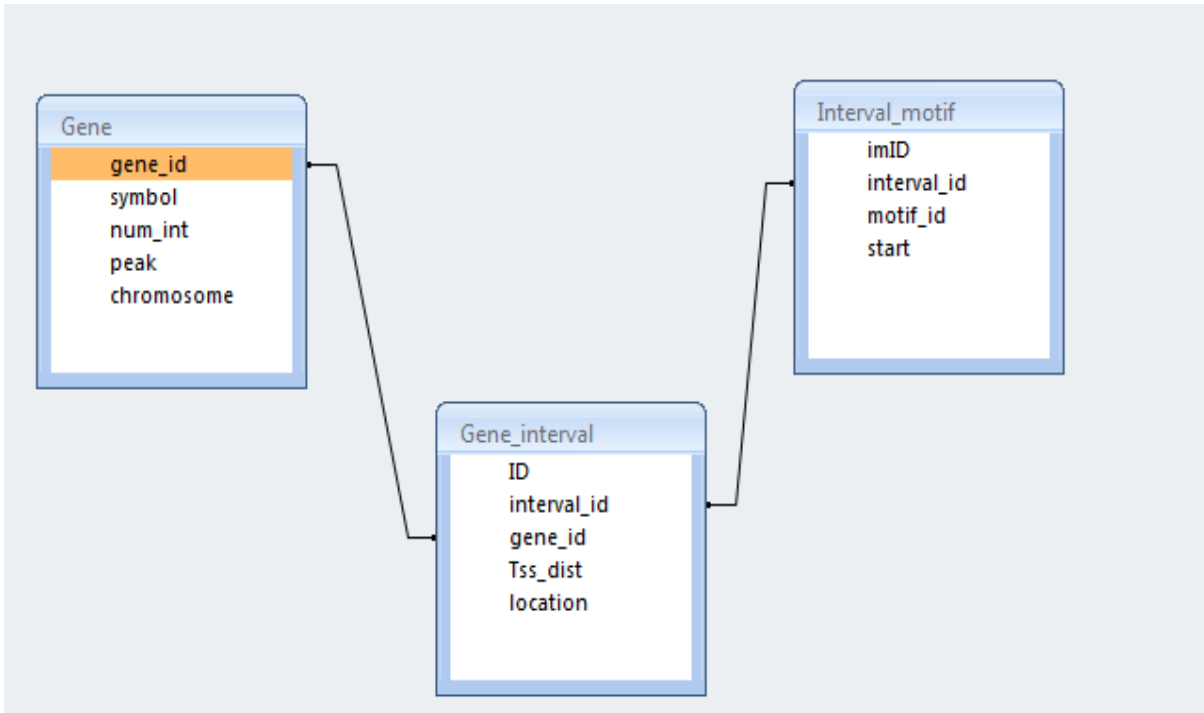


Figure 8: ER Diagram for the ChIP-on-chip database. It contains gene information (like gene ID, gene symbol, chromosome located, peak intensity) and motif information (like motif ID, distance from TSS, location, and start position of the motif).

### 3.3 SEQUENCE EXTRACTION

The locations of the 3396 ChIP-on-chip interval sequences pertaining to the IL12 replicated in a *bed* file format were employed to get the foreground and two background sets.

These interval sequences were mapped onto the UCSC Genome Browser [36] to convert them into the corresponding mouse genome mm9 build locations, and the appropriate sequences were then retrieved from the mm9 2bit file downloaded from the USCSC genome repository. First, the 2bit file was converted to a FASTA file on the command line (using instructions from the web site) [37], and then a Perl script (Appendix A) was written and run locally to extract the sequences corresponding to

the foreground and backgrounds. Essentially, the Perl script centered the start and end locations of an interval sequence (smaller than 2 kb) onto the appropriate chromosomal locations and extended equally on both sides to get the 2 kb sequence. For interval sequences already larger than 2 kb, the script trimmed the isolated sequences equally on both sides to get 2 kb. These sets of sequences containing the interval sequences were labeled the *Foreground*. We also isolated 2 kb segments of background sequences from both of the downstream (3') and upstream (5') sides of the foreground sequences, and these were labeled *Background 1* and *Background 2*, respectively.

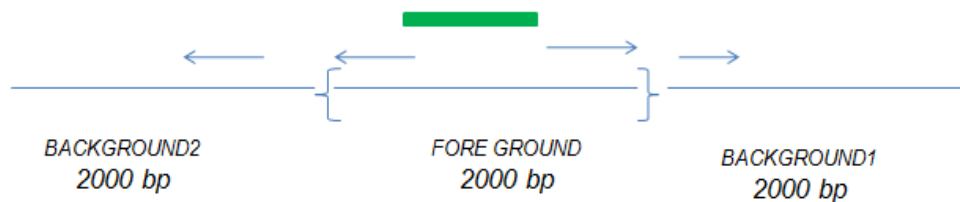


Figure9: Sequence Extraction of three regions a foreground and two backgrounds based on the interval sequence (green colored region).

A sequence of a length of 3300 kb (-3000 kb upstream and +300 kb downstream), corresponding to three sets of genes based on induction patterns, are retrieved using a sequence fetch tool called Discern [38], developed by the UITS at IUPUI, and the output is parsed using Perl code (Appendix B) to obtain the sequences in FASTA format (that is used for downstream analysis in MEME).

### 3.4 REPEAT MASKER

The regulatory sequences of the species *Mus musculus* include *repeats* in the regions that can share the same properties of motifs and hence can show up false positive in the identification of the regulatory modules. To avoid this, we masked the repeats in the obtained sequences by using a tool called “Repeat Masker” [39].

### 3.5 *de novo* MOTIF DISCOVERY

The *de novo* motif discovery is done using an online tool called MEME (Multiple Expectation Maximization for Motif Elicitation). MEME is based on an expectation-maximization motif search algorithm to discover motifs in a group of related DNA sequences. The identified motifs are represented as position probability matrices that describe the probability of each possible letter at each position in the pattern. MEME prediction results provide biologically meaningful motifs that are selected based on the information content, or the measure of motif strength in terms of conserved a position (i.e., the more conserved a position is and the rarer the conserved letters, the higher the information content is), and the number of occurrences for that motif [40].

In this work, we have used MEME to do a *de novo* motif search on the three sets of temporal induction pattern genes (no induction, transient, and sustained), and we also performed a location analysis of STAT4 sites with respect to the two potential motifs NF- $\kappa$ B and PPAR- $\gamma$ /RXR sites identified by Good et al. from their MEME analysis.






	Motif	% of Intervals	% of genes (E-value)
1		59.4%	
2		100%	100% ( $2.1 \times 10^{-181}$ )
3		5.2%	6.4% ( $5.8 \times 10^{-88}$ )
4		5.4%	6.4% ( $2.3 \times 10^{-95}$ )
5		7.8%	9.0% ( $4.1 \times 10^{-131}$ )

Figure 10: WebLogo representations from MEME analysis, where motif 2 is STAT4 and motif4 is NF-kB and motif 5 is PPAR $\gamma$ - RXR [23]

### 3.6 DNA MOTIF COMPARISON

The potential *de-novo* motifs identified from MEME are compared against many pre-existing motif databases to check for similar motifs using a web server called STAMP [41]. It queries the input of *de novo* motifs against databases of known motifs; the input motifs are aligned against the chosen database from a list. The results thus obtained give the multiple alignments of the input motifs (when two or more motifs are provided in the input), a similarity tree (when three or more motifs are provided in the input), and a ranked list of matches in the chosen dataset for each input motif [41].

### 3.7 GENE ONTOLOGY ANALYSIS

We used the GOstat tool to classify genes containing various CRMs into GO categories. GOstat is a gene ontology-based tool that is used to annotate and analyze the function of a list of genes [42]. The list of mapped genes obtained after filtering based on the threshold of the 500 bases' distance cut-off between NF- $\kappa$ B and PPAR- $\gamma$ /RXR motifs and STAT4 was submitted to the GOstat tool. Sequentially, the results in the form of molecular function and biological process annotations were obtained for the submitted gene lists with a corresponding p-value (with a cut off of 0.1).

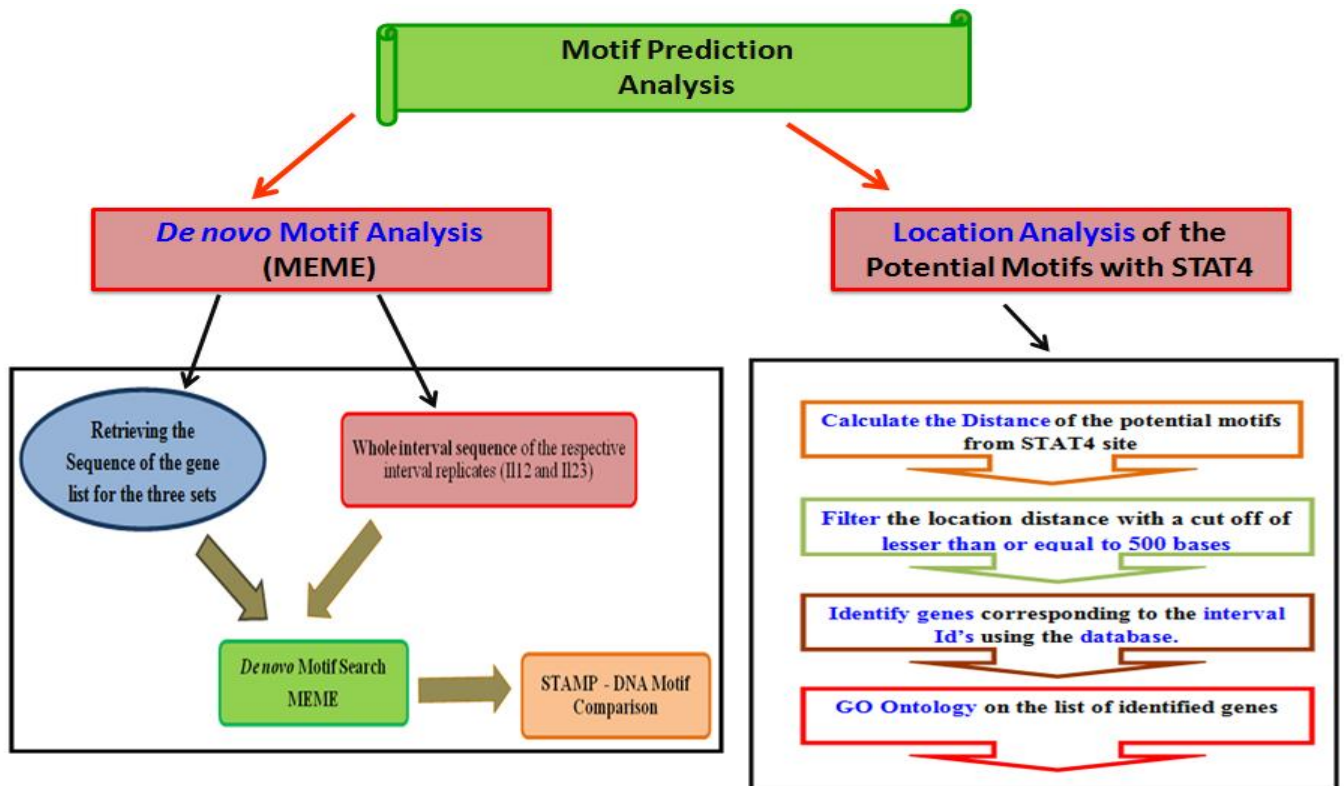


Figure 11: Work flow describing the steps involved in motif prediction analysis. This consists of two analysis a) *de novo* motif analysis and b) location analysis of potential motif with STAT4



### 3.8 CRM (CIS-REGULATORY MODULE) PREDICTION

The CRM analysis was done using a tool called MotifScanner [34]. MotifScanner is from the Toucan2 suite that analyzes regulatory and co-regulatory sequences. The PWMs (position weight matrices) serve as inputs in the TRANSFAC format. The sequences of the genes were given in a FASTA format while the test sequences were run against a species-specific background model (*Mus musculus* in our case). The tool locates motifs in the input sequences and calculates their corresponding scores (based on the probabilistic estimation of the number of hits). It displays the consensus of the enriched motifs in the sequences and also the strand on which they are present.

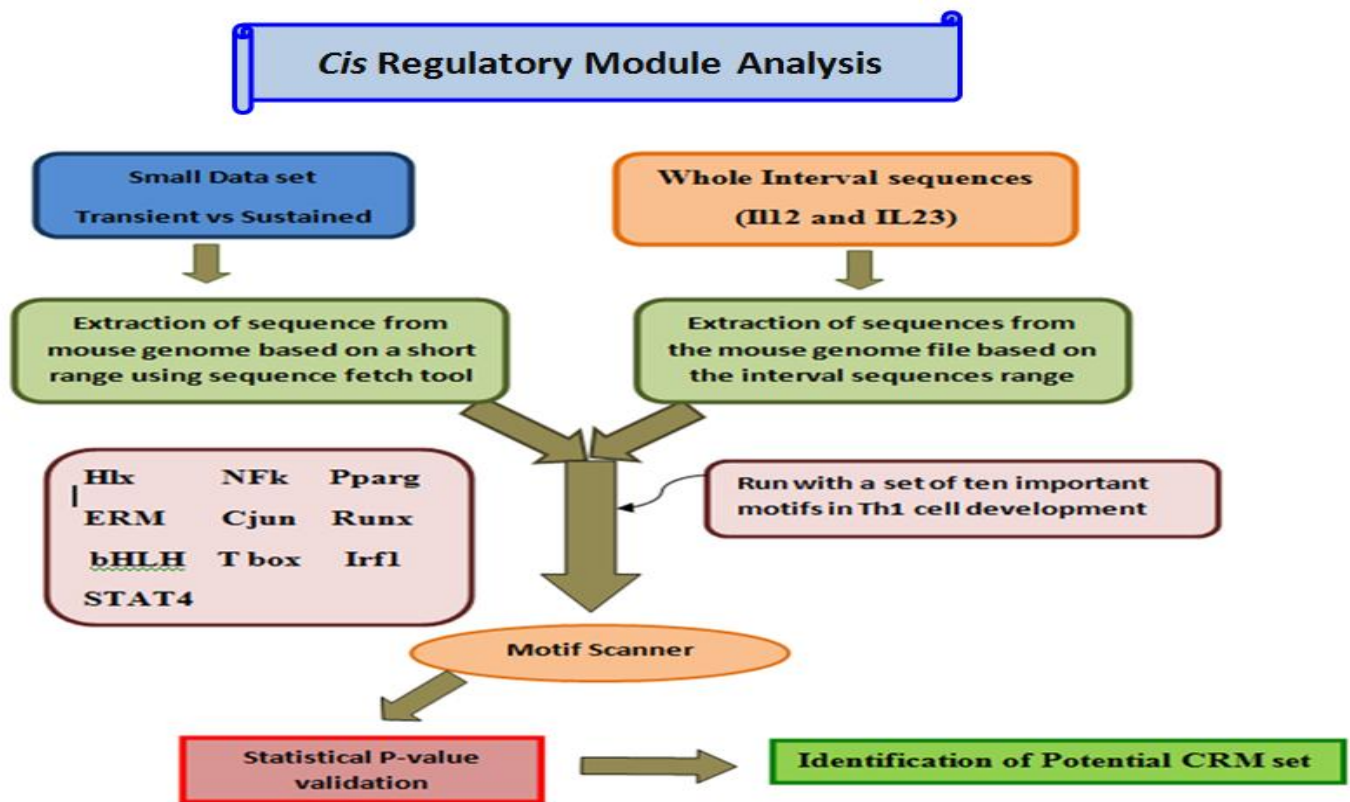


Figure 12: Work flow describing the steps involved in potential CRM prediction analysis. This analysis is done with both small data set and whole interval sequence to identify potential CRM.

### **3.9 MAPPING OF ChIP-on-chip AND ChIP-Seq DATA**

The STAT4 ChIP-seq data was mapped with STAT4 ChIP-on-chip data using Perl codes (Appendix C). The concept behind the mapping is to identify the STAT4 ChIP-seq tags that are located in a particular STAT4 ChIP-on-chip interval sequence range. Using the coordinates of interval sequence, the STAT4 ChIP-seq tags corresponding to each interval sequences are retrieved.

The interval regions pertaining to the ChIP-on-chip data has a small coverage area (between 7.5 kb upstream to 2.5 kb downstream to the TSS) whereas the tags pertaining to ChIP-seq data covers the entire mouse genome. The length of interval sequence varies from 204 bases to 4710 bases, whereas the length of ChIP-seq tags is only 200 bases. Among 3542 STAT4 ChIP-seq tags, only 553 ChIP-seq tags were mapped to 192 unique STAT4 ChIP-on-chip interval ids, whereas there were 2984 STAT4 ChIP-seq data that were not mapped or did not lie in the STAT4 ChIP-on-chip interval range.

### **3.10 COLOCALIZATION OF METHYLATION PATTERNS**

The CRM prediction analysis predicted the potential STAT4 CRM that can play a potential role in Th1 cell development (in our case the CRM identified is STAT4-PPAR- $\gamma$ -RXR CRM). The common interval ids that carry both STAT4 and PPAR $\gamma$ -RXR sites are taken, and these are mapped to their corresponding genes using the database. The coordinates of these genes are then collected from the UCSC genome browser, and subsequently, these coordinates are modified based on the length of the interval sequence. Furthermore, 2000 bp upstream and downstream were then added to these modified coordinates, thus forming three regions: one foreground where our region of interest lies and two backgrounds for validation purposes.

In co-localization analysis, we mapped the STAT4 ChIP-seq histone methylation data (K4, k27 and k36) to the three regions: the foreground and two backgrounds. This analysis provides us the information of high histone methylation patterns expressed in the three regions by obtaining the normalized tag counts [43].

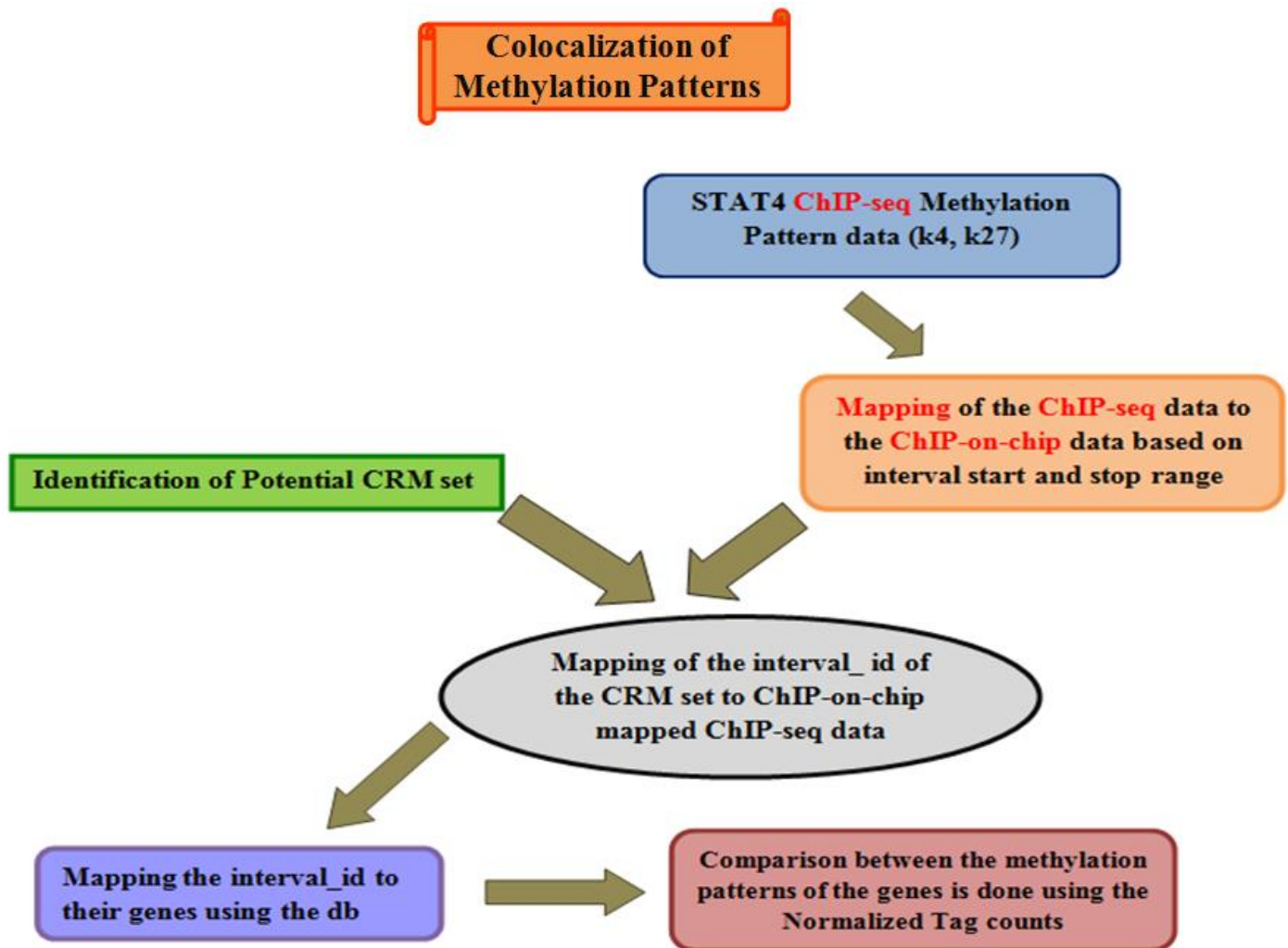


Figure 13: Work flow describing the steps involved in colocalization of methylation patterns.

### 3.11 CONSERVATION OF THE POTENTIAL CRM SET

The potential genes shortlisted for the colocalization analysis were also studied for the conservation of the STAT4-PPAR $\gamma$ -RXR CRM. For determining the conservation, we used a web server called ECR browser [44]. The organism name (in our case, *Mus musculus*) and the chromosome region are given as inputs. This gives highly conserved regions among a number of organisms and details, like the sequences of other organisms (such as chickens, dogs, human build 19, chimps, etc.), length, mapping regions with other organisms, and so forth. All these regions can be viewed in separate pop-up windows, which provide links to check the conserved transcription factor binding site (TFBS) regions in the area of interest.

The conserved TFBS regions are identified by a tool called rVista [45] from the ECR browser tool suite. The input parameters include selecting a biological species (in our case we selected vertebrates) and then setting the matrix similarity to a predefined value of 0.75, which is to ensure that it will map the matrix if it finds 75 similar with the input motifs. The output further enables us to identify the number of STAT4 and PPAR $\gamma$ -RXR TFBS of our input species (*Mus musculus*) and other organisms that displayed conservation regions.

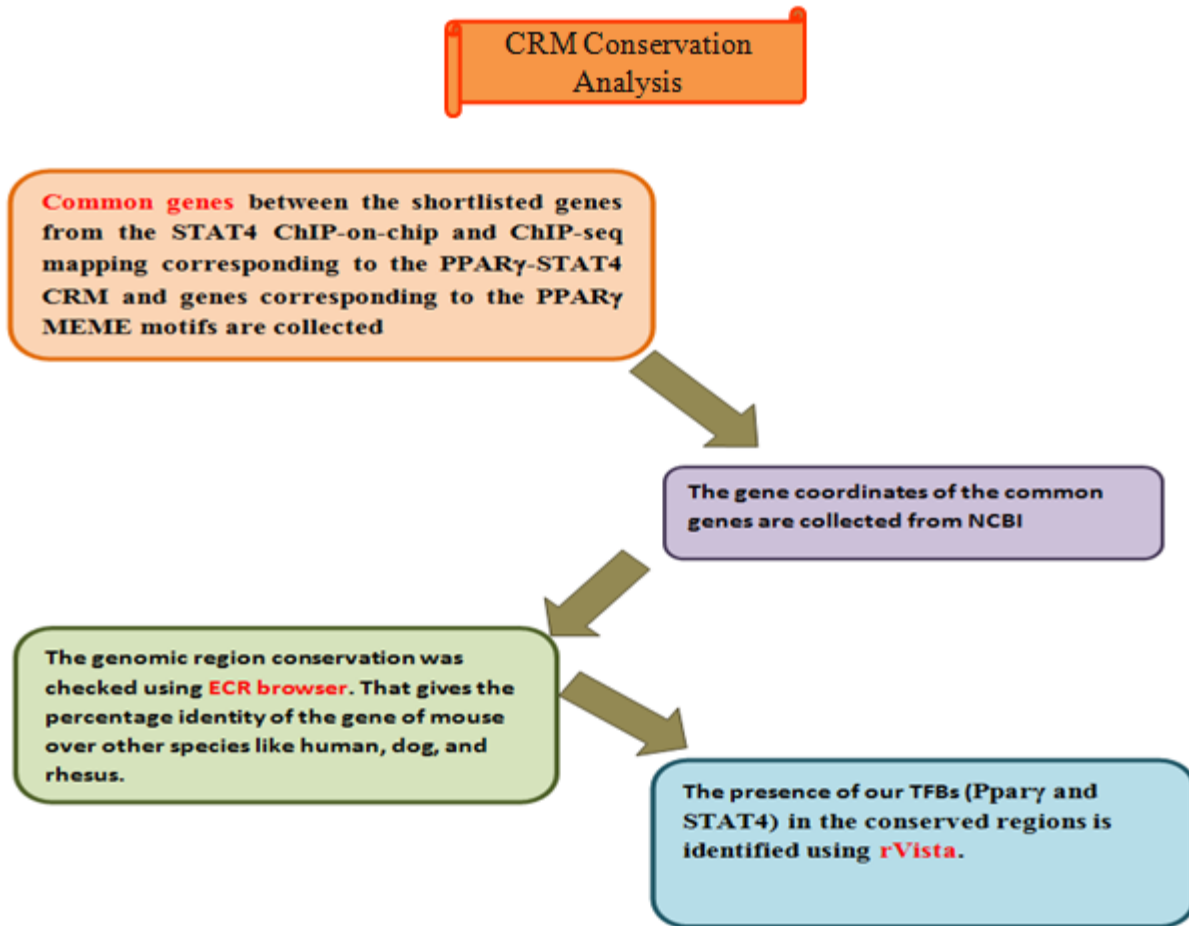


Figure 14: Work flow describing the steps involved in conservation analysis performed of the genes corresponding to potential CRM.

## CHAPTER FOUR: RESULTS

### 4.1 *de novo* MOTIF ANALYSIS

The *de novo* motif analysis performed on the two temporal induction pattern data sets, transient and sustained (Figure 15 gives a graphical representation of the binding intensity these genes), used MEME and identified few promising *de novo* motifs that were extracted using automated codes. The MEME identifies a set of 30 *de novo* motifs with their lengths ranging from 6 to 15. These motifs were further ranked based on their E-values and information contents. According to the results, a few potential *de novo* motifs in each of the temporal pattern induction sets (sustained and transient) were identified, as shown in Table 1.

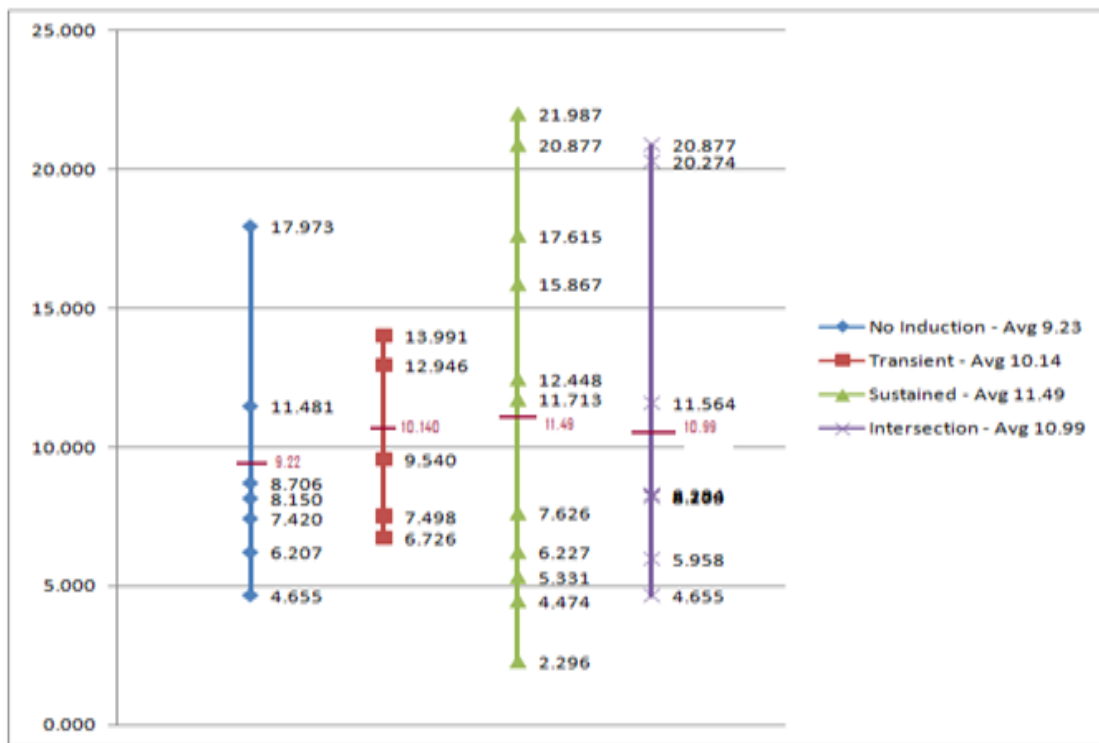


Figure 15: The four sets of genes based on the temporal induction patterns that are plotted based on their peak binding intensities.

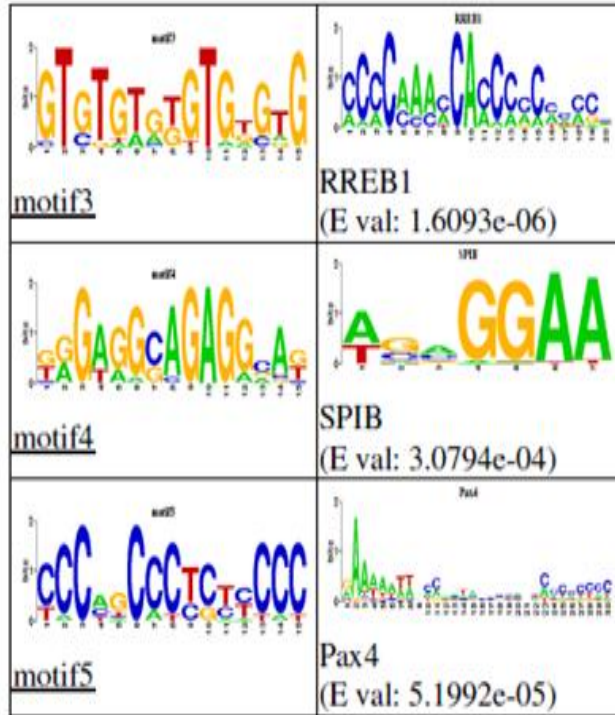
Table1.a: Potential *de novo* motif for the sustained gene set

<i>De novo</i> Motif	Motif Width	No. of Sites (out of 12 genes)	E-Value	Information Content(bits)
a				
<b>Motif 3</b>	15	11	4.7 e <sup>-003</sup>	24.7
<b>Motif 4</b>	15	12	2.2e <sup>-001</sup>	23.8
<b>Motif5</b>	11	12	6.0e <sup>+000</sup>	17.7

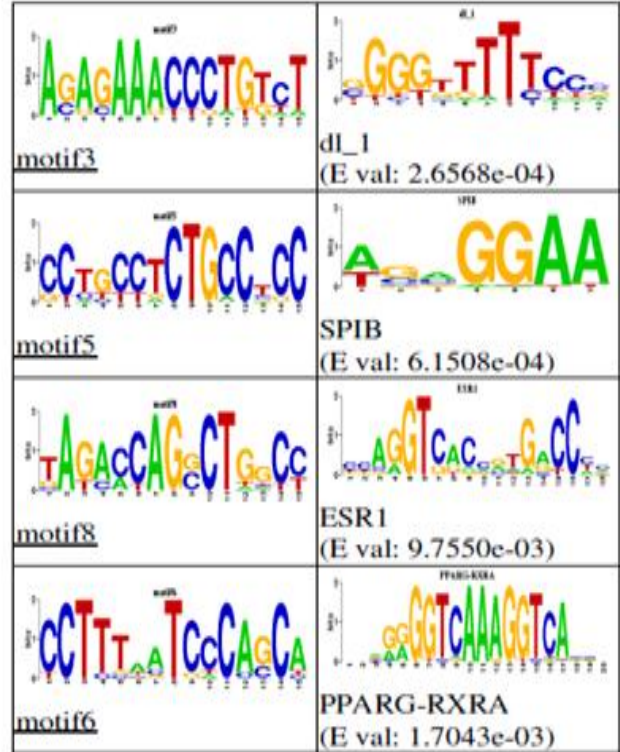
Table1.b: Potential *de novo* motif for the transient gene set.

<i>de novo</i> Motif	Motif Width	No. of Sites ( out of 10 genes)	E-Value	Information Content(bits)
<b>Motif 3</b>	15	10	3.1e <sup>-002</sup>	21.0
<b>Motif 5</b>	15	5	2.0e <sup>+002</sup>	26.1
<b>Motif6</b>	11	9	3.0e <sup>+003</sup>	18.4
<b>Motif8</b>	11	10	9.7e <sup>+003</sup>	16.8

In the transient gene set, motifs 3, 5, 6, and 8 were identified as potential *de novo* motifs, whereas in the sustained gene set we identified motifs 3, 4, and 5 as potential motifs. After shortlisting the potential *de novo* motifs from the two gene sets, the next step was to annotate them. STAMP was used to annotate these potential *de novo* motifs. The PWMs (position weight matrices) of these motifs were parsed out from the MEME result file and were given as inputs to the STAMP tool that annotated the motifs of the two gene datasets, as below:



Fig(a)



Fig(b)

Figure 16: Annotation of potential *de novo* motifs from MEME identified using STAMP a) for the sustained gene set b) for the transient gene set.

#### 4.2 LOCATION ANALYSIS OF STAT4 CRMS WITH NF-KB OR PPAR $\gamma$ /RXR SITES

In our earlier work on the CHIP-on-chip data [23], two *de novo* motifs were identified in the 1111 STAT4 interval sequences using MEME, an expectation-maximization motif search algorithm [40]. Upon a JASPAR database [30] analysis, these two motifs were characterized as being similar to the NF- $\kappa$ B and PPAR $\gamma$ /RXR sites [23]. The location distance of either of these sites from the STAT4 site was calculated, and a manual filter of an absolute distance of  $\leq 500$  was employed to restrict the



interval sequences carrying potential STAT4 CRMs. This gave us a list of interval sequences with the NF- $\kappa$ B or PPAR $\gamma$ /RXR binding site in close proximity to STAT4 in the sequence (Table 2). CRM prediction algorithms have focused on a spacing threshold of 200 bases for the relevant TFBs located together [46], and hence our approach, albeit less stringent, should find relevant CRMs. In order to look for biological relevance in the predicted sequences and CRMs, we mapped the interval sequences to their respective target genes. Table 2 shows a partial list of interval sequences that carry the PPAR $\gamma$ /RXR site, their corresponding target genes, distance from the STAT4 site, and their peak binding intensity. The binding intensities do not seem to have any correlation to the spacing distance between the two TFBs in the analyzed interval sequences.

**Table 2:** Location distance analysis of STAT4 CRMs with PPAR $\gamma$ /RXR site (partial list)

Interval	Length	Start	Distance from STAT4	abs value	Genes	Peak Intensities <sup>a</sup>
<b>IL12::1::2782</b>	396	175	44	44	Gng8,Ptgir	7.510
<b>IL12::1::2532</b>	632	185	-377	377	Centa1	4.756
<b>IL12::1::2463</b>	523	307	70	70	Cdc7	9.523
<b>IL12::1::2173</b>	518	97	-195	195	Rod1	4.257
<b>IL12::1::1998</b>	994	933	468	468	Ccn11	4.765
<b>IL12::1::2615</b>	3279	2024	494	494	<b>Gimap1,Gimap5</b>	8.834
<b>IL12::1::1296</b>	1278	496	321	321	<b>Mina,Gabrr3</b>	8.918

*a - The peak intensities are taken from the ChIP-on-chip data [23]*

### 4.3 GO ANALYSIS

Next, we mapped the lists of genes corresponding to the CRM-carrying interval sequences onto GO molecular function and biological process annotations using the Gostat tool [42]. As shown in Tables 3.A and 3.B for the PPAR- $\gamma$ -STAT4 CRM genes, a number of relevant categories, specifically with respect to immune cell development such as the molecular function, tumor necrosis factor (TNF) super family binding (condensed Table 3.B), the biological process, and lymphoid organ development (condensed Table 3.A), were significantly enriched in the input gene set compared to a mouse genome background. Some interesting genes with potential STAT4 CRMs were identified during the location analysis process, such as Mina and the Ltb-Tnf-Lta complex (Table 2), which may play a putative role in activating the STAT4 pathway for development of Th1 cells [47-48], and at least in the case of the latter gene complex, the GO analysis (Table 3.A, Table 3.B) indicated enrichment in our data set.

**Table 3.A:** Gostat Biological process. This is the result for PPAR $\gamma$ -Stat4 CRM associated genes obtained from Location analysis (partial list)

Best Gos	GO TERM	Genes	Count	Total	P-value
<b>GO:0006334</b>	Nucleosome assembly	hist1h3b hist1h1c hist1h2bb hist1h3c hist1h2ab	5	94	0.0105
<b>GO:0065004</b>	Protein-DNA complex assembly	hist1h3b hist1h1c hist1h2bb hist1h3c hist1h2ab	5	120	0.0105
<b>GO:0006333</b>	Lymphoid organ development	hist1h3b hist1h1c hist1h2bb hist1h3c hist1h2ab	5	131	0.0132
<b>GO:0032602</b>	Chemokine production	sigirr slc37a4	2	8	0.0264

**Table 3.B:** GOSTat Molecular function. This is the result for PPAR $\gamma$ -Stat4 CRM associated genes obtained from Location analysis (partial list)

Best Gos	GO TERM	Genes	Count	Total	P-value
GO:0004743	Pyruvate kinase activity	pkm2	1	2	0.0839
GO:0043120	Tumor necrosis factor binding	tnfrsf4	1	4	0.0839
GO:0004594	Pantothenate kinase activity	pank2	1	4	0.0839
GO:0005031	Tumor necrosis factor receptor activity	tnfrsf4	1	4	0.0839
GO:0003950	NAD <sup>+</sup> ADP-ribosyltransferase activity	parp6	1	12	0.1
GO:0005026	Transforming growth factor beta receptor activity, type II	amhr2	1	1	0.0839

#### 4.4 CRM ANALYSIS

A cis-regulatory module analysis was performed on the whole STAT4 binding data set from the ChIP-on-chip experiment. As described earlier, the interval sequences of the 3396 foreground and the corresponding two backgrounds (Background 1 and Background 2) were compared in this analysis. For this, a tool called MotifScanner from the Toucan2 suite [34] was used to look for TFBs located in the foreground and background sequences. The position weight matrix (PWM) of the motifs NF- $\kappa$ B, PPAR $\gamma$ /RXR, Tbox, C-jun/Ap1, HLX, Runx, and STAT4 were inputted to scan for their locations on the input sequences. The outputs also included scores based on the probabilistic estimation of the number of TFB hits. The motifs, apart from the three previously MEME-identified motifs, were included since they also played a potential role in Th1 cell development. For example, the Tbox motif was included in this analysis since the corresponding TF, viz., and T-bet have been implicated as the master regulators of IL12-mediated Th1 development [12], and the transcriptional regulatory roles of

STAT4 vs. T-bet in this biological process are not completely delineated [10]. Table 4 shows a sample output from the Foreground data indicating the locations (start and end positions and strand), enrichment scores for the STAT4, and three potential CRM motifs. For validation (next section), we also undertook another similar CRM analysis (using the same PWMs) on only one set of 3396 random sequences from the mouse genome generated with the RSAT tool [49].

**Table 4:** MotifScanner output corresponding to the foreground region (partial list) .The output from motif scanner gives the start of TFBS located on the input region along with strand and the probabilistic score.

INTERVAL_ID	START	END	SCORE	STRAND	TFBS
IL12:1:0	1148	1153	227.995	+	id "RUNX"; site "TGAGGT";
IL12:1:0	1012	1017	2345.17	-	id "RUNX"; site "TGTGGT";
IL12:1:0	1709	1718	161.251	-	id "NFkAPPAB"; site "GGAATATTCA";
IL12:1:0	1251	1257	1292.34	+	id "HLX"; site "ATAATTG";
IL12:1:0	780	786	1755.22	-	id "HLX"; site "TTAATTG";
IL12:1:0	52	58	286.024	+	id "Tbox"; site "CAAGGTG";
IL12:1:0	1934	1940	3344.32	-	id "Tbox"; site "GTAGGTG";
IL12:1:1	1374	1379	2345.17	+	id "RUNX"; site "TGTGGT";
IL12:1:1	524	529	2345.17	-	id "RUNX"; site "TGTGGT";

As expected, MotifScanner predicted a number of hits on all three sets of sequences for the four motifs. In order to control for false positives, we filtered MotifScanner outputs from all three data sets for the presence of STAT4 and the second motif (NF- $\kappa$ B, PPAR $\gamma$ /RXR, Tbox, C-jun/Ap1, HLX, or

Runx), as shown in Figure 14 for the Foreground sequences. Of the foreground sequences that carried the second motif, only about one third of them contained STAT4. This was also the case with the two background sequences, except that the frequencies of occurrence for the various pairs of motifs were consistently smaller compared to the foreground (14 – 25). We statistically validated only the interval sequences that were filtered as shown in Fig. 15 (next section). The number of predicted STAT4 sites in the foreground set (=2615) is less than what is expected, based on the MEME predictions for the whole interval sequences set (3396); this is probably due to the MotifScanner default parameters being more stringent than those in MEME.

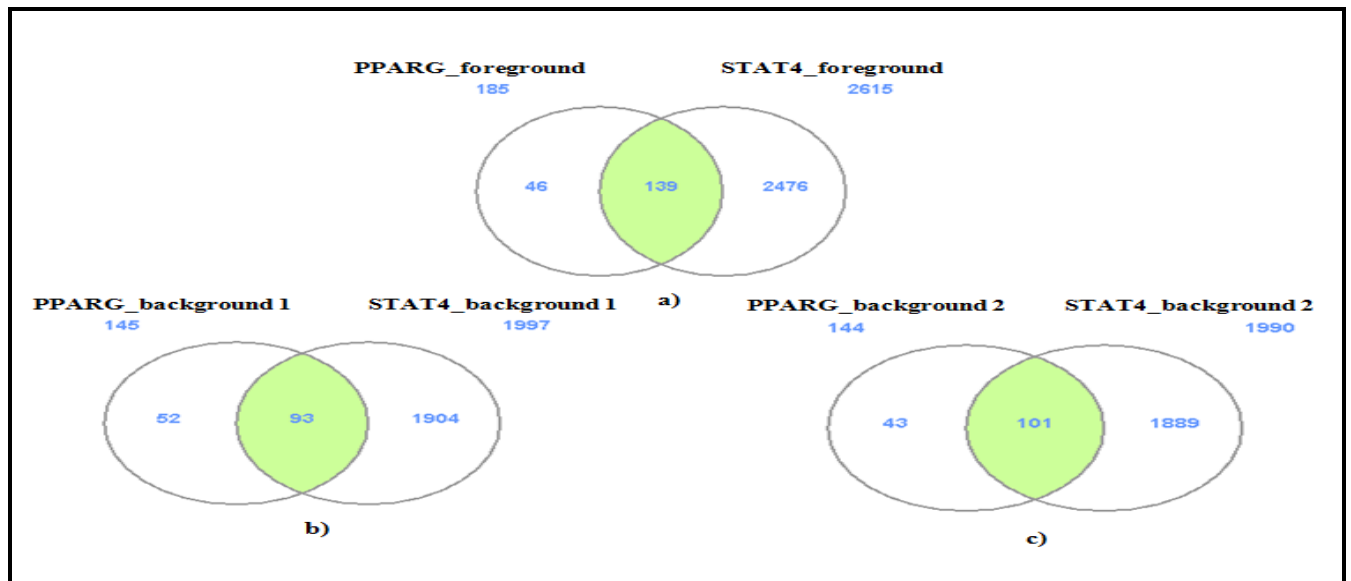


Figure 17: Venn diagram representation. Showing the overlap and unique interval ids between PPAR- $\gamma$ -Stat4 CRM in the three region foreground and two backgrounds.

#### 4.5 STATISTICAL VALIDATION OF THE PREDICTED POTENTIAL CRM SETS

We applied a statistical approach for validating the enrichment of the three CRM pairs, namely NF- $\kappa$ B-STAT4, PPAR $\gamma$ -STAT4, and Tbox-STAT4 in the foreground compared to the two backgrounds. A binomial p-value validation is done on the MotifScanner results from the foreground and the two backgrounds. Essentially, the observed frequency of motif hits is calculated for the foreground and compared to those of the background sequences. A Fisher's exact T-test was done (Table 5), and p-value was calculated using a p-value calculator. The observed count of filtered interval sequences carrying CRM pairs (Fig 15, highlighted area) is a number of binding sites of the test and a number of no binding sites, which is the total number of interval sequences (3380) minus the number of binding sites of the test. Similarly the same values are calculated for the random set of 3380 sequences generated by RSAT and filled up in the control column in Table 5. Thus, the Fisher's method is used to calculate the T-test, and the value is used in a p-value calculator to get the p-value.

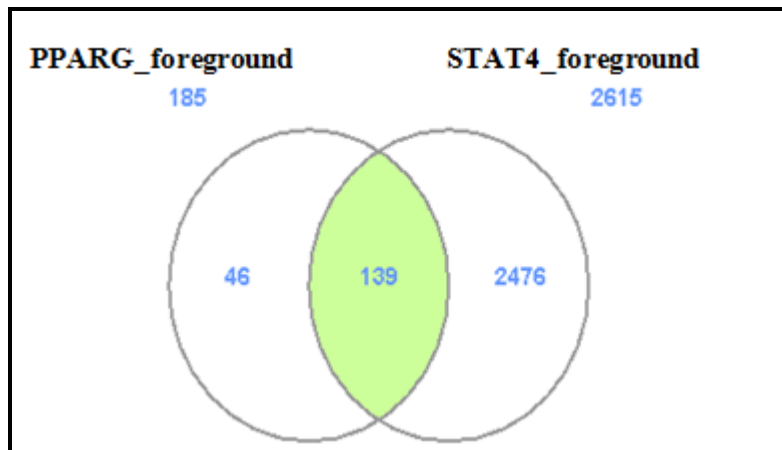


Figure 18: The highlighted region from this Venn diagram is taken as the #successes in p-value calculation. This region is the common interval region shared by PPAR $\gamma$ -STAT4 CRM.

**Table 5:** Format of Fisher’s T-Test calculation for PPAR $\gamma$ -STAT4 CRM from above Venn diagram followed by p-value calculation using the p-value calculator

<b>PPAR<math>\gamma</math>-STAT4 CRM</b>	<b>Test</b>	<b>control</b>		<b>P-value</b>
<b>Binding site</b>	139	76	215	
<b>No Binding site</b>	3241	3304	6545	0.0001
	3380	3380	6760	

Most of the CRM pairs show statistical significance in the observed enrichment seen in the foreground sequence set compared to the two neighboring backgrounds (Tables 6). There are some exceptions wherein one of the two neighboring backgrounds (Background 1 or Background 2) showed a lower p-value compared to the Foreground set (like in the case of NFK\_STAT4, where the background is enriched compared to the foreground). This may be due to a genuine enrichment of this CRM pair in the downstream sequences compared to the original ChIP-on-chip interval sequences and cannot be ruled out as a biological exception.

**Table 6:** P-value validation for CRM enrichment in Foreground sequences vs. the two Background sequences for the whole 3396 ChIP-on-chip interval sequences.

<b>CRMs</b>	<b>Background1</b>	<b>Foreground</b>	<b>Background2</b>
<b>Cjun_STAT4</b>	0.0001	0.4502	0.0001
<b>HLX_STAT4</b>	0.0001	0.0001	0.0001
<b>NFK_STAT4</b>	0.0001	0.0105	0.0001
<b>Runx_STAT4</b>	0.0001	0.0001	0.0001
<b>T-box_STAT4</b>	0.0001	0.0001	0.0001
<b>Pparg_STAT4</b>	<b>0.1859</b>	<b>0.0001</b>	<b>0.0569</b>

**Table 7:** P-value validation for CRM enrichment in Small dataset (Sustained and Transient gene list) pertaining to regions of 2100 base and 1100 base.

	Sustained		Transient	
	-1000 to +100	-2000 to +100	-1000 to +100	-2000 to +100
<b>Cjun_Stat4</b>	0.0848	0.1859	0.0001	0.0001
<b>Hlx_Stat4</b>	0.3144	0.2967	0.0063	0.0001
<b>NFK_Stat4</b>	0.8032	0.2681	0.0309	0.0001
<b>Runx_Stat4</b>	0.0985	0.2967	0.0062	0.0001
<b>Tbox_Stat4</b>	0.3323	0.2488	0.1527	0.0538
<b>Pparg_Stat4</b>	0.728	0.5431	0.6515	0.2598

#### 4.6 MAPPING ChIP-on-chip AND ChIP-Seq DATA

The STAT4 ChIP-seq data was mapped with STAT4 ChIP-on-chip (range covering the whole mouse genome) data using Perl codes (Appendix C). The interval region pertaining to the ChIP-on-chip data has a small coverage area (between 7.5 kb upstream to 2.5 kb downstream to the TSS) when compared to the ChIP-seq data that covers the entire mouse genome. The interval length of the ChIP-on-chip data varies from 204 bases to 4710 bases, whereas it is 200 bases throughout the ChIP-seq data. From Table 8, we can get a clear idea of how the ChIP-seq reads 200 bases; each of which are covered in a single interval sequence region. For example, (from Table 8) we can see that around four ChIP-seq reads are covered in a single interval (IL12::1::22) of a sequence of a length of 1146 bases. Similarly, each ChIP-on-chip interval sequence may cover two or more ChIP-seq reads corresponding to their length.



Table 8: Mapping STAT4 ChIP-Seq data with STAT- ChIP-on-chip data.

Chromosome	ChIP-seq Start	ChIP-seq End	Tag Count	ChIP-on-chip Start	ChIP-on-chip End	Interval_id
chr1	37947000	37947199	0.9	37946571	37947717	IL12::1::22
chr1	37947200	37947399	1.7	37946571	37947717	IL12::1::22
chr1	37947400	37947599	1.3	37946571	37947717	IL12::1::22
chr1	37947600	37947799	0.9	37946571	37947717	IL12::1::22
chr1	38054000	38054199	1.3	38054134	38054483	IL12::1::23
chr1	38054400	38054599	0.9	38054134	38054483	IL12::1::23

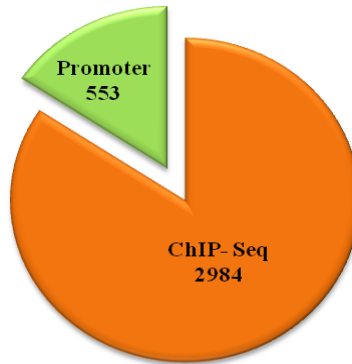


Figure 19: Pie chart displaying the distribution of ChIP-Seq data. 553 ChIP-Seq reads were found in the promoter region that corresponded to 192 ChIP-on-chip interval regions.

The results obtained can be summarized as: only 553 STAT4 ChIP-seq data mapped to the unique 192 STAT4 ChIP-on-chip interval ids, whereas there was 2984 STAT4 ChIP-seq data that was not mapped or did not lie in the STAT4 ChIP-on-chip interval range.

## 4.7 COLOCALIZATION OF STAT4 SITES WITH METHYLATION PATTERNS

The results from the statistical validation analysis predict the potential STAT4 CRM as STAT4-PPAR $\gamma$ -RXR. For performing the colocalization STAT4 sites with H3K4me and H3K27me methylation patterns, the common interval ids between STAT4 and PPAR $\gamma$ -RXR from the Toucan outputs are taken and mapped with the 192 interval sequences that were previously mapped with the STAT4 ChIP-seq data. Thus, the list of common interval ids was reduced from 139 interval ids to 12 interval ids that have corresponding ChIP-seq data. Then these 12 shortlisted interval ids were mapped to their corresponding genes using the database, giving around 18 genes. The coordinates of these mapped genes are collected from the UCSC genome browser, and the coordinates are modified based on their length of the interval sequence. The same concept was applied to retrieve the foreground and the two background sequences for Toucan analysis. After the modification of the coordinates based on the formula  $2000-l/2$  (where  $l$  represents the length of the interval sequence corresponding to the gene) and then 2000 bp upstream and downstream to these modified coordinates thus forming three regions one foreground where our region of interest lies and two background for validation purpose.

In the colocalization analysis, we mapped the STAT4 ChIP-seq histone methylation data (K4, k27 and k36) to the above three regions: a foreground and two backgrounds. This is done using Perl codes (Appendix D) by employing the same strategy that we used in ChIP-on-chip and ChIP-seq mapping. This mapping is also based on their coordinates. All the tag counts corresponding to the ChIP-seq data of each gene were summed to get the normalized tag counts pertaining to each gene. The normalized tag count of each gene for the three histone methylation patterns corresponding to the three regions are tabulated (as shown in the Table 9).

**Table 9:** H3K4me and H3K27me histone methylation patterns. Corresponding to the three regions of a foreground and two backgrounds for the 18 shortlisted genes.

Genes	Interval_id	Peak Intensities	Normalized Tag	K4			K27		
				BG1	FG	BG2	BG1	FG	BG2
<b>Riok1</b>	IL12::1::886	23.83	4.35	3.8	24.4	12.3	-	-	-
<b>Ccn1</b>	IL12::1::1998	5.38	1.70	16	13.6	4	3	-	3.2
<b>Gimap1</b>	IL12::1::2615	8.83	1.66	18.6	40.8	2.8	2.7	3	-
<b>Gimap5</b>	IL12::1::2615	8.83	1.66	18.6	40.8	2.8	2.7	3	-
<b>Ublcp1</b>	IL12::1::440	4.86	0.90	1.6	17	16.8	-	2.1	3.7
<b>Uqcr2</b>	IL12::1::2939	13.62	1.57	8.1	29.1	7.7	-	-	-
<b>Ccnt2</b>	IL12::1::111	11.81	1.47	0.7	19.5	22.9	-	-	-
<b>Acmsd</b>	IL12::1::111	11.81	1.47	0.7	19.5	22.9	-	-	-
<b>Lfng</b>	IL12::1::2539	5.28	1.60	18	20.8	5.1	-	-	-
<b>Ttyh3</b>	IL12::1::2539	5.28	1.60	18	17.4	5.1	-	-	-
<b>Pkm2</b>	IL12::1::3228	7.00	1.60	13.5	25.8	24.6	-	2.7	-
<b>Parp6</b>	IL12::1::3228	7.00	1.60	13.5	25.8	24.6	-	2.7	-
<b>Whsc2</b>	IL12::1::2386	4.99	1.60	7.9	26.7	-	-	-	-
<b>Pdcd4</b>	IL12::1::1670	13.69	1.60	3.8	17.5	9	1.7	-	-
<b>Whsc111</b>	IL12::1::3019	11.33	1.43	4.5	11.6	21.8	1.7	-	-
<b>Letm2</b>	IL12::1::3019	11.33	1.43	4.5	11.6	21.8	1.7	-	-
<b>Zfp36</b>	IL12::1::2812	9.01	1.70	21.8	26.8	26.1	-	-	-
<b>Plekhg2</b>	IL12::1::2812	9.01	1.70	21.8	26.8	26.1	-	-	-

The above normalized tags, corresponding to the three regions identifying the foreground that is the region of interest, shows a higher K4 methylation pattern compared to the two backgrounds. Moreover, the K27 methylation patterns were seen sparsely in the foreground when compared to the two backgrounds.

#### **4.8 CONSERVATION ANALYSIS**

The 18 genes in the subset from the STAT4 ChIP-on-chip and ChIP-seq mapping corresponding to the PPAR $\gamma$ -STAT4 CRM were analyzed for conservation of PPAR $\gamma$ -STAT4 CRM over other species compared to the mouse. The MEME analysis conducted by Good et al. had identified motif 5 as PPAR $\gamma$ /RXR [23]. We collected those interval sequences and mapped them to their genes, using the database we constructed that mapped to around 208 genes. In order to make our experiment more precise, the 18 PPAR- $\gamma$ -STAT4 CRM relevant genes, which were shortlisted in the previous step, were mapped with these 208 genes from MEME analysis to find the common genes between the two sets, and the number of common genes was found to be 11 genes. Therefore, these 11 genes were used for conservation analysis.

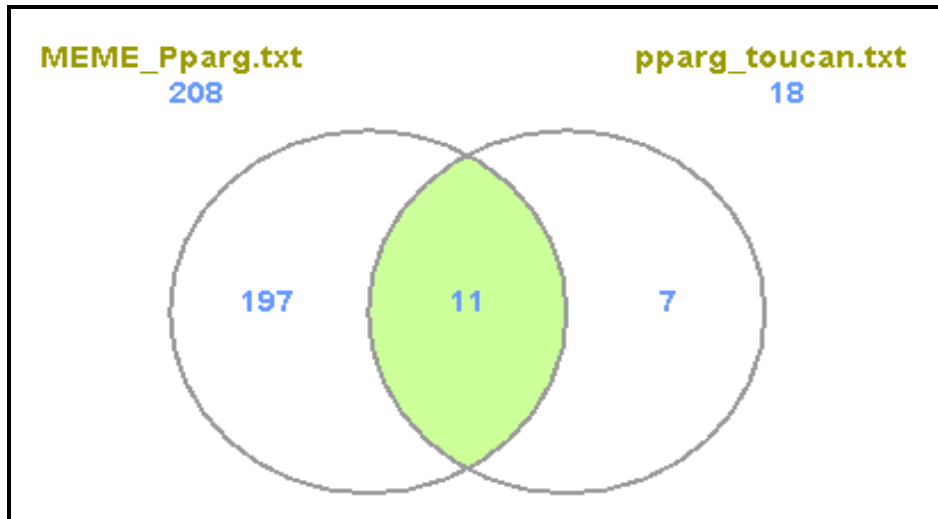


Figure 20: Venn diagram displaying the number of common genes between the genes identified in MEME analysis and the 18 subset of genes obtained from the previous analysis. The highlighted region represent that these two sets had 11 genes common between them.

Among the 11 genes, only 5 of them showed high conservation. These genes showed conservation patterns at three different regions of the 5 genes. Based on their characters, they are intergenic (a stretch of a DNA region located between clusters of genes that contain few or no genes), intronic (a stretch of a DNA region within a gene that cannot translate into proteins), and transposons and repeats (sets of repeats that can be seen at different positions within a genome of a cell). Out of the 5 genes that showed high conservation, genes that showed conservation in the intergenic region were Riok1 and Pdcd4 while the genes that had conservation patterns in the intronic region were Whsc111 and Letm2. The genes that had conservation patterns in the transposons and repeats regions were Ublcp1. The organisms that showed high conservations of PPAR $\gamma$ -STAT4 CRM (apart from mice) were humans, rhesus monkeys, and dogs.

**Table 10:** Percentage identity of the genomic region of the 5 genes corresponding to various mammalian species compared with the mouse at three regions: intergenic, intronic, and transposon & repeats.

Genes	Region	Identity with Mouse		
		Human	DOG	Rhesus
<b>Riok1</b>	Intergenic	70.23	-	67.83
<b>Pdcd4</b>	Intergenic	69.00	-	69.70
<b>Whsc1l1</b>	Intronic	71.70	71.40	-
<b>Letm2</b>	Intronic	68.95	-	71.95
<b>Ublep1</b>	Transposon and repeats	70.60	77.60	77.90

Table10 gives a brief idea of the conservation found across a few mammalian species, such as humans, dogs, and rhesus monkeys, corresponding to an identical region in mice at various locations of the genes. For example, in the case of the gene, *Riok1*, located on chromosome 13 ranging from 37552898 - 37554138 in the mouse, it corresponds to a similar region in humans located on chromosome 6 ranging from 6704809-6707492 regions with an identity of 70.23%.

Table 11: Frequency of TFBS in various species in comparison to mouse. This table gives the total number of the TFBS and the number of PPAR $\gamma$  and STAT4 TFBS separately.

Genes	Species	No. of our TFBS	No. of Our TFB's		Species	No. of our TFBS	No. of Our TFB's	
			Ppary	STAT4			Ppary	STAT4
<b>Riok1</b>	Mouse	21	3	13	Human	28	7	15
					Dog			
		28	3	16	Rhesus	33	7	17
<b>Pdcd4</b>	Mouse	19	4	11	Human	33	1	18
					Dog			
		25	4	15	Rhesus	37	2	21
<b>Whsc111</b>	Mouse	102	7	71	Human	117	7	71
		51	4	31	Dog	56	2	38
					Rhesus			
<b>Letm2</b>	Mouse	21	3	9	Human	28	3	16
					Dog			
		66	1	45	Rhesus	82	5	57
<b>Ublcp1</b>	Mouse	21	0	19	Human	15	0	12
		48	2	39	Dog	57	4	41
		48	2	38	Rhesus	46	1	36

The above table (Table 11) gives the distribution of TFBs (STAT4 and Ppar $\gamma$ ) of mice to their corresponding similar regions in other mammalian species. For instance, the gene *Riok1* has a percentage distribution of our TFBs of about 14% and 62% (PPAR $\gamma$  and STAT4TFBs) corresponding to the mouse when compared to the 25% and 54% (PPAR $\gamma$  and STAT4TFBs) of TFBs distribution in humans. Hence, the conservation analysis results show that out of 11 genes that were shortlisted for biological validation purposes, only 5 genes showed high conservation at different gene locations (intergenic, intronic, and repeats). Also, the percentage distribution of our two TFBs (STAT4 and Ppar $\gamma$ ) of these 5 genes clearly suggests that the frequency of occurrence of our two TFBs seems to have some biological relevance that can be validated with further experiments.



## 4.9 WET LAB VALIDATION

A wet lab validation was performed on the 5 genes that showed conservation from the above analysis to check for IL-12 induced stimulation. The experiment was also performed with two PPAR $\gamma$  agonists, viz. AZPC and ciglitazoneto to check for stimulation by PPAR $\gamma$ . The results were collected at a time period of 4 hours and 18 hours. Furin was used as a positive control. Out of the five genes Riok1 exhibited IL-12 induced stimulation whereas the gene Pdc4 exhibited stimulation with PPAR $\gamma$  agonist.

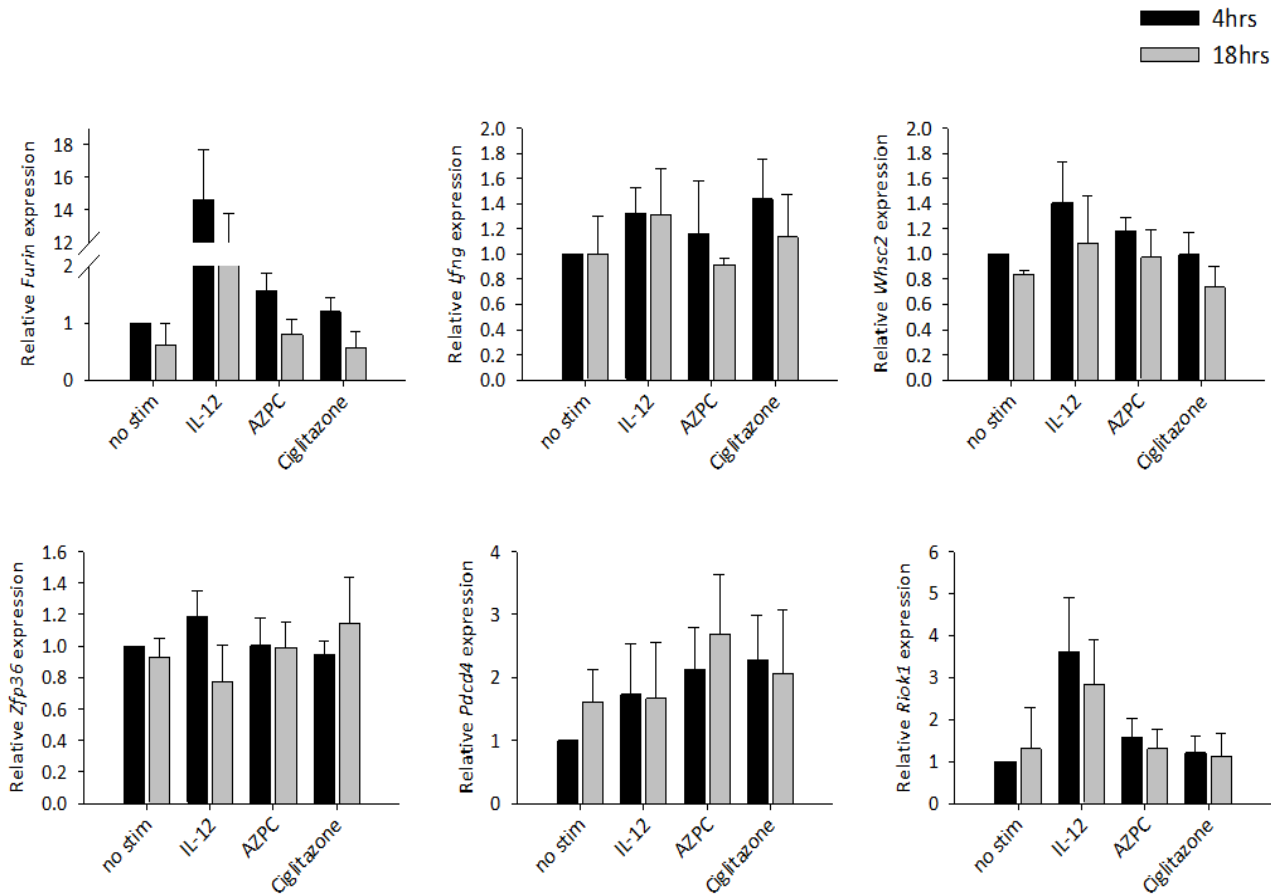


Figure 21: Graphs describing the relative expression of conserved genes that were checked for stimulation when induced by IL-12 and the two PPAR $\gamma$  agonists.

## CHAPTER FIVE: DISCUSSION

STAT4 is one of the critical transcriptional regulators along with T-bet (master regulator) in the adaptive immune response [13]. Although it is well characterized as a signaling molecule in Th1 cell development, its transcriptional regulatory role is poorly understood. The cytokine IL-12 stimulates Th1 cell differentiation, which is mediated by the STAT4 molecule at the cell surface (Figure 4) [50-51]. This signal is transformed into the nucleus for the transcriptional activity of a number of genes controlled by STAT4 binding to their regulatory regions. In this study, we have utilized STAT4-specific ChIP-on-chip data to identify the STAT4 transcriptional targets active in Th1 cell development by focusing on sets of CRMs located on these genes, and this CRM discovery may identify the nodes and edges in a TRN central to this biological process.

### **Motif prediction analysis**

In the previous work by Good et al., the STAT4-dependent genes were categorized into three groups (based on their expression patterns) as no induction, transient, and sustained. The MEME analysis was performed on the sustained and transient categories that identified a number of potential *de novo* motifs as 3 and 4 (Table1.a and Table1.b), corresponding to the two categories respectively. These motifs were shortlisted based on the low E-values and high information content. The STAMP annotation was performed in order to annotate these potential motifs (Figure12). For example, the motif 6 corresponding to the transient category was annotated as PPAR $\gamma$ -RXRA, which plays an essential role in modulating the JAK/STAT pathway that in turn modulates the differentiation of Th1 cells [29].

Good et al. in his work identified the NF- $\kappa$ B and PPAR $\gamma$ -RXR motifs as enriched in their interval

sequences from the STAT4-specific ChIP-on-chip data [23]. The location analysis results identified and characterized the NF- $\kappa$ B or PPAR $\gamma$ /RX interval sequences and their corresponding genes that were present within a separation distance of 500 bases from the STAT4 site (Table 2). The CRM is defined as two or more TF binding motifs that are present in a fairly small region of the genome (hundreds of nucleotides), and hence the reason for setting up a filter of 500 bases is to likely identify biologically meaningful CRMs. A number of genes, such as Mina, GIMAP1, and GIMAP5, which were implicated in immune cell development, were observed to belong to the filtered class of genes with PPAR $\gamma$ /RXR-Stat4 CRM (Table 2). For instance, Mina (a carrier of the NF- $\kappa$ B and PPAR $\gamma$ /RXR motifs) represses the IL-4 promoter and thereby depresses the development of T helper 2 cells [47], which is another lineage that can differentiate from the pre-Th cell in competition with the Th1 cell . Hence, it is possible that our CRM discovery in Mina may be crucial for Th1 cell development. Similarly, genes like GIMAP1 and GIMAP5 (that are closely related to each other in the GIMAP family) promote the Th1 cell lineage survival in the presence of IL-12 molecules [52]. Our analysis of these filtered genes shows immune development specific GO categories (Table 3.A and Table 3.B), providing more evidence for our CRM classifications.

### **Cis-regulatory module analysis**

As a second way to approach CRM discovery, we have used the MotifScanner [34] tool to scan for the presence of TFBs in all the STAT4 ChIP-on-chip interval sequences. We isolated an equally-sized (2000 bases) foreground and two equally-sized background sequences from each of the interval sequences corresponding to *in vivo* STAT4 binding sites (foreground) and immediately downstream and upstream sequences (backgrounds), and these sequences were subjected to motif scanning (Table 4). The rationale for this approach is to control false positives in the motif prediction. After scanning

the motifs NF- $\kappa$ B, PPAR $\gamma$ /RXR, Tbox, C-jun/Ap1, HLX, and Runx on all the STAT4-specific ChIP-on-chip interval sequences and the sustained and transient sets of genes with 2100 bases and 1100 bases, interval sequences were filtered for with pairs of motifs (CRMs) corresponding to the STAT4 and the other motifs. This determined enrichment for STAT4 CRMs in the foreground set when compared to the background sets (Figure 13). We have performed a more stringent p-value statistical validation of these enrichment results, confirming the observed enrichment of the STAT4 CRMs in the whole STAT4 interval sequence (Table 6) and in the two small gene sets (sustained and transient) pertaining to a region length of 2100 bases and 1100 bases (Table7), using Fisher's T-test followed by p-value calculations. We performed a similar statistical validation by generating a random data set corresponding to the number of sequences in our small and whole data set using RSAT. Out of all the 6 CRMs tested (excluding STAT4), the PPAR $\gamma$ /RXR-STAT4 pair showed the best enrichment in the foreground compared to the two backgrounds (Table 6), and the CRM pairs (like NF- $\kappa$ B-STAT4, C-jun-STAT4, HLX-STAT4, and Tbox-STAT4) were found to be enriched in the transient and sustained gene sets ( Table 7).

### **Validation of PPAR $\gamma$ /RXR-STAT4 with biological relevance**

Although our CRM discovery needs to be experimentally validated in the lab, our prediction results from the *de novo* motif analysis experiment and the MotifScanner experiments clearly suggest that the predicted PPAR $\gamma$ /RXR-STAT4 CRM are important players in the transcriptional regulation of Th1 cell development.

For the biological validation of this potential CRM discovered (PPAR $\gamma$ /RXR-STAT4 CRM), we performed a colocalization analysis using the STAT4-specific ChIP-seq and STAT4-specific ChIP-

seq histone methylation data. The first step was to map the STAT4 ChIP-seq tags to the STAT4 ChIP-on-chip interval sequences (Table 8). The number of ChIP-seq tags that were mapped to the ChIP-on-chip data was only 553; this is because the ChIP-on-chip covers only a promoter region of 7.5 kb upstream to 2.5 kb downstream, whereas ChIP-seq covers the whole mouse genome (Figure 15). The second step was to identify the common interval id between the identified 553 ChIP-seq and the interval ids of the PPAR $\gamma$ /RXR-STAT4 CRM from the motif scanner results, thus subset list had 12 interval ids that corresponded to 18 genes (retrieved from the pre-constructed database). In the colocalization analysis we mapped the STAT4 ChIP-seq methylation patterns to the gene regions corresponding to the foreground and two background regions obtained in a similar fashion as we obtained the foreground and two background regions in the CRM analysis (Table 9). Table 9 gives the H3K4me and H3k27me normalized tag count for each gene that shows that the H3k4me methylation patterns are high compared to the H3k27me methylation patterns. Since the H3k4me patterns are higher, these genes show a positive response in the TRN (transcription regulatory network) involving the PPAR $\gamma$ /RXR-STAT4 CRM because the correlation between the H3k4me methylation and transcription process is positive, whereas the correlation between the H3k27me methylation and transcription process is negative [53].

### **Conservation analysis**

The number of genes corresponding to the PPAR $\gamma$  motif (Motif 5), from MEME analysis performed by Good et al., was around 208 genes [23]. For conservation analysis the common genes, between the MEME PPAR $\gamma$  genes and the 18 genes identified in the colocalization analysis corresponding to the PPAR $\gamma$ /RXR-STAT4 CRM, were considered (11 genes) (Figure 16). These 11 genes were analyzed for the conservation of the PPAR $\gamma$ /RXR-STAT4 CRM over different species like humans, dogs,

rhesus monkeys, and mice. The conservation patterns were mainly identified in three regions: intergenic, intronic, and transposons and repeats. Out of the 11 genes only 5 genes were observed to have a high conservation of our PPAR $\gamma$ /RXR-STAT4 CRM (Table 10). The percentage of the occurrence of our PPAR $\gamma$ /RXR-STAT4 CRM was tabulated between the mouse and other conservations showing other species (Table 11). The result from conservation analysis of PPAR $\gamma$ /RXR-STAT4 CRM of the genes that showed biological significance in the colocalization analysis clearly depicts the involvement of the PPAR $\gamma$ /RXR-STAT4 CRM in Th1 cell development at the transcription level.

### **Limitations**

The analysis we performed includes certain limitations that needs to be taken into account. The first limitation applies to the data used in this project. The ChIP-on-chip data and ChIP-seq data were generated at different experimental conditions and also the ChIP-on-chip interval region was overlifted from mm8 mouse build to mm9 mouse build using overlift tool from UCSC . The second limitation was in the sequence retrieval step to obtain the foreground and the two background regions. The background regions might have encompassed other interval sequences that could have been the reason for some of the CRMs being enhanced in the backgrounds along with the foreground. The third limitation is that the information pertaining to the transcription factors (like PWMs) obtained from the transcription factor database may not have been annotated correctly, as per the recent update. Although the above limitations apply, the analysis performed was directed towards answering the research questions for the project.

## CHAPTER 6: CONCLUSION

This study employed a combination of bioinformatics and computational tools to understand the working of STAT4 in conjunction with the other potential motifs involved in Th1 cell development. The STAT4-specific ChIP-on-chip data generated in accordance with the TSS (transcription start site) was done so as to discover the CRMs located very closely to the TSS. These CRMs might play a pivotal role in Th1 cell development. The identification of several transcriptional regulatory motifs that may act in concert with the STAT4 binding site was the first step in determining the STAT4 CRM that may play an essential role in the Th1 cell development. Also, the genes corresponding to the NF- $\kappa$ B and PPAR $\gamma$ /RXR sites were filtered based on their location of less than 500 bases from STAT4 sites. A novel finding was the identification of the gene, Mina, which could play a significant role in Th1 cell lineage development that was one gene in the filtered subset of genes.

The study's outcome of the CRM discovery could be the initial step in decoding the transcriptional regulatory mechanisms in Th1 cell development, specifically the regulatory networks active here. The identification of PPAR $\gamma$ /RXR-STAT4 CRM as a potential CRM that may be significantly involved in Th1 cell development was provided with biological relevance using the colocalization and conservation analyses, thus giving a potential answer for our research question.

The methodology implemented could be generalized for computational identification of novel transcription regulators in Th1 cell development and other biological systems.

## REFERENCES

1. Mogensen, T.H., Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. *CLINICAL MICROBIOLOGY REVIEWS*, 2009.
2. Lewis L. Lanier, J.C.S., Do the terms innate and adaptive immunity create conceptual barriers? *Nature Reviews Immunology*, 2009. **9**: p. 302-303.
3. Smith, K.A., Medical immunology: a new journal for a new subspecialty. *Medical Immunology*, 2002(1): p. 1.
4. Satoru kitagawa, S.s., Tatsuo azuma, Jun shimizu, Toshiyuki hamaoka, and A.H. fujiwara, Heterogeneity of CD4+ T cells involved in anti-allo-class I H-2 immune response. *Journal of Immunology* 1991. **146**: p. 2513-2521.
5. Abdel Rahim A. Hamad, S.M.O.H., Michael S. Lebowitz, Ananth Srikrishnan, Joan Bieler, Jonathan Schneck, and Drew Pardoll, Potent T Cell Activation with Dimeric Peptide–Major Histocompatibility Complex Class II Ligand: The Role of CD4 Coreceptor. *J Exp Med.* , 1998. **188(9)**: p. 1633-1640.
6. Takashi Amagai, T.K., T Katsu-Iku Hirokawa, T Shin-Ichi Nishikawa, Jiro Imanishi, Yoshimoto Katsura, Dysfunction of irradiated thymus for the development of Helper t cell. *Journal of Immunology*, 1987. **139**: p. 358-364.
7. Dong, C., Helper T-cell heterogeneity: a complex developmental issue in the immune system. *Cellular & Molecular Immunology*, 2010(7): p. 163.
8. Asadullah, K., et al., The pathophysiological role of cytokines in psoriasis. *Prous Science*, 1999. **35(12)**.
9. G. Caramori, S.L., K. Ito, K. Tomita, T. Oates, E. Jazrawi, K.F. Chung, P.J. Barnes and I.M. Adcock, Expression of GATA family of transcription factors in T-cells, monocytes and bronchial biopsies. 2001. **18**: p. 466-473.
10. Thieu, V.T., et al., Signal Transducer and Activator of Transcription 4 Is Required for the Transcription Factor T-bet to Promote T Helper 1 Cell-Fate Determination. *Immunity*, 2008. **29(5)**: p. 679-690.
11. Tilo Biedermann, M.R., José M Carballido, TH1 and TH2 lymphocyte development and regulation of TH cell-mediated immune responses of the skin. *J Investig Dermatol Symp Proc.*, 2004. **9 ((1))**: p. 5-14.
12. Susanne J. Szabo, S.T.K., Gina L. Costa, Xiankui Zhang, C. Garrison Fathman, and Laurie H. Glimcher, A Novel Transcription Factor, T-bet, Directs Th1 Lineage Commitment. *cell*, 2000. **100**: p. 655-669.
13. Afkarian, M.e.a., T-bet is a STAT1-induced regulator of IL-12R expression in naive CD4+ T cells. . *Nat Immunol.* , 2002(3): p. 549-557.
14. Andrea L Wurster, T.T.a.M.J.G., The biology of Stat4 and Stat6. *Oncogene*, 2000. **19**: p. 2577-2584.
15. Rengarajan, J., S.J. Szabo, and L.H. Glimcher, Transcriptional regulation of Th1/Th2 polarization. *Immunol Today*, 2000. **21(10)**: p. 479-83.
16. Mayetri Gupta, J.S.L., De novo cis-regulatory module elicitation for eukaryotic genomes. *PNAS*, 2005. **102(20)**: p. 7079-7084.
17. Roded Sharan, A.B.-H., Gabriela G. Loots, Ivan Ovcharenko, CREME: Cis-Regulatory Module Explorer for the human genome. *Nucl Acids Res*, 2004. **32(suppl2)**: p. W253-W256.
18. Hoey T , G.M., STATs as mediators of cytokine-induced responses. *Adv Immunol.* , 1999. **71**:



- p. 145-62.
19. Grusby, M.H.K.a.M.J., Regulation of T helper cell differentiation by STAT molecules. *Journal of Leukocyte Biology*, 1998. **64**.
  20. Lund, R.J., Chen, Z., Scheinin, J. & Lahesmaa, R. , Early target genes of IL-12 and STAT4 signaling in th cells. *J Immunol.* , 2004. **172**: p. 6775-6782.
  21. Hoey, T.e.a., Distinct requirements for the naturally occurring splice forms Stat4 $\alpha$  and Stat4 $\beta$  in IL-12 responses. . *Embo J.* , 2003. **22**: p. 4237-4248.
  22. Lai Wei, G.V., Hong-Wei Sun, Wendy T. Watford, Hiroaki Takatori, Haydee L. Ramos, Hayato Takahashi, Jonathan Liang, Gustavo Gutierrez-Cruz, Chongzhi Zang, Weiqun Peng, John J. O'Shea, and Yuka Kanno, Discrete Roles of STAT4 and STAT6 Transcription Factors in Tuning Epigenetic Modifications and Transcription during T Helper Cell Differentiation. *Immunity*, 2010. **32**(6): p. 840-851.
  23. Good SR, T.V., Mathur AN, Yu Q, Stritesky GL, Yeh N, O'Malley JT, Perumal NB, Kaplan MH, **Temporal Induction Pattern of STAT4 Target Genes Defines Potential for Th1 Lineage-Specific Programming.** . *J Immunol* 2009,, 2009( 183): p. 3839-3847.
  24. O'Sullivan, A., Chang, H.C., Yu, Q. & Kaplan, M.H., STAT4 is required for interleukin-12-induced chromatin remodeling of the CD25 locus. . *J Biol Chem.*, 2004. **279**: p. 7339-7345.
  25. Letimier, F., Passini, N, Gasparian, S, Bianchi, E, Rogge, L, Chromatin remodeling by the SWI/SNF-like BAF complex and STAT4 activation synergistically induce IL-12R $\beta$ 2 expression during human Th1 cell differentiation. . *Embo J.* , 2007: p. 1292-1302
  26. Clark, R.B., The role of PPARs in inflammation and immunity. *Journal of Leukocyte Biology*, 2002. **71**: p. 388-400.
  27. Tontonoz, C.M.a.P., Linking metabolism to immunity through PPAR $\gamma$ . *Blood*, 2007(110): p. 3092-3093.
  28. Istvan Szatmari, D.T., Maura Agostini, Tibor Nagy, Mark Gurnell, Endre Barta, Krishna Chatterjee, and Laszlo Nagy, PPAR regulates the function of human dendritic cells primarily by altering lipid metabolism. *Blood*, 2007. **110**: p. 3271-3280.
  29. John J. Bright, S.K., Wanida Chearwae, and Sharmistha Chakraborty, PPAR Regulation of Inflammatory Signaling in CNS Diseases. *PPAR Research*, 2008. **volume 2008**.
  30. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B., JASPAR: an open-access database for eukaryotic transcription factor binding profiles. . *Nucleic Acids Res.*, 2004. **32**: p. D91-94
  31. E. Wingender, X.C., 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. *Nucl Acids Res*, 2000. **28**(2): p. 316-319.
  32. Xin He, X.L., Saurabh Sinha, Alignment and Prediction of cis-Regulatory Modules Based on a Probabilistic Model of Evolution. *PLoS computational Biology*, 2009. **5**(3): p. e1000299.
  33. Andra Ivan, M.S.H., Saurabh Sinha, Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs. *Genome Biology* 2008. **9**: p. R22.
  34. Aerts S, V.L.P., Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B, TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucl Acids Res*, 2005. **33**: p. 393-396.
  35. Skarstad, T.W.a.K., ChIP on Chip: surprising results are often artifacts. *BMC Genomics*, 2010. **11**: p. 414.

36. D. Karolchik, R.B., M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler and W. J. Kent, The UCSC Genome Browser Database. *Nucl Acids Res*, 2003. **31**: p. 51-54.
37. blat-Specification, u., <http://genome.ucsc.edu/goldenPath/help/blatSpec.html>.
38. Grobe, M., Get NCBI sequences for genes or specified regions. <http://discern.uits.iu.edu:8421/view-sequences.html>, 2007.
39. A.F.A. Smit, R.H.P.G.R., <http://repeatmasker.org>.
40. Bailey, T.L.E., C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers. . *Proc Int Conf Intell Syst Mol Biol.* , 1994. **2**: p. 28-36.
41. Mahony, S.a.B., P.V., STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, 2007. **35**: p. 253-258
42. Beissbarth, T.S., T.P. , GOstat: find statistically overrepresented Gene Ontologies within a group of genes. . *Bioinformatics.*, 2004. **20**: p. 1464-1465.
43. Zhibin Wang, C.Z., Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang & Keji Zhao, Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 2009. **40**: p. 897-903.
44. Ovcharenko I, N.M., Loots GG, Stubbs L., ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucl Acids Res*, 2004(32): p. 280-286.
45. Gabriela G.Loots , I.O., rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucl Acids Res*, 2004. **32**: p. W217-W221.
46. M S Hayden, A.P.W.a.S.G., NF- $\kappa$ B and the immune response. *Oncogene*, 2006(223): p. 6758-6780.
47. Okamoto, M., Van Stry, M., Chung, L., Koyanagi, M., Sun, X., Suzuki, Y., Ohara, O., Kitamura, H., Hijikata, A., Kubo, M. and Bix, M., Mina, an Il4 repressor, controls T helper type 2 bias. *Nat Immunol*, 2009. **10**( 8): p. 872-879.
48. H.Watts, T., TNF/TNFR family members in costimulation of T cell responses. *Annu. Rev. Immunol*, 2005. **23**: p. 23–68.
49. Thomas Chollier M, S.O., Turatsinze J V, Janky Rs, Defrance M, Vervisch E, Brohee S, van Helden J, RSAT: regulatory sequence analysis tools. *Nucl Acids Res*, 2008. **36**: p. 119-127.
50. Kaplan MH, S.Y., Hoey T, Grusby MJ., Impaired IL-12 responses and enhanced development of Th2 cells in Stat4-deficient mice. *Nature*, 1996. **382**: p. 174-7.
51. William E. Thierfelder, J.M.V.D., Koh Yamamoto, Ralph A. Tripp, Sally R. Sarawar, Richard T. Carson, Mark Y. Sangster, Dario A. A. Vignali, Peter C. Doherty, Gerard C. Grosveld & James N. Ihle, Requirement for Stat4 in interleukin-12-mediated responses of natural killer and T cells. *Nature*, 1996. **382**(171-174).
52. Amy Saunders, L.M.C.W., Michelle L. Janas, Amanda Hutchings, John Pascall, Christine Carter, Nicholas Pugh, Geoff Morgan, Martin Turner, and Geoffrey W. Butcher, Putative GTPase GIMAP1 is critical for the development of mature B and T lymphocytes. *Blood*, 2010. **115**: p. 3249-3257.
53. Yasuto Araki, Z.W., Chongzhi Zang, William H. Wood, Dustin Schones, Kairong Cui, Tae-Young Roh, Brad Lhotsky, Robert P. Wersto, Weiqun Peng, Kevin G. Becker, Keji Zhao , and Nan-ping Weng, Genome-wide Analysis of Histone Methylation Reveals Chromatin State-Based Regulation of Gene Transcription and Function of Memory CD8+ T Cells. *journal of Immunology*, 2009(6): p. 912-925.

## APPENDICES

### *Appendix A*

#### CODE FOR EXTRACTING SEQUENCE FROM MOUSE GENOME

Perl code to extract the sequences for the foreground and two background regions

```
#!/usr/bin/perl

use strict;

#use a hash to avoid retrieving the same IL* more than once

my %ilIdsents;

#we need to be warned if the same chromosome is being loaded more than once

my %chrNames;

open(intervalFile, 'IL12_intervals.txt');

my $lastChrName = "";

my $chrSequence = "";

while (<intervalFile>)

{

    #the regular expression is used to parse the line into

    #the four pieces (that we're loading into the four variables below)

    if (/^(\[^\t]+\)\t(\d+)\t(\d+)\t(.+)\$/)

    {

        # (four variables :-)

        my $chrName = $1;

        my $intStart = $2;
```

```

my $intEnd = $3;
my $iIdent = $4;
if ($iIdent{$iIdent})
{
    #skip this line from the interval file (we've already extracted the interval)
    next;
}
else
{
    $iIdent{$iIdent} = 1;
}
if ($chrName ne $lastChrName)
{
    print "Encountered new chromosome (".$chrName.")\n";
    #close the chr-specific output files for the previous chromosome (if needed)
    if ($lastChrName ne "")
    {
        close (outFileL);
        close (outFileC);
        close (outFileR);
    }
    #read in the sequence of the current chromosome
    print "Loading sequence for ".$chrName."...\n";

```

```

$chrSequence = "";
open(chrSeqFile,'mouse-'.$chrName.'.fa');
while(<chrSeqFile>)
{
  if (!/^>/)
  {
    chomp;
    $chrSequence .= $_;
  }
}
close (chrSeqFile);
print "Loaded sequence for " . $chrName . ".\n";
#open the output files for the left-flank, (padded?) center, and right-flank regions
open(leftOutFile,'>mouse-'.$chrName.'-left.fa');
open(centerOutFile,'>mouse-'.$chrName.'-center.fa');
open(rightOutFile,'>mouse-'.$chrName.'-right.fa');
#move "last" to current
$lastChrName = $chrName;
if ($chrNames{$chrName})
{
  print chr(7). "WARNING - Reloaded a chromosome!\n";
}
else
{

```

```

    $chrNames{$chrName} = 1; }}

#print "Determining interval start and end points for L/C/R...\n";

# note: [1,1] is of length 1, so length([i,j]) = j-i+1

my $intLength = $intEnd - $intStart + 1;

my $cIntStart = $intStart;

my $cIntEnd = $intEnd;

if ($intLength < 2000)
{
    my $intervalAdj = int((2000-$intLength)/2);

    $cIntStart -= $intervalAdj;

    $cIntEnd += $intervalAdj;

    $intLength = $cIntEnd - $cIntStart + 1;

    $cIntEnd += ($intLength % 2);

    $intLength = $cIntEnd - $cIntStart + 1;

    # $intLength should now be 2000
} #for a center [i, j], let left be [i-1999,i] and right be [j,j+1999]

my $lIntStart = $cIntStart - 1999;

my $lIntEnd = $cIntStart;

my $rIntStart = $cIntEnd;

my $rIntEnd = $cIntEnd + 1999;

#print "Determined interval start and end points for L/C/R.\n";

#determined the interval endpoints, so

# write to the files (fasta header line and then sequence data)

```

```

print leftOutFile '>'.$chrName.

    '['.$intStart.'!'.$intEnd.']-original '

    '['.$lIntStart.'!'.$lIntEnd.']-left '.$lIdent.

    "(length=2000)\n";

print leftOutFile substr($chrSequence,$lIntStart,2000)." \n";

print centerOutFile '>'.$chrName.

    '['.$intStart.'!'.$intEnd.']-original '

    '['.$cIntStart.'!'.$cIntEnd.']-center '.$lIdent.

    "(length=' $intLength. )" \n";

print centerOutFile substr ($chrSequence,$lIntStart,$intLength)." \n";

print rightOutFile '>'.$chrName.

    '['.$intStart.'!'.$intEnd.']-original '

    '['.$rIntStart.'!'.$rIntEnd.']-right '.$lIdent.

    "(length=2000)\n";

print rightOutFile substr ($chrSequence, $rIntStart, 2000)." \n";

#we're done with this line from the interval file

}}

if ($lastChrName ne "")

{

    close(outFileL);

    close(outFileC);

    close(outFileR);

}

```

```
close(intervalFile);
```

## *Appendix B*

### **CODE FOR EXTRACTING THE SEQUENCE FROM DISCERN OUTPUT**

Perl code to parse the sequences from the Discern tool

```
#!/usr/bin/perl

use strict;

use warnings;

my $output="output.txt";

#Opening the html filw

my $file = $ARGV[0];

open(FILE, $file) or die("Unable to open file");

open(OUTPUT, ">$output") or die("Unable to open file");

$i=0;

while(<FILE>)

{

my $line = $_;

chop($line);

#removing the unwanted text from the html file

if($line=~s/Get sequences for gene names, UIDs, or specified regions//g || $line=~s/\(This software is
currently in test status.\)//g ||

$line=~s/All requests processed successfully.//g)

{

$line =~ s/(<.*>)//g;

chomp($line);
```



```
}
```

```
#parsing out the gene name with the FASTA identifier at the beginning of the name
```

```
if($line=~s/.*\(.*)\s\(UID.*>$1/ || $line=~s/The\seSummary\srecord.*Gene\shumman\(.*)\s\(.*/$1/g  
or $line=~s/Gene\sname.*//g)
```

```
{
```

```
$i++;
```

```
print OUTPUT "\n";
```

```
print OUTPUT "$line\n";
```

```
}
```

```
else
```

```
{
```

```
$line =~ s/(\<.*\>)//g;
```

```
print OUTPUT "$line";
```

```
}
```

```
}
```

```
print OUTPUT "$i";
```

## *Appendix C*

### **CODE FOR MAPPING ChIP-on-chip WITH ChIP-Seq DATA**

Perl code to map ChIP-seq data with the ChIP-on-chip data

```
#!/usr/bin/perl

$output="result.txt";

open(OUTPUT, ">$output") or die("Unable to open file");

#we need to be warned if the same chromosome is being loaded more than once

# first open the ChIP-on-chip interval sequence file and take into an array

# the second step will be opening the ChIP-seq tag file and taking it into an array

# condition1: chromosome number ChIP-on-chip interval = chromosome number of ChIP-seq tags

# condition 2: the start position of ChIP tag should be less than or equal to end position

#condition 3: the stop position of the ChIP-seq tag should be greater than or

open(ChIP_chip, "IL12-2-Interval.txt");

while (<ChIP_chip>)

{

$line=$_;

chomp($line);

@interval=split(/\t/,$line);

open(ChIP_seq, "S4WTTh1.txt");

while (<ChIP_seq>)

{
```

```
$line1=$_;  
chomp($line1);  
@chipseq=split(/\t/,$line1);  
if($interval[0] eq $chipseq[0])  
{  
if($chipseq[1]<=$interval[2])  
{  
if($chipseq[2]>=$interval[1])  
{  
print OUTPUT "@chipseq\t@interval\n";  
}}}}}
```

## *Appendix D*

### **CODE FOR MAPPING METHYLATION PATTERNS FOR A SUBSET OF GENES**

Perl code to map the STAT4 ChIP-seq methylation patterns to the three regions (foreground and two backgrounds)

```
#!/usr/bin/perl

$output="output.txt";

#opening the methylation ChIP-seq tag file

my $file = $ARGV[0];

open (FILE, $file) or die("Unable to open file");

#opening the output file

open (OUTPUT, ">$output") or die("Unable to open file");

while(<FILE>)

{

my $line = $_;

chop ($line);

@interval=split (/t/,$line);

#opening the file containing the list of genes for which corresponding methylation patterns from the
#ChIP-seq methylation file should be extracted

open (FILE1, "gene.txt") or die("Unable to open file");

while(<FILE1>)

{
```

```

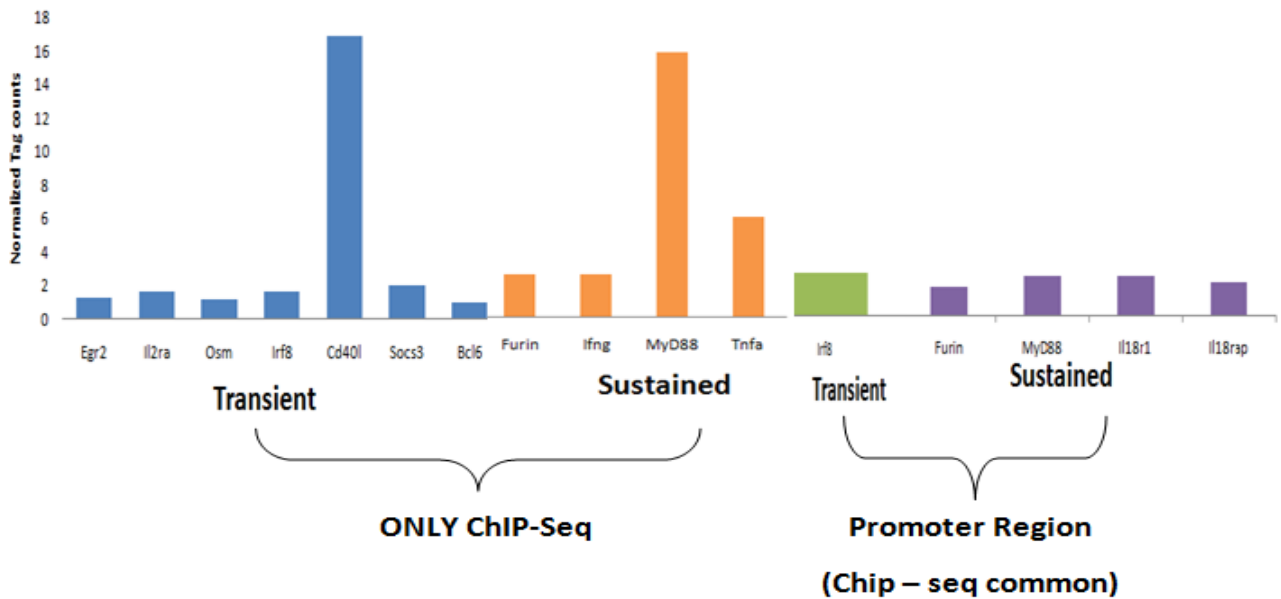
my $line1 = $_;
chop($line1);
@methyl=split (/t/,$line1");
# condition1: the start position of the gene should be less than or equal stop position of the ChIP-seq
tag corresponding to methylation patterns
#condition 2: the stop position of the gene should be greater than the start position of the ChIP-seq tag
corresponding to the methylation patterns
if($interval[1]<=$methyl[2])
{
if($interval[2]>=$methyl[1])
{
print OUTPUT
("$interval[0]\t$interval[1]\t$interval[2]\t$interval[3]\t$methyl[0]\t$methyl[1]\t$methyl[2]\n");
}}}}

```

*Appendix E*

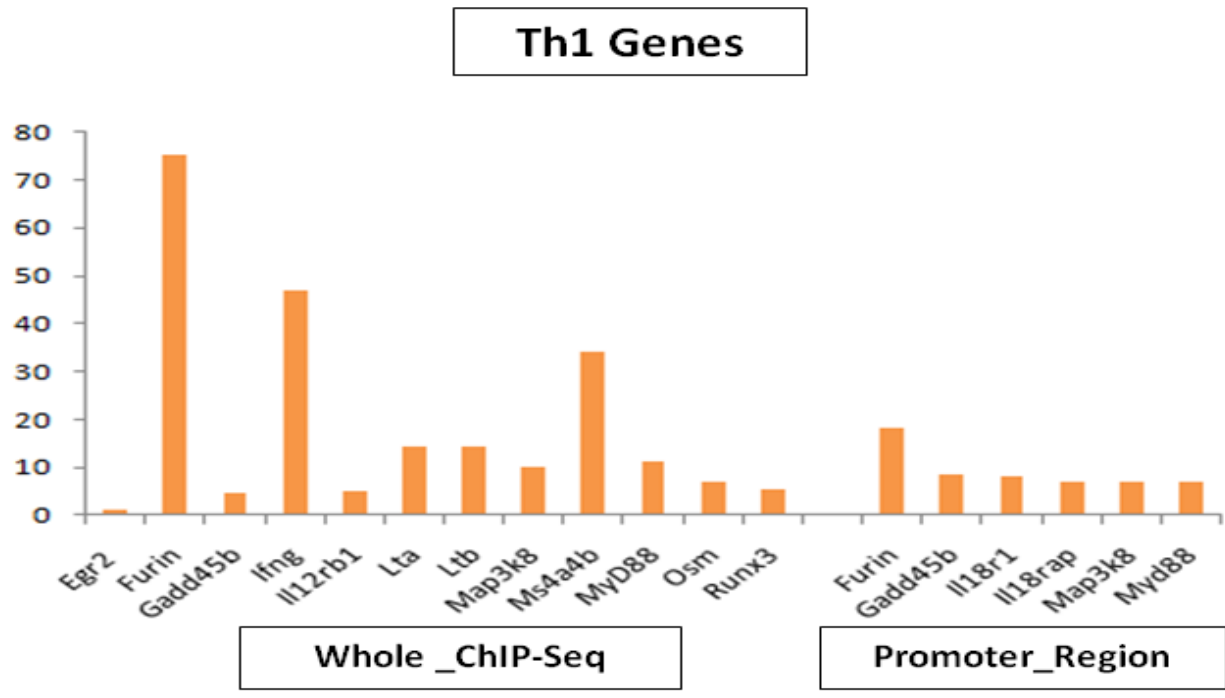
**DISTRIBUTION OF TH1 SPECIFIC GENES AND SUSTAINED AND TRANSIENT GENES  
IN THE PROMOTER REGION AND WHOLE MOUSE GENOME**

A few works were also performed in order to check the distribution of the transient, sustained, and also a few important Th1 genes from the Good et al. paper in the promoter region (ChIP-on-chip region) and in the whole genome (ChIP-seq) region.



A)

Figure A: Distribution chart for transient and sustained genes in the promoter region (common to ChIP-on-chip and ChIP-seq regions) and only ChIP-seq region apart from the promoter region



**B)**

Figure B: Distribution chart for Th1 genes in the promoter region (common to ChIP-on-chip and ChIP-seq regions) and only ChIP-seq region apart from the promoter region

## Appendix F

### CRM ANALYSIS FOR IL-23 INTERVAL SEQUENCE

We also performed the CRM analysis with the IL23 interval sequences utilizing the strategy that was applied on IL12 interval sequences and the statistical p-values, which are tabulated. From the p-value table below, we can infer that Pparg\_STAT4 CRM and Cjun\_STAT4 CRM are more enhanced in the foreground than in the two backgrounds, which can be worked on in detail to get some biological relevance of these CRMs in IL23-stimulated cell development.

Table a. P-values of TFBS enrichment in the foreground and background sequences of the IL-23 data set.

CRM	Background1	Foreground	Background2
<b>Cjun-STAT4</b>	0.463	0.066	0.216
<b>HLX_STAT4</b>	0.105	0.627	0.774
<b>NFK_STAT4</b>	0.105	0.928	0.916
<b>Runx_STAT4</b>	0.002	0.774	0.928
<b>Tbox_STAT4</b>	0.012	0.719	0.224
<b>Pparg_STAT4</b>	0.79	0.237	0.237



**Satishkumar Ranganathan Ganakammal**  
719 Indiana Avenue, suite #319, Indianapolis, IN 46202  
Phone: (317)-313-3526  
E-mail: [bioinformatics.satish@gmail.com](mailto:bioinformatics.satish@gmail.com)

---

- Proficiency in programming skills and analyzing experimental data
- Exposure to different aspects of information technology and biological systems
- Fast-learner, flexible and adaptable to different roles

### **Education**

Master of Science in Bioinformatics, **August 2010**  
Indiana University, School of Informatics, Indianapolis, IN  
**Dean's List, GPA-3.8**

Thesis: **Involved in** identifying Transcription regulator binding sites in STAT 4 mediated Th cells expression.

Bachelor of Technology in Biotechnology, **August 2008**  
Anna University, Institution-Madha Engineering College **India**  
Thesis: **Analyzed the PNA (Peptide Nucleic Acid) with various Bioinformatics tools.**

Diploma in Bioinformatics, **February 2007**  
**Sai Biosciences Research Institute, India**

### **Work Experience**

**Vanderbilt Medical Center –Biomedical Informatics, Nashville, TN** June 2010-present  
Bioinformatics System Engineer; worked with RNA-seq data, CNV data, and SNP array data and writing Perl scripts

**IU School of Informatics, Bioinformatics, IN** May 2009-Present  
Worked on High throughput and next generation sequencing data to identify Transcription Factor binding sites in STAT4 mediated Th1 cell development using PERL ,MySQL, PHP, Bioinformatics Software.  
**Contribution:** Published a paper in ACM proceeding ISB2010 and also refined the data for further analysis.

**Computational Biologist Intern, National Institute of Health, Maryland** May-July 2009  
Performed Co-evolutionary analysis for proteomics data of three interacting proteins Vwf, F8 and Adamts13  
**Contribution:** Performed all the preliminary analysis on the proteomics data.

**Teaching Assistant**, School of Informatics, IUPUI August- December2009  
Tutoring students in MySQL and PHP for the Introduction to Informatics class

**Support Center Consultant**, University Information Technology and Services September -December 2008.

### **Skill Set**

**Programming Languages:** Perl

**Databases and web design:** SQL, MySQL, Aqua data studio, PHP, HTML, MS Access

**Operating System:** UNIX, Linux, Windows, Mac

**Statistical Software:** SPSS, R (basics), SAS (basics).

**Tools:** Bowtie, TopHat, Cufflinks, Partek, GenePix array scanning software, GALAXY, Swisspdb, Rasmol, Hex 4.5, Bioedit, scientific data management tools such as elab note book (ELN), spot fire, Water Nugensis, Motif scanner, MEME suite.

**Machine Learning:** WEKA, MATLAB (basics)

**Data worked:** RNA-seq data, CNV array and SNP array data, ChIP-seq, ChIP-on-chip, Microarray data and Proteomics data

**Pipelines Familiar:** Microarray analysis, RNA-seq analysis, ChIP-on-chip and ChIP-seq analysis.

### **Biotechnological Techniques**

- Basic genetic engineering techniques :Isolation of genomic DNA from plant cells, Isolation of plasmid DNA from bacteria , isolation of RNA, Restriction analysis, Ligation reactions, Transformation of bacteria, Conjugation of bacteria, Isolation of Antibiotic Producers, Determination of Antibacterial spectrum. Electrophoresis and Southern blotting techniques

### **Project worked on**

#### **Graduate projects:**

- Constructed SVM (based on leave one out cross validation) using MATLAB and compared with the SVM constructed using WEKA Machine learning for Bioinformatics.
- Created frontend for “Regen DB”-Biological database management.
- Created a database and designed a PHP (front end) connectivity for Transcription Regulators data-Scientific Database Management course
- Created database for “Human Cancer Genes”- Introduction to Informatics course

#### **Undergraduate projects:**

- Bioinformatics analysis of PNA (Peptide Nucleic Acid) – A revolutionary mimic

### **Publication**

- Paper presented on “Cis Regulatory Module discovery in immune cell development” at the “International symposium of Biocomputing”, NIIT Calicut, India. (February 2010).
- Published an article in International Journal of Medical Engineering and Informatics Volume 1, Number 2 / 2008 on the topic “Peptide nucleic acid – a revolutionary mimic”.

### **Activities**

- **Second place in Research day Poster Presentation, IUPUI, 2008-2009**
- **Graduate Senator, ISG (Informatics Student Government), IUPUI**