

CIS REGULATORY MODULE DISCOVERY IN TH1 CELL DEVELOPMENT

Satishkumar Ranganathan Ganakammal

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree of
Master of Science in Bioinformatics,
Indiana University

December 2010

Accepted by the Faculty of Indiana University,
in partial fulfillment of the requirements for the degree of Master of Science
in Bioinformatics

**Master's Thesis
Committee**

Narayanan B. Perumal, Ph.D., Chair

Mark H. Kaplan, Ph.D.

Golnaz Vahedi, Ph.D.

© 2010

Satishkumar Ranganathan Ganakammal

ALL RIGHTS RESERVED

Dedicated to my parents

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	vii
LIST OF FIGURES.....	viii
LIST OF APPENDICES.....	ix
ACKNOWLEDGEMENTS.....	x
ABSTRACT.....	xi
CHAPTER ONE: INTRODUCTION.....	1
1.1 IMMUNE SYSTEM.....	1
1.2 CELLS INVOLVED IN IMMUNE RESPONSE.....	2
1.3 T-HELPER CELL DEVELOPMENT AND ACTIVATION.....	3
1.3.1 TRANSCRIPTION REGULATION IN TH1 CELL DEVELOPMENT... 5	5
1.4 ROLE OF CIS-REGULATORY MODULE IN TRANSCRIPTIONAL REGULATORY NETWORKS.....	6
CHAPTER TWO: BACKGROUND.....	8
2.1 STAT4 BIOLOGY.....	8
2.2 CURRENT UNDERSTANDING OF STAT4.....	11
2.3 STAT4 TARGET GENES IN TH1 DEVELOPMEN.....	12
2.4 ROLE OF PPAR γ -RXR IN T CELL DEVELOPMENT.....	13
2.5 COMPUTATIONAL METHODS INVOLVED IN CRM DISCOVERY.....	14
2.6 KNOWLEDGE GAP.....	15
2.7 RESEARCH QUESTION.....	16
CHAPTER THREE: MATERIALS AND METHODS.....	17
3.1 ChIP-on-chip AND MOUSE GENOME DATA.....	17
3.2 DATABASE FOR ChIP-on-chip DATA.....	19
3.3 SEQUENCE EXTRACTION.....	20
3.4 REPEAT MASKER.....	22
3.5 <i>de novo</i> MOTIF DISCOVERY.....	22
3.6 DNA MOTIF COMPARISON.....	23
3.7 GENE ONTOLOGY (GO) ANALYSIS.....	24
3.8 CRM PREDICTION.....	25

3.9 MAPPING ChIP-on-chip AND ChIP-Seq DATA.....	26
3.10 COLOCALIZATION OF METHYLATION PATTERNS.....	26
3.11 CONSERVATION OF POTENTIAL CRM SET.....	28
CHAPTER FOUR: RESULTS.....	30
4.1 <i>de novo</i> MOTIF ANALYSIS.....	30
4.2 LOCATION ANALYSIS OF STAT4 CRMS WITH NF-KB OR PPAR γ /RXR SITES..	32
4.3 GO ANALYSIS.....	34
4.4 CRM ANALYSIS.....	35
4.5 STATISTICAL VALIDATION OF PREDICTED POTENTIAL CRM SETS.....	38
4.6 MAPPING ChIP-on-chip AND ChIP-Seq DATA.....	40
4.7 COLOCALIZATION OF METHYLATION PATTERNS.....	42
4.8 CONSERVATION ANALYSIS.....	44
4.9 WET LAB VALIDATION.....	49
CHAPTER FIVE: DISCUSSION.....	50
CHAPTER SIX: CONCLUSIONS.....	55
REFERENCES.....	56
APPENDICES.....	59

LIST OF TABLES

Table 1.a Potential <i>de novo</i> motifs for sustained gene set.....	31
Table 1.b Potential <i>de novo</i> motifs for transient gene set.....	31
Table 2 Location distance analysis.....	33
Table 3.A GOSTat result for biological process.....	34
Table 3.B GOSTat result for molecular function.....	35
Table 4 MotifScanner output.....	36
Table 5 Model of Fishers T-test calculation.....	39
Table 6 P-value validation of CRM enrichment in whole dataset.....	39
Table 7 P-value validation of CRM enrichment in small dataset.....	40
Table 8 STAT4 ChIP-on-chip and ChIP-Seq mapping.....	41
Table 9 Colocalization of methylation patterns for the subset of genes.....	43
Table 10 Percentage identity of the genomic region.....	46
Table 11 Number of PPAR γ -STAT4 TFBS in conserved region.....	47

LIST OF FIGURES

Figure 1 Cells involved in immune response.....	3
Figure 2 T-helper cell lineage	4
Figure 3 Transcription regulation in Th1 cell	5
Figure 4 Representation of the cis-regulatory module or Cistrome unit	7
Figure 5 Biology of STAT4 involvement in Th1 cell development.....	9
Figure 6 JAK-STAT pathway representation.....	10
Figure 7 Wet lab ChIP-on-chip work flow.....	17
Figure 8 ER diagram for the ChIP-on-chip database.....	20
Figure 9 Sequence extraction concept.....	21
Figure 10 WebLogo representation from MEME analysis.....	23
Figure 11 Workflow of motif prediction analysis.....	24
Figure 12 Workflow of CRM prediction analysis.....	25
Figure 13 Workflow of colocalization of methylation patterns.....	27
Figure 14 Workflow of conservation analysis.....	29
Figure 15 Peak intensity plot of genes based on temporal induction patterns.....	24
Figure 16 STAMP annotation of MEME motifs.....	32
Figure 17 Venn diagram representation of PPAR- γ -Stat4 CRM in all three regions.....	37
Figure 18 Venn diagram representation of PPAR- γ -Stat4 CRM in the foreground.....	38
Figure 19 Pie chart representing the distribution of ChIP-Seq data.....	41
Figure 20 Venn diagram representation between MEME genes & Toucan genes.....	45
Figure 21 Plots describing the relative expression of conserved genes.....	49

LIST OF APPENDICES

Appendix A Code for extracting the sequence from mouse genome.....	59
Appendix B Code for extracting sequence from Discern output.....	64
Appendix C Code for mapping ChIP-on-chip with ChIP-Seq data.....	66
Appendix D Code for mapping ChIP-Seq methylations patterns for a subset of genes.....	68
Appendix E Distribution of genes in promoter and whole genome region.....	70
Appendix F CRM analysis for IL-23 interval sequence.....	72

ACKNOWLEDGEMENTS

I would like to take the opportunity to acknowledge some of the people who made my graduate study a memorable experience and made this thesis possible. Foremost, it is my sincere pleasure to express my deep and sincere gratitude to my advisor, Dr. Narayanan B. Perumal, for his guidance, motivation, feedback, encouragement, support, and patience during the course of my thesis. His input and efforts have been of great value for me.

I would like to thank the other members of my thesis committee, Dr. Mark H. Kaplan and Dr. Golnaz Vahedi for their time, encouragement, insightful comments, critical feedback and hard questions. I must appreciate their efforts to review my work. I also render my sincere thanks to member of Dr.Kaplan's lab for helping me with their wet lab cross validation and Dr. John O'Shea and his laboratory at NIH for providing me with his ChIP-Seq data.

I owe my sincere thanks to Indiana University for providing the financial support throughout my Master's program. Without the adequate academic preparation, my studies could not have been a successful experience. Hence, I would like to add my thanks to Dr. Mathew J. Palakal, Dr. Meeta Pradhan, Dr. Karl F. MacDorman, Dr.Malika Mahoui and other faculty and staff of the Department of Bioinformatics for their support in the course work.

I owe my loving thanks to my parents and sister for their encouragement and understanding. I would also like to thank Seth R. Good, Rahul Krishna Kollipara and James Scherschel for help in various stages of my work. My loving thanks to Rini Pauly for her help in my thesis writing and presentation. I would also like to thank my friends, Abhinita, Deepali, Anusha and the whole Indian student community at IUPUI for their support and all the fun we have had in the last three years.

ABSTRACT

Immune response enables the body to resist foreign invasions. The Inflammatory response is an important aspect in the immune response which is articulated by elements such as cytokines, APC, T-cell and B-cell, effector cell or natural killer. Of these elements, T-cells especially T-helper cells; a sub class of T-cells plays a pivotal role in stimulating the immune response by participating in various biological reactions such as, the transcription regulatory network. Transcriptional regulatory mechanisms are mediated by a set of transcription factors (TFs), that bind to a specific region (motifs or transcription factor binding sites, TFBS), on the target gene(s) controlling the expression of genes that are involved in T-helper cell mediated immune response. Eukaryotic regulatory motifs, referred to as *cis* regulatory modules (CRMs) or cistrome, co-occur with the regulated gene's transcription start site (TSS) thus, providing all the essential components for building the transcriptional regulatory networks that depends on the relevant TF-TFBS interactions. Here, we study IL-12 stimulated transcriptional regulators in STAT4 mediated T helper 1 (Th1) cell development by focusing on the identification of TFBS and CRMs using a set of Stat4 ChIP-on-chip target genes. A region containing 2000 bases of *Mus musculus* sequences with the Stat4 binding site, derived from the ChIP-on-chip data, has been characterized for enrichment of other motifs and, thus CRMs. Our experiments identify some potential motifs, (such as NF- κ B and PPAR γ /RXR) being enriched in the Stat4 binding sequences compared to neighboring background sequences. Furthermore, these predicted CRMs were observed to be associated with biologically relevant target genes in the ChIP-on-chip data set by meaningful gene ontology annotations. These analyses will enable us to comprehend the complicated transcription regulatory network and at the same time categorically analyze the IL-12 stimulated Stat4 mediated Th1 cell differentiation.