Two perspectives of cell fate determination: Systems dynamics and molecular mechanisms

Albert C Huang

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2011

Ilya Shmulevich, Chair

Steven J. Collins

William S. Noble

Program Authorized to Offer Degree: Molecular and Cellular Biology

University of Washington

Abstract

Two perspectives of cell fate determination: Systems dynamics and molecular mechanisms

Albert C Huang

Chair of the Supervisory Committee:

Cell fate determination, the process of transforming a unicellular embryonic cell into lineage-restricted specific tissues of a multicellular organism, is fundamental to the process of development. Following the division of a single zygote, how can seemingly similar neighboring cells ultimately differentiate into disparate cell fates, and maintain their differences? An understanding of this complex and intricate process is important for stem cell-based therapies and treatment of cancer.

To provide the necessary conceptual background, we review two complementary perspectives that aid our understanding of the differentiation process — a systems dynamics level view and a molecular mechanism level view. We introduce the experimental systems, the diverse types of data collected for this work, and computational methods to integrate them to arrive at a coherent model.

From a systems dynamics perspective, cell fates can be thought of as attractors in the high dimensional gene regulatory networks. With this view, we designed experiments that identified a class of genes that are "divergent" in their expression during the differentiation process. These genes are associated with neutrophil differentiation and cell cycle progression. Further promoter-based transcription factor binding site analyses reveal enrichment of factors functionally linked to cancer progression and neutrophil differentiation, suggesting a systems dynamics view has the potential to elucidate mechanistic level details.

From a molecular mechanistic perspective, we investigate how T-bet, a Th1 lineage defining transcription factor, functionally regulates its target genes. We experimentally determine genome-wide binding locations of T-bet and target genes that it regulates. We find distinct genomic signal patterns for genes that are differentially regulated compared to those that are not. We find that the chromatin-modifying and N-terminal domains of the T-bet protein are necessary for the regulation of T-bet targets. Further, T-bet targets are also differentially regulated in other T helper subtypes.

Both perspectives offer unique insights into the differentiation process. Future direction would include methods to combine both perspectives by modeling mechanistic details while keeping track of the larger dynamic "emergent" attractor properties of the network.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Chapter 1

INTRODUCTION

Violently twitching and writhing its tail, the sperm burrows head first into an awaiting egg, sparking a union that forms the zygote. Gradually, the zygote divides into multiple cells, increases in size, repositions itself, and over time, gives rise to diverse specialized tissue such as limbs, heart, neurons and red blood cells. Stunned. I was breathless the first time I saw the classic *Nova* documentary *The Miracle of Life* in grade school. Little did I know back then, the question of cell fate determination, how seemingly similar neighboring cells ultimately gain their respective cell types, would form the basis of much of my graduate school studies.

Something special must have happened at that magical moment when the sperm and egg unite. For, on the one hand, a complex and intricate chain of developmental processes requiring extraordinary coordination could unfold with great precision, yet, on the other hand, they must be robust enough to be re-enacted countless times, and conserved across all animals. One is left to wonder if there are general principles that underlie this paradoxical phenomenon.

1.1 One genome, many cell fates. How does lineage determination occur?

In 1942, esteemed biologist Conrad Waddington presented the concept of canalization, where he noted the *discrete* nature of cell types: "Animals are built up of sharply defined different tissues and not of masses of material which shade off gradually into one another ... it is difficult to persuade it to differentiate into something intermediate..." [1]. He depicted a cell fate determination process analogous to that of a

ball rolling down an "epigenetic landscape" getting funneled into an existing valley representing a cell fate.

However, the crucial question remains: as all the cells inherit the same genetic material from the zygote, what are the critical differences between the skin cells, the muscle cells, or the neurons, that enable each to obtain their respective cell fate? How can seemingly similar neighboring cells ultimately differentiate into disparate cell fates, and maintain their differences? In other words, how does lineage restriction of cell fates occur?

This is the holy grail that developmental biologists have grappled with since the inception of the field. So fundamental to our understanding of development, the insights gained have the potential to profoundly impact the treatment of diseases. Indeed, one of the most successful treatments of acute promyelocytic leukemia (APL), a type of cancer that affects white blood cells, is a compound named all-trans retinoic acid (ATRA) [2]. ATRA induces white blood cells to differentiate terminally thus curbing uncontrolled cell proliferation. Similar approaches are also pursued for brain cancer and breast cancer treatments.

Likewise, an understanding of the cell fate determination process harbors great clinical relevance for stem-cell based therapies. These therapies capitalize on stem cells' ability to self-renew and differentiate. Transplantation of an individual's own stem cells to replace damaged tissues and generate functional cells offers hope for the treatment of diverse diseases such as type 1 diabetes [3], Parkinson's disease [4], spinal cord injuries [5,6], and muscle impairments [7,8]. Understanding how seemingly similar neighboring cells gain their respective cell fates therefore holds the key to more effective clinical treatments.

1.2 Organization of the thesis

During my graduate studies, I undertook the study of cell fate determination from two complementary perspectives: a systems dynamics view and a molecular mecha-

nistic view. The experiments were conducted in two lineage branching points of the hematopoietic systems, one in the neutrophil differentiation, and the other in the T helper cell differentiation.

Chapter one and two contain the necessary background for this thesis. In chapter one, I present the overarching question of cell fate determination and motivate its clinical relevance. Brief introductions to systems dynamics and molecular mechanisms are provided, as well as materials on neutrophil and T helper cell differentiation. In chapter two, I introduce the various experimental and computational types of data I generated or collected from genomic data repositories, and describe how to combine them. Chapter three describes how a systems dynamics perspective can be utilized to help design experiments to reveal mechanistic insights. Chapter four then takes a molecular mechanistic perspective and presents how a lineage determining transcription factor, T-bet, regulates its target genes. Finally, in chapter five, I conclude with the utilities of both perspectives and suggest future efforts to unite the two perspectives.

1.3 Understanding differentiation at the genome-wide level

Global genome-wide measurements are becoming ubiquitous. Advances in high throughput measurement technologies such as cDNA microarray, high throughput sequencing, mass spectrometry, and many others are now making global measurements routine. This new deluge of data pushes the focus of understanding biological processes from the single-gene level to the genome-wide level.

In this thesis, the larger problem of cell fate determination is approached from two separate perspectives. The background for these two perspectives is introduced in this section. Though genome-wide measurements are utilized for both perspectives, they are distinct in their focus. In the systems dynamics perspective, the emphasis is on the state of the networks and their 'emergent' patterns, taking networks as a whole into consideration. The molecular mechanism perspective, on the other hand, focuses

more on mapping out mechanistic details of individual molecular factors, attempting to answer questions about the molecular machineries that produce distinct cell fates by interrogating binding locations of transcription factors, interactions with other transcription factors or epigenetic factors, the status of chromatin, and accessibility, in essence — classical molecular biology pushed to the genome-wide scale.

1.3.1 A systems dynamics perspective

Much of the theoretical foundations of the *systems dynamics* perspective we undertake was proposed by Stuart Kauffman in the 1960's from his study of the behavior of Boolean networks [9].

To see the dynamic changes of the system over time, a notion of *state* is required. Let *S(t)* be the state of the network at time *t*. $S(t) = [x_1, x_2, x_3, ..., x_n]$ represents the expression level of a network of n genes, with x_i the expression level of gene i. Gene expression at a given time then is a point in the possible N-dimensional state space. The *state space* represents the collection of all possible states of the network. For illustration purposes, we will take a simple two-gene example ($n = 2$). The expression of these two genes can be represented as a vector where S represents a point in our 2-dimensional gene expression space.

$$S(i) = [x_1, x_2] \tag{1.1}$$

where x_1 and x_2 are the gene expression levels for genes 1 and 2. State vector S of value [10, 12] then means gene 1 has an expression level of 10 and gene 2 has an expression level of 12. If expression values over time t are available, the state of the network can be monitored and represented separately. For example, expressions of the network for t=1, 2 and 3 can be represented as a series of three vectors

$$S(1) = [10, 12], S(2) = [10, 14], S(3) = [5, 10] \tag{1.2}$$

Since genes are part of a regulatory network, they can influence the expression of each other, creating dynamic changes over time. The existence of these regulatory influences prohibit parts of the state space from being occupied. For instance, using a simplified binary discretization of *On* and *Off*, in our simple two-gene example, if gene 1 represses gene 2, and gene 2 represses gene 1, then expression levels of both genes 1 and 2 cannot both be *On*, thus making the parts of the state space that represent high expression levels for both genes unavailable. This phenomenon is general to networks of genes where n is much larger, since regulatory constraints must be satisfied.

Over time, the state of the gene regulatory network can 'settle' into a stable pattern of expression called an *attractor*. In our simple example, two attractor states exist, namely, [*On*,*Off*] and [*Off*, *On*]. Nearby unstable initial states that transition into the attractor constitute the *basin of attraction*. Paths traced out by the gene regulatory network leading toward the attractor form state *trajectories*. In more complex networks with up to a hundred thousand genes, Kauffman demonstrates that hundreds of stable states, or attractors can arise naturally [10]. Further, he hypothesizes that cell types (fates) are the attractors in high dimensional space, conforming to Waddington's notion of an epigenetic landscape. More recently, Sui Huang provided experimental evidence supporting this hypothesis [11, 12].

In summary, the systems dynamics perspective is useful in that it provides an explanation to the larger question of cell fate determination, that because of the topology of the underlying gene regulatory network, attractors or cell fate automatically arise as a natural consequence.

1.3.2 A molecular mechanism perspective

Whereas the emphasis of the systems dynamics perspective is on the global properties of the network, the emphasis of the molecular mechanism perspective is on the *details of the molecular machinery* that regulates tissue-specific gene expression.

From this perspective, tissue-specific gene expression is achieved through the ac-

tions of lineage-determining transcription factors. Much like switches in a circuit, transcription factors can turn 'On' and 'Off' the expression of genes. And as the name 'lineage-determining' implies, these factors control the expression of genes specific to a lineage by activating genes inherent to a lineage and repressing genes for alternative lineages. Frequently, this is accomplished by binding to the promoter regions of the target genes, activating transcription by stabilizing the general transcription machinery, or repressing transcription by blocking access to binding sites from other competing transcription factors.

However, the genome is normally tucked away in highly organized hierarchical structures with multiple layers of control to monitor access to the underlying sequences. In order to fold the two meter long DNA into a nucleus with a diameter of 10 micrometer, DNA is wrapped around histone octomers comprised of H2A, H2B, H3, and H4 to form basic units of nucelosome, which can themselves be further packed into more condensed structures. The state of the chromatin reveals much about the process of gene regulation [13]. The DNA-histone complex, or chromatin, exhibits states that are correlated with active or repressed states of transcription. For example, covalent modification of N-terminal histone protein H3K4 with tri- methyl group (H3K4me3) is associated with active transcription and modifications of H3K27 with tri-methyl group (H3K27me3) is correlated with repression [14]. Special enzymes are required to dynamically alter these histone marks: for the deposition of methyl groups, the histone methyltransferases [15, 16]; for removal, the demethylases [17].

For a lineage-determining transcription factor to regulate its target genes, it would be beneficial to have the capacity to affect the local chromatin environment of the targets. Interestingly, the above mentioned histone demethylase and methyltransferase do not recognize specific DNA sequences that localize them to target genes. How do they know where to go? Recent experimental data suggests that lineage-determining transcription factors can bring methyltransferases and demethylases to their target genes and affect the local chromatin environment. Indeed, T-bet, a

lineage-determining transcription factor for Th1 differentiation can physically interact with demethylase Jmjd3 to remove the repressive H3K27me3 mark and recruit methyltransferase, Set7/9 to put on the active H3K27me2 mark [18].

In addition to interactions with enzymes that affect chromatin states, lineage determining transcription factors can also interact with other transcription factors to collaborate and modify an existing gene expression pattern. For example, T-bet recruits Bcl6, the lineage determining transcription factor for follicular T cells, to repress expression of Th1 target genes, *Socs1*, *Socs3*, and *Tcf7* [19]. Further, T-bet can interact with the transcription factor Runx3 to activate the expression of Th1 cytokine *Ifng* and silence Th2 cytokine *Il4* [20, 21].

Recently, rapid improvement of high-throughput sequencing technology enabled techniques such as chromatin immunoprecipitation (ChIP) and DNAse I hypersensitivity to be applied on a genome-wide scale. Along with gene expression data such as cDNA microarrays and RNA-Seq, we now have, potentially, a compendium of data spanning from physical binding of interacting transcription factors, histone modifications, chromatin accessibility, to functional consequences of mRNA expression levels. Computational methods to decipher relationships between these diverse types of data has the potential to reveal important biological insights.

1.4 Hematopoiesis as a tractable model system

Hematopoiesis, the developmental process that generates diverse blood cells such as T cells, B cells, neutrophils, and macrophages, offers an experimental system for the study of cell fate determination [22]. Compared to cells lodged in solid tissues, hematopoietic cells are discrete entities that are isolated with relative ease from blood and tissues. Many hematopoietic cells are available as *in vitro* cell cultured systems, making experimentation on them more efficient.

In this section we provide the background on two hematopoiesis transition branches, namely, the neutrophil differentiation process and the T helper cell differentiation

process. Studies of neutrophil differentiation are conducted from a systems dynamics perspective, using the HL60 cell line, while studies of T helper cell differentiation are conducted with a molecular mechanism perspective from primary mouse $CD4^+$ T cells, or T-helper cells.

1.4.1 Neutrophil differentiation

Neutrophils play a fundamental role in the fight against microbial pathogens. As the most abundant member of the white blood cells, they circulate around the body sensing and tracking chemical signals originating from microbes, then engulf and kill the microbes with the bacterialcidal granules that they carry. They release inflammation signals that instruct other immune cells to converge at the site of infection. Neutrophils are an integral part of the innate immune pathogen recognition system [23].

The neutrophil differentiation process is accompanied by a series of hallmark morphological changes. Progenitors known as myeloblasts exhibit a characteristically enlarged nuclei. They develop azurophilic granules containing myeloperoxide elastase as they become promyelocytes. A decrease of the nucleus size with a concave shape and the appearance of additional lactoferrin granules signal the transition into myelocytes and metamyelocytes. As the nucleus further decreases in size and becomes multi-lobular, neutrophils mature into segmented or polymorphonuclear granulocyte with gelatinase granules [23].

Our study of neutrophil differentiation relies on the HL60 cell culture, a human promyelocytic leukemic cell line isolated by Collins et al. from a female patient with acute promyelocytic leukemia [24]. Addition of chemicals such as All Trans-Retinoic Acid (ATRA) induces HL60 cells to terminally differentiate into neutrophils from promyelocytes. In chapter three of this thesis, we present our results of utilizing the concept that cell types are attractors in high dimensional space to help design experiments that reveal mechanistic details of neutrophil differentiation.

1.4.2 Th1 cell differentiation

$CD4^+$ T cells, otherwise known as T helper cells, are integral to the adaptive immune response. They interact with various immune cells to *help* coordinate a coherent response to pathogens. T helper cells activate macrophages to make them more phagocytic, facilitate the production of antibodies with B cells, and enhance the cytotoxic capabilities of CD8 T cells. Depending on the cytokine environment and the stimulation signals they received, naive $CD4^+$ T cells can differentiate into diverse subsets including Th1, Th2, Th17, inducible T-regulatory (iTreg), and follicular T cells (T_{FH}) [25, 26].

Each subset of the T helpers cells is tasked with defined functional roles. Th1 cells are essential for the clearance of intracellular pathogens; Th2 are important for the removal of extracellular parasites; Th17 cells are necessary for the elimination of extracellular bacteria and fungi; regulatory T cells are indispensable to establish immune tolerance; follicular T cells are requisite for the activation of B cells.

To fulfill their diverse functional roles, T helper cells secret cytokines to signal and coordinate immune responses: Th1 cells secret interferon-γ (IFNγ); Th2 cells, IL-4, IL-5, IL-9, IL-13, and IL-25; Th17 cells, IL-17a, IL-17f, and IL-22; iTreg cells, IL-10, IL-35, and TGFβ; and Follicular T cells, IL-17 and IL-21. In chapter four, we focus on how T-bet, the lineage-determining transcription factor for Th1, regulates its target genes to affect the Th1 gene expression pattern [27].

Chapter 2

INTEGRATING EXPERIMENTATION AND COMPUTATION

In chapter one, our emphasis is on the question of cell fate determination and providing conceptual background materials from *systems dynamics* and *molecular mechanisms* perspectives. In chapter two, we shift our focus towards the experimental contexts, the computational data sets, and integration methods (how we put experiments and data together).

2.1 Types of data utilized in this thesis

In this section, we introduce a series of diverse data types utilized in this thesis. Within the description of each data type, we aim to address the following three questions: 1)What is the data? 2)How is it generated? 3)Why is it needed?

2.1.1 Flow Cytometry

Flow cytometry is a way to measure physical and chemical characteristics of single cells from a population. Suspended in liquid, cells are lined up single file through the flow cytometer, giving electronic detectors diverse readings including forward scattering, side scattering, and fluorescence. When antibodies are conjugated with fluorescent molecules, the flow cytometer detects microscopic events targeted by the antibodies. In the context of our studies, we use flow cytometry to detect the presence of CD11b, a well-established cell differentiation marker for neutrophils. Thus, by looking at the CD11b expression, we can determine the stage of differentiation for a population of cells.

2.1.2 DNA Microarray

DNA microarray technology enables genome-wide survey of mRNA gene expression. With sequenced genomes and gene-finding algorithms, the mRNA parts lists become available for gene-specific probe designs. Large arrays of probes representing genes are lined up on solid surfaces and hybridized to cDNA conjugated to fluorescent molecules from biological samples. Based on fluorescent intensities of the hybridized spots, quantitative expression of individual genes can be determined.

As a general tool for gene expression, DNA microarrays are frequently utilized downstream of experimental perturbations to measure their functional effects. In our work, we used them in three experimental contexts:

1) Time point gene expression: We collect samples of RNA for a population of cells over time. Comparisons for individual genes over time reveal rich dynamic responses to the cell fate determination process. In our context, we collect gene expression measurements over the course of neutrophil and Th1 differentiation. These data sets help us identify genes that are differentially expressed during these processes.

2) Cell type gene expression: Gene expression of subsets of T helper cells — Th1, Th2, Th17, iTreg, and nTreg — are available in the Gene Expression Omnibus (GEO) data repository. We collect expression of these cell types to identify groups of genes that are preferentially regulated for a given cell type.

3) Retroviral transduction: Retroviral transduction is a way to use retroviruses to introduce foreign DNA into cells, and study the effects of this over-expression. In our experiments, we use retroviruses to express wild type and mutant T-bet proteins in $CD4^{+}$ T-$bet^{-/-}$ CD4 to study the effects of T-bet proteins.

High-throughput sequencing

Following a Moore's Law-like trajectory of technological advancement, high-throughput sequencing technologies now produce gigabases of highly accurate sequences in a few days. With the help of reference genomes and computational methods, these sequences are stitched back and aggregated to generate genome-wide views. Not surprisingly, a number of traditional molecular biology techniques have been adapted to exploit this new capability. Classic assays such as chromatin immunoprecipitation (ChIP) and DNase I hypersensitivity are extended to high-throughput settings. These two data types are incorporated in our analyses.

2.1.3 ChIP-Seq

ChIP is an experimental technique that identifies interactions between protein and DNA. Cross-linked to proteins by formaldehyde, genomic DNA sequences are sonicated and pulled down by antibodies targeting specific proteins. Traditionally detected by user-designed primer sets with polymerase chain reactions (PCR), these sequences are now detected by high-throughput sequencing, allowing investigators to survey interactions between their protein of interest and the whole genome.

1) Histone methylation: As mentioned in the previous chapter, histone methylation marks such as H3K4me3 and H3K27me3 are associated with active and repressive transcriptional states, respectively. These methylation marks reveal the epigenetic status of the genome and provide additional clues to transcriptional regulation. In our study, we investigate the relationship between various histone methylation marks and gene expression to identify genomic signals correlated with functional regulation by the transcription factor T-bet.

2) Transcription factor: Transcription factors can bind to promoter region of genes, recruit general transcriptional machineries to commence transcription. Thus, a knowledge of the transcription factor binding locations offers important cues to the genes regulated by the transcription factor of interest. In our experiment, we identify the locations of T-bet bindings to find genes regulated by T-bet. Furthermore, we

examine close-by transcription factor binding sites which mark likely candidates of co-regulation.

2.1.4 DNAse I hypersensitivity

DNase I hypersensitivity is a technique to determine regions of the genome that are open and accessible. DNases are enzymes that digest unprotected, nucleosome-free DNA sequences — regions that are frequently associated with regulatory elements. When coupled with high-throughput sequencing, this assay has the potential to reveal functional regulatory elements on a genome-wide scale. In our analysis of Th1 differentiation process, DNase I hypersensitivity data for Th1 and Th2 are downloaded from the ENCODE project to investigate relationships between regions of accessible chromatin and transcriptional regulation.

2.1.5 PhastCons score

The PhastCons sequence conservation score is a measure that sheds light on regions of the genome that are functionally important [28]. Evolution is a powerful force to preserve what works and discard what does not. Presumably regions conserved across species serve notable functions. Available from the UCSC genome repository, phastCons reports the multiple sequence alignment scores resulting from aligning 29 other vertebrate species to mouse. In our study, we examine the relationship between regions of high sequence conservation scores and transcriptional regulation.

2.2 Integrating approaches in this thesis

In addition to the differences in emphasis mentioned in the previous chapter between the systems dynamics and molecular mechanism perspectives, we present, in the next two sections, two approaches of how we integrate experimentation and computation. The two approaches are contrasting — while we take a "theory drives experimental setup" stance for the systems dynamics perspective; a "data drives computational

model building" approach is taken for the molecular mechanism perspective.

2.2.1 Use of theory to inform experimental setup

In our study of neutrophil differentiation (systems dynamics perspective), we take for granted the concept of cell types as attractors in high dimensional space, knowing that experimental evidence also supports this view. With theory in mind, we designed experiments to exploit this knowledge.

Specifically, we designed experiments to place two populations of pre-differentiated neutrophil cells on separate but comparable trajectories toward neutrophil differentiation. With the addition of all-trans retinoic acid (ATRA), HL60 promyelocytic cells proceed toward the neutrophil attractor. ATRA, in terms of attractor theory, can be thought of as a perturbation that places the HL60 gene expression in the parts of the state space that flows naturally toward the neutrophil attractor. By varying dosage and duration of the ATRA treatment, we place treated HL60 cells in two distinct paths toward neutrophil differentiation. When ATRA-treatment is removed, one dosage/duration treatment combination continues toward neutrophil differentiation, while another treatment combination reverts back to the promyelocytic attractor. Differential gene expression analysis of these two trajectories reveals a class of genes that are involved with the process of cancer development and neutrophil differentiation, hinting at the utility of a systems dynamics perspective to yield molecular level details.

2.2.2 Use of statistical machine learning techniques to discover patterns in data

In our study of Th1 differentiation (molecular mechanism perspective) we focus on the details of the molecular machinery that governs differentiation to answer the question of how a lineage-determining transcription factor regulates its target genes. In support of that, we collect diverse data sets from ChIP-Seq and DNA microarray experiments.

Our goal is to construct a statistical model that reflects the relationships and dependencies between variables within the data sets, thus allowing us to identify important attributes that signal Th1 transcriptional regulation and provide possible mechanistic level insights. This goal fits well with the framework of "supervised learning" problems, where training examples with input attributes and class labels are provided to the learning algorithm to obtain a classifier. An examination of the classifier accentuates the structural patterns within the data sets. In this process, the model produced by the machine learning algorithm binds the computation and the experimentation together.

The learning algorithm we utilize for this problem is called random forest [29]. Invented by Breiman and Cutler, it is an ensemble method that makes predictions based on voting by a collection of decision trees by "majority" rule. The random forest algorithm is highly efficient and works well with a large number of variables. Its key advantages include an estimate of the "important" variables for classification, and an internal unbiased estimate of the generalization error through a bootstrapping process.

Chapter 3

CELL FATE AS AN ATTRACTOR IN HIGH DIMENSIONAL SPACE

3.1 Abstract

The process of cellular differentiation is governed by complex dynamical biomolecular networks consisting of a multitude of genes and their products acting in concert to determine a particular cell fate. Thus, a systems level view is necessary for understanding how a cell coordinates this process and for developing effective therapeutic strategies to treat diseases, such as cancer, in which differentiation plays a significant role. Theoretical considerations and recent experimental evidence support the view that cell fates are high dimensional attractor states of the underlying molecular networks. The temporal behavior of the network states progressing toward different cell fate attractors has the potential to elucidate the underlying molecular mechanisms governing differentiation.

Using the HL60 multipotent promyelocytic leukemia cell line, we performed experiments that ultimately led to two different cell fate attractors by two treatments of varying dosage and duration of the differentiation agent all-trans-retinoic acid (ATRA). The dosage and duration combinations of the two treatments were chosen by means of flow cytometric measurements of CD11b, a well-known early differentiation marker, such that they generated two intermediate populations that were poised at the apparently same stage of differentiation. However, the population of one treatment proceeded toward the terminally differentiated neutrophil attractor while that of the other treatment reverted back toward the undifferentiated promyelocytic attractor. We monitored the gene expression changes in the two populations after their respective treatments over a period of five days and identified a set of genes

that diverged in their expression, a subset of which promotes neutrophil differentiation while the other represses cell cycle progression. By employing promoter based transcription factor binding site analysis, we found enrichment in the set of divergent genes, of transcription factors functionally linked to tumor progression, cell cycle, and development.

Since many of the transcription factors identified by this approach are also known to be implicated in hematopoietic differentiation and leukemia, this study points to the utility of incorporating a dynamical systems level view into a computational analysis framework for elucidating transcriptional mechanisms regulating differentiation.

3.2 Introduction

The process of cellular differentiation is central to our understanding of the nature of multicellular living systems, their stability in a changing environment, and how such systems fail in diseases, such as cancer [30, 31]. This developmental process of individual cells in a multicellular organism committing to their specialized phenotypic characteristics is temporally coordinated by a complex dynamical system comprised of large numbers of interacting genes and their products [32–34]. Not surprisingly, dynamical systems theory has been used to study cell differentiation [35–37].

Despite its tremendous importance, there is very little accumulated knowledge of the process of differentiation from a systems perspective and of the role of molecular programs involved in this process. Even for an agent that causes differentiation to a common recognizable state, we do not know whether the cells, as manifestations of the underlying dynamic bio-molecular systems, always follow common or different molecular paths (or system state trajectories). In the latter case, we also do not know which of those paths is the most stable and least reversible.

Since a cell's phenotype and behavior are largely determined by the activities of the genes and proteins constituting a genetic network, it follows that the rules of interactions between these elements translate directly into rules of cellular behavior.

That is, the enormous state space of a genetic network (i.e., the space of all possible configurations activities of the constituent elements) becomes reduced into a relatively small number of trajectories and steady states (attractors) of the dynamical system. Kauffman postulated that these attractor states in model networks correspond to the cell types in multicellular organisms [9, 10], and the process of differentiation corresponds to a trajectory (in the state space) leading into one of the attractors. The cellular fate is thus determined by the attractor in which the genetic network eventually ends up; this can, to a large extent, be controlled by appropriate external stimuli that place the system into different initial states. It is important to note that many trajectories ensuing from such different initial states can flow to a common attractor and thus constitute its basin of attraction.

Consider that small molecule chemicals, such as dimethyl sulfoxide (DMSO) and a host of others can induce cell differentiation in a variety of cell systems along with concomitant cellular properties [38–42]. This rather amazing fact implies the pre-existence of cellular fates that need only be selected by means of external stimuli rather than created by specific molecular events. This 'selection' of cell fates occurs by means of the inherent nature of the dynamical system to flow to an attractor when placed in some initial transient state and thus, differentiation is a process of selecting a particular attractor in a genetic regulatory network. This view has been supported experimentally by Huang *et al.* using genome-wide mRNA expression profiling [11] as well as by means of analyzing cell fates in response to generalized physical stimuli, such as cell distortion [43]. For a more extended discussion on this topic, see [10].

The homeostatic stability of a differentiated cell is a consequence of the underlying stability of the attractor – 'nearby' states, which may occur as a consequence of natural environmental variation, simply flow back to the attractor. It is known that normal cells have a balanced state of proliferation and differentiation, resulting in homeostatic stability [44, 45]. A block of normal differentiation and abnormal reversal of differentiation (sometimes called de-differentiation) [46] are believed to be

some of the hallmarks of cancer [47]. Accordingly, therapeutic strategies have been designed to facilitate cancer cells to reenter the differentiation program, often termed differentiation therapy [48, 49].

The success of such therapeutic strategies depends on our ability to systematically determine appropriate molecular 'lever points', the perturbations of which place the biomolecular system into states that are poised to differentiate. Indeed, such a strategy corresponds to placing the system in a state by means of a stimulus, such as a therapeutic agent, and allowing the system to naturally flow toward an attractor that corresponds to the desired cellular endpoint [50–52]. To identify such targets for intervention, it is necessary to characterize the underlying molecular mechanisms, such as transcriptional regulatory networks, governing the process of differentiation. Systems biology approaches, which are predicated on global measurements and data integration, are now beginning to reveal transcriptional machinery underlying complex biological processes [53–56]. The rationale behind our study was to explore whether the aforementioned systems-level view of cell fates as attractors and differentiation as a route toward an attractor, when coupled with computational systems biology approaches, is informative for elucidating the transcriptional regulatory mechanisms governing differentiation.

To this end, we have selected a well-established differentiation model, human promyelocytic leukemia cells (HL60) originally isolated by Dr. Steven Collins from a 37-year-old female acute promyelocytic leukemia (APL) patient [24]. The HL60 is a multi-potent cell line that can be stimulated to differentiate using a variety of chemical agents, including DMSO [57], all-*trans*-retinoic acid (ATRA) [58], 1,25α-dihydroxyvitamin D_3 [59], 12-O-tetradecanoylphorbol 13-acetate (TPA) [60], and granulocyte macrophage colony-stimulating factor (GM-CSF) [61]. With the addition of ATRA, the HL60 cells differentiate into neutrophils, while displaying the early differentiation marker, CD11b, which begins to be expressed within one day of treatment [62]. Although there are others, CD11b is an early differentiation marker, which

allows one to capture the initial stage of the process. The CD11b+ differentiated HL60 cells can be stained with fluorescent-labeled anti-CD11b antibody and easily recognized by commonly used flow-cytometry methods and isolated by flow-sorting for further culturing and experimentation, as we have done here. The HL60 system was also used by Huang *et al.* [11] to demonstrate the correspondence between cell fates and high-dimensional attractor states of the underlying genetic network.

One could reason as follows. If we could place the HL60 into a state from which the system would dynamically flow towards the "neutrophil" attractor, as demonstrated by Huang *et al.*, then the genes that show altered behavior along the time-course trajectory relative to unstimulated controls could be hypothesized to be implicated in the neutrophil differentiation process. This, of course, may be the case, though the interpretation is confounded by the possibility that the genes exhibiting altered behavior are responsive to the particular mechanisms activated by the stimulus used, such as ATRA. Indeed, Huang *et al.* also confronted this conceptual difficulty when they compared trajectories from ATRA-treated and DMSO-treated HL60 cells, finding that certain genes may behave differently simply as a result of different stimuli activating different biological pathways, while many other genes dynamically converge towards a common attractor, despite the system flowing from distinct starting states corresponding to ATRA and DMSO treatments [11]. To identify genes that are not stimulus dependent, but are involved in the process of neutrophil differentiation, one could then use only one treatment, but in a way that allows one to alter cellular fate, namely, terminal differentiation into neutrophils or reversion back to the undifferentiated state.

The HL60 was shown to exhibit such behavior in two separate studies both demonstrating that this differentiation process contains at least two steps in which a precommitment stage precedes the decision to differentiate. Yen *et al.* observed that with continuous exposure of ATRA at a high concentration, the HL60 proceeds through differentiation, but upon removal of the stimulus, the HL60 falls back to the undif-

ferentiated state [63]. By analogy, such a precommitment stage corresponds to a gradually sloping plateau between a valley and a mountain such that a ball sitting on this plateau would roll down into the "undifferentiated" valley in the absence of additional energy necessary to make it over the "terminally differentiated" mountain. More recently, Chang *et al.* reported a population of "primed," undifferentiated CD11b- cells upon exposure to a low dose DMSO [64]. Though these cells are negative with respect to the CD11b marker, thus considered to be "undifferentiated," upon encountering a second dose of DMSO stimulation, they exhibited an increased rate of differentiation, suggesting that the first low dose DMSO had placed them in a "primed" intermediate differentiated state.

We thus decided to determine two different treatments, both with ATRA, but with different concentrations and incubation times such that the two cell populations corresponding to these treatments would be poised at the same stage of differentiation (precommitment), but so that one population follows through to the terminally differentiated neutrophil attractor, while the other reverts back by dynamically flowing towards the undifferentiated state. The genes that would exhibit different behavior between these two trajectories would then be potentially implicated in the differentiation process.

To identify two such precommitment states, we used the percentage of CD11b+ cells at the end of a particular treatment as a measure of the stage of differentiation. We performed 80 ATRA treatments consisting of 8 doses (0.0005μM to 1 μM) and 10 durations (4 to 13 days) in triplicate and measured percentages of CD11b+ cells, relative to an isotype antibody control, using FACS analysis. Consider loci in the two-dimensional dose $\times$ duration stimulus space, where all points on a particular locus correspond to a constant fraction of CD11b+ cells. Thus, two cell populations on the same locus can be said to be at the same "stage" of differentiation at least as it pertains to CD11b. We chose two such populations, one with a higher dose and a shorter duration and the other with a lower dose and a longer duration, such that

the cells treated with the higher dose proceeded with differentiation into neutrophils while the cells treated with the lower dose reverted back to the undifferentiated state, despite both populations exhibiting the same percentage of CD11b+ cells at the end of their respective treatments. The cells were live-sorted, cultured in fresh media, and profiled every 24 hours with microarrays for five days in triplicate. This additional Fluorescence Activated Cell Sorting (FACS) step mitigates the confounding effects of cellular heterogeneity due to subpopulations that do not initiate the differentiation program (i.e. CD11b- cells). In this manner, the gene expression programs of the two cell populations, one differentiating and one reverting, could be analyzed using computational approaches.

We defined a criterion to identify genes whose behavior over time exhibits a divergence between the two treatments. It is these genes that are hypothesized to be involved in the differentiation process. We analyzed the promoters of these genes and found that they are overrepresented with known transcription factors functionally linked to myeloid differentiation, cell cycle, and development. This study points to the utility of incorporating a systems-level view of global dynamics, as distinguished from the dynamics or kinetic behavior of the individual elements of a system, into a computational analysis framework that can be used for studying transcriptional regulatory mechanisms governing a complex biological process such as differentiation.

3.3 Results

3.3.1 Two comparable dosage/duration treatment combinations lead to different macroscopic cell fate attractors

Our first goal was to determine two dosage/duration stimulation conditions that yield comparable stages of differentiated HL60 cells, with one condition ultimately leading to neutrophil differentiation and the other reverting back to the undifferentiated state. In other words, we sought to identify two perturbations that place the ATRA-treated cells in two different basins of attraction or initial states leading to two different

attractors — the promyelocyte attractor and the neutrophil attractor. This information allows us to culture large quantities of the HL60 cells under these conditions and isolate mRNA for time course experiments to examine the set of genes differentially expressed between these treatments that could explain the differences in their eventual cell fates.

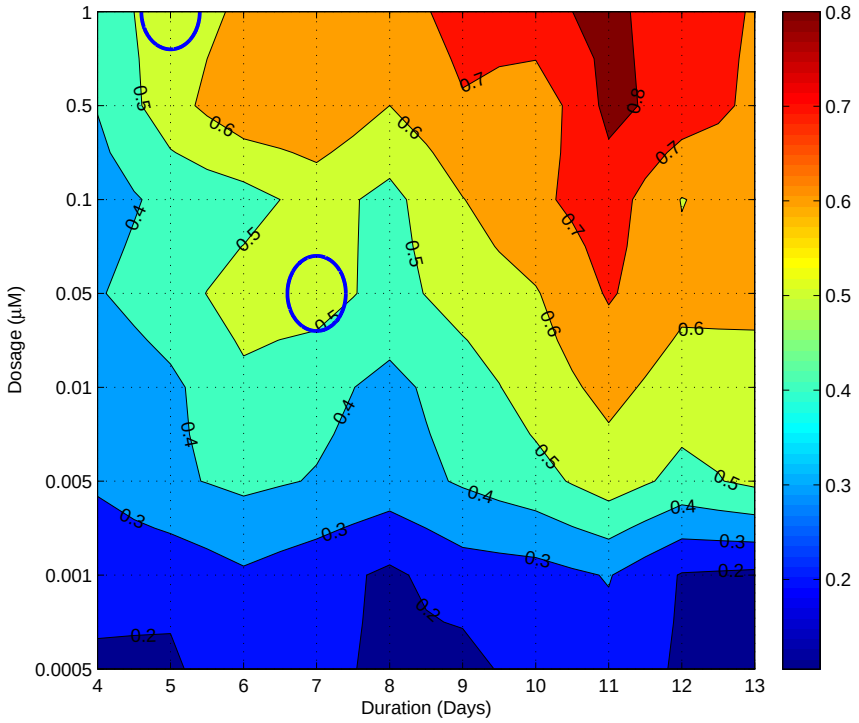

Figure 3.1: **Contour plot of percent CD11b+ cells after ATRA.** The x-axis represents the duration of ATRA treatment (Days). The y-axis represents the dosage of ATRA treatment (μM). An increase in dosage or duration leads to an increase in CD11b+ cells.

To achieve this goal, we set up a two-factor dosage and duration experiment with eight and ten levels respectively, ranging from 0.0005μM to 1 μM with 4 to 13 days of treatment. We used a well-established early marker for neutrophil differentiation, CD11b, as our surrogate for 'differentiated' and 'undifferentiated' state [62]. We measured the CD11b expression for each experimental condition and calculated the percentage of HL60 cells that are CD11b+ by comparing to the untreated samples.

Under this construction, the percentage of CD11b+ cells becomes a proxy for the developmental stage of neutrophil differentiation on a population level. The result of this dosage and duration experiment is summarized and displayed in a contour plot (Figure 3.1), showing a general trend — as the dosage or duration of ATRA treatment increases, the percentage of CD11b+ cells also increases. The treatment combinations, 0.5μM/11 Days produced the highest percentage of CD11b+ cells at 82.7%. As expected, this result conforms with our general intuition regarding ATRA treatment, that an increase in the dosage or the duration of treatments results in an increase in the percentage of differentiated CD11b+ cells. See Additional File 1 for the percentages of CD11b expression of the various treatments.

From the contour plot, we identified two treatments that both gave rise to 54 percent of CD11b positive cells, namely, 1μM/5 Days and 0.05μM/7 Days. These two treatment combinations were picked because they produced similar levels of positive CD11b expression, yet one treatment is of higher dosage with shorter duration, while the other is of lower dosage with longer duration. We grew the HL60 cells under those conditions and isolated the CD11b+ population of these cells by FACS and re-cultured these cells in fresh, ATRA-free RPMI-1640 media. We collected the re-cultured cells and isolated the mRNA for whole-genome expression analysis each day for five days. At the end of this period, we also collected cells from both treatments for Wright-Giemsa staining, a histological method that could be used to determine hematopoietic cell types based on cell morphology. The 0.05μM/7 Days treatment resembles that of the untreated HL60 cells, with clear visible nucleoli and large nuclear to cytoplasm ratios, suggesting a reversal of cell fate back to the undifferentiated HL60 state; whereas the 1μM/5 Days treatment shows morphology resembling that of differentiated neutrophils, with characteristic decreased nuclear to cytoplasmic ratio, and convolution and segmentation of the nuclei, suggesting a completion of cell fate toward the differentiated neutrophil attractor. Our observation is in accordance with the notion of a "precommitment" state previously described [64–66], whereupon the

removal of the stimulating agent, the HL60 can revert back to the undifferentiated state. Taken together, we have established a system where we identified two perturbations that place the HL60 cells in different basins of attraction, leading to different eventual macroscopic cell fates.

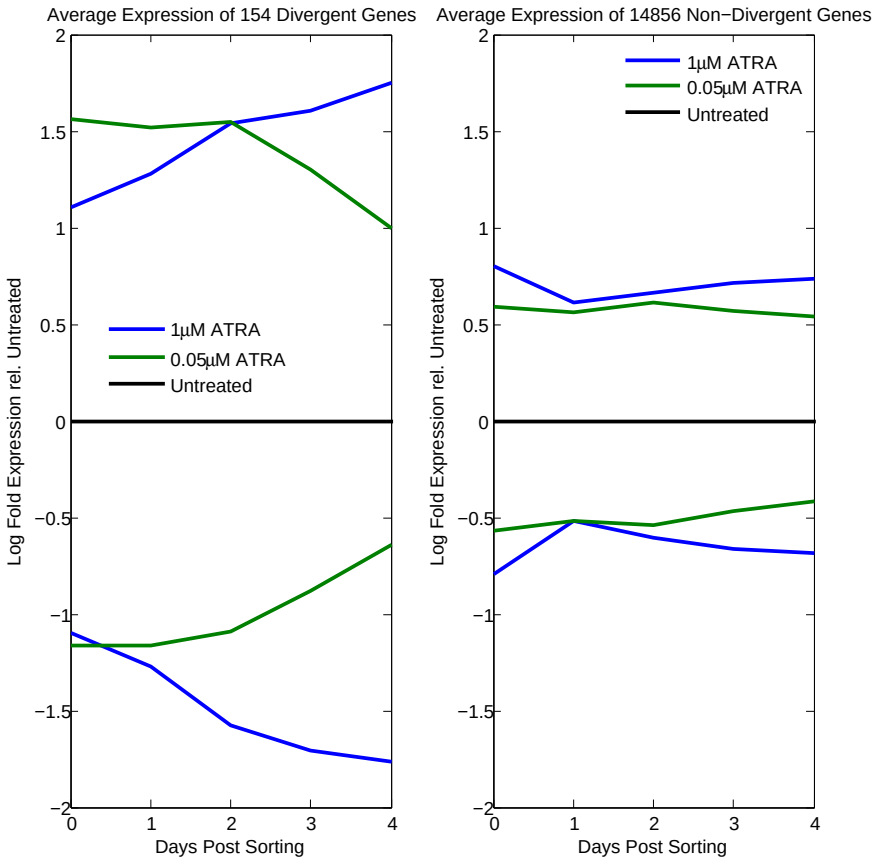

Figure 3.2: **Average mRNA Gene Expression of the Differentially Expressed Genes.** The left panel shows the divergent genes. The right panel shows the non-divergent genes. The upper panel represents the genes that are up-regulated in both 1μM/5 Day and 0.05μM/7 Days treatments relative to untreated. The lower panel represents the genes that are down-regulated in both 1μM/5 Day and 0.05μM/7 Days treatments relative to untreated.

3.3.2 A subset of the genes leading to different cell fate attractors exhibit a divergent expression pattern

To understand how the macroscopic cell fates that we observed could have arisen from these two perturbation conditions, we analyzed the gene expression profiles of the treated HL60 cells. We reasoned that we had placed the treated HL60 cells in their perspective basins of attraction when we re-cultured the sorted CD11b+ cells from these two treatments in ATRA-free media. Hence, the gene expression trajectories reflected the natural consequences of placing the HL60 cells in these specific parts of the genomic expression space. Therefore, when we looked at the gene expression profiles of differentially expressed genes between the two trajectories, the pattern we observed could potentially explain the observed macroscopic cell fate.

Interestingly, while the majority of the differentially expressed genes (relative to untreated controls) exhibit a flat and unchanging average gene expression profile under both (1μM/5 Days) and (0.05μM/7 Days) treatment conditions, there is a small subset of the genes (154, using our criterion) that exhibit a divergent gene expression profile. That is, after the high dosage/short duration treatment (1μM/5 Days), their expression levels deviate further and further away from their levels under the untreated ATRA condition, whereas their expression levels after the low dosage/long duration treatment (0.05μM/7 Days) converge toward the gene expression levels of those under untreated ATRA condition (Figure 3.2).

These divergent genes can be separated further into two distinct classes, the up-regulated, and the down-regulated genes. The up-regulated (respectively, down-regulated) genes are the ones that have elevated (respectively, repressed) expression under both high dosage/short duration and low dosage/long duration treatments relative to their expression under the untreated ATRA condition. In both cases, this display of differential expression behavior reflects the macroscopic cell fate observed, namely that the HL60 cells from the high dosage/short duration treatment continue toward differentiation whereas cells subjected to low dosage/long duration treatment

revert back toward the undifferentiated state. We hypothesized that these divergent genes participate in the selection of a particular attractor from a set of pre-existing ones. See Additional File 2 for a list of the divergent genes as well as magnitude of divergence.

3.3.3 Divergent genes promote cellular differentiation and repress cell cycle progression

After we identified the set of divergent genes and their unique gene expression patterns, we set out to investigate their known biological functions, with the goal of elucidating how these genes coordinate the transition of the HL60 cells from the promyelocyte attractor into the neutrophil attractor. In particular, cellular differentiation processes frequently entail an up-regulation of genes involved in specialization while simultaneously down-regulating genes related to proliferation and cell cycle [44, 45].

Indeed, a number of the up-regulated divergent genes are involved in the activation and specialization of neutrophils. At the top of the up-regulated divergence gene list is ankyrin repeat and SOCS box-containing 2 (ASB2), which is known to be a retinoic acid-response gene and a binding target of the promyelocytic leukemia retinoic acid receptor-alpha (RARα) oncogenic protein [67]. When ASB2 is expressed in leukemia cells, it inhibits growth and furthers myelocytic commitment, precisely as seen in the HL60 cell differentiation model system. Inherent to the process of neutrophil activation are genes promoting the homing and migration of neutrophils to the sites of inflammation. Examples of these genes among this list are orosomucoid 1 (ORM1), interleukin 8 receptor beta (IL8Rβ), and vanin2 (VNN2). ORM1 is highly expressed during acute inflammation and has been suggested as a signaling molecule that binds to L-selectin on the neutrophil cell surface to allow neutrophils to enter secondary lymphoid tissues via the high endothelial venules [68]. IL8Rβ, also known as CXCR2, is a receptor of interleukin 8 (IL8) and facilitates neutrophil migration to the site of inflammation [69]. Recently, its ligand was identified to be

the cytokine, macrophage migration inhibitory factor (MIF) [70]. Finally, expressed mainly on human neutrophils [71] and anchored on the cell surface by glycosylphosphatidylinositol (GPI), VNN2 physically associated with β_2 integrin (CD11b), the neutrophil differentiation marker for our study [72].

Besides neutrophil activation, response to inflammation is another process in which several of the up-regulated divergent genes including leukocyte immunoglobulin-like receptor subfamily B member 3 (LILRB3) and NOD9 (NLR X1) participate. Expressed LILRB3 protein binds to the MHC class I molecules on antigen-presenting cells to control inflammatory responses and cytotoxicity by transducing a negative signal that inhibits immune response and limits autoreactivity [70]. NOD9 is a member of the NLR family known to recognize microbial molecules that activate inflammatory caspases, causing cleavage and activation of inflammatory cytokines [73]. Collectively, these up-regulated divergent genes promote requisite activities of activated neutrophils.

The list of down-regulated divergent genes, on the other hand, contains many genes necessary for the progression of cell cycle. We found members of the well-known cyclin and cell division cycle (CDC) gene families [74] including cyclin d2 (CCND2), cyclin E1 (CCNE1), cell division cycle 2 (CDC2), cell division cycle 7 (CDC7), CDC28 protein kinase regulatory subunit 1B (CKS1B), and cell division cycle associated 5 (CDCA5). CCND2 forms a complex with cyclin dependent kinase 4 (CDK4) [75] and regulates cell cycle G0/G1 to S transition [76]. Similarly, CCNE1 is necessary for the progression of G1 to S transition. It interacts with cyclin dependent kinase 2 (Cdk2) to phosphorylate the target genes nuclear protein ataxia-telangiectasia locus (NPAT) and nucleophosmin, critical components of cell proliferation and DNA replication, respectively [77]. CDC2 encodes a Ser/Thr kinase and is the catalytic subunit of M-phase promoting factor (MPF), critical for G1/S and G2/M transitions [78]. CDC7 encodes a kinase that is essential for DNA replication as well as the transition between G1/S phase [79]. CKS1B binds to cyclin A for targeted degradation and passage

through the spindle checkpoint [80]. CDCA5, also known as sororin, controls the separation of sister chromatids during mitosis by stabilizing centromeric cohesin [81]. In addition to the cyclin and CDC gene families, other genes integral to cell cycle progression were also found. TTK protein kinase (TTK) is a cell cycle-regulated kinase with maximal activity during M phase, localizing to kinetochores [82]. It is required for centrosome duplication and normal progression of mitosis [83]. Kinesin family member 20A (KIF20A) accumulates in mitotic cells where it localizes to the midzone of the spindle during anaphase, and to the cleavage furrow and midbody during telophase, essential for cytokinesis to proceed [84]. Interestingly, KIF20A is a target of polo like kinase 1 (PLK1), a protein that we also identified as a divergent gene. Together, they are necessary for proper spindle assembly and function during anaphase and telophase of the cell cycle [85]. Jointly, the down-regulated divergent genes suppress the HL60 cells from progressing through the cell cycle.

To complement our literature-search approach to understanding the functions of these up and down-regulated divergent genes, we examined them separately in the context of Gene Ontology (GO) [86] enrichment analysis using GoMiner [87]. The 154 divergent genes include 48 up-regulated and 106 down-regulated divergent genes out of a total of 30729 unique genes on the Agilent array. However, at the relative conservative p-value and FDR level of 0.05 and due to a relatively small number of up-regulated divergent genes, we are only able to detect statistical significance of enrichment for the down-regulated divergent genes. The GO enriched results for the down-regulated divergent genes are listed in Table 3.1.

3.3.4 Transcription factor binding site enrichment in promoters of divergent genes

We suspected that there may be common regulatory mechanisms that control the expression of both up and down-regulated divergent genes to select for the neutrophil cell fate attractor and to efficiently activate the necessary biological processes. One common mechanism of controlling gene expression is through the regulatory actions of

transcription factors; therefore, we set out to search for enriched transcription factor binding sites (TFBS) in the promoter regions of both the up and down-regulated divergent genes simultaneously. We used a background model composed of upstream 2kb promoter sequences of all known genes and compared them to the upstream promoter sequences of the divergent genes by calculating the log likelihood ratio score to find the putative binding sites. Table 3.2 contains the sorted top 15 enriched transcription factor binding sites.

Functionally, these enriched TFBSs can be broken down into rough categories of tumor progression, cell cycle, development, or general transcription, which are processes actively engaged by the HL60 neutrophil differentiation model system. Let us consider the striking roles played by the transcription factors in tumor progression. We find TFBSs for C/EBPα, HLF, TAL1, HOXA4, MEIS1A, RAR-RXR, and FOXO4, all of which , when disregulated, have been shown to lead to leukemia. C/EBPα is crucial for the process of granulopoiesis, and its aberrant expression in acute myeloid leukemia is well-studied [88–91]. Conditional expression of C/EBPα leads to neutrophil differentiation while dominant-negative expressions of C/EBPα are found in AML patients [88]. HLF, together with E2A, forms a chimeric E2A-HLF transcription factor protein due to a translocation mutation t(17;19)(q21;p13) which occurs in a subset of acute lymphoblastic leukemias [92–94]. TAL1 is essential for early stage embryonic hematopoiesis; erythroid differentiation is associated with increased TAL1 expression, while myeloid differentiation is associated with decreassed TAL1 expression [95]. A TAL1 translocation mutation t(1;14)(p32;q11) is observed in 3% of patients with T cell acute lymphoblastic leukemia [96]. HOX genes are an important transcription factor family for hematopoiesis, the different family members of which are required to specify particular stages of hematopoietic development [97]. Further, HoxA4 promoter hypermethylation has recently been linked to mutational status in chronic lymphocytic leukemia [98]. AML/Runx1 is also linked to hematopoiesis and leukemic tumor progression [99]. It is the DNA binding el-

ement of the core binding factor (CBF) transcription complex and is required for hematopoiesis as shown in knock-out mouse studies [100]. Mutations of RUNX1 have been identified in familial platelet disorder (FPD) along with a congenital predisposition to the development of acute myeloid leukemia (AML) [101]. MEIS1, another protein with enriched TFBS, cooperates with both HOXB3 and HOXA9 to induce the transcription factor AML [102, 103] by down-regulating its expression through promoter hypermethylation in a subset of AMLs [104]. The RAR-RXR heterodimer TFBS is also enriched in the promoter regions of our divergent genes. Since the HL60 neutrophil differentiation is induced through the actions of retinoid acid, it is reasonable that we observed an enrichment of retinoid acid receptor binding sites. Though it is well-established that the chimeric fusion protein from RAR-PML translocation mutation, t(15;17), is frequently associated with acute promyelocytic leukemia [105], methylation analysis of the RARα promoter further cements its involvement [106]. Finally, FOXO4 has also been linked with acute leukemias. A translocation mutation t(X;11)(q13;q23), which fuses it with the gene MLL, was observed and cloned [107]. In summary, the list of enriched TFBSs recapitulates many important regulators of hematopoiesis, which are intimately tied to leukemia pathology, illustrating the potential utility of such systems-level experimental designs.

3.4 Discussion

In this study, we perturbed the HL60 cells into the basins of attraction of two distinct cell fate attractors using two different ATRA dosage/duration treatments such that both cell populations are poised at the same stage of differentiation. By tracking the gene expression changes en route to these cell fates, we found a subset of the differentially expressed genes that exhibited a divergent gene expression pattern, hypothesized to correspond to the observed macroscopic cell type phenotype. Literature searches identified the possible functions of these divergent genes to be involved in promoting neutrophil differentiation and repressing cell cycle progression. Analyses of

the promoter sequences of the divergent genes further showed that they are enriched with transcription factor binding sites known to be linked to hematopoiesis, tumorigenesis, cell cycle, and development, suggesting the utility of systems level analysis for deriving valuable molecular level insights.

It is worth noting that our study suffers from a number of inherent limitations. With our attractor-based experimental setup, we could not detect early onset genes that lead to the "precommitment" state. Since gene expression profiles are only recorded after the two populations of the cells have achieved similar percentages of CD11b expression, prior cellular events of interest that culminate in their perspective promyelocyte and neutrophil attractors would be missed. In addition, since our study is based on a population of cells, inherent to all microarrays studies, only measurements of the average cellular behavior are possible. Indeed, it is known that the expression kinetics on a single cell level can exhibit an all-or-none switch like behavior, unlike the seemingly gradual change of expression when measured as a population average [64]. Further, recent evidence now suggests that transcriptional noise inherent to individual cells underlies clonal heterogeneity [129]. In light of this, an analysis on the gene expression changes of individual cells flowing toward the promyelocyte and neutrophil attractors would provide valuable insights.

Our study also suggests a number of natural extensions. For example, since transcription factor binding sites frequently occur in clusters and exert their effects simultaneously, instead of looking for enrichment of individual transcription factors, one can investigate enrichment of multiple transcription factors. Another possible extension is the incorporation of protein-protein interaction networks in order to identify potential co-activators of the transcriptional complexes governing HL60 differentiation. Further, one can search for common interaction partners to multiple enriched transcription factors. Additionally, in our characterization of divergent genes, comparisons of gene expression were made on a daily basis. To mitigate the effects of measurement noise and daily fluctuations, it is possible to model the entire time

course gene expression profile for each gene by fitting a regression curve, promoting a possibly more robust identification of divergent genes.

Our study also raises an important question — can the concept of cell fate be sufficiently described by the use of one or few markers? Traditionally, cell fate has been intimately tied to the expression of cell surface receptors. However, in our study, two populations of ATRA-stimulated HL60 cells can both exhibit characteristic cell fate markers, yet be destined to have distinct cell fates, namely a promyelocyte or a neutrophil. This suggests two populations of cells may have the same "apparent" state as measured by cell surface markers while differing in other state space dimensions, ultimately leading them to disparate cell fates. Likewise, Chang *et al.* [64] also came to similar conclusions in their observation that low-dose DMSO-treated CD11b- cells are in a "primed" differentiated state as compared to DMSO-untreated CD11b- cells.

This study suggests that systems-level dynamics, such as the partitioning of the state space into distinct basins of attraction, have the potential to convey information about the molecular-level control of biological processes.

3.5 Methods

3.5.1 Cells and chemicals

Early passaged HL60 cells were generously provided by Dr. Steven Collins (Fred Hutchinson Cancer Research Center). The cells were cultured in T/25 flasks with RPMI-1640 media (Gibco) supplemented with 10% heat-inactivated fetal calf serum (Sigma). Cells were grown in media containing All-Trans Retinoic Acid (Sigma) to induce neutrophil differentiation [58]. ATRA was stored in -30 °C and dissolved in 95% ethanol to make stock solution.

3.5.2 Dosage and duration of ATRA treatments

A two-factor dosage and duration experiment was set up with eight and ten levels, respectively. HL60 cells were subjected to the following dosages (μM): 1, 0.5, 0.1,

0.05, 0.01, 0.005, 0.001, and 0.0005 in conjunction with the following durations (Days of Treatment): 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13. For example, one of the eighty treatment combinations is 1μM/4 Days, in which, we treated the HL60 cells with 1μM of ATRA for 4 Days and measured the CD11b expression of these treated cells by flow cytometry. Each of the 80 dosage/duration combinations and CD11b measurements were performed in triplicate.

3.5.3 Flow cytometry to detect surface CD11b expression

Surface CD11b is an early marker of neutrophil differentiation in HL60. Cells (10^6) were harvested, washed twice with PBS buffer, and incubated on ice for an hour with PE-conjugated CD11b antibody or its isotype control (Mouse IgG1κ) from BD Pharmingen. Cells were then washed two more times with PBS buffer and fluorescence was measured by FACSCalibur using CellQuest software (BD Biosciences). The CD11b expression levels were compared to the isotype control to correct for any non-specific binding. Untreated HL60 cells stained with the isotype control were used as background for undifferentiated cells.

3.5.4 Dosage and duration contour plot construction

One million cells were collected for each experiment. Percentages of CD11b+ expression were calculated by comparing ATRA treated HL60 samples to untreated samples. Triplicate values of the percentage of differentiation were averaged and the result was displayed as a contour plot (Figure 3.1). In constructing the dosage/duration contour plot, one round of linear interpolation was performed to obtain a finer sampling of the dosage/duration grid.

3.5.5 Wright-Giemsa stain to observe nuclear morphology

Treated and untreated cells (10^5) were harvested, washed once with PBS buffer, spun down on microscope slides using Cytospin 2 (Shandon Inc.) and air-dried. Slides were

then flooded with 2mL Modified Wright-Giemsa stain (Sigma-Aldridge) and soaked for 1 minute. 2mL deionized water was added, the slides were soaked for 3 minutes, rinsed with deionized water, and air dried.

3.5.6 Microarray experiment and analysis to measure changes in mRNA levels

Two treatment combinations that yielded similar levels of CD11b expression were identified from the dosage and duration contour plot: 1) high dosage/short duration (1μM ATRA/5 Days) and 2) low dosage/long duration (0.05μM ATRA/7 Days). HL60 cells were cultured under these two conditions and fifteen million CD11b positive cells were collected from each condition using the Cytopeia inFlux V-Gs high-speed sorter. Sorted CD11b positive cells were then re-cultured in fresh ATRA-free RPMI media. After six hours of allowing these cells to recover from the sorting process, one million cells were collected. For the next five days, one million cells were collected every twenty-four hours, culminating in a total of 5 time points —Day 0 (6 hours post-sorting) to Day 4 (102 hours post sorting). For each time point, total RNA from cell samples were isolated with Trizol and quantified using Thermo Scientific NanoDrop 1000. RNA quantity for three technical replicates was collected for each time point, except day0, day1, and day4 of the 1μM treatment, where unfortunately the RNA quantity was only sufficient for two replicates, resulting in a total of 27 microarray experiments on Agilent human whole genome oligo arrays with 44k 60-mer probes. ATRA-treated samples from each time point were labeled with the Cy5 dye, while untreated HL60 cells were labeled with the Cy3 dye for comparison. Each hybridized array was scanned with the Agilent dual laser-based scanner. Feature Extraction software version 8.0 (Agilent Technologies) was used to output the relative fluorescence intensity between the treated and untreated samples.

After the microarray experiments, the slides were scanned and the raw spot intensity (gProcessedSignal and rProcessedSignal from the feature extraction software) were used for subsequent data analysis in Matlab (The Mathworks, Inc). Of the 43931

spots on the Agilent array, spots that were designated for quality control, spots that were saturated, and spots that had signals too low to be detected were filtered out, resulting in 41509 spots remaining. The log intensity values were normalized using quantile normalization [130]. Replicate array intensity values were averaged to obtain the mean fluorescence intensity. Differentially expressed genes were picked using SAM [131] (Two class Paired). After this filtering step, 14949 differentially expressed genes remained. We compared the expression of 1μM ATRA/5 Days treatment with 0.05μM ATRA/7 Days treatment to identify genes that showed a steady daily increase of divergence in expression of 5% or larger, and designated these as divergent genes. Of these genes, those with elevated (respectively, suppressed) expression under both 1μM ATRA/5 Days treatment and 0.05μM ATRA/7 Days treatment relative to the untreated condition during the first three time points were deemed to be up-regulated (respectively, down-regulated). Comparisons of expressions were done on a daily basis, yielding a total of 176 divergent probes, 48 up-regulated and 128 down-regulated. A complete list of divergent gene probes is available in the supplementary section. Gene probes are sorted based on $\log_2$ fold divergence of Day 4 expression between 1μM ATRA/5 Days treatment and 0.05μM ATRA/7 Days treatment. The 176 divergent spots identified corresponded to 154 unique genes, 48 up-regulated genes and 106 down-regulated genes (some genes have multiple probes on the array).

3.5.7 Functional enrichment of divergent genes by Gene Ontology (GO)

Up-regulated divergent genes and down-regulated divergent genes were submitted separately to GoMiner [87] for an analysis of enriched biological processes at the p-value level of 0.05 and FDR of 0.05.

3.5.8 Searching for enriched transcription factor binding sites

Upstream 2kb promoter regions of the divergent genes were downloaded from EMBL (NCBI36) using the BioMart interface [132]. Prepackaged upstream 2kb regions of all

RefSeq genes were also downloaded using the UCSC genome browser (hg18) [133]. After filtering out duplicated sequences and sequences containing ambiguous nucleotide bases, 18827 sequences remained. 429 human TRANSFAC (Professional 9.4) matrices [134] were used to calculate the log likelihood ratio scores of transcription factor binding for the divergent genes as well as the RefSeq genes, as

$$L = \sum_{i=1}^{n} \log(P_i(M_s|s_i)) - \sum_{i=1}^{n} \log(P_i(M_b|s_i)) \tag{3.1}$$

where L is the log likelihood ratio score, M_s is the TRANSFAC matrix model, M_b is the zeroth order Markov background model with frequencies of A:0.2583 C:0.2457 G:0.2425 T:0.2535, which were calculated by counting the occurrences of the nucleotides in all promoter sequences. s_i is the ith nucleotide of the motif site under consideration, and n is the length of the motif site.

Since binding sites tend to occur in clusters in higher eukaryotes [135], attention was paid to find stretches of the DNA sequences (100bp) with large numbers of putative binding sites. Hence, log likelihood ratio scores from all 429 TRANSFAC matrices were summed at each nucleotide position for all RefSeq promoter sequences. The top 1% of these 100bp highest scoring regions were picked as cut-off values to represent regions with clusters of putative binding sites. This cut-off value was then used to search for clusters of putative binding sites in the divergent genes, resulting in 262 of these clusters of binding sites being identified for the divergent genes. Expected numbers of binding sites were calculated by counting the total number of binding sites within the high-scoring regions divided by the total number of high-scoring regions for each transcription factor. This calculation was repeated for each TRANSFAC matrix. Enriched transcription factor binding sites were then ranked by the differences between the expected values for the divergent gene set and the set of all RefSeq promoters (Table 3.2).

Table 3.1: Enriched GO Categories of Down-Regulated Divergent Genes

GO Terms and Description	Total Genes	Changed Genes	-Log10(P)	FDR
(Biological Process)				
GO:0007049 Cell cycle	597	14	5.7044	0
GO:0051301 Cell division	187	7	4.2512	0.005
GO:0022402 Cell cycle process	334	9	4.22	0.003
GO:0045739 Positive regulation of DNA repair	5	2	3.5589	0.03
GO:0051726 Regulation of cell cycle	248	7	3.4893	0.026
GO:0000082 G1 S transition of mitotic cell cycle	30	3	3.2793	0.032
GO:0009132 Nucleoside diphosphate metabolic process	7	2	3.2397	0.036
GO:0051329 Interphase of mitotic cell cycle	76	4	3.1561	0.035
GO:0006259 DNA metabolic process	373	8	3.1266	0.031
GO:0006282 Regulation of DNA repair	8	2	3.1162	0.039
GO:0051325 Interphase	79	4	3.0926	0.036
GO:0006396 RNA processing	298	7	3.0138	0.038
GO:0009262 Deoxyribonucleotide metabolic process	9	2	3.0086	0.042
GO:0007059 Chromosome segregation	38	3	2.9748	0.04

Table 3.2: Enriched Transcription Factor Binding Sites

TF Name	Experimental Link	Functional Category	Reference
OCT-1/POU2F1	Regulates Cyclin D1 w/ STAT5	Cell Cycle	[108]
SOX-9	Chondrogenesis	Development	[109]
	RAR agonist stimulates SOX9 expression	Retinoic Acid Regulated	[110]
CDP/CUTL1	Mediate cell cycle progression	Cell Cycle	[111]
AML/RUNX1	Hematopoiesis	Hematopoiesis	[112]
AR	Inhibit prostate cancer cell growth	Tumor Progression	[113]
STAT6	Induction of apoptosis	Tumor Progression	[114]
MRF2/ARID5B	Unknown	Other	
LEF1/TCF1	Neutrophil granulopoiesis	Hematopoiesis	[115]
		Cell Cycle	[116]
AP-1/JUN	Myeloid leukemia	Tumor Progression	[117]
TAL1/SCL	Lymphocytic leukemia	Tumor Progression	[118]
HLF	Lymphoblastic Leukemia	Tumor Progression	[92]
HMGIY	Myeloid metaplasia	Tumor Progression	[119]
TFCP2	Erythroid differentiation	Hematopoiesis	[120]
TBP	Required for Pol I,II,III	General Transcription	[121]
Direct Repeat 1(RAR-RXR/PPAR/HNF4)	Acute promyelocytic leukemia	Tumor Progression	[106]
	PPAR correlates with breast cancer progression	Tumor Progression	[122]
E2F	Regulates Cyclin A	Cell Cycle	[123]
FOXO4/AFX	Mixed lineage leukemia	Tumor Progression	[107]
Pax-1	Development of Vertebral Column	Development	[124]
PGR	Cell cycle dependent transcriptional activity	Cell Cycle	[125]
AP4	Interacts with AP1 to affect transcription	Other	
IPF1	Pancreas Development	Development	[126]
Ets	Melanoma	Tumor Progression	[127]
p300	Trancriptional co-activator complex	General Transcription	[128]

Chapter 4

A GENOME-WIDE ANALYSIS OF TRANSCRIPTIONAL REGULATION BY T-BET

4.1 Introduction

$CD4^+$ T cells play a paramount role in coordinating a coherent adaptive immune response. Depending on their initial encounters, may they be with extracellular, intracellular infectious pathogens, or with self antigens, the $CD4^+$ T cells secrete appropriate chemical signals to direct B cells, $CD8^+$ T cells, macrophages, eosinophils, and neutrophils to present an appropriate concerted response [25, 27, 136]. They achieve this incredible feat by differentiating into a number of functionally distinct T-helper lineages (Th1, Th2, Th17, regulatory T-cells, and follicular T-cells): In response to intracellular pathogens, naive $CD4^+$ T cells differentiate into Th1 cells [137]; extracellular parasites induce Th2 differntiation [138]; extracellular bacteria and fungi prompt Th17 differentiation [139]; regulatory T cells direct immune tolerance [140]; and follicular T-cells facilitate B cell formation [141].

The focus of this study is on the Th1 differentiation process, as orchestrated by the lineage-defining transcription factor, T-bet [142]. Its roles in the Th1 differentiation process have been the subject of many classical studies, particularly its regulation of the Th1 principle cytokine, Ifng [20, 142–144]. Nevertheless, advances in genomic technologies have enabled investigators to probe this differentiation process at the genome-wide level. A recent study identifies many T-bet targets by comparing the ChIP-Chip binding locations between Th1 and Th2 cells [145]. What remain elusive, however, are the mechanisms that T-bet utilize to regulate its target genes on a genome-wide level.

We investigate this question by dissecting it into two parts: mechanisms that are

extrinsic to the T-bet protein itself, and those that are *intrinsic*. Extrinsic to the T-bet proteins are the local genomic contexts of T-bet target genes. A thorough study of these local signals (histone methylation, transcription factor binding, DNase I hypersensitivity, and sequence conservation) has the potential to reveal regulatory insights of the target genes. Intrinsic to the T-bet protein itself are the transcriptionally regulated activities associated with T-bet protein domains. Previous studies have shown that the chromatin-modifying activity resides within the T-box binding domain and the transactivating activities resides within the N and C terminal domains. Though these domains are required for T-bet to regulate its target genes such as Ifng and Cxcr3 [18, 146, 147], it is unclear if these same mechanisms are also required to regulate other T-bet targets. Further, to study how T-bet targets are regulated in other T helper subtypes, we expand our analyses of T-bet target gene expression into Th2, Th17, iTreg, nTreg, and follicular T cells.

In this chapter, we address the key question of how T-bet functionally regulates its target genes on a genome-wide level. We first identify a set of T-bet bound target genes by comparing wild type and T-bet$^{-/-}$ CD4^{+} T cells with ChIP-Seq analysis. We then examine differentially expressed genes between wild type and T-bet$^{-/-}$ by gene expression microarrays. We compile and analyze a large number of genomic data sets to hunt for signals that are indicative of the functional regulation of T-bet targets. We study the regulation of T-bet targets by mutational analysis of T-bet protein domains and their associated functional activities. And finally, we find that T-bet target genes are also differentially regulated under other T helper lineages.

4.2 Results

4.2.1 T-bet binding is enriched within the promoter regions

To understand how a lineage-determining transcription factor such as T-bet establishes the genome-wide gene expression profile conferring the Th1 phenotype, we sought to identify the direct targets that T-bet regulates. For this purpose, naive

$CD4^+$ T cells were isolated from wild type and T-$bet^{-/-}$ mice and grown under Th1-skewing conditions for 6 days and harvested for ChIP-Seq analysis. A total of 9158 peaks were identified using MACS 1.4 (Model-based Analysis of ChIP-Seq) from two separate experiments and mapped to 7786 genes [148]. In accordance with previous convention [149], we divided the mouse genome into 4 regions: promoter (transcription start site [TSS] to upstream 10kb), exonic, intronic, and intergenic regions. Definitions of gene attributes including gene boundaries, transcript and UTR locations were downloaded from Ensembl mouse genome based on the NCBIM37 build [150].

We calculated the proportions of T-bet peaks to center in the following four genomic regions: 26% within the promoter region, 6% exonic region, 45% intronic, and 23% intergenic (Figure 4.1A). This distribution is in sharp contrast to the mouse genome itself where 2% of the genome are promoter regions, 0.2% exons, 37% intronic, and 61% intergenic regions [149]. Similar to previous reported studies, we also observed substantial enrichment of binding within the promoter region and exonic regions of the mouse genome [149, 151].

To identify the consensus binding motif for T-bet, we downloaded from the Ensembl database, repeats-masked genomic sequences that are 50 base pairs away from the T-bet summits. Within these sequences, we searched for de novo motif enrichment using MEME [152]. These sequence motifs were then aligned to known transcription factor binding matrices in TRANSFAC [153] using TOMTOM [154]. Not surprisingly, the top enriched sequence motif (T-bet) closely resembled the binding matrices of other T-box family members, including Tbx5, Tbx15, Tbx22, Tbx18, and Brachyury. The core motif of GTGxxA is preserved in all the above listed members. In addition, we also observed enriched sequence motifs that were aligned to AML/Runx, Ets, and Sp1 (Figure 4.1B). The AML/Runx family is critical for the CD4/CD8 lineage choice [155, 156], and recent studies have further shown that they are important co-factors for the specification of Th2 cells [21], Th17 cells [157], and regulatory T cells [158]. Similarly, Ets1 participates in T helper differentiation, and has been iden-

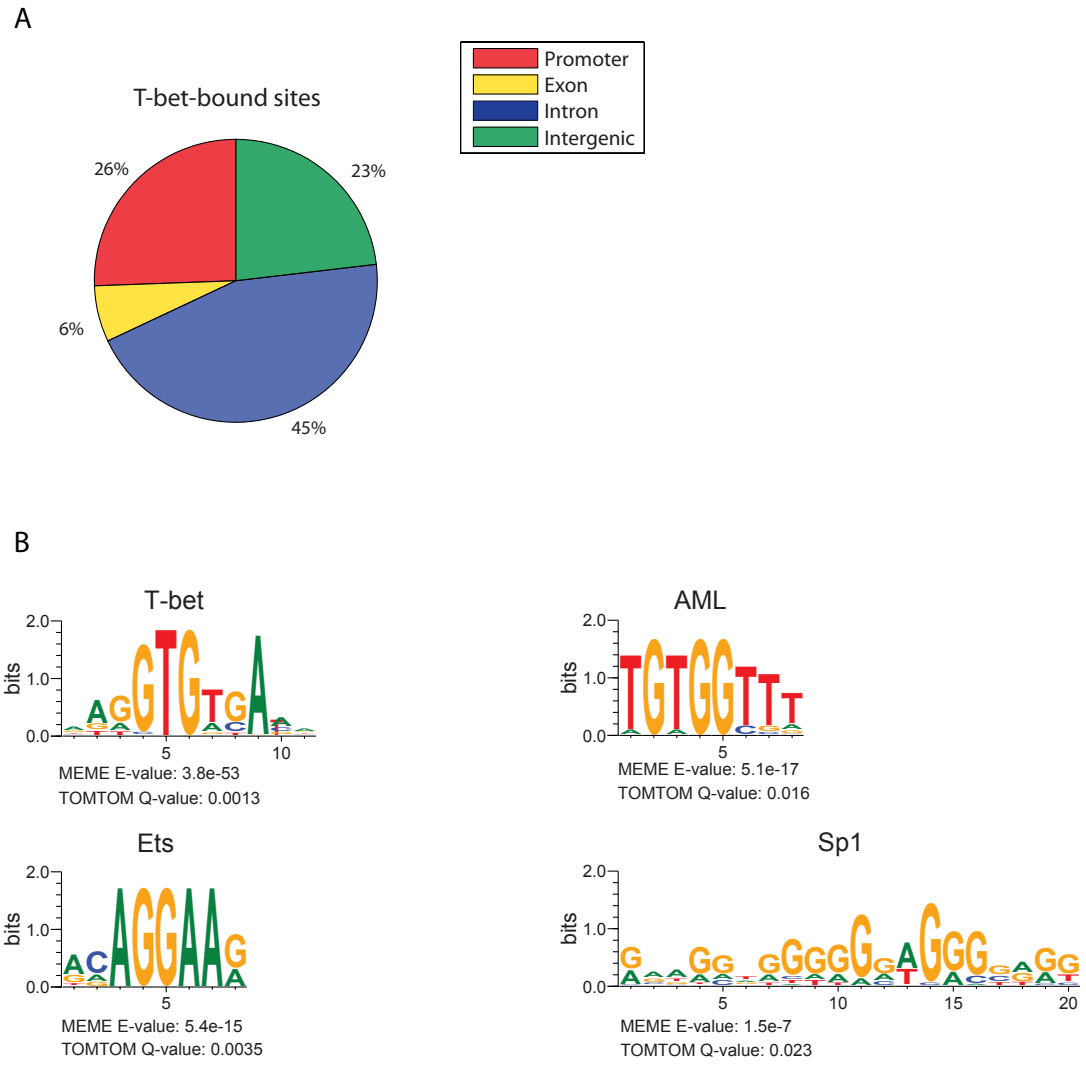

Figure 4.1: **Genomic Distribution of T-bet Binding Site and Enriched Consensus Sequence Motifs.** (A)ChIP-Seq analyses of T-bet binding sites from two independent experiments were determined by MACS 1.4 (Model-based Analysis of ChIP-Seq) and mapped to the mouse genome based on location: promoter (within 10kb upstream from the transcription start site), exon, intron, and intergenic regions. (B)Mouse genomic sequences from the summit of T-bet ChIP-Seq peaks were analyzed for enriched consensus sequence motifs by MEME. The sequence motifs were then aligned to TRANSFAC databases to uncover potential T-bet co-regulators using TOMTOM.

tified as a T-bet cofactor and negative regulator of Th17 lineage specification [159]. Though much less is known about Sp1's role in T helper cell differentiation, it has been shown to physically interact with T-bet as a member of T-bet/Gata-3 complex that regulate Th1 gene expression [160]. Taken together, we identified T-bet bound sites and found potential co-regulators by analyzing enriched sequence motifs of T-bet bound sequences.

4.2.2 Combined gene expression and ChIP-Seq data uncovers direct T-bet targets

One of the key insights from previous reports is that many T-bet bound genes are not functionally regulated by T-bet [161]; thus, we sought to identify the functional T-bet targets among the T-bet bound genes. To this end, we collected wild type and T-bet$^{-/-}$ CD4^{+} T cells grown under Th1 polarizing conditions and harvested RNA on days 1, 2, 4, and 6 to track the gene expression changes over time. Differential gene expression was determined using SAM (Significance Analysis of Microarrays) [162] for paired studies comparing wild-type and T-bet$^{-/-}$ (FDR<0.03) on all days. We observed 206 differentially expressed genes, and 7786 genes with T-bet binding (9158 binding sites). (Figure 4.2A)

We found two broad categories of genes that are bound by T-bet: functional targets that are also regulated on the RNA level, and non-functional targets that are not regulated. Interestingly, of the 7786 genes that are T-bet bound, only 117 of those are functional, suggesting that the majority of the observed T-bet binding sites actually do not have functional gene expression changes. The abundance of non-functional binding sites is in agreement with a number of published genome-wide studies [149, 161, 163] and raises the intriguing question as to why there are so many binding events, yet so few differentially expressed genes (a question that we further examine in later sections). Of the 117 functional T-bet targets, 64 are activated by T-bet and 53 repressed. Many of these functional T-bet targets are known immune response genes including: Ifng, Cxcr3, Nkg7, Ifngr1, Prf1, Klrd1,

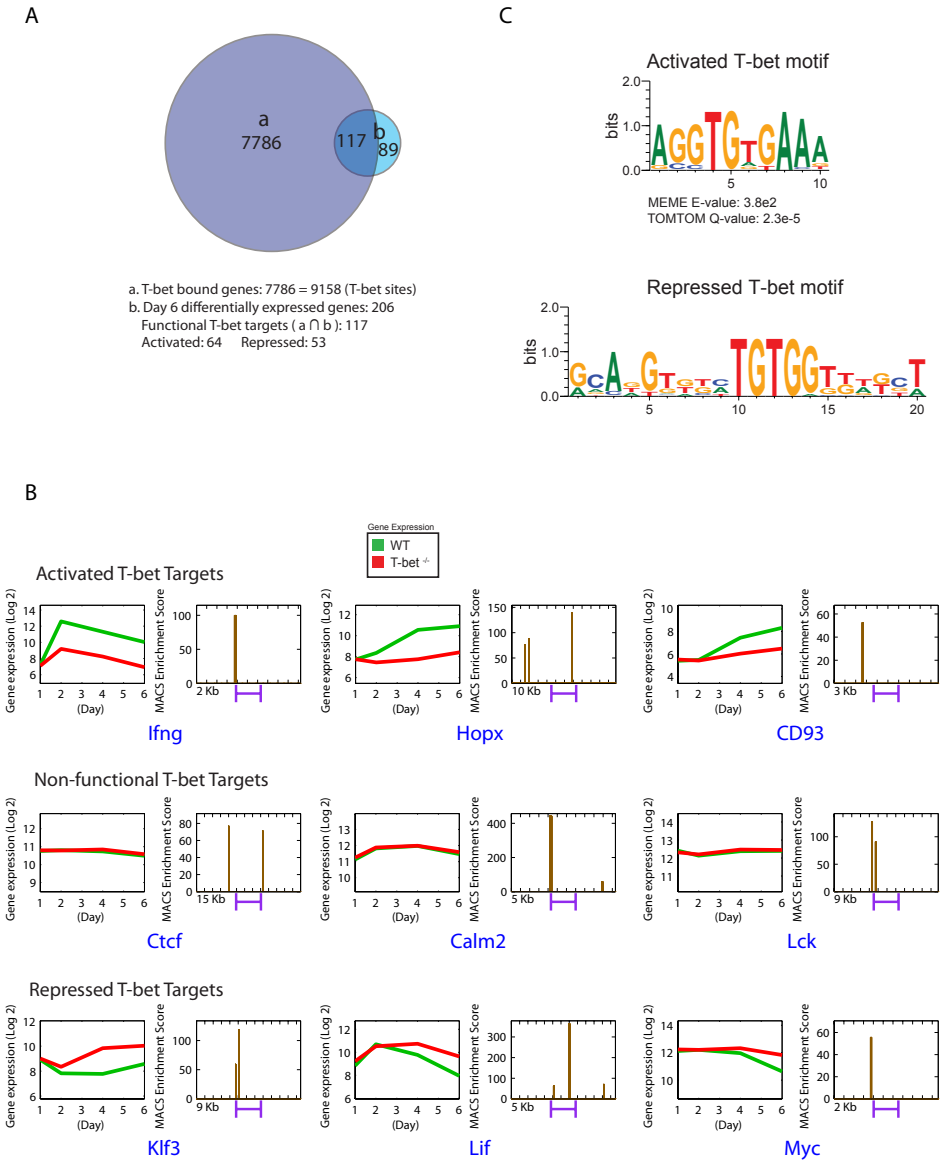

Figure 4.2: **Most T-bet bound genes are not functional.** (A)Venn diagram shows overlap between genes bound by T-bet and genes showing differential expression in wild type to T-bet$^{-/-}$ comparison. (B)Examples of T-bet bound genes that show activation, repression, and no functional consequences. Differential gene expression over time (left). Relative position of T-bet peaks to T-bet bound genes (right). Genes are directed from left to right. Gene sizes are normalized for ease of viewing. Scales are indicated by the distances between the tick marks. (C)Mouse genomic sequences from activated and repressed T-bet targets peaks analyzed by MEME identify enriched consensus binding motif. Activated T-bet targets are enriched with canonical T-box binding motif. Repressed T-bet targets are enriched with a weak T-box binding motif followed by a strong AML binding motif.

Nod1, FasL, Id3, Tcf7, Irf7, Nrp1, Cd86, and Il21. Shown in Figure 4.2B are examples of functional gene targets and non-functional gene targets with both their time point gene expression profiles as well as locations of the T-bet binding relative to the position of the genes.

To uncover potential sequence motif differences between activated and repressed functional T-bet targets, we analyzed 50 bp genomic sequences surrounding the summits of activated and repressed T-bet binding peaks and searched for enriched de novo sequence motif with MEME. Surprisingly, the canonical T-bet binding motif was only detected in the activated T-bet targets, but not the repressed T-bet targets. On the other hand, a weak T-bet binding motif followed immediately by a AML motif was detected for the repressed targets, and not the activated targets (Figure 4.2C). Our result corroborates with recent reports that link AML/Runx family proteins to T helper cell differentiation. Djuretic et al. found T-bet and Runx3 physically interact to bind collaboratively to activate expression of Ifng and silence expression of Th2 cytokine, IL4 [20], enforcing Th1 differentiation. Similarly, Lazarevic et al. found physical interactions between T-bet and Runx1(AML1) that reverses T-bet's inhibition of Th17 polarization [164]. Collectively, analysis of enriched sequence motifs between activated and repressed T-bet targets reveal key co-regulators of Th1 differentiation.

4.2.3 Stat4 binding site properties are most informative for predicting functional vs. non-functional T-bet target genes

As mentioned in the previous section, given that T-bet binds to thousands of genes across the genome, why is it that there are relatively few "functionally" regulated genes? In this section, we search for mechanisms *extrinsic* to the T-bet protein by analyzing the local genomic contexts of T-bet targets. We take a statistical machine learning approach to find distinguishing features that separate functional T-bet targets and non-functional T-bet targets.

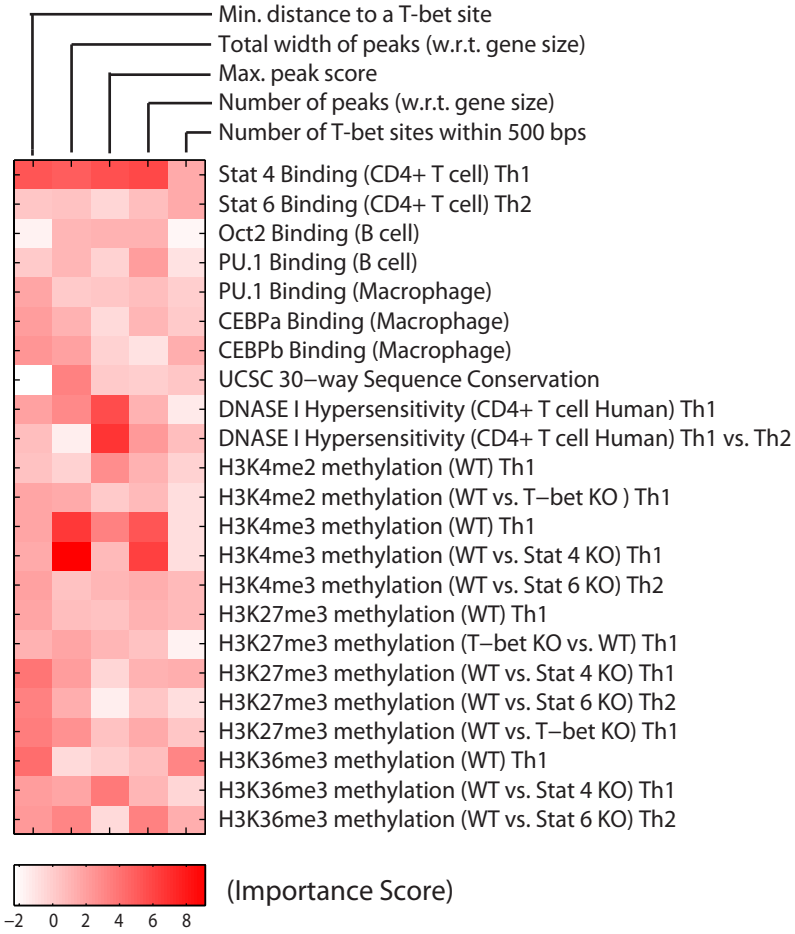

Figure 4.3: **Genomic signals from Stat4 are the most "important" predictors for a functional T-bet target gene.** Different sources of genomic data such as transcription factor binding sites, sequence conservation, DNAse hypersensitivity, and histone methylation were collected. Genomic features were then extracted from these sources of data for the prediction of "functional" T-bet targets. A random forest classifier was built to help identify patterns within these features. "Importance" scores from the random forest classifier are displayed in heatmap representation. Importance scores provide a measure of significance for the different features utilized for the classification task. The rows represent the different sources of genomic data. The columns represent the extracted features. The colors represent the magnitudes of the importance score.

We generated and collected a number of genome-wide data sets to capture the local genomic contexts, including histone methylation, transcription factor binding, DNase I hypersensitivity, and sequence conservation. From our previous study, we found T-bet induces permissive H3K4me2 histone methylation marks and remove repressive H3K27me3 marks [18]. Hence, we generated H3K4me2 and H3K27me3 ChIP-Seq data from wild type and T-bet$^{-/-}$ cells. Additionally, we downloaded H3K4me3, H3K36me3, and H3K27me3 histone methylation data for Stat4 and Stat6 from Gene Expression Omnibus (GEO) [149]. These transcription factors are key signaling molecules for Th1 and Th2 differentiation. H3K4me3 and H3K36me3 histone methylation marks correlate with active transcription while H3K27me3 marks correlate with silenced transcription [14]. We also found sequence conservation (PhastCon Multiz 30-way) [28] and Encode DNase I hypersensitivity data for Th1 and Th2 polarizing conditions [165] from the UCSC genome repository. Sequence conservation data has the potential to highlight important genomic regions selected by evolution, while DNase I hypersensitivity data reveals the open regions of the chromatin. However, the DNase I hypersensitivity data was collected in human; thus, we mapped the data from human to mouse using the lift-over tool provided by the UCSC genome repository. Finally, as negative controls, we also included ChIP-Seq data from PU.1, Oct2, CEBPα, and CEBPβ [166]. These transcription factors are not known to play important roles in T helper differentiation process, and were harvested from B cells and macrophages; therefore, we do not expect to gain significant information from them.

From the above listed types of data, we construct 23 genomic data sources(as represented in Figure 4.3 rows). For each data source, we extract 5 features (as represented in Figure 4.3 columns). The features are as followed: 1) Minimum distance between the peaks of a gene and the closest T-bet site, 2) Total width of the peaks (normalized to gene size), 3) Maximum score of the peaks, 4) Number of total peaks (normalized to gene size), and 5) Maximum number of T-bet sites within

a 500bp window. Our goal is to identify relationships between genomic features and functional T-bet regulation by extracting structural patterns within the data sets. In particular, we are interested in determining which genomic features are most informative for predicting "functional" T-bet targets. The answer to this question can potentially improve our understanding of the Th1 differentiation process.

The statistical machine learning approach we utilized is random forest [29]. It is an ensemble method that builds a collection of decision trees and outputs a classification result by majority voting. It handles class imbalance by inversely weighing misclassification cost to class sizes. Its key advantages include an output ranking of the attributes based on their "importance" in determining a classification task. Figure 4.3 is a heat map representation of the importance Z score from the random forest classifier for the 23 sources of genomic data and 5 features that we collected. The more important the attribute, the higher the score, the darker is the shade of the color red on the map.

For genes that are bound by T-bet, Stat4 binding and H3K4me3 methylation patterns are the two most distinguishing sources of data for determining whether or not T-bet functionally regulate them. All but one of the features derived from Stat4 binding are very informative. Since Stat4 is known to be integral to the signaling of Th1 differentiation [167,168], in retrospect, it is not surprising to find Stat4 binding to be the most informative source of data. In particular, the distance to a T-bet binding site seems to be the most important feature. This is encouraging because Stat4 is required by T-bet to promote Th1 differentiation [169]. The other important source of data, H3K4me3 methylation pattern, has two informative features: 1) total width of peaks and 2) number of peaks. Since these two features tend to be correlated, they both have high importance scores. Interestingly, using $Stat4^{-/-}$ methylation as a background to identify Stat4-specific histone methylation peaks improves the classification. This points to a model where Stat4 binding is the key determining event for T-bet transcriptional regulation and Stat4-induced H3K4me3 methylation

as a downstream effect.

There are also notable differences between transcription factor and histone methylation data. Transcription factors operate on specific sequence motifs, whereas histone methylation can affect large genomic areas. Unlike Stat4 peak data, the total size of H3K4me3 methylation is more important than the distance to T-bet sites. Most likely, as a transcription factor, Stat4 only binds DNA at specific sequences, whereas histone methyl transferases can operate over a much longer range, hence, the distinguishing features for the two data sources are different. There are fundamental differences between histone methylation marks as well. The total width of methylation marks is an important feature for H3K4me3, but not for H3K27me3 and H3K36me3. For a functional T-bet target, it is more important to have K27 and K36 marks close to a T-bet site, than to have long stretches of K4 methylation. Interestingly, the DNase I hypersensitivity score also seem to be indicative of functional T-bet targets. Target genes that are "really" open during Th1 differentiation process are likely to be a functional T-bet target. As expected, not all the data types are important for predicting functional T-bet targets. Genomic features derived from Oct2, PU.1, CEBPα, and CEBPβ have low importance scores.

In general, functional T-bet targets have larger percentages of H3K4me3 methylation on its gene body (3.29% vs. 1.17%). Their minimum T-bet to Stat4 binding site distance is shorter (208 bps vs. 12915 bps). They have more Stat4 binding sites (4.89 vs. 1.35). They have larger Stat4 peak widths (2.2% vs. 0.6%). They have higher maximum Stat4 peak scores (543 vs. 137). Their minimum T-bet to H3K36me3 peak distance is shorter (11176 bps vs. 19086 bps). And their minimum T-bet to H3K27me3 peak distance is shorter (14457 bps vs. 25451 bps), but only under Th2 condition, and not under Th1, suggesting that T-bet bound genes that are repressed under Th2 condition tend to be functional Th1 T-bet targets. Taken together, functional and non-functional T-bet target genes exhibit distinct genomic signal profiles. Both Stat4 and H3K4me3 binding patterns provide important infor-

mation on T-helper cell differentiation. In order for T-bet to regulate its target genes Stat4 also need to be bound, suggesting a need for co-regulator(s) to be bound for regulation to occur. Histone methylation data provides a glimpse of the epigenetic environment correlating with functional regulation of T-bet targets.

4.2.4 Chromatin-modifying and N-terminal domain are necessary for T-bet to regulate its targets

In this section, we search for mechanisms *intrinsic* to the T-bet protein itself that regulates functional T-bet targets. Our strategy is to mutate T-bet protein domains to determine domains that are necessary for regulation. Wild type and T-bet mutant constructs are over-expressed in T-bet$^{-/-}$ CD4^{+} T cells under Th1-skewing conditions. The rationale is that wild type T-bet proteins would rescue the expression of T-bet targets, while T-bet mutants would impede the expression of T-bet targets — if these mutants have been mutated in such a way that they no longer contain the functional activities necessary to regulate T-bet targets.

From our previous studies of prototypic targets such as Ifng and Cxcr3, we have generated a series of T-bet mutants that can bind to T-bet targets, but are deficient at regulating T-bet targets because key residue(s) have been mutated or an entire terminal domain has been truncated [18, 146]. These constructs include: 1) Wild type T-bet [T-bet 1-530], 2) C-terminal truncation mutant, missing transactivating potential [T-bet 1-330], 3) N-terminal truncation mutant, missing transactivating potential [T-bet 120-530], 4) Chromatin-modifying mutant that abolishes its interactions with demethylase Jmjd3 and methyltransferase Set7/9 [T-bet Y182S] (Figure 4.4A). These four T-bet constructs were cloned into MSCV 2.2, a retroviral vector with a GFP tag separated by an Internal Ribosome Entry Site (IRES).

Subsequently, these constructs were transfected into Phoenix packaging cell line along with empty vector to generate retroviruses for over-expression studies. To ensure successful transductions, expressions of wild type and mutant T-bet proteins

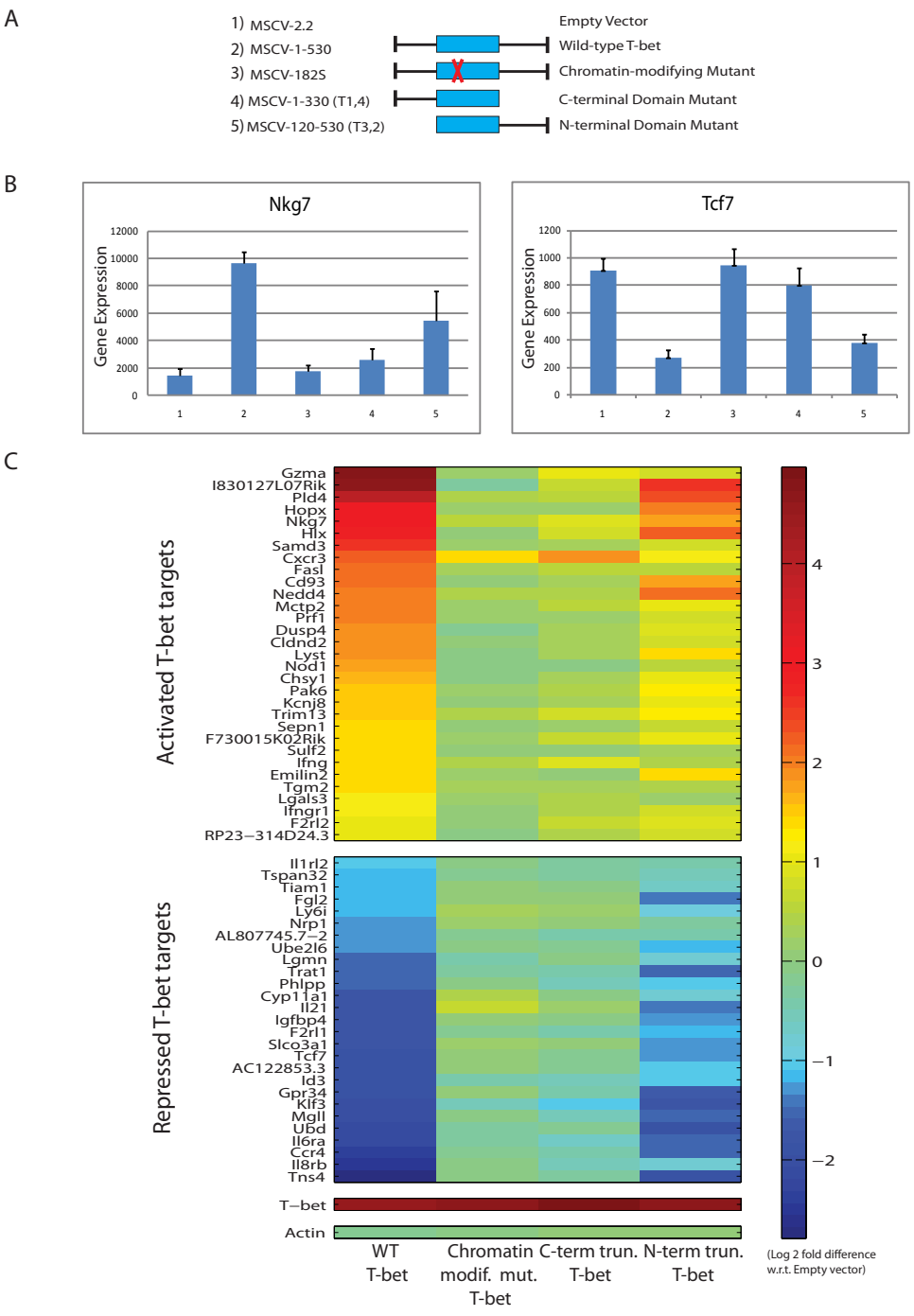

Figure 4.4: **T-bet requires both the chromatin-modifying and the c-terminal trans-activating activities to regulate its functional targets.** CD4+ T-cells from T-bet$^{-/-}$ were retrovirally transduced with T-bet mutants. (A)Schematic diagrams represent wild type and mutant T-bet proteins. T-box location (blue) and single nucleotide mutation ("X") are shown. (B)Gene expression of NKG-7 and Tcf7. Mean gene expression values from microarray experiments were shown (error bars, S.E.M.). (C) Differential gene expression comparisons between T-bet mutants and empty vector control. Activated T-bet targets (upper) and repressed targets (lower) are shown.

were verified by Western blots and only GFP+ cells were collected for our gene expression analyses. In addition, we also checked the protein expression of Ifng by flow cytometry to ensure proper expression of T-bet target. Figure 4.4B shows the effects of transducing these various constructs into T-bet$^{-/-}$ CD4^{+} T cells. NKG-7, an activated T-bet target, is minimally expressed when transduced with an empty vector, whereas over-expression of wild type T-bet activates NKG-7 expression. Both the C-terminal and chromatin-modifying mutants were unable to rescue the expression of NKG-7, while the N-terminal truncation mutant restored partial expression of NKG-7. Similarly, wild type T-bet also repressed the expression of T-bet targets. As before, both the C-terminal and chromatin-modifying mutants were unable to rescue the activity of the wild type T-bet, while the N-terminal truncation mutant contains partial rescue activity.

However, it is unclear if the requirement for chromatin-modifying and C-terminal transactivating potential persists for other functional T-bet targets as well; hence, gene expression microarray studies were performed. Figure 4.4C is a heat map showing the differential gene expression of the various constructs in comparison to empty vector. The rows represent activated or repressed functional T-bet targets with magnitudes larger than two fold. The columns represent the differential gene expressions of the various T-bet constructs. As an example, going across the columns, the gene Hopx shows activation by wild type T-bet, minimal activation by chromatin-modifying T-bet mutant and C-terminal truncation mutant, and partial activation by N-terminal truncation mutants as indicated by the colors red, green, and orange. The general trend is that the chromatin-modifying activity and the C-terminal transactivating activity are necessary for T-bet to regulate its target genes, whereas the requirement for the N-terminal domain is target dependent. This is the case for both activated and repressed functional T-bet targets. Taken together, T-bet's regulation of its targets requires both chromatin-modifying activity and C-terminal transactivating activity.

4.2.5 Functional T-bet targets are also differentially regulated in other T-helper subtypes

Thus far, we defined a set of functional T-bet targets and studied how they are regulated under Th1 skewing condition. Not surprisingly, T-bet targets are regulated under Th1 conditions. However, it is unclear if T-bet targets are also regulated under other T helper conditions. To answer this question, we extracted the expression of T-bet targets under Th1, Th2, Th17, iTreg, and nTreg conditions [149]. Shown in Figure 4.5A upper left panel, are the differential expression of functional T-bet targets under various T helper conditions compared to naive T cells. As a negative control, we also randomly sampled an equal number of genes from the same DNA microarrays (Figure 4.5A upper central panel). In addition, we performed a randomization test by repeatedly sampling sets of genes to estimate the statistical significance of our observation (Figure 4.5A upper right panel). Strikingly, functional T-bet targets are also differentially regulated under other T helper conditions. To observe a randomly selected set of genes that are as differentially expressed as the T-bet targets is an extremely unlikely event (P-value $<$ 2e-35). This highly suggests that functional T-bet targets are not unique to Th1. In addition, even in cellular conditions such as Th2 and Th17, where the expression of T-bet is repressed, T-bet target genes can still be activated or repressed, suggesting the existence of potential alternative regulators or combinatorial regulation.

To get a basis of comparison, we next analyzed the regulation of T-bet targets under a set of B cell lineage conditions. We extracted the expression of functional T-bet targets under naive B cells, plasma B cells, germinal center B cells, and memory B cells conditions [170]. Shown in Figure 4.5B are the differential expressions of functional T-bet targets compared to naive B cell (left panel), a sampling of genes under various B cell lineage conditions (center panel), and associated randomization test procedure (right panel). Significantly, the majority of functional T-bet targets are not differentially regulated under B cell conditions, suggesting that in order for

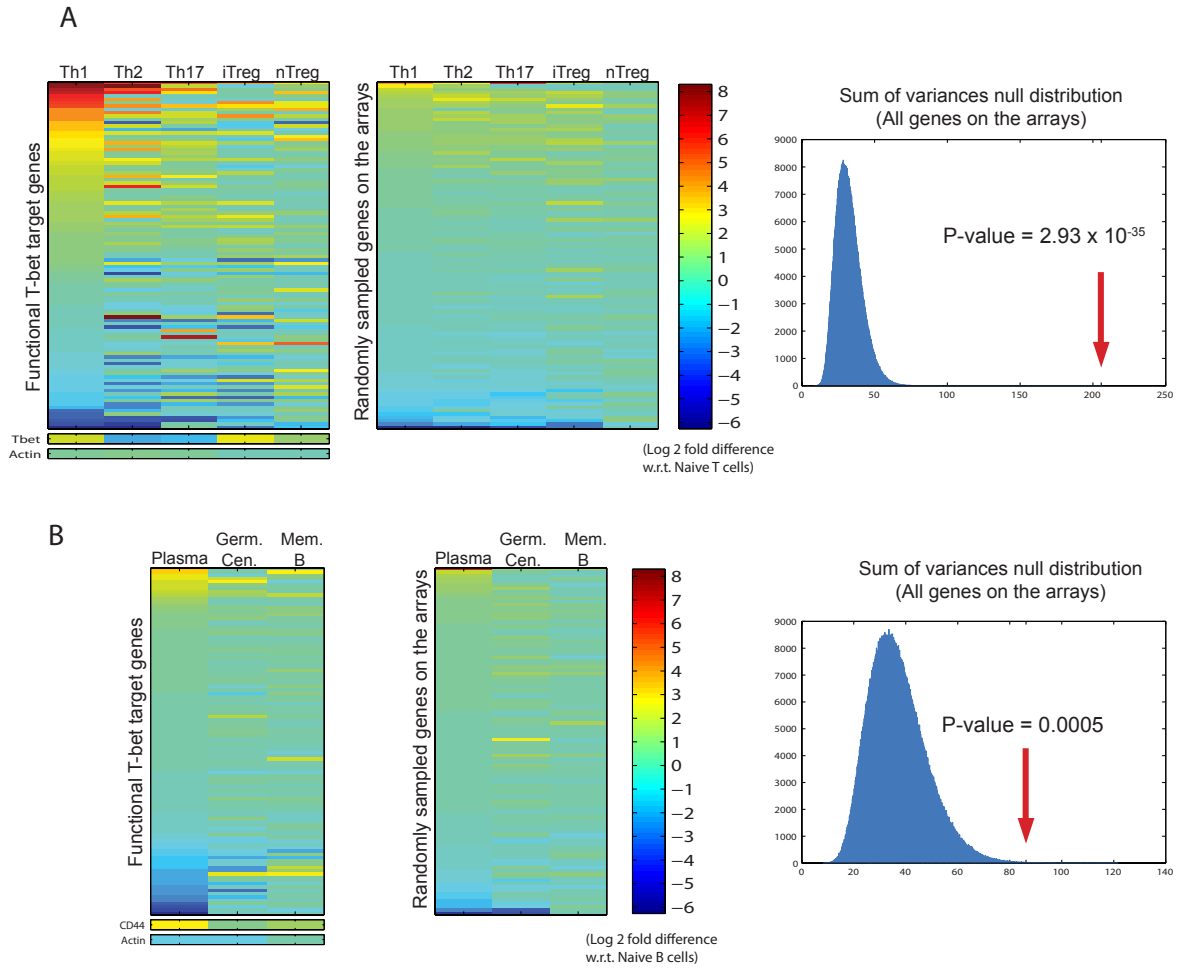

Figure 4.5: **Functional T-bet targets are also differentially regulated in other T-helper lineages.** (A)Microarray gene expression values for Th1, Th2, Th17, iTreg, and nTreg are compared to naive T cells (GSE14254). Functional T-bet targets (left), randomly selected genes (center), and statistical significance (right) are displayed. (n=100,000). (A)Microarray gene expression values for plasma B cells, germinal center B cells, and memory B cells are compared to naive B cells (GSE4142). Functional T-bet targets (left), randomly selected genes (center), and statistical significance (right) are displayed. (n=100,000).

T-bet targets to be regulated to give rise to the Th1 gene expression profile, both the cellular conditions, and the sets of targets genes involved are important. Interestingly, we also observed a small block of functional T-bet targets that are repressed in plasma B cell condition but not in germinal center B cells and memory B cells. We suspect some T-bet targets are actively repressed in naive B cell to plasma B cell transition.

4.3 Discussion

Th1 differentiation is central to the clearance of intracellular pathogens. T-bet, as the Th1 lineage-determining transcription factor, establishes the global gene expression profile that ultimately leads to the Th1 phenotype. In this study, we focused on the important question of how T-bet affects the necessary gene targets on a genome-wide level to bring forth this transition. We uncovered a set of "functional" T-bet targets by combining binding location and gene expression analyses. For T-bet to regulate these gene targets, it requires chromatin-modifying activity from the T-box and trans-activation activity from the C-terminus. Gene targets that are regulated by T-bet exhibit distinct genomic attributes distinct from those that are not. Specifically, binding pattern information from Stat4 and H3K4me3 marks are most informative. When extended beyond Th1 conditions, we found many of T-bet targets are regulated under other T helper conditions as well.

We observed an abundance of T-bet binding events across the genome, yet a relatively small percentage of T-bet bound genes are differentially expressed. This raises the intriguing question about the differences between factor-bound genes with functional consequences and those without. Stat4 binding events seem to be the most important factor for predicting functional expression of T-bet targets. This observation suggests a likely model where T-bet can bind to many locations in the genome, but for many functional targets, unless the "right" co-regulator (Stat4) is also present, transcription does not ensue. Similarly, we have previously reported that T-bet binds to target genes in multiple cell types, including B cells, CD4 T-cells,

CD8 T-cell, and NK cells, yet only a subset of T-bet bound genes exhibit functional consequences [161]. It is likely that even though T-bet bindings are detected in these cell types, T-bet does not regulate these targets unless the right co-regulating partners are also in place. This also potentially explains why we see differentially expressed T-bet targets under the various T-helper conditions much more frequently than under B cell cellular contexts. In a sense the cellular context could be thought of as the presence of a collection of T-bet co-regulators. Thus, it makes sense that under the various T helper cellular conditions, T-bet targets are more regulated than under B cell contexts, because under the B cell contexts, necessary co-regulators of T-bet are not present.

Our study paints a complex picture of regulations involving combinatorial control of transcription factors and proper setting of epigenetic states. The T-bet protein can bind to many locations in the mouse genome, where it modifies the local chromatin environment and transactivates the expression of target genes [18]. However, this regulation requires the presence of appropriate co-regulators. Subsequently, local genomic attributes of target genes are altered. T-bet target gene expression can be further modulated depending on the co-regulator(s). When T-bet is not present, other lineage-determining transcriptions likely also participate in regulating the expression of functional T-bet targets. Subsets of T-bet targets are similarly regulated under certain T helper conditions, giving rise to the complex combinatorial regulation of its target genes.

In this study, we combined binding and gene expression data types to define a set of functional T-bet targets. This approach can readily be applied to the study of other T helper lineages. If similar studies were performed for additional T helper lineages using Gata3, Foxp3, RORγt, and Bcl6, shared functional target genes between T helper lineages can be identified, and potential co-regulators teased out. Ultimately, a network-level understanding of transcriptional regulations that underlie T-helper cell differentiation would emerge. Further, this level of understanding is necessary to

modulate CD4$^+$ T cell responses that are specific to infectious pathogens, allergenic agents, and immunogenic self antigens. In summary, T helper cell differentiation is a complex process, and T-bet is but a member of this interacting web of co-regulating players that direct the expression of T helper cell targets.

4.4 Methods

4.4.1 Mice, T-cell Isolation, and Culture

CD4$^+$ T cells were isolated as previously described [161]. Briefly, spleen and lymph nodes were harvested from C57BL/6 wild type and T-bet$^{-/-}$ mice. Mouse CD4$^+$ T Cell Isolation Kit by RD Systems were used for the purification and verified by flow cytometry with over 95% purity. CD4$^+$ T Cells were activated with plate-bound αCD3(5 ug/mL) and αCD28(10 ug/mL) in the presence of α-IL4(10ug/mL) overnight in IMDM supplemented with 10% fetal bovine serum, 100IU/mL penicillin, 0.1 mg/mL streptomycin, and 2mM β-mercaptoethanol. They were then cultured under Th1 conditions with IL-2(10ng/mL) and IL-12(10ng/mL) for 3 days and expanded for another 3 days in fresh media under the same Th1 cytokine condition without αCD3 and αCD28 activation. All mouse studies were performed with the approval and by IACUC guide lines.

4.4.2 RNA Isolation and qRT-PCR

During the Th1 polarization process, CD4$^+$ T cells were harvested. Duplicates of biological samples from separate mice were collected on days 1, 2, 4, and quadruplicates on day 6. RNA was extracted by following the Qiagen RNA purification protocol (RNA Easy Mini Kit). For quantification, First Strand Superscript II Synthesis System (Invitrogen) was used to produce cDNA template. 2X qPCR Sybr Green mix (Thermo Scientific) with gene-specific primers (Table S1) amplified the templates. Actin level of each sample was used for normalization for final quantification

4.4.3 Microarray Analysis

Gene expression was first checked by traditional RT-PCR. 300 ng of total RNA were labeled. Fragmentation and labeling qualities of the cDNA were verified with Agilent Bioanalyzer prior to hybridization onto the GeneChIP Mouse Exon 1.0 ST Arrays. Gene expression values were extracted and normalized using justRMA algorithm available in R (Bioconductor package) and Log2-transformed. Gene expression profiling was performed in biological triplicates.

4.4.4 Chromatin Immunoprecipitation

Wild type and T-bet$^{-/-}$ CD4^{+} T cells were harvested 6 days post T-cell isolation. The ChIP protocol is as previously described [146, 161] with the exception of using Protein A coupled magnetic beads instead of the Pansorbin cells for pre-clearing and immunoprecipitation. Briefly, cells were cross-linked in 1% formaldehyde and the nuclei were isolated, sonicated, and then pre-cleared with Protein A (Dyna beads). Sonicated chromatins were incubated with specific and control antibodies overnight at 4° C. Chromatins were then immunoprecipitated, washed, and eluded. Cross-linking reversal was done at 65° C for 4 hours. The DNA was purified by proteinase K digestion followed by phenol/chloroform extraction and DNA precipitation. Antibodies for T-bet(Santa Cruz), H3K4me2(Millipore), and H3K4me3(Millipore) were used for the precipitation of chromatin.

4.4.5 High-Throughput Sequencing

DNA library preparation was performed by following the Illumina protocol. DNA fragment ends were repaired and ligated to illumina adaptors. The fragments were selected (300-350 bps) amplified, and sequenced on Illumina Genmoe Analyzer II. Eland pipeline was used to output sequence reads of 36 bps. Only uniquely mapped reads to mouse genome (mm9) were kept for further analysis downstream. Peak-calling was determined by MACS (Modeled-based Analysis for ChIP-Seq) at a P-value

cutoff of 10^{-5}..

4.4.6 Retroviral Transduction

Full length and mutant T-bet expression constructs were described previously [18, 146]. T-bet 1-530 (Full length), T-bet 1-331 (C-terminal truncation), T-bet 120-530(N-terminal truncation), and T-bet Y182S (domain 1 demethylase and methyl transferase mutant) were sub-cloned into the MSCV 2.2 expression vector. The MSCV2.2 vector was a gift from Dan Stetson containing GFP separated by an internal ribosomal entry site (IRES). To generate the virus, T-bet MSCV constructs were transfected into phoenix Eco packaging cell using calcium phosphate following protocols from G. Nolan. Viral soup was collected 36 hours post transfection and applied to Day 1 CD 4+ T-bet$^{-/-}$ cells and centrifuged in the presence of polybrene (5 mg/mL) at room temperature for 90 minutes under Th1 culturing condition. Viral soup was replaced with fresh Th1 culturing media on Day 2. Cells were expanded for 3 days and harvested on Day 6. GFP+ cells were sorted for gene expression analysis by qRT-PCR and microarray.

4.4.7 Transient Transfection

EL4 cells were cultured in RPMI supplemented with 10% fetal bovine serum, 100IU/mL penicillin, 0.1 mg/mL streptomycin. Transfections were performed using the AMAXA nucleofection system following manufacturer's protocol. 4 million cells were used for each transfection with program O-17 and solution V. Cells were harvested 18-22 hours post transfection.

4.4.8 Data Collection and Pre-processing

In addition to the T-bet, H3K4me2, and H3K27me3 ChIP-Seq datasets we generated a number of genome-wide datasets were collected and extracted as features for our classifier that distinguishes between functional and non-functional T-bet targets.

From GEO, we downloaded STAT4 and STAT6 binding sites, H3K4me3, H3K27me3, and H3K36me3 histone methylation sites (GSE22104) and 2) CEBPα, CEBPβ, Oct2, and PU.1 binding sites (GSE21512). From UCSC genome repository, we downloaded: 3) Th1 and Th2 DNase I hypersensitivity sites (UW DNase I HS) and 4) Sequence Conservation (Multiz 30-way PhastCon scores). A total of 23 sources of data are collected. When the short sequence reads are available, peaks were determined by MACS at P-value threshold of 10^{-5}.. Otherwise, we used the peaks provided by the investigators of the studies, after we visually verified the quality of their peak-calling outputs. For each of the 23 types of peak data sources, 5 features were extracted: A) Minimum distance to the closest T-bet site, B) Total width of the peaks (normalized to gene size), C) Maximum score of the peaks, D) Number of total peaks (normalized to gene size), and E) Maximum number of T-bet sites within a 500bp window. A total of 115 (23 x 5) features were constructed for predicting if a bound T-bet target gene is functional or non-functional.

4.4.9 Classifier Construction and Prediction

We constructed a Random Forest [29] classifier to help identify the structural patterns that distinguish between functional and non-functional T-bet targets. Functional targets are bound by T-bet and have differential gene expression between WT and T-bet KO, whereas non-functional targets are bound by T-bet but have a differential gene expression of less than 1.2 fold. Differential gene expression is determined by Significance Analysis of Microarrays (SAM) FDR < 0.03 [162]. To handle the imbalance of training examples between the functional targets and the non-functional targets, the classes were weighted inversely proportional to their class sizes. 30 variables were split at each node. 500 trees were grown in the forest for training and testing.

4.4.10 Motif Enrichment Analysis

To find de novo enriched motifs that are in close proximity to T-bet binding sites, we downloaded sequences that are 100 bps from the summit of the T-bet peaks from Ensembl [150] with repeats masked. T-bet peaks with the highest enrichment scores as reported by MACS were scanned by MEME [152]. Enriched motifs were submitted to TOMTOM [154] to identify the transcription factor associated with the enriched motifs. Sequences associated with up-regulated and down-regulated peaks were also analyzed by MEME and mapped back to transcription factor using TOMTOM. To identify potential discriminative transcription factor binding partners, 100 bp sequences in close proximity to T-bet peak summits of up-regulated and down-regulated T-bet targets were scanned with transcription factor matrices in TRANSFAC [153] using FIMO from MEME suite at p-value cutoff of 10^{-5}.

Chapter 5

CONCLUSION

Graduate training has been a journey, and I am fortunate to have had the rare opportunity to study a research question from two distinct perspectives: Most students get only one view. This endeavor stretches my thinking and forces me to question what is true, while reconciling differences between the perspectives. The question of cell fate determination — how seemingly similar neighboring cells ultimately differentiate into disparate cell fates, and maintain their differences — presents a fertile ground to explore diverse subjects such as developments of neutrophils and T helper cells; designing and conducting experiments; and constructing and interpreting computational models.

In chapter one, the question of cell fate determination is presented from both *systems dynamics* and *molecular mechanism* perspectives. Brief introductory material on neutrophil and T helper cell differentiation is provided. In chapter two, the discussion is shifted to experimental data types and ways of tying experimentation and computation together. Chapter three contains our study of neutrophil differentiation from a systems dynamics perspective. Chapter four includes a study of Th1 differentiation from a molecular mechanism angle.

For our systems dynamics study, the novelty is in designing experiments that exploit the key point that cell fates are attractors in high dimensional space — use theory to inform experimental design. This allows us to hone in on a class of genes that exhibit "divergent" gene expression profiles and are associated with neutrophil differentiation and cancer, elucidating molecular level details.

Though not directly addressed in our study, the systems dynamics perspective is grounded on a conceptually rich formal framework rooted in dynamical systems

theory [171–173]. Not only is this framework powerful for studying development, it has been applied to comprehend a network's emerging properties such as criticality [174], robustness [175], noise [12], heterogeneity of cell types [176], cancer [177,178] and stem cell behaviors [179]. Indeed, it is rare to find a unifying perspective that encompasses all these phenomena.

Coming from a molecular mechanism perspective, we focus on details of the molecular machinery that underlies differentiation. We begin by defining a set of genes that are functionally regulated by T-bet. Then, to hunt for extrinsic mechanisms regulating T-bet targets, we collect a number of data sets and use statistical machine learning techniques to find important features that signal Th1 differentiation; for intrinsic mechanism, we over-express mutant T-bet proteins and study these mutants' abilities to regulate T-bet targets. In the end, we found T-bet target genes to be regulated under other T helper conditions as well.

With the advancement of genome-wide technologies and the releases of consortium data such as those from ENCODE [165], high-throughput data has become more accessible and the usage of machine learning techniques to analyze biological data more popular. Data sets that include histone methylation, transcription factor binding, DNase-Seq, DNA methylation, RNA-Seq, and gene annotation make the problem of heterogeneous data integration a very active area of research, as investigators grapple to put all the information together to gain biological knowledge. And over time, the "molecular maps" that detail regulatory relationships between genes should improve.

Both perspectives offer valuable insights into the process of cell fate determination. Naturally, one is left to wonder if the two perspectives can be merged or combined. One exciting future direction could be a coupling between the two approaches. First, take advantage of the molecular mechanism approach: interrogate patterns within the abundant genomic data to construct detailed models of regulation (gene regulatory networks). Second, use these molecular maps as the basis for a systems dynamics approach to construct simulation models to track state changes of the systems. For

example, one might combine transcription factor binding data with time point gene expression data to construct a model of a gene regulatory network for a module of genes. The model can then be used to find attractor states for the system. Further, one can take additional steps to identify "lever" node(s) that could perturb the system to jump from one attractor state to another as possible intervention measures so as to alter the cellular state.

BIBLIOGRAPHY

[1] Waddington C: **Canalization of development and the inheritance of acquired characters**. *Nature* 1942, **150**:563–565.

[2] Castaigne S, Chomienne C, Daniel MT, Ballerini P, Berger R, Fenaux P, Degos L: **All-trans retinoic acid as a differentiation therapy for acute promyelocytic leukemia. I. Clinical results**. *Blood* 1990, **76**:1704–1709.

[3] Voltarelli JC, Couri CE, Stracieri AB, Oliveira MC, Moraes DA, Pieroni F, Coutinho M, Malmegrim KC, Foss-Freitas MC, Simoes BP, Foss MC, Squiers E, Burt RK: **Autologous nonmyeloablative hematopoietic stem cell transplantation in newly diagnosed type 1 diabetes mellitus**. *JAMA* 2007, **297**:1568–1576.

[4] Kim JH, Auerbach JM, Rodriguez-Gomez JA, Velasco I, Gavin D, Lumelsky N, Lee SH, Nguyen J, Sanchez-Pernaute R, Bankiewicz K, McKay R: **Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease**. *Nature* 2002, **418**:50–56.

[5] Vadivelu S, Platik MM, Choi L, Lacy ML, Shah AR, Qu Y, Holekamp TF, Becker D, Gottlieb DI, Gidday JM, McDonald JW: **Multi-germ layer lineage central nervous system repair: nerve and vascular cell generation by embryonic stem cells transplanted in the injured brain**. *J. Neurosurg.* 2005, **103**:124–135.

[6] McDonald JW, Liu XZ, Qu Y, Liu S, Mickey SK, Turetsky D, Gottlieb DI, Choi DW: **Transplanted embryonic stem cells survive, differentiate and promote recovery in injured rat spinal cord**. *Nat. Med.* 1999, **5**:1410–1412.

[7] Goodell MA, Jackson KA, Majka SM, Mi T, Wang H, Pocius J, Hartley CJ, Majesky MW, Entman ML, Michael LH, Hirschi KK: **Stem cell plasticity in muscle and bone marrow**. *Ann. N. Y. Acad. Sci.* 2001, **938**:208–218.

[8] Jackson KA, Majka SM, Wang H, Pocius J, Hartley CJ, Majesky MW, Entman ML, Michael LH, Hirschi KK, Goodell MA: **Regeneration of ischemic cardiac muscle and vascular endothelium by adult stem cells**. *J. Clin. Invest.* 2001, **107**:1395–1402.

[9] Kauffman SA: **Metabolic stability and epigenesis in randomly constructed genetic nets**. *Journal of Theoretical Biology* 1969, **22**(3):437–467.

[10] Kauffman SA: *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York 1993.

[11] Huang S, Eichler G, Bar-Yam Y, Ingber DE: **Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network**. *Physical Review Letters* 2005, **94**(12):128701.

[12] Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S: **Transcriptome-wide noise controls lineage choice in mammalian progenitor cells**. *Nature* 2008, **453**:544–547.

[13] Strahl BD, Allis CD: **The language of covalent histone modifications**. *Nature* 2000, **403**:41–45.

[14] Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells**. *Cell* 2006, **125**:315–326.

[15] Lee JH, Tate CM, You JS, Skalnik DG: **Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex**. *J. Biol. Chem.* 2007, **282**:13419–13428.

[16] Muller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, Miller EL, O'Connor MB, Kingston RE, Simon JA: **Histone methyltransferase activity of a Drosophila Polycomb group repressor complex**. *Cell* 2002, **111**:197–208.

[17] Nottke A, Colaiacovo MP, Shi Y: **Developmental roles of the histone lysine demethylases**. *Development* 2009, **136**:879–889.

[18] Miller SA, Huang AC, Miazgowicz MM, Brassil MM, Weinmann AS: **Coordinated but physically separable interaction with H3K27-demethylase and H3K4-methyltransferase activities are required for T-box protein-mediated activation of developmental gene expression**. *Genes & development* 2008, **22**(21):2980–93.

[19] Oestreich KJ, Huang AC, Weinmann AS: **The lineage-defining factors T-bet and Bcl-6 collaborate to regulate Th1 gene expression patterns**. *J. Exp. Med.* 2011, **208**:1001–1013.

[20] Djuretic IM, Levanon D, Negreanu V, Groner Y, Rao A, Ansel KM: **Transcription factors T-bet and Runx3 cooperate to activate Ifng and silence Il4 in T helper type 1 cells**. *Nature immunology* 2007, **8**(2):145–53.

[21] Naoe Y, Setoguchi R, Akiyama K, Muroi S, Kuroda M, Hatam F, Littman DR, Taniuchi I: **Repression of interleukin-4 in T helper type 1 cells by Runx/Cbf beta binding to the Il4 silencer**. *The Journal of experimental medicine* 2007, **204**(8):1749–55.

[22] Orkin SH, Zon LI: **Hematopoiesis: an evolving paradigm for stem cell biology**. *Cell* 2008, **132**:631–644.

[23] Borregaard N: **Neutrophils, from marrow to microbes**. *Immunity* 2010, **33**:657–670.

[24] Gallagher R, Collins S, Trujillo J, McCredie K, Ahearn M, Tsai S, Metzgar R, Aulakh G, Ting R, Ruscetti F, Gallo R: **Characterization of the continuous, differentiating myeloid cell line(HL-60) from a patient with acute promyelocytic leukemia**. *Blood* 1979, **54**(3):713–733.

[25] Murphy KM, Reiner SL: **The lineage decisions of helper T cells**. *Nature reviews. Immunology* 2002, **2**(12):933–44.

[26] Zhu J, Paul WE: **CD4 T cells: fates, functions, and faults**. *Blood* 2008, **112**(5):1557–69.

[27] Zhu J, Paul WE: **Heterogeneity and plasticity of T helper cells**. *Cell research* 2010, **20**:4–12.

[28] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome research* 2005, **15**(8):1034–50.

[29] Breiman L: **Random Forests**. *Machine Learning* 2001, **45**:5–32.

[30] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell*. Garland Science, 5 edition 2007.

[31] Gilbert S: *Developmental Biology*. Sinauer Associates Inc, 8 edition 2006.

[32] Di Cara A, Garg A, De Micheli G, Xenarios I, Mendoza L: **Dynamic simulation of regulatory networks using SQUAD**. *BMC Bioinformatics* 2007, **8**:462.

[33] Newman S, Christley S, Glimm T, Hentschel H, Kazmierczak B, Zhang Y, Zhu J, Alber M: **Multiscale models for vertebrate limb development**. *Curr. Top. Dev. Biol.* 2008, **81**:311–340.

[34] Takagi H, Kaneko K: **Dynamical systems basis of metamorphosis: diversity and plasticity of cellular states in reaction diffusion network**. *J. Theor. Biol.* 2005, **234**:173–186.

[35] Kaneko K, Yomo T: **Isologous diversification: a theory of cell differentiation**. *Bull. Math. Biol.* 1997, **59**:139–196.

[36] Qu K, Abi Haidar A, Fan J, Ensman L, Tuncay K, Jolly M, Ortoleva P: **Cancer onset and progression: a genome-wide, nonlinear dynamical systems perspective on onconetworks**. *J. Theor. Biol.* 2007, **246**:234–244.

[37] Mendoza L: **A network model for the control of the differentiation process in Th cells**. *BioSystems* 2006, **84**:101–114.

[38] Yu Z, Quinn P: **Dimethyl sulphoxide: a review of its applications in cell biology**. *Biosci. Rep.* 1994, **14**:259–281.

[39] Kulyk W, Hoffman L: **Ethanol exposure stimulates cartilage differentiation by embryonic limb mesenchyme cells**. *Exp. Cell Res.* 1996, **223**:290–300.

[40] Woodbury D, Schwarz E, Prockop D, Black I: **Adult rat and human bone marrow stromal cells differentiate into neurons**. *J. Neurosci. Res.* 2000, **61**:364–370.

[41] Kalf G, O'Connor A: **The effects of benzene and hydroquinone on myeloid differentiation of HL-60 promyelocytic leukemia cells**. *Leuk. Lymphoma* 1993, **11**:331–338.

[42] Constantinou A, Huberman E: **Genistein as an inducer of tumor cell differentiation: possible mechanisms of action**. *Proc. Soc. Exp. Biol. Med.* 1995, **208**:109–115.

[43] Huang S, Ingber D: **Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks**. *Exp. Cell Res.* 2000, **261**:91–103.

[44] Chen J, Mandel E, Thomson J, Wu Q, Callis T, Hammond S, Conlon F, Wang D: **The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation**. *Nat. Genet.* 2006, **38**:228–233.

[45] Garcia A, Vega M, Boettiger D: **Modulation of cell proliferation and differentiation through substrate-dependent changes in fibronectin conformation**. *Mol. Biol. Cell* 1999, **10**:785–798.

[46] Sell S: **Cellular origin of cancer: dedifferentiation or stem cell maturation arrest?** *Environ. Health Perspect.* 1993, **101 Suppl 5**:15–26.

[47] Huang S, Ingber D: **Cell tension, matrix mechanics, and cancer development**. *Cancer Cell* 2005, **8**:175–176.

[48] Sell S: **Stem cell origin of cancer and differentiation therapy**. *Crit. Rev. Oncol. Hematol.* 2004, **51**:1–28.

[49] Pollett J, Corsi K, Weiss K, Cooper G, Barry D, Gharaibeh B, Huard J: **Malignant transformation of multipotent muscle-derived cells by concurrent differentiation signals**. *Stem Cells* 2007, **25**:2302–2311.

[50] Shmulevich I, Dougherty E, Zhang W: **Gene perturbation and intervention in probabilistic Boolean networks**. *Bioinformatics* 2002, **18**:1319–1331.

[51] Datta A, Choudhary A, Bittner M, Dougherty E: **External control in Markovian genetic regulatory networks: the imperfect information case**. *Bioinformatics* 2004, **20**:924–930.

[52] Datta A, Choudhary A, Bittner M, Dougherty E: **Control Approaches for Probabilistic Gene Regulatory Networks**. *IEEE Signal Processing Magazine* 2007, **24**:54–63.

[53] Gilchrist M, Thorsson V, Li B, Rust A, Korb M, Roach J, Kennedy K, Hai T, Bolouri H, Aderem A: **Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4**. *Nature* 2006, **441**:173–178.

[54] Miller J, Oldham M, Geschwind D: **A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging**. *J. Neurosci.* 2008, **28**:1410–1420.

[55] Bonneau R, Facciotti M, Reiss D, Schmid A, Pan M, Kaur A, Thorsson V, Shannon P, Johnson M, Bare J, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang D, Diruggiero J, Johnson C, Hood L, Baliga

N: **A predictive model for transcriptional control of physiology in a free living cell**. *Cell* 2007, **131**:1354–1365.

[56] Ramsey S, Klemm S, Zak D, Kennedy K, Thorsson V, Li B, Gilchrist M, Gold E, Johnson C, Litvak V, Navarro G, Roach J, Rosenberger C, Rust A, Yudkovsky N, Aderem A, Shmulevich I: **Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics**. *PLoS Comput. Biol.* 2008, **4**:e1000021.

[57] Collins S, Ruscetti F, Gallagher R, Gallo R: **Terminal differentiation of human promyelocytic leukemia cells induced by dimethyl sulfoxide and other polar compounds**. *Proc. Natl. Acad. Sci. U.S.A.* 1978, **75**:2458–2462.

[58] Breitman T, Selonick S, Collins S: **Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid**. *Proc. Natl. Acad. Sci. U.S.A.* 1980, **77**:2936–2940.

[59] McCarthy D, San Miguel J, Freake H, Green P, Zola H, Catovsky D, Goldman J: **1,25-dihydroxyvitamin D3 inhibits proliferation of human promyelocytic leukaemia (HL60) cells and induces monocyte-macrophage differentiation in HL60 and normal human bone marrow cells**. *Leuk. Res.* 1983, **7**:51–55.

[60] Rovera G, Santoli D, Damsky C: **Human promyelocytic leukemia cells in culture differentiate into macrophage-like cells when treated with a phorbol diester**. *Proc. Natl. Acad. Sci. U.S.A.* 1979, **76**:2779–2783.

[61] Tomonaga M, Golde D, Gasson J: **Biosynthetic (recombinant) human granulocyte-macrophage colony-stimulating factor: effect on normal bone marrow and leukemia cell lines**. *Blood* 1986, **67**:31–36.

[62] Drayson MT, Michell R, Durham J, Brown G: **Cell proliferation and CD11b expression are controlled independently during HL60 cell differentiation initiated by 1,25 α-dihydroxyvitamin D_3 or all-trans-retinoic acid**. *Experimental Cell Research* 2001, **266**:126–134.

[63] Yen A, Reece S, Albright K: **Dependence of HL-60 myeloid cell differentiation on continuous and split retinoic acid exposures: precommitment memory associated with altered nuclear structure**. *J. Cell. Physiol.* 1984, **118**:277–286.

[64] Chang H, Oh P, Ingber D, Huang S: **Multistable and multistep dynamics in neutrophil differentiation**. *BMC Cell Biol.* 2006, **7**:11.

[65] Yen A, Reece S, Albright K: **Membrane origin for a signal eliciting a program of cell differentiation**. *Exp. Cell Res.* 1984, **152**:493–499.

[66] Yen A, Reece S, Albright K: **Control of cell differentiation during proliferation. II. Myeloid differentiation and cell cycle arrest of HL-60 promyelocytes preceded by nuclear structural changes**. *Leuk. Res.* 1985, **9**:51–71.

[67] Heuze M, Guibal F, Banks C, Conaway J, Conaway R, Cayre Y, Benecke A, Lutz P: **ASB2 is an Elongin BC-interacting protein that can assemble with Cullin 5 and Rbx1 to reconstitute an E3 ubiquitin ligase complex**. *J. Biol. Chem.* 2005, **280**:5468–5474.

[68] Gunnarsson P, Levander L, Pahlsson P, Grenegard M: **The acute-phase protein alpha 1-acid glycoprotein (AGP) induces rises in cytosolic Ca2+ in neutrophil granulocytes via sialic acid binding immunoglobulin-like lectins (siglecs)**. *FASEB J.* 2007, **21**:4059–4069.

[69] Catusse J, Liotard A, Loillier B, Pruneau D, Paquet J: **Characterization of the molecular interactions of interleukin-8 (CXCL8), growth related oncogen alpha (CXCL1) and a non-peptide antagonist (SB 225002) with the human CXCR2**. *Biochem. Pharmacol.* 2003, **65**:813–821.

[70] Bernhagen J, Krohn R, Lue H, Gregory J, Zernecke A, Koenen R, Dewor M, Georgiev I, Schober A, Leng L, Kooistra T, Fingerle-Rowson G, Ghezzi P, Kleemann R, McColl S, Bucala R, Hickey M, Weber C: **MIF is a noncognate ligand of CXC chemokine receptors in inflammatory and atherogenic cell recruitment**. *Nat. Med.* 2007, **13**:587–596.

[71] Sendo D, Takeda Y, Ishikawa H, Sendo F, Araki Y: **Localization of GPI-80, a beta2-integrin-associated glycosylphosphatidyl-inositol anchored protein, on strongly CD14-positive human monocytes**. *Immunobiology* 2003, **207**:217–221.

[72] Yoshitake H, Takeda Y, Nitto T, Sendo F, Araki Y: **GPI-80, a beta2 integrin associated glycosylphosphatidylinositol-anchored protein, concentrates on pseudopodia without association with beta2 integrin during neutrophil migration**. *Immunobiology* 2003, **208**:391–399.

[73] Ting J, Willingham S, Bergstralh D: **NLRs at the intersection of cell death and immunity**. *Nat. Rev. Immunol.* 2008, **8**:372–379.

[74] Hartwell L, Culotti J, Reid B: **Genetic control of the cell-division cycle in yeast. I. Detection of mutants**. *Proc. Natl. Acad. Sci. U.S.A.* 1970, **66**:352–359.

[75] Ewen M, Sluss H, Sherr C, Matsushime H, Kato J, Livingston D: **Functional interactions of the retinoblastoma protein with mammalian D-type cyclins**. *Cell* 1993, **73**:487–497.

[76] Inaba T, Matsushime H, Valentine M, Roussel M, Sherr C, Look A: **Genomic organization, chromosomal localization, and independent expression of human cyclin D genes**. *Genomics* 1992, **13**:565–574.

[77] Mazumder S, DuPree E, Almasan A: **A dual role of cyclin E in cell proliferation and apoptosis may provide a target for cancer therapy**. *Curr Cancer Drug Targets* 2004, **4**:65–75.

[78] Draetta G, Beach D: **The mammalian cdc2 protein kinase: mechanisms of regulation during the cell cycle**. *J. Cell Sci. Suppl.* 1989, **12**:21–27.

[79] Tenca P, Brotherton D, Montagnoli A, Rainoldi S, Albanese C, Santocanale C: **Cdc7 is an active kinase in human cancer cells undergoing replication stress**. *J. Biol. Chem.* 2007, **282**:208–215.

[80] Wolthuis R, Clay-Farrace L, van Zon W, Yekezare M, Koop L, Ogink J, Medema R, Pines J: **Cdc20 and Cks direct the spindle checkpoint-independent destruction of cyclin A**. *Mol. Cell* 2008, **30**:290–302.

[81] Diaz-Martinez L, Gimenez-Abian J, Clarke D: **Regulation of centromeric cohesion by sororin independently of the APC/C**. *Cell Cycle* 2007, **6**:714–724.

[82] Stucke V, Sillje H, Arnaud L, Nigg E: **Human Mps1 kinase is required for the spindle assembly checkpoint but not for centrosome duplication**. *EMBO J.* 2002, **21**:1723–1732.

[83] Fisk H, Mattison C, Winey M: **Human Mps1 protein kinase is required for centrosome duplication and normal mitotic progression**. *Proc. Natl. Acad. Sci. U.S.A.* 2003, **100**:14875–14880.

[84] Hill E, Clarke M, Barr F: **The Rab6-binding kinesin, Rab6-KIFL, is required for cytokinesis**. *EMBO J.* 2000, **19**:5711–5719.

[85] Sumara I, Gimenez-Abian J, Gerlich D, Hirota T, Kraft C, de la Torre C, Ellenberg J, Peters J: **Roles of polo-like kinase 1 in the assembly of functional mitotic spindles**. *Curr. Biol.* 2004, **14**:1712–1722.

[86] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat. Genet.* 2000, **25**:25–29.

[87] Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data**. *Genome Biology* 2003, **4**(4):R28.

[88] Pabst T, Mueller B, Harakawa N, Schoch C, Haferlach T, Behre G, Hiddemann W, Zhang D, Tenen D: **AML1-ETO downregulates the granulocytic differentiation factor C/EBPalpha in t(8;21) myeloid leukemia**. *Nat. Med.* 2001, **7**:444–451.

[89] Kirstetter P, Schuster M, Bereshchenko O, Moore S, Dvinge H, Kurz E, Theilgaard-Monch K, Mansson R, Pedersen T, Pabst T, Schrock E, Porse B, Jacobsen S, Bertone P, Tenen D, Nerlov C: **Modeling of C/EBPalpha mutant acute myeloid leukemia reveals a common expression signature of committed myeloid leukemia-initiating cells**. *Cancer Cell* 2008, **13**:299–310.

[90] Castilla L: **C/EBPalpha in leukemogenesis: a matter of being in the right place with the right signals**. *Cancer Cell* 2008, **13**:289–291.

[91] Pabst T, Mueller B, Zhang P, Radomska H, Narravula S, Schnittger S, Behre G, Hiddemann W, Tenen D: **Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia**. *Nat. Genet.* 2001, **27**:263–270.

[92] Devaraj P, Foroni L, Sekhar M, Butler T, Wright F, Mehta A, Samson D, Prentice H, Hoffbrand A, Secker-Walker L: **E2A/HLF fusion cDNAs and the use of RT-PCR for the detection of minimal residual disease in t(17;19)(q22;p13) acute lymphoblastic leukemia**. *Leukemia* 1994, **8**:1131–1138.

[93] LeBrun D: **E2A basic helix-loop-helix transcription factors in human leukemia**. *Front. Biosci.* 2003, **8**:s206–222.

[94] Inukai T, Hirose K, Inaba T, Kurosawa H, Hama A, Inada H, Chin M, Nagatoshi Y, Ohtsuka Y, Oda M, Goto H, Endo M, Morimoto A, Imaizumi M, Kawamura N, Miyajima Y, Ohtake M, Miyaji R, Saito M, Tawa A, Yanai F, Goi K, Nakazawa S, Sugita K: **Hypercalcemia in childhood acute lymphoblastic leukemia: frequent implication of parathyroid hormone-related peptide and E2A-HLF from translocation 17;19**. *Leukemia* 2007, **21**:288–296.

[95] Begley C, Green A: **The SCL gene: from case report to critical hematopoietic regulator**. *Blood* 1999, **93**:2760–2770.

[96] Hall M, Curtis D, Metcalf D, Elefanty A, Sourris K, Robb L, Gothert J, Jane S, Begley C: **The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12**. *Proc. Natl. Acad. Sci. U.S.A.* 2003, **100**:992–997.

[97] Abramovich C, Humphries R: **Hox regulation of normal and leukemic hematopoietic stem cells**. *Curr. Opin. Hematol.* 2005, **12**:210–216.

[98] Strathdee G, Sim A, Parker A, Oscier D, Brown R: **Promoter hypermethylation silences expression of the HoxA4 gene and correlates with IgVh mutational status in CLL**. *Leukemia* 2006, **20**:1326–1329.

[99] Okuda T, van Deursen J, Hiebert S, Grosveld G, Downing J: **AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis**. *Cell* 1996, **84**:321–330.

[100] Sasaki K, Yagi H, Bronson R, Tominaga K, Matsunashi T, Deguchi K, Tani Y, Kishimoto T, Komori T: **Absence of fetal liver hematopoiesis in mice deficient in transcriptional coactivator core binding factor beta**. *Proc. Natl. Acad. Sci. U.S.A.* 1996, **93**:12359–12363.

[101] Michaud J, Wu F, Osato M, Cottles G, Yanagida M, Asou N, Shigesada K, Ito Y, Benson K, Raskind W, Rossier C, Antonarakis S, Israels S, McNicol A, Weiss H, Horwitz M, Scott H: **In vitro analyses of known and novel RUNX1/AML1 mutations in dominant familial platelet disorder with predisposition to acute myelogenous leukemia: implications for mechanisms of pathogenesis**. *Blood* 2002, **99**:1364–1372.

[102] Calvo K, Knoepfler P, Sykes D, Pasillas M, Kamps M: **Meis1a suppresses differentiation by G-CSF and promotes proliferation by SCF: potential**

mechanisms of cooperativity with Hoxa9 in myeloid leukemia. *Proc. Natl. Acad. Sci. U.S.A.* 2001, **98**:13120–13125.

[103] Thorsteinsdottir U, Kroon E, Jerome L, Blasi F, Sauvageau G: **Defining roles for HOX and MEIS1 genes in induction of acute myeloid leukemia**. *Mol. Cell. Biol.* 2001, **21**:224–234.

[104] Lasa A, Carnicer M, Aventin A, Estivill C, Brunet S, Sierra J, Nomdedeu J: **MEIS 1 expression is downregulated through promoter hypermethylation in AML1-ETO acute myeloid leukemias**. *Leukemia* 2004, **18**:1231–1237.

[105] Melnick A, Licht J: **Deconstructing a disease: RARalpha, its fusion partners, and their roles in the pathogenesis of acute promyelocytic leukemia**. *Blood* 1999, **93**:3167–3215.

[106] Chim C, Wong S, Pang A, Chu P, Lau J, Wong K, Kwong Y: **Aberrant promoter methylation of the retinoic acid receptor alpha gene in acute promyelocytic leukemia**. *Leukemia* 2005, **19**:2241–2246.

[107] Borkhardt A, Repp R, Haas OA, Leis T, Harbott J, Kreuder J, Hammermann J, Henn T, Lampert F: **Cloning and characterization of AFX, the gene that fuses to MLL in acute leukemias with a t(X;11)(q13;q23)**. *Oncogene* 1997, **14**:195–202.

[108] Magne S, Caron S, Charon M, Rouyez MC, Dusanter-Fourt I: **STAT5 and Oct-1 Form a Stable Complex That Modulates Cyclin D1 Expression**. *Molecular and Cellular Biology* 2003, **23**(24):8934–8945.

[109] Guimont P, Grondin F, Dubois C: **Sox9-dependent transcriptional regulation of the proprotein convertase furin.** *American journal of physiology. Cell physiology* 2007, **293**:C172–183.

[110] Afonja O, Raaka B, Huang A, Das S, Zhao X, Helmer E, Juste D, Samuels H: **RAR agonists stimulate SOX9 gene expression in breast cancer cell lines: evidence for a role in retinoid-mediated growth inhibition**. *Oncogene* 2002, **21**:7850–7860.

[111] Michl P, Knobel B, Downward J: **CUTL1 is phosphorylated by protein kinase A, modulating its effects on cell proliferation and motility**. *J. Biol. Chem.* 2006, **281**:15138–15144.

[112] Hart S, Foroni L: **Core binding factor genes and human leukemia.** *Haematologica* 2002, **87**(12):1307–1323.

[113] Moehren U, Papaioannou M, Reeb C, Hong W, Baniahmad A: **Alien interacts with the human androgen receptor and inhibits prostate cancer cell growth.** *Molecular Endocrinology* 2007, **21**(5):1039–1048.

[114] Gooch J, Christy B, Yee D: **STAT6 mediates interleukin-4 growth inhibition in human breast cancer cells.** *Neoplasia* 2002, **4**(4):324–31.

[115] Skokowa J, Cario G, Uenalan M, Schambach A, Germeshausen M, Battmer K, Zeidler C, Lehmann U, Eder M, Baum C, Grosschedl R, Stanulla M, Scherr M, Welte K: **LEF-1 is crucial for neutrophil granulocytopoiesis and its expression is severely reduced in congenital neutropenia**. *Nat. Med.* 2006, **12**:1191–1197.

[116] Albanese C, Wu K, D'Amico M, Jarrett C, Joyce D, Hughes J, Hulit J, Sakamaki T, Fu M, Ben-Ze'ev A, Bromberg J, Lamberti C, U UV, Gaynor R, Byers S, Pestell R: **IKKalpha regulates mitogenic signaling through transcriptional induction of cyclin D1 via Tcf.** *Molecular biology of the cell* 2003, **14**(2):585–99.

[117] Zada A, Singh S, Reddy V, Elsasser A, Meisel A, Haferlach T, Tenen D, Hiddemann W, Behre G: **Downregulation of c-Jun expression and cell cycle regulatory molecules in acute myeloid leukemia cells upon CD44 ligation**. *Oncogene* 2003, **22**:2296–2308.

[118] Brown L, Cheng JT, Chen Q, Siciliano MJ, Christ W, Buchanan G, Baer R: **Site-specific recombination of the tal-1 gene is a common occurence in human T cell leukemia**. *EMBO J.* 1990, **9**:3343–3351.

[119] Andrieux J, Demory J, Dupriez B, Quief S, Plantier I, Roumier C, Bauters F, Lai J, Kerckaert J: **Dysregulation and overexpression of HMGA2 in myelofibrosis with myeloid metaplasia.** *Genes Chromosomes Cancer* 2004, **39**:82–87.

[120] Chae J, Kim C: **CP2 binding to the promoter is essential for the enhanced transcription of globin genes in erythroid cells**. *Mol. Cells* 2003, **15**:40–47.

[121] Rowlands T, Baumann P, Jackson SP: **The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria**. *Science* 1994, **264**:1326–1329.

[122] Jiang W, Douglas-Jones A, Mansel R: **Expression of peroxisome-proliferator activated receptor-gamma (PPARgamma) and the PPARgamma co-activator, PGC-1, in human breast cancer correlates with clinical outcomes.** *International journal of cancer* 2003, **106**(5):752–757.

[123] Mudryj M, Devoto S, Hiebert SW, T TH, Pines J, Nevins JR: **Cell cycle regulation of the E2F transcription factor involves an interaction with cyclin A**. *Cell* 1991, **65**:1243–1253.

[124] Smith C, Tuan R: **Human PAX gene expression and development of the vertebral column.** *Clinical orthopaedics and related research* 1994, **302**:241–250.

[125] Narayanan R, Edwards D, Weigel N: **Human progesterone receptor displays cell cycle-dependent changes in transcriptional activity.** *Molecular and Cellular Biology* 2005, **25**(8):2885–2898.

[126] Schwitzgebel V, Mamin A, Brun T, Ritz-Laser B, Zaiko M, Maret A, Jornayvaz F, Theintz G, Michielin O, Melloul D, Philippe J: **Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1.** *The Journal of clinical endocrinology and metabolism* 2003, **88**(9):4398–4406.

[127] Keehn C, Smoller B, Morgan M: **Expression of the ets-1 proto-oncogene in melanocytic lesions.** *Modern pathology* 2003, **16**(8):772–777.

[128] Hecht A, Vleminckx K, Stemmler M, van Roy F, Kemler R: **The p300/CBP acetyltransferases function as transcriptional coactivators of beta-catenin in vertebrates**. *EMBO Journal* 2000, **19**:1839–1850.

[129] Chang H, Hemberg M, Barahona M, Ingber D, Huang S: **Transcriptome-wide noise controls lineage choice in mammalian progenitor cells**. *Nature* 2008, **453**:544–547.

[130] Bolstad BM, Irizarry RA, Åstrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.

[131] Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of Nationall Acadamy of Sciences U S A.* 2001, **98**(9):5116–5121.

[132] Hubbard TJP, Aken BL, Beal1 K, Ballester1 B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Overduin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007**. *Nucleic Acids Research* 2007, **Database**:D1–D8.

[133] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Research* 2002, **12**(6):996–1006.

[134] Wingender E, Dietze D, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites**. *Nucleic Acids Research* 1996, **24**:238–241.

[135] Berman B, Nibu Y, Pfeiffer B, Tomancak P, Celniker S, Levine M, Rubin G, Eisen M: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(2):757–62.

[136] Zhu J, Paul WE: **CD4+ T cell plasticity-Th2 cells join the crowd**. *Immunity* 2010, **32**:11–3.

[137] Mosmann TR, Cherwinski H, Bond MW, Giedlin MA, Coffman RL: **Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins**. *Journal of immunology* 1986, **136**(7):2348–57.

[138] Ansel KM, Djuretic I, Tanasa B, Rao A: **Regulation of Th2 differentiation and Il4 locus accessibility**. *Annual review of immunology* 2006, **24**:607–56.

[139] Miossec P, Korn T, Kuchroo VK: **Interleukin-17 and type 17 helper T cells**. *The New England journal of medicine* 2009, **361**(9):888–98.

[140] Fontenot JD, Gavin MA, Rudensky AY: **Foxp3 programs the development and function of CD4+CD25+ regulatory T cells**. *Nature immunology* 2003, **4**(4):330–6.

[141] Nurieva RI, Chung Y, Martinez GJ, Yang XO, Tanaka S, Matskevitch TD, Wang YH, Dong C: **Bcl6 mediates the development of T follicular helper cells**. *Science* 2009, **325**(5943):1001–5.

[142] Szabo SJ, Kim ST, Costa GL, Zhang X, Fathman CG, Glimcher LH: **A novel transcription factor, T-bet, directs Th1 lineage commitment**. *Cell* 2000, **100**(6):655–69.

[143] Chang S, Aune TM: **Histone hyperacetylated domains across the Ifng gene region in natural killer cells and T cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(47):17095–100.

[144] Shnyreva M, Weaver WM, Blanchette M, Taylor SL, Tompa M, Fitzpatrick DR, Wilson CB: **Evolutionarily conserved sequence elements that positively regulate IFN-gamma expression in T cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(34):12622–7.

[145] Jenner RG, Townsend MJ, Jackson I, Sun K, Bouwman RD, Young RA, Glimcher LH, Lord GM: **The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes**. *Proc. Natl. Acad. Sci. U.S.A.* 2009, **106**:17876–17881.

[146] Lewis MD, Miller SA, Miazgowicz MM, Beima KM, Weinmann AS: **T-bet's ability to regulate individual target genes requires the conserved T-box domain to recruit histone methyltransferase activity and a separate family member-specific transactivation domain**. *Molecular and cellular biology* 2007, **27**(24):8510–21.

[147] Miller SA, Mohn SE, Weinmann AS: **Jmjd3 and UTX play a demethylase-independent role in chromatin remodeling to regulate T-box family member-dependent gene expression**. *Molecular cell* 2010, **40**(4):594–605.

[148] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol.* 2008, **9**:R137.

[149] Wei L, Vahedi G, Sun HW, Watford WT, Takatori H, Ramos HL, Takahashi H, Liang J, Gutierrez-Cruz G, Zang C, Peng W, O'Shea JJ, Kanno Y: **Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation**. *Immunity* 2010, **32**(6):840–51.

[150] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM: **Ensembl 2011**. *Nucleic acids research* 2011, **39**(Database issue):D800–6.

[151] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nature methods* 2007, **4**(8):651–7.

[152] Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 1994, **2**:28–36.

[153] Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic acids research* 2003, **31**:374–8.

[154] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs**. *Genome biology* 2007, **8**(2):R24.

[155] Collins A, Littman DR, Taniuchi I: **RUNX proteins in transcription factor networks that regulate T-cell lineage choice**. *Nature reviews. Immunology* 2009, **9**(2):106–15.

[156] Egawa T: **Runx and ThPOK: a balancing act to regulate thymocyte lineage commitment**. *Journal of cellular biochemistry* 2009, **107**(6):1037–45.

[157] Zhang F, Meng G, Strober W: **Interactions among the transcription factors Runx1, RORgammat and Foxp3 regulate the differentiation of interleukin 17-producing T cells**. *Nature immunology* 2008, **9**(11):1297–306.

[158] Sakaguchi S, Yamaguchi T, Nomura T, Ono M: **Regulatory T cells and immune tolerance**. *Cell* 2008, **133**(5):775–87.

[159] Moisan J, Grenningloh R, Bettelli E, Oukka M, Ho IC: **Ets-1 is a negative regulator of Th17 differentiation**. *The Journal of experimental medicine* 2007, **204**(12):2825–35.

[160] Chen GY, Osada H, Santamaria-Babi LF, Kannagi R: **Interaction of GATA-3/T-bet transcription factors regulates expression of sialyl Lewis X homing receptors on Th1/Th2 lymphocytes**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(45):16894–9.

[161] Beima KM, Miazgowicz MM, Lewis MD, Yan PS, Huang TH, Weinmann AS: **T-bet binding to newly identified target gene promoters is cell type-independent but results in variable context-dependent functional effects**. *The Journal of biological chemistry* 2006, **281**(17):11992–2000.

[162] Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(9):5116–21.

[163] Good SR, Thieu VT, 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**. *Journal of immunology* 2009, **183**(6):3839–47.

[164] Lazarevic V, Chen X, Shim JH, Hwang ES, Jang E, Bolm AN, Oukka M, Kuchroo VK, Glimcher LH: **T-bet represses T(H)17 differentiation by preventing Runx1-mediated activation of the gene encoding RORgammat**. *Nature immunology* 2011, **12**:96–104.

[165] Thomas DJ, Rosenbloom KR, Clawson H, Hinrichs AS, Trumbower H, Raney BJ, Karolchik D, Barber GP, Harte RA, Hillman-Jackson J, Kuhn RM, Rhead BL, Smith KE, Thakkapallayil A, Zweig AS, Haussler D, Kent WJ: **The ENCODE Project at UC Santa Cruz**. *Nucleic acids research* 2007, **35**(Database issue):D663–7.

[166] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Molecular cell* 2010, **38**(4):576–89.

[167] Hsieh CS, Macatonia SE, Tripp CS, Wolf SF, O'Garra A, Murphy KM: **Development of TH1 CD4+ T cells through IL-12 produced by Listeria-induced macrophages**. *Science* 1993, **260**(5107):547–9.

[168] Kaplan MH: **STAT4: a critical regulator of inflammation in vivo**. *Immunologic research* 2005, **31**(3):231–42.

[169] Thieu VT, Yu Q, Chang HC, Yeh N, Nguyen ET, Sehra S, Kaplan MH: **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):679–90.

[170] Luckey CJ, Bhattacharya D, Goldrath AW, Weissman IL, Benoist C, Mathis D: **Memory T and memory B cells share a transcriptional program of self-renewal with long-term hematopoietic stem cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(9):3304–9.

[171] Zhou JX, Huang S: **Understanding gene circuits at cell-fate branch points for rational cell reprogramming**. *Trends Genet.* 2011, **27**:55–62.

[172] Foster DV, Foster JG, Huang S, Kauffman SA: **A model of sequential branching in hierarchical cell fate determination**. *J. Theor. Biol.* 2009, **260**:589–597.

[173] Huang S: **Cell lineage determination in state space: a systems view brings flexibility to dogmatic canonical rules**. *PLoS Biol.* 2010, **8**:e1000380.

[174] Nykter M, Price ND, Aldana M, Ramsey SA, Kauffman SA, Hood LE, Yli-Harja O, Shmulevich I: **Gene expression dynamics in the macrophage exhibit criticality**. *Proc. Natl. Acad. Sci. U.S.A.* 2008, **105**:1897–1900.

[175] Huang S: **Reprogramming cell fates: reconciling rarity with robustness**. *Bioessays* 2009, **31**:546–560.

[176] Huang S: **Non-genetic heterogeneity of cells in development: more than just noise**. *Development* 2009, **136**:3853–3862.

[177] Huang S: **On the intrinsic inevitability of cancer: From foetal to fatal attraction**. *Semin. Cancer Biol.* 2011, **21**:183–199.

[178] Huang S, Ernberg I, Kauffman S: **Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective**. *Semin. Cell Dev. Biol.* 2009, **20**:869–876.

[179] Huang S: **Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways**. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 2011, **366**:2247–2259.