University of North Dakota
UND Scholarly Commons

Theses and Dissertations | Theses, Dissertations, and Senior Projects

January 2022

The Architecture And Dynamics Of Gene Regulatory Networks Directing Cell-Fate Choice During Murine Hematopoiesis

Joanna Handzlik

Follow this and additional works at: https://commons.und.edu/theses

Recommended Citation

Handzlik, Joanna, "The Architecture And Dynamics Of Gene Regulatory Networks Directing Cell-Fate Choice During Murine Hematopoiesis" (2022). *Theses and Dissertations*. 4262.
https://commons.und.edu/theses/4262

This Dissertation is brought to you for free and open access by the Theses, Dissertations, and Senior Projects at UND Scholarly Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UND Scholarly Commons. For more information, please contact und.commons@library.und.edu.

THE ARCHITECTURE AND DYNAMICS OF GENE REGULATORY NETWORKS DIRECTING CELL-FATE CHOICE DURING MURINE HEMATOPOIESIS

by

Joanna Elżbieta Handzlik
Bachelor of Science, National and Kapodistrian University of Athens, 2012
Master of Science, National and Kapodistrian University of Athens, 2017

A Dissertation

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Grand Forks, North Dakota

May
2022

Copyright 2022 Joanna E. Handzlik

DocuSign Envelope ID: 6E0D068E-48DD-4D7F-826D-974FE3365968

Name: Joanna E. Handzlik

Degree: Doctor of Philosophy

This document, submitted in partial fulfillment of the requirements for the degree from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

DocuSigned by:
Manu
E0D49834CCAC470...
Dr. Manu

DocuSigned by:
Diane Darland
09715EE0C123422...
Dr. Diane Darland

DocuSigned by:
Dr. Junguk Hur
1C49E6598068480...
Dr. Junguk Hur

DocuSigned by:
760EF86F7CCE4FF...
Dr. Yen Lee Loh

DocuSigned by:
Turk Rhen
32ECCCC21FF5488...
Dr. Turk Rhen

This document is being submitted by the appointed advisory committee as having met all the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

DocuSigned by:
Chris Nelson
2E0AF088C733403...
Chris Nelson
Dean of the School of Graduate Studies

5/2/2022
Date

PERMISSION

Title	The Architecture and Dynamics of Gene Regulatory Networks Directing Cell-Fate Choice During Murine Hematopoiesis
Department	Biology
Degree	Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Joanna E. Handzlik
5/1/2022

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The completion of my doctoral studies was possible thanks to many great people who helped me through this long journey and deserve a hefty load of acknowledgments.

First and foremost, I thank my adviser Manu for guiding me through this endeavor with much care and unwavering patience. Manu's expertise and depth of knowledge in so many different fields has been for me truly inspirational and motivational throughout these years.

I thank my committee members, Diane Darland, Junguk Hur, Yen Lee Loh, and Turk Rhen for their time, support, and expertise. Yen Lee and Manu, thank you for the Friday meetings where we brainstormed about the why's and how's of biology and the universe.

I thank all the labs participating in the Epigenetics Group at the UND School of Medicine and Health Sciences for informative meetings and great discussions over the years.

I thank all the current and past members of Manu Lab for constructive lab meetings and fruitful discussions. I thank Andrea Repele for collaboration and for being a European friend on "the other side of the ocean". I thank my friend and former neighbor Tapas Bhattacharyya for encouragement and for staying in touch.

I thank my old roommate and dear friend Fatima Kuehn for two years of facing together the challenges of pursuing a PhD.

I thank all my friends in Greece for decades of friendship, constant support

and communication that made everything easier. Dimitris Polychronopoulos, for always being there and for the reminders of siga siga (σιγά σιγά) when I needed to take things "slowly, slowly". Nikos Papaioannou, for positive energy, laughs, and support. Thanks also to, Cleio Antoniou, Anta Fotiou, Argiris Van–Brussel, and many others who have been with me even as we are so far away.

I thank my family in Poland and Germany for support and specially my grandmother Maria Jakubowska in Poland, for being there every day with a cheerful and witty character, always encouraging me to pursue knowledge.

I thank my fiancé Steve Nelson for unwavering support and for making this journey hundreds of times easier and more enjoyable.

To my mother and my brother

ABSTRACT

Mammals produce hundreds of billions of new blood cells every day through a process known as hematopoiesis. Hematopoiesis starts with stem cells that develop into all the different types of cells found in blood by changing their genome-wide gene expression. The remodeling of genome-wide gene expression can be primarily attributed to a special class of proteins called transcription factors (TFs) that can activate or repress other genes, including genes encoding TFs. TFs and their targets therefore form recurrent networks called gene regulatory networks (GRNs). GRNs are crucial during physiological developmental processes, such as hematopoiesis, while abnormalities in the regulatory interactions of GRNs can be detrimental to the organisms. To this day we do not know all the key components that comprise hematopoietic GRNs or the complete set of their regulatory interactions. Inference of GRNs directly from genetic experiments is low throughput and labor intensive, while computational inference of comprehensive GRNs is challenging due to high processing times.

This dissertation focuses on deriving the architecture and the dynamics of hema- topoietic GRNs from genome-wide gene expression data obtained from high-resolu- tion time-series experiments. The dissertation also aims to address the technical challenge of speeding up the process of GRN inference. Here GRNs are inferred and modeled using gene circuits, a data-driven method based on Ordinary Differential Equations (ODEs). In gene circuits, the rate of change of a gene product depends on regulatory influences from other genes encoded as a set of parameters that are inferred from time-series data.

A twelve-gene GRN comprising genes encoding key TFs and cytokine receptors involved in erythrocyte-neutrophil differentiation was inferred from a high-resolution time-series dataset of the *in vitro* differentiation of a multipotential cell line. The inferred GRN architecture agreed with prior empirical evidence and predicted novel regulatory interactions. The inferred GRN model was also able to predict the outcome of perturbation experiments, suggesting an accurate inference of GRN architecture. The dynamics of the inferred GRN suggested an alternative explanation to the currently accepted sequence of regulatory events during neutrophil differentiation. The analysis of the model implied that two TFs, C/EBPα and Gfi1, initiate cell-fate choice in the neutrophil lineage, while PU.1, believed to be a master regulator of all white-blood cells, is activated only later. This inference was confirmed in a single-cell RNA-Seq dataset from mouse bone marrow, in which PU.1 upregulation was preceded by C/EBPα and Gfi1 upregulation.

This dissertation also presents an analysis of a high-temporal resolution genome-wide gene expression dataset of *in vitro* macrophage-neutrophil differentiation. Analysis of these data reveal that genome-wide gene expression during differentiation is highly dynamic and complex. A large-scale transition is observed around 8h and shown to be related to wide-spread physiological remodeling of the cells. The genes associated by myeloid differentiation mainly change during the first 4 hours, implying that the cell-fate decision takes place in the first four hours of differentiation.

The dissertation also presents a new classification-based model-training tech-

nique that addresses the challenge of the high computational cost of inferring GRNs. This method, called Fast Inference of Gene Regulation (FIGR), is demonstrated to be two orders magnitude faster than global non-linear optimization techniques and its computational complexity scales much better with GRN size.

This work has demonstrated the feasibility of simulating relatively large realistic GRNs using a dynamical and mechanistically accurate model coupled to high-resolution time series data and that such models can yield novel biological insight. Taken together with the macrophage-neutrophil dataset and the computationally efficient GRN inference methodology, this work should open up new avenues for modeling more comprehensive GRNs in hematopoiesis and the broader field of developmental biology.

CHAPTER 1

Introduction

All the information and instructions for building a complete, fully functioning organism are encoded within the genome of the fertilized oocyte. During development, the zygote slowly morphs into an adult body that is made up of trillions of cells and hundreds of different cell types with nearly all cells bearing identical DNA. Cells start their life as stem or progenitor cells and through the process of cellular differentiation change their gene expression programs and progressively mature to acquire the distinct morphology and function of their fate. How the genetic code is translated into the vast array of cellular identities is one of the fundamental questions in developmental and molecular biology.

A special class of proteins called transcription factors (TF) have the ability to regulate the expression of other genes, including cell identity genes and those encoding TFs themselves. In interacting and regulating each other's expression, TFs form networks known as Gene Regulatory Networks (GRNs) that are particularly important for making cell-fate choices and for cellular differentiation. The topology and architecture of GRNs are still largely unknown for most developmental systems. Understanding the underpinnings of such networks would provide insights into normal development and abnormalities in disease. This dissertation will focus on hematopoiesis, the process of forming all of the cell types found in blood, as a model to study cell-fate choice and differentiation. This dissertation presents a novel GRN model simulating the differentiation of hematopoietic progenitors into neutrophils and erythrocytes that revises the causality of regu-

latory events in neutrophil development. It also describes a machine learning algorithm that considerably speeds up the process of inferring such models from genome-wide gene expression time-series data, and investigates the differentiation of hematopoietic progenitors into neutrophils and macrophages by the analysis of a high temporal resolution genome-wide gene expression dataset.

1.1 MAMMALIAN DEVELOPMENT

Living organisms begin their life as a single cell, a zygote, and through the process of continuous development and growth, they acquire adult morphology with trillions of organized and cooperating cells. Even unicellular organisms develop and change, despite being somewhat less complex than their multicellular descendants. Animal development in a broad sense is the process by which genotype is decoded to create phenotype and can be approached and studied at different levels of system organization, from cells to entire ecosystems (Gilbert & Baressi, 2016).

1.1.1 Main stages of development

Mammalian development can be divided into two main phases, the embryonic and post-embryonic phase. During embryogenesis the zygote is transformed into a fully formed body, while during post-embryonic development the organism grows and maintains its homeostasis.

Embryogenesis is the first part of mammalian development that starts with fertilization and ends with birth. Immediately after fertilization, the zygote un-

dergoes a series of rapid mitotic divisions without overall growth called cleavage that results in a formation of smaller cells called blastomeres. By the end of cleavage, the blastomeres form a hollow sphere called the blastocyst (Gilbert & Baressi, 2016). The inner wall of the blastocyst is lined with a collection of cells called the inner cell mass that will give rise to all the different cell types of the fetus (Zakrzewski et al., 2019). These cells, if derived before implantation, can be differentiated *in vitro* into all fetal cell types and are referred to as embryonic stem cells (ESC).

The next fundamental stage of animal embryogenesis is gastrulation, during which the blastoderm is reorganized from a hollow sphere to a multilayered structure called the gastrula. The gastrula is organized into the three germ layers, endoderm, mesoderm, and ectoderm. The cells in each layer proliferate, differentiate, and mature, giving rise to different tissues and organs through a process called organogenesis. The endoderm gives rise to gastrointestinal, respiratory, and urinary systems and many endocrine glands. The mesoderm forms the notochord, axial skeleton, cartilage, connective tissue, trunk muscles, kidneys, and blood. The ectoderm gives rise to the nervous system, epidermis and various neural crest-derived tissues (Kiecker et al., 2016). During gastrulation and the subsequent organogenesis, the embryo's cells organize into diverse configurations that emerge as tissues of defined form and type. This process involves both morphogenesis, mechanical changes in cell and tissue shape, and gene regulation, which dictates cell-fate decisions and patterning. These two processes are dependent on each other, as morphogenesis can induce changes in gene expression and vice

versa (Shahbazi, 2020).

Development does not stop with birth but continues throughout adulthood, during which senescent cells are replenished by new ones in order to maintain tissue homeostasis. Every day hundreds of billions of cells in the adult human body die and are replaced by new ones. This process of maintenance and repair of tissues is coordinated by stem cells, a special type of cell capable of self-renewing and differentiating into multiple lineages (Klein & Simons, 2011; Häving et al., 2021).

1.1.2 Developmental processes

Development is driven by many different processes that cause cells to grow, proliferate, differentiate, change shape, communicate with each other, migrate, and die. Each of these central developmental processes occurs in precise temporal and spatial manner during the lifetime of a cell. Most cells increase in size and content before they divide into two daughter cells. Cell proliferation is an increase in the numbers of cells through cell division (Kaldis, 2016; Gilbert & Baressi, 2016). During embryogenesis, cells rapidly proliferate and differentiate, or change their fate, to produce the many specialized types of cells that make up different tissues and organs of multicellular animals. During cellular differentiation, cells acquire lineage specific characteristics through the activation of appropriate gene expression programs. For example, each B cell produces a B-cell receptor (BCR) having a unique antigen binding site. When a naïve or memory B cell is activated by an antigen, it proliferates and differentiates into an antibody-secreting effector cell

(Alberts et al., 2002). Cellular differentiation is described in greater detail in Sections 1.3 and 1.4.2.

Migration is an important process shaping development and stem cells have the natural ability to migrate to distant locations in the embryo where they specialize to form different tissues. Stem cells are also one of the few types of cells that migrate in adults. The hematopoietic stem cell (HSC), the progenitor of all the different cell types found in blood, is undoubtedly one of the best examples of migratory stem cells. HSCs reside in the bone marrow but can egress from the bone marrow into circulation, and subsequently extravasate into tissues or ingress into the bone marrow at a different location (de Lucas et al., 2018).

Cells can change shape due to internal mechanical forces such as those exerted by the cytoskeleton or external mechanical forces exerted on the cell from neighboring cells or the extra-cellular matrix (ECM) (Paluch & Heisenberg, 2009). For example, the reorganization of the cytoskeletal architecture that transforms a "cuboidal" epithelial cell into an elongated mesenchymal cell with migratory properties is the defining characteristic of the epithelial-to-mesenchymal transition (EMT) important for both normal development but also for metastasis in cancer (Nelson et al., 2008; Lamouille et al., 2014; Serrano-Gomez et al., 2016; Lai et al., 2020).

Cell-cell communication, the exchange of chemical signals such as growth factors, neurotransmitters, morphogens, hormones, or cytokines, is crucial in embryogenesis and adult homeostasis (Basson, 2012; Hwang, 2013). Intercellular signals are classified according to the distance they traverse. Paracrine signals

act on cells locally through the secretion of signaling proteins (ligands) into the ECM where the signal elicits fast but short-lived responses due to the fast degradation of paracrine ligands. Endocrine signals occur between distant cells, are transported within the blood, and produce a slower but long-lasting response. Autocrine signals are produced by cells that themselves can uptake that signal while direct or juxtacrine signaling involves neighboring cells exchanging the signals by direct contact through the gap junctions. Signaling molecules such as cytokines usually act by binding to cell-surface receptors and initiating certain intracellular signal-transduction pathways that ultimately lead to the activation of signaling effector TFs that drive gene expression programs causing the cell to change its behavior.

Lastly, programmed cell death, or apoptosis, is an important mechanism that controls cell number and eliminates unneeded, infected, mutated, or damaged cells in particular tissues and times (Vaux & Korsmeyer, 1999; Arya & White, 2015). For example, the webbing between toes and fingers or vestigial tails are removed by apoptosis during human embryogenesis (Gilbert & Baressi, 2016).

1.2 CELL-FATE CHOICE

One of the most important processes in development, and the focus of this dissertation, is the specialization of cells into the hundreds of types required for a functioning animal. Also referred to as cell-fate specification, cell-fate choice, or differentiation, this specialization occurs during both embryonic and adult development. While cell-fate choice in embryogenesis starts with pluripotent ESCs,

development is associated with multipotent adult stem cells (ASCs) in adults. ASCs have been identified in most mammalian tissues and are responsible for the homeostasis and the continuous repair and regeneration of tissues (Pekovic & Hutchison, 2008).

The microenvironment of stem cells, known as the stem-cell niche, plays a central role in the maintenance and differentiation of stem cells. Cell-cell and cell-matrix interactions, signaling by soluble molecules, and physical and mechanical stimuli determine whether stem cells will divide symmetrically (symmetric renewal) to produce two stem cells, divide asymmetrically and produce either one stem cell and one differentiated cell (asymmetric division), or two differentiated cells (symmetric commitment) (Redondo et al., 2017; Wang et al., 2018). For example, asymmetrical division of germinal stem cells (GSCs) in the germarium of the *Drosophila* ovary occurs when one of the GSC daughter cells loses the contact with the cap cells of the niche and differentiates into a mature follicle cell while the daughter that remains attached to the cap cells retains its stem cell properties (Panchal et al., 2017). In blood, signaling molecules such as prostoglandin E2 have been shown to increase HSC numbers in the zebrafish aorta-gonad-mesonephros region (North et al., 2007), and the self-renewal and the myelo-lymphoid maturation of HSCs are supported by bone marrow stromal cells in the mouse HSC niche (Seita & Weissman, 2010).

In most cases, it is thought that stem cells adopt a particular fate in a deterministic manner, by the virtue of their lineage or in response to environmental cues. In some situations, however, stem cells choose their fate stochastically, regard-

less of their surroundings or history (Losick & Desplan, 2008). Stochastic cell-fate choice can be observed in the ommatidia, the individual eyes that constitute the compound eye of *Drosophila*. In each ommatidium, a stochastic choice is made in one of the eight photoreceptor cells (called R7) to become one of the two possible cell types. Once this choice is made, the R7 cell instructs the photoreceptors lying underneath it (called R8) to express either a blue-sensitive or a green-sensitive rhodopsin photopigment. Each ommatidium makes its choice independently although the flipping of the coin is biased, as the ratio of blue to green subtypes is 30:70 (Losick & Desplan, 2008).

1.3 CELL-FATE CHOICE AND GENE EXPRESSION

Animal development involves numerous intricate processes all operating in a timely and coordinated manner. The complexity of developing organisms apparently increases in time due to the emergence of new cell types, tissues, and organs, their complex spatial organization, and the emergence of physiological processes. It is a profoundly difficult task to analyze and understand all the details of animal development. However, one problem—that of cell-fate choice—can be understood in terms of a simple organizing principle: the phenotypic differences between cell types can ultimately be attributed to differences in their genome-wide gene expression programs.

According to the central dogma of molecular biology, DNA is transcribed into mRNA and mRNA is translated into proteins that determine phenotypes (Koonin, 2012). Since every somatic cell is a descendent of the zygote, nearly all cell types

have identical genomes (Gilbert & Baressi, 2016). Given that, with rare exceptions, all somatic cells have identical genomes and that proteins are encoded in the DNA, how can different cell types produce different sets of proteins to elicit different phenotypes? A neuron for example is morphologically and functionally radically different from a lymphocyte, despite having the same genome (Uzman, 2003).

The answer to this question, of course, is that even though the DNA sequence of all genes are present in all cell types, genes are expressed at different levels in different cells, creating distinct genome-wide mRNA and protein expression programs or signatures. Cell-fate choice or differentiation is the initial appearance of different gene expression patterns. Differentiation is followed by commitment, the establishment of changes in transcriptional and epigenetic programs that may not be reversed (Ladewig et al., 2013; Gilbert & Baressi, 2016).

1.3.1 Gene regulatory networks

Given that phenotypic differences in cell types arise from differences in genome-wide gene expression, the regulation of gene expression is central to cell-fate choice. Gene expression is controlled by a group of proteins known as transcription factors (TFs) that regulate transcription of their target genes in space and time. TFs regulate transcription by recognizing short motifs or patterns in DNA sequence and binding to DNA in upstream, intronic, or downstream noncoding regions of target genes. The binding sites of TFs usually occur close to each other in noncoding regions called *cis* regulatory elements (CREs), which include pro-

moters and enhancers depending on whether the CRE is proximal or distal to the transcription start site respectively (Mitsis et al., 2020). After TFs bind to CREs, they usually interact with other TFs bound to the CRE, and recruit co-factors to ultimately promote or inhibit the recruitment of RNA polymerase II and act as either activators or repressors respectively. The importance of TFs in cell-fate choice and commitment is best illustrated by trans- or dedifferentiation studies. In a breakthrough experiment, Takahashi and Yamanaka showed that fibroblasts from the adult mouse can be reprogrammed to pluripotent stem cells, called induced pluripotent stem cells (iPSCs), by the enforced expression of only four TFs, Oct3/4, Sox2, c-Myc, and Klf4, now known as the Yamanaka factors (Takahashi & Yamanaka, 2006).

Besides regulating the expression of genes that confer cell-type specific characteristics, TFs also regulate each other's gene expression, forming gene regulatory networks (GRNs). GRNs are usually represented as graphs where nodes are genes and edges are the regulatory relationships between genes. Depending on the level of description, the edges can be directed or undirected, or have weights signifying the strength and the nature of regulation. GRNs integrate the information encoded in a cell's genotype and internal state and the environment to regulate the downstream physiological responses. Decoding the architecture and function of GRNs is central to understanding cell-fate choice. Section 1.5 describes the structure of hematopoietic GRNs and different ways of representing and modeling them.

1.4 HEMATOPOIESIS

Hematopoiesis, from Greek αίμα, "blood" and ποίησις "creation", is the process of blood cell formation. Hematopoiesis is an ongoing process during embryogenesis and adulthood, whilst the hematopoietic system is among the first complex tissues to form in the incipient embryo. The functional, spatial, and temporal characteristics of the hematopoietic system change radically among different developmental stages before the system stabilizes in the bone marrow and thymus in adults. During adulthood, $\sim 10^{11}$ blood cells are replenished every day, and the newly generated cells derive from HSCs that replicate on average once every 40 weeks in humans (Catlin et al., 2011). The requirement for the precise regulation and maintenance of the HSC pool coupled with the demand for a continuous supply of an enormous number of blood cells presents a conundrum that has sparked extensive investigations of the hematopoietic system over the last several decades (Seita & Weissman, 2010). Moreover, the dysregulation of the differentiation and proliferation of hematopoietic cells can lead to numerous hematologic cancers, such as lymphoma, leukemia, or myeloma (Yamashita et al., 2020), which has further prompted the interest in understanding hematopoiesis.

1.4.1 Bone marrow microenvironment

The HSC resides in a highly complex ecosystem inside the bone marrow, called the HSC niche, which promotes the survival and long-term maintenance of the HSC pool (Laurenti & Göttgens, 2018). Precise maintenance and regulation of HSCs in the bone marrow is crucial for the survival of an organism, and self-renewal and

multilineage differentiation of HSCs are strictly regulated by numerous cellular components in the bone marrow microenvironment (Man et al., 2021). In adults, the active or red bone marrow is responsible for the generation of blood cells. Active bone marrow is a nutrient-dense and spongy tissue located in the cavities of cancellous bone. It contains thin branching fibers of reticular connective tissue, hematopoietic cords or islands of cells, and marrow adipose tissue. The bone marrow also includes various types of vessels such as arteries, arterioles, capillaries, and sinusoidal capillaries (sinusoids), that are involved in the transport of cells, nutrients, oxygen, and waste products (Shahrabi et al., 2016).

The HSC niche is defined by the presence of extracellular growth factors, such as stem cell factor (SCF), CXC-chemokine ligand 12 (CXCL12), and thrombopoietin, that support the maintenance of HSCs (Crane et al., 2017). The binding of SCF to receptor tyrosine kinase c-Kit that HSCs express on their surface, leads to its autophosphorylation and the transduction of signals promoting proliferation, migration, survival, and differentiation of hematopoietic progenitors (Lennartsson & Rönnstrand, 2012). CXCL12 promotes HSC maintenance and retention in the bone marrow by signaling through CXC-chemokine receptor 4 (CXCR4) (Zou et al., 1998). Thrombopoietin activates signaling by myeloproliferative leukemia protein (MPL; also known as TPOR) on HSCs.

Several supporting cell types and bone marrow components are known to regulate the niche and produce the extracellular signals required for HSC maintenance. Perivascular stromal cells or CXCL12-abundant reticular cells (CAR cells), and endothelial cells synthesize both CXCL12 and SCF in bone marrow. SCF is

mainly expressed by perivascular stromal cells that are associated with sinusoidal blood vessels throughout the bone marrow and to a lesser extent by endothelial cells (Sugiyama et al., 2006). Osteoblasts, or bone-forming cells, also produce CXCL12 but at approximately 1000-fold lower levels than perivascular stromal cells (Crane et al., 2017). Macrophages promote HSC quiescence *in vivo* by producing transforming growth factor β1 (TGFβ1) (Zhao et al., 2014; Yamazaki et al., 2009). Thrombopoietin is expressed at high levels in the liver and to a lesser extent in kidney, with limited expression in the bone marrow under normal circumstances, and it is currently unknown whether HSC maintenance is promoted by thrombopoietin that is produced locally in bone marrow or at distant sites (Crane et al., 2017). Although nerve fibers, and the associated Schwann cells, are not required for the maintenance of HSCs in bone marrow, they regulate the daily circadian rhythm of HSC mobilization from bone marrow into the blood, perhaps by regulating the cyclical expression of CXCL12 by stromal cells (Crane et al., 2017). Whether there are perivascular domains that serve as niches for different types of hematopoietic progenitors, or whether other growth factors and molecules take part in HSC maintenance remain open questions about the regulation of HSCs by its niche (Crane et al., 2017).

1.4.2 Hematopoietic differentiation

The concept of the stem cell was first defined by studies of hematopoiesis, especially the pioneering and highly influential work of McCulloch and Till in the 1960s. With a series of experiments on bone marrow transplantation in heavily ir-

radiated mice they discovered a very small sub-population of donor bone marrow cells that had the ability to generate multiple types of myeloerythroid cells and the ability to self-replicate. These findings introduced the two defining criteria of stem cells i.e. multi-potency and self-renewal (Becker et al., 1963; Siminovitch et al., 1963; Till & McCulloch, 1961; Wu et al., 1968). Today we know that HSCs are the only type of cell that can differentiate into all functional blood cell types and also give rise to identical HSC daughter cells (Seita & Weissman, 2010).

The quest for the identification and isolation of HSCs from mouse bone marrow started in the 1980s with landmark studies that utilized fluorescence activated cell sorting (FACS) technologies and monoclonal antibodies (Spangrude et al., 1988). Cells express a distinct assortment of proteins and lipids on their plasma membrane and these cell-surface markers can be used to distinguish between different cell types (Gundry et al., 2008). For example, it was demonstrated that cells that are Thy-1^{low}, that is, they express low levels of Thy-1 or CD90, Lin $^{-}$, which implies that they do not express lineage specific markers, and Sca-1^{+}, that is, they express Sca-1, are capable of reconstituting the entire hematopoietic system for more than 3 months when transplanted into lethally irradiated mice (Spangrude et al., 1988; Seita & Weissman, 2010). This combination of cell-surface marker expression thus serves as the definition for HSCs with long-term reconstitution capacity (LT-HSCs). Since then the definition of LT-HSCs has been refined as cells that can reconstitute hosts for 16 weeks and for an additional 16 weeks in secondary transplantation. The cell-surface phenotype of LT-HSCs has also been further refined to include c-Kit, CD34, CD150, Flt3, and CD48 so that some in-

vestigators isolate LT-HSCs as $CD150^{+}CD48^{-}CD34^{+}Lin^{-}Sca\text{-}1^{+}Kit^{+}$ cells) (Ali et al., 2017).

The mammalian hematopoietic system produces many distinct blood cell types. The erythrocytes and megakaryocytes/platelets comprise the red blood cell lineage. Monocytes/macrophages, granulocytes such as neutrophils, basophils, eosinophils, and mast cells make up the myeloid lineage. Lymphoid cells consist of T and B lymphocytes, natural killer cells, and dendritic cells. In addition to the terminally differentiated cell types found in blood, there are several intermediate cell types that exist transiently during differentiation. If a cell can produce all blood cell types and engraft transiently in primary (and in some cases secondary) transplants, they are referred to as Intermediate-Term (IT-) HSCs, Short-Term (ST-) HSCs or Multipotent Progenitors (MPPs) depending on the length and robustness of the graft produced (Laurenti & Göttgens, 2018).

Experiments characterizing the progenitor populations and their differentiation potential downstream of the HSC (Morrison & Weissman, 1994; Christensen & Weissman, 2001) have resulted in a hierarchical hematopoietic model depicted usually as a tree (Fig. 1.1A). The HSC resides at the top of the tree, which then branches out and segregates the lymphoid lineage potential from the myeloid, erythroid, and megakaryocytic lineages, followed by a number of further branching steps on either side of the tree progressing from multipotent to bipotent and finally to unipotent progenitor cells. The common myeloid progenitor (CMP) gives rise to the more restricted megakaryocytic erythroid progenitor (MEP) and the granulocyte monocyte progenitor (GMP) that produce the megakaryocytic-

erythroid and granulocyte-monocyte lineages respectively. The common lymphoid progenitor (CLP) gives rise to lymphoid lineages comprising T, B, and natural killer cells. This hematopoietic differentiation model is still used in many textbooks today (Laurenti & Göttgens, 2018). Inclusion of additional surface markers and the outcomes of subsequent transplantation experiments led to a revision of the previous hierarchy, in which the multipotent progenitor compartment was further divided into distinct sub-populations and the myeloid and lymphoid lineages diverge after the red-blood cell lineage has branched off (Fig. 1.1B). More recent studies utilizing single-cell RNA sequencing (scRNA-Seq) do not support the existence of distinct intermediate progenitor populations and suggest that differentiation likely occurs as a continuous process, with gradual changes in cellular state as cells traverse from the HSC to unipotent progenitors (Fig. 1.1C) (Paul et al., 2015; Nestorowa et al., 2016; Velten et al., 2017).

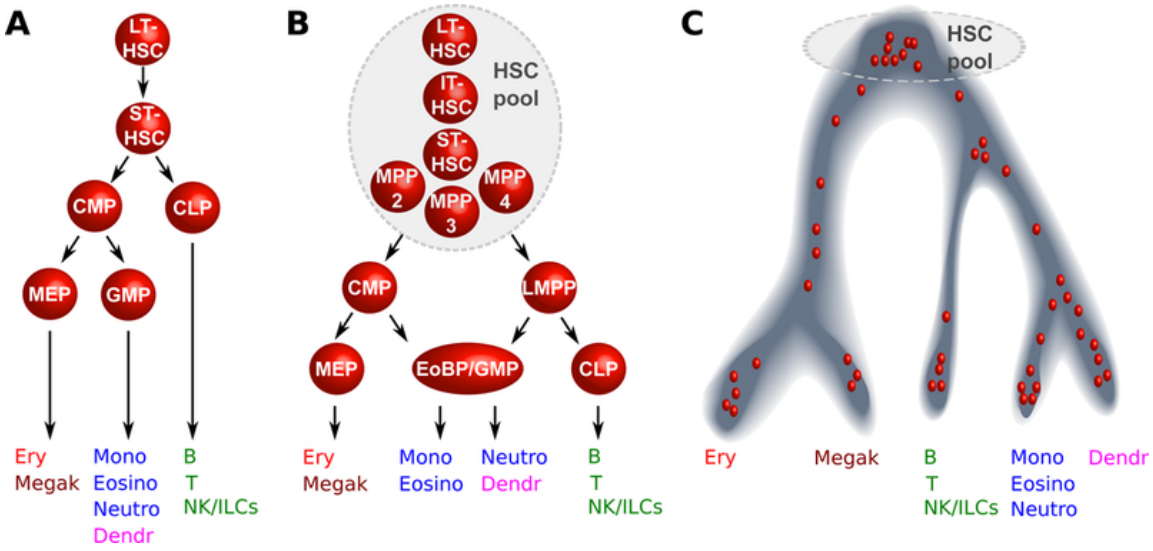


Figure 1.1: Conceptual evolution of hierarchical models of hematopoiesis. (A) A model of hematopoietic differentiation (early 2000 to 2005). HSCs represent a homogeneous population downstream of which the first lineage bifurcation separates the erythro-myeloid and lymphoid branches via the common myeloid progenitors (CMP) and common lymphoid progenitors (CLP) populations. The CMP gives rise to further downstream progenitors, the megakaryocytic erythroid progenitor (MEP) and the granulocyte monocyte progenitor (GMP). (B) An alternative model of hematopoiesis (2005-2015). The HSC pool is heterogeneous and contains distinct sub-populations both in terms of self-renewal (vertical axis) and differentiation (horizontal axis). The myeloid and lymphoid branches remain associated further down in the hierarchy via the lymphoid-primed multipotential progenitor (LMPP) population and the GMP compartment is shown to be fairly heterogeneous, additionally containing the eosinophil basophil progenitor (EoBP). (C) Another one alternative hematopoietic model (from 2016 onward). scRNA-Seq data indicate a continuum of differentiation. Each red dot represents a single cell and its location along the differentiation trajectory. Used with permission of Springer Nature (Laurenti & Göttgens, 2018).

In parallel to investigations of the lineage relationships between different blood

cell types, a great deal of effort has been made to understand the genetic basis of these cell-fate decisions. The main model of cell-fate choice in hematopoiesis involves the ideas of multi-lineage priming and lineage cross antagonism. Multi-lineage priming is the simultaneous low-level expression of genes associated with multiple lineages in progenitor cells (Laslo et al., 2006a; Enver et al., 2009; van Galen et al., 2014). Lineage cross antagonism is the repression of gene expression programs of alternative lineages by TFs expressed in a particular lineage. Together, multi-lineage priming and cross antagonism imply that alternative lineage programs are competing in progenitor cells and cell-fate choice is made when one gene expression program prevails while the rest of the programs are extinguished.

The most famous model of cell-fate specification by priming and cross antagonism is drawn from the erythrocyte-leukocyte decision process. GATA1 and PU.1 (encoded by *Spi1*) are required for the production of mature cells in megakaryocyte/erythroid and granulocyte-macrophage lineages respectively and can reprogram cells towards their cognate lineages upon over-expression (Graf & Enver, 2009a). GATA1 and PU.1 inhibit each other's expression (Zhang et al., 2000; Nerlov et al., 2000) while also autoactivating their own expression (Tsai et al., 1991; Chen et al., 1995). GATA1 represses PU.1 and its target genes by binding and interacting with PU.1 and preventing the recruitment of its coactivator c-Jun while simultaneously inhibiting histone H3K9 acetylation that turns the genes on. PU.1 represses GATA1 and its targets by outcompeting a histone acetyltransferase (C-terminal binding protein or CBP) and recruiting histone H3K9 methyltransferases that create a repressed chromatin structure (Burda et al., 2010). Auto-

activation is thought to stabilize and reinforce the decision once it has been made. GATA1 and PU.1 positive auto-regulation stabilizes erythroid (Nishikawa et al., 2003) and myeloid fate (Okuno et al., 2005) respectively, while the triad of Gata2, Tal1/Scl, and Fli1 is thought to stabilize the stem/progenitor state in mice (Pimanda et al., 2007; Narula et al., 2010). This motif, two TFs that repress each other while also auto-activating their own expression, is referred to as a bistable switch, since mathematical models representing it have two stable solutions (Huang et al., 2007a). The bistable switch is such an appealing framework that such switches have been hypothesized for nearly every lineage decision (Graf & Enver, 2009b).

Recent long-term live tracking study has however questioned the popular GATA1-PU.1 bistable switch as a prime driver for the erythroid/myeloid fate choice (Hoppe et al., 2016). The study tracked hematopoietic progenitors derived from mice in which GATA1 and PU.1 had been tagged with fluorescent proteins. The study failed to detect any cells simultaneously expressing both GATA1 and PU.1 at low levels, a key requirement of the bistable switch model, suggesting that the bistable switch was reinforcing a decision already made instead of making the decision.

The bistable switch model requires some initial asymmetry in order to create biased expression patterns. Two alternative models have been proposed to explain the emergence of this asymmetry. The instructive or deterministic model posits that signaling molecules like cytokines provide the initial asymmetry or "instruct" progenitors to differentiate into specific lineages (Bhoopalan et al., 2020). In the stochastic or permissive model, lineage commitment and terminal differentiation are determined in a cell-intrinsic manner whereas cytokines provide per-

missive growth and survival signals (Robb, 2007a). In the instructive model, cytokine signals are transduced to the cells ultimately activating certain pathways that lead to cell-fate decisions. It is supported by *in vitro* experiments on bipotential GM-colony forming cells (GM-CFCs), which develop into macrophages when cultured with M-CSF and into granulocytes when cultured with stem cell factor or G-CSF (Robb, 2007b). The stochastic model is supported by experiments that show that cell fate is independent of the identity of the cytokine. For example, a human GM-CSF receptor transgene expressed in EPOR-null fetal liver cells causes erythropoiesis and not granulopoiesis when the cells are stimulated by GM-CSF, implying that the erythroid potential of the cells is independent of EPOR (Robb, 2007b). More recent, single-cell tracking studies have conclusively demonstrated an instructive role for cytokines by showing that bipotential cells differentiate in response to cytokines without significant cell death, a necessary condition for the stochastic model (Rieger et al., 2009).

Extensive studies on stem cells and HSCs have revealed many different processes and molecules participating in and driving cell fate decisions, but the snapshots we get from the experiments have not enabled us so far to see the entire "developmental movie" that takes place inside and outside cells. We have the idea that cell decisions are made on the basis of intrinsic cellular states, which are guided by gene expressions programs which in turn are influenced by cellular stochasticity or extrinsic cues like external signaling and cell's microenvironment. How exactly differentiation proceeds in time and how the key components during differentiation work together to produce rich differentiation outcome are still

largely unknown. Although experimental studies can determine individual players, it is difficult to ascertain how all these disparate components work together to decide cell fate. Over the past decades computational and mathematical modeling has emerged as a means to understand the differentiation in depth, to build unified theories that can explain diverse experimental results, and provide a broad and global view of cellular differentiation.

1.5 MODELING HEMATOPOIESIS

A mathematical model is a simplified but quantitative representation of a complex system. One of the first and most famous examples of mathematical models was Newton's laws of motion applied to the solar system. Newton's model was a simplified version of reality because it did not account for all the details of the universe but rather focused only on the motions of the planets and considered them as single points, with properties such as mass, position, and velocity. The abstraction of a system that reduces it to its essential characteristics is a fundamental aspect of modeling. Models are simplifications and approximations of the real world, such as the "spherical cows" [1] in the humorous metaphor used by the physicists when describing highly simplified scientific models of complex phenomena. The famous aphorism "all models are wrong, but some are useful" coined by a statistician George E. P. Box, acknowledges the fact that while models

[1] The phrase comes from a joke that spoofs the oversimplified assumptions about the systems used in theoretical physics. The joke talks about a dairy farmer who wanted to increase the milk production at his farm and sought the help of a theoretical physicist at the local university. After carefully studying the problem, the physicist tells the farmer, "I have a solution, but it only works if we assume a spherical cow."

will never encompass all the complexity of reality, sound mathematical models are nevertheless very useful. Newton's laws of motion describe with high accuracy the motions of objects in the classical realm while mathematical modeling of infectious disease has provided among others the means of predicting the scale of disease transmission or providing advance warning during the outbreak of epidemics.

1.5.1 Levels of modeling

Biological systems such as hematopoiesis are inherently complex. Each individual cell is a microcosm of millions of molecules interacting with each other. Most models of hematopoiesis reduce that complexity by focusing on the most fundamental aspect of cell-fate differentiation and commitment, which is gene regulation. The spatial and temporal pattern of a gene's expression is determined primarily by the regulation of its transcription and/or translation. Gene expression is regulated by TFs, non-coding RNAs, and other regulatory molecules through promoters and enhancers. The regulators of genes are themselves regulated by other regulators, often by the products of their own targets, forming gene regulatory networks (GRNs). These processes can be modeled by two broad categories of models. In coarse-grained GRN models (Huang et al., 2007a; Laslo et al., 2006b), the details of gene regulation, such as DNA sequence and TF binding sites are not represented and genes are treated like black boxes (Fig. 1.2). Sequence-based models (Bertolino et al., 2016; Segal et al., 2008; Kim et al., 2013), on the other hand, include such details. Most modeling in hematopoiesis has employed

coarse-grained GRN models and we focus on those in this dissertation.

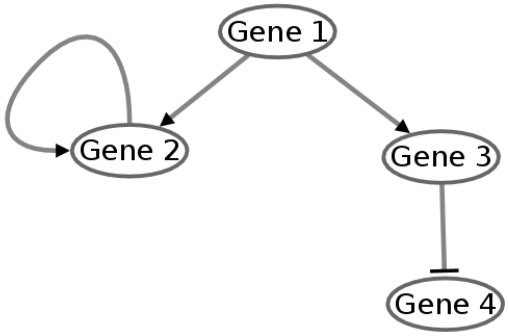


Figure 1.2: An example of a GRN with four genes. The genes in the network are represented as nodes, while the interactions between the genes are represented as edges. Gene activation is indicated with pointed arrows and repression is indicated with flat-headed arrow. In this GRN, Gene1 activates Gene 2 and 3, Gene 2 autoactivates itself, and Gene 3 represses Gene 4. In some GRN graphs, edges bear weights that signify the strength of interactions between the genes.

1.5.2 The problem of GRN inference

Despite their central biological roles, both the architecture and the function of GRNs are poorly defined in most developmental systems. One difficulty in the inference of GRNs using genetic analysis is low throughput of such experiments. High-throughput biochemical approaches, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) also have limitations such as the lack of suitable affinity reagents for all TFs and the large number of combinations of TFs and cellular states that have to be analyzed. The recent revolution in DNA sequencing technologies and the progressive reduction of sequencing costs has en-

abled an alternative approach to GRN inference during cellular differentiation. RNA sequencing costs are low enough that is possible to measure abundances of the entire transcriptome—which defines the state of a cell—at high spatial and temporal resolution. Training mechanistic models on such time-series data using machine learning approaches allows for the inference or "reverse-engineering" of GRNs.

1.5.2.1 Training data for inferring GRNs

Inferring GRNs by training mechanistic models can be accomplished by measuring GRN state using either protein concentration or mRNA concentration as readouts, although the latter is used most commonly. Microarrays, RNA-seq, or single-cell RNA-seq (scRNA-seq) experiments allow the monitoring of the mRNA levels of thousands of genes simultaneously either in bulk or at a single-cell level. Such experiments can be classified into two groups: time-series experiments and perturbation experiments. The former are used to understand dynamic processes in the cell while the latter reveal genetic interconnections by identifying genes affected by treatments such as knockouts, knockdowns, or overexpression of TFs. An important part of time-series experiments are inducible cell differentiation systems, which are cell lines that can be induced to differentiate along different lineages *in vitro*. More comprehensive studies combine time-series data with perturbations and such data are utilized in Chapters 2 and 4.

1.5.2.2 Estimating GRN parameters by model training

Nearly all GRN models have a number of free parameters whose values have to be chosen in order to simulate gene expression. The number of free parameters usually depends on the number of genes in the model. In densely interconnected networks of n nodes, the total number of parameters grows as $O(n^2)$. There are generally three approaches available for setting the values of the free parameters. In the first approach (Ackers et al., 1982; Reinitz & Vaisnys, 1990), the values are chosen based on empirical measurement of the biophysical and biochemical properties of the system. In the second approach, "qualitative modeling" (Huang et al., 2007a; Laslo et al., 2006b; Chickarmane et al., 2009; Tyson et al., 2011), the parameter space is systematically sampled to identify regions that produce gene expression patterns in qualitative agreement with those observed empirically. In the third approach, "data-driven modeling" or "reverse engineering" (Jaeger et al., 2004c; Manu et al., 2009; Aluru, 2005; Handzlik & Manu, 2022), the values of the parameters are determined by fitting to quantitative gene expression data. One modeling method, gene circuits (Reinitz & Sharp, 1995a; Jaeger et al., 2004b; Manu et al., 2009; Handzlik & Manu, 2022), combines the inference of network topology and genetic architecture and parameter values into a single fitting procedure. Once the parameters have been estimated, the GRN models can be used for mechanistically investigating the biological system through the simulation of the wildtype GRN as well as after *in silico* perturbation.

1.5.3 Approaches to coarse-grained GRN modeling

Depending on the level of abstraction, the availability and type of empirical data, the prior knowledge of the biological system, and the purpose of the study, there are different approaches to coarse-grained GRN modeling. If the goal of the study is to represent simple relations between genes, a qualitative model representing the topology and interactions between genes is sufficient, but if *in silico* predictions are of interest, then quantitative modeling is imperative.

GRN models can be either dynamic or static, depending on whether they simulate the evolution of gene product concentrations in time or not. Static models are limited to simulating GRNs at equilibrium and therefore cannot model the transient behavior of the system. Dynamic models, in contrast, are capable of simulating transient behavior relevant for modeling transient processes such as cell-fate decisions during development. Dynamic models are also more computationally demanding than static ones since they require the solution of differential equations. GRN models can also be deterministic or stochastic. The latter simulate random fluctuations of gene expression in individual cells while the former are limited to simulating the average gene expression in a population of cells. In stochastic models, gene expression is described by random variables which follow a probability distribution. Even when all the parameters and initial conditions are the same, the expression of a gene will be different in each independent stochastic simulation, while a deterministic model produces exactly the same output in each simulation. Additionally, GRNs can be synchronous, operating under the assumption that all states of the genes are updated at the same time, or asyn-

chronous, where one variable at a time is updated (Aluru, 2005).

1.5.3.1 Differential Equation Models

The most detailed models of any complex dynamical system, like GRNs, are based on Ordinary Differential Equations (ODEs) that describe how the rates of change of state variables depend on the values of the state variables as well as extrinsic inputs. ODE models are deterministic and capable of describing non-linear dynamics of complex systems such as hematopoiesis. ODE GRN models represent the concentration of gene product i, where $i \in 1, \cdots, n$, by time-dependent variables $x_i(t),\ t \in T$, where $T \subseteq \mathbb{R}_{\geq 0}$ is a closed time interval and n is the number of genes. $x_i : T \to \mathbb{R} \geq 0$ are non-negative real numbers since concentrations cannot take negative values. The time evolution of $x_i(t)$ is given by the solution of a system of ODEs

$$\frac{dx_i}{dt} = f_i(\boldsymbol{x}) \quad i \in 1, \cdots, n, \quad \mathbf{x}(t=0) = \mathbf{x}_0, \tag{1.1}$$

where $\boldsymbol{x} = x_1, \cdots, x_n{}'$, the function $f_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ represents the regulatory interactions that determine the rate of change $x_i(t)$ (de Jong & Geiselmann, 2005), and $\mathbf{x}_0$ are the initial concentrations of the gene products. The specific functional form of f_i is determined from hypotheses about the regulatory architecture of the GRN and the biochemical principles of gene regulation. $f_i(\boldsymbol{x})$ usually has a non-linear sigmoidal form since the rate of transcription varies between zero and maximum determined by the recruitment of RNA Pol II to promoters (Aluru, 2005). Eq. 1.1

can be rewritten as a vector equation

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}), \tag{1.2}$$

where $\boldsymbol{f} = f_1, \cdots, f_n{}'$ (de Jong, 2002). Different forms of f representing GRNs (Eq. 1.2) have been proposed in various studies. A novel representation of a complex dynamical ODE system describing differentiation of hematopoietic progenitor cells is presented in section 2.2.1.1.

An example of an ODE model of a GRN with two genes *a* and *b* and gene products A and B forming a mutual-inhibition network, represented as a circuit rather than classical network graph, is presented in Fig. 1.3. x_a and x_b represent the concentrations of proteins A and B. The rate of change in the concentration of x_a equals the difference of the synthesis rate, $\kappa_a h^-(x_b, \theta_b, m_b)$, and the degradation term $\gamma_a x_a$. κ_a is the maximum synthesis rate of protein A, while the repression of gene *a* by protein B is given by the Hill function $h^-(x_b, \theta_b, m_b)$ where $0 \leq h^- \leq 1$. $h^- = 1$ for $x_b = 0$ and h^- asymptotically reaches 0 when $x_b \to +\infty$. For high values of x_b the h^- approximates zero which reduced significantly the synthesis rate κ_a, whereas for low values of x_b the gene a is expressed at synthesis rate close to κ_a. The steepness of h^- depends on the cooperativity parameter m_b and the inflection point of the curve is given by θ_b. For $m_b > 1$ the Hill function has a sigmoidal form that is often observed experimentally and when $x_b = \theta_b$ the expression of gene *a* reaches half of its maximum synthesis rate κ_a. The second term of the differential equation, $\gamma_a x_a$, represents the degradation of protein A as a first order reaction. Therefore degradation is not regulated in this example. The differential equation

for x_b has an analogous interpretation (de Jong & Geiselmann, 2005).

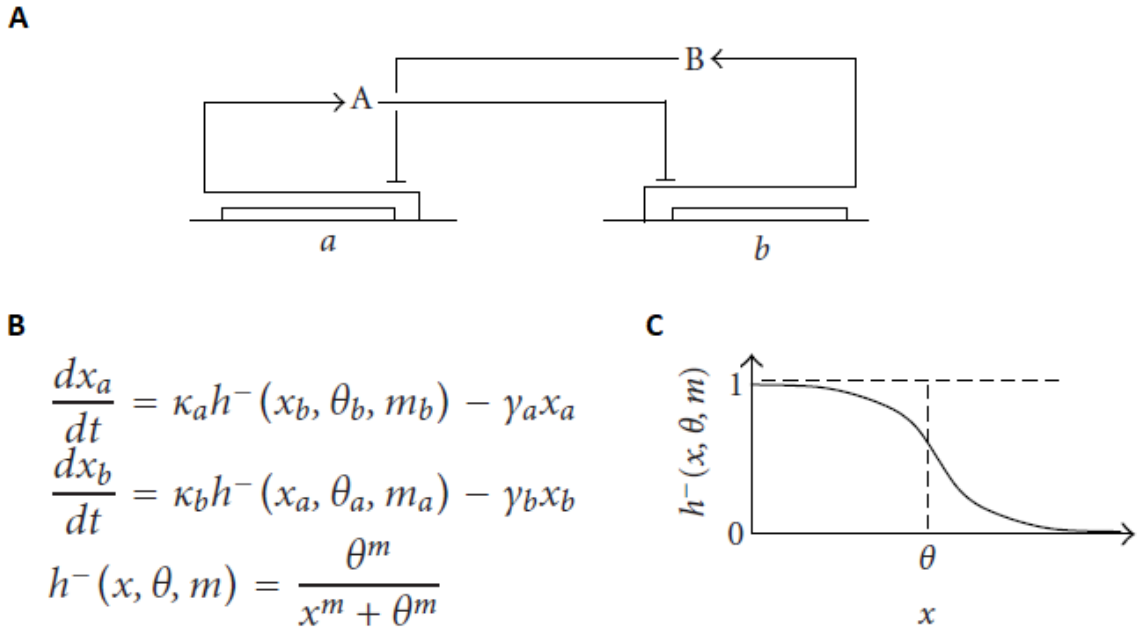


Figure 1.3: An example GRN with the topology of a bistable switch. ***A.*** The representation of GRN as a circuit with genes a and b, and their protein products A and B that repress each other's gene expression ***B.*** A system of ODEs with the first and second equation describing the time evolution of the protein concentrations of gene a and b respectively. The rate of change in protein concentration depends on the difference between the synthesis and degradation rates. The synthesis term is the product of the maximum synthesis rate κ and the Hill function defined in the third equation. ***C.*** Graphical representation of the Hill function, which takes a sigmoidal form for $m > 1$. Adapted from (de Jong & Geiselmann, 2005).

Because of the non-linearity of $\mathbf{f}$, equations such as Eq. 1.2 usually cannot be solved analytically and the solutions are approximated numerically using methods such as Runge-Kutta (Bulirsch & Stoer, 1992; Magnusson et al., 2017; Fehr et al., 2019; Ma et al., 2020; Handzlik et al., 2021).

Gene circuit models. Gene circuits (Reinitz & Sharp, 1995b) are deterministic models utilizing coupled ODEs or PDEs (Eq. 1.1) with a specific choice of f to represent the synthesis and degradation of gene products. Furthermore, GRN architecture is encoded in the values of the free parameters of gene circuits so that they can infer GRN architecture when trained with gene expression time series data. In gene circuit models Eq. 1.1 takes a special form where the time evolution of the concentrations of the mRNA or protein product, $x_g(t)$, of genes $g = 1, \cdots, G$ is described according to G coupled ordinary differential equations

$$\frac{dx_g}{dt} = R_g \, S\left(\sum_{f=1}^{G} T_{gf} x_f + h_g\right) - \lambda_g x_g \tag{1.3}$$

where R_g is the maximum synthesis rate of product g. T_{gf} are genetic interconnectivity coefficients describing the regulation of gene g by the product of gene f. Positive and negative values of T_{gf} imply activation and repression of gene g by gene f respectively. The threshold h_g determines the basal synthesis rate, and λ_g is the degradation rate of product g. $S(u)$ is the regulation-expression function, which controls the level of synthesis relative to the maximum synthesis rate of the gene R_g and which depends on the regulatory input $u = \sum_{f=1}^{G} T_{gf} x_f + h_g$. $S(u)$ takes values between 0 and 1. A commonly utilized form of the regulation-expression function (Reinitz & Sharp, 1995b; Jaeger et al., 2004a) is the sigmoid function (Fig. 1.4)

$$S(u) = \sigma(u) = \frac{1}{2}\left(\frac{u}{\sqrt{1+u^2}} + 1\right). \tag{1.4}$$

It has been shown that gene circuits are capable of representing arbitrarily complex gene expression patterns (Vakulenko & Grigoriev, 2003).

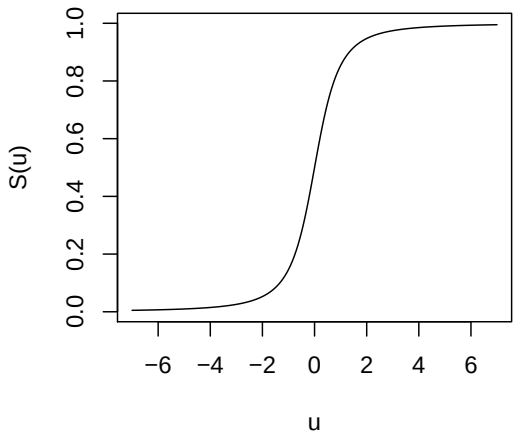


Figure 1.4: An example of a sigmoid function. Sigmoid function $S(u) = \frac{1}{2}\left(\frac{u}{\sqrt{1+u^2}} + 1\right)$ where $-\infty < u < +\infty$ and $0 < S(u) < 1$. When S is applied to GRN models such as the one described in Eq. 1.3, then u is perceived as a regulatory input. If regulatory input u is positive, then $S(u) > \frac{1}{2}$, and if u is negative, then $S(u) < \frac{1}{2}$. For $u = 0$, $S(u) = \frac{1}{2}$ and synthesis rate reaches half of its maximum synthesis rate R.

1.5.3.2 Non-differential-equation based models

Numerous methods and algorithms other than ODE models have been employed for the modeling of GRNs. Some of them include statistical methods such as Boolean, Bayesian, or graphical networks, and machine learning algorithms such as neural networks (de Jong, 2002; Cahan et al., 2014; Sanchez-Castillo et al., 2017; Hamey et al., 2017; Shu et al., 2021). Some of the most commonly used ones are

presented below.

Boolean Networks. The simplest dynamic network models that are capable of exhibiting some of the biological and systems-level properties of real gene networks are the Boolean networks. Boolean networks were first used by Kauffman in the 1970s as a framework for modeling biological networks (Kauffman, 1974). Boolean networks contain a set of n state variables $X = \{x_1, x2, \cdots, x_n\}, \quad x_i \in \mathbb{B}$. The state variables can assume only a discrete number of values (Davidson et al., 2002; Theiffry et al., 1993; Sánchez & Thieffry, 2001; Collombet et al., 2017b; Bonzanni et al., 2013). In their most common realization, state variables can only take two values, 0 or 1, which are analogous to inactive (0) or active (1) genes in the context of GRNs. The state of a Boolean network at time t is defined by a vector $\vec{x(t)} = (x_1(t), \cdots, x_n(t))$. In Boolean network models, time is considered to be discrete so that the variables evolve in a finite number of steps. Each variable is updated from one time point to the next $x_i(t) \mapsto \mathrm{x}_i(t+1)$ by a Boolean function $x_i(t+1) = f_i(\vec{x(t)}), \quad f_i : \mathbb{B}^n \to \mathbb{B}$ (Schwab et al., 2020). Although Boolean network GRN models are usually constructed by hand curation of experimental evidence or literature (Davidson et al., 2002; Bonzanni et al., 2013; Collombet et al., 2017b), they may also be reverse-engineered by fitting to gene expression data (Aluru, 2005; Schwab et al., 2020). A major downside of Boolean networks is that, even though the number of transcripts or protein molecules varies smoothly (Tusi et al., 2018), gene expression has to be discretized into a finite number of values by a user-defined threshold. Such thresholds are arbitrary, may not have

a sound biological basis, and reduce the reproducibility of such models. Many genes present complex behaviour such as continuous changes in expression or oscillations. The discretization of gene states for such genes could lead to a loss of information about gene expression dynamics and erroneous conclusions.

Bayesian Networks Bayesian Networks are a class of graphical probabilistic models. A Bayesian Network is an annotated acyclic graph $G(X, E)$ where the nodes, $x_i \in X$, are random variables representing gene expression and the edges indicate the dependencies between the nodes. The random variables are drawn from conditional probability distributions $P(x_i|Pa(x_i))$, where $Pa(x_i) \subseteq X$ is the set of parents for each node. Given its parents, each variable is independent of its non-descendants. Each Bayesian network uniquely specifies a decomposition of the joint distribution over all variables down to the conditional distributions of the nodes: $P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|Pa(x_i))$. The conditional probability distributions are usually modeled by Gaussian distributions assuming a linear dependence of the child node's mean expression on the mean expression of the parents (Friedman et al., 2000). Given genome-wide gene expression data, the goal is to infer the GRN topology or architecture as well as the weights through which a child node's expression depends on its parents' expression (Aluru, 2005). One limitation of Bayesian networks is that they are static and only model GRNs at equilibrium. Additionally, Bayesian networks cannot model GRNs containing feedback loops since they are limited to acyclic graphs.

1.6 HEMATOPOIETIC GRN MODELS

Numerous GRN models have been built to simulate hematopoietic differentiation over the past decade or so using diverse approaches. Hematopoietic GRN models have largely focused on transcriptional interactions between the core regulators of hematopoietic differentiation. The earliest models of hematopoiesis focused on 2 or 3 key regulators in a bistable-switch framework (Laslo et al., 2006a; Chickarmane et al., 2009; Narula et al., 2010). These early efforts captured some key properties of hematopoietic cell-fate specification, but clearly oversimplified the complexities of multifactor regulatory interactions (Göttgens, 2015). More recent work has attempted to model larger, more realistic GRNs (Collombet et al., 2017a; Hamey et al., 2017), but has largely relied on the simplified and computationally inexpensive Boolean network framework. Dynamic hematopoietic GRN models constructed so far can be divided into three main classes: Boolean network models, qualitative differentiation equation models (Section 1.5.2.2), and data-driven differential equation models.

Hand-curated and dynamical Boolean network comprised of 11 TFs (*Runx1, Smad6, Spi1, Eto2, Gata1, Scl, Fli1, Erg, Zfpm1, Gata2, Hhex*) active in early HSCs simulated the direct transitions of the HSC state to B-cell, monocyte, natural killer (NK), CD4 (T cells), CD8 (cytotoxic T cells), CD4-activated, CD8-activated, granulocyte, and erythroid states. This model predicted a novel negative regulation of Fli1 by Gata1 which was further experimentally verified (Bonzanni et al., 2013). Another hand-curated and dynamical Boolean network with 21 core component and regulators collected from the literature analysis and ChIP-seq datasets simu-

lated the cytokine-induced B-cell and macrophage differentiation from multipotent progenitors (Collombet et al., 2017a). An example of an even larger dynamical hematopoietic network comes from a study of data-driven Boolean network simulating the differentiation of HSCs into the lymphoid-primed multipotent progenitors comprised of 29 TFs and a separate Boolean network simulating the differentiation of HSCs into megakaryocyte-erythroid lineage with 31 TFs (Hamey et al., 2017).

Qualitative differentiation equations models that do not fit the model to data but instead find parameter sets that match the qualitative behavior of the system have been typically utilized to model cell-fate decisions as bistable switches. Examples of such networks are the 2-gene bistable switch between *Gata1* and *Spi1* simulating erythroid and myeloid/monocytic differentiation (Huang et al., 2007b), a 2-gene bistable switch between *Egr (Egr1/Egr2/Nab2)* and *Gfi1* simulating the differentiation of macrophage and neutrophil lineages (Laslo et al., 2006a) or the 4-gene GRN with *Gata1, Spi1, Fog1, Cebpa* simulating the erythroid-myeloid lineage decisions (Chickarmane et al., 2009).

Data-driven differential equation models are trained on the data to infer the values of their unknown parameters. A 3-gene GRN comprising *Gata1, Gata2* and *Spi1* was inferred to simulate the differentiation of FDCP-mix cells into erythrocytes (May et al., 2013a). A bigger network of 12 TFs (*Copeb, Ets1, Gata3, Irf4, Jun, Maf, Myb, Nfatc3, Nfkb1, Rela, Stat3, Usf2*) was inferred for the differentiation of naïve T cells into Th2 helper T-cells (Magnusson et al., 2017).

Even though many hematopoietic models have been built, only two used the

mechanistically accurate differential equation framework and were data-driven. The first (May et al., 2013a) modeled only a 3-gene network in one lineage (erythroid). The other one modeled a larger, more realistic 12-gene GRN but only in one lineage (Magnusson et al., 2017). Models that are simultaneously 1) mechanistically accurate (ODE-based), 2) realistic (large GRNs), and 3) capable of simulating cell-fate choice, that is, differentiation into two or more lineages, have not been developed yet.

1.7 THE GOALS AND STRUCTURE OF THIS DISSERTATION

Due to the complexity of cell-fate decisions, the size and interconnectedness of the GRNs, and the technical challenges of assaying and modeling them, the architecture of hematopoietic GRNs and the causality of transcriptional and regulatory events remain poorly understood. The goals of this dissertation are to use gene circuits and high temporal resolution time-series data to determine GRN architecture and causality during hematopoietic cell-fate choice, develop methods to enable mechanistic modeling of larger, more realistic GRNs, and to analyze high temporal resolution genome-wide gene expression datasets to understand regulatory events during differentiation and enable more comprehensive GRN modeling in the future.

To determine the architecture and causality of GRNs during the erythroid-neutrophil development, a gene circuit built on high-temporal data from differentiating in vivo hematopoietic progenitors is presented in Chapter 2. This chapter is a published study of a twelve-gene data-driven gene circuit model capturing

the complex interactions of key components during the differentiation of progenitor cells into two lineages. Although it was possible to infer the aforementioned network using global non-linear optimization methods, it required many weeks of computation time. A practical solution to the high computational cost of inferring large GRNs from high-resolution gene expression data is presented in Chapter 3. This chapter includes work that was a part of a manuscript describing a novel classification-based method for the inference of gene circuit parameters that is two orders of magnitude faster than brute-force approaches. The inference of accurate hematopoietic GRNs depends on high temporal resolution gene expression data. Such datasets are relatively rare in the field (May et al., 2013b) but are essential for reverse engineering GRNs. Chapter 4 of this dissertation describes a high-resolution time-series dataset of genome-wide gene expression during *in vitro* macrophage-neutrophil differentiation. Chapter 5 summarizes the dissertation and discusses future directions for this research.

BIBLIOGRAPHY

Ackers, G. K., Johnson, A. D., & Shea, M. A. (1982). Quantitative model for gene-regulation by lambda-phage repressor. *Proceedings of the National Academy of Sciences USA*, *79*, 1129–1133.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell*. Boca Raton, FL: CRC Press, 4 ed.

Ali, M. A. E., Fuse, K., Tadokoro, Y., Hoshii, T., Ueno, M., Kobayashi, M., Nomura, N., Vu, H. T., Peng, H., Hegazy, A. M., Masuko, M., Sone, H., Arai, F., Tajima, A., & Hirao, A. (2017). Functional dissection of hematopoietic stem cell populations with a stemness-monitoring system based on ns-gfp transgene expression. *Scientific Reports*, *7*(1), 11442.

Aluru, S. (2005). *Handbook of Computational Molecular Biology*. Chapman and Hall /CRC.

Arya, R., & White, K. (2015). Cell death in development: Signaling pathways and core mechanisms. *Semin. Cell Dev. Biol.*, *39*, 12–19.

Basson, M. A. (2012). Signaling in cell differentiation and morphogenesis. *Cold Spring Harb. Perspect. Biol.*, *4*(6), a008151–a008151.

Becker, A. J., McCulloch, E. A., & Till, J. E. (1963). Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*, *197*(4866), 452–454.

Bertolino, E., Reinitz, J., & Manu (2016). The analysis of novel distal cebpa enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification. *Dev Biol*, *413*(1), 128–44.

Bhoopalan, S. V., Huang, L. J.-S., & Weiss, M. J. (2020). Erythropoietin regulation of red blood cell production: from bench to bedside and back. *F1000Res.*, *9*, 1153.

Bonzanni, N., Garg, A., Feenstra, K. A., Schütte, J., Kinston, S., Miranda-Saavedra, D., Heringa, J., Xenarios, I., & Göttgens, B. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics*, *29*(13), i80–8.

Bulirsch, R., & Stoer, J. (1992). *Introduction to Numerical Analysis*. New York: Springer-Verlag, second ed.

Burda, P., Laslo, P., & Stopka, T. (2010). The role of pu.1 and gata-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia*, *24*(7), 1249–1257.

Cahan, P., Li, H., Morris, S. A., Lummertz da Rocha, E., Daley, G. Q., & Collins, J. J. (2014). Cellnet: network biology applied to stem cell engineering. *Cell*, *158*(4), 903–915. 25126793[pmid].

Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P., & Abkowitz, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo. *Blood*, *117*(17), 4460–4466. 21343613[pmid].

Chen, H., Ray-Gallet, D., Zhang, P., Hetherington, C. J., Gonzalez, D. A., Zhang, D. E., Moreau-Gachelin, F., & Tenen, D. G. (1995). PU.1 (spi-1) autoregulates its expression in myeloid cells. *Oncogene*, *11*(8), 1549–1560.

Chickarmane, V., Enver, T., & Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput. Biol.*, *5*(1), e1000268.

Christensen, J. L., & Weissman, I. L. (2001). Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. *Proc. Natl. Acad. Sci. U. S. A.*, *98*(25), 14541–14546.

Collombet, S., van Oevelen, C., Ortega, J. L. S., Abou-Jaoude, W., Stefano, B. D., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017a). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences*, *114*(23), 5792–5799.

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017b). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc Natl Acad Sci U S A*, *114*(23), 5792–5799.

Crane, G. M., Jeffery, E., & Morrison, S. J. (2017). Adult haematopoietic stem cell niches. *Nature Reviews Immunology*, *17*(9), 573–590.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Rust, A. G., Pan, Z. j., Schilstra, M. J., Clarke, P. J. C., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., & Bolouri, H. (2002). A genomic regulatory network for development. *Science*, *295*(5560), 1669–78.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, *9*(1), 67–103. PMID: 11911796.

de Jong, H., & Geiselmann, J. (2005). *Modeling and simulation of genetic regulatory networks by ordinary differential equations*. Genomic Signal Processing and Statistics.

de Lucas, B., Pérez, L. M., & Gálvez, B. G. (2018). Importance and regulation of adult stem cell migration. *J. Cell. Mol. Med.*, *22*(2), 746–754.

Enver, T., Pera, M., Peterson, C., & Andrews, P. W. (2009). Stem cell states, fates, and the rules of attraction. *Cell Stem Cell*, *4*(5), 387–397.

Fehr, D. A., Handzlik, J. E., Manu, & Loh, Y. L. (2019). Classification-based inference of dynamical models of gene regulatory networks. *G3 (Bethesda)*, *9*(12), 4183–4195.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*(3-4), 601–620. PMID: 11108481.

Gilbert, S. F., & Baressi, M. J. F. (2016). *Developmental Biology*. Sinauer Associates, 11 ed.

Göttgens, B. (2015). Regulatory network control of blood stem cells. *Blood*, *125*(17), 2614–2620.

Graf, T., & Enver, T. (2009a). Forcing cells to change lineages. *Nature*, *462*(7273), 587–594.

Graf, T., & Enver, T. (2009b). Forcing cells to change lineages. *Nature*, *462*(7273), 587–94.

Gundry, R. L., Boheler, K. R., Van Eyk, J. E., & Wollscheid, B. (2008). A novel role for proteomics in the discovery of cell-surface markers on stem cells: Scratching the surface. *Proteomics Clin. Appl.*, *2*(6), 892–903.

Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K., & Göttgens, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, *114*(23), 5822–5829.

Handzlik, J. E., Loh, Y. L., & Manu (2021). Dynamic modeling of transcriptional gene regulatory networks. *Methods Mol. Biol.*, *2328*, 67–97.

Handzlik, J. E., & Manu (2022). Data-driven modeling predicts gene regulatory network dynamics during the differentiation of multipotential hematopoietic progenitors. *PLOS Computational Biology*, *18*(1), 1–31.

Häving, A. L., Windmöller, B. A., Knabbe, C., Kaltschmidt, B., Kaltschmidt, C., & Greiner, J. F. W. (2021). Between fate choice and self-renewalheterogeneity of adult neural crest-derived stem cells. *Frontiers in Cell and Developmental Biology*, *9*.

Hoppe, P. S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K. D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Gambardella, A., Ahmed, N., Etzrodt, M., Coutu, D. L., Rieger, M. A., Marr, C., Strasser, M. K., Schauberger, B., Burtscher, I., Ermakova, O., Bürger, A., Lickert, H., Nerlov, C., Theis, F. J., & Schroeder, T. (2016). Early myeloid lineage choice is not initiated by random pu.1 to gata1 protein ratios. *Nature*, *535*(7611), 299–302.

Huang, S., Guo, Y., May, G., & Enver, T. (2007a). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, *305*, 695–713.

Huang, S., Guo, Y.-P., May, G., & Enver, T. (2007b). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, *305*(2), 695–713.

Hwang, I. (2013). Cell-cell communication via extracellular membrane vesicles and its role in the immune response. *Mol. Cells*, *36*(2), 105–111.

Jaeger, J., Blagov, M., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Surkova, S., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004a). Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster. *Genetics*, *167*(4), 1721–1737.

Jaeger, J., Blagov, M., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Surkova, S., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004b). Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. *Genetics*, *167*, 1721–1737.

Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004c). Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, *430*, 368–371.

Kaldis, P. (2016). Quo vadis cell growth and division? *Frontiers in Cell and Developmental Biology*, *4*.

Kauffman, S. A. (1974). The large scale structure and dynamics of gene control circuits: An ensemble approach. *The Journal of Theoretical Biology*, *44*, 167–190.

Kiecker, C., Bates, T., & Bell, E. (2016). Molecular specification of germ layers in vertebrate embryos. *Cell. Mol. Life Sci.*, *73*(5), 923–947.

Kim, A.-R., Martinez, C., Ionides, J., Ramos, A. F., Ludwig, M. Z., Ogawa, N., Sharp, D. H., & Reinitz, J. (2013). Rearrangements of 2.5 kilobases of noncoding dna from the drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic. *PLoS Genet*, *9*(2), e1003243.

Klein, A. M., & Simons, B. D. (2011). Universal patterns of stem cell fate in cycling adult tissues. *Development*, *138*(15), 3103–3111.

Koonin, E. V. (2012). Does the central dogma still stand? *Biology Direct*, *7*(1), 27.

Ladewig, J., Koch, P., & BrΓijstle, O. (2013). Leveling waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nature Reviews Molecular Cell Biology*, *14*(4), 225–236.

Lai, X., Li, Q., Wu, F., Lin, J., Chen, J., Zheng, H., & Guo, L. (2020). Epithelial-mesenchymal transition and metabolic switching in cancer: Lessons from somatic cell reprogramming. *Frontiers in Cell and Developmental Biology*, *8*.

Lamouille, S., Xu, J., & Derynck, R. (2014). Molecular mechanisms of epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.*, *15*(3), 178–196.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., & Singh, H. (2006a). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, *126*(4), 755–766.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., & Singh, H. (2006b). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, *126*(4), 755–66.

Laurenti, E., & Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. *Nature*, *553*(7689), 418–426. 29364285[pmid].

Lennartsson, J., & Rönnstrand, L. (2012). Stem cell factor receptor/c-kit: From basic science to clinical implications. *Physiological Reviews*, *92*(4), 1619–1649. PMID: 23073628.

Losick, R., & Desplan, C. (2008). Stochasticity and cell fate. *Science*, *320*(5872), 65–68.

Ma, B., Fang, M., & Jiao, X. (2020). Inference of gene regulatory networks based on nonlinear ordinary differential equations. *Bioinformatics*, *36*(19), 4885–4893.

Magnusson, R., Mariotti, G. P., Köpsén, M., Lövfors, W., Gawel, D. R., Jörnsten, R., Linde, J., Nordling, T. E. M., Nyman, E., Schulze, S., Nestor, C. E., Zhang, H., Cedersund, G., Benson, M., Tjärnberg, A., & Gustafsson, M. (2017). Lassima network inference toolbox for genome-wide mechanistic modeling. *PLOS Computational Biology*, *13*(6), 1–19.

Man, Y., Yao, X., Yang, T., & Wang, Y. (2021). Hematopoietic stem cell niche during homeostasis, malignancy, and bone marrow transplantation. *Frontiers in Cell and Developmental Biology*, *9*.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009). Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biology*, *7*, e1000049. Doi:10.371/journal.pbio.1000049.

May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen, D. G., Peterson, C., & Enver, T. (2013a). Dynamic analysis of gene expression and

genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, *13*(6), 754–768.

May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen, D. G., Peterson, C., & Enver, T. (2013b). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, *13*(6), 754–68.

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, . P., George, & Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression (review). *World Acad Sci J*, *2*(1), 3–8.

Morrison, S. J., & Weissman, I. L. (1994). The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity*, *1*(8), 661–673.

Narula, J., Smith, A. M., Gottgens, B., & Igoshin, O. A. (2010). Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate. *PLoS Comput. Biol.*, *6*(5), e1000771.

Nelson, C. M., Khauv, D., Bissell, M. J., & Radisky, D. C. (2008). Change in cell shape is required for matrix metalloproteinase-induced epithelial-mesenchymal transition of mammary epithelial cells. *J. Cell. Biochem.*, *105*(1), 25–33.

Nerlov, C., Querfurth, E., Kulessa, H., & Graf, T. (2000). GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood*, *95*(8), 2543–2551.

Nestorowa, S., Hamey, F. K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N. K., Kent, D. G., & Göttgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, *128*(8), e20–e31. 27365425[pmid].

Nishikawa, K., Kobayashi, M., Masumi, A., Lyons, S. E., Weinstein, B. M., Liu, P. P., & Yamamoto, M. (2003). Self-association of gata1 enhances transcriptional activity in vivo in zebra fish embryos. *Mol. Cell. Biol.*, *23*(22), 8295–8305.

North, T. E., Goessling, W., Walkley, C. R., Lengerke, C., Kopani, K. R., Lord, A. M., Weber, G. J., Bowman, T. V., Jang, I.-H., Grosser, T., FitzGerald, G. A., Daley, G. Q., Orkin, S. H., & Zon, L. I. (2007). Prostaglandin e2 regulates vertebrate haematopoietic stem cell homeostasis. *Nature*, *447*(7147), 1007–1011.

Okuno, Y., Huang, G., Rosenbauer, F., Evans, E. K., Radomska, H. S., Iwasaki, H., Akashi, K., Moreau-Gachelin, F., Li, Y., Zhang, P., Göttgens, B., & Tenen, D. G. (2005). Potential autoregulation of transcription factor PU.1 by an upstream regulatory element. *Mol. Cell. Biol.*, *25*(7), 2832–2845.

Paluch, E., & Heisenberg, C.-P. (2009). Biology and physics of cell shape changes in development. *Current Biology*, *19*(17), R790–R799.

Panchal, T., Chen, X., Alchits, E., Oh, Y., Poon, J., Kouptsova, J., Laski, F. A., & Godt, D. (2017). Specification and spatial arrangement of cells in the germline stem cell niche of the drosophila ovary depend on the maf transcription factor traffic jam. *PLoS Genet.*, *13*(5), e1006790.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., & Amit, I. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, *163*(7), 1663–1677.

Pekovic, V., & Hutchison, C. J. (2008). Adult stem cell maintenance and tissue regeneration in the ageing context: the role for a-type lamins as intrinsic modulators of ageing in adult stem cells and their niches. *J. Anat.*, *213*(1), 5–25.

Pimanda, J. E., Ottersbach, K., Knezevic, K., Kinston, S., Chan, W. Y. I., Wilson, N. K., Landry, J.-R., Wood, A. D., Kolb-Kokocinski, A., Green, A. R., Tannahill, D., Lacaud, G., Kouskoff, V., & Göttgens, B. (2007). Gata2, fli1, and scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(45), 17692–17697.

Redondo, P. A., Pavlou, M., Loizidou, M., & Cheema, U. (2017). Elements of the niche for adult stem cell expansion. *J. Tissue Eng.*, *8*, 2041731417725464.

Reinitz, J., & Sharp, D. H. (1995a). Mechanism of *eve* stripe formation. *Mechanisms of Development*, *49*, 133–158.

Reinitz, J., & Sharp, D. H. (1995b). Mechanism of eve stripe formation. *Mech. Dev.*, *49*(1-2), 133–158.

Reinitz, J., & Vaisnys, J. R. (1990). Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of cooperativity. *The Journal of Theoretical Biology*, *145*, 295–318.

Rieger, M. A., Hoppe, P. S., Smejkal, B. M., Eitelhuber, A. C., & Schroeder, T. (2009). Hematopoietic cytokines can instruct lineage choice. *Science*, *325*(5937), 217–8.

Robb, L. (2007a). Cytokine receptors and hematopoietic differentiation. *Oncogene*, *26*(47), 6715–6723.

Robb, L. (2007b). Cytokine receptors and hematopoietic differentiation. *Oncogene*, *26*(47), 6715–23.

Sánchez, L., & Thieffry, D. (2001). A logical analysis of the *Drosophila* gap-gene system. *The Journal of Theoretical Biology*, *211*, 115–141.

Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C., & Huang, Y. (2017). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*, *34*(6), 964–970.

Schwab, J. D., Kühlwein, S. D., Ikonomi, N., Kühl, M., & Kestler, H. A. (2020). Concepts in boolean network modeling: What do they all mean? *Computational and Structural Biotechnology Journal*, *18*, 571–582.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., & Gaul, U. (2008). Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, *451*, 535–540.

Seita, J., & Weissman, I. L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, *2*(6), 640–653.

Serrano-Gomez, S. J., Maziveyi, M., & Alahari, S. K. (2016). Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. *Molecular Cancer*, *15*(1), 18.

Shahbazi, M. N. (2020). Mechanisms of human embryo development: from cell fate to tissue shape and back. *Development*, *147*(14). Dev190629.

Shahrabi, S., Rezaeeyan, H., Ahmadzadeh, A., Shahjahani, M., & Saki, N. (2016). Bone marrow blood vessels: Normal and neoplastic niche. *Oncology reviews*, *10*(2), 306–306. 27994770[pmid].

Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., & Ma, J. (2021). Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, *1*(7), 491–501.

Siminovitch, L., Mcculloch, E. A., & Till, J. E. (1963). The distribution of colony-forming cells among spleen colonies. *J. Cell. Comp. Physiol.*, *62*(3), 327–336.

Spangrude, G. J., Heimfeld, S., & Weissman, I. L. (1988). Purification and characterization of mouse hematopoietic stem cells. *Science*, *241*(4861), 58–62.

Sugiyama, T., Kohara, H., Noda, M., & Nagasawa, T. (2006). Maintenance of the hematopoietic stem cell pool by CXCL12-CXCR4 chemokine signaling in bone marrow stromal cell niches. *Immunity*, *25*(6), 977–988.

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–676.

Theiffry, D., Colet, M., & Thomas, R. (1993). Formalization of regulatory networks: A logical method and its automatization. *Mathematical Modelling and Scientific Computing*, *2*, 144–151.

Till, J. E., & McCulloch, E. A. (1961). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.*, *14*(2), 213–222.

Tsai, S. F., Strauss, E., & Orkin, S. H. (1991). Functional analysis and in vivo footprinting implicate the erythroid transcription factor GATA-1 as a positive regulator of its own promoter. *Genes Dev.*, *5*(6), 919–931.

Tusi, B. K., Wolock, S. L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J. R., Klein, A. M., & Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, *555*(7694), 54–60. 29466336[pmid].

Tyson, J. J., Baumann, W. T., Chen, C., Verdugo, A., Tavassoly, I., Wang, Y., Weiner, L. M., & Clarke, R. (2011). Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nat Rev Cancer*, *11*(7), 523–32.

Uzman, A. (2003). Molecular biology of the cell (4th ed.): Alberts, b., johnson, a., lewis, j., raff, m., roberts, k., and walter, p. *Biochemistry and Molecular Biology Education*, *31*(4), 212–214.

Vakulenko, S., & Grigoriev, D. (2003). Complexity of gene circuits, pfaffian functions and the morphogenesis problem. *Comptes Rendus de l'Académie des Sciences (Paris), Série I*, *337*, 721–724.

van Galen, P., Kreso, A., Wienholds, E., Laurenti, E., Eppert, K., Lechman, E. R., Mbong, N., Hermans, K., Dobson, S., April, C., Fan, J.-B., & Dick, J. E. (2014). Reduced lymphoid lineage priming promotes human hematopoietic stem cell expansion. *Cell Stem Cell*, *14*(1), 94–106.

Vaux, D. L., & Korsmeyer, S. J. (1999). Cell death in development. *Cell*, *96*(2), 245–254.

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A. G., & Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.*, *19*(4), 271–281.

Wang, Y., Tian, H., Cai, W., Lian, Z., Bhavanasi, D., Wu, C., Sato, T., Kurokawa, M., Wu, D., Fu, L., Wang, H., Shen, H., Liang, D., & Huang, J. (2018). Tracking hematopoietic precursor division ex vivo in real time. *Stem Cell Research & Therapy*, *9*(1), 16.

Wu, A. M., Till, J. E., Siminovitch, L., & McCulloch, E. A. (1968). Cytological evidence for a relationship between normal hemotopoietic colony-forming cells and cells of the lymphoid system. *J. Exp. Med.*, *127*(3), 455–464.

Yamashita, M., Dellorusso, P. V., Olson, O. C., & Passegué, E. (2020). Dysregulated haematopoietic stem cell behaviour in myeloid leukaemogenesis. *Nat. Rev. Cancer*, *20*(7), 365–382.

Yamazaki, S., Iwama, A., Takayanagi, S.-I., Eto, K., Ema, H., & Nakauchi, H. (2009). TGF-beta as a candidate bone marrow niche signal to induce hematopoietic stem cell hibernation. *Blood*, *113*(6), 1250–1256.

Zakrzewski, W., DobrzyĎski, M., Szymonowicz, M., & Rybak, Z. (2019). Stem cells: past, present, and future. *Stem Cell Research & Therapy*, *10*(1), 68.

Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K. A., Mueller, B. U., Narravula, S., Torbett, B. E., Orkin, S. H., & Tenen, D. G. (2000). PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*, *96*(8), 2641–2648.

Zhao, M., Perry, J. M., Marshall, H., Venkatraman, A., Qian, P., He, X. C., Ahamed, J., & Li, L. (2014). Megakaryocytes maintain homeostatic quiescence and promote post-injury regeneration of hematopoietic stem cells. *Nat. Med.*, *20*(11), 1321–1326.

Zou, Y. R., Kottmann, A. H., Kuroda, M., Taniuchi, I., & Littman, D. R. (1998). Function of the chemokine receptor CXCR4 in haematopoiesis and in cerebellar development. *Nature*, *393*(6685), 595–599.

CHAPTER 2

Data-driven modeling predicts gene regulatory network dynamics during the differentiation of multipotential hematopoietic progenitors

Cellular differentiation during hematopoiesis is guided by GRNs comprising TFs and the effectors of cytokine signaling. The genetic architecture-the type and strength of regulatory interconnections-and the dynamics of the majority of GRNs are still poorly understood. This chapter comes from a published study (Handzlik & Manu, 2022) presenting the inference of the architecture and the dynamics of a twelve-gene GRN including key TFs and cytokine receptors from transient gene expression patterns. Gene circuit was used as a mathematical model for the derivation of the GRN from hematopoietic progenitor cells differentiating into erythrocytes and neutrophils.

2.1 INTRODUCTION

Cell-fate decisions during hematopoiesis are thought to be made by transcriptional gene regulatory networks (GRNs) (Orkin & Zon, 2008; Laslo et al., 2008, 2006), which are comprised of genes that influence each others' expression through their products. The genetic architecture, by which we mean the regulators of genes, whether each regulator activates or represses, as well as the quantitative strength of regulation, of hematopoietic GRNs is not fully understood. Hematopoietic cell-fate choice has often been interpreted in the context of a simple network motif, the bistable switch (Huang et al., 2007; Enver et al., 2009; Laslo et al., 2006).

In the bistable switch model, two TFs repress each others' expression and cell-fate is chosen in a cell-autonomous manner when small stochastic fluctuations cause the system to shift to one of two steady states corresponding to the alternative cell fates. For example, the choice between the red- and white-blood cell fates is thought to be made by mutual repression between two transcription factors (TFs), Gata1 and PU.1 (encoded by *Spi1*) (Huang et al., 2007). Similar bistable switches have been proposed for other binary cell-fate choices in hematopoiesis (Laslo et al., 2008) and more generally in development (Graf & Enver, 2009).

A number of recent developments suggest that the bistable switch framework might be insufficient to explain cell-fate choice and that hematopoietic GRNs have a densely interconnected architecture. Network reconstructions based on genome-wide gene expression data reveal large modules of co-regulated genes (Novershtern et al., 2011) and genome-wide TF binding data show that most regulatory regions are co-bound by multiple TFs (Wilson et al., 2010a; Nègre et al., 2011). A second issue is that the bistable-switch hypothesis is anchored in a developmental sequence of discrete binary cell-fate decisions with well-defined intermediate progenitors. Single-cell RNA sequencing data imply however that cellular states during hematopoiesis are situated along a continuum and may not involve binary decisions (Velten et al., 2017; Tusi et al., 2018). Bistable switches, such as Gata1-PU.1, were inferred from genetic and biochemical analyses conducted at steady state, which lack information about the dynamics and causality of events. For instance, tracking the expression dynamics of fluorescently tagged Gata1 and PU.1 in live cells suggests that rather than initiating lineage choice, the divergent expression of

the two proteins is itself a consequence of as-yet-unknown upstream regulatory events (Hoppe et al., 2016). Finally, the cell-autonomous bistable-switch framework cannot integrate and account for the instructive influence of cytokines on hematopoietic differentiation (Mossadegh-Keller et al., 2013; Rieger et al., 2009).

Here we take an alternative approach to inferring the genetic architecture and dynamics of the red- and white-blood cell-fate decision. Our approach utilizes a data-driven predictive modeling methodology called gene circuits (Reinitz & Sharp, 1995; Fehr David A. et al., 2019). Gene circuits determine the time evolution of protein or mRNA concentrations using coupled nonlinear ODEs in which synthesis is represented as a switch-like function of regulator concentrations. The data can be derived from a wide variety of experiments, ranging from genome-wide studies of unperturbed development to narrower studies involving targeted perturbations. The values of the free parameters define the regulatory influences among the genes in the network. Gene circuits do not presuppose any particular scheme of regulatory interactions, but instead determine it by estimating the values of the parameters from quantitative data using global nonlinear optimization techniques (Chu et al., 1999; Kozlov et al., 2012; Gursky et al., 2004; Abdol et al., 2017). Gene circuits infer not only the topology of the GRN but also the type, either activation or repression, and strength of interactions. Most importantly, the inference procedure yields ODE models that can be used to interrogate the dynamics and causality of regulatory events during differentiation as well as to simulate and predict the consequences of developmental perturbations (Jaeger et al., 2004; Manu et al., 2009b,a; Wu et al., 2015).

We inferred the genetic architecture and gene regulation dynamics underlying red- and white-blood cell differentiation using gene circuit models comprising 12 genes. The gene circuits included receptors and effectors of cytokine signaling in addition to well-known lineage specifying TFs, such as Gata1 and PU.1, so that they could incorporate the potential influence of cytokines. The gene circuits were trained on publicly available high temporal resolution genome-wide gene expression data acquired during the differentiation of an inducible cell line, FDCP-mix (Huang et al., 2007; May et al., 2013), into erythrocytes and neutrophils. Most of the inferred pair-wise regulatory interactions were consistent with available empirical evidence. The models also correctly predicted the effect of knock-out, knock-down, and overexpression of key TFs both qualitatively and quantitatively. The genetic architecture of the models, instead of being hierarchical, is densely interconnected and features extensive cross-repression between genes expressed in different lineages. Furthermore, analysis of the model showed that *Spi1* upregulation occurred in the latter half of neutrophil differentiation, which was driven instead by two other TFs expressed in the neutrophil lineage, C/EBPα and Gfi1. We tested this prediction of the model by inspecting the sequence of gene upregulation during neutrophil differentiation in a single-cell RNA-seq dataset (Tusi et al., 2018) from mouse bone marrow. These data confirmed that *Cebpa* and *Gfi1* upregulation precede that of *Spi1 in vivo*.

2.2 RESULTS

2.2.1 Data-driven modeling of gene expression dynamics during the differentiation of FDCP-mix cells

2.2.1.1 Gene Circuit Models

A gene circuit (Manu et al., 2009b) computes the time evolution of mRNA concentrations of a network of interacting genes by solving the coupled ordinary differential equations (ODEs)

$$\frac{dx_i^l}{dt} = R_i S\left(\sum_{j=1}^{N} T_{ij}x_j^l + b_i c^l + h_i\right) - \lambda_i x_i^l, \tag{2.1}$$

where $x_i^l(t)$ is the concentration of the mRNA of gene i at time t in lineage (or condition) l, and N is the total number of genes in the model. The synthesis rate depends on the concentrations of a gene's regulators through sigmoidal regulation-expression function $S(u) = \frac{1}{2}\left(u/\sqrt{(u^2+1)}+1\right)$. $S(u)$ determines the fraction of the maximum synthesis rate R_i attained by the gene given the total regulatory input $u = \sum_{j=1}^{N} T_{ij}x_j^l + b_i c^l + h_i$. The first term of u, $\sum_{j=1}^{N} T_{ij}x_j^l$, represents the regulation of gene i by the other genes in the network. Positive and negative values of T_{ij} signify activation and repression of gene i by gene j respectively. The regulation of gene i by factors specific to the condition l that have not been explicitly represented in the model is described by the second term of u, $b_i c^l$, where c^l is -1, 0, or 1 for neutrophil, progenitor, and erythroid conditions respectively. The threshold h_i determines the basal synthesis rate and λ_i is the degradation rate of

mRNA for gene i. Training gene circuit models on quantitative gene expression data results in estimates of the values of these parameters. Estimates of the genetic inteconnectivity coefficients (T_{ij}) allows the inference of the genetic architecture of the GRN.

The sigmoid regulation-expression function allows the synthesis rate of a target to change with a regulator's concentration in either a gradual or a sharp manner, depending on the magnitude of the genetic interconnectivity T_{ij}. If the magnitude is small, then the synthesis rate will change gradually as the regulator's concentrations vary. If the magnitude is large, small changes in regulator concentration can lead to sharp changes in synthesis rate. In the latter scenario, sharp changes occur when the total regulatory input crosses zero and hence the regulator does not have a fixed threshold concentration, which can vary depending on the contributions of other regulators.

2.2.1.2 Specification of a gene circuit model for erythrocyte-neutrophil differentiation

We constructed a gene circuit model comprising 12 main lineage-specifying TFs and cytokine receptors implicated in erythrocyte-neutrophil differentiation. Among them, *Tal1* and *Gata2* are expressed in pluripotent stem cells and are necessary for the differentiation of multiple lineages including erythrocytes (Cantor & Orkin, 2002; Doré et al., 2012; Vicente et al., 2012; Shivdasani et al., 1995; Mikkola et al., 2003; Huang et al., 2009). *Gata1*, its partner *Zfpm1*, which encodes the Fog1 protein, and *Klf1* are necessary for erythroid and megakaryocytic differentiation (Cantor & Orkin, 2001; Laslo et al., 2008; Starck et al., 2003; Stachura et al., 2006; Porcher

et al., 1996; Mancini et al., 2012; Siatecka & Bieker, 2011). All white blood-cell lineages are absent in the bone marrow of *Spi1*$^{-/-}$ knockout mice (Scott et al., 1994), the products encoded by *Cebpa* and *Gfi1* specify the neutrophil cell fate (Zhang et al., 1997; Laslo et al., 2006), and the TF encoded by *Stat3* acts downstream of GCSF signaling (Tian et al., 1996). Previous work has suggested that the expression level of cytokine receptors can influence the activation of lineage specifying TFs (Palani & Sarkar, 2008). We included three genes, *Epor*, *Csf3r*, and *Il3ra*, encoding the cytokine receptors Epor, GCSF-R, and the alpha subunit of the IL3 receptor respectively (Robb, 2007) in order to detect such potential regulatory mechanisms. Although all of these genes are well-known participants in erythrocyte-neutrophil differentiation, the precise genetic architecture of the network remains to be determined.

While there are other genes known to be important for the specification of these cell fates, we limited the number of genes to 12 in order to minimize the risk of overfitting and to complete model fitting in a reasonable amount of time. Increasing the number of genes increases the number of free parameters in the model and these extra degrees of freedom increase the chances that the model will be overfit, resulting in poor predictive ability. With the training data used here, the 12 gene model has a three-fold excess of datapoints over free parameters, which makes overfitting unlikely.

2.2.1.3 *Time-series data for training the gene circuits*

We trained the gene circuit on May *et al.*'s high temporal-resolution dataset (May et al., 2013) of genome-wide gene expression during erythrocyte-neutrophil differentiation. May *et al*. utilized FDCP-mix cells (Stachura et al., 2006) which are maintained in a multipotent state in the presence of IL3 and can be induced to differentiate into erythrocytes or neutrophils by culturing in low IL3, Epo, and hemin or GCSF and SCF respectively. In the rest of the paper, we refer to the culture of FDCP-mix cells in low IL3, Epo, and hemin as erythrocyte conditions and culturing in GCSF and SCF as neutrophil conditions. The dataset comprises genome-wide gene expression measurements at 30 time points during the 7-day course of differentiation towards either cell fate, with sampling frequency reducing from once every two hours during the first day to once in three days during the last three days.

The trajectories of gene expression for the modeled genes (Fig. 2.1) exhibit rich temporal dynamics. Whereas the expression of some genes, such as *Klf1* and *Gfi1*, diverges between erythroid and neutrophil conditions during the first few hours of differentiation, the expression of genes such as *Il3ra*, *Gata2*, and *Spi1* does not diverge until 2-3 days into the differentiation. Besides timing, the genes differ also in the magnitude of change during the course of differentiation. Although all the genes change expression significantly over 7 days, *Il3ra* is upregulated $\sim$2-fold in the neutrophil condition while *Csf3r* is upregulated $\sim$230-fold in the neutrophil condition. With the exception of *Gata2*, the expression patterns of the remaining genes are consistent with those in murine bone marrow at a qualitative level

(SFig. 2.1). While *Gata2* is upregulated in both conditions in FDCP-mix cells, it is downregulated along both the erythrocyte and neutrophil lineages in data from bone marrow. Lastly, all genes except *Gata2* demonstrate an "either-or" pattern of regulation in FDCP-mix cells, being upregulated in one condition, while being downregulated in the other (Fig. 2.1).

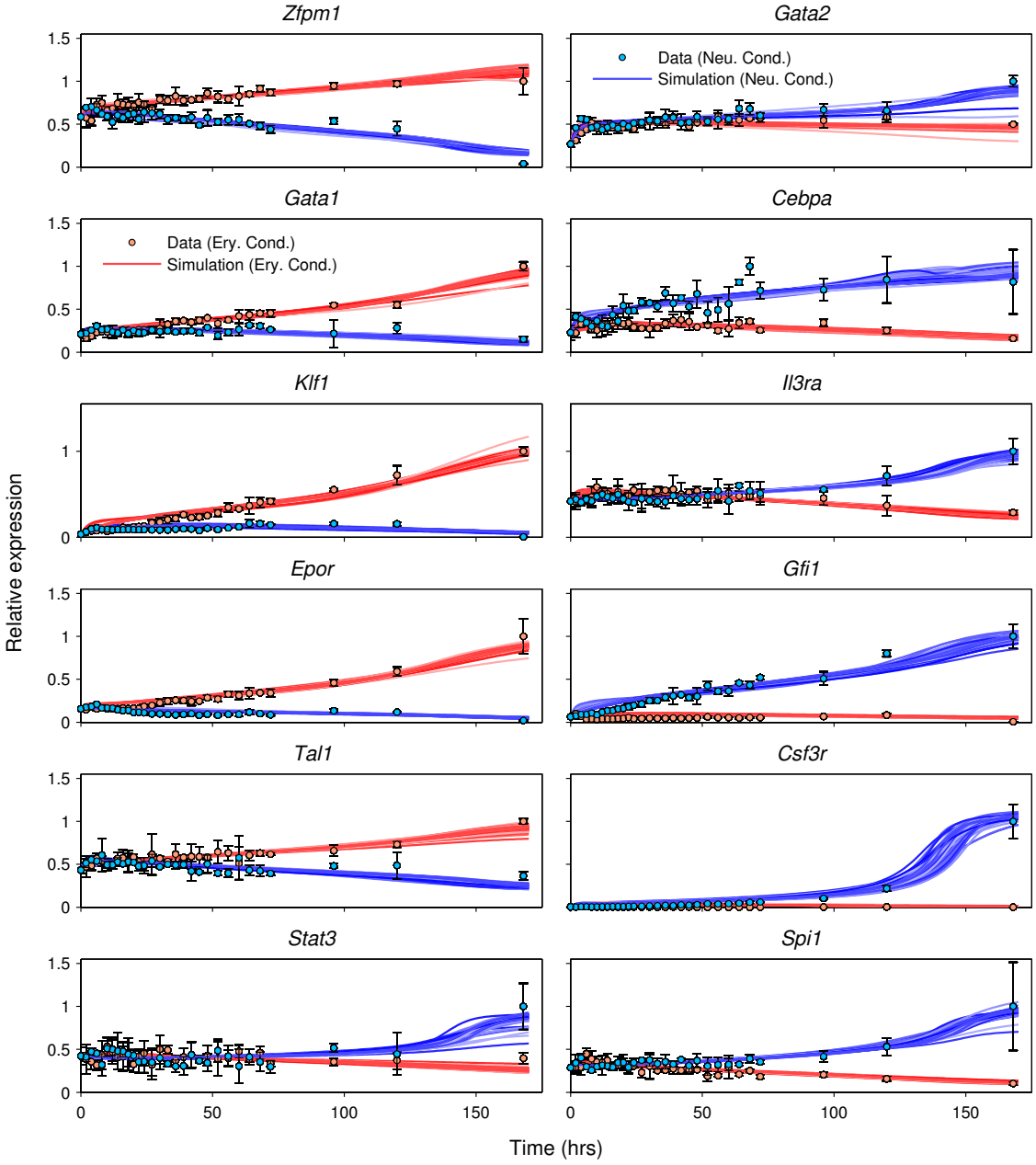


Figure 2.1: Gene expression time-series data vs. model output. Mean microarray gene expression measurements and model output for the 12 modeled genes are plotted as circles and lines respectively. Here, and in the following figures, relative expression of a gene is given by the ratio of its expression to its maximum expression across all conditions and time points. Errors bars show standard deviation over 3 replicates. The output of the 71 models that met the goodness-of-fit criteria (Section 2.4) are shown simultaneously. Data and model output for FDCP-mix cells cultured in low IL3, Epo, and hemin, referred to as the erythrocyte condition hereafter, are shown in red. Data and model output for FDCP-mix cultured in GCSF and SCF, referred to as the neutrophil condition hereafter, are shown in blue.

2.2.1.4 *Training the gene circuits on time-series gene expression data*

We trained the gene circuits on May et al.'s time-series data, with initial conditions specified by gene expression in progenitor cells, using a global nonlinear optimization method called Parallel Lam Simulated Annealing (PLSA; Chu et al., 1999; Manu et al., 2009b). PLSA is a stochastic method and results in a distinct set of parameters each time a gene circuit is inferred from data. In order to ensure that our analysis was not influenced by any idiosyncrasy of a particular model, we inferred 100 independent gene circuit models and chose 71 that met our goodness-of-fit of criteria (Section 2.4) for further analysis.

2.2.1.5 *Simulation of the GRN during FDCP-mix erythrocyte-neutrophil differentiation*

The output of the 71 analyzed gene circuits agreed with the training data within experimental error for all 12 genes and the vast majority of time points (Fig. 2.1). The sole exception was that the models did not reproduce a spike in *Cebpa* expression occurring around the 70 hour time point, although it is unclear whether this spike is genuine or the result of experimental error. Models trained on randomly shuffled data fit the data poorly (Section 2.4), implying that the fits to the empirical data are statistically significant (SFig. 2.2). We also checked how sensitive the model is with respect to perturbations in initial conditions and found that model output was robust to perturbations of up to $\pm 70\%$ (SFig. 2.3). Consistent with the general agreement with the data, the models' outputs reproduce all the essential dynamical features of the data—the either-or differential expression, gene-specific timing of expression divergence, and gene-to-gene variation in the dynamic range

of expression.

2.2.2 Gene circuits predict the consequences of genetic perturbations

Having obtained gene circuits that are able to quantitatively reproduce the observed time series data, we next tested whether the models could predict the outcomes of experimental treatments *de novo*, that is, without being trained on the data from the experiments. We simulated two kinds of experiments using the gene circuits. The first class are knockouts of *Gata1* and *Spi1*, experiments that were not carried out in FDCP-mix cells but in mice or other cell types. One should not expect the model to predict the outcomes of such knockout experiments at a quantitative level since the model was neither trained on the data from these cell types nor were all of its state variables measured in the experiments. Therefore, we compare model predictions with the results of knockout experiments at a qualitative level. The second class of experiments involved the knockdown or overexpression of key gene products followed by genome-wide expression profiling conducted by May *et al.* in FDCP-mix cells (May et al., 2013). Simulation of these perturbations may be compared to experiments at a quantitative level since they share the experimental system and all of the model's state variables were measured.

2.2.2.1 Simulation of Spi1 *and* Gata1 *knockout*

We simulated *Spi1* knockout by setting its initial expression and maximum synthesis rate to zero (Section 2.4). The consequences of this perturbation differed

by condition (Fig. 3.2). In erythrocyte conditions, although the change was more rapid in the mutant, the expression of all genes moved in the same direction and attained very similar values on day 7 as the wildtype. The model predicted therefore, that erythrocyte differentiation is largely unperturbed in *Spi1* mutants, which matches experimental observations from *Spi1* knockout mice (Scott et al., 1994). Gene expression temporal profiles differed markedly between mutant and wildtype in neutrophil conditions however, and changed very little from their initial values. A lack of change in gene expression implies that cells are arrested in a progenitor state in the *Spi1* mutant during neutrophil differentiation. This prediction is supported by the observations that *Spi1* knockout mice lack mature white-blood cells (Scott et al., 1994) and that their bone marrow contains IL3-dependent granulocyte-monocyte progenitors (GMPs) (Walsh et al., 2002; Dahl et al., 2003b), while disruption of *Spi1* in mouse granulocyte/monocyte-committed progenitors prevents their maturation but not proliferation (Iwasaki et al., 2005).

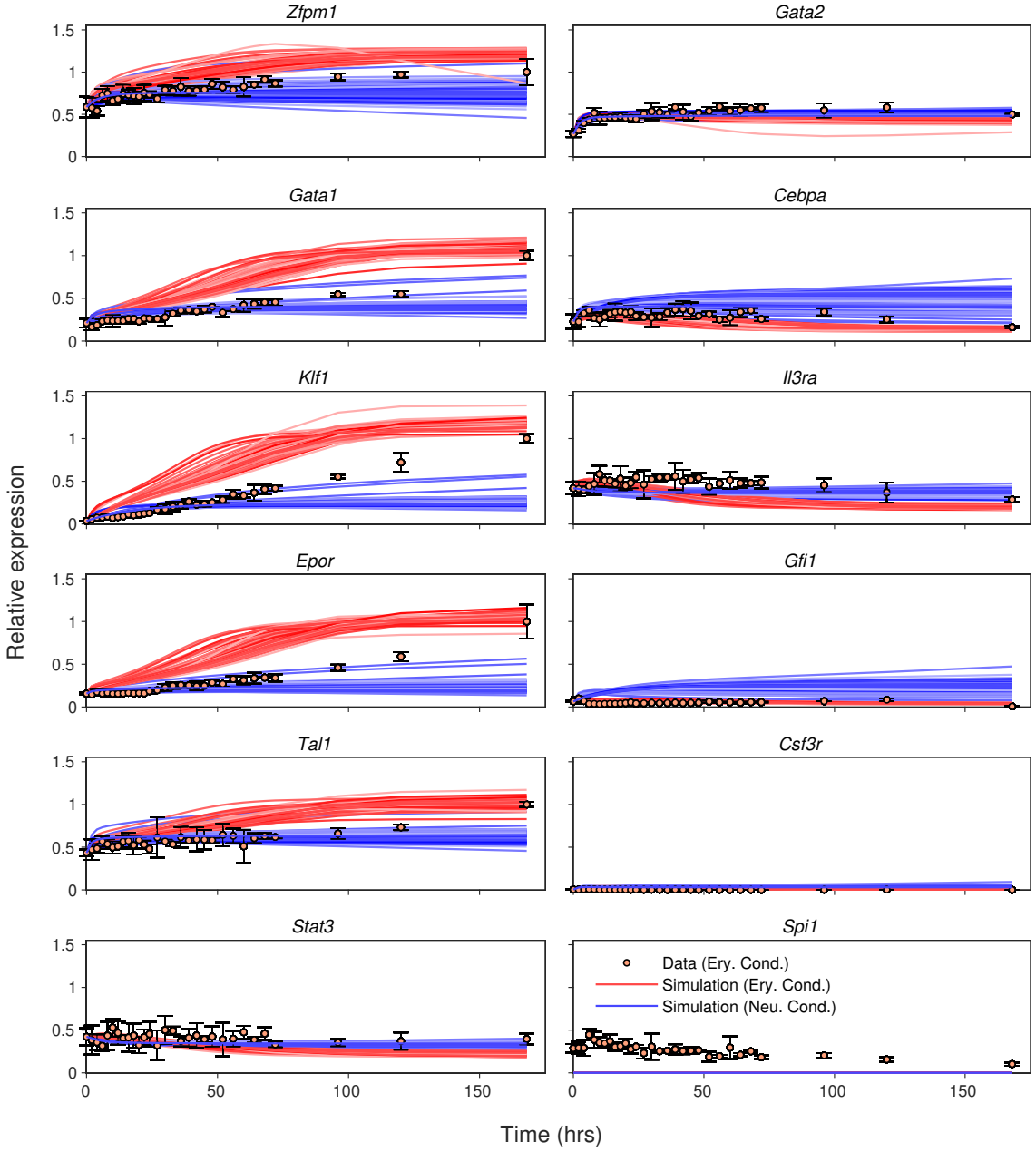


Figure 2.2: Simulation of *Spi1* knockout. *Spi1* knockout was simulated in all 71 models that met the goodness-of-fit criteria. Their output is plotted as lines. The symbols and colors are the same as Fig 2.1.

The results of *Gata1* knockout (Fig S2.4) were opposite to those of *Spi1* knockout. In neutrophil conditions, the expression of all genes changed in the same direction and reached the same endpoints as in wildtype, albeit more rapidly, implying that neutrophil differentiation is not affected by *Gata1* mutation. In erythrocyte conditions, however, gene expression of all genes did not change much from initial conditions, implying an arrest in the progenitor state. These predictions match the empirical results that embryonic stem cells (ESCs) lacking *Gata1* undergo developmental arrest at the proerythroblast stage (Weiss et al., 1994) and that *Gata1*-null ESCs cultured in the presence of Epo resemble proerythroblasts (Kitajima et al., 2006).

2.2.2.2 Simulation of knockdown and overexpression experiments in FDCP-mix cells

We simulated the knockdown of *Spi1* and *Gata2* in FDCP-mix cells and compared model output to the changes in gene expression observed in experiment (May et al., 2013). Since the knockdown was performed in self-renewing IL3 conditions, we set the lineage condition parameter to zero (Section 2.4) and simulated knockdown by reducing the synthesis rate of either gene and computing the solution until equilibrium was achieved. Since the knockdown efficiency achieved in the experiment is unknown, we set the synthesis rate to a value that results in a fold change in the expression of the targeted gene—*Spi1* or *Gata2*—that matches the empirically observed value. Therefore, we "fit" the knockdown model to the expression of the targeted gene to predict the changes in the expression of the remaining eleven genes. Finally, this analysis—and all subsequent analyses—were

performed using one representative model (model #66) out of the 71 that matched the goodness-of-fit criteria (Section 2.4).

There is strong agreement between prediction and observation for *Spi1* knockdown (Fig. 2.3). Consistent with the well-known regulatory role of PU.1, the model predicted the upregulation and downregulation of the erythrocyte and neutrophil lineage genes respectively, which matched the pattern of gene expression observed in the experiment. The only exception was *Il3ra*, which was predicted by the model to be slightly downregulated but in fact did not change in expression. In contrast to the results with *Spi1*, the model was unable to predict the consequences of *Gata2* knockdown (Fig. 2.3), suggesting that aspects of *Gata2*'s regulation were inferred poorly by model training. This is corroborated by the fact that many of the Gata2-related regulatory parameters were poorly constrained (Fig. 2.4).

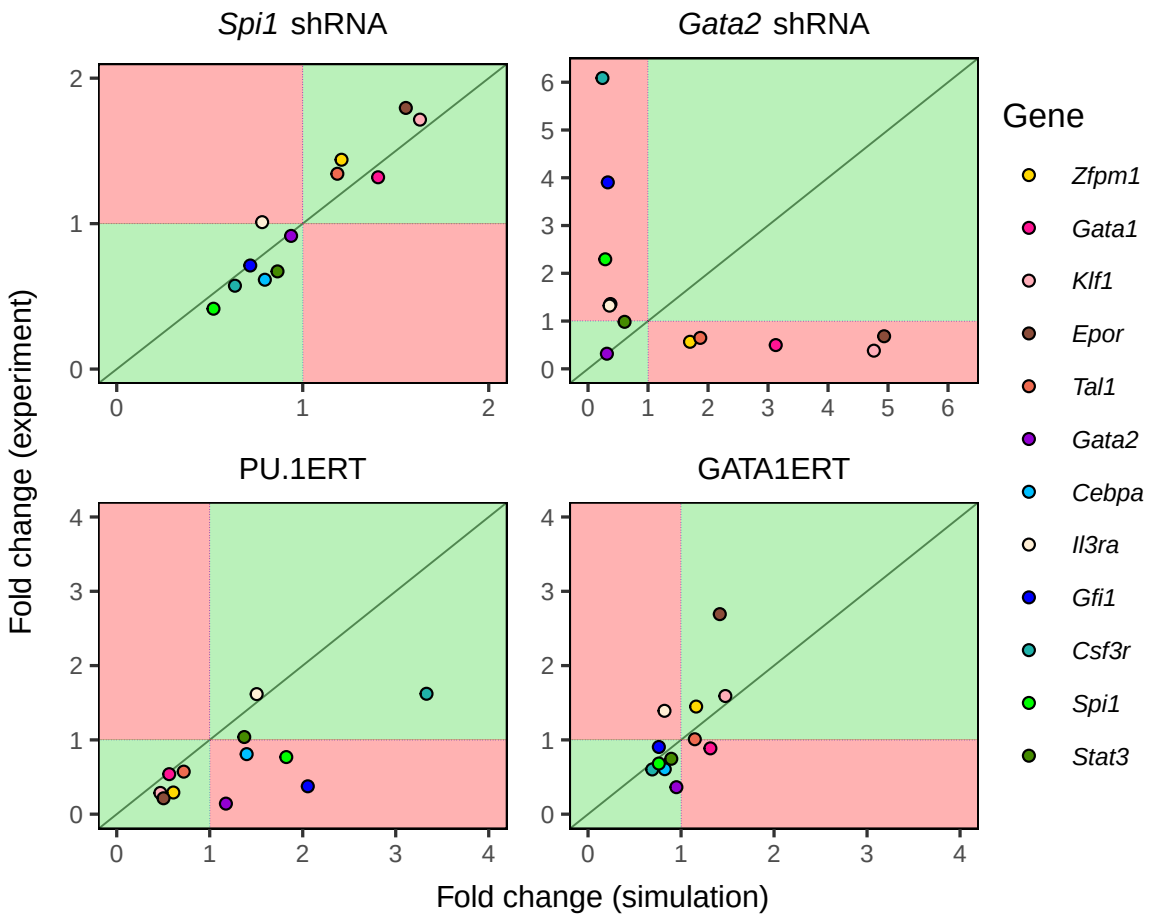


Figure 2.3: Simulation of knockdown and overexpression of key transcription factors in FDCP-mix cells. The fold change in gene expression in simulations of *Spi1* and *Gata2* knockdown (top two panels) or PU.1 and Gata1 overexpression (bottom two panels) is plotted against the fold change observed in experiment. The dotted lines correspond to no change so that points in the green quadrants indicate qualitative agreement and points in the red quadrants indicate qualitative disagreement between prediction and observation. The green line represents a perfect quantitative agreement between prediction and observation.

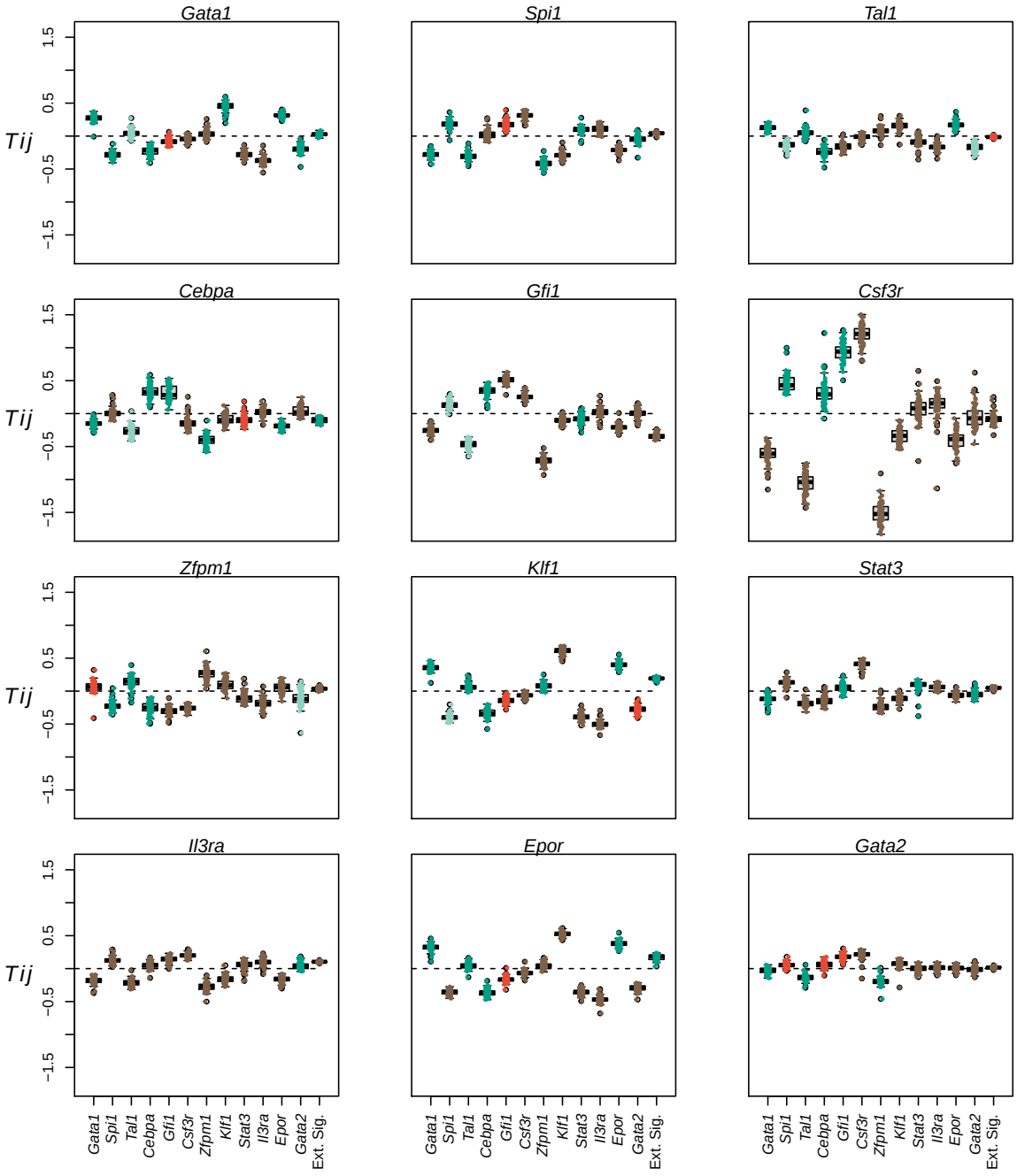


Figure 2.4: Inferred genetic architecture. The distribution of each genetic interconnectivity parameter (T_{ij}) over the ensemble of 71 models is shown as a box plot. The distribution of each regulatory parameter representing the influence of cytokine conditions (b_i) is shown as a box plot ("Ext. Sig."). In the box plots, the box lines are the first quartile, median, and third quartile. The whiskers extend to the most extreme values lying within 1.5 times the interquartile range. Individual parameter values inferred by the models are shown as circles overlaid on the box plots.

(continued)

Figure 2.4: Inferred genetic architecture (continued). Each panel shows the regulation of a particular target. Positive and negative values of Tij indicate activation and repression respectively. Positive values of bi indicate activation by Epo and repression by GCSF while negative values indicate activation by GCSF and repression by Epo. Activation is inferred if the first quartile of the distribution is positive, while repression is inferred if the third quartile is negative. The type of regulation is considered to be poorly constrained when the interquartile range spans negative and positive values. The parameters whose inferred sign agrees with prior empirical evidence (STable 2.2) are marked as dark green while those that are contradictory are marked as red. The parameters for which there is empirical evidence for interaction but the type of interaction, activation or repression, is not known are marked as light green. The parameters for which we were unable to find experimental evidence, the experiments yielded negative results, or the sign was unconstrained are marked as brown. https://doi.org/10.1371/journal.pcbi.1009779.g004

The overexpression experiments were simulated differently than knockdown experiments since the induction of the ERT fusion proteins by OHT does not change their mRNA expression directly but changes their TF activity, instead. Since the genetic interconnectivity matrix elements parameterize the activity of the TFs in gene circuits, we simulated the induction of ERT protein activity by OHT by adding a bias term to the total regulatory input of each gene. The bias term of each gene is proportional to the interconnectivity element through which the gene is regulated by the overexpressed gene (Section 2.4). Similar to the knockdown experiments, the proportionality constant is unknown and was determined by fitting the overexpression model to the expression of one of the genes. Finally, we did not fit to the expression of the overexpressed gene since the observed

mRNA includes an unknown contribution from the ERT fusion transgene.

The model was able to correctly predict the change in expression of all the genes except *Il3ra* in the GATA1ERT experiment (Fig. 2.3). The quantitative agreement between model prediction and experiment was also good with the exception of *Epor*, for which a ∼1.5-fold upregulation was predicted while a ∼3-fold upregulation was observed. In the PU.1ERT experiment, the model predicted the change in expression of all genes except *Cebpa*, *Gfi1*, and *Gata2*. Whereas the model predicted an upregulation of these genes upon PU.1 overexpression, these genes were found to be downregulated in the actual experiment. The downregulation of *Gfi1* and *Cebpa* observed in experiment is inconsistent with the known role of PU.1 as an activator of these white-blood cell lineages genes (Cooper et al., 2015; Bertolino et al., 2016; Repele et al., 2019; Wilson et al., 2010b) as well as their downregulation upon *Spi1* knockdown. This inconsistency could be the result of PU.1 overexpression promoting a macrophage gene expression program by repressing neutrophil genes indirectly via *Egr1/2* (Laslo et al., 2006) or *Irf8* (Olsson et al., 2016). The misprediction of *Cebpa* and *Gfi1* expression in PU.1ERT could, therefore, be a consequence of omitting macrophage lineage genes in the model.

2.2.3 Erythrocyte-neutrophil GRN architecture is non-hierarchical and evolves in time

Having verified that the inferred models have predictive ability, we next determined the architecture of the GRN implied by the values of the genetic interconnectivity parameters, T_{ij}. T_{ij} determines how the product of gene j regulates gene

i, where positive or negative values denote activation or repression respectively. The distributions of the majority of interconnectivity parameters across the ensemble of 71 analyzed models were well constrained and distinguishable as either activation or repression (Fig. 2.4 and Table S1). For example, the positive values of $T_{Gata1 \to Gata1}$ (Fig. 2.4A) and $T_{Spi1 \to Spi1}$ (Fig. 2.4B) in all but one model implies that both genes autoactivate while the negative values of $T_{Gata1 \to Spi1}$ (Fig. 2.4B) and $T_{Spi1 \to Gata1}$ (Fig. 2.4A) in all analyzed models implies that the two genes repress each other. We compared the inferred genetic interconnections to published empirical evidence (Fig. 2.4 and STable 2.2). The model inferred the correct role, activation or repression, for 58 of the 69 interconnections that we found empirical evidence for. The vast majority of the interconnections have not been previously examined and the model, therefore, implies novel inferences about the genetic architecture of the network.

The experimental evidence was inconclusive or conflicting in some instances (STable 2.2). Notably, the model inferred that Gfi1 activates *Spi1*, upregulated during FDCP-mix neutrophil differentiation, and represses *Gata1*, *Klf1*, and *Epor*, genes downregulated during FDCP-mix neutrophil differentiation (Fig. 2.1). These model inferences are supported by single-cell RT-qPCR data that show that *Gfi1* expression is positively correlated with *Spi1* expression in GMPs, LMPPs, and HSCs, while it is negatively correlated with *Gata1* expression in HSCs and GMPs (Moignard et al., 2013). Furthermore, Gfi1 is known to cooperate with C/EBPϵ to activate neutrophil genes (van der Meer et al., 2010; Khanna-Gupta et al., 2005). Contradicting the model's inferences and the above evidence, *Spi1* is upregulated

in MPPs from *Gfi1*$^{-/-}$ mice (Hock et al., 2003a; Spooner et al., 2009) while *Gata1*, *Klf1*, and *Epor* are downregulated in bone-marrow cells from *Gfi1*$^{-/-}$ mice (Kim et al., 2014). The conflicting evidence and lack of agreement between the model and data may be a result of the pleiotropic roles that *Gfi1* play in both HSC maintenance and neutrophil development (Hock et al., 2003b). As noted in the previous section, the regulatory parameters of *Gata2*, another gene acting pleiotropically in HSCs, the erythroid-megakaryocytic lineage, and the myeloid lineage (Cantor & Orkin, 2002; Iwasaki et al., 2006), were poorly or incorrectly inferred. These inconsistencies were, however, a small proportion of the total inferences and there is overall good agreement between model inference and empirical evidence (Fig. 2.4 and STable 2.1).

The genetic architecture of the network, in fact, changes in time since the strength of the regulation of one gene by the products of another gene depends on the concentration of the latter, which evolves during the differentiation process. In order to gain insight into this "dynamical GRN", we computed the time-dependent regulatory contribution, given by the product of the genetic interconnectivity parameters by the concentrations of the cognate regulators ($T_{ij} \cdot x_j^l(t)$), for all pairs of regulators and targets in the model. The GRN may then be represented as a graph in which each gene is a node and the type—activation or repression—and time-dependent strength of regulation between each gene pair is an edge (Fig. 2.5).

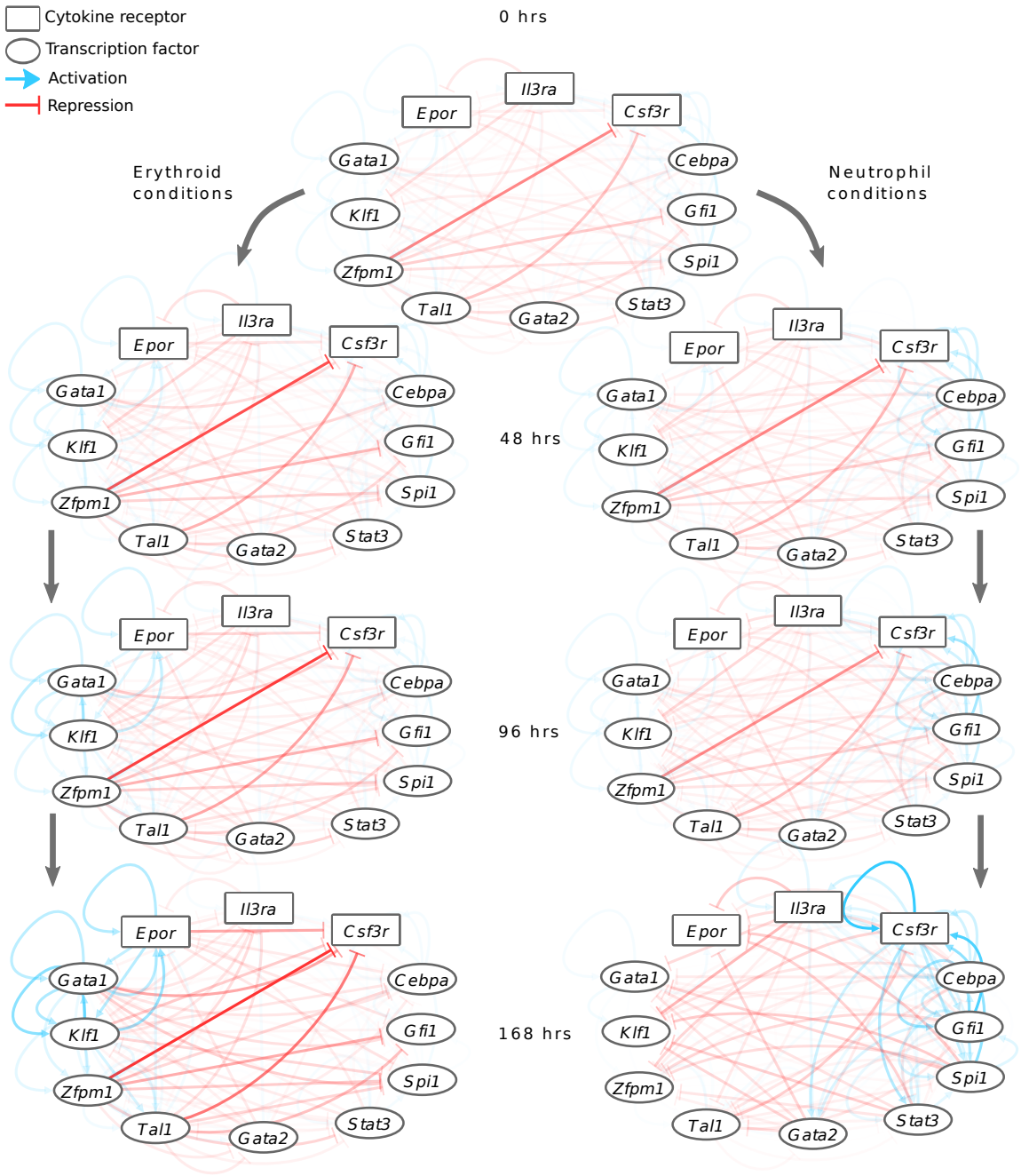


Figure 2.5: The time evolution of the inferred GRN. The GRN is depicted as a graph at different time points during differentiation in both erythrocyte and neutrophil conditions. The contributions of each regulator to the regulation of its targets, given by the product of the pairwise genetic interconnectivity parameter and the regulator's concentration, are shown as edges from the regulator to the targets. Blue and red edges correspond to activation and repression respectively, while the opacity of the lines indicates the strength of regulation. The maximum opacities of activation and repression have been normalized to 1 separately.

The erythrocyte-neutrophil network inferred by the model from FDCP-mix data is densely interconnected with genes associated with the erythrocyte lineage repressing genes of the neutrophil lineage and vice versa. This conclusion is in agreement with other analyses based on genome-wide gene expression data (Novershtern et al., 2011) and contrasts the view that the genetic architecture consists of a hierarchy of bistable switches (Laslo et al., 2008). The time evolution of the network reveals two broad principles. First, there is a preponderance of repressive interactions at earlier time points during the differentiation suggesting that the cell-fate decision is dictated by loss of repression rather than a gain of activation. Conversely, activation between co-expressed genes gains prominence at later time points, suggesting that activation mainly reinforces the decision once it has been made.

2.2.4 Gene circuits predict that C/EBPα and Gfi1 drive neutrophil development in FDCP-mix cells

How each gene in the network is regulated is, as discussed earlier, not static but changes as the concentrations of its regulators evolve in time during differentiation. We reasoned that the temporal dynamics of gene regulation could provide insight into the causality of the regulatory events underlying differentiation. The temporal dynamics of gene regulation can be analyzed by "looking under the hood" of the gene circuit model and decomposing the total regulatory input for each gene into contributions from individual regulators (Fig. 2.6; see Section 2.4 for details). In Figure 2.6, the total regulatory input (dotted black line) is plotted

in time. A gene is at half its maximum activation when the total regulatory input is zero and thus the time at which this happens (black vertical lines) serves as a marker to order the sequence in which genes turn on or off as differentiation proceeds. The contributions of repressors and activators are shown as shaded sections above and below the total regulatory input respectively. The regulators accounting for the up- or down-regulation of a gene can be determined by noting their contribution to the change in the total regulatory input. For example, the bulk of the change in *Cebpa*'s regulatory input from the start of neutrophil differentiation to reaching half-max expression is the result of autoactivation (light blue) and activation by Gfi1 (dark blue; Fig. 2.6).

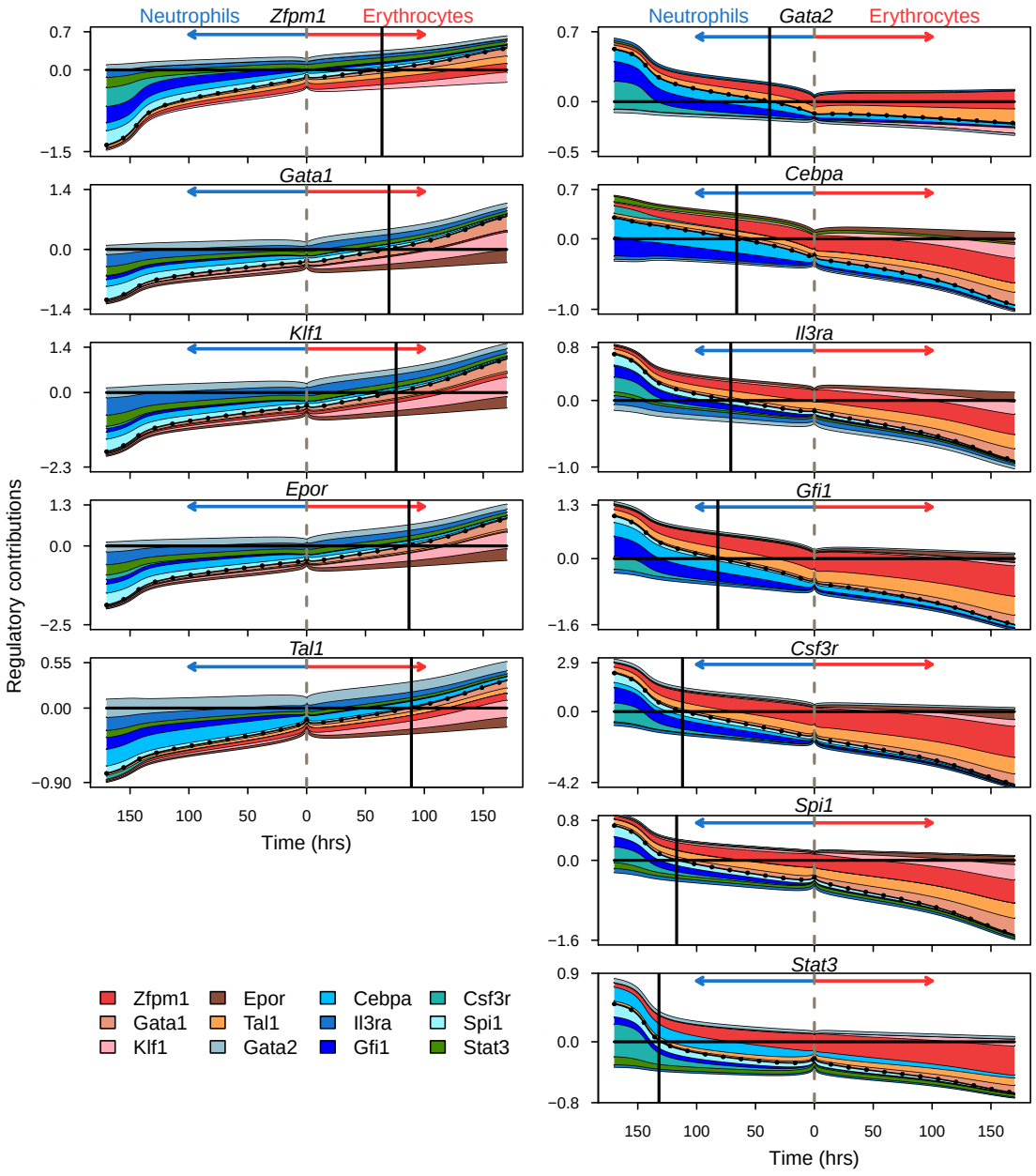


Figure 2.6: The dynamics of gene regulation during differentiation. The total regulatory input (u) is plotted as the dotted black line. The colored layers show the regulatory contribution of individual regulators. See Section 2.4 for the definitions of total regulatory input and regulatory contributions. The contributions of repressors and activators are shown above and below the dotted line respectively. The vertical dashed line in the center corresponds to uninduced FDCP-mix cells at the start of differentiation.

(continued)

Figure 2.6: The dynamics of gene regulation during differentiation (continued). Regulatory contributions during erythrocyte and neutrophil differentiation are shown to the right and left of the dashed line respectively. The vertical black line marks the time when the total regulatory input crosses zero so that synthesis occurs at half its maximum rate (Section 2.4).

Several observations can be made regarding the temporal dynamics of gene regulation during erythrocyte-neutrophil differentiation (Fig. 2.6). All the genes are in a partially repressed state, since the negative contribution from repressors is greater than the positive contribution from activators, in undifferentiated FDCP-mix cells. This is reminiscent of multilineage transcriptional priming (Laslo et al., 2006; Chickarmane et al., 2009; Paul et al., 2015)—the low-level expression of genes from multiple lineages in multipotential progenitors. What accounts for the repression varies by the target gene. Genes downregulated in neutrophil conditions, *Gata1*, *Zfpm1*, *Klf1*, *Tal1*, and *Epor*, are repressed by several genes of small effect. Genes downregulated in erythrocyte conditions, however, *Spi1*, *Cebpa*, *Gfi1*, *Stat3*, *Gata2*, *Il3ra*, and *Csf3r*, are mainly repressed by a combination of *Zfpm1* and *Tal1*.

During erythrocyte differentiation, all the upregulated genes are activated more or less simultaneously since they reach half-max activation in a short $\sim$30 hour window (Fig. 2.6). Upregulation of the genes involves both the loss of repression as well as increased activation (Fig. 2.6). The three main activating influences are *Gata1*, *Klf1*, and *Epor*. The first two are well-known activators of erythrocyte genes, while the activating influence of *Epor* implies that upregulation of the re-

ceptor's gene expression provides positive feedback, indirectly, to the TFs driving erythroid differentiation.

In contrast to erythrocyte differentiation, the sequence of activation of genes during neutrophil differentiation is spread out over ∼100 hours (Fig. 2.6). Surprisingly, *Spi1* is one of the last genes in the activation sequence, reaching half-max activation around day 5 of the differentiation process, while *Gata2* and *Cebpa* are the first ones to be activated. Unlike erythrocyte differentiation, during which three activators provided activation throughout the process, the genes accounting for activation change in time and with the target gene.

PU.1 provides activation during the later stages of the differentiation once it has increased in expression. This is consistent with the observations that PU.1 acts primarily in a concentration-dependent manner (Huang et al., 2007; DeKoter & Singh, 2000; Dahl et al., 2003b) and that conditional *Spi1* knockout in adult bone marrow does not eliminate granulopoiesis but instead results in the development of immature granulocytes (Dakic et al., 2005). Csf3r also provides activation at late time points. *Cebpa* and *Gfi1* together account for most of the early activation of the genes upregulated during neutrophil differentiation in FDCP-mix cells (Fig. 2.6). Although *Gfi1* expression is positively correlated with genes upregulated during neutrophil differentiation in FDCP-mix cells (Fig. 2.1) and with *Spi1* in GMPs (Moignard et al., 2013), Gfi1 is known to function primarily as a repressor in MPPs and the lymphoid and myeloid lineages (van der Meer et al., 2010; Hock et al., 2003a; Spooner et al., 2009; Dahl et al., 2003a). The activation role inferred here for Gfi1 during the neutrophil differentiation could result from indirect reg-

ulation of its targets. Another reason for the Gfi1 acting as activator is the high level of similarity between the expression of *Cebpa* and *Gfi1* in the training data (Fig. 2.1) that renders the two factors interchangeable in the model. C/EBPα is known to directly activate itself, *Spi1*, *Csf3r*, and *Gfi1* during neutrophil differentiation (STable 2.1; Leddin et al., 2011; Ma et al., 2014; Legraverend et al., 1993; Cooper et al., 2015; Bertolino et al., 2016; Repele et al., 2019; Zhang et al., 1997; Smith et al., 1996). We conclude therefore that the activation of neutrophil targets by Gfi1 inferred by the model could, in fact, represent the activity of C/EBPα. Taken together this analysis implies that neutrophil development in FDCP-mix cells is driven by C/EBPα and potentially Gfi1 acting indirectly (Li et al., 2010), which activate *Spi1* at later time points.

2.2.5 *Cebpa* and *Gfi1* expression precedes *Spi1* upregulation in the neutrophil lineage in mouse bone-marrow hematopoietic progenitor cells

We sought confirmation of the sequence of gene activation implied by our model of FDCP-mix cell differentiation in an independent experimental system. We analyzed Tusi *et al.*'s single-cell RNA-seq (scRNA-Seq) data from Kit$^+$ mouse bone-marrow HPCs. Although scRNA-Seq data are static snapshots of the progression of cell states during steady-state hematopoiesis, it is possible to infer the order of cell states under a few assumptions. Weinreb *et al.* developed Population Balance Analysis (PBA) (Weinreb et al., 2018b), which computes the probability of transitions between the cell states—defined by genome-wide gene expression—observed in single-cell gene expression data and hence the probability that an

intermediate cell state will evolve into some terminal cell fate (Fig. 2.7A). Cell states corresponding to multipotential progenitors—the origin of the differentiation process—and committed unilineage progenitors—the termini of the differentiation process—are identified by the expression of marker genes. PBA assumes that there are no oscillations in cellular state so that the dynamics are governed by a potential function of cellular state and cells always move from higher to lower potential (Weinreb et al., 2018b, Fig. 2.7C). Under this assumption, it is possible to order the cells in developmental time by arranging them in order of decreasing potential (see Weinreb et al., 2018b, for details).

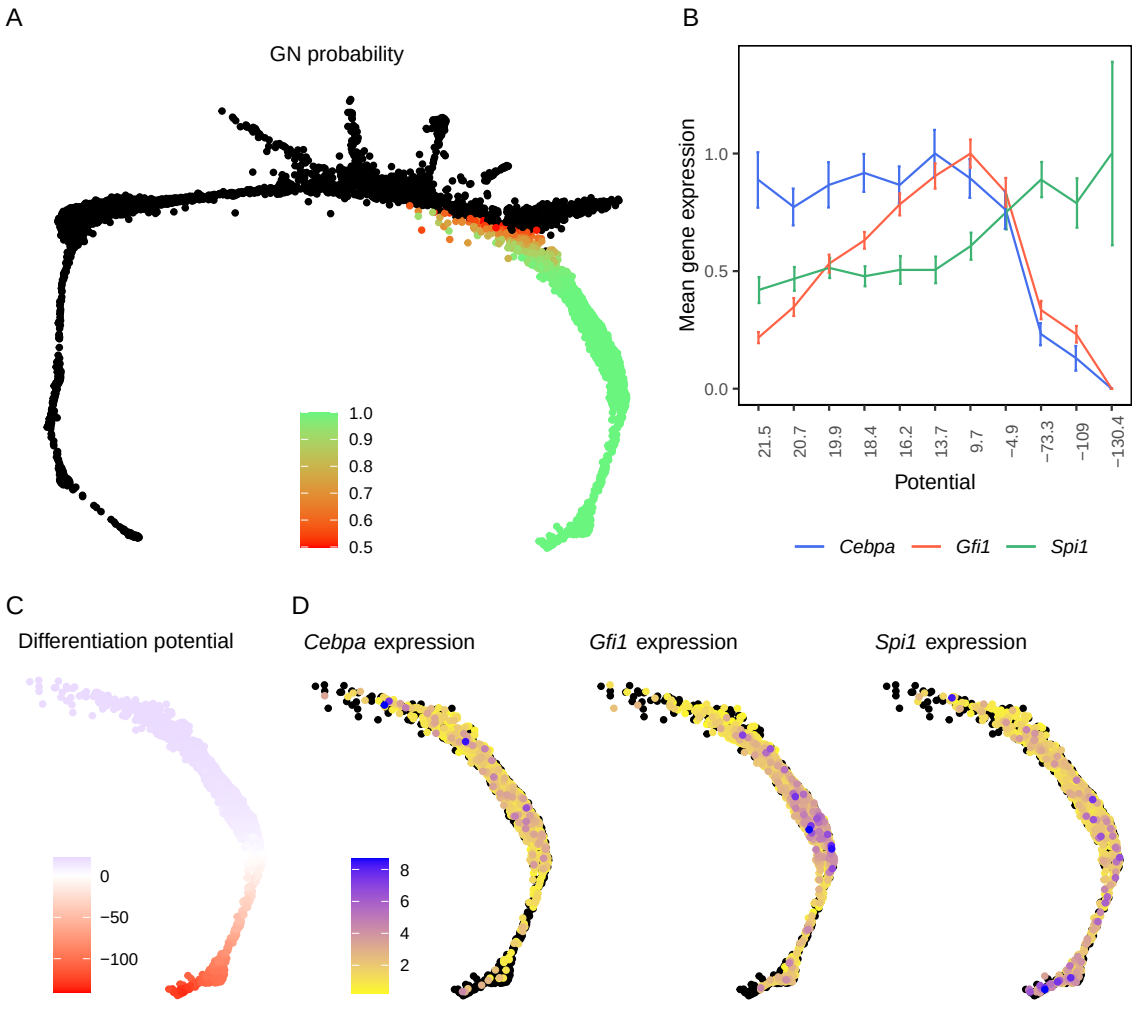


Figure 2.7: The expression of *Cebpa*, *Gfi*, and *Spi1* in individual hematopoietic progenitor cells from murine bone marrow. Panels **A**, **C**, and **D** are SPRING plots (Weinreb et al., 2018a) of Tusi *et al.*'s scRNA-Seq dataset (Tusi et al., 2018) of mouse bone-marrow derived Kit$^+$ progenitors. Each point corresponds to an individual cell and cells are arranged as a k-nearest-neighbor (knn) graph according to their pairwise distances in gene expression space (Weinreb et al., 2018a). **A**. The probability of a cell adopting the neutrophil fate, as computed by the PBA algorithm, is shown as a color map if the probability is greater than 0.5. Cells with neutrophil probability less than 0.5 are shown as black dots. **B**. The mean expression of *Cebpa*, *Gfi1*, and *Spi1* in cells binned according to their potential (shown in panel C). Each bin contains 141 cells. The expression of each gene has been normalized relative to its maximum expression over the bins. The error bars show standard error. **C**. The potential landscape of the cells fated to be neutrophils is shown as a color map and orders the cells according to their maturity or developmental age. **D**. The expression of *Cebpa*, *Gfi1*, and *Spi1* is shown as a color map. Cells with no detected transcripts are plotted in black.

We profiled the expression of *Cebpa*, *Gfi1*, and *Spi1* in Tusi *et al.*'s dataset by identifying cells having a high probability of becoming neutrophils based on the fate probabilities assigned to them by PBA (Fig. 2.7A). The potential decreases with increasing neutrophil probability (Fig. 2.7B) and it is possible to visualize how gene expression changes with developmental age at a single-cell level (Fig. 2.7D) by following the direction of decreasing potential. Since single-cell read counts have considerable cell-to-cell variability, we also divided the potential into 11 bins containing an equal number of cells and averaged the expression over the cells in each bin (Fig. 2.7B). *Spi1*, although expressed at lower levels at the earlier stages, changes relatively little until bin 6. *Spi1* is upregulated subsequently and reaches its maximum expression in bin 9 and maintains that level until the latest stage captured in this dataset. This temporal progression of *Spi1* expression is consistent with the patterns observed through live imaging of the PU.1 protein—it is expressed at low levels in HSCs and is upregulated during myeloid differentiation (Hoppe et al., 2016)—and the differentiation of FDCP-mix cells into neutrophils (Fig. 2.1).

The scRNA-Seq data also show that *Cebpa* and *Gfi1* expression precedes the granulocyte-specific upregulation of *Spi1* (Fig. 2.7B,D). *Cebpa* is already at its maximum level at the highest potential or earliest developmental stage. *Gfi1* rises rapidly at earlier stages and peaks at bin 7. *Spi1* levels in bin 10 are greater than in bin 6 (*Cebpa* peak; Welch's one-sided two-sample t-test $p = 0.004$) or bin 7 (*Gfi1* peak; $p = 0.04$). Interestingly, both *Cebpa* and *Gfi1* are downregulated to lower levels in the latest developmental stages. These inferred temporal patterns of gene

expression during the granulocytic differentiation of bone-marrow HPCs are consistent with our model's predictions that *Cebpa* and *Gfi1* are expressed earlier than and activate *Spi1* during neutrophil development.

2.3 DISCUSSION

Despite our knowledge of the main genes effecting hematopoietic cell-fate decisions, their genetic architecture as well as the causality of their regulation is not fully understood. Here we have taken the approach, complementary to empirical genetic analyses, of learning the genetic architecture by training gene circuit models on gene expression time-series data. We trained a comprehensive model comprising 12 genes encoding TFs and cytokine signaling components on a high-temporal resolution dataset (May et al., 2013). The correct predictions of the consequences of genetic perturbations at a quantitative level support the biological accuracy of the model. Similarly, we demonstrated through a detailed comparison with literature that the model correctly inferred the nature, activation or repression, of most known pairwise interactions. Our analysis implies that the genetic architecture of the erythrocyte-neutrophil decision is non-hierarchical and highly interconnected. There are extensive repressive interactions between genes from alternative lineages, while there is positive feedback from cytokine receptors. Furthermore, the gene circuit approach goes beyond static GRNs, and reveals their dynamics during the FDCP-mix cell differentiation process. We found that repressive interactions dominate at the earliest stages of the cell-fate decision while activation gains importance only at later stages. Finally, we show through

model analysis followed by validation in an independent scRNA-seq dataset (Tusi et al., 2018) that *Cebpa* and, possibly, *Gfi1* contribute to neutrophil development by upregulating *Spi1* and other downstream genes.

Hematopoietic cell-fate decisions have been modeled by two main approaches so far. In the first approach, the GRN is modeled using ODEs (Huang et al., 2007; Laslo et al., 2006; Li & Wang, 2013; Hong et al., 2012) and the quantitative values of parameters are fixed by an exhaustive search of the parameter space to find regions that reproduce the qualitative behavior of the GRN. Such models have been mostly limited to 2–3 well-known "master" regulators, perhaps due to their relatively high computational expense. The second approach circumvents the high computational expense of ODEs by constructing logical or Boolean models that are more comprehensive and include 11–20 genes (Bonzanni et al., 2013; Collombet et al., 2017). The two approaches are similar in that the genetic architecture implemented by the models is based on prior empirical evidence.

The gene circuits built here differ from previous bistable-switch models in a number of ways. First, while bistable-switch models are constructed assuming a certain genetic architecture—mutual repression between two genes and autoactivation—gene circuits do not impose any interaction scheme beforehand but instead learn it from data. Gene circuits therefore offer an independent means of decoding the genetic architecture to supplement, but also to potentially refine, what we know from purely empirical approaches. The utility of this is illustrated by the fact that the gene circuits independently inferred the mutual antagonism between PU.1 and Gata1 and autoactivation of each gene that is baked into bistable-switch

models, while diverging from them in also inferring that other factors, such as *Cebpa*, contribute to *Spi1* upregulation. Second, the gene circuits constructed here are more comprehensive, simulating a GRN of 12 genes compared to previous much smaller models (Huang et al., 2007; May et al., 2013; Chickarmane et al., 2009) without resorting to Boolean networks that assume that gene expression is restricted to a few discrete levels. Third, while previous models and gene circuits differ in the precise switch-like function employed, this difference is unlikely to matter since the parameter values are inferred by fitting.

Analysis of gene regulation dynamics in the model followed by validation in an independent dataset (Tusi et al., 2018) led us to the insight that *Cebpa* and *Gfi1* are upregulated earlier than *Spi1* and drive the activation of *Spi1* and other neutrophil genes in FDCP-mix cells. *Spi1* has been thought to reside at the top of the hierarchy (Cantor & Orkin, 2002; Laslo et al., 2006, 2008; Huang et al., 2007; Graf & Enver, 2009; May et al., 2013) of white-blood cell genes since *Spi1* knockout mice lack all white-blood cells (Scott et al., 1994). Additionally, evidence that PU.1 inhibits Gata1 (Zhang et al., 2000) and vice versa (Nerlov et al., 2000) led to a model in which *Gata1* and *Spi1* form a bistable switch that decides the fate, while all the other genes are downstream targets of Gata1 or PU.1 (Graf & Enver, 2009). However, the causal role of Gata1 and PU.1 in erythro-myeloid differentiation has been questioned recently by experiments in which Gata1 and PU.1 expression was monitored in differentiating HSPCs (Hoppe et al., 2016). These experiments failed to detect an intermediate stage where cells co-expressed low amounts of both Gata1 and PU.1, which is a necessary condition for the fate deci-

sion to be driven by the genes' mutual repression. Furthermore, in cells destined for a myeloid fate, PU.1 was expressed at a constant level before being upregulated during the later stages of commitment while Gata1 remained undetectable throughout. This observation suggested that some factor or factors other than Gata1, unknown heretofore, drive PU.1 upregulation during myeloid differentiation. Our analysis therefore implicates *Cebpa* and, potentially, *Gfi1* as candidate upstream factors driving PU.1 upregulation during myeloid differentiation.

The activation of *Spi1* by *Cebpa* inferred here helps provide a link in the chain of causation leading to neutrophil maturation during FDCP-mix cell differentiation. The upregulation of *Spi1* is discernable only $\sim$50 hours after GCSF treatment and reaches its peak on day 7 (Fig. 2.1), which is consistent with the pattern observed in mouse bone marrow (Fig. 2.7). *Cebpa* is known to be upregulated by GCSF treatment (Scott et al., 1992; Dahl et al., 2003b; Bertolino et al., 2016; Repele et al., 2019). The C/EBPα protein is phosphorylated downstream of GCSF signaling (Jack et al., 2009) and autoactivates *Cebpa* transcription by binding to its promoter (Legraverend et al., 1993) and enhancers (Cooper et al., 2015; Bertolino et al., 2016; Repele et al., 2019). *Cebpa*, therefore, is a direct target of GCSF signaling, gets upregulated soon after GCSF treatment and activates *Spi1* subsequently. The late upregulation of *Spi1* could be reconciled with its mutant phenotype—the absence of all white-blood cells—if it were necessary for the activation of all white-blood cell genes, including those characteristic of neutrophil function. *Spi1* could then be seen as a hub which integrates input from lineage-specifying genes such as *Cebpa* and coordinates the expression of downstream functional genes.

Gene regulation during differentiation is dynamic; the contributions of the regulators modulating a gene's transcription and the overall balance of activation and repression change as the regulators' concentrations vary in time. Gene circuits, being dynamical models, allow us to determine how regulatory control varies in time both at the level of individual target genes (Fig. 2.6) and more broadly at the network level (Fig. 2.5). Our analysis indicates that, both at the individual and global levels, repression dominates over activation at earlier stages of erythrocyte-neutrophil differentiation. As a result, all the genes in the network are partially repressed and expressed at low levels in progenitors. The data support this inference. Each gene in the network is upregulated by at least two-fold in one lineage or the other (Fig. 2.1), which implies that the expression level observed in the progenitors is significantly below that of an actively transcribed gene.

The predominance of repression in the earlier stages implies, in turn, that the divergence of gene expression during differentiation is driven by relief of repression rather than by activation. This is similar to the idea of lineage priming (Hu et al., 1997; Laslo et al., 2006; Huang et al., 2007; Yoshida et al., 2010; Chickarmane et al., 2009; Chang et al., 2008) in the bistable switch model (Huang et al., 2007; Laslo et al., 2006), where genes from alternative lineages are expressed at low levels and repress each others' expression in progenitors. Our model differs from the bistable switch model in two ways. First, whereas cell fate is selected by the initial concentrations of the two genes in the bistable switch model, cytokines select the fate by exerting asymmetric effects on each gene in the gene circuits modeled here (Section 2.2). The second difference is that many more genes participate in cross-

antagonism than Gata1 and PU.1 as hypothesized in the bistable switch model.

The overall balance shifts in favor of activation at later stages of differentiation, leading to the establishment of positive feedback loops between genes co-expressed in the same lineage. Of note is the activation of lineage-specific TFs by cytokine receptors. In the model, *Csf3r*, which codes for GCSFR, provides substantial activation to most of the genes upregulated in the neutrophil condition, while *Epor* performs a similar function in the erythrocyte condition (Fig. 2.6). As discussed above, *Cebpa* is known to be downstream of GCSF signaling as are other myeloid TFs (Tian et al., 1996). Similarly, EpoR phosphorylates and activates Gata1 through the PI3K-AKT pathway (Zhao et al., 2006) and Epo signaling positively regulates several erythroid genes (Rogers et al., 2008; Deindl et al., 2014; Chiba et al., 1993; Rogers et al., 2012). Cytokine receptor-mediated positive feedback has been shown to generate bistability in a model of Epo-dependent *Gata1* activation (Palani & Sarkar, 2008), resulting in greater sensitivity to Epo cytokine concentration. The positive feedback loops inferred in this bigger GRN might also result in bistability or multistability and sharp responses to cytokine concentration, a possibility that awaits confirmation through non-linear stability analysis (Hirsch et al., 2004).

Despite its general success in predicting the consequences of genetic perturbations, the model was unable to do so for *Gata2* knockdown (Fig. 2.3) implying that *Gata2*-related inferences are incorrect for both FDCP-mix and *in vivo* differentiation. The model predicted nearly the exact opposite of the observed effects. The neutrophil lineage genes were predicted to be downregulated about two-fold,

when in fact they were upregulated 1.2–4 fold, while erythrocyte lineage genes were predicted to be upregulated instead of being downregulated about two-fold (Fig. 2.3). These mispredictions may be traced to the fact that Gata2-related parameters were not inferred with much certainty during fitting. 4 of 12 of the interconnectivity parameters (T_{ij}) where Gata2 is the regulator and 4 of 12 of the interconnectivity parameters where *Gata2* is the target are indistinguishable from zero among the gene circuits that met goodness-of-fit criteria (Fig. 2.4 and STable 2.1). This implies that the goodness-of-fit was insensitive to the type, activation or repression, of those interconnections. The uncertainty about how Gata2 regulates its targets and how it is regulated itself likely arises from the fact that there is almost no divergence in *Gata2* expression between the erythrocyte and neutrophil conditions (Fig. 2.1), with differences discernible only at one time point out of thirty. The lack of different patterns of expression in the two conditions means that the *Gata2* data do not bear sufficient information to constrain *Gata2*'s regulatory parameters. Similarly, some of the inferences, such as the activation of *Spi1* and repression of *Gata1*, *Epor*, and *Klf1* by Gfi1 (Fig. 2.4 and STable 2.2), that did not match empirical data probably resulted from a lack of training data from MPPs, monocytes, and lymphocytic progenitors, where Gfi1 exerts the experimentally observed effects (Hock et al., 2003b; Li et al., 2010; Kim et al., 2014). This limitation of the gene circuit methodology—that the training dataset may not contain sufficient information to accurately infer certain regulatory parameters—may be overcome by experimental designs that either sample differentiation trajectories in a larger number of conditions and cell types or after genetic perturbations.

In gene circuits, the interconnection between a pair of genes can represent both direct and indirect regulation of one by the other. Furthermore, gene circuits as implemented here do not include higher-order interactions such as the regulation of targets by a Fog1-Gata1 complex. These design choices have both advantages and disadvantages. On the one hand, this flexibility leads to inferred GRNs that are not completely specified mechanistically. We could not hope to delineate GRNs with biochemical details relying exclusively on gene circuits. On the other hand, this very flexibility also makes predictive modeling of GRN dynamics feasible. Although biochemically detailed models of intracellular signaling (Palani & Sarkar, 2008) and gene regulation (Bertolino et al., 2016; Repele et al., 2019) have been constructed for individual pathways and enhancers, it is currently not possible to model multiple signaling pathways or the gene regulation of multiple genes simultaneously. The challenges involved in constructing comprehensive but biochemically detailed models are many; the components are yet to be completely delineated, it is impractical to measure all the biochemical parameters, learning them from data leads to highly underdetermined problems, and the computational cost of such models would be prohibitive. Gene circuits, by coarse-graining much of the biochemical detail allow the construction of more complete models that are predictive in spite of a lack of biochemical detail.

The gene circuits derived here, being deterministic models trained on bulk gene expression data, are unable to account for stochasticity in gene expression or the effects of cellular heterogeneity in FDCP-mix populations. These limitations could potentially result in erroneous inferences of two types. First, the model may

be overfit to the average initial conditions so that actual initial concentrations in single cells result in qualitatively different outcomes. Although we ensured that the inferred models were not fragile to errors of up to 70% in the mean expression of individual genes, it is possible that the inferred models produce non-biological outcomes in the presence of errors in the expression of multiple genes. Second, it is possible that the observed changes in averaged gene expression are not the result of gene expression modulation in single cells but that of changes in the sizes of phenotypically distinct subpopulations. Population heterogeneity could therefore lead to incorrect inferences about gene regulation. Most of the connections inferred here likely have a sound basis since a large proportion of them agreed with genetic and biochemical manipulations that are not confounded by cellular heterogeneity (Fig 2.4 and STable 2.2). Furthermore, the handful of genes whose expression has been monitored live are clearly regulated at a single cell level (Hoppe et al., 2016; Rieger et al., 2009; Mossadegh-Keller et al., 2013). Single-cell RNA-Seq data (Tusi et al., 2018) also support the view that gene expression is changing at a single cell level and not as a result of varying proportions of admixed cellular subpopulations. This evidence does not rule out more complex scenarios where both single-cell and population-level processes contribute to the observed changes in mean gene expression and stochastic models trained on single-cell data would be necessary to uncouple these effects.

Our results show that the temporal dynamics of gene expression bear information about the genetic architecture underlying cell-fate choice. With a few exceptions such as the segmentation system of *Drosophila* (Surkova et al., 2008), our

current knowledge of the genetic architecture of most developmental systems is based on genetic analyses carried out at end points. Coupling gene circuits with high temporal resolution time series data is a viable complementary approach to decode the genetic architecture and reveal the causality of events during differentiation. One potential drawback of this approach is the cost of sequencing. However, the cost of sequencing is expected to decline exponentially over time (Muir et al., 2016) and is not likely to be a limitation in the future. Another concern is the high computational cost of fitting the gene circuits, which entails the use of parallel computers. This challenge was recently overcome by an algorithm called Fast Inference of Gene Regulation (FIGR) (Fehr David A. et al., 2019) that is much more computationally efficient and can infer models on a consumer-grade computer in a reasonable amount of time. We anticipate that with these improvements, it will be possible to collect time series datasets that span multiple hematopoietic lineages and genetic backgrounds and use the gene circuit approach to comprehensively decode the genetic architecture of hematopoietic cell-fate decisions.

2.4 MATERIALS AND METHODS

2.4.1 Gene Circuit Model of Erythroid-Neutrophil differentiation

The initial conditions were given by the mRNA concentrations in progenitor cells. Equations 2.1 were solved numerically using the Bulirsch-Stöer adaptive step-size solver to an accuracy of 10^{-3} as described previously (Manu et al., 2009a).

2.4.2 Training Data

The gene circuit was trained on May *et al.*'s genome-wide gene expression time-series dataset (GEO GSE49991; May et al., 2013) acquired during the differentiation of FDCP-mix cells into erythrocytes or neutrophils. See (May et al., 2013) for the details of data processing and cross-sample normalization. The expression level of each gene was further normalized against its maximum expression in either condition for model training and visualization.

2.4.3 Optimization by Parallel Lam Simulated Annealing (PLSA)

The parameters of Equation 2.1 were inferred by minimizing the cost function

$$E = \sum_{i,m,l} \left(x_i^l(t_m) - \hat{x}_i^l(t_m) \right)^2 + \text{Penalty}, \tag{2.2}$$

where $x_i^l(t_m)$ and $\hat{x}_i^l(t_m)$ are model output and data respectively for gene i in lineage/condition l at time t_m. The penalty is a weighted regularization term that limits the search space or magnitude of the regulatory parameters T_{ij}, b_i, and h_i. The penalty is given by

$$\text{Penalty} = \begin{cases} \exp(\Pi) - \exp(1), & \text{if } \Pi > 1 \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

where

$$\Pi = \sum_i \Lambda_i \left(\sum_j (T_{ij} \hat{x}_j^{\max})^2 + (b_i^{\max})^2 + h_i^2 \right).$$

$\hat{x}_j^{\max}$ is the maximum expression of gene j observed in the dataset (John et al., 1995) and $b_i^{\max}$ is the maximum value of b_i^l over all conditions l. Λ_i controls the magnitude of the regulatory parameters of gene i. Λ_i was set to 0.1 for all genes except *Csf3r*, for which Λ_i was set to 0.01. This allowed *Csf3r*'s regulatory parameters to have larger values, which was necessary for the model to be able to recapitulate the large dynamic range of *Csf3r* expression data (Fig 2.1).

The cost function (Eq. 2.2) was minimized using parallel Lam simulated annealing (PLSA)—simulated annealing with the Lam cooling schedule (Lam & Delosme, 1988)—running in parallel (Chu et al., 1999) as described previously (Manu et al., 2009b). PLSA was carried out on 10 CPUs (Intel Xeon E5-2643 v3 cores) in parallel.

2.4.4 Selection of gene circuits for analysis

Since PLSA is a stochastic method (Chu et al., 1999), each optimization attempt results in different values of inferred parameters and hence in a distinct gene circuit model. In order to evaluate their reproducibility, we repeated the optimization to obtain 100 different gene circuits. The root mean square (RMS) score,

$$\text{RMS} = \sqrt{\frac{E}{N_d}}, \tag{2.4}$$

where N_d is the total number of data points, was used to measure the goodness-of-fit of each gene circuit model. We chose 71 gene circuits having RMS scores lower than 0.06, corresponding to an average error of 6% in expression levels. Models

with higher RMS scores showed qualitative defects in their expression patterns compared to data.

2.4.5 Significance of fits

The optimization problem for the 12-gene circuit is overdetermined, having 720 data points and 192 free parameters, and the risk of overfitting is minimal. Nevertheless, we checked whether the model fits captured temporal patterns inherent in the data or whether the degrees of freedom were so numerous that the model could fit randomized non-biological data equally well. We randomized the data in a manner that preserved the dynamic range of the real data while creating non-biological temporal expression patterns and tested the ability of gene circuits to fit the latter compared to the former. For each gene, we created chimerical temporal expression patterns by combining erythrocyte training data up to the 96 hour time point with neutrophil data at later time points and vice versa. In each synthetic dataset, 10 of 12 genes were given chimerical expression patterns while the other two retained the original training data. 66 such synthetic datasets were generated for each combination of 10 genes (Algorithm 1). 10 gene circuits were trained per dataset resulting in a total of 660 gene circuits. The RMS scores of the resultant gene circuits were compared to the 100 gene circuits trained on the real data. The statistical significance of the differences between the RMS scores of gene circuits trained on random and real data was determined using the Wilcoxon rank sum test with continuity correction.

Algorithm 1 Gene expression swapping between two conditions

C_i: combination i for 10 out of 12 genes
c_i: 2 genes not included in C_i
$time\_points$: 30 sampled differentiation time points, $time\_points \in \{0, 2, 4, ..., 96, 120, 168\}$
$x_{ery}(i, g, t) \leftarrow$ gene expressions in erythrocyte condition for gene combination i, gene g, and time point t
$x_{neu}(i, g, t) \leftarrow$ gene expressions in neutrophil condition for gene combination i, gene g, and time point t
$x_{eryW}(i, g, t)$: empty $66 \times 12 \times 30$ array for storing swapped and normal gene expressions in erythrocyte condition for gene combination i, gene g, and time point t
$x_{neuW}(i, g, t)$: empty $66 \times 12 \times 30$ array for storing swapped and normal gene expressions in neutrophil condition for gene combination i, gene g, and time point t
$file_i$: output file for writing swapped and normal expressions for gene combination i
for i in 1:$^{12}C_{10}$ **do**
  for gene g in C_i **do**
    for t in $time\_points$ **do**
      if $t < 96$ **then**
        $x_{eryW}(i, g, t) \leftarrow x_{ery}(i, g, t)$
        $x_{neuW}(i, g, t) \leftarrow x_{neu}(i, g, t)$
      else
        $x_{eryW}(i, g, t) \leftarrow x_{neu}(i, g, t)$
        $x_{neuW}(i, g, t) \leftarrow x_{ery}(i, g, t)$
      end if
    end for
  end for
  for gene g in c_i **do**
    for t in $time\_points$ **do**
      $x_{eryW}(i, g, t) \leftarrow x_{ery}(i, g, t)$
      $x_{neutW}(i, g, t) \leftarrow x_{neu}(i, g, t)$
    end for
  end for
  for gene g in (C_i and c_i) **do**
    for t in $time\_points$ **do**
      WRITE($file_i$, $x_{eryW}(i, g, t)$, $x_{neuW}(i, g, t)$)
    end for
  end for
end for

2.4.6 The sensitivity of the model to initial conditions

The initial concentration of each gene was perturbed by $\pm 10, 20, 30, 50, 70\%$, one gene at a time. Model 66 was run with the perturbed initial conditions (120 simulations) and the RMS for each simulation was calculated.

2.4.7 Simulation of perturbation experiments

Gata1 and *Spi1* knockout was simulated by setting their initial concentrations and mRNA synthesis rates R_i to zero.

To simulate the knockdown and overexpression experiments carried out by (May et al., 2013) in FDCP-mix cells, we chose one representative model from the 71 that had met the goodness-of-fit criteria. For each model, we determined the number of regulatory parameters (T_{ij}) that had the same sign as the majority of the models. Of the 7 models having the largest number of regulatory parameters aligning with the consensus, one model, model #66, was chosen for perturbation simulations.

The knockdown *Spi1* or *Gata2* in FDCP-mix cells was simulated by decreasing the maximum synthesis rate of the gene, R_{Spi1} or R_{Gata2}, respectively. Since the efficiency of the knockdown achieved in the specific experiments was unknown, we chose the value of R_{Spi1} or R_{Gata2} so that the simulated expression of *Spi1* or *Gata2* matched the empirical values respectively. The simulations therefore could be said to predict the expression of only 11 of the 12 genes.

In the PU.1ERT and GATA1ERT experiments, 4-hydroxy-tamoxifen (OHT) treatment did not directly modulate the amount of *Spi1* or *Gata1* mRNA but instead

increased the activity of the constitutively expressed PU.1ERT and GATA1ERT fusion proteins. We simulated the increase in the activity of PU.1 or Gata1 by introducing a constant bias term B_i in the total regulatory input u of each gene i,

$$u = \sum_{j=1}^{N} T_{ij} x_j^l + b_i c^l + h_i + B_i.$$

The bias term is proportional to the genetic interconnectivity parameter corresponding to the regulation of each gene by PU.1 or Gata1 so that $B_i = T_{i \leftarrow Spi1} \cdot \beta_{Spi1}$ or $B_i = T_{i \leftarrow Gata1} \cdot \beta_{Gata1}$ respectively. The proportionality constants β_{Spi1} and β_{Gata1} represent the additional amount of active PU.1 and Gata1 induced by OHT respectively. Similar to the knockdown experiments, the efficiency of activation achieved in the overexpression experiments was unknown and we chose the values of the proportionality constants to match the observed expression of 1 of 12 genes. We did not however fit to the observed expression of the overexpressed gene since it stems from a mixture of mRNAs transcribed from the endogenous locus and the constitutively expressed ERT fusion gene. Instead we chose the values of the proportionality constants so the simulations matched the observed expression of *Gata1* in the PU.1ERT and *Spi1* in the GATA1ERT experiments respectively.

The simulations were carried out with $c^l = 0$ to simulate the progenitor condition since the experiments had been conducted in undifferentiated FDCP-mix cells. The simulations were compared to experimental data at equilibrium. The GRN was simulated for 1000 hours to allow the solution to reach equilibrium. The

ratio of each gene's expression in the perturbed condition to its expression in the unperturbed condition was computed to determine the fold change predicted by the simulation. This was compared to the empirical fold change, computed as the ratio of gene expression in treated cells to gene expression in control cells.

2.4.8 Analysis of gene regulation dynamics

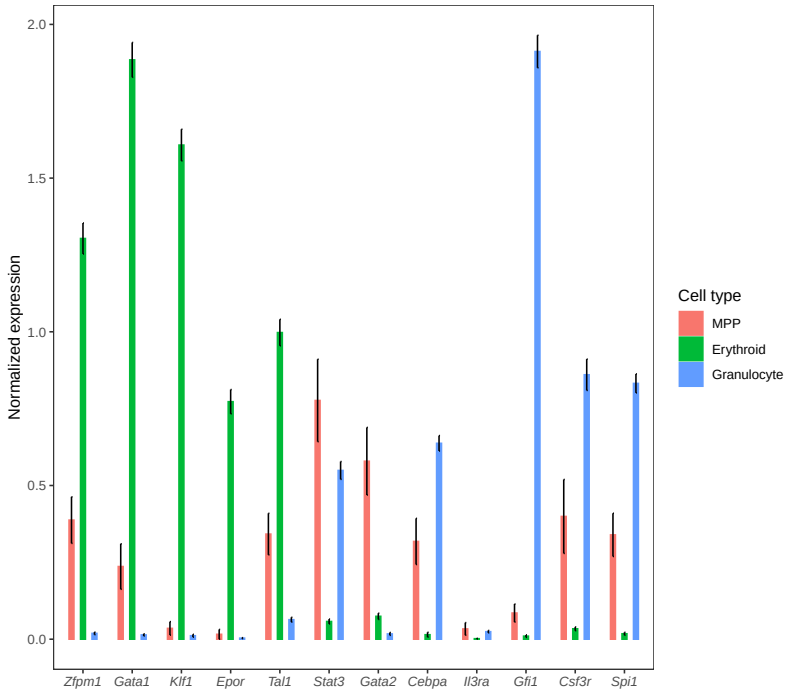
The contribution of individual regulators to the activation or repression of a target was determined by decomposing the total regulatory input $u = \sum_{j=1}^{N} T_{ij}x_j^l + b_i c^l + h_i$ into its individual terms. The contribution of regulator j to the regulation of gene i was determined by computing $T_{ij}x_j^l(t)$, where T_{ij} is the genetic interconnectivity of the two genes and $x_j^l(t)$ is the model solution for the mRNA concentration of gene j at time t and condition l. Since the mRNA concentrations vary in time, the relative contributions of the regulators to the activation or repression of any target also vary in time. When the total regulatory input crosses 0, that is $u = 0$, the regulation-expression $S(u) = \frac{1}{2}\left(u/\sqrt{(u^2+1)}+1\right) = \frac{1}{2}$ and the mRNA is synthesized at half the maximum rate (Eq. 2.1). The time at which different genes achieve half-maximum expression was used to order their activation in time.

2.4.9 Visualization of Tusi *et al.*'s scRNA-Seq data

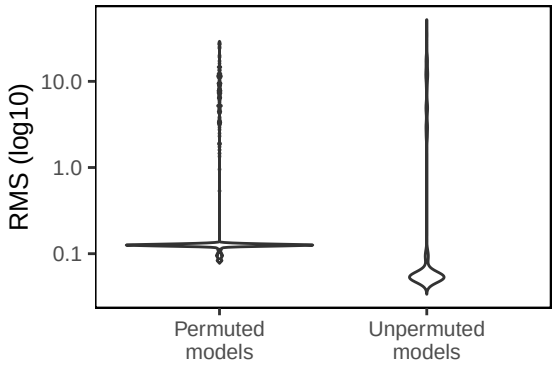
The expression of *Cebpa*, *Gfi1*, and *Spi1* in individual Kit$^+$ hematopoietic progenitors cells from mouse bone marrow (GEO GSE49991; Tusi et al., 2018) was visualized as follows. The cells were arranged in 2D space as a k-nearest-neighbor (knn)

graph according to their pairwise distances in gene expression space (SPRING algorithm; Weinreb et al., 2018a). The potential landscape and the probability of each cell to adopt a given fate were given by Population Balance Analysis (PBA; see Weinreb et al., 2018b, for details). Genome-wide normalized gene expression counts, the PBA potential, the PBA lineage probability, and the 2D SPRING coordinates of each cell were obtained from
https://kleintools.hms.harvard.edu/paper_websites/tusi_et_al/.

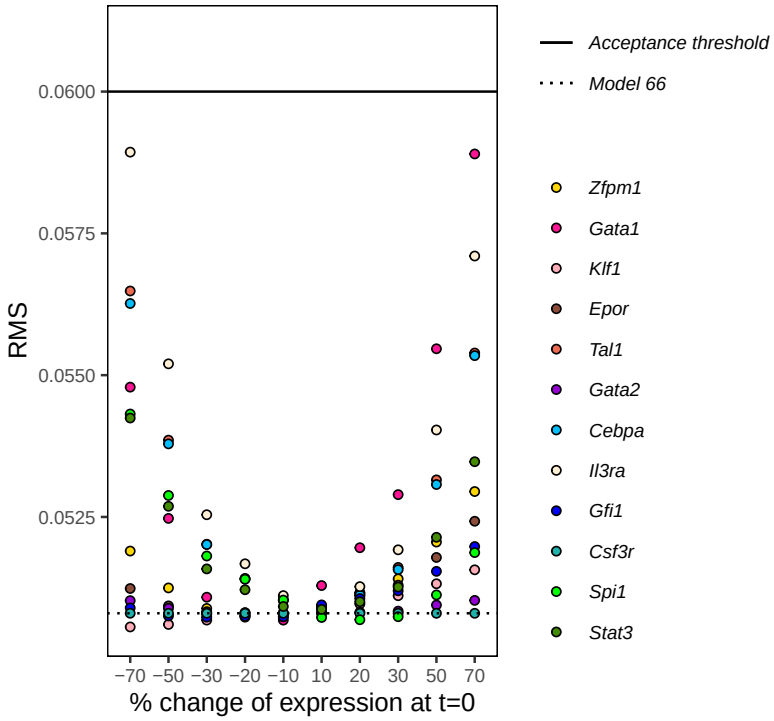
2.5 SUPPLEMENTARY DATA



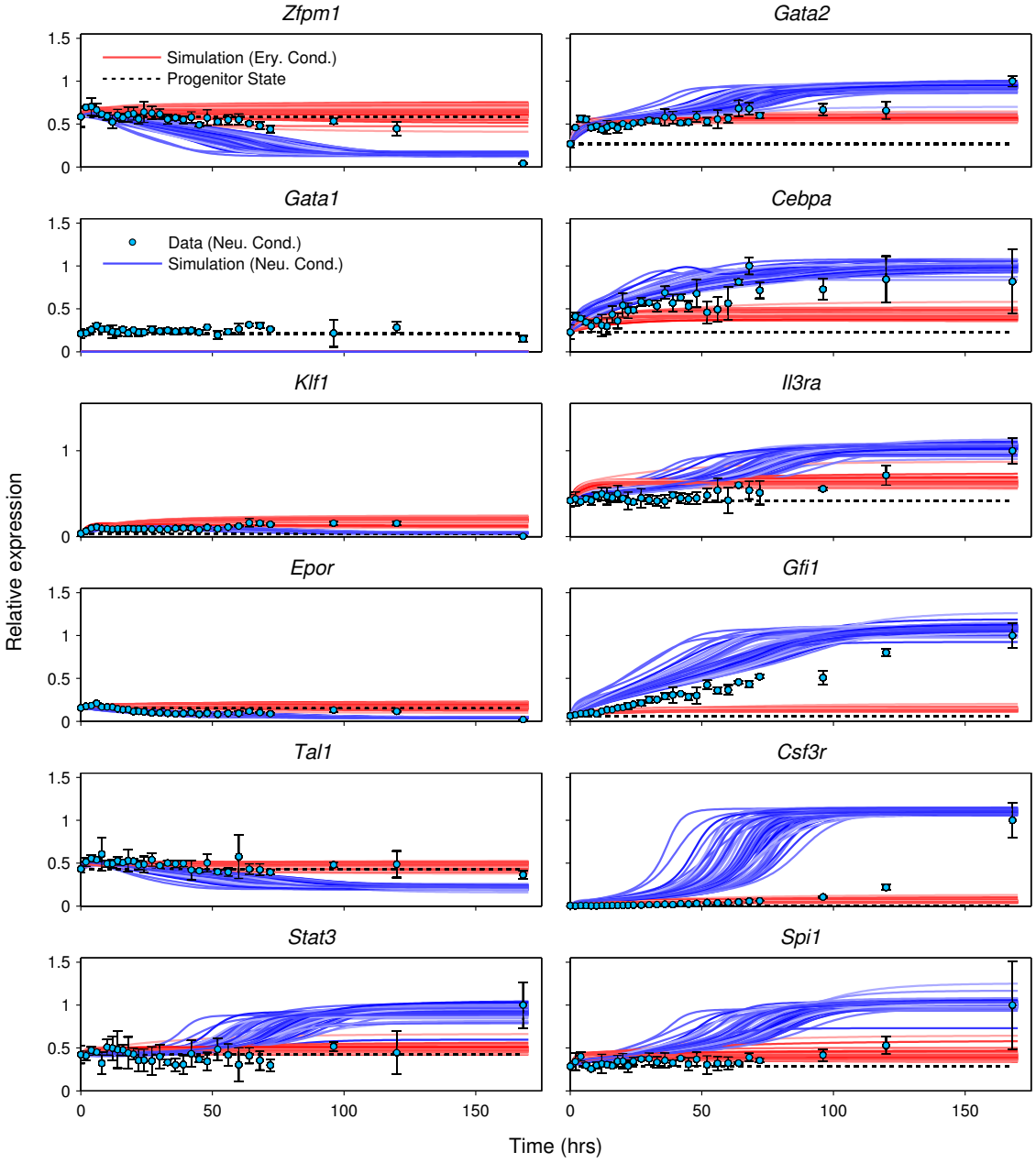
SFigure 2.1: Expression of the modeled genes in the Tusi *et al.* scRNA-Seq dataset. The average expression in MPPs, erythroid progenitors, and granulocytic progenitors is shown for the modeled genes. Erythroid and granulocytic progenitors were identified as having a PBA erythroid and granulocytic probability (see 2.4) greater than 0.9 respectively. MPPs were identified as cells having a low PBA probability (< 0.2) of belonging to any lineage. Error bars show the standard error of the mean.



SFigure 2.2: The significance of gene circuit fits. The distributions of the RMS scores of gene circuits trained on real data (Unpermuted models) or on randomized synthetic data (Permuted models) are shown as violin plots. The scores were compared using the Wilcoxon ranksum test with continuity correction ($p = 3.8 \times 10^{-8}$).



SFigure 2.3: Sensitivity of the model to initial conditions. Model 66 was run with the initial conditions perturbed one gene at a time. The x-axis is the magnitude of the perturbation. The y-axis is the RMS. The perturbed gene is indicated by the color of the points. The dotted line is the RMS of model 66 and the black line is the goodness-of-fit threshold RMS.



SFigure 2.4: Simulation of *Gata1* knockout. *Gata1* knockout was simulated in all 71 models that met the goodness-of-fit criteria. Their output is plotted as lines. The symbols and colors are the same as in Fig. 2.1.

STable 2.1: The values of the parameters of the gene circuit models that met the goodness-of-fit criteria

Columns correspond to parameters while rows correspond to models. T_{ij} are shown as T_Gene$i$_Gene$j$, b_i are shown as b_Gene$i$, h_i are shown as h_Gene$i$, R_i are shown as R_Gene$i$, and λ_i are shown as lambda_Gene$i$.
https://doi.org/10.1371/journal.pcbi.1009779.s005 (TSV)

STable 2.2: Comparison of model predictions with published experimental evidence

Each row compares a model prediction about a genetic interconnectivity parameter T_{ij}, representing the regulation of gene i by gene j, with published experimental evidence. Comparisons of the same parameter to multiple papers are listed in separate rows. T_{ij} is listed as T_Gene$i$_Gene$j$. The prediction column lists the type of regulation inferred by the model. It shows activation or repression when the first quartile of the distribution of the inferred parameter is positive or if the third quartile of the distribution is negative respectively (Fig. 2.4). The prediction column shows "sign not constrained" when the interquartile range spans negative and positive values. The experiment column lists that type of interaction established in the paper. If the paper describes evidence only of binding but not whether the target is activated or repressed, then the entry is "binding". Negative experimental results are listed as "no effect found". The entry in the experiment column is "not found" if we were not able to find any published tests of the parameter in question. The "Status of prediction" columns lists whether the evidence matches the prediction or not. "confirmed" implies agreements, while "incorrect prediction" implies disagreement. An asterisk indicates that conflicting experimental evidence was found. Conflicting evidence was found for the regulation of *Gata1*, *Spi1*, and *Gata2* by Gfi1 (Moignard et al., 2013). Situations where no evidence was found or the paper reported negative results are listed as "undetermined". The "Type of evidence" column classifies the evidence as genetic, protein-protein interaction, *cis* regulation, or functional *cis* regulation. Genetic evidence involves genetic manipulation of the predicted regulator followed by a characterization of the target's expression and usually cannot distinguish between direct and indirect effects. Protein-protein interaction implies biochemical evidence of direct protein-protein interactions. *cis* regulatory evidence indicates direct interactions by identifying regulatory elements or binding sites potentially bound by the predicted regulator but does not establish a functional relationship between binding and the expression of the target gene. Functional *cis* regulation goes a step further and manipulates the binding sites and measures reporter or target expression to provide evidence that the binding of the regulator has functional impacts. The "Organism/Cells" and "Citation" columns list the organism or cells in which the interaction was tested and the DOI URL for the paper respectively. https://doi.org/10.1371/journal.pcbi.1009779.s006 (TSV)

BIBLIOGRAPHY

Abdol, A. M., Cicin-Sain, D., Kaandorp, J. A., & Crombach, A. (2017). Scatter search applied to the inference of a development gene network. *Computation*, *5*(2).

Bertolino, E., Reinitz, J., & Manu (2016). The analysis of novel distal cebpa enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification. *Dev Biol*, *413*(1), 128–44.

Bonzanni, N., Garg, A., Feenstra, K. A., Schütte, J., Kinston, S., Miranda-Saavedra, D., Heringa, J., Xenarios, I., & Göttgens, B. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics*, *29*(13), i80–8.

Cantor, A. B., & Orkin, S. H. (2001). Hematopoietic development: a balancing act. *Curr Opin Genet Dev*, *11*(5), 513–9.

Cantor, A. B., & Orkin, S. H. (2002). Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*, *21*(21), 3368–76.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., & Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, *453*(7194), 544–7.

Chiba, T., Nagata, Y., Kishi, A., Sakamaki, K., Miyajima, A., Yamamoto, M., Engel, J. D., & Todokoro, K. (1993). Induction of erythroid-specific gene expression in lymphoid cells. *Proceedings of the National Academy of Sciences*, *90*(24), 11593–11597.

Chickarmane, V., Enver, T., & Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol*, *5*(1), e1000268.

Chu, K. W., Deng, Y., & Reinitz, J. (1999). Parallel simulated annealing by mixing of states. *The Journal of Computational Physics*, *148*, 646–662.

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc Natl Acad Sci U S A*, *114*(23), 5792–5799.

Cooper, S., Guo, H., & Friedman, A. D. (2015). The +37 kb cebpa enhancer is critical for cebpa myeloid gene expression and contains functional sites that bind scl, gata2, c/ebp, pu.1, and additional ets factors. *PLoS One*, *10*(5), e0126385.

Dahl, R., Walsh, J. C., Lancki, D., Laslo, P., Iyer, S. R., Singh, H., & Simon, M. C. (2003a). Regulation of macrophage and neutrophil cell fates by the pu.1 c/ebpa ratio and granulocyte colony-stimulating factor. *Nature Immunology*, *4*(10), 1029–1036.

Dahl, R., Walsh, J. C., Lancki, D., Laslo, P., Iyer, S. R., Singh, H., & Simon, M. C. (2003b). Regulation of macrophage and neutrophil cell fates by the pu.1:c/ebpalpha ratio and granulocyte colony-stimulating factor. *Nat Immunol*, *4*(10), 1029–36.

Dakic, A., Metcalf, D., Di Rago, L., Mifsud, S., Wu, L., & Nutt, S. L. (2005). PU.1 regulates the commitment of adult hematopoietic progenitors and restricts granulopoiesis. *J. Exp. Med.*, *201*(9), 1487–1502.

Deindl, P., Klar, M., Drews, D., Cremer, M., Gammella, E., Gassmann, M., & Dame, C. (2014). Mice over-expressing human erythropoietin indicate that erythropoietin enhances expression of its receptor via up-regulated gata1 and tal1. *Haematologica*, *99*(10), e205–e207.

DeKoter, R. P., & Singh, H. (2000). Regulation of b lymphocyte and macrophage development by graded expression of pu.1. *Science*, *288*(5470), 1439–41.

Doré, L. C., Chlon, T. M., Brown, C. D., White, K. P., & Crispino, J. D. (2012). Chromatin occupancy analysis reveals genome-wide gata factor switching during hematopoiesis. *Blood*, *119*(16), 3724–33.

Enver, T., Pera, M., Peterson, C., & Andrews, P. W. (2009). Stem cell states, fates, and the rules of attraction. *Cell Stem Cell*, *4*(5), 387–97.

Fehr David A., J. E., Handzlik, Manu, . ., & Loh, Y. L. (2019). Classification-based inference of dynamical models of gene regulatory networks. *G3: Genes, Genomes, Genetics*, *9*(12), 4183–4195.

Graf, T., & Enver, T. (2009). Forcing cells to change lineages. *Nature*, *462*(7273), 587–94.

Gursky, V. V., Jaeger, J., Kozlov, K. N., Reinitz, J., & Samsonova, A. M. (2004). Pattern formation and nuclear divisions are uncoupled in *Drosophila* segmentation: comparison of spatially discrete and continuous models. *Physica D*, *197*, 286–302.

Handzlik, J. E., & Manu (2022). Data-driven modeling predicts gene regulatory network dynamics during the differentiation of multipotential hematopoietic progenitors. *PLOS Computational Biology*, *18*(1), 1–31.

Hirsch, M., Smale, S., & Devaney, R. (2004). *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Boston: Academic Press.

Hock, H., Hamblen, M. J., Rooke, H. M., Traver, D., Bronson, R. T., Cameron, S., & Orkin, S. H. (2003a). Intrinsic requirement for zinc finger transcription factor gfi-1 in neutrophil differentiation. *Immunity*, *18*(1), 109–120.

Hock, H., Hamblen, M. J., Rooke, H. M., Traver, D., Bronson, R. T., Cameron, S., & Orkin, S. H. (2003b). Intrinsic requirement for zinc finger transcription factor gfi-1 in neutrophil differentiation. *Immunity*, *18*(1), 109–120.

Hong, T., Xing, J., Li, L., & Tyson, J. J. (2012). A simple theoretical framework for understanding heterogeneous differentiation of cd4+ t cells. *BMC Syst Biol*, *6*, 66.

Hoppe, P. S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K. D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Gambardella, A., Ahmed, N., Etzrodt, M., Coutu, D. L., Rieger, M. A., Marr, C., Strasser, M. K., Schauberger, B., Burtscher, I., Ermakova, O., Bürger, A., Lickert, H., Nerlov, C., Theis, F. J., & Schroeder, T. (2016). Early myeloid lineage choice is not initiated by random pu.1 to gata1 protein ratios. *Nature*, *535*(7611), 299–302.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., & Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev*, *11*(6), 774–85.

Huang, S., Guo, Y., May, G., & Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, *305*, 695–713.

Huang, Z., Dore, L. C., Li, Z., Orkin, S. H., Feng, G., Lin, S., & Crispino, J. D. (2009). Gata-2 reinforces megakaryocyte development in the absence of gata-1. *Mol Cell Biol*, *29*(18), 5168–80.

Iwasaki, H., Mizuno, S.-i., Arinobu, Y., Ozawa, H., Mori, Y., Shigematsu, H., Takatsu, K., Tenen, D. G., & Akashi, K. (2006). The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes Dev*, *20*(21), 3010–21.

Iwasaki, H., Somoza, C., Shigematsu, H., Duprez, E. A., Iwasaki-Arai, J., Mizuno, S.-i., Arinobu, Y., Geary, K., Zhang, P., Dayaram, T., Fenyus, M. L., Elf, S., Chan, S., Kastner, P., Huettner, C. S., Murray, R., Tenen, D. G., & Akashi, K. (2005). Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood*, *106*(5), 1590–1600.

Jack, G. D., Zhang, L., & Friedman, A. D. (2009). M-csf elevates c-fos and phospho-c/ebpalpha(s21) via erk whereas g-csf stimulates shp2 phosphorylation in marrow progenitors to contribute to myeloid lineage specification. *Blood*, *114*(10), 2172–80.

Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004). Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, *430*, 368–371.

John, R., Eric, M., & H., S. D. (1995). Model for cooperative control of positional information in drosophila by bicoid and maternal hunchback. *Journal of Experimental Zoology*, *271*(1), 47–56.

Khanna-Gupta, A., Zibello, T., Idone, V., Sun, H., Lekstrom-Himes, J., & Berliner, N. (2005). Human neutrophil collagenase expression is c/ebp-dependent during myeloid development. *Experimental Hematology*, *33*(1), 42–52.

Kim, W., Klarmann, K. D., & Keller, J. R. (2014). Gfi-1 regulates the erythroid transcription factor network through Id2 repression in murine hematopoietic progenitor cells. *Blood*, *124*(10), 1586–1596.

Kitajima, K., Zheng, J., Yen, H., Sugiyama, D., & Nakano, T. (2006). Multipotential differentiation ability of gata-1-null erythroid-committed cells. *Genes & Development*, *20*(6), 654–659.

Kozlov, K., Surkova, S., Myasnikova, E., Reinitz, J., & Samsonova, M. (2012). Modeling of gap gene expression in drosophila kruppel mutants. *PLoS Comput Biol*, *8*(8), e1002635.

Lam, J., & Delosme, J.-M. (1988). An efficient simulated annealing schedule: Derivation. Tech. Rep. 8816, Yale Electrical Engineering Department, New Haven, CT.

Laslo, P., Pongubala, J. M. R., Lancki, D. W., & Singh, H. (2008). Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin Immunol*, *20*(4), 228–35.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., & Singh, H. (2006). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, *126*(4), 755–66.

Leddin, M., Perrod, C., Hoogenkamp, M., Ghani, S., Assi, S., Heinz, S., Wilson, N. K., Follows, G., Schönheit, J., Vockentanz, L., Mosammam, A. M., Chen, W., Tenen, D. G., Westhead, D. R., Göttgens, B., Bonifer, C., & Rosenbauer, F. (2011). Two distinct auto-regulatory loops operate at the pu.1 locus in b cells and myeloid cells. *Blood*, *117*(10), 2827–38.

Legraverend, C., Antonson, P., Flodby, P., & Xanthopoulos, K. G. (1993). High level activity of the mouse ccaat/enhancer binding protein (c/ebp alpha) gene promoter involves autoregulation and several ubiquitous transcription factors. *Nucleic Acids Res*, *21*(8), 1735–42.

Li, C., & Wang, J. (2013). Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: Landscape and biological paths. *PLoS Comput Biol*, *9*(8), e1003165 EP –.

Li, H., Ji, M., Klarmann, K. D., & Keller, J. R. (2010). Repression of id2 expression by gfi-1 is required for b-cell and myeloid development. *Blood*, *116*(7), 1060–9.

Ma, O., Hong, S., Guo, H., Ghiaur, G., & Friedman, A. D. (2014). Granulopoiesis requires increased c/ebpα compared to monopoiesis, correlated with elevated cebpa in immature g-csf receptor versus m-csf receptor expressing cells. *PLOS ONE*, *9*(4), 1–14.

Mancini, E., Sanjuan-Pla, A., Luciani, L., Moore, S., Grover, A., Zay, A., Rasmussen, K. D., Luc, S., Bilbao, D., O'Carroll, D., Jacobsen, S. E., & Nerlov, C. (2012). Fog-1 and gata-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. *The EMBO Journal*, *31*(2), 351–365.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009a). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. *PLoS Computational Biology*, *5*, e1000303. Doi:10.1371/journal.pcbi.1000303.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009b). Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biology*, *7*, e1000049. Doi:10.371/journal.pbio.1000049.

May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen, D. G., Peterson, C., & Enver, T. (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, *13*(6), 754–68.

Mikkola, H. K. A., Klintman, J., Yang, H., Hock, H., Schlaeger, T. M., Fujiwara, Y., & Orkin, S. H. (2003). Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia scl/tal-1 gene. *Nature*, *421*(6922), 547–51.

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., SchΓtte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., Jacobsen, S. E., de Bruijn, M. F., & GΓ¶ttgens, B. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biology*, *15*(4), 363–372.

Mossadegh-Keller, N., Sarrazin, S., Kandalla, P. K., Espinosa, L., Stanley, E. R., Nutt, S. L., Moore, J., & Sieweke, M. H. (2013). M-csf instructs myeloid lineage fate in single haematopoietic stem cells. *Nature*, *497*(7448), 239–43.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., & Gerstein, M. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol*, *17*, 53.

Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R. F., Chen, J., Hwang, L., Cheng, C., Auburn, R. P., Davis, M. B., Domanus, M., Shah, P. K., Morrison, C. A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild,

N. A., Grundstad, A. J., Hanley, D., MacAlpine, D. M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R. L., Ren, B., Posakony, J. W., Kellis, M., & White, K. P. (2011). A cis-regulatory map of the drosophila genome. *Nature*, *471*(7339), 527–31.

Nerlov, C., Querfurth, E., Kulessa, H., & Graf, T. (2000). Gata-1 interacts with the myeloid pu.1 transcription factor and represses pu.1-dependent transcription. *Blood*, *95*(8), 2543–51.

Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T., Frampton, G. M., Drake, A. C. B., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J. W., Liefeld, T., Smutko, J. S., Chen, J., Friedman, N., Young, R. A., Golub, T. R., Regev, A., & Ebert, B. L. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, *144*(2), 296–309.

Olsson, A., Venkatasubramanian, M., Chaudhri, V. K., Aronow, B. J., Salomonis, N., Singh, H., & Grimes, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, *537*(7622), 698–702.

Orkin, S. H., & Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, *132*(4), 631–44.

Palani, S., & Sarkar, C. A. (2008). Positive receptor feedback during lineage commitment can generate ultrasensitivity to ligand and confer robustness to a bistable switch. *Biophys J*, *95*(4), 1575–89.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., & Amit, I. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, *163*(7), 1663–77.

Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F. W., & Orkin, S. H. (1996). The t cell leukemia oncoprotein scl/tal-1 is essential for development of all hematopoietic lineages. *Cell*, *86*(1), 47 – 57.

Reinitz, J., & Sharp, D. H. (1995). Mechanism of *eve* stripe formation. *Mechanisms of Development*, *49*, 133–158.

Repele, A., Krueger, S., Bhattacharyya, T., Tuineau, M. Y., & Manu (2019). The regulatory control of cebpa enhancers and silencers in the myeloid and red-blood cell lineages. *PLoS One*, *14*(6), e0217580.

Rieger, M. A., Hoppe, P. S., Smejkal, B. M., Eitelhuber, A. C., & Schroeder, T. (2009). Hematopoietic cytokines can instruct lineage choice. *Science*, *325*(5937), 217–8.

Robb, L. (2007). Cytokine receptors and hematopoietic differentiation. *Oncogene*, *26*(47), 6715–23.

Rogers, H., Wang, L., Yu, X., Alnaeeli, M., Cui, K., Zhao, K., Bieker, J. J., Prchal, J., Huang, S., Weksler, B., & Noguchi, C. T. (2012). T-cell acute leukemia 1 (tal1) regulation of erythropoietin receptor and association with excessive erythrocytosis. *Journal of Biological Chemistry*, *287*(44), 36720–36731.

Rogers, H. M., Yu, X., Wen, J., Smith, R., Fibach, E., & Noguchi, C. T. (2008). Hypoxia alters progression of the erythroid program. *Experimental Hematology*, *36*(1), 17–27.

Scott, E. W., Simon, M. C., Anastasi, J., & Singh, H. (1994). Requirement of transcription factor pu.1 in the development of multiple hematopoietic lineages. *Science*, *265*(5178), 1573–7.

Scott, L., Civin, C., Rorth, P., & Friedman, A. (1992). A novel temporal expression pattern of three C/EBP family members in differentiating myelomonocytic cells. *Blood*, *80*(7), 1725–1735.

Shivdasani, R. A., Mayer, E. L., & Orkin, S. H. (1995). Absence of blood formation in mice lacking the t-cell leukaemia oncoprotein tal-1/scl. *Nature*, *373*(6513), 432–4.

Siatecka, M., & Bieker, J. J. (2011). The multifunctional role of eklf/klf1 during erythropoiesis. *Blood*, *118*(8), 2044–2054.

Smith, L., Hohaus, S., Gonzalez, D., Dziennis, S., & Tenen, D. (1996). Pu.1 (spi-1) and c/ebp alpha regulate the granulocyte colony- stimulating factor receptor promoter in myeloid cells. *Blood*, *88*(4), 1234–1247.

Spooner, C. J., Cheng, J. X., Pujadas, E., Laslo, P., & Singh, H. (2009). A recurrent network involving the transcription factors pu.1 and gfi1 orchestrates innate and adaptive immune cell fates. *Immunity*, *31*(4), 576–86.

Stachura, D. L., Chou, S. T., & Weiss, M. J. (2006). Early block to erythromegakaryocytic development conferred by loss of transcription factor gata-1. *Blood*, *107*(1), 87–97.

Starck, J., Cohet, N., Gonnet, C., Sarrazin, S., Doubeikovskaia, Z., Doubeikovski, A., Verger, A., Duterque-Coquillaud, M., & Morle, F. (2003). Functional cross-antagonism between transcription factors fli-1 and eklf. *Mol Cell Biol*, *23*(4), 1390–402.

Surkova, S., Kosman, D., Kozlov, K., Manu, Myasnikova, E., Samsonova, A., Spirov, A., Vanario-Alonso, C. E., Samsonova, M., & Reinitz, J. (2008). Characterization of the *Drosophila* segment determination morphome. *Developmental Biology*, *313*(2), 844–862.

Tian, S. S., Tapley, P., Sincich, C., Stein, R. B., Rosen, J., & Lamb, P. (1996). Multiple signaling pathways induced by granulocyte colony-stimulating factor involving activation of jaks, stat5, and/or stat3 are required for regulation of three distinct classes of immediate early genes. *Blood*, *88*(12), 4435–44.

Tusi, B. K., Wolock, S. L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J. R., Klein, A. M., & Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, *555*(7694), 54–60. 29466336[pmid].

van der Meer, L. T., Jansen, J. H., & van der Reijden, B. A. (2010). Gfi1 and gfi1b: key regulators of hematopoiesis. *Leukemia*, *24*(11), 1834–1843.

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A. G., & Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol*, *19*(4), 271–281.

Vicente, C., Conchillo, A., García-Sánchez, M. A., & Odero, M. D. (2012). The role of the gata2 transcription factor in normal and malignant hematopoiesis. *Critical Reviews in Oncology/Hematology*, *82*(1), 1 – 17.

Walsh, J. C., DeKoter, R. P., Lee, H.-J., Smith, E. D., Lancki, D. W., Gurish, M. F., Friend, D. S., Stevens, R. L., Anastasi, J., & Singh, H. (2002). Cooperative and antagonistic interplay between pu.1 and gata-2 in the specification of myeloid cell fates. *Immunity*, *17*(5), 665–676.

Weinreb, C., Wolock, S., & Klein, A. M. (2018a). Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, *34*(7), 1246–1248.

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., & Klein, A. M. (2018b). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, *115*(10), E2467–E2476.

Weiss, M. J., Keller, G., & Orkin, S. H. (1994). Novel insights into erythroid development revealed through in vitro differentiation of gata-1 embryonic stem cells. *Genes & Development*, *8*(10), 1184–1197.

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P. M., Kinston, S., Ouwehand, W. H., Dzierzak, E., Pimanda, J. E., de Bruijn, M. F. T. R., & Göttgens, B. (2010a). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, *7*(4), 532–44.

Wilson, N. K., Timms, R. T., Kinston, S. J., Cheng, Y.-H., Oram, S. H., Landry, J.-R., Mullender, J., Ottersbach, K., & Gottgens, B. (2010b). Gfi1 expression is controlled by five distinct regulatory regions spread over 100 kilobases, with scl/tal1, gata2, pu.1, erg, meis1, and runx1 acting as upstream regulators in early hematopoietic cells. *Mol Cell Biol*, *30*(15), 3853–63.

Wu, H., Manu, Jiao, R., & Ma, J. (2015). Temporal and spatial dynamics of scaling-specific features of a gene regulatory network in drosophila. *Nat Commun*, *6*, 10031.

Yoshida, T., Ng, S. Y.-M., & Georgopoulos, K. (2010). Awakening lineage potential by ikaros-mediated transcriptional priming. *Curr Opin Immunol*, *22*(2), 154–60.

Zhang, D. E., Zhang, P., Wang, N. D., Hetherington, C. J., Darlington, G. J., & Tenen, D. G. (1997). Absence of granulocyte colony-stimulating factor signaling and neutrophil development in ccaat enhancer binding protein alpha-deficient mice. *Proc Natl Acad Sci U S A*, *94*(2), 569–74.

Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K. A., Mueller, B. U., Narravula, S., Torbett, B. E., Orkin, S. H., & Tenen, D. G. (2000). Pu.1 inhibits gata-1 function and erythroid differentiation by blocking gata-1 dna binding. *Blood*, *96*(8), 2641–8.

Zhao, W., Kitidis, C., Fleming, M. D., Lodish, H. F., & Ghaffari, S. (2006). Erythropoietin stimulates phosphorylation and activation of gata-1 via the pi3-kinase/akt signaling pathway. *Blood*, *107*(3), 907–15.

CHAPTER 3

Classification-based Inference of Dynamical Models of Gene Regulatory Networks

In the previous chapter, it was shown how the genetic architecture of a GRN can be inferred by fitting gene circuits to gene expression time-series data using global non-linear optimization methods such as simulated annealing (SA). Parallel Lam SA (Chu et al., 1999; Manu et al., 2009b) required $\sim$6 hours to train one 12-gene circuit on 10 processors. PLSA is a stochastic optimization method with many optimization parameters that needs to be executed hundreds of times to obtain confident networks, therefore long computational times can be impractical in the inference of big GRNs. This chapter presents a novel classification-based inference approach called Fast Inference of Gene Regulation (FIGR) (Fehr David A. et al., 2019) to determining in an ultra-fast manner gene circuit parameters. FIGR exploits the switch-like nature of gene regulation by breaking the gene circuit inference problem into two simpler optimization problems. In (Fehr David A. et al., 2019) it was demonstrated that FIGR is faster than a global non-linear optimization by a factor of 600 during the inference of gap system in *Drosophila melanogaster*. Additionally, in FIGR the computational complexity scales much better with GRN size, making the inference of GRNs more practical. This chapter presents parts of the paper (Fehr David A. et al., 2019) where I performed an accuracy comparison between FIGR and SA run on synthetic GRNs.

3.1 INTRODUCTION

Development is controlled by gene regulatory networks (GRNs) that integrate extrinsic signals and intrinsic cell state to make decisions about cell fate (Levine & Davidson, 2005; Davidson & Levine, 2008). Modeling of GRNs is an important approach to understanding a wide variety of developmental processes such as pattern formation (Manu et al., 2009b,a; Verd et al., 2018; Balaskas et al., 2012), cell-fate specification (Hamey et al., 2017; May et al., 2013), pluripotency and cell-fate reprogramming (Collombet et al., 2017; Li & Wang, 2013), oncogenesis (Tyson et al., 2011), and regeneration (Pietak et al., 2019). Over the past decade or so, it has become clear that developmental GRNs comprise tens to hundreds of densely interconnected genes (Davidson et al., 2002a; Novershtern et al., 2011) rather than a few so-called master regulators. Moreover, developmental GRNs are wired recursively since the genes encoding transcription factors (TFs) are themselves regulated by other TFs or indirectly by non-TF gene products (Palani & Sarkar, 2008; Kueh et al., 2013). Their large size and high interconnectivity make the modeling of developmental GRNs a challenging problem.

Coupled ordinary or partial differential equations (ODEs or PDEs) are a natural choice for modeling GRNs since GRNs are nonlinear dynamical systems (Manu et al., 2009a; Weston et al., 2018; Li & Wang, 2013; Laslo et al., 2006) whose time evolution depends on their state. The state is defined by the concentrations of gene products and the equations are parameterized by constants with a biochemical or biophysical underpinning, such as synthesis and degradation rates and binding constants. Estimating the values of these parameters is necessary for sim-

ulating the time evolution of GRN state but direct *in vivo* biochemical measurement of the large numbers of parameters involved is generally infeasible if not outright impossible. One approach to estimating parameter values is to search in parameter space for broad regions that reproduce the qualitative behavior of the system (Huang et al., 2007; Laslo et al., 2006; Li & Wang, 2013; Hong et al., 2012). The other approach (Reinitz & Sharp, 1995; May et al., 2013) to parameter estimation is data-driven, that is, parameter values are inferred by fitting the ODEs or PDEs to quantitative observations of GRN state sampled in space and/or time. In inferring parameters from quantitative data, data-driven differential equation modeling of GRNs provides a framework for understanding developmental cellular decisions at a quantitative and predictive level.

Here we focus on a specific data-driven and predictive ODE modeling framework, termed *gene circuits*, that has been particularly successful in inferring and modeling developmental GRNs from spatiotemporal protein (Jaeger et al., 2004a; Manu et al., 2009b,a; Kozlov et al., 2012; Hengenius et al., 2011) or mRNA (Crombach et al., 2012) data. Gene circuits determine the time evolution of protein or mRNA concentrations using coupled nonlinear ODEs in which synthesis is represented as a switch-like function of regulator concentrations. The values of the free parameters define the regulatory influences among the genes in the network. Gene circuits do not presuppose any particular scheme of regulatory interactions, but instead determine it by estimating the values of the parameters from quantitative data using optimization. Gene circuits infer not only the topology of the GRN but also the type, either activation or repression, and strength of interac-

tions. Most importantly, the inference procedure yields ODE models that can be used to simulate and predict developmental perturbations (Jaeger et al., 2004b; Manu et al., 2009b,a; Wu et al., 2015; Verd et al., 2018).

Despite its successes, the gene circuit method suffers from the drawback that parameter inference is computationally expensive. Efficient optimization methods, such as steepest descent (Gursky et al., 2004) are guaranteed to find the global minimum only if the cost function, usually the sum of squared differences between model output and data, is convex—has a unique minimum—which is not the case in such problems. This implies that the only practical approach currently available for fitting gene circuit models is global nonlinear optimization with techniques such as simulated annealing (SA; Kirkpatrick et al., 1983; Lam & Delosme, 1988a,b), that minimize the cost function by searching the high-dimensional parameter space stochastically. Not only do global nonlinear optimization methods need to make millions of cost function evaluations in order to find the minimum, but each evaluation is itself quite costly since it involves solving a set of coupled differential equations. Furthermore, the computational cost scales poorly, as $O(G^3)$, with gene number G, since G ODEs are solved in each function evaluation and the number of cost function evaluations required is proportional to the number of parameters ($O(G^2)$). High computational cost and poor scalability have hamstrung the application of the gene circuit method to larger networks or more broadly in development. Gene circuits have only been inferred for relatively small networks so far (Reinitz & Sharp, 1995; Manu et al., 2009b; Cotterell & Sharpe, 2010; May et al., 2013; Verd et al., 2018).

One approach to speeding up the inference procedure has been to explore different global optimization methods such as evolutionary algorithms (Kozlov & Samsonov, 2009; Kozlov et al., 2012) and scatter search (Abdol et al., 2017). Alternative global optimization methods do not circumvent the problem of high computational cost since each cost function evaluation still involves the solution of coupled ODEs. Another important strategy for inferring gene circuits in a reasonable amount of time has been the development of parallel optimization algorithms such as parallel Lam simulated annealing (pLSA; Chu et al., 1999) and Differential Evolution Entirely Parallel (DEEP; Kozlov & Samsonov, 2009; Kozlov et al., 2012), including attempts at reducing communication overhead (Jostins & Jaeger, 2010; Lou & Reinitz, 2016). Although parallel methods reduce the absolute amount of time required to infer a gene circuit of a given size, they nevertheless suffer from the scaling problem.

In this paper we present an alternative approach, FIGR (Fast Inference of Gene Regulation), for determining gene circuit parameters that is significantly more computationally efficient than global nonlinear optimization. Our algorithm exploits the observation that the inference of the connectivity of a given gene can be rephrased as a supervised learning problem: to find a hyperplane in state space that classifies observations into two groups, one where the gene is ON and the other where the gene is OFF. Our algorithm determines whether a gene is ON or OFF at a given observation point by computing the time derivative of concentrations in a numerically robust manner. It then performs classification using either logistic regression or support vector machines (SVM) to determine the equation

of the switching hyperplane. The genetic interconnectivity can then be computed from the coefficients of the hyperplane equation in a straightforward manner. We have implemented the algorithm in MATLAB and tested its ability to recover the genetic interconnectivity of random GRNs of up to 50 genes from simulated data. The algorithm works as expected and recovers parameters accurately, provided that sufficient data are available. We also demonstrate the ability of our method to correctly infer the gap gene regulatory network of *Drosophila melanogaster* from empirical data. We observed a ∼600-fold speed up relative to simulated annealing on the gap gene problem.

3.2 MATERIALS AND METHODS

3.2.1 Validation of FIGR with synthetic data

3.2.1.1 Generation of synthetic data from random gene circuits

Random gene circuits were generated and simulated to generate synthetic data as follows. The synthesis rates R_g and degradation rates λ_g were drawn uniformly from the interval $[0.5, 2]$. We chose the genetic interconnectivity coefficients T_{gf} and threshold h_g such that the switching hyperplane passed through a random point $\mathbf{x}^{\text{cen}}$ drawn uniformly from the bounding hypercube ($0 < x_g^{\text{cen}} < \frac{R_g}{\lambda_g}$), and the normal to the switching hyperplane ($\widehat{\mathbf{T}}_g$) was drawn uniformly from the unit G-sphere, where G is the number of genes in the GRN. We generated N trajectories starting at random initial position $\mathbf{x_n}(t = 0)$ drawn uniformly from the bounding hypercube by integrating the Glass equations without diffusion (Eq. 3.3) using

MATLAB's `ode45` solver. We stored the values of these functions $x_{ng}(t_k)$ at N_t timepoints equally subdividing the interval of the simulation ($[0, 2]$) to serve as synthetic data for both FIGR and SA.

3.2.1.2 Inference with FIGR

Gene circuits were inferred using FIGR as described in Section FIGR: Classification-based inference. The user-defined options and parameters utilized in this study are provided in STable 3.1.

3.2.1.3 Inference with SA

SA was carried out with gene circuit C code in serial largely as described previously (Manu et al., 2009b) save for a few modifications. The quality factor λ was set to 0.001 and the averaging parameters λ_u and λ_v were set to 200 and 1000 respectively. The stopping criterion was 0.001. The parameter controlling the search space of the regulatory parameters Λ was set to 0.1. The search space of R_g and λ_g were set to $(0.4, 2.1)$. The Glass equations (Eq. 3.3) were solved using a 4th order Runge-Kutta solver. Since SA is a stochastic method, different optimization runs yield slightly different gene circuits. The inferences were carried out in 5 replicates, each starting for a random set of initial parameter values. For each synthetic dataset, several replicates having low RMS could be identified. The circuit with the lowest RMS was chosen for further analysis.

3.2.2 Data Availability

SFig. 3.1 shows the fraction of genetic interconnectivity signs inferred correctly from synthetic data. SFig. 3.3 shows the training error of SA-inferred gene circuits. SFig. 3.3 shows the inference of h_g and kinetic parameters from synthetic data. STable 3.1 lists user-defined options and parameters utilized in FIGR code. SFrame 3.5 describes an alternative method for determining kinetic parameters in FIGR. FIGR Source code is freely available at http://github.com/mlekkha/FIGR.

3.3 RESULTS

3.3.1 Gene circuit models of GRNs

We consider a GRN of G genes whose state at time t is defined by the concentrations of the gene products $x_g(t)$, $g = 1, 2, \cdots, G$. We assume that the GRN functions cell autonomously, that is, the expression of the genes is independent of the state of other cells. Gene circuits (Reinitz & Sharp, 1995) describe the time evolution of $x_g(t)$ according to G coupled ordinary differential equations,

$$\frac{dx_g}{dt} = R_g\, S\left(\sum_{f=1}^{G} T_{gf} x_f + h_g\right) - \lambda_g x_g, \tag{3.1}$$

where R_g is the maximum synthesis rate of product g. T_{gf} are genetic interconnectivity coefficients describing the regulation of gene g by the product of gene f. Positive and negative values of T_{gf} signify activation and repression of gene g by gene f respectively. The threshold h_g determines the basal synthesis rate, and

λ_g is the degradation rate of product g. Nominally, all genes in the model also function as regulators, so that both g and f run over the range $1, 2, 3, \ldots, G$. Sometimes such gene networks include upstream regulators that are not themselves influenced by other gene products represented in the model. For example, in the *Drosophila* segmentation gene network, maternal proteins such as Bicoid activate the zygotically expressed genes, but are not regulated by their targets (Akam, 1987). An upstream regulator g can be represented by setting $T_{gf} = 0$ for all f.

$S(u)$ is the regulation-expression function, which determines the fraction of the maximum synthesis rate attained by the gene given the total regulatory input $u = \sum_{f=1}^{G} T_{gf} x_f + h_g$. $S(u)$ is required to have a switch-like dependence on u and to take values between 0 and 1. If the total regulatory input has large positive values, $u \gg 0$, as a result of high activator concentrations, low repressor concentrations, or both, $S(u) \sim 1$ and the gene product is synthesized at the maximum rate R_g. If the total regulatory input has large negative values, $u \ll 0$, so that $S(u) \sim 0$, the gene product is not synthesized. One sigmoid function that satisfies these properties,

$$S(u) = \sigma(u) = \frac{1}{2}\left(\frac{u}{\sqrt{1+u^2}} + 1\right), \tag{3.2}$$

has been utilized almost exclusively in previous studies (Reinitz & Sharp, 1995; Jaeger et al., 2004a; Manu et al., 2009b; Kozlov et al., 2012). However, any function that satisfies these rather general properties is a valid regulation-expression function.

3.3.1.1 *Glass networks*

In what follows, we show that one choice of the regulation-expression function permits a radical simplification of the gene circuit inference problem. If the regulation-expression function is chosen to be the Heaviside function,

$$S(u) = \Theta(u) = \begin{cases} 0 & \text{if} \quad u < 0 \\ 1 & \text{if} \quad u \geq 0. \end{cases}$$

the resulting differential equations (Eq. 3.1) are piece-wise linear and are referred to as Glass networks (Glass & Kauffman, 1973; Edwards, 2000; Glass & Pasternack, 1978; Mestl et al., 1996).

Using the state vector $\mathbf{x} = (x_1, x_2, \ldots, x_G)$ to represent a point in the G-dimensional state space of the model and the vector $\mathbf{T}_g$ to represent the gth row of the genetic interconnectivity matrix, the Glass equations may be written as

$$\frac{dx_g}{dt} = R_g \, \Theta(\mathbf{T}_g \cdot \mathbf{x} + h_g) - \lambda_g x_g, \qquad g = 1, 2, 3, \ldots, G. \tag{3.3}$$

The gene may said to be "ON" or "OFF" depending on whether the gene product is being synthesized or not respectively. Equation (3.3) implies that

$$\text{gene } g \text{ is} \begin{cases} \text{ON if} & \mathbf{T}_g \cdot \mathbf{x} + h_g > 0 \\ \text{OFF if} & \mathbf{T}_g \cdot \mathbf{x} + h_g < 0. \end{cases} \tag{3.4}$$

Thus the "gene g ON" and "gene g OFF" configurations are separated in state

space by the hyperplane defined by the equation $\mathbf{T}_g \cdot \mathbf{x} + h_g = 0$. We call this the *switching hyperplane*. $\mathbf{T}_g$ is the normal to the switching hyperplane and $\mathbf{T}_g \cdot \mathbf{x} + h_g$ is the perpendicular distance of any point $\mathbf{x}$ to the hyperplane. Furthermore,

$$x_g(t) = \begin{cases} x_g(0)e^{-\lambda_g t} + \frac{R_g}{\lambda_g}(1 - e^{-\lambda_g t}) & \text{if ON for } t \geq 0 \\ x_g(0)e^{-\lambda_g t} & \text{if OFF for } t \geq 0. \end{cases} \tag{3.5}$$

Equations (3.4) and (3.5) imply that for Glass networks, the *regulatory parameters*, $\mathbf{T}_g$ and h_g, and the *kinetic parameters*, R_g and λ_g, are separable. The former determine the switching hyperplane, while the latter determine the trajectories on either side of the hyperplane. Figure 3.1D,G shows examples of the switching hyperplanes and trajectory of a two-gene gene circuit (Fig. 3.1A,B) having a stable spiral equilibrium solution.

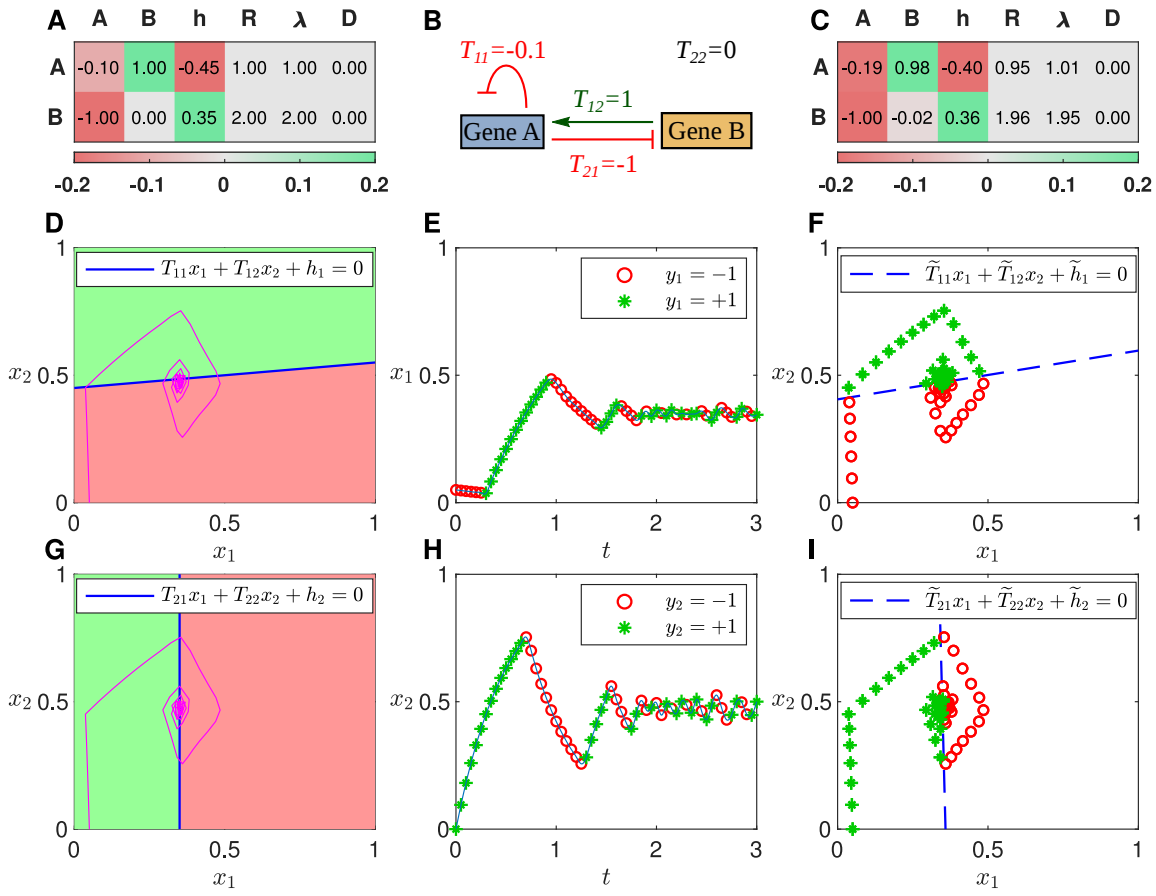


Figure 3.1: Classification-based inference of an example gene circuit. **A**. Theoretical parameter values are listed by row for each gene. The **T** matrix is shown in the first two columns, one column per regulator. Green (red) indicates activation (repression). **B**. Schematic of the theoretical gene circuit. **C**. Parameters inferred by FIGR. **D,G**. Trajectory in state space (purple) overlaid upon the Heaviside regulation-expression function. Green (red) is ON (OFF). The switching hyperplane is plotted as a blue line. Switching hyperplanes for genes A and B are showing in panels D and G respectively. **E,H**. Sampled gene expression trajectories and assignment of ON/OFF state for genes A (panel E) and B (panel H). Trajectories are numerical solutions of Equation (3.3). Detected ON or OFF state (Section Determining ON/OFF state) is indicated with green stars or red circles respectively. **F,I**. Switching hyperplane (dashed blue line) inferred using Logistic regression for genes A (panel F) and B (panel I). Sampled trajectories annotated with ON/OFF state are plotted in state space.

3.3.2 FIGR: Classification-based inference

Let the expression of each gene be measured at N_t time points $t_e, e = 1, \cdots, N_t$, along trajectories starting from $n = 1, \cdots, N$ initial conditions. The goal of GRN inference is to estimate the values of the gene circuit parameters $\tilde{T}_{gf}$, $\tilde{h}_g$, $\tilde{R}_g$, and $\tilde{\lambda}_g$ given the measurements $x_{ng}(t_e)$.

In FIGR, we exploit the separability of the regulatory ($\mathbf{T}_g$, h_g) and kinetic (R_g, λ_g) parameters to break up the inference problem into two distinct tractable subproblems. For inferring the parameters of any given gene, we classify the data points into two classes—one in which the gene's product is being synthesized (ON class) and the other in which the product is not being synthesized (OFF class). The regulatory parameters are inferred by determining the optimal $G-1$ dimensional hyperplane separating the two classes. The kinetic parameters can be inferred either by fitting the piece-wise linear Glass equations to estimates of the rate of change of gene product concentrations or by fitting Equation 3.5 to the gene product concentration time series.

3.3.2.1 Determining ON/OFF state

We will assume that the gene product concentration, including initial concentration, is bounded by the maximum concentration determined by the synthesis and degradation rates, that is,

$$0 \leq x_g < \frac{R_g}{\lambda_g}, \qquad g = 1, 2, 3, \ldots, G. \tag{3.6}$$

Let

$$y_g \equiv (\mathbf{T}_g \cdot \mathbf{x} + h_g) = \pm 1 \tag{3.7}$$

represent the ON/OFF state of gene g. Then,

$$\frac{dx_g}{dt} = \begin{cases} R_g - \lambda_g x_g > 0 & \text{if} \quad \mathbf{T}_g \cdot \mathbf{x} + h_g > 0 \\ -\lambda_g x_g \leq 0 & \text{if} \quad \mathbf{T}_g \cdot \mathbf{x} + h_g < 0. \end{cases} \tag{3.8}$$

This implies that the ON/OFF state of a gene can be determined by ascertaining the sign of the *velocity*, $v_g = \frac{dx_g}{dt}$.

$$y_g \equiv (\mathbf{T}_g \cdot \mathbf{x} + h_g) = \frac{dx_g}{dt}, \qquad g = 1, 2, 3, \ldots, G. \tag{3.9}$$

Gene expression data, such as those obtained from immunofluorescence or high-throughput sequencing, inevitably contain noise. If the gene expression level is close to its maximum ($x_g \approx \mathbf{R}_g/\lambda_g$) or minimum level ($x_g \approx 0$), $\frac{dx_g}{dt}$ is theoretically close to zero, but noise causes $\frac{dx_g}{dt}$ to fluctuate, which might be interpreted as spurious switching events. To avoid this problem, we identify a gene's ON/OFF state as follows. If the gene expression level x_g is increasing (decreasing) at a rate greater than a user-supplied *velocity threshold* v_g^c, then the gene is classified as ON (OFF). Otherwise, if the expression level is above (below) a user-supplied *expression threshold* x_g^c, then the gene is classified as ON (OFF). This can be summarized

as

$$y_g = \begin{cases} \frac{dx_g}{dt} & \frac{dx_g}{dt} \geq v_g^c \\ \\ (x_g - x_g^c) & \frac{dx_g}{dt} < v_g^c. \end{cases} \tag{3.10}$$

In our implementation of FIGR, cubic smoothing splines are fit to time series data and differentiated to estimate velocity. Figure 3.1E,H illustrates the determination of y_g for an example two-gene network.

3.3.2.2 Determining regulatory parameters

Within the Glass model, the ON/OFF state of a particular target gene g, whose index we shall omit from now on, is given by $y = (\mathbf{T} \cdot \mathbf{x} + h)$. Suppose that gene product concentrations have been sampled P times, in time and in one or more conditions or cell types. The gene ON/OFF state y_p is determined for each experimentally measured state vector $\mathbf{x}_p$, $p = 1, 2, 3, \ldots, P$, according to the method described above (Section Determining ON/OFF state). Then, the regulatory parameters can be inferred by finding $\tilde{\mathbf{T}}$ and $\tilde{h}$ such that

$$y_p = \left(\tilde{\mathbf{T}} \cdot \mathbf{x}_p + \tilde{h}\right) \tag{3.11}$$

is satisfied for as many p as possible. Inferring the regulatory parameters therefore reduces to the problem of linear binary classification (Hastie et al., 2009).

There are many well-known supervised learning algorithms for linear binary

classification. We have used both support vector machines (SVM) and logistic regression. An SVM finds a hyperplane buffered by the biggest possible margin such that the number of points $\mathbf{x}_p$ belonging to each class, "gene ON" or "gene OFF", is maximized on opposite sides of the margin zone. This can be accomplished by minimizing the cost function

$$\chi(\mathbf{T}, h, \lambda) = \lambda \mathbf{T}^2 + \sum_{p=1}^{P} L(y_p, \mathbf{x}_p), \tag{3.12}$$

where the first term is a regularization penalty that maximizes the margin. The second term is the hinge loss function, $L(y_p, \mathbf{x}_p) = \max\left(0, 1 - y_p(\mathbf{T} \cdot \mathbf{x}_p + h)\right)$, which is non-zero only for points that transgress their class boundary, each such point contributing an amount proportional to its distance from the margin. The parameter λ is used to choose the relative weight of the penalty and loss terms. Two-class logistic regression models the posterior probability of the ON/OFF state of a point as a logit transformation of its distance from the switching hyperplane. The optimal switching hyperplane can be found by minimizing Equation (3.12) with a Binomial deviance loss function $L(y_p, \mathbf{x}_p) = \log\left\{1 + e^{-y_p(\mathbf{T} \cdot \mathbf{x}_p + h)}\right\}$. Figure 3.1C,F,I illustrate binary classification for an example two-gene network.

Minimization of Equation (3.12) is a convex optimization problem, which can be solved by quadratic programming or the Newton-Raphson method quite efficiently, even for large G. This is the key benefit of the separation of regulatory and kinetic parameters enabled by the Glass equations.

3.3.2.3 *Determining kinetic parameters*

Having identified the ON/OFF state of a gene, y_p, for P measurements of its concentration, x_p, the Glass equations (Eq. 3.3) can be rewritten as

$$v_p = \begin{cases} R - \lambda x_p & \text{if } y_p = +1, \\ -\lambda x_p & \text{if } y_p = -1. \end{cases} \tag{3.13}$$

The velocities $v = \frac{dx}{dt}$ are estimated by differentiating cubic smoothing splines fit to the time series data (Section Determining ON/OFF state). Thus, for any particular gene, Eq. (3.13) takes the form of P equations that are linear in the two unknowns R and λ. This is an overdetermined linear system, so best estimates for R and λ can be extracted by least-squares linear regression. In practice, the error in the spline, and hence in v, is the largest when a gene is switching states. We therefore exclude a user-configurable number of time points nearest to switching events. This method is implemented as the "`slope`" method of FIGR. Alternatively, R and λ can also be determined by fitting Equation 3.5 to the concentration data (see File S1).

3.3.3 Validation of FIGR on synthetic data

As a first test of FIGR, we tested its ability to recover known parameters from synthetic data. In each test, 100 randomly generated gene circuits (Section Validation of FIGR with synthetic data) were simulated using the Glass equations (Eq. 3.3). For each gene circuit, N trajectories starting from random initial starting points

were computed and sampled at N_t time points to obtain synthetic time series data resulting in $N_O = N \times N_t$ observations per simulation. The quality of the inference depends not only on the effectiveness of the method but also on how well determined the inference problem is. A gene circuit of G genes has $N_p = G(G+3)$ parameters. If $N_O \gg N_p$, then the problem is well determined and the accuracy of the inference depends primarily on the effectiveness of the algorithm. On the other hand if $N_O \sim N_p$ then there isn't a sufficient amount of data to infer the parameters accurately irrespective of the effectiveness of the algorithm. We checked how effective FIGR is at inferring parameters of gene circuits of various sizes by exploring different combinations of the number of free parameters and the number of data points. With the exception of the 50-gene network, we also inferred the parameters with SA (Section Inference with SA) to serve as a point of reference. Inference of 100 random 20-gene networks took 5 days on 500 CPUs with SA, and hence it was impractical to infer 50-gene networks.

The inferred parameter values were compared to the known values by computing the discrepancies between them. From the viewpoint of correctly predicting a gene's ON/OFF state, the accuracy of individual genetic interconnectivity coefficients T_{gf} is less important than the accuracy with which the switching hyperplane has been inferred. Accordingly, we judged the accuracy of the genetic interconnectivity matrix by computing the magnitude of the vector difference between the unit normals of the theoretical ($\mathbf{T}_g$) and inferred ($\tilde{\mathbf{T}}_g$) hyperplanes, $\delta_T = \|\tilde{\mathbf{T}}_g - \mathbf{T}_g\|$. When the angle between the unit normals is small, δ_T gives the angle between them. $\delta_T = \sqrt{2}$ implies that the inferred hyperplane is orthogonal

to the theoretical one, and $\delta_T = 2$ is the maximum value possible, implying that the two normals are in exactly opposite directions and the assignment of ON/OFF state has been reversed. We also computed the null distribution of δ_T which results from choosing the inferred unit normal at random uniformly on the unit G-sphere. The discrepancies in the other parameters were computed as $\delta_h = |\tilde{h}_g - h_g|$, $\delta_R = |\tilde{R}_g - R_g|$, and $\delta_\lambda = |\tilde{\lambda}_g - \lambda_g|$, where $\tilde{h}_g$, $\tilde{R}_g$, and $\tilde{\lambda}_g$ are inferred parameter values.

In the first set of simulations, we simulated networks of size ranging from two to fifty genes (Fig. 3.2A–E) with $N = 100$ trajectories sampled at $N_t = 21$ time points. The switching hyperplane is inferred with high accuracy ($\delta_T < 0.1$ or 6 angle to the theoretical normal) for the vast majority of two-gene random gene circuits, showing that FIGR is capable of recovering the true values of the parameters if a sufficient amount of data is available. As the size of the gene circuit, and consequently the number of free parameters, increases, the accuracy declines. However the inference is still fairly accurate for 20-gene networks since 75% of the inferred hyperplanes have $\delta_T < 0.5$ or less than a 30 angle to the theoretical hyperplane. Inferring the signs of the genetic interconnectivity coefficients T_{gf}, that is, whether a regulator activates or represses a target, is an important goal in gene circuit analysis. FIGR achieves 90% accuracy in inferring the signs of T_{gf} for 20-gene networks (SFig. 3.3). For 50-gene networks, the accuracy of quantitative inference is quite low and most inferred hyperplanes have $\delta_T > 0.5$ or larger than a 30 angle to the theoretical hyperplane. This is not entirely surprising since a 50-gene network has 2,650 free parameters, while the model is being inferred from

only 2,100 observations. Although 50-gene networks are inferred with poor accuracy, the δ_T distribution is significantly better than the null distribution, suggesting that the inferred parameters still contain information about GRN structure. In fact, the signs of 82% interconnectivity coefficients are inferred correctly (SFig. 3.1) suggesting that FIGR still works reasonably well at a qualitative level in a highly underdetermined problem.

Hyperplanes inferred with SA also show a trend of decreasing accuracy with increasing network size, suggesting that declining accuracy is a result of progressive reduction in the determinacy of the problem rather than an intrinsic inability of FIGR to infer larger networks. In fact, in nearly all cases, SA's inferences were more variable than FIGR, and were slightly less accurate than FIGR on the 20-gene problem. The relatively lower accuracy of SA is not a result of poor fitting since the RMS of most of the random gene circuit fits is less than 0.04 ($\sim$4% error; SFig. 3.3).

In the second set of simulations, we simulated random 20-gene networks, but varied the number of sampled time points N_t (Fig. 3.2F,G) or the number of trajectories N (Fig. 3.2H–J). Increasing or decreasing the number of time points to 41 or 11 respectively had a minimal effect on the quality of the inference of the switching hyperplanes by FIGR or SA (compare to Fig. 3.2D). This suggested that 11 time points were sufficient to reliably estimate the genetic interconnection coefficients. In contrast, decreasing the number of trajectories progressively reduced accuracy and both FIGR and SA inferences were indistinguishable from the null distribution when only 10 trajectories were sampled. Once again, this is not sur-

prising since the 460 free parameters of a 20-gene network are being inferred from only 210 observations. These results imply that while increasing temporal resolution beyond a certain point provides diminishing gains in accuracy, the number of trajectories or conditions the trajectories are sampled from is a crucial parameter influencing the quality of the inference.

The inference of h_g (SFig. 3.3) was quite accurate and behaved like the inference of the switching hyperplanes. Increasing gene network size or reducing the amount of data tended to reduce accuracy, although the effects were less pronounced than what was observed while inferring switching hyperplanes. The kinetic parameters were also inferred accurately by FIGR (SFig. 3.3). The accuracy of both R_g and λ_g increased with increasing number of time points but did not depend on the number of genes or trajectories. This can be understood as a consequence of the separation of regulatory and kinetic parameters in Glass gene circuits—the inference of the kinetic parameters occurs independently for each gene and depends only on the sampling frequency.

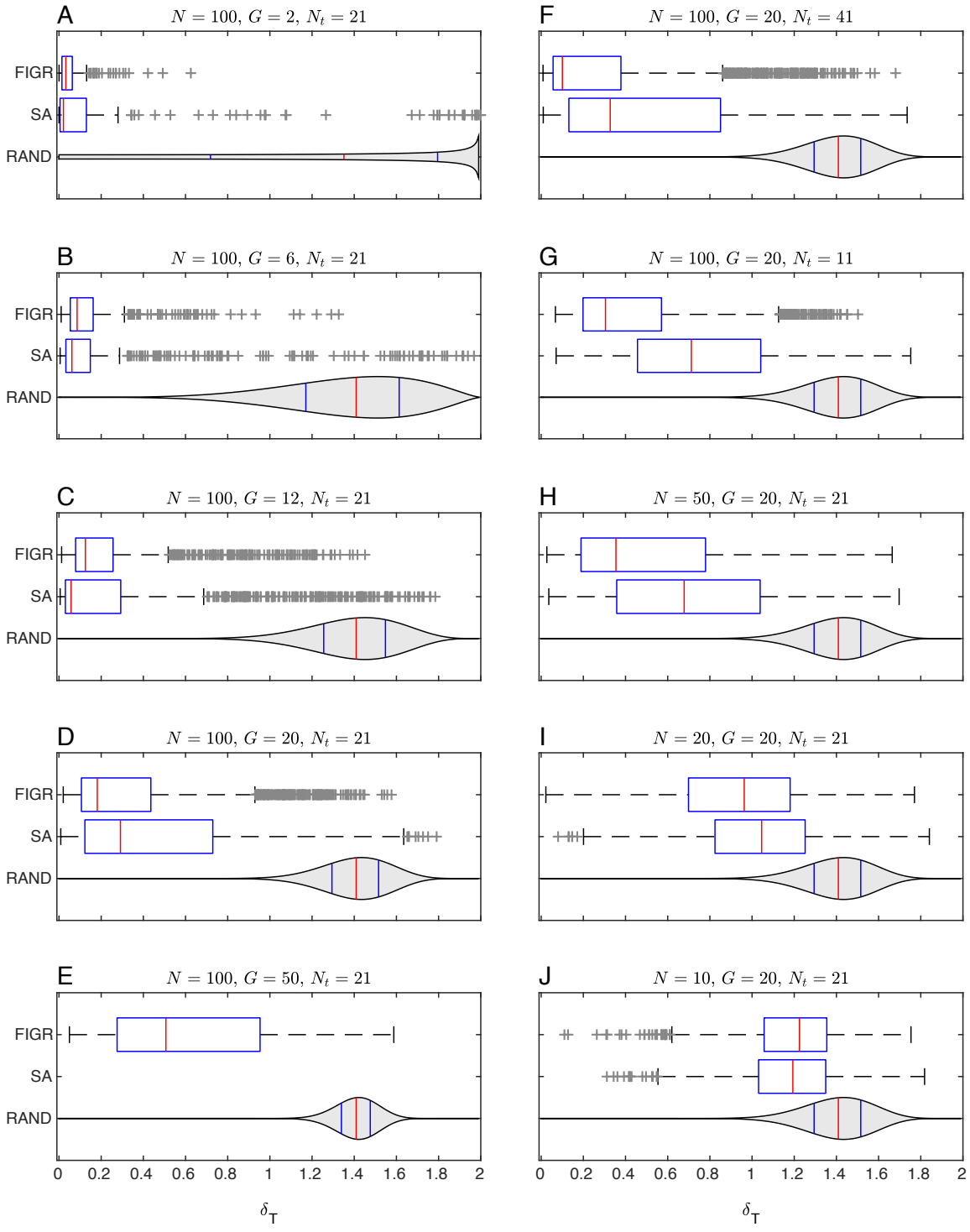


Figure 3.2: Inference of genetic interconnectivity coefficients from synthetic data. The distribution of the discrepancy between inferred and theoretical switching hyperplanes, $\delta_T = \|\tilde{\mathbf{T}}_g - \mathbf{T}_g\|$, in inferring 100 random gene circuits with FIGR or SA is shown as boxplots. $\tilde{\mathbf{T}}_g$ and $\mathbf{T}_g$ are unit normals to the inferred and theoretical hyperplanes.

(continued)

Figure 3.2: Inference of genetic interconnectivity coefficients from synthetic data (continued). Note that each boxplot is constructed from $100G$ discrepancies since each random parameter set contains G switching hyperplanes, where G is the number of genes. The box lines are the first quartile, median, and the third quartile. The whiskers extend to the most extreme values lying within 1.5 times the interquartile range, and any datapoints outside the whiskers are shown as crosses. The blue violin plot ("RAND") shows the null discrepancy distribution that would be obtained, $P_0(\delta_T) \propto {\delta_T}^{G-2}(1-\frac{\delta^2}{4})^{\frac{G-3}{2}}$, if $\tilde{\mathbf{T}}_g$ were not inferred but instead picked randomly from a uniform distribution over the surface of the G-sphere. The width of the violin plot is proportional to $P_0(\delta_T)$, and blue and red vertical lines indicate quartiles and median respectively. **A–E**. Number of trajectories $N = 100$ and number of timepoints $N_t = 21$. The number of genes G was varied between 2 and 50 for FIGR and 2 and 20 for SA, since SA was impractical for $G = 50$. **F,G**. Number of trajectories $N = 100$ and number of genes $G = 20$. The number of timepoints N_t was varied between 41 (panel F) and 11 (panel G). **H–J**. Number of genes $G = 20$ and number of timepoints $N_t = 21$. The number of trajectories N_t was varied between 10 and 50.

Although the inference is quite accurate, the discrepancies are not zero for most gene circuits and can be fairly large for a small number of gene circuits, even in the 2-gene case. This results from constraints imposed by the intrinsic dynamics of gene circuits and finite sample size. For instance, trajectories move away from the switching hyperplane for autoactivating genes. In this case, the initial conditions act as support vectors for the inferred hyperplane, which then strongly depends upon the random sample of starting points. Another situation that arises is that of a hyperplane that divides the bounding hypercube into ON and OFF regions in a lopsided manner. Since initial points are sampled uniformly, this results in too few sampled points in the vicinity of the hyperplane and poor

inference. Given that FIGR was at least as accurate as SA, these failure modes are not specific to the inference methodology but likely represent fundamental limitations of inferring differential equations models. These considerations are also valid when inferring GRNs from empirical data. Such insights and their implications for parameter identifiability will be described elsewhere. Notwithstanding these constraints, our analysis demonstrates that FIGR is capable of inferring parameters quantitatively when provided with a sufficient number of data points and qualitatively (signs of T_{gf}) even when the problem is underdetermined as in the 50-gene case.

3.4 DISCUSSION

Gene circuits (Reinitz & Sharp, 1995; Jaeger et al., 2004a; Manu et al., 2009b) provide many unique advantages for inferring and modeling developmental GRNs. The differential equations are biologically realistic in representing gene regulation as a nonlinear switch-like function of TF concentrations. Gene circuits not only infer the topology of the network but the directionality (causality), sign (activation/repression), and strength of regulatory interconnections. Most importantly, gene circuits are not limited to inference but are capable of accurately simulating and predicting gene expression patterns. Finally, the use of differential equations allows gene circuits to compute transient solutions, an important factor in simulating development since fate determination can occur before equilibrium is reached (Manu et al., 2009a; Simcox & Sang, 1983). Despite the promise held by gene circuits, their application, as of other data-driven differential equation mod-

els, has been limited to smaller networks so far. Analysis of larger networks is limited to correlative approaches (Margolin et al., 2006; Segal et al., 2003) that neither infer causality nor simulate or predict the time evolution of GRN state.

A major challenge in broader application of gene circuits is the high computational expense of inferring the free parameters from time series data. Currently, the approach for inferring parameter values (Chu et al., 1999; Reinitz & Sharp, 1995; Kozlov et al., 2012; Abdol et al., 2017) is to solve ("integrate") the ODEs to obtain trajectories, compare with experimental trajectories, and refine parameters using global optimization techniques such as SA. This procedure is slow and expensive because it requires performing multidimensional optimization on a complicated cost function $\chi^2(\{T, h, R, \lambda\})$ with many local minima and each function evaluation involves solving a system of ODEs. Moreover, the computational complexity grows rapidly ($O(G^3)$) so that global optimization approaches for gene circuits scale poorly with G.

In contrast, FIGR directly attempts to fit the differential equations, which describe how the velocities v_g depend upon the concentrations x_g. $\mathbf{T}_g$ and h_g are determined using binary classification (support vector machines or logistic regression). Both of these algorithms reduce to quadratic programming, and thence to convex optimization. Subsequently, R_g and λ_g can be determined by linear regression against velocities or non-linear regression against concentrations using the piece-wise analytical solutions of the ODEs, which are even simpler optimization problems. Each inference can be completed in a fraction of a second on a consumer-grade computer, even with interpreted MATLAB code.

The computational efficiency of FIGR does not come at the expense of accuracy. In testing the recovery of known parameters from synthetic data (Fig. 3.2, S1, and S3), we found that FIGR and SA had comparable accuracy for smaller gene circuits, while FIGR had slightly higher accuracy than SA for 20-gene networks. We speculate that the lower accuracy of SA results from "sloppiness" (Gutenkunst et al., 2007; Ashyraliyev et al., 2008; Kozlov et al., 2012)—insensitivity of model output to certain genetic interconnectivity coefficients. If a certain parameter gives similar solutions over some interval, then SA can infer any value in the interval rather than the true value. This insensitivity can result from the compensatory and redundant roles many parameters play in the model (Ashyraliyev et al., 2008; Kozlov et al., 2012). For example, a high expression level can be achieved by having higher activating genetic interconnection coefficients, by having higher synthesis rates, or by having lower degradation rates. The higher accuracy of FIGR could perhaps be attributed to the separation of the regulatory and kinetic parameters, which limits the opportunities available for redundant parameters to produce similar solutions.

In representing synthesis as a binary ON/OFF choice, Glass equations (Eq. 3.3) are similar to Boolean or logical models, which have been applied to a broad range of developmental GRNs (Theiffry et al., 1993; Sánchez & Thieffry, 2001; Davidson et al., 2002b; Thieffry & Sánchez, 2003; Collombet et al., 2017; Bonzanni et al., 2013). Given this similarity between Boolean models and Glass equations, FIGR should be readily applicable to a large class of GRN modeling problems. Moreover, Glass equations relax the assumption made in logical models—that genes

are expressed at a small number of discrete levels—to allow expression at any arbitrary level. This makes Glass models more general than Boolean models and capable of simulating transient dynamics during development.

Although in our tests FIGR was shown to be at least as effective as, and much faster than, SA, it does have a few limitations. First, in order to estimate the velocity and determine ON/OFF state reliably, FIGR requires that the data be sampled sufficiently frequently in time. Roughly speaking, during each time period in which a gene is in a particular state (ON or OFF), its product concentration would have to be sampled at least three times in order to ascertain the velocity and state. FIGR therefore would not be suitable for datasets that have been sampled sparsely in time. Methods reliant on solving the ODEs will, in contrast, attempt to fit the trajectories to a sparsely sampled dataset, even if the actual inference achieved is poor (Fig. 3.2J).

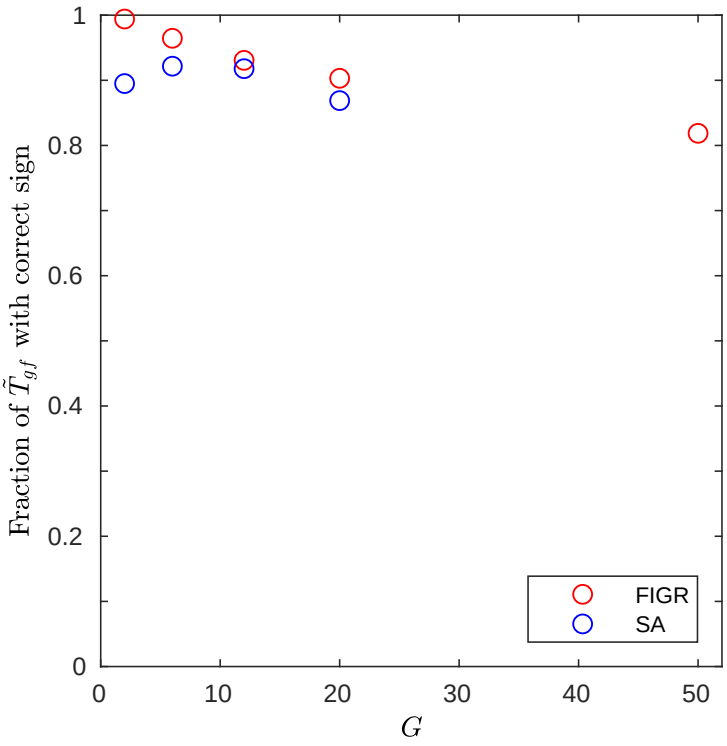
Besides the problem of computational efficiency, the broader application of gene circuits, and indeed all nonlinear differential equation models, is limited by a lack of understanding of parameter identifiability. Most commonly, *a posteriori* confidence intervals for the parameter estimates are computed (Ashyraliyev et al., 2008). Such calculations are based on the strong assumption that the solution is linear in the parameters and that the measurement errors are normally distributed. *a posteriori* parameter identifiability analysis also does not provide any hints to improve experimental design for achieving better identifiability in future studies.

Although we have not directly addressed the problem here, we anticipate that

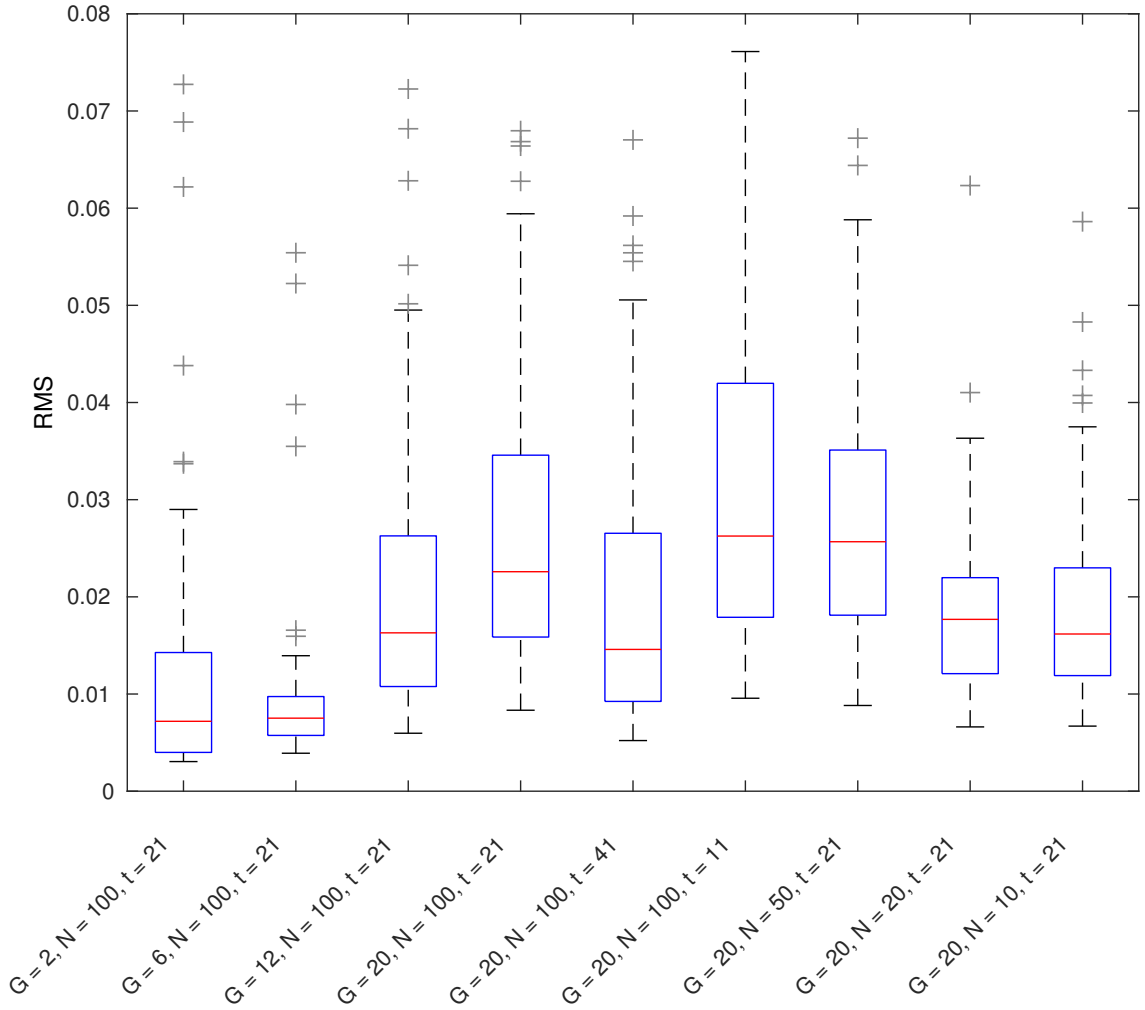
conceiving of and visualizing gene circuit inference as a classification problem will lead to insights into parameter identifiability. For example, it is evident from state space plots (Fig. 3.1F,I) that sampling gene expression trajectories closer to the true switching hyperplane of the gene will lead to less "wiggle room" for the inferred hyperplane and result in more accurate parameter inference. This implies that datasets that measure gene expression near steady state, for example in differentiated cell types, are unlikely to lead to accurate parameter inference irrespective of the number of data points sampled or the precision of the experiment. Sampling transient trajectories densely when genes are turning ON or OFF is the best strategy for accurate parameter inference. Another less obvious implication is that it is easier to infer the regulation of negatively autoregulated genes than positively autoregulated ones. Trajectories move toward or away from the switching hyperplane for negatively or positively autoregulated genes respectively, making it more likely that sampled data points will lie near the hyperplane in the former. This analysis will be reported elsewhere.

In summary, we have exploited features of the mathematical structure of gene circuits to break a difficult optimization problem into a series of two, much simpler, optimization problems. We have demonstrated that FIGR is effective on synthetic as well as experimental data from a biologically realistic GRN. We have validated the inferred gap gene model by comparing its parameters against models inferred with SA as well as comparing its output against experimental data. The improvement in computational efficiency and scalability should allow the inference of much larger GRNs than was possible previously.

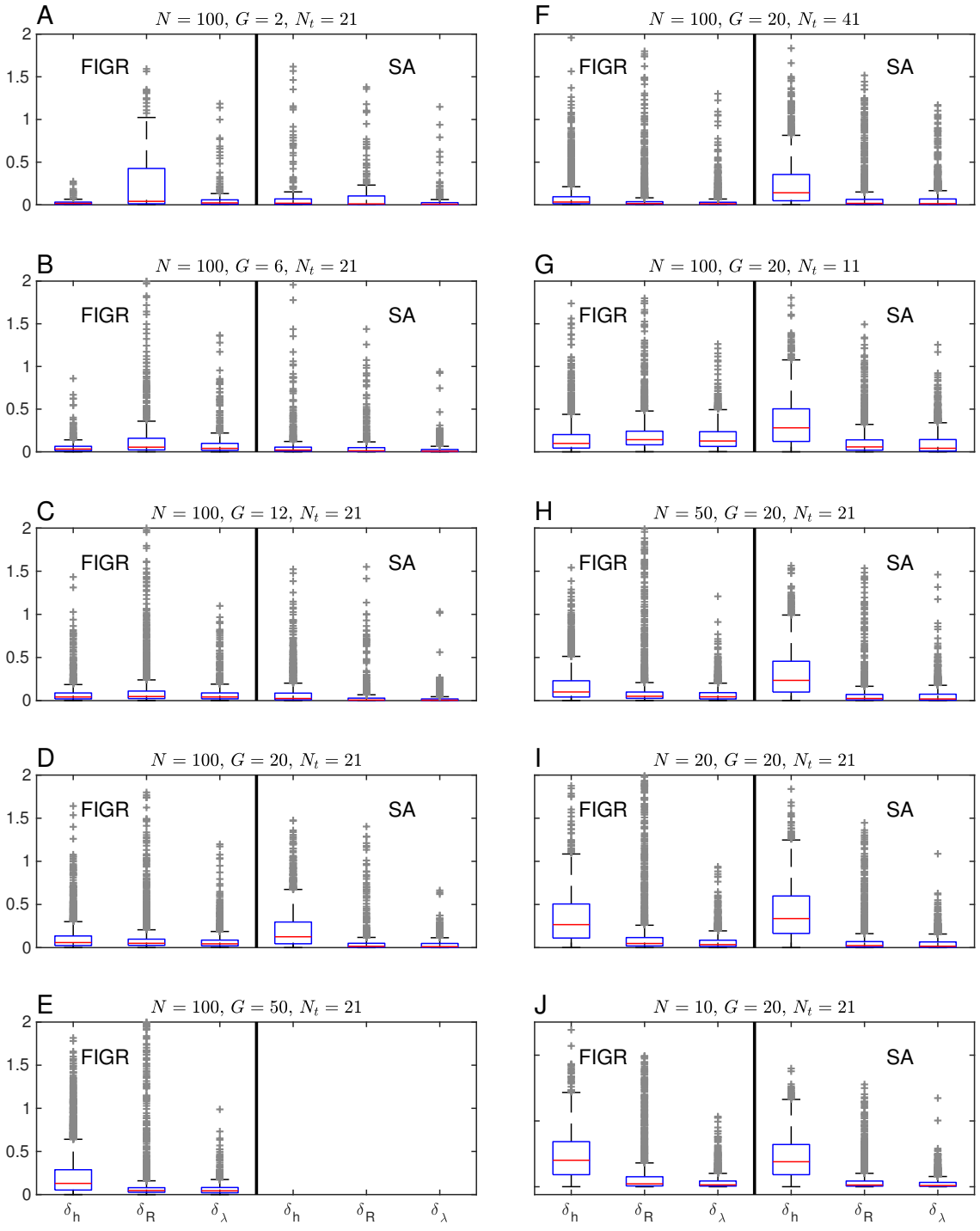
3.5 SUPPLEMENTARY DATA



SFigure 3.1: Fraction of genetic interconnectivity signs inferred correctly from synthetic data. The fraction of genetic interconnectivity coefficients for which the sign was inferred correctly $(sgn(\tilde{T}_{gf}) = sgn(T_{gf}))$ is shown. $\tilde{T}_{gf}$ and T_{gf} are inferred and theoretical genetic interconnectivity coefficients. $N = 100$ and $N_t = 21$. SA was not run for $G = 50$ since it was impractical.



SFigure 3.2: Training error of SA-inferred gene circuits. Boxplots show root mean square (RMS) error of gene circuits inferred from 100 synthetic datasets using SA. For each combination of the number of genes G, the number of trajectories N , and the number of timepoints N_t , synthetic datasets corresponding to 100 random parameter sets were generated. SA was used to infer 5 replicate gene circuits from each dataset. The replicate with the lowest RMS is included in the plot. The box lines are the first quartile, median, and the third quartile. The whiskers extend to the most extreme values lying within 1.5 times the interquartile range, and any datapoints outside the whiskers are shown as dots.



SFigure 3.3: Inference of h, g, and kinetic parameters from synthetic data. The discrepancy between inferred and theoretical values of h_g, R_g, and g from 100 random parameter sets is shown as boxplots. Note that each boxplot is constructed from $100G$ parameter values since each random parameters set contains G values of h_g, R_g , and g , where G is the number of genes.

(continued)

SFigure 3.3: Inference of h, g, and kinetic parameters from synthetic data (continued). The box lines are the first quartile, median, and the third quartile. The whiskers extend to the most extreme values lying within 1.5 times the interquartile range, and any datapoints outside the whiskers are shown as crosses. AE. Number of trajectories N = 100 and number of timepoints $N_t = 21$. The number of genes G was varied between 2 and 50 for FIGR and 2 and 20 for SA. F,G. Number of trajectories $N = 100$ and number of genes $G = 20$. The number of timepoints N_t was varied between 41 (panel F) and 11 (panel G). HJ. Number of genes $G = 20$ and number of timepoints $N_t = 21$. The number of trajectories N t was varied between 10 and 50. SA was not run for $G = 50$ since it was impractical.

Alternative method for determining kinetic parameters In the diffusion-less case, in addition to fitting the Glass equations to velocity data (Eq. 3.13), the kinetic parameters R and λ can also be determined by fitting Glass equation solutions (Eq. 3.5) to the concentration time series data. We identify time intervals during which all y are either positive or negative so that

$$x_m(t_k) = \begin{cases} x_m(t_0)e^{-\lambda\Delta t_k} + \frac{R}{\lambda}(1 - e^{-\lambda\Delta t_k}) & \text{if } y_m(t_k) = +1 \quad \forall k, \\ x_m(t_0)e^{-\lambda\Delta t_k} & \text{if } y_m(t_k) = -1 \quad \forall k, \end{cases} \tag{3.14}$$

where m and k index the time intervals and the time points lying inside a particular interval respectively. Within a particular interval, $x_m(t_k)$ is the concentration at the kth time point, $x_m(t_0)$ is the initial concentration, and $\Delta t_k = t_k - t_0$ is the time elapsed from the start of the interval. Equations 3.14 are $P \gg 2$ non-linear equations with two unknowns, R and λ, and can be fit relatively easily using off-the-shelf non-linear optimization methods. We used MATLAB's `lsqnonlin`

STable 3.1: User-defined options and parameters utilized in FIGR code

User-defined options and parameters utilized in FIGR code The spline smoothing parameter is passed to the spline-fitting `csaps` function of MATLAB. It takes values between 0 and 1, where 1 implies no smoothing while 0 results in a straight-line fit

			Synthetic Data Tests	Gap Gene Inference
Determining regulatory parameters				
Spline smoothing parameter for determining velocities	`splinesmoothing`	$[0, 1]$	1	0.01
Velocity threshold for determining on/off state	`slopethresh` (v_g^c)	≥ 0	0.01	1
Expression threshold for determining on/off state	`exprthresh` (x_g^c)	> 0	0.2	100
Determining kinetic parameters				
Method for determining the kinetic parameters	`Rld_method`	`'slope'` `'kink'` `'conc'`	`'slope'`	`'kink'`
Determining kinetic parameters by "slope" method				
Margin to exclude unreliable velocity estimates near maxima and minima of the time series	`Rld_tsafety`	≥ 0	3	NA
Determining kinetic parameters by "kink" method				
Spline smoothing parameter for identifying spatial expression domains and border positions	`spatialsmoothing`	$[0, 1]$	NA	0.5
Expression threshold above which points are included in fitting the kink equations, expressed as fraction of maximum domain expression	`minborder_expr_ratio`	$(0, 1)$	NA	0.01

function that implements a Trust-Region algorithm. This is implemented as the "`conc`" method of FIGR.

BIBLIOGRAPHY

Abdol, A. M., Cicin-Sain, D., Kaandorp, J. A., & Crombach, A. (2017). Scatter search applied to the inference of a development gene network. *Computation*, *5*(2).

Akam, M. (1987). The molecular basis for metameric pattern in the *Drosophila* embryo. *Development*, *101*, 1–22.

Ashyraliyev, M., Jaeger, J., & Blom, J. G. (2008). Parameter estimation and determinability analysis applied to drosophila gap gene circuits. *BMC Syst Biol*, *2*, 83.

Balaskas, N., Ribeiro, A., Panovska, J., Dessaud, E., Sasai, N., Page, K. M., Briscoe, J., & Ribes, V. (2012). Gene regulatory logic for reading the sonic hedgehog signaling gradient in the vertebrate neural tube. *Cell*, *148*(1-2), 273–84.

Bonzanni, N., Garg, A., Feenstra, K. A., Schütte, J., Kinston, S., Miranda-Saavedra, D., Heringa, J., Xenarios, I., & Göttgens, B. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics*, *29*(13), i80–8.

Chu, K. W., Deng, Y., & Reinitz, J. (1999). Parallel simulated annealing by mixing of states. *The Journal of Computational Physics*, *148*, 646–662.

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc Natl Acad Sci U S A*, *114*(23), 5792–5799.

Cotterell, J., & Sharpe, J. (2010). An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol. Syst. Biol.*, *6*(1), 425.

Crombach, A., Wotton, K. R., Cicin-Sain, D., Ashyraliyev, M., & Jaeger, J. (2012). Efficient reverse-engineering of a developmental gene regulatory network. *PLoS Comput Biol*, *8*(7), e1002589.

Davidson, E. H., & Levine, M. S. (2008). Properties of developmental gene regulatory networks. *Proc Natl Acad Sci U S A*, *105*(51), 20063–6.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Rust, A. G., Pan, Z. j., Schilstra, M. J., Clarke, P. J. C., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., & Bolouri, H. (2002a). A genomic regulatory network for development. *Science*, *295*(5560), 1669–78.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Schilstra, M. J., Clarke, P. J., Rust, A. G., Pan, Z., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., & Bolouri, H. (2002b). A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Developmental Biology*, *246*, 162–190.

Edwards, R. (2000). Analysis of continuous-time switching networks. *Physica D: Nonlinear Phenomena*, *146*(1–4), 165 – 199.

Fehr David A., J. E., Handzlik, Manu, . ., & Loh, Y. L. (2019). Classification-based inference of dynamical models of gene regulatory networks. *G3: Genes, Genomes, Genetics*, *9*(12), 4183–4195.

Glass, L., & Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *The Journal of Theoretical Biology*, *39*, 103–129.

Glass, L., & Pasternack, J. S. (1978). Stable oscillations in mathematical models of biological control systems. *Journal of Mathematical Biology*, *6*(3), 207.

Gursky, V. V., Jaeger, J., Kozlov, K. N., Reinitz, J., & Samsonova, A. M. (2004). Pattern formation and nuclear divisions are uncoupled in *Drosophila* segmentation: comparison of spatially discrete and continuous models. *Physica D*, *197*, 286–302.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology*, *3*(10), 1–8.

Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K., & Göttgens, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, *114*(23), 5822–5829.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hengenius, J. B., Gribskov, M., Rundell, A. E., Fowlkes, C. C., & Umulis, D. M. (2011). Analysis of gap gene regulation in a 3d organism-scale model of the drosophila melanogaster embryo. *PLOS ONE*, *6*(11), 1–12.

Hong, T., Xing, J., Li, L., & Tyson, J. J. (2012). A simple theoretical framework for understanding heterogeneous differentiation of cd4+ t cells. *BMC Syst Biol*, *6*, 66.

Huang, S., Guo, Y., May, G., & Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, *305*, 695–713.

Jaeger, J., Blagov, M., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Surkova, S., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004a). Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. *Genetics*, *167*, 1721–1737.

Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004b). Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, *430*, 368–371.

Jostins, L., & Jaeger, J. (2010). Reverse engineering a gene network using an asynchronous parallel evolution strategy. *BMC Systems Biology*, *4*(1), 17.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Kozlov, K., & Samsonov, A. (2009). Deep - differential evolution entirely parallel method for gene regulatory networks. In V. Malyshkin (Ed.) *Parallel Computing Technologies*, (pp. 126–132). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kozlov, K., Surkova, S., Myasnikova, E., Reinitz, J., & Samsonova, M. (2012). Modeling of gap gene expression in drosophila kruppel mutants. *PLoS Comput Biol*, *8*(8), e1002635.

Kueh, H. Y., Champhekhar, A., Nutt, S. L., Elowitz, M. B., & Rothenberg, E. V. (2013). Positive feedback between pu.1 and the cell cycle controls myeloid differentiation. *Science*.

Lam, J., & Delosme, J.-M. (1988a). An efficient simulated annealing schedule: Derivation. Tech. Rep. 8816, Yale Electrical Engineering Department, New Haven, CT.

Lam, J., & Delosme, J.-M. (1988b). An efficient simulated annealing schedule: Implementation and evaluation. Tech. Rep. 8817, Yale Electrical Engineering Department, New Haven, CT.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., & Singh, H. (2006). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, *126*(4), 755–66.

Levine, M., & Davidson, E. H. (2005). Gene regulatory networks for development. *Proc Natl Acad Sci U S A*, *102*(14), 4936–42.

Li, C., & Wang, J. (2013). Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: Landscape and biological paths. *PLoS Comput Biol*, *9*(8), e1003165 EP –.

Lou, Z., & Reinitz, J. (2016). Parallel simulated annealing using an adaptive resampling interval. *Parallel Comput.*, *53*, 23–31.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009a). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. *PLoS Computational Biology*, *5*, e1000303. Doi:10.1371/journal.pcbi.1000303.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009b). Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biology*, *7*, e1000049. Doi:10.371/journal.pbio.1000049.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7 Suppl 1*, S7.

May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen,

D. G., Peterson, C., & Enver, T. (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, *13*(6), 754–68.

Mestl, T., Lemay, C., & Glass, L. (1996). Chaos in high-dimensional neural and gene networks. *Phys. D*, *98*(1), 33Ű52.

Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T., Frampton, G. M., Drake, A. C. B., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J. W., Liefeld, T., Smutko, J. S., Chen, J., Friedman, N., Young, R. A., Golub, T. R., Regev, A., & Ebert, B. L. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, *144*(2), 296–309.

Palani, S., & Sarkar, C. A. (2008). Positive receptor feedback during lineage commitment can generate ultrasensitivity to ligand and confer robustness to a bistable switch. *Biophys J*, *95*(4), 1575–89.

Pietak, A., Bischof, J., LaPalme, J., Morokuma, J., & Levin, M. (2019). Neural control of body-plan axis in regenerating planaria. *PLOS Computational Biology*, *15*(4), 1–35.

Reinitz, J., & Sharp, D. H. (1995). Mechanism of *eve* stripe formation. *Mechanisms of Development*, *49*, 133–158.

Sánchez, L., & Thieffry, D. (2001). A logical analysis of the *Drosophila* gap-gene system. *The Journal of Theoretical Biology*, *211*, 115–141.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, *34*(2), 166–76.

Simcox, A. A., & Sang, J. H. (1983). When does determination occur in *Drosophila* embryos? *Developmental Biology*, *97*, 212–221.

Theiffry, D., Colet, M., & Thomas, R. (1993). Formalization of regulatory networks: A logical method and its automatization. *Mathematical Modelling and Scientific Computing*, *2*, 144–151.

Thieffry, D., & Sánchez, L. (2003). Dynamical modelling of pattern formation during embryonic development. *Current Opinion in Genetics and Development*, *13*, 1–5.

Tyson, J. J., Baumann, W. T., Chen, C., Verdugo, A., Tavassoly, I., Wang, Y., Weiner, L. M., & Clarke, R. (2011). Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nat Rev Cancer*, *11*(7), 523–32.

Verd, B., Clark, E., Wotton, K. R., Janssens, H., Jimnez-Guri, E., Crombach, A., & Jaeger, J. (2018). A damped oscillator imposes temporal order on posterior gap gene expression in drosophila. *PLOS Biology*, *16*(2), 1–24.

Weston, B. R., Li, L., & Tyson, J. J. (2018). Mathematical analysis of cytokine-induced differentiation of granulocyte-monocyte progenitor cells. *Front. Immunol.*, *9*, 2048.

Wu, H., Manu, Jiao, R., & Ma, J. (2015). Temporal and spatial dynamics of scaling-specific features of a gene regulatory network in drosophila. *Nat Commun*, *6*, 10031.

CHAPTER 4
Neutrophil-macrophage differentiation in PUER cells

Chapter 2 presented a data-driven model simulating a 12-gene GRN for erythrocyte-neutrophil differentiation inferred by training on a high temporal resolution dataset (May et al., 2013). Inferring comprehensive and accurate GRN models requires frequent sampling of gene expression over the entire course of differentiation. Besides the requirement for frequent sampling, one of the insights gained in Chapter 2 was that a limited training dataset may not contain sufficient information to accurately infer regulatory parameters for certain genes. For example, the poorly constrained parameters of *Gata2* might have been the result of *Gata2* having very similar expression patterns in the two conditions. Potential remedies for the lack of information in a dataset would be to either supplement with genetic perturbation data or to include high temporal resolution data from other lineages during training, to increase the amount of information for constraining parameters. This chapter presents the analysis of a high-resolution time-series gene expression dataset from the *in vitro* differentiation of macrophages and neutrophils. These data provide an extra layer of information on gene expression dynamics in differentiating neutrophils and macrophages that should help us further understand the causality of regulatory events and develop more comprehensive models. The differentiation experiment was performed by Andrea Repele. I performed the data pre-processing and analysis.

4.1 INTRODUCTION

Macrophages and neutrophils are developmentally closely related cell types that share a common progenitor, the granulocyte-monocyte progenitor (GMP) (Görgens et al., 2013; Keohane, 2020). PU.1, a member of the *ets* family of TFs, plays a crucial role in the development of white blood cells (Scott et al., 1994). PU.1 is expressed in a cell-type specific manner and different levels of PU.1 expression are associated with different types of white blood cells. High levels of PU.1 produce macrophages, while intermediate levels of PU.1 induce the differentiation of neutrophils (Scott et al., 1994; Zhang et al., 1997; Dahl et al., 2003). Another key TF regulating macrophage-neutrophil differentiation is *Cebpa*. *Cepba*$^{-/-}$ mice have neutropenia while enforced expression of *Cebpa* in B-cell precursors reprograms them into macrophages (Xie et al., 2004). Based on overexpression experiments in bone-marrow progenitors from *Spi1*$^{-/-}$ mice, the macrophage-neutrophil decision is thought to be determined by the ratio of PU.1 and C/EBPα proteins, with a larger ratio favoring the macrophage fate and a smaller ratio favoring the neutrophil fate. PU.1 and C/EBPα are thought to promote the expression of *Egr1/Egr2/Nab2* and *Gfi1* respectively that repress the alternative fate. This GRN has been modeled as a bistable switch (Laslo et al., 2006). Genetic perturbations and chromatin immunoprecipitation followed by sequencing from single murine cells revealed another bistable switch regulating macrophage and neutrophil cell-fates, this time comprised of mutually repressive Irf8 and Gfi1 (Olsson et al., 2016). Aside from these few key regulators thought to govern the cell-fate decision, the larger GRN remains uncharacterized (Wang et al., 2020). Furthermore, despite in-

tensive study in this area, the how these TFs initiate the macrophage or neutrophil gene expression programs and especially the temporal sequence and causality of events remains poorly understood (Keightley et al., 2017; Guanglan et al., 2020).

In order to discover the larger GRNs controlling the macrophage-neutrophil decision and the initiation of the macrophage and neutrophil gene expression programs, we acquired a high temporal resolution RNA-Seq dataset of *in vitro* macrophage-neutrophil differentiation. We utilized PU.1 estrogen receptor (PUER) cells, an important model system for macrophage-neutrophil differentiation (Dahl et al., 2003; Bertolino et al., 2016; Repele et al., 2019a). PUER cells are an IL-3 dependent hematopoietic progenitor cell line derived from the fetal liver of PU.1$^{-/-}$ mice in which PU.1 has been reintroduced after fusion to the ligand binding domain of the estrogen receptor (Walsh et al., 2002). The estrogen receptor domain is preferentially regulated by tamoxifen (OHT), which when present in the cells allows the PUER fusion protein to act as a TF. PUER cells can be maintained indefinitely as bipotential progenitors by culturing in IL-3 media and can be differentiated into macrophages by OHT treatment. PUER cells can also be differentiated into neutrophils by substituting G-CSF for IL-3 and inducing with OHT.

PUER cells were differentiated into macrophages and neutrophils and genome-wide gene expression was assayed by RNA-Seq at 29 timepoints along both lineages over the course of seven days (Fig. 4.8). The analysis of the RNA-seq data revealed rich temporal gene expression patterns during the differentiation of PUER cells and found thousands of differentially expressed genes (DEGs) between the endpoints of differentiation. Correlations between samples and principal com-

ponents analysis reveal there is a sudden large-scale transition occurring around $8-12$h after OHT addition. Furthermore, genes expressed differentially between consecutive time points were also identified. The first 12h of differentiation had the highest number of DEGs, suggesting that the differentiation has the highest momentum at early time points. Moreover, gene ontology (GO) analysis of early DEGs revealed that comparisons until 4h were enriched in myeloid differentiation GO terms but the 4h/8h and 8h/12h comparisons were not enriched in terms for myeloid differentiation but for metabolic and ribosome biogenesis terms. These findings suggest that the first observed transition around 8h corresponds to large-scale physiological reprogramming and that the cell-fate decision is made in the first few hours of differentiation.

4.2 RESULTS

4.2.1 Global changes in the gene expression landscape during myeloid differentiation

As a first step we sought to uncover temporal patterns of changes in genome-wide gene expression. We computed Pearson correlation of genome-wide gene expression between all pairs of time points (Fig. 4.1). As one would expect, the correlation coefficient is higher for nearby time points and reduces as the difference between the time points increases. The largest effect is that of G-CSF pre-treatment (-48h vs. 0h; Fig. 4.1A), which can be understood as the result of the large time difference of 48 hours and the important role that G-CSF plays in the growth and maturation of hematopoietic cells (Franzke, 2006). Unexpectedly, we found that

time points could be divided into at least two groups, early (< 12 hours) and late (≥ 12 hours), so that timepoints have much higher correlation within each group than to timepoints from the other group. This is most easily discerned in the macrophage differentiation in IL-3 conditions (Fig. 4.1B). For example, the 16h timepoint has very low correlation, $0 \geq r < 0.2$, with the 2h timepoint in the early group, occurring only 14 hours earlier, but has very high correlation, $r > 0.8$, with the similarly spaced 32h timepoint in the late group. Although post-OHT timepoints have higher correlation overall, a similar pattern is discernible in the G-CSF condition. That the differentiation can be divided into two phases suggests that there is a large-scale transition in genome-wide gene expression patterns occurring around $8 - 12$ hours.

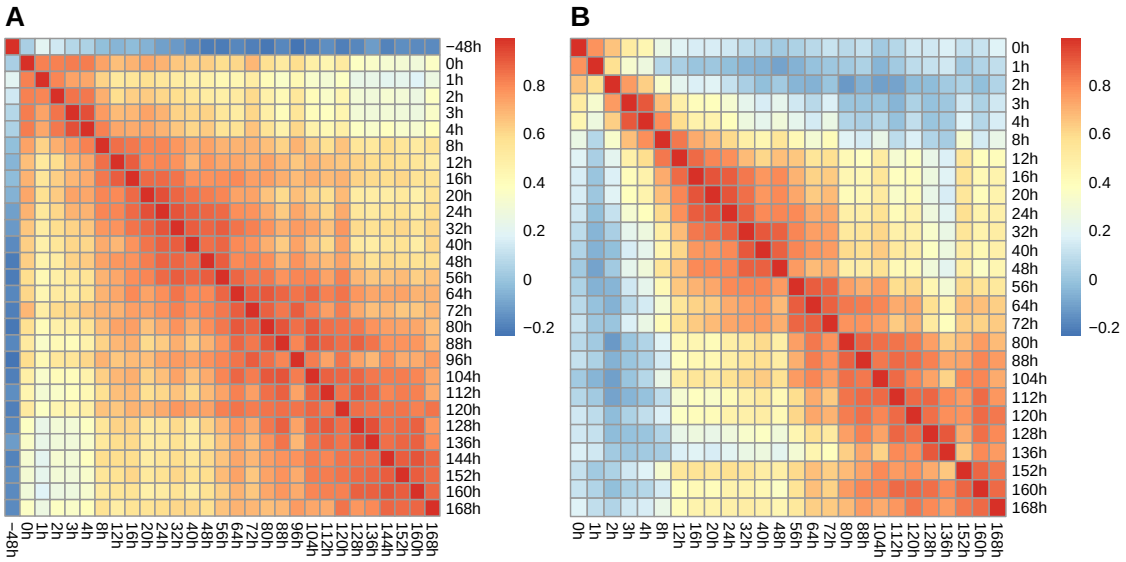


Figure 4.1: Heatmap of genome-wide gene expression correlation. Pearson correlation coefficient between all pairs of timepoints in G-CSF (A) and IL-3 (B) conditions. The Pearson correlation coefficient was calculated between the mean genome-wide gene expression of all pairs of time points. The expression of each gene was scaled between 0 and 1 beforehand.

We further characterized the global patterns of gene expression time evolution using principal component analysis (PCA) (Fig. 4.2). The first two principal component axes accounted for 45% of the total variance in the data, hinting that there are two main effects and that the two-dimensional space is a good approximation to the high-dimensional gene expression space. The G-CSF and IL-3 conditions follow distinct differentiation trajectories and are separated by a large shift along the PC1 axis, occurring during the G-CSF pre-treatment, which implies that the first principal component corresponds to the effect of G-CSF. In both the G-CSF and IL-3 conditions, the trajectories move along the PC2 axis after OHT addition

at 0h, suggesting that the second principal component corresponds to the effect of PU.1 and time. The displacement between the 8h and 12h IL-3 time points is the second largest after that of G-CSF pre-treatment, which implies a very large rate of change in genome-wide gene expression given that it occurs in only 4 hours compared to the 48 hour duration of G-CSF pre-treatment. Similar but smaller jumps are observed between the 4h and 8h G-CSF and the 72h and 80h IL-3 time points, while there is a sharp reversal of the direction of movement at the 136h IL-3 timepoint. These jumps corroborate the inference drawn from the Pearson correlation analysis (Fig. 4.1) that there is a large-scale transition in the pattern of genome-wide gene expression around $8 - 12$ hours after OHT induction, and indicate that there are other such transitions occurring at later stages of the differentiation as well.

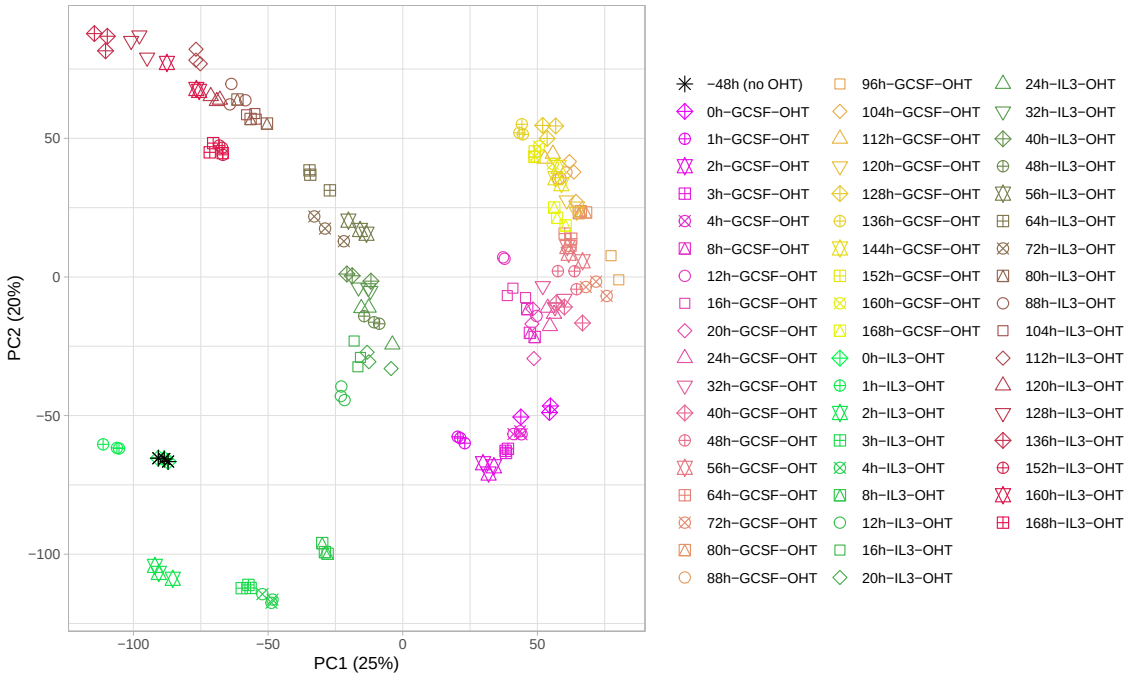


Figure 4.2: Principal components analysis of all the samples. Gene expression was standardized to have zero mean and unit variance. The samples are plotted along the first two principal components, PC1 and PC2, that account for 45% of the total variance.

4.2.2 Diversity of temporal patterns of transient gene expression

In order to gain insight into the genome-wide transitions occurring during the course of differentiation we next analyzed the temporal patterns of the expression of individual genes. We enriched for genes likely to be regulated during the differentiation process by first identifying genes expressed differentially between the end points, undifferentiated PUER cells, -48h G-CSF or 0h IL-3, and 7-day OHT treated samples, 168h G-CSF or IL-3 (Fig. 4.3A,D). 43% and 27% of genes were differentially expressed between the end points in G-CSF and IL-3 conditions re-

spectively. The neutrophil differentiation is the compounded effect of G-CSF and OHT treatments and one may discern between the two by identifying genes differentially expressed because of G-CSF pre-treatment, by comparing -48h samples to 0h G-CSF samples, and those differentially expressed before and after 7 days of OHT treatment, by comparing 0h to 168h G-CSF samples (Fig. 4.3B,C). Consistent with both the correlation analysis and PCA, more genes are differentially expressed due to G-CSF pre-treatment (37%) than due to OHT treatment (29%), even though the latter is conducted over a larger time interval. There is significant overlap in the genes differentially expressed in the two conditions with common DEGs comprising 73% and 44% of all the genes differentially expressed between IL-3 endpoints and G-CSF endpoints respectively (Fig. 4.3E,F).

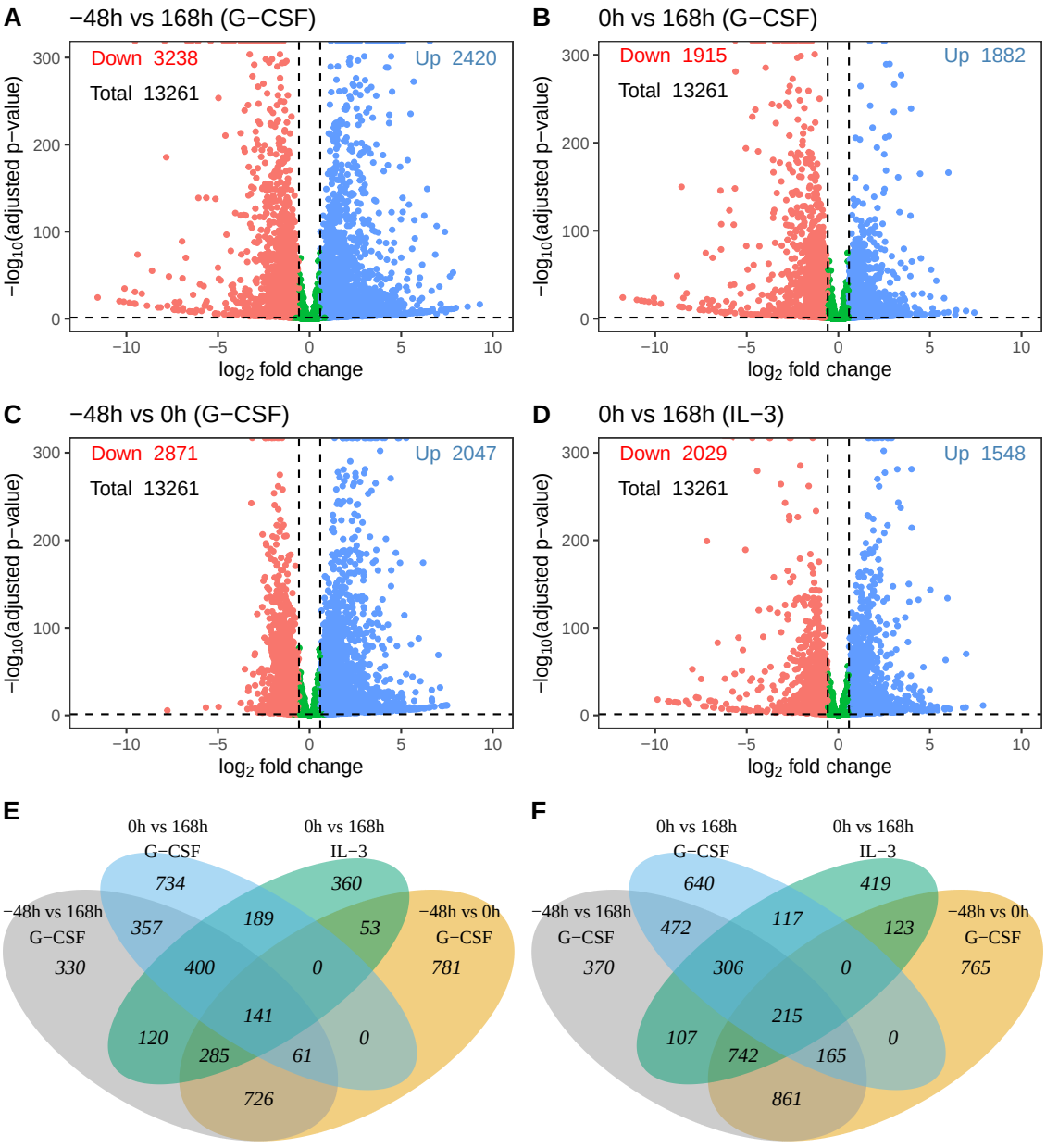


Figure 4.3: The identification of genes expressed differentially between the endpoints of the differentiation. A–D) Scatter plots of p-value vs. fold change for all the genes. The p-value and fold change thresholds used to identify DEGs (Section 2.4) are shown as horizontal and vertical dashed lines respectively. (A) Comparison of undifferentiated PUER cells (−48h) with cells treated with OHT for 7 days in G-CSF conditions (168h).

(continued)

Figure 4.3: The identification of genes expressed differentially between the endpoints of the differentiation. (continued). (B) Comparison of PUER cells pre-treated with G-CSF for 48 hours (0h) with cells treated with OHT for 7 days in G-CSF conditions (168h). (C) Comparison of undifferentiated PUER cells (-48h) with those pre-treated with G-CSF for 48 hours (0h). (D) Comparison of undifferentiated PUER cells (-48h) with those treated with OHT for 7 days in IL-3 conditions (168h). (E,F) Overlap of DEGs for selected time points for up-regulated (E) and down-regulated (F) genes.

The temporal expression patterns of the differentially expressed genes are very diverse and show extensive transient regulation, in which expression at the start and end of differentiation is similar but is modulated in the middle (Fig. 4.4). In order to better reveal broader patterns, the genes were clustered hierarchically according to the similarity of their temporal expression patterns. Several patterns are noticeable. Consistent with all the previous analyses, G-CSF pre-treatment exerts significant effect on the gene expression, with a large number of genes turning off and a smaller but still sizeable group turning on at 0h in the G-CSF treatment (Fig. 4.4A). Another significant shift in the gene expression occurs around the $8-12$h time point, when a large number of genes are upregulated and a smaller number of genes are downregulated. Furthermore, the number of genes coordinately regulated in this manner is greater in the IL-3 condition than the G-CSF condition. Also discernible in the IL-3 condition, but less so in the G-CSF condition, are several waves of transient gene upregulation and downregulation during the first 8 hours of differentiation, with different groups of genes peaking at different timepoints. To summarize, the temporal gene expression patterns further

corroborate the transitions inferred previously, show that the first $8 - 12$ hours of differentiation involve rapid changes in expression, and reveal extensive transient and coordinated regulation of groups of genes.

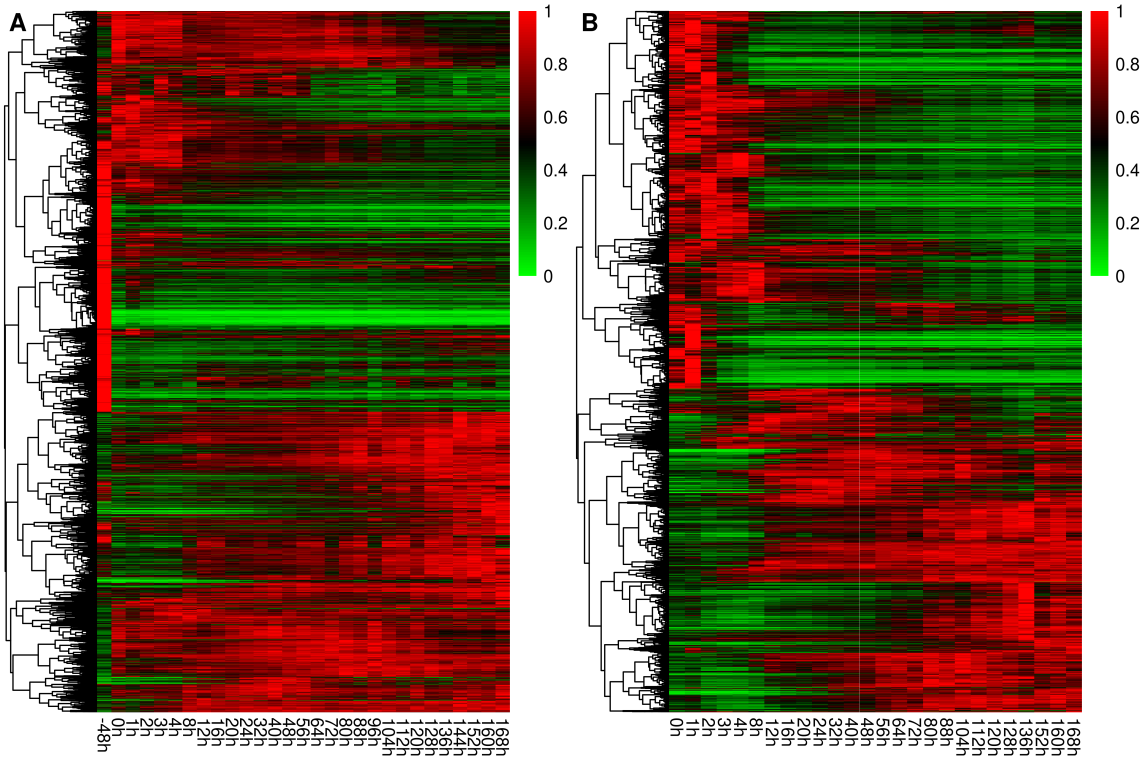


Figure 4.4: Transient expression of differentially expressed genes. The expression of all differentially expressed genes, scaled between 0 and 1, is shown as a color map. Genes are clustered hierarchically according to the Pearson correlation of scaled temporal expression patterns. (A) G-CSF. Plotted genes are the union of DEGs identified in the comparisons between $-48h$ and $168h$, $-48h$ and $0h$, and $0h$ and $168h$ (Fig. 4.3A–C). (B) IL-3. DEGs identified in the comparison between $0h$ and $168h$.

4.2.3 Gene expression changes most rapidly during the earliest stages of the differentiation

Temporal gene expression patterns of differentially expressed genes suggested that most of the changes in genome-wide gene expression are concentrated in the first 12 hours of the differentiation. Next, we sought to check whether this was the case for all genes. We determined the genes differentially expressed between consecutive timepoints and utilized the number of such genes as a measure of the rate of change (Fig. 4.5). The timing of sharp gene expression shifts during the differentiation may indicate important moments when crucial lineage decisions are made. Most DEGs between consecutive time points are detected before the 12h timepoint in both conditions. This implies that rapid, early changes during the first 12 hours after OHT induction are general feature and not just restricted to genes differentially expressed between the endpoints of the differentiation. Additionally, there is very little overlap between DEGs detected between different pair of consecutive timepoints, implying that genes are undergoing rapid but short-lived, transient changes during the early stages of the differentiation (Fig. 4.5B,C).

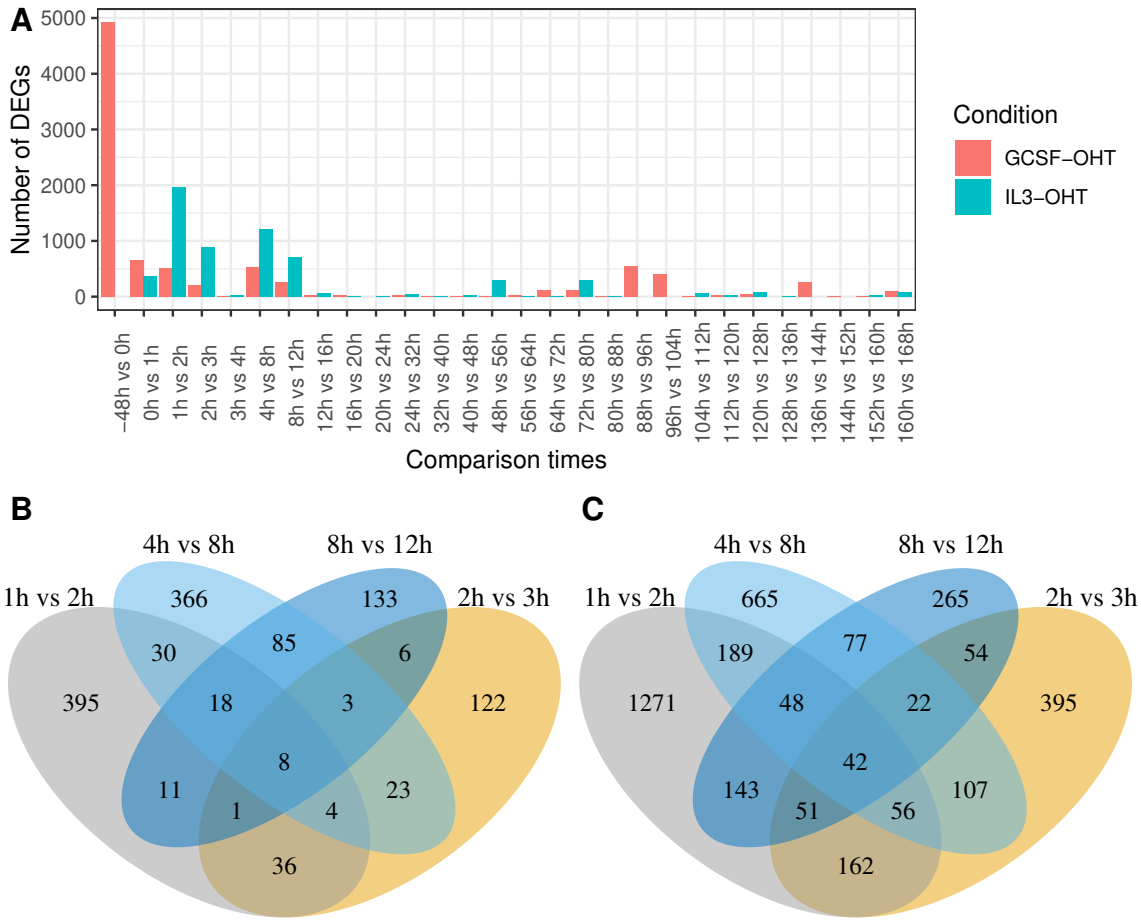


Figure 4.5: The number of differentially expressed genes between consecutive time points. (A) The number of DEGs detected between each pair of consecutive time points. The overlap between the DEGs detected at different time points in (B) G-CSF and (C) IL-3 condition.

4.2.4 Functional enrichment analysis of early DEGs

We performed functional enrichment analysis to determine the biological processes the early DEGs play a role in. We determined gene ontology (GO) terms enriched ($\alpha = 0.05$) in the DEGs ascertained between consecutive early time points

(Tables 4.1 and 4.2). The enriched GO terms show clear temporal patterns. As expected by the long duration of the G-CSF pre-treatment and the large effect of the cytokine (Fig. 4.2), a diverse array of terms ranging from intracellular signaling, to blood cell differentiation, to the cell cycle are enriched in the set of genes differentially expressed between the -48h and 0h G-CSF timepoints. Consistent with the induction of immediate early genes (Tian et al., 1996), the DEGs from the earliest timepoints, $0 - 2$h, are enriched in the GO terms for intracellular signaling, for example "inactivation of MAPK activity" or "negative regulation of Ras protein signal transduction". Following this, GO terms related to development and cellular differentiation progressively rise in rank during the $1 - 3$h time period, suggesting that the cell-fate decision is made very early during the differentiation process. Although biological functions related to myeloid differentiation, highlighted in light gray, are common, GO terms related to the differentiation of non-myeloid cell lineages, highlighted in light cyan, are also enriched, reflecting the pleiotropic roles that most hematopoietic TFs play in multiple lineages.

The GO terms related to signal transduction, development, and cellular differentiation are no longer present in the list of terms for comparisons between 4h and 8h or between 8 and 12h. Interestingly, these terms were absent from the entire list of enriched GO terms, not only from the top 50. Instead, GO terms related to physiological processes such metabolism, ribosome biogenesis, ion homeostasis, and cell-cell adhesion were highly ranked. This suggests that the cells initiate the remodeling of physiological processes as early as 8h in the course of the differentiation to implement a decision that's already been made. Furthermore, this

analysis identifies the $8 - 12$h transition observed above (Figs. 4.1, 4.2, 4.4) as the initiation of large-scale physiological changes in response to differentiation cues. Finally, comparisons between later timepoints were also not enriched in myeloid differentiation and cell-fate commitment GO terms but were instead linked to various physiological processes, strengthening the interpretation that the cell-fate decision is made within the first 4 hours of OHT induction.

Table 4.1: GO analysis for biological functions performed on DEGs estimated for $-48h$ vs $0h$, $0h$ vs $1h$, $1h$ vs $2h$, $2h$ vs $3h$, $4h$ vs $8h$, and $8h$ vs $12h$ in G-CSF condition. Table shows the first 50 GO terms with the lowest adjusted p-values in ascending order. GO terms highlighted in light gray correspond to myeloid differentiation and in light cyan to functions related to the blood development or differentiation of various cell lineages.

-48h vs 0h	0h vs 1h	1h vs 2h
cell-cell adhesion	inflammatory response	negative regulation of MAP kinase activity
regulation of cell adhesion	regulation of inflammatory response	peptidyl-threonine dephosphorylation
G protein-coupled receptor signaling pathway	cellular response to molecule of bacterial origin	negative regulation of protein serine/threonine kin. act.
negative regulation of cell cycle	cellular response to lipopolysaccharide	regulation of MAP kinase activity
negative regulation of cell cycle process	cellular response to biotic stimulus	cellular response to fibroblast growth factor stimulus
adaptive immune response	interleukin-1 production	inactivation of MAPK activity
meiotic cell cycle	regulation of angiogenesis	negative regulation of MAPK cascade
microtubule cytoskeleton organization	blood vessel morphogenesis	response to fibroblast growth factor
regulation of cell cycle process	blood vessel development	negative regulation of phosphorylation
epithelial cell differentiation	interleukin-1 beta production	negative regulation of phosphorus metabolic process
leukocyte cell-cell adhesion	regulation of cytokine production	negative regulation of phosphate metabolic process
lymphocyte activation	regulation of cell adhesion	negative regulation of protein phosphorylation
regulation of cytokine production	angiogenesis	regulation of myeloid cell differentiation
meiotic cell cycle process	vasculature development	peptidyl-tyrosine dephosphorylation
taxis	regulation of vasculature development	cellular response to growth factor stimulus
T cell activation	regulation of interleukin-1 production	response to growth factor
negative regulation of cell cycle phase transition	regulation of interleukin-1 beta production	regulation of protein serine/threonine kinase activity
chemotaxis	inactivation of MAPK activity	regulation of myeloid leukocyte differentiation
regulation of cell activation	response to lipopolysaccharide	negative regulation of locomotion
nuclear division	response to molecule of bacterial origin	negative regulation of cellular component movement
lymphocyte mediated immunity	response to bacterium	negative regulation of transferase activity
mitotic cell cycle phase transition	regulation of defense response	regulation of hemopoiesis
inflammatory response	cell chemotaxis	leukocyte differentiation

Continued on next page

-48h vs 0h	0h vs 1h	1h vs 2h
negative regulation of mitotic cell cycle phase trans.	ossification	myeloid cell differentiation
cell cycle phase transition	positive regulation of smooth muscle cell prolif.	negative regulation of protein modification process
regulation of cell-cell adhesion	heat generation	inflammatory response
extracellular matrix organization	peptidyl-tyrosine dephosphorylation	regulation of epithelial cell differentiation
extracellular structure organization	positive regulation of cytokine production	skeletal system development
lymphocyte differentiation	muscle cell proliferation	myeloid leukocyte differentiation
regulation of mitotic cell cycle	regulation of transcription from RNA Pol. II prom.	regulation of leukocyte differentiation
response to bacterium	myeloid leukocyte migration	epithelial cell differentiation
leukocyte mediated immunity	cell-cell adhesion	ERK1 and ERK2 cascade
regulation of cell cycle phase transition	regulation of smooth muscle cell proliferation	negative regulation of protein kinase activity
T cell differentiation	cellular response to lipid	positive regulation of leukocyte differentiation
meiotic nuclear division	regulation of MAPK cascade	negative regulation of kinase activity
positive regulation of cytokine production	leukocyte migration	MAPK cascade
adaptive immune response based on somatic recomb.	regulation of system process	negative regulation of ERK1 and ERK2 cascade
regulation of mitotic cell cycle phase transition	smooth muscle cell proliferation	regulation of ERK1 and ERK2 cascade
cell fate commitment	regulation of vasoconstriction	regulation of cell adhesion
chromosome organization involved in meiotic cell cycle	negative regulation of signal transduction in abs.	positive regulation of programmed cell death
regulation of leukocyte activation	negative regulation of extrinsic apoptotic sign.	signal transduction by protein phosphorylation
mitotic cell cycle checkpoint	regulation of cell-cell adhesion	cellular response to drug
meiosis I cell cycle process	positive regulation of cell migration	regulation of small GTPase mediated signal transd.
organelle fission	MAPK cascade	chemotaxis
regulation of mononuclear cell proliferation	response to chemokine	response to radiation
regulation of adaptive immune response	cellular response to chemokine	positive regulation of hemopoiesis
regulation of leukocyte cell-cell adhesion	regulation of DNA-templated transcription in resp.	taxis
angiogenesis	peptidyl-threonine dephosphorylation	regulation of inflammatory response
positive regulation of immune response	positive regulation of cell motility	regionalization
blood vessel morphogenesis	signal transduction by protein phosphorylation	chemokine-mediated signaling pathway

2h vs 3h	4h vs 8h	8h vs 12h
endocytosis	ribosome biogenesis	pyruvate metabolic process
regulation of lipid localization	fructose 6-phosphate metabolic process	carbohydrate catabolic process
positive regulation of lipid localization	ribonucleoprotein complex biogenesis	nucleoside diphosphate phosphorylation
blood vessel development	cellular response to hypoxia	nucleotide phosphorylation
cellular response to hypoxia	cellular chemical homeostasis	glycolytic process
cellular response to decreased oxygen levels	cellular metal ion homeostasis	ATP generation from ADP
vasculature development	peptidyl-proline hydroxylation to 4-hydroxy-L-proline	ADP metabolic process
telencephalon development	cellular response to decreased oxygen levels	purine nucleoside diphosphate metabolic process
regulation of lipid transport	regulation of cell adhesion	purine ribonucleoside diphosphate metabolic pr.
response to ischemia	rRNA processing	nucleoside diphosphate metabolic process
cellular response to oxygen levels	positive regulation of cytosolic calcium ion conc.	ribonucleoside diphosphate metabolic process
positive regulation of myotube differentiation	cellular cation homeostasis	generation of precursor metabolites and energy
connective tissue development	divalent inorganic cation homeostasis	cellular response to oxygen levels
inactivation of MAPK activity	response to hypoxia	cellular response to hypoxia
lipid localization	metal ion homeostasis	cellular response to decreased oxygen levels
positive regulation of DNA-binding trans. fac. activity	rRNA metabolic process	carbohydrate metabolic process

Continued on next page

2h vs 3h	4h vs 8h	8h vs 12h
positive regulation of lipid transport	glial cell activation	response to oxygen levels
positive regulation of cell migration	cellular divalent inorganic cation homeostasis	response to hypoxia
regulation of epithelial cell migration	myeloid leukocyte activation	response to decreased oxygen levels
mesenchyme development	peptidyl-proline hydroxylation	monocarboxylic acid metabolic process
regulation of myotube differentiation	cellular ion homeostasis	ATP metabolic process
positive regulation of epithelial cell migration	calcium ion homeostasis	purine nucleotide metabolic process
positive regulation of vasculature development	response to decreased oxygen levels	purine-containing compound metabolic process
positive regulation of cell motility	positive regulation of Rho protein signal transduction	peptidyl-proline hydroxyl.
positive regulation of cell development	cellular response to oxygen levels	hexose metabolic process
cell morphogenesis involved in differentiation	positive regulation of cytosolic calcium ion conc.	nucleotide metabolic process
positive regulation of striated muscle cell differentiation	microglial cell activation	nucleoside phosphate metabolic process
positive regulation of locomotion	leukocyte activation involved in inflammatory response	ribose phosphate metabolic process
positive regulation of cellular component movement	cellular calcium ion homeostasis	peptidyl-proline hydroxylation
regulation of endothelial cell migration	protein homotetramerization	glycolytic process through fructose-6-phosphate
fatty acid transport	regulation of cytosolic calcium ion concentration	purine ribonucleotide metabolic process
epithelial cell migration	import across plasma membrane	glucose metabolic process
tissue migration	macrophage activation	monosaccharide metabolic process
epithelium migration	response to oxygen levels	cellular chemical homeostasis
positive regulation of nervous system development	embryo implantation	ribonucleotide metabolic process
long-chain fatty acid transport	sequestering of calcium ion	protein hydroxylation
endothelial cell migration	cation homeostasis	metal ion homeostasis
pallium development	positive regulation of cytokine production	cellular metal ion homeostasis
plasminogen activation	neuroinflammatory response	leukocyte homeostasis
positive regulation of myoblast fusion	cell chemotaxis	divalent inorganic cation homeostasis
positive regulation of angiogenesis	regulation of smooth muscle cell proliferation	nucleobase-containing small molecule metab. proc.
regulation of fatty acid transport	positive regulation of leukocyte migration	cellular cation homeostasis
regulation of ion transport	regulation of cytokine production	cellular divalent inorganic cation homeostasis
response to hypoxia	regulation of glial cell migration	cellular ion homeostasis
cerebral cortex development	cell-substrate adhesion	blood circulation
interleukin-1 beta production	inorganic ion homeostasis	ion homeostasis
blood vessel morphogenesis	smooth muscle cell proliferation	glycolytic process through glucose-6-phosphate
positive regulation of nucleotide metabolic process	metanephric nephron development	hexose catabolic process
positive regulation of purine nucleotide metabolic proc.	positive regulation of calcium ion import	cellular carbohydrate metabolic process
response to decreased oxygen levels		cation homeostasis

Table 4.2: GO analysis for biological functions performed on DEGs estimated for $0h$ vs $1h$, $1h$ vs $2h$, $2h$ vs $3h$, $4h$ vs $8h$, and $8h$ vs $12h$ in IL-3 condition. Table shows the first 50 GO terms with the lowest adjusted p-values in ascending order. GO terms highlighted in light gray correspond to myeloid differentiation and in light cyan to functions related to the blood development or differentiation of various cell lineages.

0h vs 1h	1h vs 2h	2h vs 3h
inactivation of MAPK activity	regulation of cell adhesion	ovulation cycle
negative regulation of protein phosphorylation	negative regulation of cell differentiation	cell junction organization
negative regulation of protein serine/threonine kin. act.	negative regulation of protein phosphorylation	negative regulation of interleukin-6 production
negative regulation of MAP kinase activity	blood vessel development	ovulation cycle process
negative regulation of ERK1 and ERK2 cascade	vasculature development	negative regulation of response to external stimulus
cellular response to lipopolysaccharide	positive regulation of cell adhesion	regulation of epithelial cell differentiation
negative regulation of MAPK cascade	leukocyte differentiation	epithelial cell differentiation
cellular response to molecule of bacterial origin	blood vessel morphogenesis	postsynapse organization
response to lipopolysaccharide	regulation of MAPK cascade	regulation of hormone secretion
response to bacterium	regulation of cellular response to growth factor stimulus	synapse organization
negative regulation of phosphorylation	epithelial cell differentiation	actin cytoskeleton organization
negative regulation of phosphorus metabolic process	pattern specification process	regulation of system process
negative regulation of phosphate metabolic process	negative regulation of phosphorylation	regulation of peptide hormone secretion
response to molecule of bacterial origin	MAPK cascade	negative regulation of cell population proliferation
epithelial cell differentiation	lymphocyte differentiation	striated muscle cell development
negative regulation of intracellular signal transduction	muscle structure development	negative regulation of locomotion
cellular response to biotic stimulus	response to growth factor	rhythmic process
cellular response to lipid	negative regulation of protein modification process	hormone secretion
negative regulation of Ras protein signal transduction	signal transduction by protein phosphorylation	muscle cell development
negative regulation of inflammatory response	negative regulation of catalytic activity	metanephros development
negative regulation of small GTPase mediated sig. tr.	T cell differentiation	hormone transport
negative regulation of protein modification process	cellular response to growth factor stimulus	bone development
inflammatory response	transmembrane receptor protein ser./thr. kin.	organ growth
lens fiber cell differentiation	negative regulation of ossification	heart development
regulation of ERK1 and ERK2 cascade	regulation of T cell differentiation	muscle cell apoptotic process
regulation of epithelial cell differentiation	negative regulation of vasculature development	regulation of ion transport
negative regulation of protein kinase activity	skeletal system development	polysaccharide biosynthetic process
response to lipid	negative regulation of cell population proliferation	peptidyl-threonine dephosphorylation
negative regulation of transferase activity	regulation of vasculature development	cardiocyte differentiation
ERK1 and ERK2 cascade	regulation of lymphocyte differentiation	muscle cell differentiation
negative regulation of kinase activity	response to transforming growth factor beta	peptide hormone secretion
positive regulation of leukocyte differentiation	negative regulation of protein ser./thre. kin.	female gonad development
negative regulation of response to external stimulus	regulation of protein serine/threonine kinase activity	amino acid transmembrane transport
negative regulation of growth	negative regulation of phosphorus metabolic process	cellular polysaccharide biosynthetic process
regulation of epidermis development	negative regulation of phosphate metabolic process	vasculature development
peptidyl-threonine dephosphorylation	regulation of epithelial cell differentiation	regulation of release of cytochrome c from mitoch.

Continued on next page

0h vs 1h	1h vs 2h	2h vs 3h
lens development in camera-type eye	negative regulation of cell adhesion	regulation of transmembrane transport
learning or memory	response to decreased oxygen levels	
regulation of defense response	cellular response to transf. growth factor beta stimulus	
cellular response to interferon-beta	ERK1 and ERK2 cascade	
regulation of MAP kinase activity	taxis	
negative regulation of locomotion	regionalization	
negative regulation of interleukin-1 production	peptidyl-tyrosine dephosphorylation	
keratinocyte proliferation	heart morphogenesis	
epidermis development	cell surface receptor sig. pathway inv. in cell-cell sig.	
regulation of protein serine/threonine kinase activity	positive regulation of leukocyte differentiation	
negative regulation of transcription from RNA pol. II pr.	chemotaxis	
negative regulation of epithelial to mesenchymal trans.	coronary vasculature development	
positive regulation of lymphocyte differentiation	negative regulation of MAPK cascade	
positive regulation of hemopoiesis	regulation of transm. receptor protein ser./thr. ki.	

4h vs 8h	8h vs 12h
cell-cell adhesion	ribosome biogenesis
ion homeostasis	rRNA processing
G protein-coupled receptor signaling pathway	rRNA metabolic process
cellular divalent inorganic cation homeostasis	cellular response to hypoxia
divalent inorganic cation homeostasis	response to hypoxia
cellular calcium ion homeostasis	ncRNA processing
metal ion transport	response to decreased oxygen levels
cellular metal ion homeostasis	cellular response to decreased oxygen levels
cellular cation homeostasis	ribonucleoprotein complex biogenesis
calcium ion homeostasis	response to oxygen levels
cellular ion homeostasis	cellular response to oxygen levels
regulation of cytosolic calcium ion concentration	ncRNA metabolic process
regulation of cell activation	maturation of LSU-rRNA from tricist.
inorganic ion transmembrane transport	maturation of LSU-rRNA
inflammatory response	pyruvate metabolic process
regulation of leukocyte activation	protein localization to nucleolus
positive regulation of leukocyte activation	maturation of SSU-rRNA
cellular chemical homeostasis	carbohydrate catabolic process
inorganic cation transmembrane transport	amine metabolic process
positive regulation of cell activation	glycolytic process through fructose-6-phosphate
regulation of cell adhesion	purine nucleoside diphosphate metabolic process
cation transmembrane transport	purine ribonucleoside diphosphate metabolic process
metal ion homeostasis	maturation of 5.8S rRNA from tricistronic rRNA transcr.
inorganic ion homeostasis	nucleotide phosphorylation
cation homeostasis	nucleoside diphosphate metabolic process
regulation of cell-cell adhesion	ribonucleoside diphosphate metabolic process
regulation of ion transport	cellular amine metabolic process
positive regulation of cell adhesion	ADP metabolic process
positive regulation of lymphocyte activation	hexose metabolic process

Continued on next page

4h vs 8h	8h vs 12h
regulation of leukocyte cell-cell adhesion	maturation of 5.8S rRNA
T cell activation	maturation of SSU-rRNA from tricistronic rRNA trans.
positive regulation of leukocyte cell-cell adhesion	nucleoside diphosphate phosphorylation
regulation of ion transmembrane transport	glycolytic process
regulation of lymphocyte activation	ATP generation from ADP
leukocyte cell-cell adhesion	peptidyl-proline hydroxylation to 4-hydroxy-L-proline
positive regulation of T cell activation	glycolytic process through glucose-6-phosphate
regulation of transmembrane transport	monosaccharide metabolic process
lymphocyte activation	ribosomal small subunit biogenesis
positive regulation of cell-cell adhesion	polyamine metabolic process
alpha-amino acid metabolic process	polyamine biosynthetic process
chemotaxis	negative regulation of smooth muscle cell proliferation
extracellular matrix organization	carbohydrate metabolic process
extracellular structure organization	dicarboxylic acid metabolic process
import into cell	cellular biogenic amine metabolic process
positive regulation of cytosolic calcium ion conc.	glucose catabolic process
taxis	amine biosynthetic process
cellular amino acid metabolic process	peptidyl-proline hydroxylation
regulation of T cell activation	
regulation of metal ion transport	
blood vessel morphogenesis	

The genes differentially expressed during the first 4 hours of differentiation and enriched in GO terms related to myeloid differentiation are likely to play a role in the cell-fate decision. We focused on the genes differentially expressed between consecutive timepoints from -48h to 3h that were also members of the 5 gene ontologies related to myeloid differentiation (highlighted in gray in Tables 4.1 and 4.2). Genes were clustered on the similarity of their G-CSF temporal expression patterns measured by the Pearson correlation coefficient (Fig. 4.6). Remarkably, even though these genes were identified by virtue of being differentially expressed at earlier stages, most continue to change in expression and achieve their zenith or nadir at later timepoints. A second remarkable property

is that the temporal expression patterns in the two conditions bear a striking resemblance to each other, implying that the difference in the macrophage and neutrophil phenotypes is generated by a small number of genes. Two groups of genes that could be candidates for the divergence of the two lineages may be discerned. The first, group I, comprises genes strongly induced by G-CSF pre-treatment that decay gradually after OHT induction. Although this group of genes is also induced in IL-3 conditions, the activation occurs later and is weaker than the activation caused by G-CSF treatment. Group II genes are downregulated strongly and irreversibly by G-CSF pre-treatment but are transiently induced by OHT during the first 4h of IL-3 differentiation. For both of these groups of genes, the G-CSF and IL-3 conditions differ mainly in transient gene expression during the first 4h, suggesting that the decision is made by short-lived differences very early in the differentiation.

We further narrowed our attention to the TFs amongst the genes expressed differentially between early timepoints and associated with myeloid differentiation GO terms. Ten co-expression clusters or modules were extracted from the hierarchical cluster run on scaled G-CSF gene expressions (Fig. 4.7). Clusters I & II are expressed in undifferentiated PUER cells but are downregulated in both G-CSF and IL-3 conditions. Many of these TFs, such as Gata2, Gata3, Zfpm1, Pou2f2 and Tal1, are known to be involved in the specification of non-myeloid lineages (May et al., 2013; Cantor & Orkin, 2002). Clusters IV & V comprise transiently upregulated genes known to play a role in neutrophil specification such as Gfi1, Myb, Cebpe, and Id2 (Hock et al., 2003; Li et al., 2010; Ward et al., 2000). Cluster X is

comprised of TFs such as Egr1 that peak in expression within an hour of OHT induction in IL-3 conditions and are known to be necessary for macrophage cell-fate specification. Clusters IX and VIII feature a mixture of neutrophil-associated (e.g. Cebpb, Rara, Rarg) and macrophage-associated (e.g. Jun, Fos, Junb) TFs; the common theme connecting them is that the expression of these TFs peaks in the latter half of the differentiation.

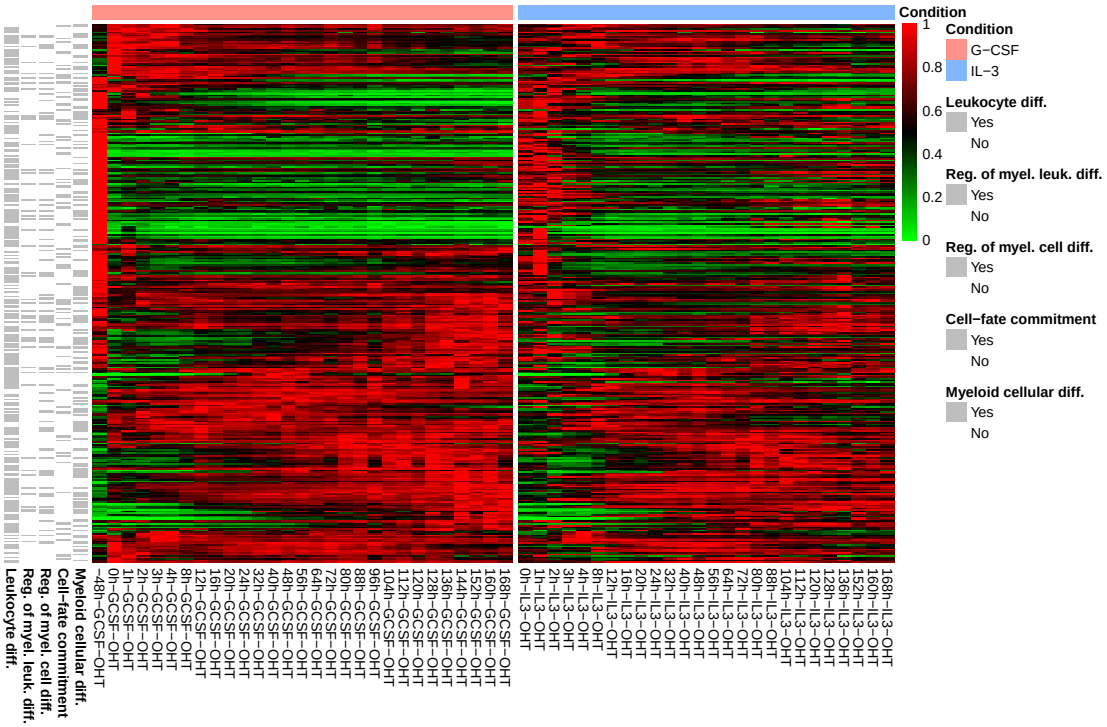


Figure 4.6: The expression of genes differentially expressed during the first 4 hours of OHT treatment and associated with the myeloid differentiation GO terms "Leukocyte differentiation", "Regulation of myeloid leukocyte differentiation", "Regulation of myeloid cell differentiation", "Cell-fate commitment", and "Myeloid cellular differentiation". Each row corresponds to DEG. Genes were clustered on the similarity of their G-CSF temporal expression patterns measured by the Pearson correlation coefficient. Left and right part of the heatmap correspond to the G-CSF and IL-3 condition respectively. Gray marks left to each gene signify GO terms the gene belongs to.

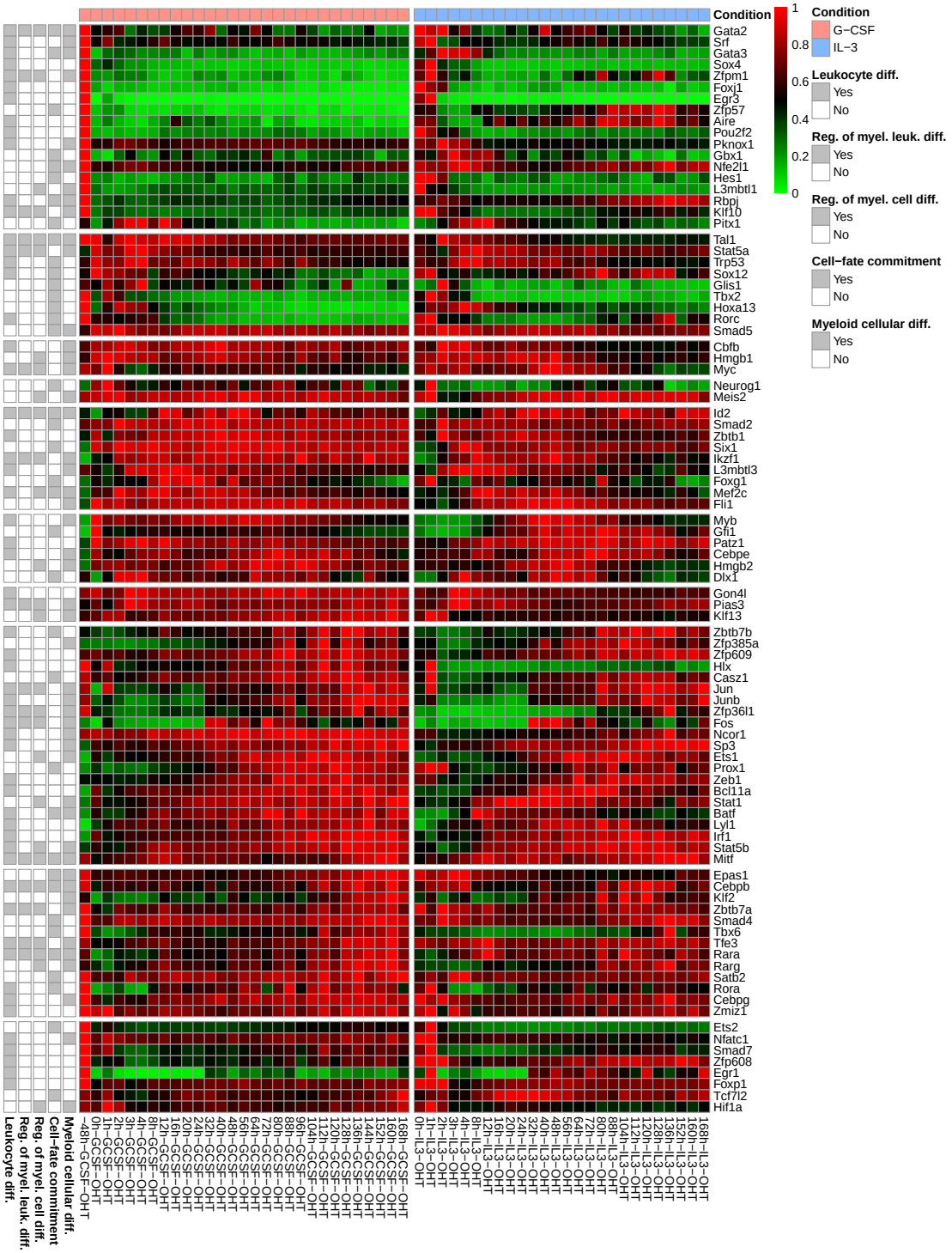


Figure 4.7: Heatmap of gene expression in selected differentially expressed TFs. The expression of TFs differentially expressed during the first 4 hours of OHT treatment and associated with the myeloid differentiation GO terms "Leukocyte differentiation",

(continued)

Figure 4.7: Heatmap of gene expression in selected differentially expressed TFs. (continued) "Regulation of myeloid leukocyte differentiation", "Regulation of myeloid cell differentiation", "Cell-fate commitment", and "Myeloid cellular differentiation". The TFs were clustered on the similarity of their G-CSF temporal expression patterns measured by the Pearson correlation coefficient. Each row corresponds to DEG. Genes were clustered on the similarity of their G-CSF temporal expression patterns measured by the Pearson correlation coefficient. Left and right part of the heatmap correspond to the G-CSF and IL-3 condition respectively. Gray squares left to each gene signify GO terms the gene belongs to. The heatmap was divided into 10 co-expression clusters.

4.3 DISCUSSION

As was shown in Chapter 2, time series datasets from differentiation experiments are a powerful method for inferring GRNs and clarifying the causality of gene regulatory events during development. While such data are acquired *in vitro* and may be only an approximation to the *in vivo* phenomena, they provide a window into dynamics that is not available in static snapshots of developmental processes. While pseudotime approaches (Tusi et al., 2018; Weinreb et al., 2018) have been used to infer the developmental sequence of cells in scRNA-Seq data, it is impossible to determine the rate of change of expression. Furthermore, *in vitro* differentiation allow one to conduct carefully controlled experiments to minimize the effect of animal-to-animal and technical variation. Despite these strengths, high-resolution time series datasets are fairly uncommon and we are only aware of one dataset (May et al., 2013) in hematopoiesis, the one utilized in Chapter 2, that has been published so far.

In our analysis of erythrocyte-neutrophil differentiation (Chapter 2), the regulatory interconnections were poorly constrained for *Gata2*. This failure of the inference methodology was traced to the upregulation of *Gata2* in both lineages so that its expression did not carry any information about regulation. A potential remedy for this problem is to acquire gene expression data from other hematopoietic lineages—the greater the number of sampled gene switching events, the greater the amount of information to constrain model parameters. Furthermore, most hematopoietic regulators act in a pleiotropic manner (Rothenberg, 2014; Laslo et al., 2008; Friedman, 2007) and building more comprehensive models requires time series data from multiple lineages.

In this chapter, we have described the generation and analysis of the second high-temporal resolution dataset in hematopoiesis. Our technical goals were to densely sample in time while ensuring high technical and biological reproducibility. In order to accomplish this, we utilized a liquid handling workstation to extract RNA in a high throughput manner while ensuring a high level of technical reproducibility. We employed a rigorous quality control scheme to ensure low genomic DNA contamination, which confounds the quantification of the RNA, and a high level of RNA integrity (median RIN was 9.9). Finally, all the samples were sequenced to an average depth of $\sim$30 million reads per sample to ensure that low abundance transcripts, such as those encoding TFs, were estimated reliably. Our analysis of the data suggests that these measures were effective. Most genes have a low coefficient of variation and we can successfully detect small changes in gene expression. The data have a dynamic range of five orders of magnitude,

with transcripts detected between $\sim 10^1$ (e.g. *Gata3*) and $\sim 10^6$ (e.g. *Lyz2*). The high temporal resolution and technical quality of the data appear to have paid off as we were able to observe transient phenomena that haven't been described previously.

Our analyses suggest that there is a large-scale sudden change in gene expression, reminiscent of phase transitions, occurring around $8-12$h after the induction of PU.1 by OHT. We confirmed this conclusion in multiple different analyses: the Pearson correlation of genome-wide gene expression (Fig. 4.1), PCA (Fig. 4.2), and visualization of differentially expressed genes (Fig. 4.5). Furthermore there is another transition occurring around 80h. Similar to our data, it has been observed that there is relatively low rate of change between 12h and 48 hours and from 72h to 168h in the reprogramming of B cells into macrophages by the enforced expression of *Cebpa* (Choi et al., 2021). This suggests perhaps that these transitions are a general phenomenon and not an idiosyncrasy of the PUER system. However, it wasn't clear whether the jumps between 0h and 12h and 48h and 72h during B cell transdifferentiation were the result of the relatively large time intervals, 12h and 24h, or a significantly higher velocity of gene expression change. The much higher temporal resolution of our data unambiguously establishes that the jumps are the result of increased velocity—the shift between 8h and 12h in IL-3 conditions is the second largest shift after the one induced by G-CSF pre-treatment but occurs in 4h instead of 48h.

The second main conclusion from our analysis is that the cell-fate decision appears to have been made by the time of the $8-12$h transition. GO terms related

to myeloid differentiation were in the top 50 terms enriched amongst the differentially expressed genes in the comparisons between consecutive time points up to 3h. However, in the comparisons between 4h and 8h and 8h and 12h, myeloid differentiation terms were no longer in the top 50. Instead, terms related to metabolic processing and ribosome biogenesis were highly enriched, suggesting that the cells are already remodeling their physiology by 12h and the transition corresponds to this physiological change.

This dataset is a rich resource and the analysis reported here probably scratches the surface of the biological insights harbored within. One of the main goals of future work would be clarifying the causality of event at a finer granularity both in time and at the level of genes. These data could also be combined with TF footprinting assays such as ATAC-Seq (Buenrostro et al., 2013) or DNase-Seq (Pique-Regi et al., 2011) assays to distinguish between direct and indirect effects and create a "blow-by-blow" description of cellular differentiation. Finally, the high reproducibility and quantitative nature of the data enable future gene circuit models of macrophage-neutrophil differentiation or of multiple lineages, if combined with other datasets (May et al., 2013; Tusi et al., 2018).

4.4 METHODS

4.4.1 PUER cell culture

PUER cells were cultured according to standard procedures (Repele et al., 2019b). PUER cells were routinely maintained in complete Iscove's Modified Dulbecco's Glutamax medium (IMDM; Gibco, 12440061) supplemented with 10% FBS, 50μM

β-mercaptoethanol and 5ng/mL IL-3 (Peprotech, 213-13).

4.4.2 PUER differentiation

PUER cells were expanded in T-75 flasks prior to the initiation of differentiation. G-CSF pre-treatment was initiated by washing the cells 3 times with PBS and then seeding in 48-well plates at a concentration of 5×10^5 cells/ml in PUER cell culture medium in which IL-3 had been replaced by 10ng/mL Granulocyte Colony Stimulating Factor (G-CSF; Peprotech, 300-23). At the same time, the PUER cells destined for the macrophage differentiation were seeded in 48-well plates at a concentration of 5×10^5 cells/ml in IL-3 PUER cell culture medium. After 48 hours of G-CSF pre-treatment, the neutrophil differentiation was commenced by the addition of 100nM 4-hydroxy-tamoxifen (OHT; Sigma, H7904-5MG). The macrophage differentiation was initiated at the same time by the addition of 200nM OHT. In this chapter, we regard time zero (0h) of differentiation as occurring just before the addition of OHT. With this starting point, the initiation of G-CSF pre-treatment occurs immediately after -48h while cells pre-treated with G-CSF for 48h but not yet induced by OHT are at 0h. Since both treatments start with uninduced PUER cells, the data for the -48h time point of the neutrophil differentiation and 0h time point of the macrophage differentiation are derived from the same samples and are identical. Half the medium was replaced and fresh OHT was added at 40h, 88h, and 136h since OHT converts from the Z isomer to the E isomer having 100-fold lower activity in cell culture media.

4.4.3 Sample collection

In addition to undifferentiated PUER cells, which correspond to the -48h neutrophil and 0h macrophage timepoints, samples were collected after 48h G-CSF pre-treatment (0h neutrophil), every hour for the first four hours, every four hours for the first day, and every eight hours until the end of the seventh day (Fig. 4.8). 4 biological replicates were collected for each timepoint. Cells had been seeded into one 48-well plate for each time point so that the samples for a timepoint could be collected without disturbing the remaining samples. Since the differentiation produces adherent cells, the cells were detached with trypsin using standard protocols for all timepoints after 24h. The cells were transferred into a 96-well plate, which was centrifuged at 1500rpm for 5 min, the majority of the medium was aspirated, the cell pellet was snap-frozen in liquid nitrogen, and stored at -80C until RNA extraction.

4.4.4 Total RNA extraction, quality control, and spike in of ERCC standards

Total RNA was extracted on a Bio-Mek FX$^{\text{P}}$ liquid handling workstation (Beckman Coulter) using the RNAdvance Tissue total RNA isolation kit (Beckman Coulter, A32649) in a 96-well format following the manufacturer's protocol. Genomic DNA contamination was assessed by reverse transcribing the RNA with and without reverse-transcriptase and detecting GAPDH using qPCR. The number of additional cycles required to reach a threshold (ΔC_t) was utilized to assess the fraction of genomic DNA ($2^{-\Delta C_t}$) and only samples with less than 1% genomic DNA were utilized. The quality of the RNA was assessed by capillary gel electrophoresis

on the Agilent 2100 Bioanalyzer using the Eukaryote Total RNA Nano kit (Agilent, 5067-1511). Only samples with RNA integrity numbers (RIN) greater than or equal to 9.5 were utilized, although there was only one sample with a RIN of 9.5 and the median RIN was 9.9. The concentration of RNA was determined using the Qubit fluorometer and the Qubit RNA High Sensitivity kit (Invitrogen, Q32855). With the exception of 8 samples with lower RNA yield, the samples were standardized to a mass of $1,875ng$ in a $25\mu l$ volume. The samples with lower yield were standardized to a mass of $1,437.5ng$ in a $25\mu l$ volume. $3.75\mu l$ or $2.88\mu l$ of a $1:100$ dilution of the External RNA Control Consortium (ERCC) ExFold RNA Spike-In mix (Invitrogen, 4456739) was added to the high- and low-yielding samples respectively.

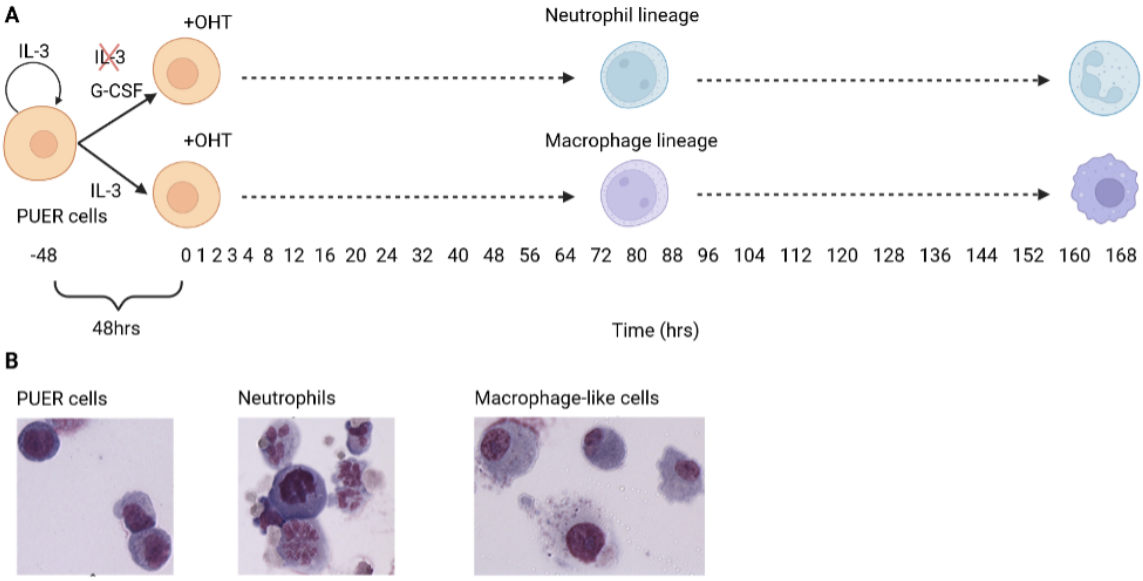


Figure 4.8: The PUER differentiation experiment and sampling scheme. (A) PUER progenitor cells are maintained in IL-3 condition. G-CSF pre-treatment for neutrophil differentiation was initiated at -48h by substituting G-CSF for IL-3 in the culture medium. Neutrophil and macrophage differentiation was initiated at 0h by adding OHT to the G-CSF and IL-3 cultures respectively. The numbers indicate the timepoints at which the cells were sampled for RNA-seq. (B) Wright Giemsa stains of PUER cells in uninduced IL-3 (progenitor), 7 days after OHT induction in G-CSF (neutrophils), and 7-days after OHT induction in IL-3 (macrophage) conditions. Uninduced cells have a blast morphology with high nucleocytoplasmic ratio. Cells induced in G-CSF condition have segmented nuclei, while induction in IL-3 results in cells with vacuolated cytoplasm and low nucleocytoplasmic ratio.

4.4.5 Library preparation and RNA sequencing

Illumina libraries were prepared by Novogene Corporation Inc. (Chula Vista, CA) using the NEB Ultra II RNA Library Prep Kit for Illumina according to manufacturer protocols. The libraries were sequenced on an Illumina Novaseq 6000 S2 2 $\times$

150 bp flow cell to an average depth of 29.24×10^6 raw reads per sample for a total of 4.82^9 reads. The sequencing provider filtered the reads to remove ones containing Illumina adaptors, or more than 10% indeterminate bases ("N"s), or having more than half bases with a phred score below 5. The remaining "clean" reads, representing 95.52% of raw reads, were processed further as described below.

4.4.6 Alignment and quantification

To assess the expression level of each gene from the sequence data, the RNA-seq reads have to be mapped to either genome or transcripts. This mapping involves several challenges such as the high computational cost of aligning billions of reads, the ambiguity in assigning reads to alternative splice isoforms, and the non-uniform sampling of reads due to sequence-selectivity of sequencers, GC-content bias, and other technical biases. We utilized `Salmon` (Patro et al., 2017), a tool for processing RNA-seq data, which overcomes these challenges. `Salmon` infers the percentage of nucleotides corresponding to a particular transcript present in the sample and models the probability of observing the sequenced fragments parameterized by the unknown abundance of each transcript, including the effects of various biases described above. The abundances are then estimated using Maximum Likelihood approach. `Salmon` can be used to both align reads and quantify transcript abundances ("quasi-mapping" mode) or to quantify a set of reads that have already been aligned to a transcriptome ("quantification" mode). We utilized the former. We utilized the *Mus musculus* GRCm38 genome primary assembly and GRCm38.100 transcript annotation for mapping and quantifying

the reads.

4.4.7 Normalization of reads

To compare estimated transcript abundances between samples, they must be normalized to correct for variation in library size from sample to sample due to technical factors. A major hindrance to library size normalization is that a few highly expressed genes can dominate the library size, masking the effect of technical variation (Gierlinski et al., 2015). We utilized the `estimateSizeFactors()` function of the DESEq2 R package, which computes the ratio of each gene's expression in each sample to the geometric mean of the expression across samples and determines the normalization factor of each sample as the median of these ratios, which is robust to the influence of a few highly expressed genes.

4.4.8 Outlier detection

RNA-seq read counts are influenced by the many stages of the experiment, such as sample collection, extraction of RNA, library preparation, and sequencing. Detection and removal of outlier samples is essential for an accurate comparisons between samples (George et al., 2015). The detection of outlier samples was performed with the help of the Principal Component Analysis (PCA). PCA was run on normalized samples, in both conditions, considering only genes with mean expression above 5 reads in all samples. Genes with mean expression of < 5 reads in all samples are referred hereafter as lowly expressed genes. For each time point and replicate i from a total of N replicates, the z-scores of the first two

principal components (PCs) were calculated as $|z_i|^k = |\frac{{PC_i}^k - \frac{\sum_{i=1}^{N} {PC_i}^k}{N}}{\frac{\sum_{i=1}^{N} ({PC_i}^k - \frac{\sum_{i=1}^{N} {PC_i}^k}{N})^2}{N}}|$, where $k = 1, 2, 3, 4$ are the top four PCs that explain 96% of the total variation in the data. If a replicate i had a z-score $|z_i| > 2$ for any of the first four PCs, this replicate was considered as an outlier. No outliers were detected in the data using this method.

This method does not detect unusual non-replicate samples. In an effort to visually identify potentially unusual samples, we first removed lowly expressed genes and clustered all samples hierarchically using Pearson correlation as a distance metric after scaling each gene's expression between 0 and 1 (Figs. 4.9 and 4.10). All three replicates for $96h$ and $144h$ visually appeared to be significantly different than neighboring samples. Given that these timepoints were immediately preceded by supplementation with OHT (see above), we excluded these two timepoints from further analyses. The level of variation in the total gene expression for times close to $96h$ and $144h$ further justified the exclusions of those samples (Fig. 4.11).

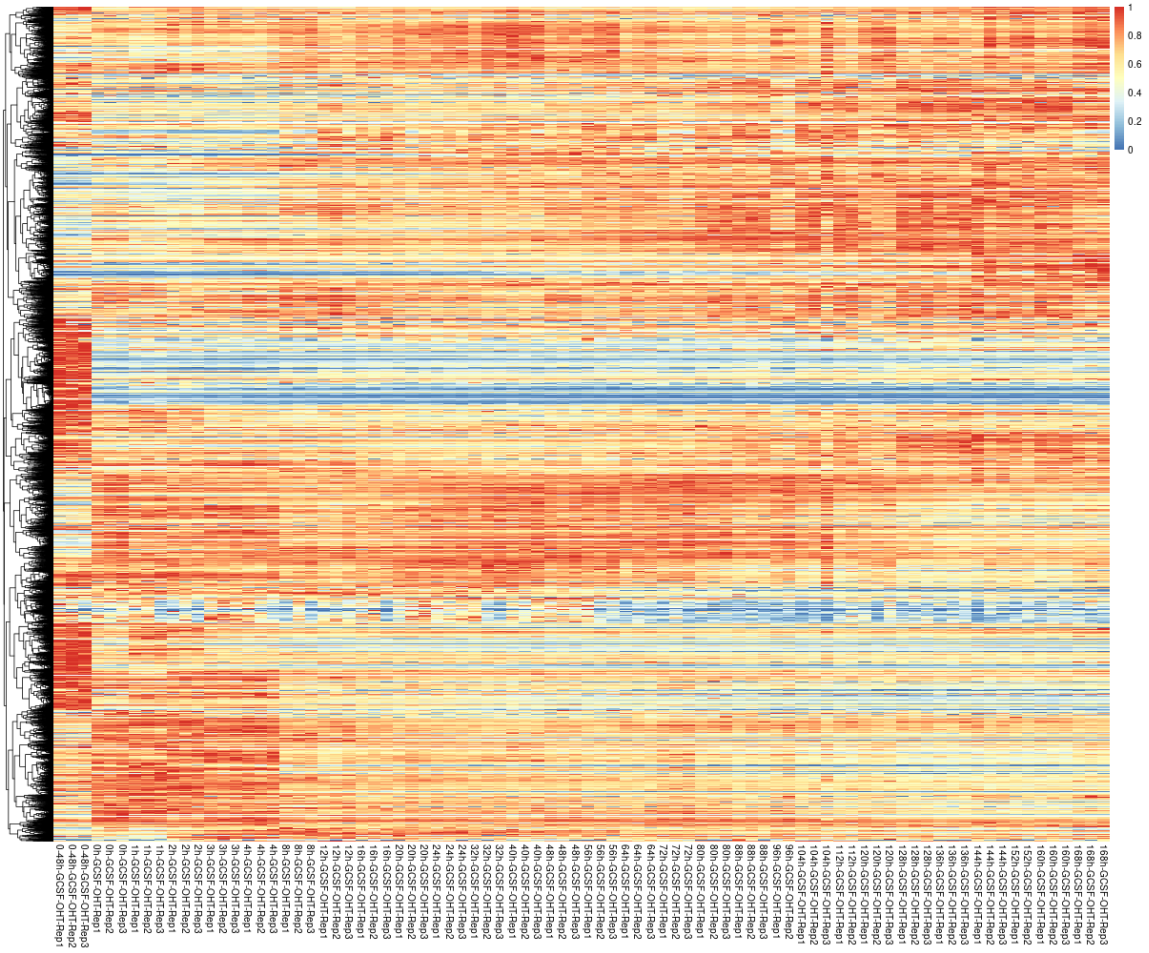


Figure 4.9: Hierarchical clustering of G-CSF samples using Pearson correlation as a distance measure. Color indicates the expression of a gene (y-axis) at each timepoint (x-axis) scaled from 0 to 1.

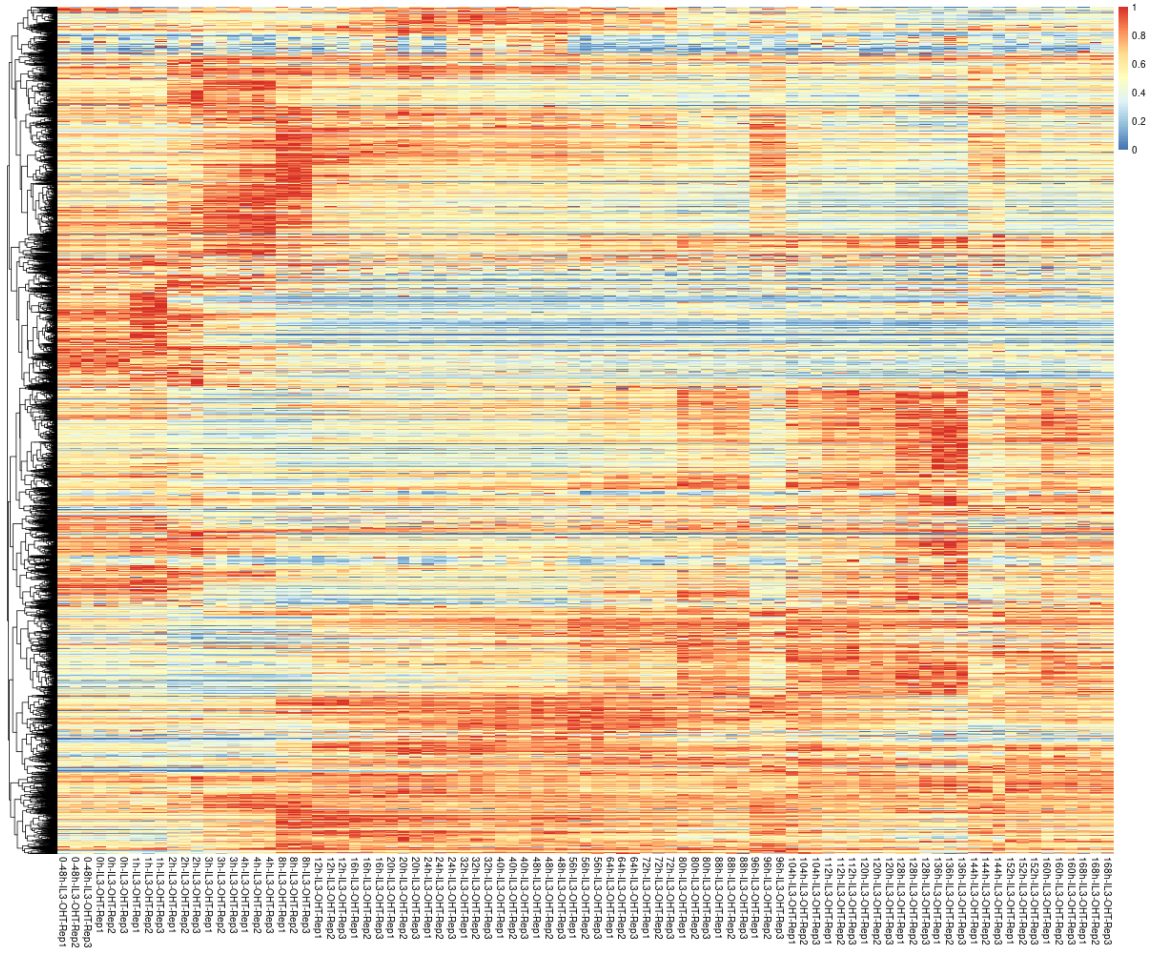


Figure 4.10: Hierarchical clustering of IL-3 samples using Pearson correlation as a distance measure. Color indicates the expression of a gene (y-axis) at each timepoint (x-axis) scaled from 0 to 1.

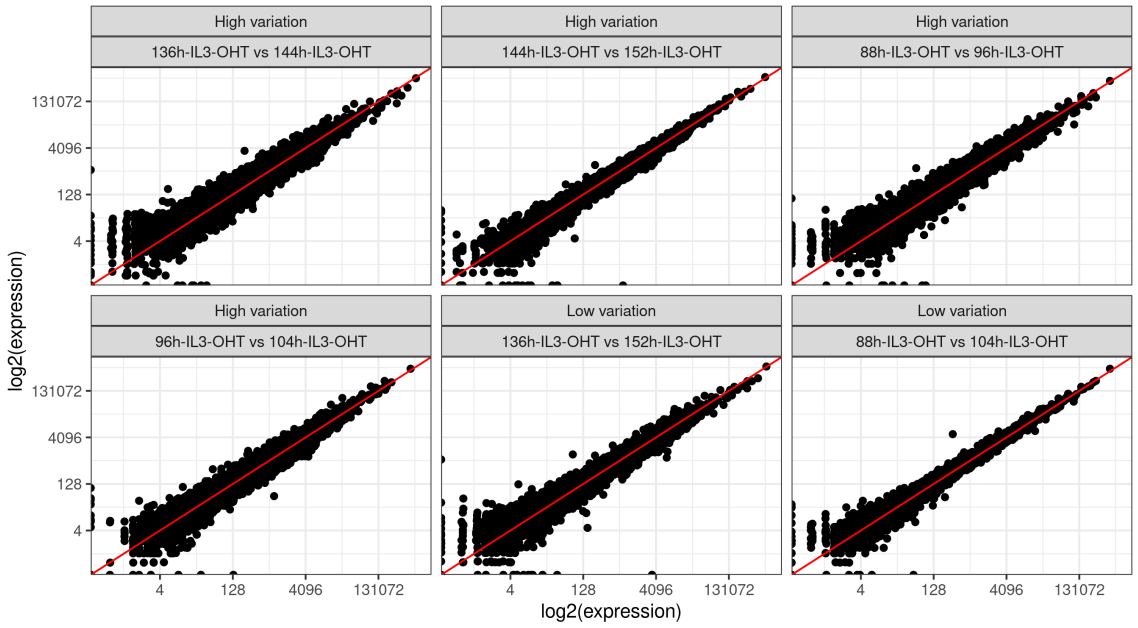


Figure 4.11: Correlation in gene expression between selected samples. Correlation in gene expression between suspected outliers and immediate neighbors (88h vs 96h, 96h vs 104h, 136h vs 144h, and 144h vs 152h). Correlation between the timepoint immediately preceding and the timepoint immediately following the suspected outlier (88h vs 104h and 136h vs 152h).

4.4.9 Correlation between samples

The Pearson correlation coefficient of mean genome-wide gene expression at each timepoint was computed between each pair of timepoints. Lowly expressed genes were excluded from this analysis. The expression of genes was scaled between 0 and 1. The `pheatmap` R function was used to plot the correlation coefficient as heat maps without clustering the timepoints.

4.4.10 Principal component analysis

Principal component analysis (PCA) was carried out with the `prcomp` R function. Lowly expressed genes were excluded from this analysis. The parameter "scale" was set to true so that the gene expression at each timepoint was scaled to have zero mean and unit variance to avoid highly expressed genes from unduly influencing the principal components.

4.4.11 Hierarchical clustering and gene expression heatmaps

Samples or timepoints were clustered hierarchically on gene expression scaled between 0 and 1 using Pearson correlation as a similarity metric. Lowly expressed genes were excluded from the analysis. The `pheatmap` R function was used to cluster genes and plot expression as a heatmap.

4.4.12 Differential gene expression analysis

Differential expression analysis was conducted using the DESeq2 (Love et al., 2014) R package. DESeq2 uses a Generalized Linear Model (GLM) to model the read counts given a design matrix and associated coefficients. The read counts are modeled with the negative binomial distribution whose mean depends on the independent variables through the log link function. The estimates of the coefficients are used to determine the fold change between conditions. DESeq2 uses the Wald test statistic to determine the p-values of estimated fold change. Adjustment for multiple testing is performed with the Benjamini-Hochberg procedure. We utilized an adjusted p-value threshold of 0.05 and $log2$ fold change (FC) threshold

of 0.58 ($\pm 50\%$ change) to infer differentially expressed genes (DEGs).

4.4.13 Functional analysis

`clusterProfiler` was used as a functional analysis tool to identify over-represented Gene Ontology (GO) terms associated with the DEGs. GO categorizations or terms comprise a universal and consistent description and roles of genes and gene products (Ashburner et al., 2000). We utilized Biological Process (BP) GO classification. Significant gene list contained DEGs with adjusted p-value < 0.05 and $|log_2\text{FC}| \geq 0.58$ while the background gene list contained all annotated genes used in the analysis.

BIBLIOGRAPHY

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, *25*(1), 25–29.

Bertolino, E., Reinitz, J., & Manu (2016). The analysis of novel distal cebpa enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification. *Dev Biol*, *413*(1), 128–44.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat Methods*, *10*(12), 1213–8.

Cantor, A. B., & Orkin, S. H. (2002). Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*, *21*(21), 3368–76.

Choi, J., Lysakovskaia, K., Stik, G., Demel, C., Sding, J., Tian, T. V., Graf, T., & Cramer, P. (2021). Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. *eLife*, *10*, e65381.

Dahl, R., Walsh, J. C., Lancki, D., Laslo, P., Iyer, S. R., Singh, H., & Simon, M. C. (2003). Regulation of macrophage and neutrophil cell fates by the pu.1:c/ebpalpha ratio and granulocyte colony-stimulating factor. *Nat Immunol*, *4*(10), 1029–36.

Franzke, A. (2006). The role of G-CSF in adaptive immunity. *Cytokine Growth Factor Rev.*, *17*(4), 235–244.

Friedman, A. D. (2007). Transcriptional control of granulocyte and monocyte development. *Oncogene*, *26*(47), 6816–28.

George, N. I., Bowyer, J. F., Crabtree, N. M., & Chang, C.-W. (2015). An iterative leave-one-out approach to outlier detection in rna-seq data. *PLOS ONE*, *10*(6), 1–10.

Gierlinski, M., Cole, C., Schofield, P., Schurch, N. J., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G., Owen-Hughes, T., Blaxter, M., & Barton, G. J.

(2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, *31*(22), 3625–3630.

Görgens, A., Radtke, S., Horn, P. A., & Giebel, B. (2013). New relationships of human hematopoietic lineages facilitate detection of multipotent hematopoietic stem and progenitor cells. *Cell cycle (Georgetown, Tex.)*, *12*(22), 3478–3482. 24189527[pmid].

Guanglan, L., Wenke, H., & Wenxue, H. (2020). Transcription factor PU.1 and immune cell differentiation (review). *Int. J. Mol. Med.*, *46*(6), 1943–1950.

Hock, H., Hamblen, M. J., Rooke, H. M., Traver, D., Bronson, R. T., Cameron, S., & Orkin, S. H. (2003). Intrinsic requirement for zinc finger transcription factor gfi-1 in neutrophil differentiation. *Immunity*, *18*(1), 109–120.

Keightley, M.-C., Carradice, D. P., Layton, J. E., Pase, L., Bertrand, J. Y., Wittig, J. G., Dakic, A., Badrock, A. P., Cole, N. J., Traver, D., Nutt, S. L., McCoey, J., Buckle, A. M., Heath, J. K., & Lieschke, G. J. (2017). The pu.1 target gene zbtb11 regulates neutrophil development through its integrase-like hhcc zinc finger. *Nature Communications*, *8*(1), 14911.

Keohane, E. (2020). *Rodak's hematology : clinical principles and applications*. St. Louis, Missouri: Elsevier.

Laslo, P., Pongubala, J. M. R., Lancki, D. W., & Singh, H. (2008). Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin Immunol*, *20*(4), 228–35.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., & Singh, H. (2006). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, *126*(4), 755–66.

Li, H., Ji, M., Klarmann, K. D., & Keller, J. R. (2010). Repression of id2 expression by gfi-1 is required for b-cell and myeloid development. *Blood*, *116*(7), 1060–9.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, *15*(12), 550.

May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen,

D. G., Peterson, C., & Enver, T. (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, *13*(6), 754–68.

Olsson, A., Venkatasubramanian, M., Chaudhri, V. K., Aronow, B. J., Salomonis, N., Singh, H., & Grimes, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, *537*(7622), 698–702. 27580035[pmid].

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14*(4), 417–419. 28263959[pmid].

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Res*, *21*(3), 447–55.

Repele, A., Krueger, S., Bhattacharyya, T., Tuineau, M. Y., & Manu (2019a). The regulatory control of cebpa enhancers and silencers in the myeloid and red-blood cell lineages. *PLoS One*, *14*(6), e0217580.

Repele, A., Krueger, S., Bhattacharyya, T., Tuineau, M. Y., & Manu (2019b). The regulatory control of cebpa enhancers and silencers in the myeloid and red-blood cell lineages. *PLOS ONE*, *14*(6), 1–24.

Rothenberg, E. V. (2014). Transcriptional control of early t and b cell developmental choices. *Annu Rev Immunol*, *32*, 283–321.

Scott, E. W., Simon, M. C., Anastasi, J., & Singh, H. (1994). Requirement of transcription factor pu.1 in the development of multiple hematopoietic lineages. *Science*, *265*(5178), 1573–7.

Tian, S. S., Tapley, P., Sincich, C., Stein, R. B., Rosen, J., & Lamb, P. (1996). Multiple signaling pathways induced by granulocyte colony-stimulating factor involving activation of jaks, stat5, and/or stat3 are required for regulation of three distinct classes of immediate early genes. *Blood*, *88*(12), 4435–44.

Tusi, B. K., Wolock, S. L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J. R., Klein, A. M., & Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, *555*(7694), 54–60. 29466336[pmid].

Walsh, J. C., DeKoter, R. P., Lee, H. J., Smith, E. D., Lancki, D. W., Gurish, M. F., Friend, D. S., Stevens, R. L., Anastasi, J., & Singh, H. (2002). Cooperative and antagonistic interplay between pu.1 and gata-2 in the specification of myeloid cell fates. *Immunity*, *17*(5), 665–76.

Wang, L., Gao, S., Wang, H., Xue, C., Liu, X., Yuan, H., Wang, Z., Chen, S., Chen, Z., de Thé, H., Zhang, Y., Zhang, W., Zhu, J., & Zhou, J. (2020). Interferon regulatory factor 2 binding protein 2b regulates neutrophil versus macrophage fate during zebrafish definitive myelopoiesis. *Haematologica*, *105*(2), 325–337.

Ward, A. C., Loeb, D. M., Soede-Bobok, A. A., Touw, I. P., & Friedman, A. D. (2000). Regulation of granulopoiesis by transcription factors and cytokine signals. *Leukemia*, *14*(6), 973–90.

Weinreb, C., Wolock, S., & Klein, A. M. (2018). Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, *34*(7), 1246–1248.

Xie, H., Ye, M., Feng, R., & Graf, T. (2004). Stepwise reprogramming of b cells into macrophages. *Cell*, *117*(5), 663–76.

Zhang, D. E., Zhang, P., Wang, N. D., Hetherington, C. J., Darlington, G. J., & Tenen, D. G. (1997). Absence of granulocyte colony-stimulating factor signaling and neutrophil development in ccaat enhancer binding protein alpha-deficient mice. *Proc Natl Acad Sci U S A*, *94*(2), 569–74.

CHAPTER 5

Summary

Cell-fate decisions and cellular differentiation are remarkable and highly complex developmental processes during which progenitor and stem cells choose their fate and gradually acquire the characteristics of mature cells. This dissertation used hematopoiesis, i.e., the process of blood cell formation, as a model to study cellular differentiation and cell-fate decisions. A fundamental aspect of hematopoiesis, and differentiation in general, is the complex dynamics of gene expression that drives cell-fate decisions, commitment, and maturation of the cells. TFs play a crucial role in the process of hematopoiesis by regulating each other's expression, and so form complex and dynamic GRNs. Precise knowledge of the architecture of GRNs and their dynamics could uncover the main drivers and progression of regulatory events during differentiation. Despite their importance, most GRNs remain poorly understood since traditional genetic experiments are low throughput, labor intensive, and are usually conducted at steady state.

In this dissertation, I applied the gene circuit methodology (Reinitz & Sharp, 1995) to infer hematopoietic GRNs from high temporal resolution genome-wide gene expression data. In gene circuits, the architecture of a GRN is not specified beforehand but is encoded in the values of the free parameters, so that it may be learned by fitting the model to gene expression time-series data. Being an *in-silico* method, gene circuits provide an easier and faster approach to GRN inference than targeted genetic experiments while also being capable of simulating the dynamics of biological processes such as differentiation.

Most models of hematopoietic GRNs have been confined to a few TFs (Laslo et al., 2006; Huang et al., 2007; Chickarmane et al., 2009; Narula et al., 2010; May et al., 2013) and assume a bistable or a tristable switch architecture. The fact that TFs bind thousands of locations in the genome and that there is widespread co-regulation (Novershtern et al., 2011; Wilson et al., 2010; Vierstra et al., 2020) suggests that GRNs are comprised of tens to hundreds of genes and are highly interconnected. Chapter 2 presented a comprehensive 12-gene dynamic model trained on time-series gene expression data from the *in vitro* differentiation of erythrocytes and neutrophils. The model was able to quantitatively predict the consequences of perturbation experiments and demonstrated positive feedback loops from cytokines receptors. Importantly, the analysis of the model lead to interesting observations about the erythrocyte-neutrophil development. According to the analysis, *Gfi1* and *Cebpa* act as upstream regulators of *Spi1* and other key TFs during neutrophil differentiation, and the cell-fate decision is driven by early repression reinforced by later activation. The inference of the regulatory causality was made possible by the coupling of gene circuit to time-series data. The resultant 12-gene GRN is based on a mechanistically accurate ODE formulation that is capable of simulating the differentiation in two lineages and is a significant advance over previous modeling attempts that were either restricted to a single lineage, used the non-mechanistic Boolean formalism, or were restricted to GRNs of only $2-3$ genes (Bonzanni et al., 2013; Collombet et al., 2017; Hamey et al., 2017; Magnusson et al., 2017).

A few issues arise during the inference of GRNs from time-series gene ex-

pression data. One issue is the high computational cost of model training, which usually requires access to and expertise in high-performance computing, particularly during the phase of selecting the proper set of genes or finding the optimal optimization conditions. For example, fitting 100 replicate gene circuit models in Chapter 2 took about 2.5 days on 100 CPUs, with each fit carried out in parallel on 10 CPUs. This computational cost proves to be particularly heavy in the initial phases of building a model when multiple rounds of model fitting are required to determine optimal conditions such as the Λ parameter in the penalty term (Eq. 2.3) in simulated annealing (Chu et al., 1999; Manu et al., 2009). Chapter 3 of this dissertation presented a novel method called FIGR that allows GRN inference on a desktop computer in a matter of seconds. FIGR is a binary classification-based method that takes advantage of the switch-like nature of gene regulation to divide the GRN inference problem into two simpler optimization problems that are much easier to solve. For example, FIGR inferred the gap gene system of *Drosophila melanogaster* 600x faster than SA (Fehr et al., 2019). The speed and user friendliness of FIGR could help advance modeling of GRNs in hematopoiesis and beyond in the future.

Another problem in modeling GRNs is the subjectivity involved in choosing the genes to model. In Chapter 2, the genes were selected based on their importance in erythrocyte-neutrophil differentiation as described in the current literature. One immediate issue with that approach is that not all genes have been studied equally and important regulators might have been omitted. Therefore, methods for recognizing important genes in differentiation independently of prior

knowledge are required.

Accurate inference of GRNs using gene circuits requires high temporal resolution measurements of gene expression during differentiation. Time-resolved data are crucial for avoiding overfitting during model inference. Having sufficient measurements restricts the parameter space and mitigates the problem of parameter uncertainty. Very few high-resolution time-series experiments of differentiating hematopoietic cells that capture the full dynamics of gene expression have been performed so far (May et al., 2013). Chapter 4 presented one such dataset of high-resolution time-series gene expression measurement of *in vitro* differentiation of macrophages and neutrophils. We showed that the high temporal resolution reveals a "phase transitions" at 12 after induction, when there are sudden large-scale changes in the transcriptome of PUER cells, and that the cell-fate decision is made very early and involves transient rather than permanent changes in gene expression. Further improvement of gene circuit accuracy requires high-resolution time-series data from many different lineages to allow bigger and better constrained gene circuits. This dataset therefore will enable future gene circuit models of macrophage-neutrophil differentiation or of multiple lineages when combine with other datasets (May et al., 2013).

The rapid development in single-cell RNA sequencing (scRNA-Seq) technology coupled to fluorescent-activated cell sorting (FACS) and single-cell lineage tracing has enabled the detailed analysis of *in vivo* lineage relationships. scRNA-Seq data could potentially be a valuable resource for training gene circuits in the future since they do not suffer from potential artefacts of *in vitro* experiments

such as the exposure of cells to non-physiological levels of cytokines, oxygen, glucose, or other metabolites (Rodriguez-Fraticelli & Camargo, 2021). In addition to these technical improvements, single-cell resolution data could enable a new class of gene circuits that can simulate heterogeneous mixtures of populations and stochastic gene expression (Paul et al., 2015; Nestorowa et al., 2016; Velten et al., 2017). Gene circuits presented in this dissertation have been limited to population averages but could be modified using stochastic modeling frameworks (van Kempen & van Vliet, 2000) although they would have to take into account the lack of time information and the low depth of coverage of scRNA-Seq data (Lähnemann et al., 2020).

In this dissertation I have demonstrated that coupling gene circuit models of relatively large GRNs with high-temporal resolution gene expression data is feasible and yields mechanistic models that provide insight into GRN architecture and the causality of events during development. The speedup of FIGR makes the inference of even larger GRNs computationally feasible, while the dataset and analysis presented in Chapter 4 enables the modeling of macrophage-neutrophil differentiation while also providing novel insights into the timing of cell-fate choice. Taken as a whole this study opens up avenues for future research where scRNA-Seq data or time-series datasets from multiple lineages could enable a unified gene circuit model capable of simulating the differentiation of all blood cell types from the hematopoietic stem cell. Such a model, if successful, could then also be applied to other developmental systems to gain broad and general insights in the genetic architectures underlying development.

BIBLIOGRAPHY

Bonzanni, N., Garg, A., Feenstra, K. A., Schütte, J., Kinston, S., Miranda-Saavedra, D., Heringa, J., Xenarios, I., & Göttgens, B. (2013). Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. *Bioinformatics*, *29*(13), i80–8.

Chickarmane, V., Enver, T., & Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol*, *5*(1), e1000268.

Chu, K. W., Deng, Y., & Reinitz, J. (1999). Parallel simulated annealing by mixing of states. *The Journal of Computational Physics*, *148*, 646–662.

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., & Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc Natl Acad Sci U S A*, *114*(23), 5792–5799.

Fehr, D. A., Handzlik, J. E., Manu, & Loh, Y. L. (2019). Classification-based inference of dynamical models of gene regulatory networks. *G3 (Bethesda)*, *9*(12), 4183–4195.

Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K., & Göttgens, B. (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proceedings of the National Academy of Sciences*, *114*(23), 5822–5829.

Huang, S., Guo, Y.-P., May, G., & Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, *305*(2), 695–713.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., & Schönhuth, A.

(2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., & Singh, H. (2006). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, *126*(4), 755–66.

Magnusson, R., Mariotti, G. P., Köpsén, M., Lövfors, W., Gawel, D. R., Jörnsten, R., Linde, J., Nordling, T. E. M., Nyman, E., Schulze, S., Nestor, C. E., Zhang, H., Cedersund, G., Benson, M., Tjärnberg, A., & Gustafsson, M. (2017). Lassima network inference toolbox for genome-wide mechanistic modeling. *PLOS Computational Biology*, *13*(6), 1–19.

Manu, Surkova, S., Spirov, A. V., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. *PLoS Computational Biology*, *5*, e1000303. Doi:10.1371/journal.pcbi.1000303.

May, G., Soneji, S., Tipping, A. J., Teles, J., McGowan, S. J., Wu, M., Guo, Y., Fugazza, C., Brown, J., Karlsson, G., Pina, C., Olariu, V., Taylor, S., Tenen, D. G., Peterson, C., & Enver, T. (2013). Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, *13*(6), 754–68.

Narula, J., Smith, A. M., Gottgens, B., & Igoshin, O. A. (2010). Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate. *PLoS Comput. Biol.*, *6*(5), e1000771.

Nestorowa, S., Hamey, F. K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N. K., Kent, D. G., & Göttgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, *128*(8), e20–e31. 27365425[pmid].

Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T., Frampton, G. M., Drake, A. C. B., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J. W., Liefeld, T., Smutko, J. S., Chen, J., Friedman, N., Young, R. A., Golub, T. R., Regev, A., & Ebert, B. L. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, *144*(2), 296–309.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., & Amit, I. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, *163*(7), 1663–77.

Reinitz, J., & Sharp, D. H. (1995). Mechanism of eve stripe formation. *Mech. Dev.*, *49*(1-2), 133–158.

Rodriguez-Fraticelli, A. E., & Camargo, F. (2021). Systems analysis of hematopoiesis using single-cell lineage tracing. *Curr. Opin. Hematol.*, *28*(1), 18–27.

van Kempen, G., & van Vliet, L. (2000). Background estimation in non linear image restoration. *Journal of Optical Society of America*, *A17*(3), 425–433.

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A. G., & Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol*, *19*(4), 271–281.

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, *583*(7818), 729–736.

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P. M., Kinston, S., Ouwehand, W. H., Dzierzak, E., Pimanda, J. E., de Bruijn, M. F. T. R., & Göttgens, B. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, *7*(4), 532–44.