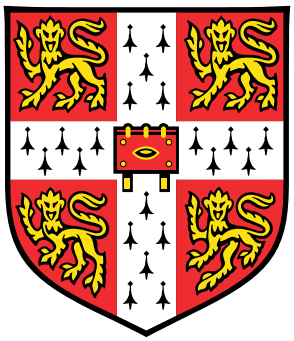
Synthesising executable gene regulatory networks in haematopoiesis from single-cell gene expression data



Steven Woodhouse

Department of Haematology

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College

October 2016

Abstract

A fundamental challenge in biology is to understand the complex gene regulatory networks which control tissue development in the mammalian embryo, and maintain homoeostasis in the adult. The cell fate decisions underlying these processes are ultimately made at the level of individual cells. Recent experimental advances in biology allow researchers to obtain gene expression profiles at single-cell resolution over thousands of cells at once. These single-cell measurements provide snapshots of the states of the cells that make up a tissue, instead of the population-level averages provided by conventional high-throughput experiments. The aim of this PhD was to investigate the possibility of using this new high resolution data to reconstruct mechanistic computational models of gene regulatory networks.

In this thesis I introduce the idea of viewing single-cell gene expression profiles as states of an asynchronous Boolean network, and frame model inference as the problem of reconstructing a Boolean network from its state space. I then give a scalable algorithm to solve this synthesis problem. In order to achieve scalability, this algorithm works in a modular way, treating different aspects of a graph data structure separately before encoding the search for logical rules as Boolean satisfiability problems to be dispatched to a SAT solver.

Together with experimental collaborators, I applied this method to understanding the process of early blood development in the embryo, which is poorly understood due to the small number of cells present at this stage. The emergence of blood from Flk1+ mesoderm was studied by single cell expression analysis of 3934 cells at four sequential developmental time points. A mechanistic model recapitulating blood development was reconstructed from this data set, which was consistent with known biology and the bifurcation of blood and endothelium. Several model predictions were validated experimentally, demonstrating that HoxB4 and Sox17 directly regulate the haematopoietic factor Erg, and that Sox7 blocks primitive erythroid development.

A general-purpose graphical tool was then developed based on this algorithm, which can be used by biological researchers as new single-cell data sets become available. This tool can

deploy computations to the cloud in order to scale up larger high-throughput data sets.

The results in this thesis demonstrate that single-cell analysis of a developing organ coupled with computational approaches can reveal the gene regulatory networks that underpin organogenesis. Rapid technological advances in our ability to perform single-cell profiling suggest that my tool will be applicable to other organ systems and may inform the development of improved cellular programming strategies.

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. Specific details of work arising through collaboration are given in Chapters 4 and 6. The contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. The total length of the main body of this dissertation including figure legends is 31,786 words and therefore does not exceed the limit of 60,000 words for such a dissertation.

Steven Woodhouse
October 2016

Acknowledgements

I would like to thank my supervisors, Bertie Göttgens and Jasmin Fisher, who have taught me a lot about biology, haematology, and the process of doing scientific research in general. I also want to thank Nir Piterman who acted as a third advisor for the technical parts of my PhD and more than once noticed when I was doing something very far from optimally. I thank Vicki Moignard: the main results of this PhD came from a collaboration with Vicki, who had already begun the embryonic blood development study when I started my PhD. Also, the members of the two labs I spent my PhD in, the Haematopoietic Stem Cell Lab at the Cambridge Institute for Medical Research and the Programming Principles and Tools group at Microsoft Research Cambridge. Finally, I thank my family for their support throughout my time at Cambridge.

This work was supported by a Microsoft Research PhD Scholarship.

Contents

List of Figures

Papers arising from this PhD

1. **Transcriptional hierarchies regulating early blood cell development.** Victoria Moignard, Steven Woodhouse, Jasmin Fisher and Berthold Göttgens. (2013). *Blood Cells, Molecules and Diseases*.

2. **Building an ENCODE-style data compendium on a shoestring.** David Ruau, Felicia Ng, Nicola Wilson, Rebecca Hannah, Evangelia Diamanti, Patrick Lombard, Steven Woodhouse and Berthold Göttgens. (2013). *Nature Methods*.

3. **Single cell analyses of regulatory network perturbations using enhancer targeting TAL Effectors suggest novel roles for PU.1 during haematopoietic specification.** Adam C. Wilkinson, Viviane K. S. Kawata, Judith Schütte, Xuefei Gao, Stella Antoniou, Claudia Baumann, Steven Woodhouse, Rebecca Hannah, Yosuke Tanaka, Gemma Swiers, Victoria Moignard, Jasmin Fisher, Shimauchi Hidetoshi, Marloes R. Tijssen, Marella F. T. R. de Bruijn, Pentao Liu and Berthold Göttgens. (2014). *Development*.

4. **Decoding the regulatory network of early blood development from single-cell gene expression measurements.** Victoria Moignard*, Steven Woodhouse*, Laleh Haghverdi, Josh Lilly, Yosuke Tanaka, Adam C. Wilkinson, Florian Buettner, Iain C. Macaulay, Wajid Jawaid, Evangelia Diamanti, Shin-Ichi Nishikawa, Nir Piterman, Valerie Kouskoff, Fabian J. Theis, Jasmin Fisher and Berthold Göttgens. (2015). *Nature Biotechnology*. * Equal contribution

5. **Synthesising executable gene regulatory networks from single-cell gene expression data.** Jasmin Fisher, Ali Sinan Koksal, Nir Piterman and Steven Woodhouse. (2015). *Computer Aided Verification (CAV)*.

6. **Processing, visualising and reconstructing network models from single cell data.** Steven Woodhouse, Victoria Moignard, Berthold Göttgens and Jasmin Fisher. (2015). *Immunology & Cell Biology*.

7. **Single cell analysis of T cell differentiation reveals three distinct states with massive acceleration of proliferation.** Valentina Proserpio, Andrea Piccolo, Liora Haim-Vilmovsky, Tapio Lönnberg, Jhuma Pramanik, Kedar Natarajan, Weichao Zhai, Valentine Svensson, Gozde Kar, Xiuwei Zhang, Giacomo Donati, Melis Kayikci, Jurij Kotar, Andrew N. McKenzie, Ruddy Montandon, Oliver Billker, Steven Woodhouse, Pietro Cicuta, Mario Nicodemi, Sarah A. Teichmann. (2016). *Genome Biology*.

8. **BTR: training asynchronous Boolean models using single-cell expression data.** Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher and Berthold Göttgens. (2016) *BMC Bioinformatics*.

Chapter 1

Introduction

Section 1.3 of this introduction was published in Woodhouse et al. (2015).

Uncovering and understanding the gene regulatory networks (GRNs) which underlie development and homeostasis is a central issue in molecular cell biology. These GRNs control the self-renewal and differentiation capabilities of the stem cells that maintain adult tissues, and become perturbed in diseases such as cancer. They also specify the complex developmental processes that lead to the initial formation of tissues in the embryo. Understanding how to effectively control GRNs can lead to important insights for the programmed generation of clinically-relevant cell types important for regenerative medicine, as well as into the design of molecular therapies to target cancerous cells.

As biological data becomes more accurate and becomes available in larger volumes, researchers are increasingly adopting concepts from computer science to the modelling and analysis of living systems. Formal methods have been successfully applied to gain insights into biological processes and to direct the design of new experiments. New single-cell resolution gene expression measurement technology provides an exciting opportunity for modelling biological systems at the cellular level. Single-cell gene expression profiles provide a snapshot of the true states that cells can reach in the real experimental system, a level of detail which has not been available before, suggesting it may be possible to reconstruct mechanistic computational models of gene regulatory network function directly from data. A major challenge for researchers is to move beyond established methods for the analysis of population average data, to new techniques that take advantage of this single-cell resolution data.

1.1 Haematopoiesis

Haematopoiesis is one of the paradigmatic systems for studying mammalian stem cell biology, due to the ease of access to the blood and bone marrow, and the development of sophisticated techniques for the purification and functional characterisation of stem and progenitor cells (Orkin and Zon (2008)).

In humans, it has been estimated that close to a trillion blood cells are generated every day from the small pool of haematopoietic stem cells (HSCs) which are responsible for maintaining the adult blood system (Ogawa (1993)). HSCs have the ability to both self-renew and to differentiate through a hierarchy of intermediate progenitor cells to the mature cells of all blood lineages (Bryder et al. (2006); Foster et al. (2009); Heng and Painter (2008); Schütte et al. (2012)). Much has been learnt about this process from the study of model organisms, in particular the mouse (Spangrude et al. (1988); Till and McCulloch (1961)). In the mouse, HSCs are functionally defined as cells that are capable of long-term reconstitution of the hematopoietic system of a lethally irradiated recipient animal.

In addition to being a model system for adult stem cell biology, the study of HSCs is also highly clinically relevant. Many leukaemias and haematological malignancies are caused by disruptions to normal cellular decision making, leading to an imbalance in the numbers of different types of blood cell; and bone marrow transplantation represented the first, and still dominant form of stem cell therapy.

While much has been learned about the lineages of the blood system, and the importance of specific transcriptional regulators in normal haematopoiesis and in the development of malignancies, little is known about how these factors are integrated into a wider gene regulatory network that controls cellular decision making. Our understanding of how the hematopoietic system first develops during embryogenesis is also far from complete.

1.1.1 Development of the haematopoietic system

The ontogeny of the haematopoietic system has been studied in detail for over 100 years. Genetic and imaging studies have revealed a complex process that occurs at different developmental time points and at different locations in the embryo and fetus (reviewed in Moignard et al. (2013b)).

Development of the haematopoietic system proceeds in a series of distinct waves. In the mouse, the first, primitive, wave occurs on embryonic day (E)7.5 towards the end of gas-

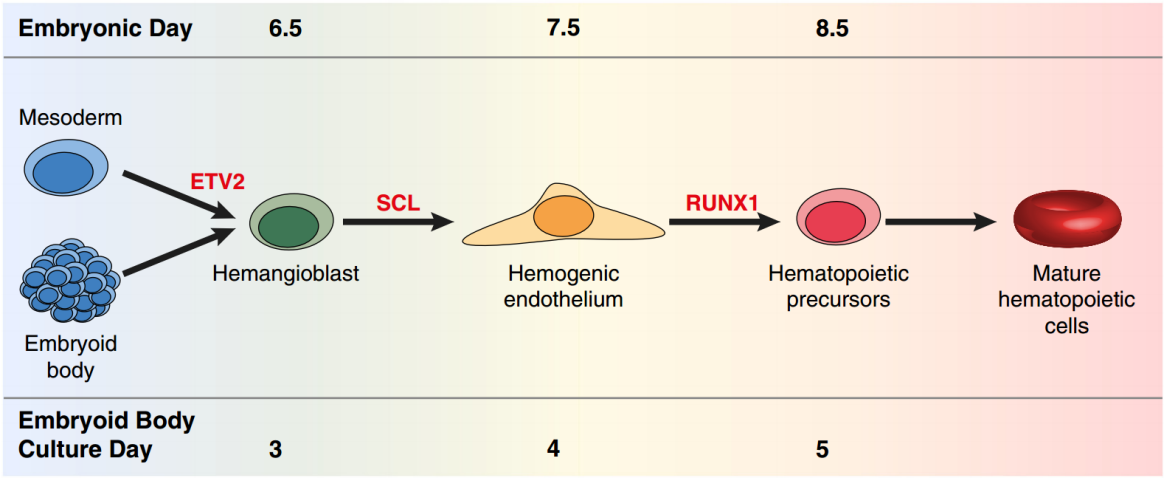


Figure 1.1 Schematic of hematopoietic development *in-vivo* in the developing mouse embryo and *in-vitro* in embryonic stem cell cultures. Hematopoietic cells derive from the embryonic mesoderm through a hemangioblast intermediate, with the transcription factor Etv2 implicated in its emergence and/or commitment. Hemangioblast-like cells can also be identified when embryonic stem cells are induced to differentiate, either through embryoid bodies in suspension or in adherent culture. The transcription factor Scl then regulates the transition from the hemangioblast to the hemogenic endothelium, both *in-vivo* and in embryonic stem cell cultures. Clusters of hematopoietic cells form adjacent to the hemogenic endothelium from which hematopoietic precursor cells bud out into the blood vessels, in a process termed endothelial-to-hematopoietic transition, which is regulated by Runx1 (Moignard et al. (2013b)).

trulation, and primarily generates primitive erythrocytes to supply oxygen to the rapidly growing embryo (Baron et al. (2013); Moignard et al. (2013b)). This is followed by a definitive wave, which can be further divided into the production of multipotent erythro-myeloid progenitors with limited potential for expansion on E8.25, and the emergence of the true HSCs that will go on to populate the bone marrow and maintain the blood system throughout adult life, on E10.5. While embryonic stem cell models are able to recapitulate key aspects of this process, de-novo generation of HSCs *in-vitro* from pluripotent stem cells is still not possible. The study of the gene regulatory networks and signalling pathways involved in the developmental of the blood system is therefore an active topic of study in molecular haematopoiesis.

1.1.1.1 Primitive haematopoiesis

In the mouse, gastrulation begins after the implantation of the embryo at E6.5, with the formation of the three germ layers (mesoderm, ectoderm and endoderm) that will go on to give rise to the different tissues and organs of the embryo. The blood is one of the first tissues to develop, with primitive erythroid progenitors developing at E7.0-7.5 outside of the embryo-proper from mesodermal progenitors in the yolk sac which express the VEGF receptor, Flk1 (Lux et al. (2008); Moignard et al. (2015); Shalaby et al. (1997)).

Differentiated primitive erythrocytes can be detected by E8.0, which differ from adult-type definitive erythrocytes in the existence of a nucleus, their larger size, and the expression of foetal/embyronic instead of adult globins (Baron et al. (2013)). Megakaryocytes and macrophages can also be detected developing in the yolk sac by this stage (Frame et al. (2013)). Primitive erythrocytes continue to mature after entering circulation at around E8.5, and will eventually enucleate (Baron et al. (2013); Frame et al. (2013); Moignard et al. (2013b)). Development of primitive erythrocytes is dependent on the transcription factors Scl (also known as Tal1), Gata1, Gata2, Lmo2 and EKLF, but, unlike definitive erythrocytes, not on Runx1, c-Myb or Zbp89 (Baron et al. (2013)).

1.1.1.2 Definitive haematopoiesis and the emergence of HSCs

Following primitive haematopoiesis in the yolk sac, there is a second, "transient" wave of haematopoiesis, which also occurs independently of HSCs. This wave begins at around E8.25, primarily in the yolk sac but also in the para-aortic splanchnopleura, the aorta-gonad-mesonephros (AGM) region, the vitelline and umbilical arteries, the placenta and the heart (Frame et al. (2013)). These hematopoietic progenitor cells have definitive erythroid and myeloid potential, but do not have long-term reconstitution capability. After emergence, the transient progenitors localise to the fetal liver at E10, where they differentiate. This second wave of haematopoiesis provides the mid- to late-stage embryo with blood cells and is required for survival until birth, by which time the adult haematopoietic system is established. HSC-independent progenitors with lymphoid potential can also be detected in the yolk sac at around E9.5 (Böiers et al. (2013); Yoshimoto et al. (2011, 2012)). These cells are believed to be distinct from the transient erythro-myeloid progenitors.

Finally, hematopoietic stem cells emerge on E10.5 in very small numbers, migrating to the foetal liver by E12.5 and finally to the bone marrow shortly before birth to establish the adult haematopoietic system (Swiers et al. (2013a,b)). HSCs first emerge from the wall of

the dorsal aorta in the AGM region, and then later at multiple sites in the embryo (Medvinsky and Dzierzak (1996)).

Transient progenitors and HSCs emerge from an endothelial source, through a process known as the endothelial-to-hematopoietic transition (Garcia-Porrero et al. (1995); Jaffredo et al. (1998)) which has recently been visualised in real-time using time-lapse imaging (Eilken et al. (2009)). This hemogenic endothelium itself develops from a precursor known as the hemangioblast, a multipotent progenitor which also has the potential to give rise to vascular smooth muscle (Lancrin et al. (2009)). The hemangioblast develops from mesodermal cells which express the VEGF receptor, Flk1. Knockouts studies have revealed some of the transcriptional regulators involved in this linear process (Moignard et al. (2013b), Figure 1.1). The development of the hemangioblast from Flk1 mesoderm is dependent on the transcription factor Etv2 (Kataoka et al. (2013); Moignard et al. (2013b); Wareing et al. (2012)), while Scl is required for the transition from the hemangioblast to the hemogenic endothelium (D'Souza et al. (2005)). The endothelial-to-hematopoietic transition is regulated by Runx1 (Chen et al. (2009); Swiers et al. (2013a,b)).

1.1.2 Maintenance of the adult haematopoietic system

The adult haematopoietic system is maintained by a small pool of haematopoietic stem cells which reside primarily in the bone marrow but can also be found circulating in the blood stream (Figure 1.2). HSCs can make the decision to quiesce, to self-renew or to differentiate through a hierarchy of progressively more lineage-restricted progenitor cells to produce all of the mature adult blood cell types, from oxygen-carrying erythrocytes and platelet-producing megakaryocytes to the cells of the innate and adaptive immune systems. In order to maintain homoeostasis, these cell fate decisions must be carefully regulated in order to produce the correct ratio of each of the mature cell types while maintaining the stem cell pool.

Stem cell decision making is regulated by both internal gene regulatory networks and external cytokines which feed information into the internal regulatory program via signalling pathways. Years of molecular haematology research has identified many of the transcription factors involved in the internal HSC gene regulatory network, the importance of which is highlighted by the fact that their forced expression can commit a stem cell to a specific lineage choice, while their absence can result in the depletion of specific lineages, and by the fact that they are often mutated or dysregulated in leukaemia. For example, forced expression of the transcription factor Gata1 is sufficient to drive haematopoietic progenitors toward

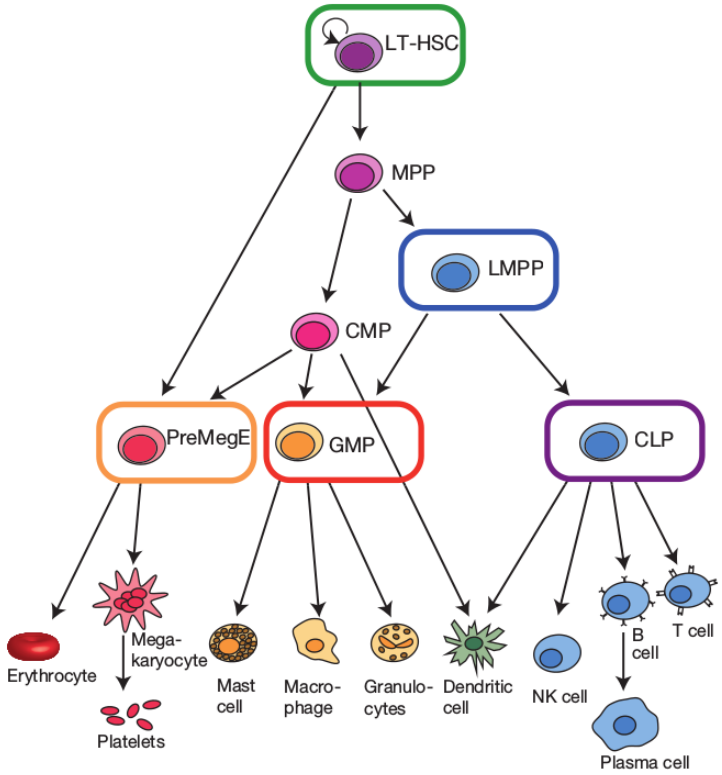


Figure 1.2 Schematic of the haematopoietic hierarchy. Multipotent stem and progenitor cells in purple, megakaryocyte–erythroid lineage in red, other myeloid lineages in orange, and lymphoid cells in blue. (Moignard et al. (2013a)).

an erythroid/megakaryocytic fate (Heyworth et al. (2002); Kulessa et al. (1995)), while PU.1 promotes alternative myeloid fates (Galloway et al. (2005); Rhodes et al. (2005)). Loss of Gfi1 results in the absence of neutrophil progenitors (Hock et al. (2003)). Scl is often translocated in T-cell acute lymphoblastic leukaemia (Robb et al. (1995)), while Pax5 is often deleted in B-cell leukaemia (Medvedovic et al. (2011)) and the Runx1(AML1)–ETO fusion protein is associated with acute myeloid leukaemia (Mulloy et al. (2002)).

The most dramatic demonstration of the power of transcription factors to control cellular state is the conversion of mature cell types back to a pluripotent state reminiscent of the embryonic stem cell (known as an induced pluripotent stem cell, or iPSC) by introduction of the transcription factors Oct3/4, Sox2, c-Myc and Klf4, for which Yamanaka received the 2012 Nobel Prize in Physiology or Medicine (Takahashi and Yamanaka (2006)). Similar reprogramming is possible within the haematopoietic system. For example, B and T cells can be reprogrammed to macrophages by expression of the transcription factor C/EBPα (Laiosa et al. (2006); Xie et al. (2004)). Recent work has shown that introduction of Gata2, cFos,

Gfi1b and Etv6 into mouse fibroblasts; Hoxa9, Erg, Rora, Sox4 and Myb into human myeloid-restricted precursors; or Run1t1, Hlf, Lmo2, Prdm5, Pbx1, and Zfp37 into committed mouse lymphoid and myeloid progenitors can give rise to HSC-like cells (Doulatov et al. (2013); Pereira et al. (2013); Riddell et al. (2014)).

Another active topic of study is the stem cell microenvironment, or niche, of haematopoietic stem cells. Research suggests that the interaction of HSCs with osteoblasts and vascular cells in the bone marrow modulates self-renewal and quiescence (Kiel and Morrison (2006); Kiel et al. (2007); Sipkins et al. (2005); Yoshihara et al. (2007); Zhang et al. (2003)), and that abnormal niches can be involved in the development of leukaemias (Azizidoost et al. (2015); Evans and Calvi (2015); Perry and Li (2007); Schepers et al. (2013)).

1.2 Gene Regulatory Networks

As mentioned above, much has been learnt from studying leukemic patients and from loss-of-function and over-expression experiments about the transcriptional regulators which are important to the development and maintenance of the haematopoietic system, but relatively little is known about how these regulators are integrated into a wider gene regulatory network and how this network executes the complex program of cell fate decision making.

1.2.1 Regulation of gene expression

Gene regulatory networks are built from non-protein-coding regulatory DNA elements — promoters and enhancers — and the transcription factors and epigenetic regulators which interact with these elements and with each other in order to control gene expression and therefore the identity and function of the cell. In mammals, the DNA which encodes each of the genes of the organism along with these regulatory elements is billions of base pairs long ($\sim$2.8 billion bp in mouse) and separated into multiple chromosomes. The human genome, if it was laid out, would be nearly two metres in length (Ball (2003)). In order to compact such a huge amount of genetic material into the cell nucleus, which has a diameter of only a few micrometres, the DNA is supercoiled around proteins called histones to form structures called nucleosomes (Felsenfeld and Groudiner (2003)). Each nucleosome consists of two copies each of four core histone proteins — H2A, H2B, H3, and H4 — wrapped by around 146 base pairs of DNA (Ball (2003); Khorasanizadeh (2004); Kornberg and Lorch (1999); Luger et al. (1997)).

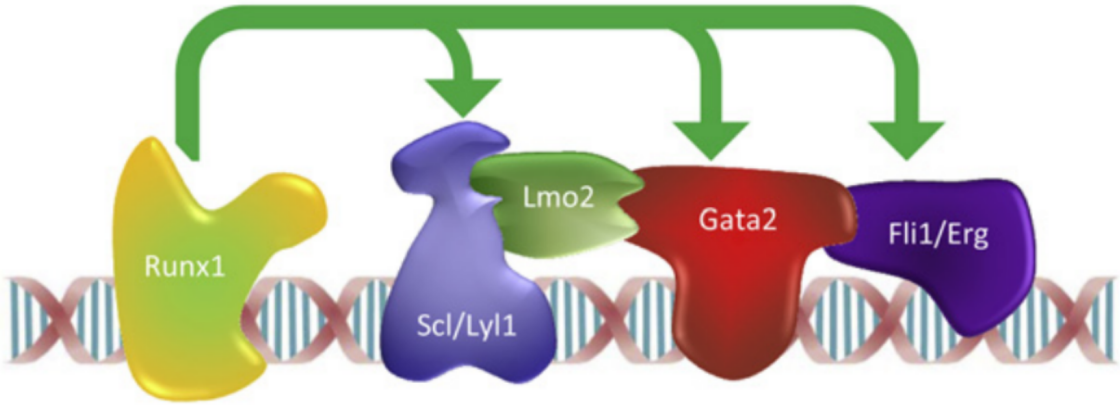


Figure 1.3 Scl, Lyl1, Lmo2, Gata2, Runx1, Erg, Fli1 form a heptad of transcription factors which bind together in a complex in a haematopoietic progenitor cell line (Wilson et al. (2010)).

These histones have N-terminal tails which are subject to epigenetic modifications, such as acetylation and methylation, which regulate the accessibility of the wound DNA to transcriptional regulators and RNA polymerase (Bannister and Kouzarides (2011); Calo and Wysocka (2013); Dawson et al. (2012)). For example, acetylation of histone 3 lysine 27 (H3K27Ac) is enriched at regulatory elements of genes which can be actively transcribed, while trimethylation of histone 3 lysine 27 (H3K27me3) is associated with epigenetically silenced genes (Alberts et al. (2002); Bernstein et al. (2005); Creyghton et al. (2010); Kouzarides (2007); Rada-Iglesias et al. (2011); Schübeler et al. (2004)). The DNA itself is also subject to epigenetic modifications. Methylation of a promoter region is associated with silenced gene expression (Suzuki and Bird (2008)). Both of these forms of epigenetic regulation can be inherited through cell divisions and can persist in daughter cells (Cedar and Bergman (2011); Klose and Bird (2006)).

Transcription factors are modular proteins which bind to regulatory elements in order to activate or repress gene expression. Transcription factors account for a large proportion of the protein-coding genes in the mammalian genome, with 1700-1900 of the 20000-25000 genes in the human genome predicted to be transcription factors (Messina et al. (2004)). It has been estimated from an integrative study of population microarray data that around 150 to 300 transcription factors are expressed in a given human tissue, accounting for around 6% of the transcriptome (Ravasi et al. (2010); Vaquerizas et al. (2009)).

Transcription factors can be characterised by their DNA binding domains, which recognise

and bind to short (4-10 base pairs) DNA sequences, known as motifs. The haematopoietic factors Erg, Fli1 and PU.1 are Ets-factors, one of the largest families of transcription factors in mouse and human (Sharrocks (2001)). These factors have an 85-amino acids DNA binding domain with a winged helix-turn-helix structure (Donaldson et al. (1996); Liang et al. (1994)) which binds to the Ets motif, the GGAW DNA sequence (Sharrocks (2001)). Scl and Lyl1 are basic helix-loop-helix factors (Begley and Green (1999)) which bind the Ebox motif (CANNTG). The Gata family features two zinc finger domains which bind to the WGATAR consensus motif (Ko and Engel (1993)). Runx1 is a core-binding factor which recognises and binds to the Runt binding motif (TGYGGT) (Ito et al. (2015)).

Transcription factors regulate gene expression through their transactivation domains, which bind to accessory proteins that can initiate, prevent, or modulate transcription (Spitz and Furlong (2012); Vaquerizas et al. (2009)). Other factors are believed to act as "pioneer factors", recruiting chromatin remodelling enzymes which deplete nucleosomes, open up chromatin and make DNA accessible to other transcription factors which can subsequently bind to regulatory elements and mediate transcription (Chen et al. (2014b); Iwafuchi-Doi and Zaret (2014); Zaret and Carroll (2011)). Transcription factors may also be involved in recruiting histone and DNA modifying enzymes to lay down or erase epigenetic marks.

1.2.1.1 Combinatorial transcription factor activity

Transcription factors rarely work alone, and instead form multi-factor complexes which function together to regulate gene expression. These complexes can be formed via direct protein-protein interactions between factors which are bound to nearby motifs (Figure 1.3). Some transcription factors, such as Lmo2, are unable to bind DNA directly, and so rely completely on partner factors to recruit them to regulatory regions. Binding of a transcription factor can therefore occur even if the motif it recognises is not present.

Examination of transcription factor binding patterns through ChIP-sequencing experiments followed by co-immunoprecipitation assays to establish protein-protein interactions identified a heptad of factors (Scl, Lyl1, Lmo2, Gata2, Runx1, Erg, Fli1) which bind together in a complex in a haematopoietic progenitor cell line (Wilson et al. (2010), Figure 1.3). A well characterised transcription factor complex is the AFF-2, c-Jun, IRF-3/IRF-7, NFκB complex which regulates expression of interferon-β upon viral infection. Binding of all factors together is required before transcription is activated (Carey (1998); Merika and Thanos (2001); Panne et al. (2007)). Another example of combinatorial transcription factor regulation from haematopoiesis is found at the Scl+19 regulatory element, where the presence

and precise spacing of one Gata and two Ets motifs is required for Scl expression (Ng et al. (2014); Pimanda et al. (2007)).

The combined activity of multiple transcription factors at regulatory elements allows the cell to execute more complex cell fate decisions, where multiple spatial and temporal inputs are fed into a regulatory element and combinatorial logic determines whether the exact conditions are met for the target gene to be expressed, or not (Istrail and Davidson (2005)). The target gene may in turn code for a transcription factor, which then feeds back into regulatory elements, forming a complex gene regulatory network with non-linear logic and feedback loops (Bonzanni et al. (2013); Krumsiek et al. (2011)). This paradigm governs the expression of the stripe patterns during the segmentation of the Drosophila melanogaster embryo, for example (Wilczynski and Furlong (2010)).

Regulatory elements may execute "AND", "OR", or dominant repressing "NOT" logic, and more complex combinations of these (Istrail and Davidson (2005); Peter et al. (2012)). To experimentally determine the logic which controls the expression of a gene requires the identification of its regulatory elements the factors which bind to these elements, and the mutation of each binding motif to prevent binding, both individually and in all possible combinations. Additional perturbations may be required to prevent formation of complexes and to distinguish between factors when multiple factors are able to bind to the same site. This is extremely time-consuming, and a computational method which could predict gene regulatory logic from gene expression data alone would have the potential to massively speed up this process.

1.2.1.2 Transcriptional machinery

In eukaryotes, RNA is transcribed by RNA polymerase I, II or III depending on the type of the RNA; with RNA polymerase II (Pol II) responsible for transcribing protein-coding genes to mRNA (Butler and Kadonaga (2002); Goodrich and Tjian (2010); Juven-Gershon and Kadonaga (2010); Kadonaga (2012); Sandelin et al. (2007)). Transcription factors recruit Pol II to core promoters, regions of DNA that are found at -30, -75 and -90 base pairs upstream from the transcription start site. There Poll II assembles together with six general transcription factors that recognise core promoter motifs and perform essential functions such as the unwinding of DNA — TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH — to form the pre-initiation complex.

Other proteins are involved in this process — coactivators and corepressors such as p300 which modulate the rate of transcription (Teufel et al. (2007); Vo and Goodman (2001)), and

positive and negative elongation factors which regulate the release of the polymerase from the transcriptional start site to allow it to move along the DNA and begin copying DNA to RNA (Adelman and Lis (2012)).

After transcription, further layers of control add to the complexity of gene regulation: post-transcriptional modification of RNA including splicing, whereby introns are removed and exons are joined (Barash et al. (2010)); RNA editing (Pachter (2012)); down-regulation by non-coding RNAs such as miRNA (Chen et al. (2004); Rodriguez et al. (2007); Shivdasani (2006); Stadler et al. (2010); Thai et al. (2010); Xiao et al. (2007)); translation of mRNA to an amino acid chain to form protein; and post-translational modification of the synthesised protein.

1.2.1.3 Enhancers

Complicating the study of mammalian gene expression is the fact that the trans-regulatory elements which control the activity of a gene can be located kilobases up- or down-steam of the transcriptional start site, with recent research suggesting they could even lie on different chromosomes. These elements, called enhancers, loop to physically interact with the promoter and to bring their bound transcription factors into contact with the pre-initiation complex (Hughes et al. (2014); Shlyueva et al. (2014); Spitz and Furlong (2012)). Several proteins and protein complexes have been found to be involved in this process, including the mediator complex (Kagey et al. (2010)); cohesin, which forms rings to stabilise enhancer/promoter interaction (Peric-Hupkes and van Steensel (2008); Seitan and Merkenschlager (2012)); and the transcription factor CTCF (Herold et al. (2012); Phillips and Corces (2009)). Genes can be regulated by multiple enhancers, enhancers can regulate multiple genes, and enhancers generally show tissue or cell-type specific activity.

Despite these factors making the identification of regulatory elements difficult, many enhancers important to the regulation of haematopoiesis are now known, and their activity has begun to be characterised. Potential enhancers can be identified by comparative genomics, identifying regions of DNA which are conserved throughout evolution (Donaldson et al. (2005); Göttgens et al. (2000)); by DNaseI hypersensitivity assays, which reveals DNA accessible to transcription factor binding (Meissner et al. (2008); Song and Crawford (2010)); and by the identification of combinatorial binding of multiple transcription factors and of activating histone modifications, through ChIP-seq (Wilson et al. (2010))). Chromatin conformation capture and related technologies can reveal chromatin looping (Hughes et al. (2014); Lieberman-Aiden et al. (2009); Patwardhan et al. (2009)). Once a putative

enhancer has been identified, it can be tested for activity by cloning the DNA sequence next to a minimal promoter and a reporter gene (Bonzanni et al. (2013)). Tissue-specificity can then be assessed in transgenic embryos (Schütte et al. (2012)). Characterisation of the regulatory module can be carried out by the effect of individual and combined transcription factor binding site mutagenesis in cell lines (Lelieveld et al. (2015); Moignard et al. (2015); Ng et al. (2014); Wilkinson et al. (2014)).

Scl expression in haematopoiesis is regulated by at least two distal enhancers, located at +19 and -4 (Göttgens et al. (2004)) kilobases. The Lmo2 promoter and Lmo2-75, Lmo2-70 and Lmo2-25 enhancers all show activity individually in haematopoietic tissues (Landry et al. (2009)). The Runx1 promoter alone is not active in haematopoietic tissues, but shows activity when combined with the Runx1+23 enhancer (Bee et al. (2009)). Enhancers for other key haematopoietic factors, including Erg, Fli1, Gata2, Gfi1b, and PU.1 (Okuno et al. (2005)) have also been identified and partially characterised.

1.2.2 Modelling gene regulatory networks

Gene regulatory networks control the dynamic expression pattern of genes, ensuring that the correct genes are expressed at the correct times and places at each stage of development, and determining the final, mature cell types that eventually develop. The complex non-linear interactions and feedback loops in these networks mean that mathematical and computational models are required in order to understand the dynamics of differentiation.

In 1957 Waddington proposed the epigenetic landscape, a metaphor for how multipotent cells differentiate by becoming progressively more lineage restricted (Goldberg et al. (2007); Waddington (1957)). In this metaphor, cells are thought of as marbles placed at the top of a valley. As the cells differentiate, they roll down the slopes of the valley and finally come to rest at the lowest points. These points represent the final mature differentiated cell types.

This picture has become the basis for understanding the uni-directional process of development. In the modern formulation, cellular states are defined by the level of expression of each gene, and the dynamics of the system are given by the activation and repression relations between genes (Huang et al. (2005); Kauffman (1969)). If the system is allowed to evolve for long enough, it will eventually end up in one of possibly several attractors, where it will remain. These attractors are thought of as stable cell types. An attractor can be a single stable state, or it can be cyclic, consisting of a series of states which are continually transitioned though.

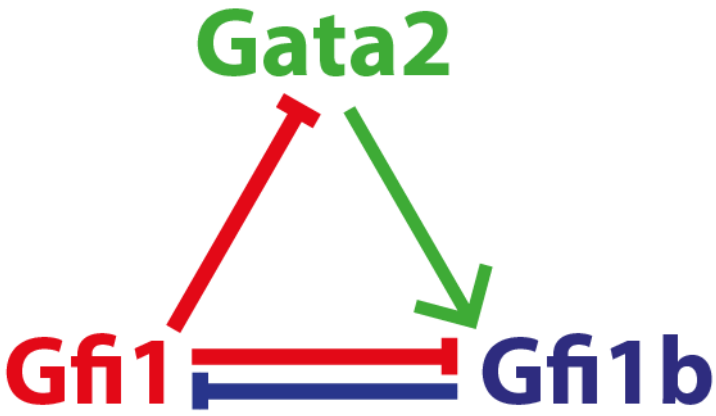


Figure 1.4 Gata2, Gfi1 and Gfi1b form a regulatory triad (Moignard et al. (2013a)).

1.2.2.1 Network motifs

Early work on modelling haematopoiesis focused on identifying small "network motifs" consisting of a few regulatory factors and investigating their effect on cellular decision making, both experimentally and through mathematical modelling. The theory of network motifs has been studied in detail by Alon, who has identified and experimentally characterised several classes of network building blocks that recur throughout biological networks and across organisms, more often than would be expected at random (Alon (2007)).

Motifs that have been identified to be active in haematopoiesis include a double-negative loop between Gata1 and Pu.1 in adult haematopoiesis, in which Gata1 and Pu.1 repress each other and activate their own expression (Arinobu et al. (2007); Chickarmane et al. (2009); Duff et al. (2012); Huang et al. (2007); Monteiro et al. (2011); Roeder and Glauche (2006); Wontakal et al. (2012)). Once activated, this motif locks the cell into one of two classes of states: Gata1 on, Pu.1 off; and Gata1 off, Pu.1 on. This bifurcation dynamic potentially explains why Pu.1 expression in the erythroid–myeloid lineage triggers monocytic differentiation, while Gata1 expression triggers erythroid and megakaryocytic differentiation (discussed above). Pu.1 expression would activate the expression of downstream monocytic genes, and, by repressing Gata1, prevent Gata1 activation of erythroid and megakaryocytic genes (and vice-versa).

Another example is the fully connected Scl-Gata2-Fli1 triad, activated in the specification of HSCs in the embryo, and thought to be a central player in the haematopoietic stem cell network (Pimanda et al. (2007)). In this motif, the three transcription factors cooperate to positively regulate each other and maintain mutual expression, by binding to the Gata2-3,

Fli1+12, and and Scl+19 distal enhancers. This motif is thought to be rare in prokaryoctes, but exists in other stem cell systems, for example the Nanog-Oct4-Sox2 triad in embryonic stem cells (Boyer et al. (2005); Chickarmane et al. (2006)). It has therefore been suggested that this motif is involved in maintaining the stem cell state, and that down-regulation of any member of the triad by an external factor, for example Gata2 repression by Gata1, results in exit of the stem cell state and promotes differentiation.

A mathematical model of the Scl-Gata2-Fli1 triad was built, using ordinary differential equations (ODEs) to model continuous change of protein levels over time, and a thermodynamic analysis of enhancer activity levels to estimate kinetic parameters for TF-DNA and TF-TF binding affinities (Narula et al. (2010)). Mathematical analysis of this model reveals that it exhibits bistability, where initial activation locks the triad into a self-sustaining "ON" state, and a repressor protein must persist for a significant period of time in order to switch the system back into an "OFF" state.

A third example is the Gata2-Gfi1-Gfi1b triad active in HSCs (identified via gene expression profiling of single-cell cells and examining gene expression correlation, Moignard et al. (2013a)). In this motif, which resembles a "type 2 coherent feedforward loop", Gata2 modulates the mutual inhibition of Gfi1 and Gfi1b (Figure 1.4). It has been suggested that this motif may be involved in differentiation and exit of the stem cell state. Ultimately, however, decision making cannot be understood by studying network motifs in isolation, and instead it must be understood how they fit into larger gene regulatory networks. Pu.1-Gata1 antagonism, for example, has been shown to be context-dependent (Monteiro et al. (2011); Sugiyama et al. (2008)).

1.2.2.2 Developmental gene regulatory networks

Wider gene regulatory networks in development are better understood in invertebrate model organisms, such as Drosophila and the sea urchin. Perhaps the best understood gene regulatory network is the network governing endomesodermal specification from early cleavage up to gastrulation in the purple sea urchin embryo (*Strongylocentrotus purpuratus*). This network has been studied for over thirty years by the Davidson lab (Damle and Davidson (2012); Davidson (2006, 2010); Peter and Davidson (2011)). The result of these studies is a Boolean computational model which consists of a Boolean variable (a value of 1 represents "expressed", and 0 "unexpressed") for each of the 45 regulatory genes expressed during this process, together with a logical equation which specifies how the gene is regulated by the other genes (Peter et al. (2012)). Each gene is only expressed when the correct combina-

tion of regulatory inputs is expressed. The model also incorporates spatial and signalling information.

The model can be executed, starting from an experimentally-determined initial state where genes expressed in the earliest developmental stage are turned on. The dynamics of the model proceeds in a series of synchronous time steps where the value of each gene is updated based upon the value of its inputs from the previous state. At several places in the regulatory logic time delays are incorporated to allow states to depend on the value of older states than the previous state. This is to account for differences in the time it takes factors to be transcribed, translated and find their binding sites. Remarkably, the predicted temporal and spatial gene expression patterns which result from execution of the model were found to be in near complete agreement with observed experimental data. Only 2 out of 33 measured genes were found to be expressed in incorrect spatial domains, and only 39 elements in a 45×106 temporal expression matrix were in disagreement with experimental data. This demonstrates that a systems-level mechanistic understanding of development can emerge from a Boolean model based purely on regulatory logic.

A second class of predictions comes from *in-silico* perturbations, in which a gene's regulatory logic is changed so that it is always expressed (over-expression) or never expressed (knock out). The results can then be compared to experimental perturbations. Here again the model was found to be in agreement with experimental data. The model is therefore able to explain nearly all existing experimental data and is a powerful tool to generate predictions which can be used to design new experiments.

1.2.2.3 Stem cell gene regulatory networks

Gene regulatory networks active in stem cells can be expected to have qualitatively different topology and dynamics to those active during development. Developmental gene regulatory networks can follow a principle of "forward-momentum", where the cell is driven from a multipotent state to a final differentiated stable state by transiting through a series of progressively more restricted states. Once initiated, this process can happen fairly autonomously, although external stimuli may be integrated via signalling at each stage. Stem cells, on the other hand, have two seemingly conflicting properties: stability of the multipotent state to support self-renewal, and plasticity for differentiation.

In an elegant theoretical work, Suzuki, Furusawa and Kaneko investigated all possible 5-gene networks in a class of ODE models with cell division and where a single regulatory factor quickly diffuses completely freely between all cells, introduced to model cell-to-

cell communication (Suzuki et al. (2011)). Of the 145,269,760 gene regulatory networks considered, 15,145 had differentiation dynamics, and just 231 showed both unidirectional differentiation and indefinite self-renewal. In these 231 models, the cell begins in a stem cell state where gene expression levels oscillate. The cell then divides to produce more stem cells which continue to exhibit oscillatory gene expression, but where the phases of oscillations are not synchronised between cells. Cell-to-cell communication then is the trigger which forces cells to differentiate, by driving cells out of the oscillating stem cell attractor. Crucially, only cells at a certain phase of oscillation are sensitive enough to this perturbation to differentiate, and so a pool of self-renewing stem cells is always maintained while differentiated cells are also produced.

This theory suggests that oscillatory dynamics is key to stem cell gene regulatory networks, and that stem cell GRNs can be expected to feature oscillation-generating motifs such as negative feedback loops. This is in line with evidence from single-cell studies that demonstrate stem cells show significant heterogeneity in gene expression (Canham et al. (2010); Chambers et al. (2007); Chang et al. (2008); Hayashi et al. (2008); Huang (2009); Macarthur and Lemischka (2013); Moignard et al. (2013a); Toyooka et al. (2008); Warren et al. (2006)), and that haematopoietic cytokines directly induce differentiation and lineage choice rather than only promoting survival of cells that have already committed to a particular lineage as a result of internal decision making (Mossadegh-Keller et al. (2013); Rieger et al. (2009); Sarrazin et al. (2009); Thalheimer et al. (2014)). Several transcription factors have also been found to show oscillatory expression dynamics in imaging studies - for example, Hes1, Nanog, Rex1, Stella and Hex in embryonic stem cells (Chambers et al. (2007); Hayashi et al. (2008); Hirata et al. (2002); Kageyama et al. (2007); Kalmar et al. (2009); Kobayashi and Kageyama (2010, 2011); Kobayashi et al. (2009); MacArthur et al. (2012); Miyanari and Torres-Padilla (2012)).

1.2.2.4 Haematopoietic gene regulatory network models

Recently, two asynchronous Boolean network models of haematopoietic gene regulatory networks have been built, one modelling the common myeloid progenitor and one the haematopoietic stem cell. In an asynchronous Boolean network, each transcription factor is represented by a Boolean variable together with a Boolean update rule that specifies its regulatory logic. Unlike the synchronous Boolean model described above for sea urchin development, the dynamics of these models proceed by a series of single–gene changes. At each update, a gene is chosen uniformly random and its value is updated based upon the value of all genes. This allows stochastic decision making to be incorporated, and means

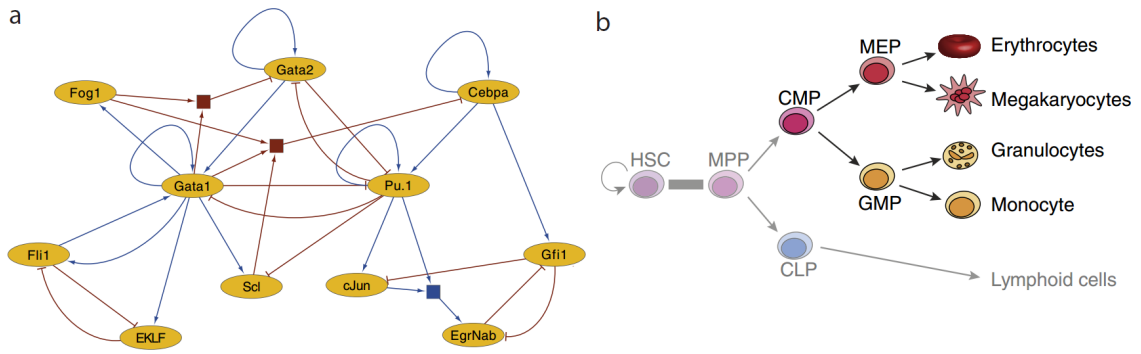


Figure 1.5 A Boolean model of the core transcription factor network active in common myeloid progenitors. (a) Visual representation of the common myeloid progenitor regulatory network model as encoded by the Boolean update rules from Krumsiek et al. (2011). Blue edges represent activation and red edges repression. Square boxes connecting edges represent AND operations. (b) Schematic of the adult hematopoietic hierarchy in bone marrow. The CMP is regulated by the network in (a) to produce multiple outputs: the granulocyte–monocyte progenitor which gives rise to granulocytes, monocytes and other myeloid cells, and the megakaryocyte–erythroid progenitor, which produces erythrocytes and megakaryocytes. HSC, hematopoietic stem cell; MPP, multipotent progenitor; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; MEP, megakaryocyte–erythroid progenitor; GMP, granulocyte–monocyte progenitor.

that different executions of the same model can result in different outcomes.

Krumsiek and Marr et al. constructed an asynchronous Boolean network model of the core transcription factor network active in common myeloid progenitors, following a comprehensive literature survey (Krumsiek et al. (2011)). This model contains 11 haematopoietic transcription factors (Figure 1.5). The complex combinatorial logic governing the interactions between transcription factors is encoded as Boolean update rules using the logical functions And, Or and Not. For example, Gata2 positively regulates its own expression, and is inhibited by Gata1 and Fog1. As both Gata1 and Fog1 are required to repress Gata2 expression, they are combined using And in the Boolean update rule for Gata2. Computational analysis of this model, beginning from an initial transcription factor expression state representing the common myeloid progenitor, revealed an acyclic, 232-element hierarchical state space which recapitulated the steps of myeloid differentiation. This state space contained four terminal stable states, which were found to be in good agreement with microarray expression profiles of megakaryocytes, erythrocytes, granulocytes and monocytes. Once it had been established that this model recapitulated myeloid differentiation, further analyses based upon perturbations to the network were conducted. This analysis demonstrated that

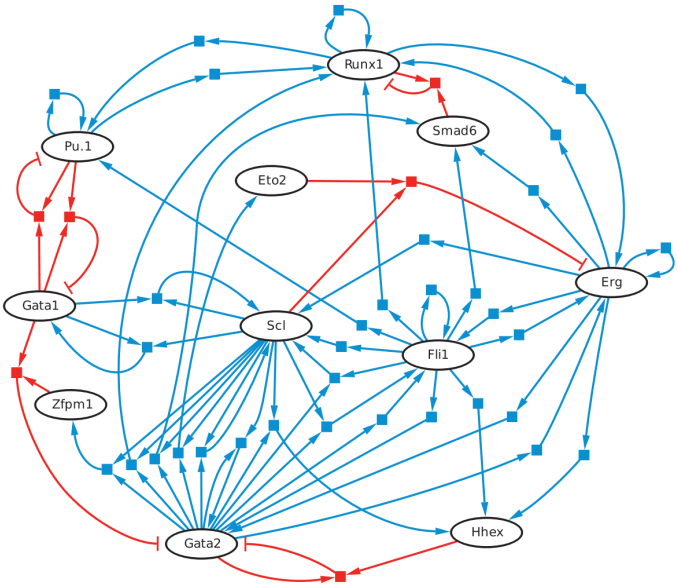


Figure 1.6 A Boolean model of the core transcription factor network active in haematopoietic stem cells, from Bonzanni et al. (2013). Blue edges represent activation and red edges repression. Square boxes connecting edges represent AND operations.

in-silico knockouts were able to reproduce known experimental lineage depletion results and that *in-silico* overexpression reproduced known experimental reprogramming results.

Bonzanni, Garg and Feenstra et al. built a similar model consisting of another set of 11 haematopoietic transcription factors together with experimentally-determined regulatory logic that governs their expression in HSCs (Bonzanni et al. (2013), 1.6). Interestingly, this model exhibited an oscillating "stem-cell" attractor consisting of 32 connected states, which cannot be exited without external intervention. This is in line with the theory of oscillating stem cell gene regulatory networks discussed above. Forced expression of transcription factors is able to commit the model to exit from this attractor and to reach one of 11 stable states, which again were found to be in good experimental agreement with expression profiles of mature blood cell types. Analysis of this model led to the prediction of a new repressive regulatory link between Gata1 and Fli1, which was then subsequently validated. Without this link, expression of Gata1 alone is not sufficient to drive the model out of the stem cell attractor and to differentiate to the erythroid stable state, although it is known that Gata1 overexpression leads to erythro-myeloid commitment experimentally. After introducing this new link, the model behaves as expected. These results on modelling adult haematopoiesis using asynchronous Boolean networks are highly encouraging, and should form the basis of future work. Currently, no similar model exists for developmental

haematopoiesis.

Predictions about the modes of interaction between genes resulting from computational analysis can be tested experimentally through a range of assays. For example, if analysis of a model predicts that gene X is activated by gene A, a ChIP (Chromatin ImmunoPrecipitation) assay can be used to assess whether the protein coded for by A binds to a regulatory region of X. Then, perturbations which prevent the binding of A to this region can be introduced, and the effect that this has on the expression of X can be examined.

1.2.2.5 Abstraction

An important issue that must be considered when attempting to model a system is the level of detail which the system will be represented at. Biological systems have been modelled at different levels of abstraction. At the lowest level are stochastic process models which attempt to capture the precise biochemical events inside a cell, given by chemical master equations (Paulsson (2004); Pedraza and Paulsson (2008); Sjöberg et al. (2009); Van Kampen (2007); Wilkinson (2012)). These chemical events are fundamentally stochastic, driven by random fluctuations of molecules present at very low concentrations (and therefore modelled by discrete rather than continuous variables) and by Brownian motion. Stochastic process models can be simulated using the Gillespie algorithm and analysed using mathematical tools such as the fluctuation–dissipation theorem (Becskei et al. (2005); Chandler and Percus (1988); Gillespie (1977); Paulsson (2004); Sjöberg et al. (2009); Van Kampen (2007)). This class of models has been used to construct small "toy" systems in order to conduct fundamental studies of stochasticity in gene expression (Amir et al. (2007); Elf et al. (2003); Hilfinger and Paulsson (2011); Hilfinger et al. (2012); Huh and Paulsson (2011a,b); Ozbudak et al. (2005); Paulsson (2004, 2005a,b); Pedraza and Paulsson (2008); Zhou et al. (2005)), such as obtaining limits on the suppression of fluctuations by negative feedback loops (Grönlund et al. (2011, 2013); Lestas et al. (2010)).

At a higher level are ordinary differential equation models, which abstract away the discrete copy number of molecules and model molecular concentrations as continuous variables, and reactions as continuous changes in concentrations over time (Elowitz and Leibler (2000); Krumsiek et al. (2010); Mischnik et al. (2014); Narula et al. (2010); Wilhelm (2009)). Analysis of these models is more tractable than analysis of stochastic processes, allowing them to scale to model larger systems, but their dynamics are deterministic and continuous and so fail to capture the stochastic nature of gene expression.

Importantly, the fundamental studies of stochastic models have highlighted how many dif-

ferent parameter values and completely different models can equally fit the same data. For example, fluctuations in protein abundances, which are usually assumed to be due to stochasticity in transcription and translation, can be equally well explained by a model where noise is due to unequal partitioning of molecules between daughter cells at cell division (Huh and Paulsson (2011a,b); Landgraf et al. (2012)). These results can be taken as an argument against trying to build exact quantitative physical models of gene regulatory networks. Given that the processes we are interested in are under-determined, with many unknown reactions, and that kinetic parameters for binding, transcription, translation and degradation rates are unavailable and currently difficult to measure, we should instead try to build more abstract models that capture the essential qualitative properties of the system, and operate at a level for which experimental data is available.

The remarkable success of Boolean models, both in modelling the development of the sea urchin embryo and in predicting haematopoietic cell states from regulatory logic, motivates their use in modelling gene regulatory networks. Asynchronous Boolean networks abstract away details of transcription, translation and molecular binding reactions and represent the status of each modelled substance as either active (on) or inactive (off), while retaining the stochastic nature of events, and capturing the regulatory logic determining whether a gene is activated or not by Boolean update functions. The simplicity of these models means that they can scale to much larger systems, and that powerful computational techniques can be applied to analyse them. It also means that it may be possible to automatically reconstruct them directly from single-cell gene expression data.

1.3 Computational analysis of high–dimensional single–cell gene expression data

Recent advances in protocols, microfluidics technology, and a reduction in costs have opened up a new field of single-cell genomics. This new field promises to provide insights into cellular identity and decision making over more conventional bulk population data, which averages over the properties thousands of cells and therefore obscures the state of individual cells (Moignard and Göttgens (2014)). Single-cell qPCR can simultaneously measure the level of expression of tens to hundreds of genes, while the newer technique of single-cell RNA-sequencing can sample the whole transcriptome.

After experimental measurement, data must firstly be processed and normalised to ensure correct interpretation. Once these steps have been carried out subsequent analysis can be

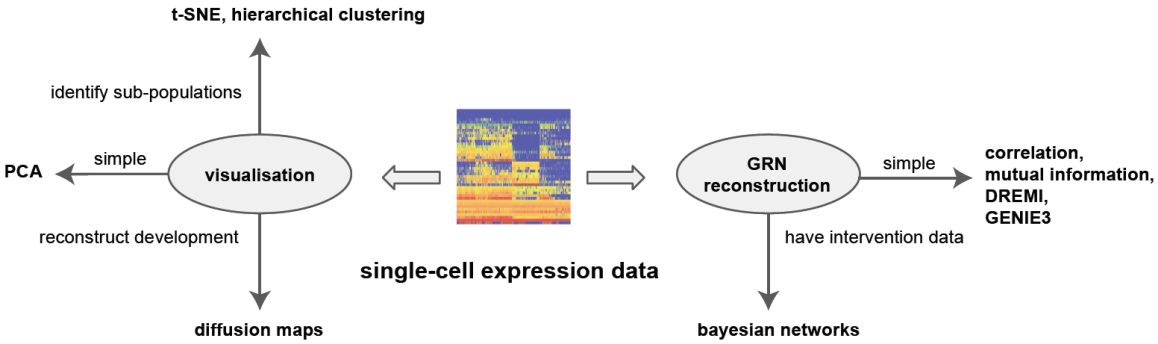


Figure 1.7 Overview of different single-cell analyses.

applied to answer specific biological questions.

Typically one of the first questions a researcher will want to ask about their single-cell expression data set is whether interesting sub-populations with characteristic gene expression profiles can be identified (Amir et al. (2013); Buganim et al. (2012); Dalerba et al. (2011); Jaitin et al. (2014); Moignard et al. (2013a); Wilson et al. (2015)). These sub-populations might represent previously unidentified cell types or cells with an abnormal phenotype. For example, in a study of the immune system, two separate populations might correspond to activated and naive cells, or in a patient sample, to cancerous and healthy cells (Mahata et al. (2014); Patel et al. (2014); Shalek et al. (2014); Spitzer et al. (2015)). Once identified, the sub-populations can be isolated and investigated further. Population-level gene expression data, on the other hand, would average out the differences between these groups, giving a representative view of neither.

Once structure has been identified, the researcher can investigate potential biological processes that have been captured in the data. Often, the data are representative of a developmental or differentiation time-course, with early cells such as stem cells or early progenitors progressing to more mature cells (Bendall et al. (2014); Moignard et al. (2015); Trapnell et al. (2014)). In this case, the single cell profiling data set can be used for gene regulatory network reconstruction. I will describe several techniques for reconstructing regulatory networks (Figure 1.7). Some of these methods have been adapted from analyses of population data, and some have been specifically developed to take advantage of single-cell resolution data.

1.3.1 qPCR on the Fluidigm BioMark

The Fluidigm BioMark platform uses microfluidics devices to scale back reagent and sample requirements, thereby facilitating thousands of parallel qPCR reactions and allowing up to 96 genes to be assayed in a single cell. Initial data processing takes place using the Fluidigm Real-Time PCR Analysis Software. Like conventional qPCR, the BioMark outputs Ct values, and the software allows sample and assay names to be assigned along with the quality thresholds, baseline correction methods and Ct thresholds used to calculate the final Ct values.

Next, expression values that fall outside of the linear range of the BioMark HD or the assays are excluded from further analysis. To do this, a limit of detection (LOD) is calculated from standard curves for each primer set as the last Ct value at which amplification can be reliably and repeatedly detected (Livak et al. (2013); Trapnell et al. (2014)). Ct values higher than the LOD, as well as samples where the amplification has failed entirely or where the amplification curves have failed quality control are usually given the same value as the limit of detection and treated as not detected.

Additional filtering can be used to exclude whole genes or samples. For example, genes may be excluded where there is amplification in typically >10% of no template controls, and where the amplification level in no template controls is too similar to that of single cells to be sure that the expression in the cells is real. In published studies, cells have been excluded from the analysis based on a number of criteria, including lack of expression of key or housekeeping genes, expression of no or low numbers of cells, or where the expression of particular genes differs significantly from the population (Buganim et al. (2012); MacArthur et al. (2012); Moignard et al. (2013a); Pina et al. (2012)), although these can also occur due to the choice of genes and transcriptional bursting rather than due to a poor quality or missing cell.

Single cell expression data are typically log-normally distributed so it is useful to view data on a Log2 scale. The final step of processing therefore converts the data either to ΔCt values normalized against one or more housekeeping genes which exhibit stable expression across the populations (Buganim et al. (2012); Guo et al. (2010); MacArthur et al. (2012); Moignard et al. (2013a, 2015); Pina et al. (2012); Swiers et al. (2013a)), or as the Log2 expression above the LOD (PCR cycles above background; Log2Ex Guo et al. (2013); Stahlberg et al. (2011)). Log2Ex values can be further normalized to remove variability due to factors such as cell size (Livak et al. (2013)).

1.3.2 Single cell RNAseq

Single cell RNAseq (scRNAseq) has recently come to the fore for transcriptomics due to increases in multiplexing and concurrent decreases in price. Compared with qPCR, it offers the potential to study the entire transcriptome rather than a specific set of pre-selected genes, so has a much wider potential for discovery. However, there are many current challenges both for processing samples and analyzing data (Macaulay and Voet (2014); Stegle et al. (2015)).

There are many different scRNAseq protocols which can capture different aspects of the transcriptome depending on the priming and reverse transcription (RT) methods used. Typically, either the 5' or 3' end of the transcript is captured (Hashimshony et al. (2012); Islam et al. (2011)), although some methods can capture entire transcripts (Picelli et al. (2013); Tang et al. (2009)). Samples are multiplexed using indexed primers during library preparation, with 96 to 384 individual cells sequenced per lane of a flow cell. After sequencing, samples are deconvoluted based on index sequences, and normalised read counts are generated for further analysis. Alternatively, short and unique DNA sequences (unique molecular identifiers, UMI) can be incorporated into every transcript during the RT step to act as barcodes to enable molecule counting. Regardless of how many times a transcript-UMI pair is sequenced, it can only have come from a single mRNA within the cell and so is only counted once, with the total number of UMIs per transcript summed to give an absolute expression count for each gene (Kivioja et al. (2011)). However, this currently only allows for the sequencing of the 3' end of the transcript, providing information about expression levels but not splicing.

Quality check of samples is an important step before downstream analysis. An important quality control method for scRNAseq is the inclusion of extrinsic standards to facilitate normalization and comparison between single cells. Typically, RNA standards of known concentration and sequence, such as the External RNA Control Consortium (ERCC) set of 92 artificial RNA molecules (Jiang et al. (2011), are spiked into the reverse transcription step. These molecules should be amplified uniformly across samples, so can be used to estimate RT efficiency, technical variation in library preparation and to indicate which genes show real biological variation as well as technical noise. Spikes can additionally be used to identify cells with degraded RNA, for example where the percentage of mapped reads is particularly low compared with reads mapped to spike molecules. Other important metrics which are used for quality control and to discard poor-quality cells include the fraction of reads mapped to mitochondrial genes (a large fraction is believed to be indicative of the cell undergoing apoptosis — Islam et al. (2014); Stegle et al. (2015)). Principal component anal-

ysis (discussed later), can also be used to identify outlier cells, based upon the assumption that good-quality cells will cluster together while poor-quality cells will be isolated (Stegle et al. (2015).

Samples undergo initial quality control prior to alignment, with tools originally developed for bulk RNA-seq such as fastqc which monitor sequencing quality, GC nucleotide content, sequence length and so on. Reads are assigned to individual cells based on their indexes, the sequencing adapters are trimmed off and the resultant sequences are mapped to a reference transcriptome using existing alignment tools such as TopHat (Trapnell et al. (2009)), Star (Dobin et al. (2013)) or GSNAP (Wu and Nacu (2010)). Tools such as HTseq (Anders et al. (2014)) are then used to generate read counts per gene. Further quality control, as discussed above, can then be carried out. Normalisation is required to account for differences in sequencing depth between samples, which is calculated from the total mappable reads and the ratio of mapped reads to those coming from spike molecules. However, adequate normalization of scRNAseq data is an ongoing challenge (Stegle et al. (2015)) as much is still unknown about technical variation in library preparation and sequencing bias towards particular transcripts.

1.3.3 Visualisation

High-dimensional data sets can be hard to visualise. A two or three dimensional data set can be directly plotted to try to reveal structure in the data (Figure 1.8). This is not possible with high dimensional data such as a single-cell gene expression data set, which has a dimension corresponding to each measured gene. In the field of machine learning, a number of clustering and dimensionality reduction techniques have been developed to help aid visualisation of high-dimensional data (Bishop (2006); Hastie et al. (2009)). Clustering algorithms attempt to group data points into subsets called clusters, where data points within a cluster are more similar to each other than to points from different clusters. Dimensionality reduction algorithms attempt to transform the high-dimensional data set into a lower-dimensional (2 or 3) representation that can then be directly plotted and visualised.

1.3.3.1 Hierarchical clustering

Agglomerative hierarchical clustering has been used to identify sub-populations in single-cell data (Guo et al. (2013); Moignard et al. (2013a)). Rather than seeking to identify a predetermined number of clusters, the algorithm recursively builds a hierarchical represen-

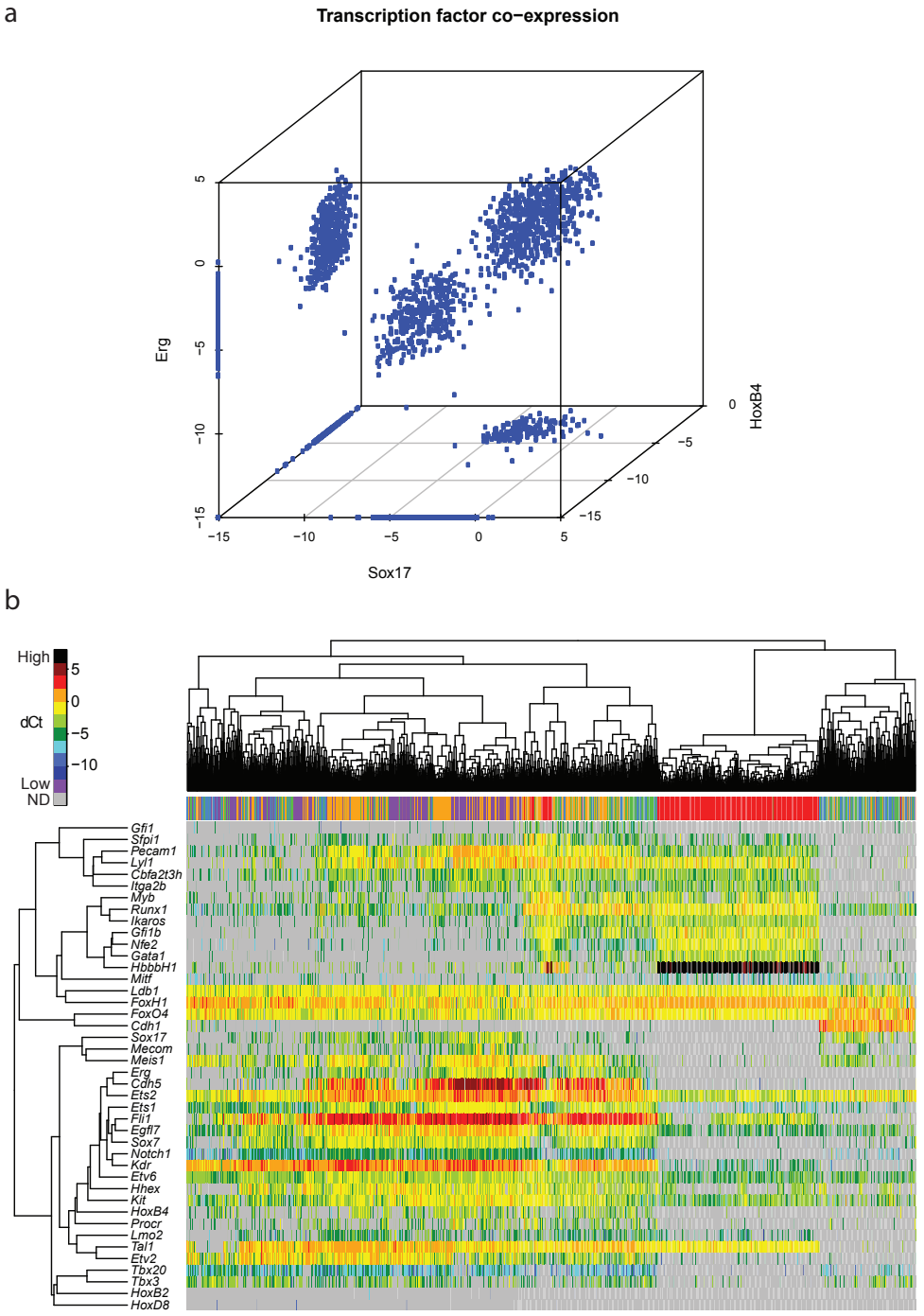


Figure 1.8 High-dimensional data can be hard to visualise. (a) Plotting the expression of three genes against each other to try to uncover their relationship. (b) Hierarchical clustering of a high-dimensional single-cell qPCR data set with 40 genes and 3934 cells. Rows represent genes and columns represent cells. Left-hand side colour bar shows measured ΔCt level of expression of genes. Top colour bar shows cell types—blood cell progenitors (red) fall into one large cluster while other cell types separate into two more large clusters and do not separate by cell type.

tation of the data where each level organises the data into a different number of clusters. This makes the algorithm useful for exploratory analysis.

At the beginning of the algorithm each data point is placed into its own cluster. Then, at each subsequent step the two most similar clusters from the previous iteration are merged into one. The algorithm terminates when all of the data lies in a single cluster (Hastie et al. (2009)). The results of hierarchical clustering can be plotted as a heat map (a coloured representation of the data matrix, reorganised according to the clustering) with a dendrogram, which is a binary tree showing the hierarchical neighbour relationships between clusters. As we go to higher levels in the dendrogram, the dissimilarity between merged clusters increases. By examining the reorganised expression matrix, and the cell types and gene expression patterns of closely placed points, natural clusters can often be discerned by eyea (Figure 1.8)

Before hierarchical clustering can be performed, two measures of similarity need to be specified: a notion of distance between pairs of data points, and a notion of distance between clusters (the linkage criterion), defined in terms of the distance between data points. For the distance between data points, the Euclidean, Manhattan or Spearman correlation distance can be used. For the linkage criterion between clusters A and B, one distance is the nearest neighbour distance (known as single linkage), which is the distance between the point in A and the point in B which are most similar. A second distance is the farthest neighbour (known as complete linkage), which is the distance between the point in A and the point in B which are least similar.

Care must be taken when interpreting the results of hierarchical clustering, keeping in mind that different choices of dissimilarity measure and linkage criterion will result in different hierarchies, and that the algorithm will always impose a hierarchy on the data whether or not one truly exists.

Many other clusterings algorithms exist, but hierarchical clustering and related methods stand out in their utility for exploratory visualisation. Spectral clustering is closely related to diffusion maps (Nadler and Galun (2007)) (discussed later). DBSCAN is a very commonly used algorithm which groups together points with many nearby neighbours (Ester et al. (1996)). K-means clustering clustering places each point into the cluster with the closest mean, but requires the desired number of clusters to be specified a-priori (and is therefore best used for classification after using another method for exploratory visualisation) (MacQueen (1967)).

The SPADE algorithm was introduced specifically for the analysis of single-cell data, and is

based upon firstly applying hierarchical clustering, and then linking clusters together using a minimum spanning tree to infer developmental progression, while taking into account the existence of rare cell populations via density-dependent downsampling (Qiu et al. (2011)). The BackSPIN algorithm is conceptually similar to hierarchical clustering, but seeks to avoid noise from cell dissimilarity caused by uninformative genes. It works by sorting the gene expression matrix through cell-cell and gene-gene similarity (Zeisel et al. (2015)). Grün et al. recently introduced an algorithm, RaceID, designed specifically for identification of rare cell types in single cell data (Grün et al. (2015)).

1.3.3.2 Principal component analysis

The most ubiquitous tool used for dimensionality reduction is principal component analysis (PCA). PCA is used to find a projection of the data onto a smaller linear subspace, such that the variance of the projected data is maximised (the data points are spread out as much as possible) (Bishop (2006); Hastie et al. (2009)).

PCA finds a sequence of uncorrelated best linear approximations of the data, which are ordered in decreasing order of variance and are known as principal components. The first two or three of these components can be retained and plotted as a scatter plot to perform dimensionality reduction (Guo et al. (2013); Kumar et al. (2014); Moignard et al. (2013a)). Equivalently, PCA can viewed as an instance of the Multidimensional Scaling (MDS) algorithm with Euclidean distances. MDS attempts to preserve all pairwise distances between data points in the high-dimensional space, as best as possible (Borg and Groenen (2005)). In general, the method will fail to preserve all pairwise distances perfectly. For example, in a ten dimensional data set, up to 11 data points may be mutually equidistant, while there is no way to accurately represent this in a three dimensional plot.

The advantages of principal component analysis are its simplicity, its computational efficiency and its direct interpretation in terms of linear combinations of genes. A disadvantage is that it fails to capture non-linear structure in the data. Single-cell gene expression data in particular can be expected to be highly non-linear (Figure 3d). Manifold learning and graph-based visualisations, discussed next, attempt to address this weakness. Non-linear generalisations of PCA also exist, most notably kernel PCA, which also falls into the class of manifold learning algorithms (Schölkopf et al. (1998)).

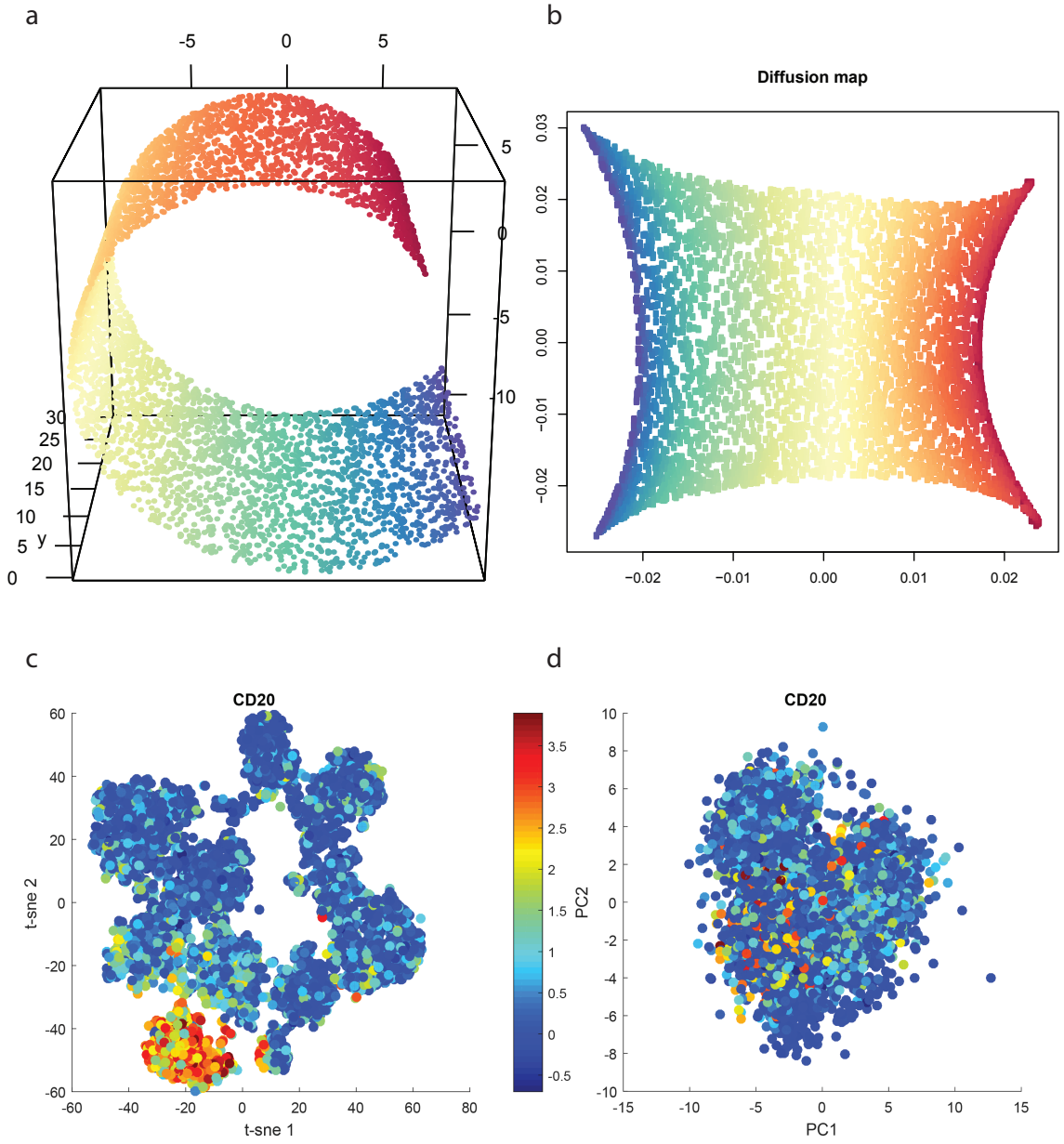


Figure 1.9 Manifold learning. (a) A two-dimensional curving manifold embedded in three dimensions. (b) Diffusion map applied to 'unfold' the manifold to a rectangle, giving one possible way of representing the three-dimensional data in two dimensions. (c) t-SNE separates bone marrow cells measured by cytometry into different immune cell types. Points are coloured by CD20 expression, a B-cell cell-surface lineage marker. (d) PCA, a linear projection method, fails to separate between the different immune subtypes on the first two principal components.

1.3.3.3 Non-linear manifold learning

In general, there is no way to represent a high-dimensional data set in a lower dimensional space without discarding information. Different dimensionality reduction tools therefore aim to embed the data in a way that preserves some particular property of interest. I will focus the remainder of this section on two non-linear dimensionality reduction methods that have recently been used to visualise single cell gene expression data: t-Distributed Stochastic Neighbor Embedding (t-SNE) and diffusion maps.

t-SNE aims to preserve the pairwise distance between points, but (unlike MDS/PCA) only between those points which are very close neighbours in the high dimensional space, focusing only on preserving local structure rather than attempting to preserve pairwise distances between all points (Maaten and Hinton (2008)). This allows the global structure of the embedding to become non-linear, as distances at different regions of the embedding are allowed to correspond differently to distances in the high dimensional space. Diffusion maps attempt to reconstruct the global non-linear connectivity of the data from a local random walk on the data, and place points close together in the low-dimensional map if they are connected by many short paths in the high-dimensional space.

Diffusion maps and t-SNE belong to a class of techniques known as manifold learning algorithms. Manifold learning is based on the hypothesis that the dimensionality of the data under consideration is only artificially high, and that rather than being uniformly distributed throughout the high dimensional space it actually lies on a lower dimensional non-linear manifold that curves through the high dimensional space (Figure 1.9). This manifold hypothesis seems particularly appropriate for single-cell gene expression data as the expression states that a cell can take are highly constrained by an underlying gene regulatory network. A cellular state therefore has relatively few degrees of freedom in terms of the states it can immediately progress to, an idea that was formalised in Waddington's epigenetic landscape (Goldberg et al. (2007)). This landscape can also be expected to be non-linear because of complex gene interactions, waves of gene expression and positive and negative feedback loops in the gene regulatory network. PCA can be considered as a linear manifold learning algorithm, that assumes data lies on a linear hyperplane.

The aim of a non-linear manifold learning algorithm is to reconstruct the geometry of the low-dimensional manifold the data lies on from the only information we have: the similarities between data points. Key to these algorithms is the idea that it is local distances, similarities between nearby points that are important for reconstructing this geometry.

1.3.3.4 t-SNE

t-SNE defines a Gaussian probability distribution over pairs of data points in the high-dimensional space, that captures the pairwise similarity of points. The probability of a pair being chosen is high if the points are very similar in terms of their high-dimensional gene expression profiles, and very close to zero if they are dissimilar. A second distribution over pairs of points in the low dimensional embedding is then defined, this time as a Student's t-distribution. Points are placed on the two- or three- dimensional plot, and the discrepancy between these two probability distributions (the Kullback–Leibler divergence) is iteratively minimised via a gradient descent optimisation method, shifting points around until this discrepancy reaches a minimum (Maaten and Hinton (2008)).

A disadvantage of t-SNE is that it can be slow to compute. For this reason, a Barnes–Hut approximation algorithm has been developed which can scale better to larger data sets (van der Maaten (2014)). t-SNE has been used very successfully to dissect heterogeneity in leukemia samples using single-cell mass cytometry data (Amir et al. (2013)), and to identify an improved cell-sorting strategy for hematopoietic stem cells by separating true stem cells from non-stem cells in combined single-cell qPCR and single-cell indexed flow cytometry data (Wilson et al. (2015)).

1.3.3.5 Diffusion maps

Unlike t-SNE, which tends to pull data apart into separate clusters, diffusion maps tend to organise the data into a single continuous manifold and are therefore particularly appropriate when the data is sampled from a developmental or differentiation process that we wish to reconstruct (Figure 3). The algorithm was first introduced in the context of biology by Haghverdi, Buettner, and Theis, adapting it to deal with uncertainties or missing measurement values in qPCR data, and adding density normalisation to cope with heterogeneities in data sampling (Haghverdi et al. (2015); Moignard et al. (2015)).

Diffusion maps are based upon the idea of reconstructing the global geometry of the data set by constructing and iterating a random walk on the data points, and attempt to accurately approximate the so-called "diffusion distance" between data points when mapping to a lower-dimensional space (Coifman et al. (2005); Nadler et al. (2007)). This diffusion distance is small if there are many high-probability short paths connecting the two points, and large if the points are connected only by long paths or low-probability transitions. When reducing to a lower-dimensional space, the diffusion algorithm attempts to place points with

a low diffusion distance nearby in the map.

The diffusion map algorithm works by constructing a transition matrix on the data, where the probability of jumping from one data point to another in one step is high if the two data points are similar in the high-dimensional space. If the points are dissimilar, this probability is very close to zero. It then approximates the diffusion distance in lower dimensional space without explicitly iterating the random walk, which would be computationally expensive. The diffusion map algorithm is computationally efficient, and, because it integrates over all paths, robust to noise, unlike some manifold learning algorithms. For a review of other manifold learning approaches, see Lin et al. (2015).

1.3.4 Network reconstruction

Visualisation of gene expression data is an important first step towards understanding a developmental or homeostatic process. However, to gain a full understanding of the underlying biology we need to establish mechanistic models of gene regulatory networks.

1.3.4.1 Statistical relationships between genes

When trying to infer regulatory interactions between genes one of the most obvious things to look for is correlation in gene expression levels. If there is a strong correlation between two genes, this may indicate that one directly regulates the other. Performing this analysis on all possible pairs and selecting strong and statistically significant relationships results in a relevance network, which is an undirected graph where edges between genes indicate a potential interaction (Figure 1.10). There are two types of edge: positive edges where strong positive correlation indicates a potential activation and negative edges where strong negative correlation indicates a potential repression (Butte and Kohane (2003)).

The standard Pearson correlation coefficient is a measure of the linear dependence between two variables. As genes may not exhibit a linear relationship the Spearman rank correlation is generally preferred. Spearman correlation measures how well the relationship between the two variables can be fit by a monotonic function. A measure from information theory called mutual information is more general still and can capture more complex relationships.

These statistical relationships can scale to huge data sets, and have been successfully applied to find previously unknown regulatory links in single cell data (Moignard et al. (2013a)). However, relevance networks are undirected and can be very dense, with almost all gene

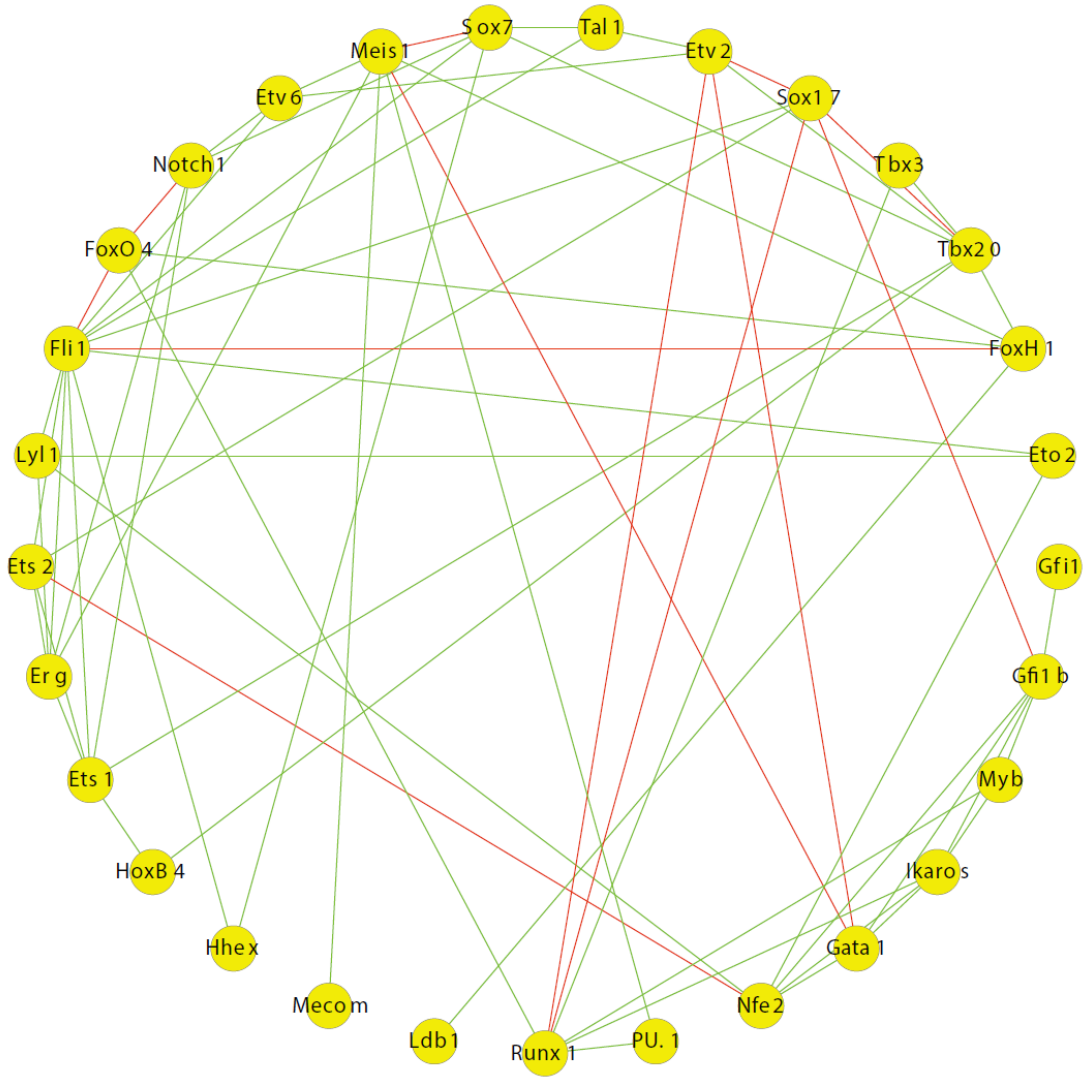


Figure 1.10 Relevance network obtained from partial correlation analysis. Green: activation; red: repression.

pairs showing significant correlation. Partial correlation attempts to address this second issue, by calculating correlation after first controlling for the effect of all other genes, and therefore retaining the links that are most likely to be direct interactions (Figure 1.10).

To understand another problem, consider two subpopulations, one of which expresses gene A but not gene B, the other expresses gene B but not A. Correlation would suggest a very strong negative link between the two genes, although there is no strong reason to believe they directly regulate each other. One way to address this is to compute the correlation only on cells which coexpress the genes of interest (Pina et al. (2015)).

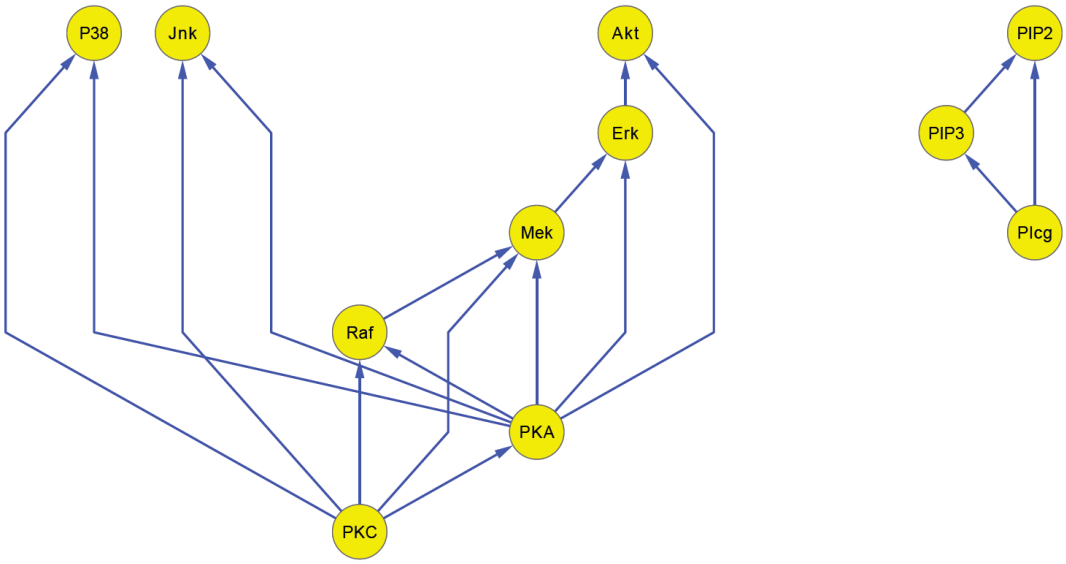


Figure 1.11 Bayesian network for T-cell signalling.

Other methods for detecting statistical signals in gene expression data exist. One notable method is GENIE3, which constructs random forests of decision trees (Huynh-Thu et al. (2010)). GENIE3 was best performer in the DREAM5 Network Inference challenge for population data (Marbach et al. (2012)), and has been applied to single-cell data (Ocone et al. (2015a)). A reweighted mutual information measure known as DREMI, specifically designed for single-cell data has recently been introduced (Krishnaswamy et al. (2014)).

1.3.4.2 Learning Bayesian networks

A Bayesian network over a set of variables X (which is our case represent genes) is defined by a directed acyclic graph G that represents conditional independence relations between variables, coupled with a set P of local probability distributions associated with each variable. Together, the graph and local probability distributions define the global joint probability distribution for X (Heckerman (1996)). Any two variables in a Bayesian network are conditionally independent, given the value of their parents.

Although Bayesian networks are often used to represent causal relationships, care must be taken when interpreting them this way. A directed edge from x to y does not necessarily imply that x is causally dependent on y, only that they are not conditionally independent. Often an equivalent graph structure is equally compatible with the data: $a \rightarrow b \rightarrow c$ and $a \leftarrow b \leftarrow c$

represent the same conditional independence relationships (but $a \rightarrow b \leftarrow c$ does not). To establish causal relationships, Bayesian networks often must be coupled with interventional data.

Given a Bayesian network, inference can be performed in the model to predict the effect of perturbations on the probability distributions of downstream genes. It is the directed acyclic graph and conditional independence structure of Bayesian networks which allows efficient inference to be performed, and permits efficient learning of models from data.

When we learn the structure and parameters of a Bayesian network, we attempt to find a model that induces a probability distribution that fits the data as closely as possible. There are two general approaches to learning the structure of a Bayesian network. The first, constraint-based learning, is based upon using statistical tests to directly recover conditional independence relationships in the data. A graph which satisfies these relationships is then constructed.

In the second approach, score-based learning, each candidate network is assigned a score which measures how well it fits the data (Teyssier and Koller (2012)). We then try to maximise this score. There are a super exponential number of such candidate networks, and no clear way to efficiently find an optimal structure. Instead, most algorithms apply a local search. We start from a random network or a network which encodes our prior knowledge, and use a local search algorithm such as greedy Hill climbing, tabu search or simulated annealing, stopping when we are unable to find a better candidate network. At each step in the search we add, delete, or reverse the direction of an edge, being careful not to introduce any directed cycles, and assess whether the change improves the score of the candidate network. Another approach is to use Markov Chain Monte Carlo methods to sample a large number of high-scoring networks, and then take a network structure which is an average of these models. The scoring function is usually chosen to penalise complicated network structures and favour simple ones, balancing the conflicting goals of a close match to the data and a simple model, and helping to avoid over-fitting.

Bayesian networks were first applied in the context of genomics by Friedman et al. (2000) to infer networks from population microarray data. They have since been applied by Sachs et al. (2005) to reconstruct signalling networks from single cell flow cytometry data taken from primary human T cells. Sachs et al. measured 11 phosphorylated proteins and phospholipids in 5400 individual cells spanning nine different conditions. Seven of these conditions directly perturbed variables of the network by activating or inhibiting phosphorylation. The differences between these perturbed populations were then used to infer causality. Data were first discretised to 3 levels (low, medium and high expression), and then learning algo-

rithms were applied to construct a Bayesian network which was subsequently successfully validated against existing literature.

One of the key insights of this paper is the need for perturbation data to reconstruct an accurate Bayesian network from single cell data. If we have two correlated variables, X and Y, and we find that direct inhibition of X affects the value of Y and that direct inhibition of Y does not affect X, we can conclude that Y is downstream of X. A learning algorithm can then often determine the direction of additional edges downstream of the perturbed variables, even when these edges were not directly perturbed.

Bayesian networks have been successfully applied to dissect connections between components of signalling pathways. However, they do suffer from two drawbacks that limit their application to reconstruction of wider gene regulatory networks. Firstly, as concluded by the Sachs study, to infer accurate networks the single cell data needs to be coupled with intervention data. Generating such intervention data is very time consuming and often impractical, and cannot be done without disturbing the wild-type system that we are supposed to be studying. Secondly, Bayesian networks are acyclic, and have no feedback. Feedback is a crucial component of gene regulatory networks.

1.4 Solving combinatorial problems

Some computational problems, such as the ones dealt with in this thesis, are combinatorial in nature, in that they involve objects with a large number of possible configurations that grows exponentially as the problem size is increased. There is often no obvious way to proceed directly to a solution without exploring a large space of candidate configurations (Knuth (2016)).

Sometimes an algorithm for these problems can be found, with a worst-case running time that scales polynomially with the size of the problem (Agrawal et al. (2004); Edmonds (1965); Karmarkar (1984); Kasteleyn (1963); Valiant (2006, 2008)). In other cases, we can prove a problem is NP-complete, implying that such an algorithm is unlikely to exist since it would imply a polynomial time algorithm for all of the thousands of other known NP-complete problems (Arora and Barak (2009)). This would resolve the famous P vs. NP problem and earn the discoverer of the algorithm a million dollars from the Clay Mathematics Institute. Detection of a stable state attractor of a Boolean network, one of the problems discussed below, is NP-complete (Akutsu et al. (1999); Tamura and Akutsu (2008)).

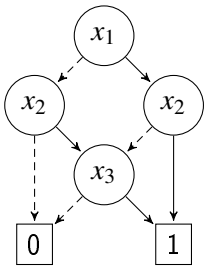


Figure 1.12 A BDD for the majority function $(x_1 \wedge x_2) \vee (x_2 \wedge x_3) \vee (x_3 \wedge x_1)$ (Knuth (2016)).

Computer scientists have developed techniques for the efficent exploration and symbolic manipulation of combinatorial spaces. The synthesis algorithm introduced in chapter 2 of this thesis uses a Boolean Satisfiability solver as a sub procedure. Algorithms for finding attractors of Boolean networks employ Binary Decision Diagrams or Boolean Satisfiability solvers.

1.4.1 Binary Decision Diagrams

A Reduced Ordered Binary Decision Diagram (ROBDD, or simply BDD) is a rooted, directed acyclic graph with one or two terminal nodes of out-degree zero labelled 0 or 1, and with all other nodes having out-degree two and labelled with a variable u (Bryant (1986); Knuth (2016)). A BDD is ordered — variables always occur in the same order along any path from root to terminal, and reduced — the left and right branches of a node cannot lead to the same node, and there are no two distinct nodes n and n' with isomorphic subgraphs.

A BDD is essentially a compressed representation of the truth table of a Boolean function, with each path to a root node representing an evaluation of the function. A left branch at a variable represents an assignment of 0 to that variable, and a right branch represents an assignment of 1. The value of the function is given by the terminal node that the path ends at. Given a fixed ordering for the variables of a function, there is a unique BDD representation for that function. An example BDD for the majority function $(x_1 \wedge x_2) \vee (x_2 \wedge x_3) \vee (x_3 \wedge x_1)$ is shown in Figure 1.12.

BDDs allow for the efficient representation and manipulation of sets of objects and relations on objects. Although in the worst case a BDD requires exponential space to represent all the solutions to a function, in many practical cases they allow the symbolic exploration of state spaces that would be impossible to represent explicitly.

1.4.2 Boolean Satisfiability

Often, the most efficent method for solving a combinatorial problem is to translate it to a symbolic representation in Boolean propositional logic, and then treat it as an instance of the Boolean Satisfiability problem (SAT), exploiting highly optimised SAT solvers (Barrett et al. (2009); Knuth (2016); Konev and Lisitsa (2014)). SAT is the canonical NP-complete problem (Cook (1971)), and solvers run in exponential time in the number of variables in the worst case. However, modern SAT solvers run surprisingly quickly on many real-world instances, and are now routinely used in industrial applications such as verification of hardware and software (Biere et al. (1999); Kaivola et al. (2009)).

SAT
Input: A Boolean formula in conjunctive normal form $C_1 \wedge C_2 \cdots \wedge C_n$, where each clause C_i is a disjunction of literals (a literal is a variable, x_j, or its negation, $\neg x_j$), $l_1 \vee \cdots \vee l_m$.
Question: Is there an assignment of the n variables that satisfies all clauses (causes each to evaluate to true)?

A straightforward way to solve this problem is using a classical backtracking algorithm. At each step in the algorithm we select a previously unchosen variable x_i and set its value to 0 or 1. We stop when we have successfully assigned a value to each variable without causing a conflict that leads to a clause becoming unsatisfiable. If we introduce a conflict, we backtrack to a previous stage and reverse a variable assignment. If all variable assignment choices lead to conflicts, the formula is unsatisfiable. This simple backtracking algorithm fails to take advantage of unit clauses, which consist of only one literal which therefore is forced to be 1 in order to make the clause satisfiable. Branching on unit clauses is unnecessary.

The DPLL (Davis–Putnam–Logemann–Loveland, Davis et al. (1962)) algorithm extends this backtracking procedure with unit propogation, which detects when assignments lead to a clause becoming a unit clause – only one literal remains unassigned, and for the clause to be satisfied this literal must be assigned to 1. This literal can therefore be assigned to the value 1, which in turn may cause other clauses to become unit clauses. Iterating this procedure can dramatically increase the speed of SAT solving on real-world instances. Unit propogation can both satisfy clauses and lead to conflicts.

Modern SAT solvers used for industrial applications are based upon the CDCL (Conflict-Driven Clause Learning) algorithm, which adds clause learning and non-chronological back-

```
function CDCL(F)
    assignment ← UNITPROPOGATION(F, ∅)
    if assignment = CONFLICT then
        return UNSAT
    end if
    while NOT(ALLVARIABLESASSIGNED(F, assignment)) do
        assignment ← PICKBRANCHINGVARIABLE(F, assignment)
        assignment ← UNITPROPOGATION(F, assignment)
        if assignment = CONFLICT then
            (learntClause, assignment) ← CONFLICTANALYSIS(F, assignment)
            F ← F ∪ {learntClause}
            if assignment = CONFLICT then
                return UNSAT
            end if
        end if
    end while
    return (SAT, assignment)
end function
```

Figure 1.13 Conflict-Driven Clause Learning algorithm for SAT.

tracking (Eén and Sörensson (2004); Marques-Silva et al. (2009); Marques Silva et al. (1996); Moskewicz et al. (2001); Zhang (1997)). This algorithm is shown in Figure 1.13. For each conflict that is generated during the search, CDCL constructs a new clause which identifies the root cause of the conflict. This new clause, which is implied by the existing clauses, is added to the formula and guides the search away from encountering the same conflict again. Conflict clauses are also analysed and used to backtrack multiple levels, to the earliest assignment choice that led to the conflict. Together, these two features can prune a large portion of the search space, allowing the algorithm to find a solution or prove unsatisfiability faster.

Further increasing the efficiency of CDCL-based SAT solvers is an active area of research. The SAT competition is held annually, evaluating solvers on a range of benchmarks. Topics of research include the development of heuristics for when learnt clauses should be discarded in order to save memory and to speed up propagation (Audemard and Simon (2009, 2012); Biere (2014)), and the design of data structures to efficiently implement the unit propogation and backtracking procedures (Eén and Sörensson (2004)).

There are many important restrictions and generalisations of Boolean satisfiability. Propositional Horn clauses are Boolean formulas of the restricted form $\neg x_1 \vee \neg x_2 \vee \ldots \vee \neg x_i \vee x_{i+1}$.

The problem of deciding whether a set of propositional Horn clauses is satisfiable, known as HORNSAT, is *P*-complete, meaning that it is solvable in polynomial time and that every other problem with a polynomial time algorithm can be efficiently reduced to it. HORNSAT can be solved in linear time in the total number of occurrences of literals by unit propogation (Allender et al. (2005); Dowling and Gallier (1984); Scutella (1990)). Satisfiability modulo theories (SMT) solvers extend Boolean propositional logic with additional *theories*, such as integers, linear arithmetic over real numbers, or arrays (Moura et al. (2007); Sebastiani (2007)). These solvers usually work by integrating a SAT solver with theory-specific solvers. Satisfiability in more expressive logics has found application in the verification of software. Separation logic is used for expressing and verifying safety properties of programs which can directly manipulate memory (Antonopoulos et al. (2014); Berdine et al. (2004); Brotherston and Kanovich (2014); Brotherston et al. (2014); Reynolds (2002)). First-order Horn clauses with linear integer arithmetic and uninterpreted functions have been used in the verification of procedural and multi-threaded programs (Beyene et al. (2013); Bjø rner et al. (2013); Rümmer et al. (2014)).

1.4.3 Formal verification and synthesis of computer programs

In computer science, synthesis is a general term for the counterpart of verification. In verification, a hand-built model or computer program is given, along with a specification of how it ought to behave. Then the model is checked to ensure it satisfies the specification. This can often be done automatically through model checking algorithms which make use of BDDs or SAT solvers that check all possible executions of the program satisfy the specification. In synthesis, a specification is given and a model is *automatically* generated that satisfies this specification (Pnueli and Rosner (1989); Vardi (2008)).

In recent years much progress has been made on the usage of SAT and SMT solvers for synthesis. Essentially, the existence of a program that solves a certain problem is posed as a satisfiability query. Then, a solver tries to search for a solution to the query, which corresponds to a program. For example, Srivastava et. al. show that the capabilities of SMT solvers to solve quantified queries enable the search for conditions and code fragments that match a given specification (Srivastava et al. (2010, 2013)). Similarly, Solar-Lezama et. al. build a framework for writing programs with “holes” and letting a search algorithm find proper implementations for them (Solar-Lezama et al. (2005)). Beyene et. al. have shown how constraint solving can also be used in the context of infinite-state reactive programs (Beyene and Rybalchenko (2014)).

1.5 Aims of this PhD

New single-cell resolution gene expression measurement technology provides snapshots of the gene expression states of that cells that make up a biological tissue, a level of detail which has not been available before. The aim of this PhD was to investigate the possibility of using this new high resolution data to reconstruct mechanistic computational models of gene regulatory networks, which could then be tested experimentally, and used to make useful predictions. This aim led to several objectives:

1. To develop and implement an algorithm for the reconstruction of executable models of gene regulatory networks from single-cell gene expression data.
2. To apply this algorithm to a new data set covering 3934 single cells measured during early embryonic blood development, in order to reconstruct a predictive model of primitive haematopoiesis and generate new biological insights.
3. To develop a user-friendly and efficient graphical tool which can be used by biologists to reconstruct gene regulatory network models from new single-cell gene expression data sets as they become available.

Chapter 2

Boolean Networks

2.1 Definition

An *asynchronous Boolean network* (ABN) is $B(V,U)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of *variables*, and $U = \{u_1, u_2, \ldots, u_n\}$ is a set of Boolean *update functions*. For every $u_i \in U$ we have $u_i : \{0,1\}^n \to \{0,1\}$ associated with variable v_i. A *state* of the system is a map $s : V \to \{0,1\}$. We say that an update function u_i is *enabled* at state s if $u_i(s) \neq s(v_i)$, i.e. applying the update function u_i to state s changes the value of variable v_i.

State $s' = (d'_1, d'_2, \ldots, d'_n)$ is a *successor* of state $s = (d_1, d_2, \ldots, d_i, \ldots, d_n)$ if for some i we have that u_i is enabled, $d'_i = u_i(s)$, and for all $j \neq i$ we have $d'_j = d_j$. That is, we get to the next state s', by non-deterministically selecting an enabled update function u_i and updating the value of the associated variable: $s' = (d_1, d_2, \ldots, u_i(d_i), \ldots, d_n)$. If no update function is enabled, $s' = s$.

An ABN induces a labelled transition system $T = (N,R)$, where N is the set of 2^n states of the ABN, and $R \subseteq N \times V \times N$ is the successor relation. Each transition (s_1, v_i, s_2) is labelled with the variable v_i such that $s_1(v_i) \neq s_2(v_i)$.

The *undirected state space* of an ABN is an undirected graph $S = (N,E)$, where each vertex $n \in N$ is uniquely labelled with a state s of the Boolean network, and there is an edge $\{s_1, s_2\} \in E$ iff s_1 and s_2 differ in the value of exactly one variable, v. The edge $\{s_1, s_2\}$ is labelled with v. In general, an undirected state space does not have to include all 2^n states induced by a Boolean network.

An ABN $B(V,U)$ *induces* a *directed state space* on an undirected state space $S = (N,E)$.

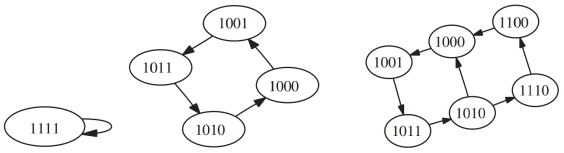


Figure 2.1 Attractors in an asynchronous Boolean network: a stable state, and two different types of loop.

Consider the transition system $T = (2^V, R)$ of $B(U,V)$. Then, the induced directed state space is $S' = (N,A)$, where $(s_1,s_2) \in A$ implies that there is a variable v_i such that $(s_1,v_i,s_2) \in R$. We say that (s_1,s_2) is *compatible* with u_i, if $s_2(v_i) = u_s(s_1)$, and for every $j \neq i$ we have $s_2(v_j) = s_1(v_j)$.

2.2 Attractors

Since a Boolean network has a finite number of states, any execution eventually converges to either a single stable state or a cycle of states, called an *attractor* (Figure 2.1). Formally, an attractor is a set of states S such that for all $s \in S$, we have that $F(s,T) = S$, where $F(s,T)$ is the set of states reachable from s in the transition system of the ABN, T.

BDD and SAT-based algorithms have been introduced for identifying the attractors of Boolean networks. Finding a stable state s of a Boolean network is easily encoded as a SAT problem: $(u_1(s) \leftrightarrow s(v_1)) \wedge \cdots \wedge (u_n(s) \leftrightarrow s(v_n))$. To find more complex attractors we need to identify cycles in the transition relation of the Boolean network.

For identifying attractors in asynchronous Boolean networks, Garg et al. introduced a BDD-based algorithm, shown in Figure 2.2 (Garg et al. (2008)). This algorithm works by manipulating a BDD representing the transition relation of the ABN and BDDs representing sets of states. Starting from an arbitrary initial state, the algorithm explores all states which are reachable from this state, and all states which can reach this state, by iteratively applying the transition relation forwards and backwards. An attractor has been found if the forward reachable states are contained in the backward reachable states. The explored states are removed from the state space and the process is repeated from another arbitrary state, until the entire state space has been explored. This algorithm was subsequently improved by Zheng et al. (Zheng et al. (2013)).

```
function ALLATTRACTORS(transitionBDD)
    terminalStates ← ∅
    while transitionBDD ≠ false do
        s ← RANDOMINITIALSTATE(transitionBDD)
        fr ← FORWARDREACHABLESTATES(transitionBDD, s)
        br ← BACKWARDREACHABLESTATES(transitionBDD, s)
        if fr ∧¬ br = false then
            terminalStates ← {fr}∪ terminalStates
        end if
        transitionBDD ← transitionBDD ∧¬ (s ∨ br)
    end while
    return terminalStates
end function
```

Figure 2.2 Attractor finding algorithm from Garg et al. (2008).

Chapter 3

Proposed algorithm

3.1 Introduction

Part of this chapter was published as Fisher et al. (2015).

Here I address the problem of automatically constructing such Boolean network models directly from data. If we think of single-cell gene expression profiles as the state space of an asynchronous Boolean network, can we identify the underlying gene regulatory logic that could have generated this data?

I encode the matching of an asynchronous Boolean network to a state space as a synthesis problem and use constraint (satisfiability) solving techniques for answering the synthesis problem. The synthesised network has to match the data in two aspects. First, the resulting network should try to minimise transitions to expression points that are not part of the sampled data. Second, the resulting network should allow for a progression through the state space in a way that matches the flow of time through the different experiments that produced the data. A direct encoding of this problem into a satisfiability problem does not scale well. I suggest a modular search that handles parts of the state space and the network and does not need to reason about the entire network at once.

In this thesis I consider two test cases. First, I try to reconstruct an existing asynchronous Boolean network from its state space. I am able to reconstruct Boolean rules from the original network. Second, in chapter 4 I apply this technique to experimental data derived from a single-cell resolution study of embryonic blood cell development. The network that is produced by my technique matches known dependencies and suggests interesting novel predictions. Some of these predictions were validated experimentally by collabora-

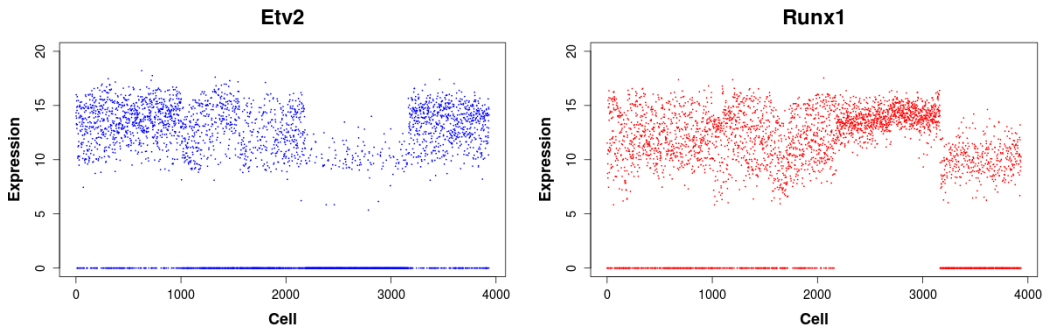


Figure 3.1 Single-cell gene expression measurements for two genes, in 3934 cells. Points at zero indicate expression of the gene was below detection level in that cell. Points above zero indicate the level of expression that was detected.

tors. Chapter 4 will discuss how the method was used to analyse this biological data set. In this chapter, I will focus on the algorithmic aspects of the method. However, in the next section I will briefly introduce the experimental data set which motivated the development of this algorithm.

3.2 Viewing single-cell gene expression data as the state space of a Boolean network

Single-cell gene expression experiments produce gene expression profiles for individually measured cells. Each of these gene expression profiles is a vector where each element gives the level of expression of one gene in that cell. Figure 3.1 plots the level of the genes *Etv2* and *Runx1* over 3934 cells.

Experimental collaborators performed such gene expression profiling on five batches of cells taken from four sequential developmental time points of a mouse embryo. For each time point, the experiment aimed to capture every cell with the potential to develop into a blood cell, providing a comprehensive single–cell resolution picture of the developmental timecourse of blood development. This resulted in a data set of 3934 cell measurements. Full details of this experiment and our analysis can be found in chapter 4. This data set is the first of its kind, attempting to capture an entire tissue's worth of progenitor cells across a developmental time course. This level of coverage of the potential cell state space is required for our approach to accurately recover gene regulatory networks, and typically requires the measurement of thousands of cell profiles. Later I will introduce a synthetic data set of a

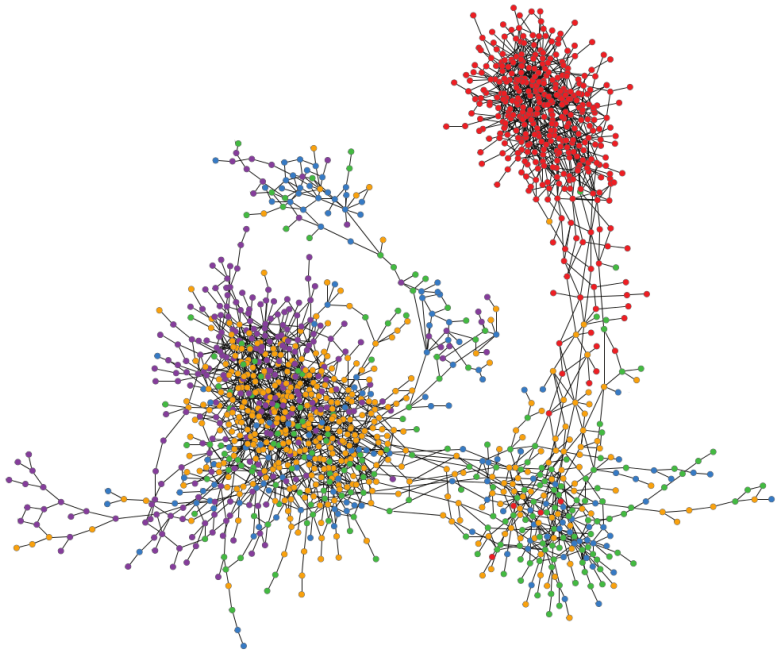


Figure 3.2 State graph. Node colours correspond to the time point at which a state was measured. States from the earliest of the time points are coloured blue, and blood progenitor states from the last time point are coloured red.

few hundred cell states in order to illustrate how our approach works, but I would like to stress that to be usable on real experimental data our algorithm needs to be able to scale thousands of cell states.

For each of 3934 cells, the level of expression of 33 transcription factor genes was measured. Expression levels are non-negative real numbers, where the value 0 indicates that the given gene is unexpressed in the cell (see Figure 3.1).

The key idea introduced in this thesis is to view this gene-expression data as a sample from the state-space of an asynchronous Boolean network. In the past, manually curated Boolean networks have been successfully used to recapitulate experimental results (Bonzanni et al. (2013); Kazemzadeh et al. (2012); Krumsiek et al. (2011)). Such Boolean networks were hand–constructed from biological knowledge that has accumulated in the literature over many years. Here, I aim to produce such Boolean networks automatically, directly from gene expression data, by employing synthesis techniques. I aim to produce a Boolean network that can explain the data and can be used to inform biological experiments for uncovering the nature of gene regulatory networks in real biological systems.

In order to convert the data into a format that can be viewed as a Boolean network state space, I first discretise expression values to binary, assigning the value 1 to all non-zero gene expression measurements. A value of zero corresponds to the discovery threshold of the equipment used to produce the data. Discretising the 3934 expression profiles in this way yields 3070 unique binary states, where every state is a vector of 33 Boolean values corresponding to the activation/inactivation level of each of 33 genes in a given cell. In an asynchronous Boolean network, transitions correspond to the change of value of a single variable. Hence, I next look for pairs of states that differ by only one gene (that is, the Hamming distance between the two vectors is 1). An analysis of the connected components of this graph shows that one connected component contains 44% of the states. Note that in a random sample of 3934 elements from a space of 2^{33}, the chance of seeing repeats or neighbours with Hamming–distance 1 is negligible.

To efficiently construct this state transition graph $G = (N, E)$ on variables $V = \{v_1, v_2, \ldots, v_n\}$ with $n \leq 64$ and states S where each state $s \in S$ is a map $s : V \to \{0, 1\}$ and each node $n \in N$ is uniquely labelled with a state $s \in S$, we can represent each state $s \in S$ in bitvector form as a 64 bit integer, and then flip each bit in turn to generate each of the possible n neighbours $\{s_1, s_2, \ldots, s_n\}$. If $s_i \in S$ then we create an edge $\{s, s_i\}$.

A plot of the graph of the largest connected component is given in Figure 3.2. We add an edge for every Hamming–distance 1 pair and cluster together highly connected nodes. The colours of nodes correspond to the developmental time the measurements was taken. Note that there is a clear separation between the earliest developmental time point and the latest one. This representation already suggests a clear change of states over the development of the embryo, with separate clusters identifiable and obvious fate transitions between clusters.

We wish to find an asynchronous Boolean network that matches this graph. For that we impose several restrictions on the Boolean network. Connections between states correspond to a change in the value of one gene, however, we do not know the direction of the change. Thus, we search simultaneously for directions and update functions of the different genes that satisfy the following two conditions: states from the earliest developmental time point should be able to evolve, through a series of single–gene transitions, to the states from the latest developmental time point. Secondly, the update functions must minimise the number of transitions that lead to additional, unobserved states, that were not measured in the experiment.

Gene	Update function
Gata2	$Gata2 \wedge \neg(Pu.1 \vee (Gata1 \wedge Fog1))$
Gata1	$(Gata1 \vee Gata2 \vee Fli1) \wedge \neg Pu.1$
Fog1	$Gata1$
EKLF	$Gata1 \wedge \neg Fli1$
Fli1	$Gata1 \wedge \neg EKLF$
Scl	$Gata1 \wedge \neg Pu.1$
Cebpa	$Cebpa \wedge \neg(Scl \vee (Fog1 \wedge Gata1))$
Pu.1	$(Cebpa \vee Pu.1) \wedge \neg(Gata1 \vee Gata2)$
cJun	$Pu.1 \wedge \neg Gfi1$
EgrNab	$(Pu.1 \wedge cJun) \wedge \neg Gfi1$
Gfi1	$Cebpa \wedge \neg EgrNab$

Figure 3.3 Boolean update functions for a manually curated network.

3.3 Example: reconstructing an ABN from its state space

I first illustrate my synthesis method using an example. I take an existing Boolean network, construct its associated state space, and then use this state space as input to my synthesis method in order to try to reconstruct the Boolean network that we started with.

Krumsiek *et. al.* introduce a Boolean network model of the core regulatory network active in common myeloid progenitor cells (Krumsiek et al. (2011)). Their network is based upon a comprehensive literature survey. It includes a set of 11 Boolean variables (corresponding to genes) and a Boolean update function for each variable (Figure 3.3).[1] The model is given a well-defined initial starting state, representing the expression profile of the common myeloid progenitor, and computational analysis reveals an acyclic, hierarchical state space of 214 states with four stable state attractors (Figure 3.4).

These stable attractors are in agreement with experimental expression profiles of megakaryocytes, erythrocytes, granulocytes and monocytes; four of the mature myeloid cell types that develop from common myeloid progenitors.

We treat the state space of this Boolean network as we would treat experimental data, forgetting all directionality information, and connecting all states which differ in the expression of only one gene by an undirected edge (Figures 3.4 and 3.5, where each edge is labelled with the single gene that changes in value between the states it connects). We would now like to reconstruct the Boolean network given in Figure 3.3 from this undirected state space.

For each gene, we would like to assign a direction to each of its labelled edges (or decide that it does not exist), in a way that is compatible with a Boolean update function. For example, in Figure 3.5, we may orient the *Pu.1*-labelled edge between states 97 and 95 in the direction $s_{97} \rightarrow s_{95}$, in the direction $s_{95} \rightarrow s_{97}$, or decide that this is not a possible update.

[1] The function of *Cebpa* is modified from that in Krumsiek et al. (2011) to match the format I assume.

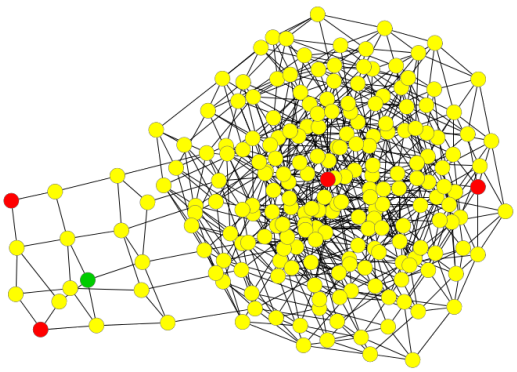


Figure 3.4 Boolean network state space. Initial state is coloured green, stable states red.

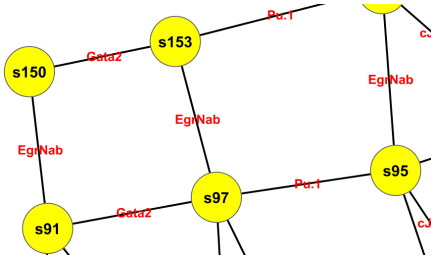


Figure 3.5 Close-up of Boolean network state space.

We also allow the edge to be directed in both directions. If $s_{97} \rightarrow s_{95}$, we want a Boolean update function $u_{Pu.1}$ that takes state s_{97} to state s_{95}. Since there is no *Pu.1*–labelled edge leaving state s_{150}, we can also add the constraint that $u_{Pu.1}$ takes s_{150} to s_{150}.

We also add reachability constraints that restrict which edges are included and their orientation. Since the state space was constructed starting from a well-defined initial state, we would like to enforce the constraint that each non-initial state ought to be reachable by some directed path from the initial state. Since cell development proceeds hierarchically and unidirectionally, I favour short paths over long paths. This eliminates routes that seem biologically implausible, for example routes that cross a fate transition and then return to where they began. It also reduces the space of paths we have to search through. By increasing the lengths of allowed paths, we can increase the number of considered solutions.

The results of applying my technique are shown in Figure 3.6. The method reconstructs the Boolean update functions for all but one gene (*EgrNab*), in some cases uniquely identifying the original function. I note that when multiple solutions are found for an update function, these solutions, while not exact, all provide useful regulatory information that could be verified experimentally. For example, both solutions for *Scl* successfully predict *Scl*'s activation by *Gata1*, although one of the two solutions omits its repression by *Pu.1*.

Gene	Synthesised update functions	Comments
Gata2	$Gata2 \wedge \neg(Fog1 \vee Pu.1)$ $Gata2 \wedge \neg(Fog1 \vee (Pu.1 \wedge Cebpa))$ $Gata2 \wedge \neg(Fog1 \vee (Pu.1 \wedge Gata2))$ $Gata2 \wedge \neg(Gata2 \wedge (Pu.1 \vee Fog1)$ $\boldsymbol{Gata2} \wedge \neg(\boldsymbol{Pu.1} \vee (\boldsymbol{Gata1} \wedge \boldsymbol{Fog1}))$ $Gata2 \wedge \neg(Pu.1 \vee (Gata2 \wedge Fog1))$	
Gata1	$(Gata1 \vee Cebpa) \wedge \neg Pu.1$ $(Gata2 \vee Fog1) \wedge \neg Pu.1$ $(Gata1 \vee Gata2) \wedge \neg Pu.1$ $(\boldsymbol{Gata1} \vee \boldsymbol{Gata2} \vee \boldsymbol{Fli1}) \wedge \neg \boldsymbol{Pu.1}$ Other functions of the form $(X \vee Y \vee Z) \wedge \neg Pu.1$	
Fog1	$\boldsymbol{Gata1}$	Unique
EKLF	$\boldsymbol{Gata1} \wedge \neg \boldsymbol{Fli1}$	Unique
Fli1	$\boldsymbol{Gata1} \wedge \neg \boldsymbol{EKLF}$	Unique
Scl	$Gata1$ $\boldsymbol{Gata1} \wedge \neg \boldsymbol{Pu.1}$	
Cebpa	$Cebpa \wedge \neg(Fog1 \vee Scl)$ $Cebpa \wedge \neg(Cebpa \wedge (Scl \vee Fog1))$ $Cebpa \wedge \neg(Fog1 \wedge (Scl \vee Cebpa))$ $Cebpa \wedge \neg(Fog1 \vee (Scl \wedge Gata1))$ $Cebpa \wedge \neg(Fog1 \vee (Scl \wedge Gata2))$ $Cebpa \wedge \neg(Gata1 \wedge (Fog1 \vee Scl)$ $Cebpa \wedge \neg(Scl \vee (Fog1 \wedge Cebpa)$ $\boldsymbol{Cebpa} \wedge \neg(\boldsymbol{Scl} \vee (\boldsymbol{Fog1} \wedge \boldsymbol{Gata1})$	
Pu.1	$Pu.1 \wedge \neg Gata2$ $(Pu.1 \wedge Cebpa) \wedge \neg Gata2$ $Pu.1 \wedge \neg(Gata1 \vee Gata2)$ Other functions of the form $Pu.1 \wedge \neg(Gata2 \vee X)$ $Pu.1 \wedge \neg(Gata2 \wedge Cepba)$ $Pu.1 \wedge \neg(Gata2 \wedge Pu.1)$ $Cebpa \wedge \neg(Gata1 \vee Gata2)$ $Cebpa \wedge \neg(Gata2 \vee Fog1)$ $(\boldsymbol{Cebpa} \vee \boldsymbol{Pu.1}) \wedge \neg(\boldsymbol{Gata1} \vee \boldsymbol{Gata2})$ $(Cebpa \wedge Pu.1) \wedge \neg(Gata1 \vee Gata2)$ Other functions of the form $(Cebpa \vee X) \wedge \neg(Gata2 \vee Y)$ Other functions of the form $(Pu.1 \vee X) \wedge \neg(Gata2 \vee Y)$ Other functions of the form $(Cebpa \wedge Pu.1) \wedge \neg(Gata2 \vee X)$	
cJun	$\boldsymbol{Pu.1} \wedge \neg \boldsymbol{Gfi1}$	Unique
EgrNab	$(cJun \vee Gata1) \wedge \neg Gfi1$	Incorrect with shortest paths
Gfi1	$\boldsymbol{Cebpa} \wedge \neg \boldsymbol{EgrNab}$	Unique

Figure 3.6 Synthesised update functions.

3.4 Formal definition of the problem

Our synthesis problem can be stated as follows: we are given an undirected state space S over a given set of variables V. We would like to extract a set of Boolean update functions that induce a directed state space from S such that each of the states in S are reachable from a given set of initial states. We also want to ensure that no additional, undesired states not in S are reachable, by ruling out transitions which 'exit' the state space.

More formally, we are given a set of variables $V = \{v_1, v_2, \ldots, v_n\}$, an undirected state space $S = (N, E)$ over V, and a set $I \subseteq N$ of *initial* vertices.

We would like to find an update function $u_i : \{0,1\}^n \to \{0,1\}$ for each variable $v_i \in V$, such

that the following conditions hold. Let $U = \{u_i \mid v_i \in V\}$ be the set of update functions.

1. Every non-initial vertex $s \in N - I$ is reachable from some initial vertex $s_i \in I$ by a directed path in the directed state space induced by $B(V,U)$ on S.
2. For every variable $v_i \in V$, let N_i be the set of states without an outgoing v_i-labelled arc. For every i we require that for each $s \in N_i$, $u_i(s) = s(v_i)$.

3.4.1 Generalising the definition to partial data

Since I intend to apply this method in an experimental setting, where we only have an incomplete sample from the possible states of the system, I relax this definition to extend it to partial data. Instead of requiring that *every* state is reachable from those initial states that we have measured, we only require that a set of *final* states are reachable. Instead of requiring that every undesired transition is ruled out, we seek to maximise the number of such transitions which are eliminated. This is formally stated next.

As before, we are given a set of variables $V = \{v_1, v_2, \ldots, v_n\}$, an undirected state space $S = (N,E)$ over V, and a designated set $I \subseteq N$ of *initial* vertices. In addition, we are given a designated set $F \subseteq N$ of *final* vertices, along with a *threshold* t_i for each variable $v_i \in V$. The threshold t_i specifies how many undesired transitions must be ruled out.

We would like to find an update function $u_i : \{0,1\}^n \to \{0,1\}$ for each variable $v_i \in V$, such that the following conditions hold. Let $U = \{u_i \mid v_i \in V\}$ be the set of update functions.

1. Every final vertex $s_f \in F$ is reachable from some initial vertex $s_i \in I$ by a directed path in the directed state space induced by $B(V,U)$ on S.
2. For every variable $v_i \in V$, let N_i be the set of states without an outgoing v_i-labelled arc. For every i the number of states $s \in N_i$ such that $u_i(s) = s(v_i)$ is greater or equal to t_i.

In the remainder of the text, I refer to condition 1 as the *reachability condition* and condition 2 as the *threshold condition*.

I restrict the search to update functions of the form $f_1 \wedge \neg f_2$, where f_i is a monotone Boolean formula (contains $\wedge$ and $\vee$ gates, but no negation). The variables of f_1 are *activators* of f and the variables of f_2 are *repressors*. This restriction was chosen so that repression dominates activation, and was made after discussion with biologist colleagues and consultation of the literature (e.g., Bonzanni et al. (2013); Krumsiek et al. (2011)).

3.5 A direct encoding

We start with a direct encoding of the search for a matching Boolean network. The search is parameterised by the shape of update functions (how many activators and how many repressors each function has), the length of paths from initial states to final states, and the thresholds for each variable. By increasing the first two parameters and decreasing the last we can explore all possible Boolean networks.

3.5.1 Possible update functions

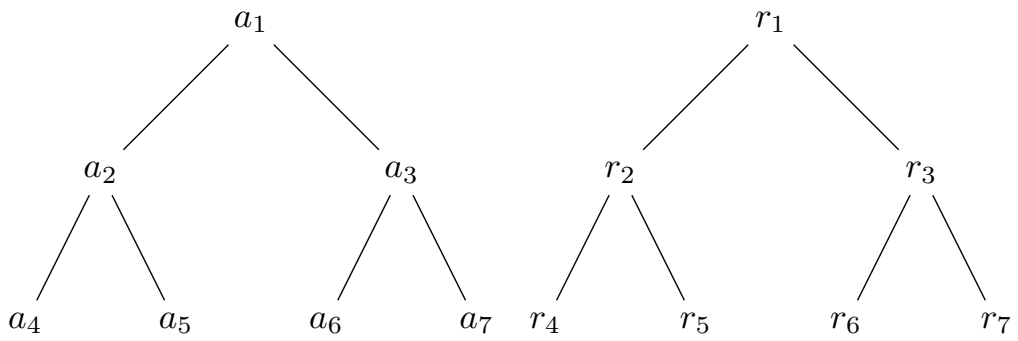


Figure 3.7 Boolean formulae $f = f_1 \wedge \neg f_2$ are represented as a pair of binary trees in the SAT encoding. Each bitvector a_i represents a variable or gate in f_1 and each bitvector r_i represents a variable or gate in f_2.

In order to represent the Boolean update function for gene v_i, $u_i = f_1 \wedge \neg f_2$, we use a bitvector encoding. We represent a monotone Boolean formula f_j of up to 4 inputs as a depth 2 binary tree encoded by a set of bitvectors, $\{a_1, a_2, \ldots, a_7\}$ (see Figure 3.7). Each bitvector $a_i \in \{a_1, a_2, a_3\}$ represents a Boolean operator: AND, OR, or LEFT, and each bitvector $a_i \in \{a_4, \ldots, a_7\}$ represents a variable $v_i \in V$. The LEFT operator returns its left argument.

For example, the formula $v_1 \wedge (v_2 \vee v_3)$ is represented by the solution $a_1 = \mathsf{AND}, a_2 = \mathsf{OR}, a_3 = v_1, a_4 = v_2, a_5 = v_3$. The formula $v_1 \vee v_2$ is represented by the solution $a_1 = \mathsf{LEFT}, a_2 = \mathsf{OR}, a_3 = v_1, a_4 = v_2$.

We search for functions up to a maximum number of activators, A_i, and a maximum number of repressors, R_i. For example, to search for functions with only one activator, we add constraints to fix $a_1 = a_2 = a_3 = \mathsf{LEFT}$. In order to allow the function to have zero repressors, we allow r_1 to also take the value NOTHING.

To encode the application of function u_i to a state s, $u_i(s)$, we add implications which unwrap the bitvector encoding of u_i to the constituent variables and logical operators; substituting values, $s(v_j)$, for variables, v_j, and directly mapping operators to logical constraints in the Boolean satisfiability formula:

$$\bigwedge_{j=1}^{3} \left(a_j = \mathsf{AND} \rightarrow i_{a_j} \leftrightarrow i_{a_{l(j)}} \wedge i_{a_{r(j)}} \right) \bigwedge$$

$$\bigwedge_{j=1}^{3} \left(a_j = \mathsf{OR} \rightarrow i_{a_j} \leftrightarrow i_{a_{l(j)}} \vee i_{a_{r(j)}} \right) \bigwedge$$

$$\bigwedge_{j=1}^{3} \left(a_j = \mathsf{LEFT} \rightarrow i_{a_j} \leftrightarrow i_{a_{l(j)}} \right) \bigwedge$$

$$\bigwedge_{j=4}^{7} \left(a_j = v_k \rightarrow i_{a_j} \leftrightarrow s(v_k) \right)$$

where each i_{a_j} is an intermediate variable that stores the result of intermediary computations, and $l(i) = i \times 2$ and $r(i) = l(i) + 1$.

The value of the application $u_i(s)$ is given by

$$\left(r_1 = \mathsf{NOTHING} \rightarrow i_{a_1} \right) \bigwedge$$

$$\left(r_1 \neq \mathsf{NOTHING} \rightarrow i_{a_1} \wedge \neg i_{r_1} \right)$$

For example, the application of the function $(v_1 \vee v_2) \wedge \neg v_3$ to the state s_1 is mapped to $(s_1(v_1) \vee s_1(v_2)) \wedge \neg s_1(v_3)$.

3.5.2 Ensuring reachability

To enforce the global reachability condition we consider all of the underlying directed edges in the undirected state space $S = (N, E)$, and their associated single–gene transitions.

Recall that we require every final vertex to be reachable from some initial vertex by a directed path in the directed state space induced on S by the Boolean network. That is, we require that every final vertex is reachable by a directed path, and that every v_j-labelled edge along this path is compatible with its associated update function, u_j.

To enforce this we add constraints that track the compatibility of edges with update functions and define reachability recursively. We consider reachability by paths up to a maximum length: recall that we consider shorter paths to be more biologically likely. By iteratively increasing the length of the paths considered, we can obtain all satisfying models.

We introduce a pair of Boolean variables e_{ij}, e_{ji} for each v_i-labelled undirected edge $\{s_i, s_j\} \in E$, which track the value of the application of u_i to s_i and to s_j (and the compatibility of the underlying directed edges (s_i, s_j) and (s_j, s_i) with u_i). e_{ij} is true iff $u_i(s_i) = s_j(v)$.

We introduce an integer given by a bitvector encoding, r_n, for each node $n \in N$. Bitvector r_n encodes the fact that node n is reachable from an initial node in r_n steps, up to some maximum encodable value $2^{|r_n|} - 1$. Bitvector r_n is given a value of -1 to indicate that n is not reachable in this maximum number of steps.

Reachability is then defined inductively:

1. Initial nodes are reachable in zero steps: for every $i \in I$, $r_i = 0$.
2. A non–initial node s_i is reachable in M steps if there is a compatible incoming edge (s_j, s_i) from another node s_j, and s_j is itself reachable in fewer than M steps. That is, for every $n = s_j \in N - I$ and $m = s_i \in N$ such that $\{s_i, s_j\} \in E$ we have $e_{ij} \rightarrow r_m < r_n$. We also have that non–initial nodes cannot be reached in zero steps: For every $n \in N - I$, $r_n = -1 \vee r_n > 0$.

Finally, we add a constraint that every final node $n \in F$ is reachable from some initial node: $r_n \neq -1$.

3.5.3 Enforcing the threshold condition

We enforce the threshold condition for each update function as follows.

Consider an update function $u_i : V \to \{0,1\}$. We say that a node $s \in N_i$ is *negatively matched* by u_i if $u_i(s) = s(v_i)$. That is, by using u_i as the update function of variable v_i, u_i does not change the value of v_i from node s. We are searching for an update function such that a maximum number of nodes from N_i are negatively matched.

We add a variable, m_{is} for each node $s \in N_i$ to record whether u_i negatively matches s. We then add a constraint demanding that the number of negatively matched nodes is greater than or equal to the threshold: $\sum_{s \in N_i} m_{is} \geq t_i$.

We search for satisfying assignments to the constraint variables encoding the representation of the Boolean update functions u_i for all v_i in V. The resulting synthesised Boolean network is the combination of these update functions.

Unfortunately, in practice the direct encoding of the search does not scale to handle our experimental data (see Figure 3.8). In the next section I suggest a compositional way to solve the problem.

3.6 A compositional algorithm

I now introduce my compositional algorithm, which scales better than the direct encoding given above. The problem of synthesising a Boolean network from the data is partitioned to three stages. Crucially, we avoid searching for a complete Boolean network and consider parts of the network that can be constructed independently.

3.6.1 Pruning the set of possible edges and possible update functions

We start by building a directed graph from the given undirected state space $S = (N, E)$, by considering which of the underlying directed edges in E are compatible with some Boolean update function, and pruning those that are not. We consider each underlying directed edge (s_1, s_2) and (s_2, s_1) of each of the v_i-labelled undirected edges $\{s_1, s_2\}$ in E independently. At the same time, we prune any Boolean update function which is not compatible with the threshold condition (condition 2, 3.4.1).

To do this, we maintain a set of *candidate functions*, C, and a set of *unprocessed edges*, U. Initially C is empty and U contains every underlying directed edge (s_1, s_2) and (s_2, s_1) of each of the v_i-labelled undirected edges $\{s_1, s_2\}$ in E. We lazily construct an explicit representation of each possible Boolean update function u_i and evaluate it at each state

without an outgoing v_i-labelled arc to check whether it passes the threshold condition. If u_i passes, we add it to C and then we evaluate it at each directed edge $(s_1, s_2) \in U$. If $u_i(s_1) = s_2(v_i)$ we remove (s_1, s_2) from U.

When this phase terminates, we are left with a directed graph, where there is an edge (s_1, s_2) if there exists a compatible update function for that edge. We have eliminated edges which have no compatible update function, and cannot participate in the reachability condition. We are also left with a set of candidate update functions for each variable. On the experimental data set from Chapter 4, this phase prunes up to 50% of the possible edges for a gene, and can prune over 99% of the possible update functions.

3.6.2 Ensuring reachability

We now come to the only part of the algorithm that considers the edges of all variables together, in order to enforce the global reachability condition (condition 1, 3.4.1).

We construct, for each pair of initial nodes $i \in I$ and final nodes $f \in F$, the shortest path p_{if} from i to f in the directed graph that was built in the previous phase of the algorithm. These paths can be computed via a breadth–first search. The longest such path on the example data set from Section 3.3 has length 10. The longest such path on the experimental data set from Chapter 4 has length 32.

Due to the edge pruning of the previous phase of the algorithm, if there is no path to a final node f, this implies that there are no satisfying models (at the given threshold and function size parameters). Otherwise, our reachability condition will be enforced by fixing a set of directed edges P_i for each variable $v_i \in V$ corresponding to these shortest paths. We will then require that the update function we search for, u_i, is compatible with each of the edges in P_i.

We choose, for each final node f, one path $p_f = p_{if}$ from one of the initial nodes i. By fixing this path, we ensure that f is reachable from an initial node. We define $p_f|_i$ as the set of v_i-labelled edges in the path p_f. We define P_i, the v_i-labelled edges which must be fixed to ensure reachability via the chosen paths, as the the set of v_i-labelled edges in p_f for each final node f:

$$P_i = \bigcup_{f \in F} \{(s_1, s_2) \mid (s_1, s_2) \in p_f|_i\}$$

By considering only the edges in P_i, we can search for an update function for v_i independently of all other variables, while ensuring the global reachability condition holds.

3.6.3 Final update functions

We can now search for the update function of variable v_i, u_i, independently of all other variables. We fix the v_i-labelled edges computed in the previous phase and encode the search for u_i as a Boolean satisfiability problem.

As before we add constraints to encode the representation of u_i. We fix each of the v_i-labelled edges $(s_1, s_2) \in P_i$ to establish reachability, by adding a conjunction requiring that u_i is compatible with each of them: $u_i(s_1) = s_2(v_i)$. We also add constraints that fix u_i as one of the candidate functions left over after the pruning phase of Section 3.6.1. Importantly, this means that we no longer have to enforce the threshold condition.

We search for satisfying assignments of the constraint variables encoding u_i, using an ALL-SAT procedure to extract all possible update functions for variable v_i. This gives rise to a set of update functions per variable and a set of Boolean networks from the product of the set of update functions per variable.

We note that this final phase of the algorithm can fail to find update functions for a variable v_i, because there are no possible update functions compatible with all of the path edges P_i that were computed in the previous phase. That is, while each edge in P_i is individually compatible with some update function, there may be no update function that is compatible with every edge in P_i. In order to cope with this limitation, we can extract the minimal unsatisifiable core of the Boolean formula, and search for replacement paths that exclude incompatible combinations of edges. This step can be iterated until satisfying solutions are found for all variables, or until no path can be found, implying that there are no valid models.

By extending our search from the shortest paths between initial and final node pairs in the directed graph to the k-shortest paths between pairs and incrememtally increasing k (Yen (1971)), we can increase the number of possible update functions that we consider. In the limit, we will obtain all satisfying models.

An implementation of my algorithm, which is written in F# and uses Z3 as the satisfiability solver, is available at https://github.com/swoodhouse/SCNS-Toolkit. In Figure 3.8 I present experimental results from running my implementation of the direct encoding from Section 3.5 and compositional algorithm on four data sets: the small synthetic data set from

Data set	Genes	States	Direct (seconds)	Compositional (seconds)
CMP (synthetic)	11	214	25	10
Blood stem cells	22	613	OUT OF MEMORY	191
Embryonic (66%)	33	956	OUT OF MEMORY	136
Embryonic (full)	33	1448	OUT OF MEMORY	244

Figure 3.8 Performance of direct encoding and compositional algorithm on example data sets.

Section 3.3, the large embryonic experimental data set from Section 3.2, and a second experimental data set covering blood stem cells. I also show results from rerunning on the embryonic data set with a third of states removed. All experiments were performed on an Intel Core i5 @ 1.70GHz with 8GB of RAM.

While the direct encoding synthesised a matching Boolean network on the small synthetic data set, it cannot scale to the real experimental data sets, quickly running out of memory. The compositional algorithm, on the other hand, can scale to handle real data sets of the sort produced by my experimental collaborators.

Chapter 4

Application to haematopoetic data

This chapter was published as Moignard et al. (2015), and was joint work resulting from a close collaboration with Victoria Moignard, who performed single-cell gene expression experiments. The diffusion map implementation introduced for this study was developed by Laleh Haghverdi and Florian Buettner, and is described in further detail in Haghverdi et al. (2015).

4.1 Introduction

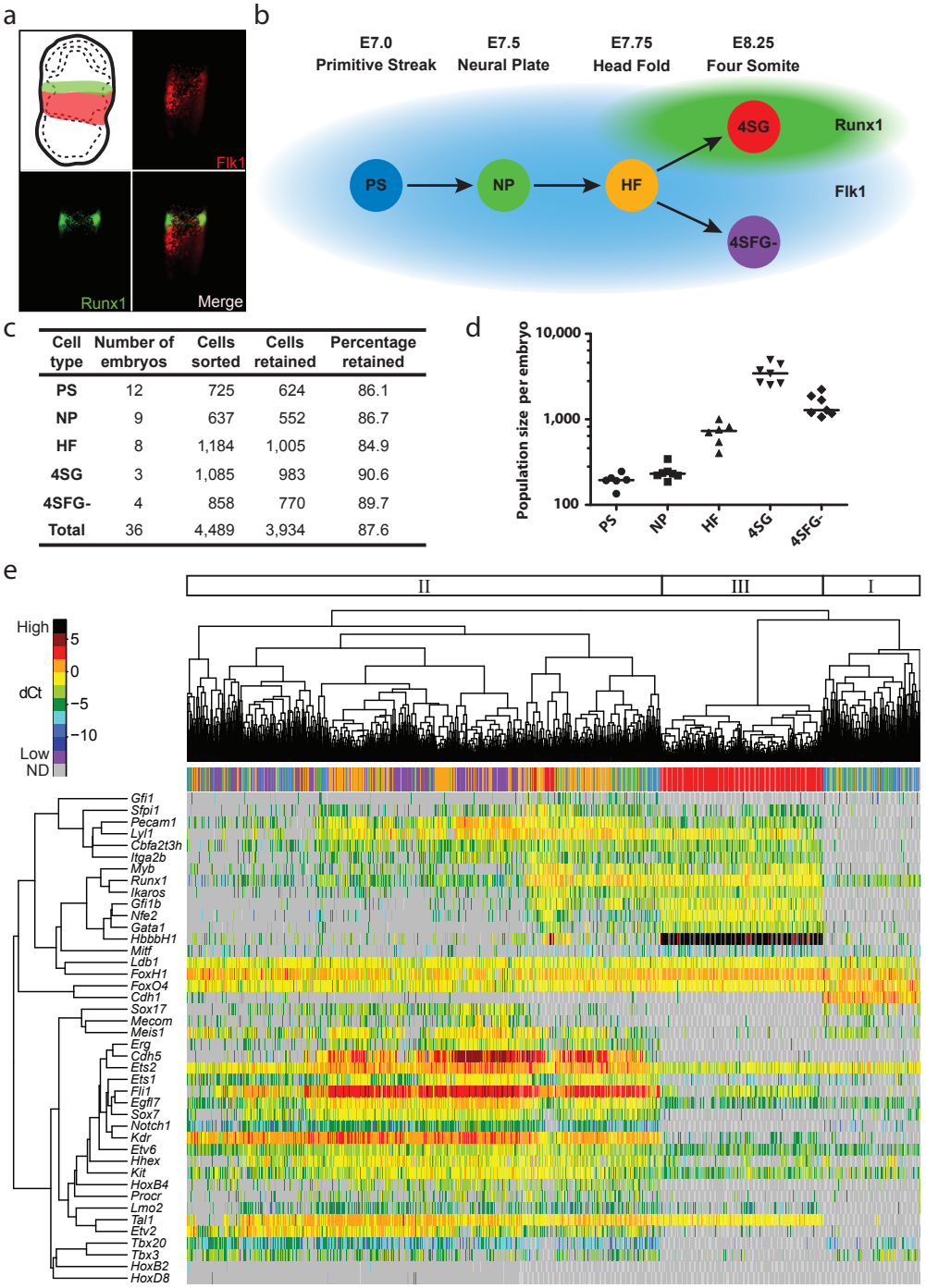
Blood has long served as a model to study organ development owing to the accessibility of blood cells and the availability of markers for specific cell populations. Blood development initiates at gastrulation from multipotent Flk1+ (encoded by *Flk1*, also known as *Kdr*) mesodermal cells, which initially have the potential to form blood, endothelium and smooth muscle cells (Shalaby et al. (1995, 1997)). Blood development represents one of the earliest stages of organogenesis, as the production of primitive erythrocytes is required to support the growing embryo. Single-cell gene expression analysis has already been successfully applied to study the earliest stages of preimplantation mouse and human development (Guo et al. (2010); Xue et al. (2013); Yan et al. (2013)), to identify lineage commitment (Pina et al. (2012)) and transcriptional regulatory (Moignard et al. (2013a)) events in blood, and, more recently, to probe the emergence of hematopoietic stem cells (HSCs) from the hemogenic endothelium of the dorsal aorta (Swiers et al. (2013a)).

Here we report *in-vivo* gene expression analysis of early blood development at the single-cell level, focusing on transcription factors as regulators of cell fate. Using qRT-PCR, we

analyzed >40 genes in 3934 cells with blood and endothelial potential from five populations at four sequential stages of post-implantation mouse development between embryonic day (E)7.0 and E8.25. We adapted the diffusion plot methodology previously reported in non-biological contexts (Coifman et al. (2005)) for dimensional reduction of single-cell data, where pseudotemporal ordering of individual cells revealed a putative developmental hierarchy branching toward both blood and endothelial-like fates. To discover the underlying regulatory network, we developed a single-cell network synthesis algorithm for synthesising executable Boolean network models from binary single-cell expression states, which correspond to the on and off patterns of transcription factor expression. Using this method we identified a core network of 20 highly connected transcription factors, which could reach eight stable states representing blood and endothelium. We validated model predictions to demonstrate that Sox7 blocks primitive erythroid development, and Sox and Hox factors directly regulate expression of the HSC regulator, *Erg*. Our algorithm therefore opens up network reconstruction for other systems without the requirement for prior knowledge of regulatory interactions.

4.2 Capturing cells with blood potential during gastrulation

The first wave of primitive hematopoiesis originates from Flk1+ mesoderm (Guo et al. (2010); Lux et al. (2008); Yan et al. (2013))), with all hematopoietic potential in the mouse contained within the Flk1+ population from E7.0 onwards. Although some blood progenitor cells lose Flk1 expression just before the onset of circulation (Ding et al. (2013)), previous work using a *LacZ* reporter knocked into the *Runx1* locus showed that hematopoietic potential remains confined to the Runx1 + fraction (Tanaka et al. (2012)), which was confirmed with a GFP reporter driven by the Runx1 +23 enhancer, which reproduces Runx1 expression (Swiers et al. (2013a)). Using Flk1 expression in combination with a Runx1-ires-GFP reporter mouse (Lorsbach et al. (2004)) therefore allowed us to capture cells with blood potential at distinct anatomical stages across a time course of mouse development (Figure 4.1a,b). Single Flk1+ cells were flow sorted at E7.0 (primitive streak, PS), E7.5 (neural plate, NP) and E7.75 (head fold, HF) stages. We subdivided E8.25 cells into putative blood and endothelial populations by isolating GFP+ cells (four somite, 4SG) and Flk1+ GFP cells (4SFG), respectively (Figure 4.1b). Cells were sorted from multiple embryos at each time point, with 3934 cells going on to subsequent analysis (Figure 4.1c). Total cell numbers and numbers of cells of appropriate phenotypes (Figure 4.1d) present in each embryo were



Cell type	Number of embryos	Cells sorted	Cells retained	Percentage retained
PS	12	725	624	86.1
NP	9	637	552	86.7
HF	8	1,184	1,005	84.9
4SG	3	1,085	983	90.6
4SFG-	4	858	770	89.7
Total	36	4,489	3,934	87.6

Figure 4.1 Single-cell gene expression analysis of early blood development. (a) Flk1 and Runx1 staining in E7.5 mesoderm and blood band, respectively. (b) Single cells sorted from five populations at four anatomically distinct stages from E7.0–E8.25. (c) Quantification of cells sorted and retained for analysis after quality control. (d) Quantification of Flk1+, GFP+ or Flk1+ GFP cells in embryos at each time point from FACS data. Line indicates median. (e) Unsupervised hierarchical clustering was performed using the Spearman correlation and complete linkage for the normalized gene expression of the 33 transcription factors and 7 marker genes in all cells. Shown is the level of expression for each gene in every cell (see key).

estimated from fluorescence-activated cell sorting (FACS) data, indicating that for the first three stages, more than one embryo equivalent of Flk1+ cells was collected.

We next quantified the expression of 33 transcription factors involved in endothelial and hematopoietic development (Moignard et al. (2013b)), nine marker genes, including the embryonic globin *Hbb-bH1* and cell surface markers such as *Cdh5* (VE-Cadherin) and *Itga2b* (CD41), as well as four reference housekeeping genes in all 3934 cells using microfluidic qRT-PCR technology (Moignard et al. (2013a)), which resulted in >150,000 quantitative expression scores.

4.3 Development of blood progenitor cells is not synchronized

Unsupervised hierarchical clustering of the 33 transcription factor and 9 marker genes across all 3934 cells revealed three major clusters (Figure 4.1e). Cluster I was small and comprised mostly PS and NP cells. It did not express blood-associated genes, but showed low expression of some endothelial genes and high expression of *Cdh1* (E-cadherin), likely representing mesodermal cells at the primitive streak (Thiery et al. (2009)). Cluster II contained the greatest number of cells and included most of the PS, NP, HF and 4SFG cells, was characterized by endothelial gene expression, and contained subclusters with elevated expression of hemogenic endothelial genes, such as *Cdh5*, or hematopoietic genes such as *Gfi1*, indicating that this cluster contains a continuum of cells maturing from mesodermal to hematopoietic and endothelial fates. Cluster III was formed by most of the E8.25 Runx1 GFP+ 4SG cells, and had robust expression of hematopoietic genes (including *Hbb-bH1*, *Gata1*, *Nfe2*, *Gfi1b*, *Ikzf1* (Ikaros) and *Myb*), and low expression of endothelial genes (*Erg*, *Sox7*, *Sox17*, *Hoxb4*, *Cdh5*). The mixing of cells from different anatomical stages by hierarchical clustering analysis therefore suggested that developmental maturation of single cells in early mesodermal cell populations is asynchronous, with cells at multiple stages expressing similar combinations of developmental regulators. This is consistent with the gradual ingression of cells through the primitive streak and lineage commitment during gastrulation.

Principal component analysis (PCA) of the expression values of all 3934 cells confirmed the large-scale mixing of cells from different anatomical stages, with only 4SG cells forming a stage-specific group (Figure 4.2a)). The PCA was retrospectively colored to show which embryo each cell belongs to (Figure 4.2b)), to determine whether this mixing is the result of developmental asynchrony within embryos or differences in maturation between different

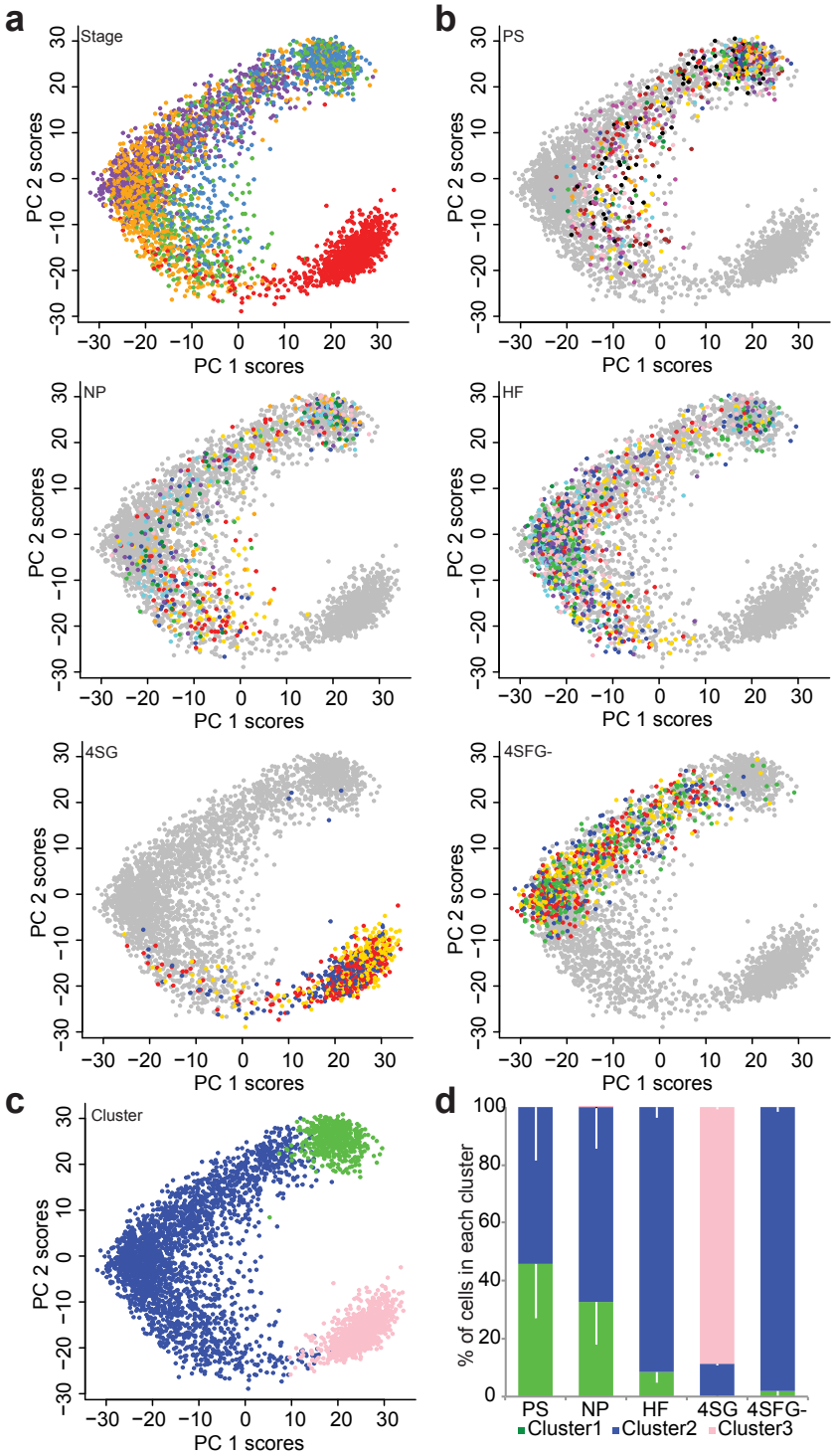


Figure 4.2 Development is asynchronous. (a) PCA of the 3934 cells, coloured retrospectively according to the stage from which they were sorted. Blue, PS; green, NP; orange, HF; red, 4SG; purple, 4SFG-. (b) For each stage, the cells from different embryos are shown on the PCA as different colours (cells from other stages shown in grey). (c) PCA coloured according to the clusters cells belong to. Green, I; Blue, II; Pink, III. (d) For each embryo, the percentage of cells in clusters I, II and III was calculated. The mean and standard deviation was then calculated for each cluster in each stage.

embryos classified as being of the same anatomical stage. We quantified the percentage of cells from each embryo belonging to clusters I, II and III, as identified by hierarchical clustering (Figure 4.1e and 4.2c,d). This showed that cells collected from each embryo at the PS, NP and HF stages were distributed across clusters I and II, with the earlier stages showing a greater bias toward cluster I than later stages. These results are therefore consistent with a model whereby cells representing both early and later stages along the differentiation trajectory toward blood are present throughout the PS, NP and HF time points, captured as snapshot measurements in our high-throughput, single-cell expression profiling.

A proportion of Flk1+ cells will give rise to mesodermal lineages other than blood and endothelium, and the extent to which they emerge over time and contribute to the variability would need to be analyzed using different gene sets. Notably, however, >50% of PS, NP and HF cells expressed both Flk1 and Runx1 at the mRNA level, highlighting the presence of Flk1+ cells with hemogenic potential (Swiers et al. (2013a); Tanaka et al. (2012)) from the earliest time points. Analysis of 50-cell pools from the PS, NP and HF stages by RNA-seq showed graded expression increases of hematopoietic and endothelial genes from the E7.0 to the E7.5 and E7.75 samples. This is entirely consistent with the continuous emergence of blood-specified cells deduced from our single-cell data, as an increase in the proportion of cells expressing a given gene between stages will increase population-averaged expression measurements. Key mesodermal and cardiac genes, by contrast, showed graded downregulation in the pooled-cell RNA-seq. These graded expression changes over time are not consistent with a discrete on or off switch at a specific developmental time point, but could again be due to gradual changes in the proportion of cells expressing the marker genes, similar to our observations from single-cell analysis of blood and endothelial genes. Alternatively, quantitative changes in expression levels within a constant proportion of cardiac-specified cells would similarly result in a change in the overall expression level of a population and cannot be excluded from the pooled-cell RNA-seq. Therefore, our results indicate, at least for cells destined to become blood and endothelium, that these cells arise at all stages of the analyzed time course rather than in a synchronized fashion at one precise time point, consistent with the gradual nature of gastrulation. Notably, only single-cell analysis over a developmental time-course has the power to reveal the contribution to cellular heterogeneity made by unsynchronized maturation of individual cells.

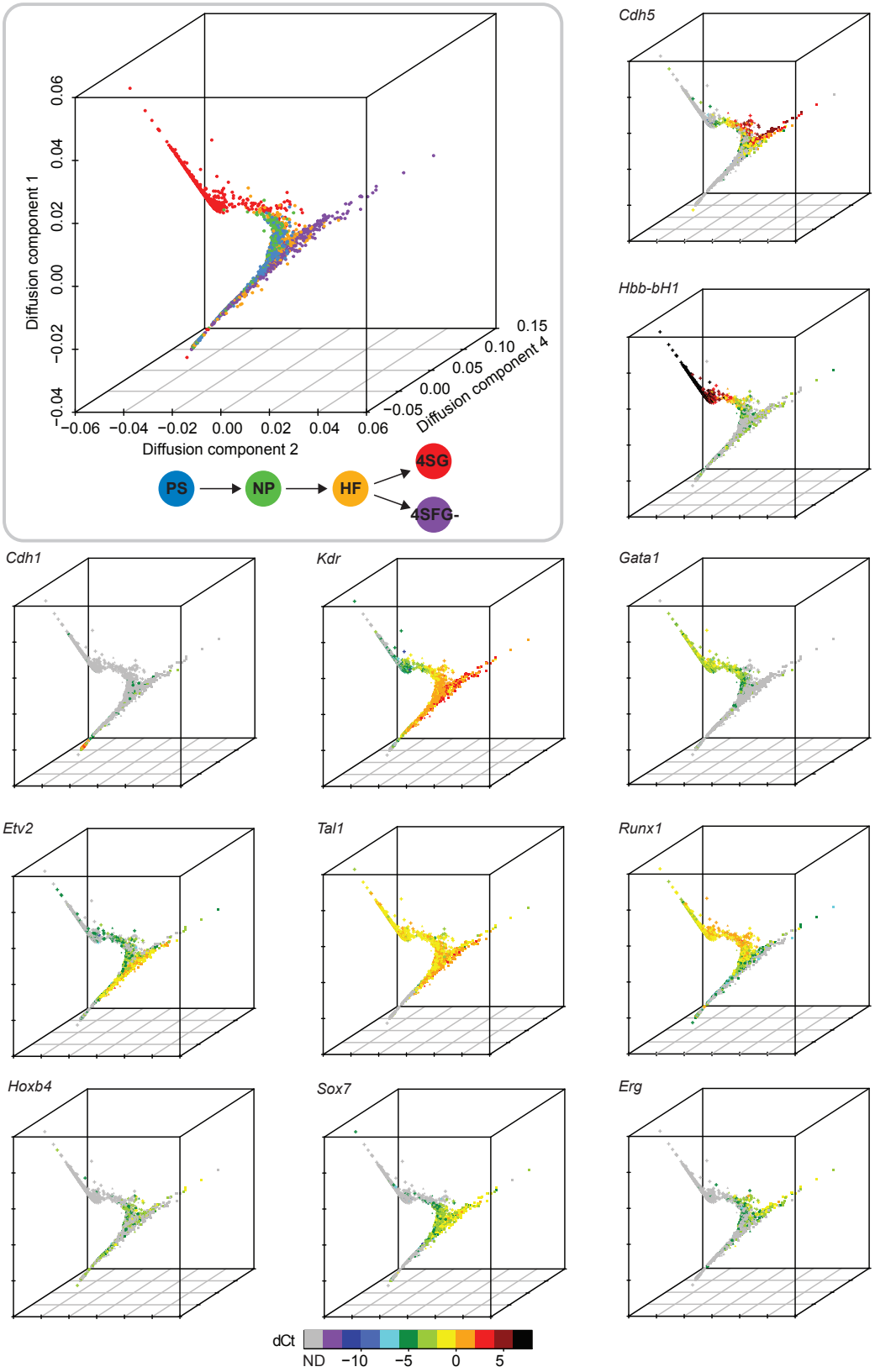


Figure 4.3 Diffusion plots identify developmental trajectories. Diffusion plot of all 3934 cells calculated from the expression of 33 transcription factors and seven marker genes (top left). The expression levels of individual genes were then overlaid onto the diffusion plot to highlight patterns of expression.

4.4 Diffusion maps identify developmental trajectories

To identify and visualize putative developmental trajectories from the PS to 4S stages in the single-cell gene expression data, we developed a computational approach for dimension reduction. Our method is based on the concept of diffusion distances, which can be interpreted as a metric for objects (here, cells) that are related to each other through a gradual but stochastic, diffusion-like process, such as cellular differentiation. In brief, similarities between all 3934 cells are calculated based on their gene expression patterns, and then visualized globally in a three-dimensional map (Figure 4.3). The resulting components span a low-dimensional diffusion-space, in which distance reflects how similar cells are in terms of their diffusion distance, and can be inferred to represent developmental time.

Although there is extensive mixing between PS, NP, HF and 4SFG populations in the diffusion plot, there is a general progression in the cell stages present in different regions of the plot from largely early E7.0 PS and E7.5 NP cells through the later HF cells to the E8.25 4SG cells that form a homogeneous cluster, in line with the expected developmental progression of the blood system or 4SFG cells. Furthermore, we observed that whereas the E8.25 Flk1+ Runx1-GFP (4SFG) cells mostly mix with earlier Flk1+ cells, a subset that was not identified by clustering or PCA branches off. This branch expresses endothelial and hemogenic endothelial genes (*Cdh5*, *Erg*, *Itga2b*, *Pecam1* (CD31), *Sox7*, *Fli1*) with lower to absent expression of *Etv2* and *Runx1*. This observation is consistent with the known bifurcation of blood and endothelium (reviewed in Costa et al. (2012)) and the downregulation of *Runx1* in more mature endothelial cells (Samokhvalov et al. (2007)). Genes that mark early, intermediate and late stages of blood development showed dynamic expression across the diffusion map (Figure 4.3), with *Cdh1* expressed first, followed by *Cdh5* and then the embryonic globin *Hbb-bH1*. The transcription factors *Etv2*, *Tal1* (Scl), *Runx1* and *Gata1* were expressed in a pattern consistent with their known sequential roles during the development of hemangioblasts through to erythroid cells (Chen et al. (2009); Fujiwara et al. (1996); North et al. (1999); Robb et al. (1995); Schlaeger et al. (2005); Shivdasani et al. (1995); Sumanas et al. (2008); Wareing et al. (2012)). Dynamic expression patterns were also observed for other transcription factors not previously recognized as major regulators of primitive hematopoiesis, including *Erg*, *Sox7* and *Hoxb4*. The diffusion map method therefore represents an attractive approach for ordering cells in developmental time, identifying patterns of expression for key regulators and bifurcation events not readily found with standard algorithms.

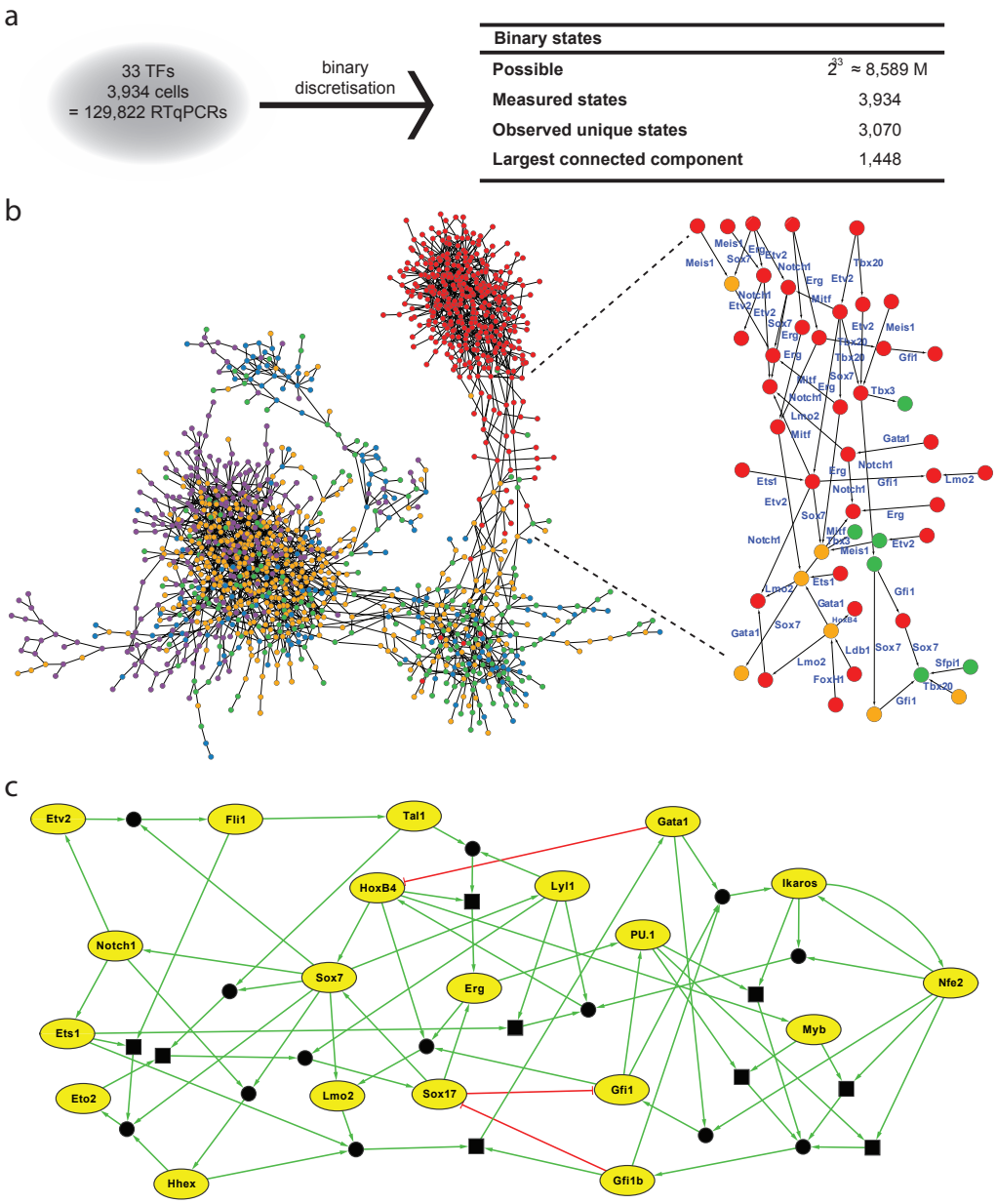


Figure 4.4 Regulatory network synthesis from single-cell expression profiles. (a) Discretization of 3934 expression profiles for 33 transcription factors yields 3070 unique binary states, 1448 of which can be connected by single-gene changes to yield a state graph. (b) Representation of resulting state graph, colored by first embryonic stage appearing in each state. Blue, PS; green, NP; orange, HF; red, 4SG; purple, 4SFG. Magnification of fate transition toward 4SG states, with, for example, Sox7 expression switching off along all routes. (c) Representation of synthesized asynchronous Boolean network models for core network of 20 transcription factors. Red edges indicate activation; blue edges indicate repression. Square boxes represent AND operations. Circles connecting edges indicate multiple update rules.

4.5 Synthesis of a network model for early blood development

The correspondence between the diffusion map and known developmental timelines suggested that the measured expression changes reflect developmental trajectories and might be exploited to define the regulatory networks that drive mesodermal cells toward a hematopoietic fate. Cell fate decisions have been modeled successfully using state space analysis of asynchronous Boolean regulatory network models (Bonzanni et al. (2013); Krumsiek et al. (2011)). In this approach, each gene is associated with a Boolean variable (1 or 0), which represents whether the gene is expressed or not expressed, respectively, in the cell. Each gene is also given a Boolean update rule that specifies how its expression value changes over time owing to regulation by other genes. Boolean network dynamics are then modeled by a series of asynchronous single-gene changes, and state space analysis reveals the final stable states of the model. We were interested in the inverse problem: if we think of the single-cell expression profiles as the state space of a Boolean network, can we identify the underlying gene regulatory logic? Although single-cell data have been used to refine static networks curated from the literature (Xu et al. (2014)), to our knowledge Boolean rules have not been derived directly from single-cell expression data without a priori knowledge of the structure of the network. To tackle this complex question of revealing the molecular changes underpinning cell state transitions, we developed an algorithmto synthesise Boolean networks based on single-gene transitions in our data.

We first discretized all 3934 single-cell expression profiles to binary states and connected those states that differ in the expression of only one gene. The threshold for binary discretization was determined as described in the Materials and Methods. This yielded a connected state transition graph of 1448 expression states, connected by single- gene transitions (Figure 4.4a,b). The number of times each state occurs is indicated in Figure 4.5. The probability of seeing even one repeated state or neighbor in the potential space of 2^{33} spaces is negligible, illustrating the nonrandom nature of the data. Most states that corresponded to the Runx1-GFP+ 4SG cells clustered together at one end of the state transition graph, whereas states corresponding to cells from other time points were dispersed between two additional clusters. Likely developmental transitions were revealed, with specific genes consistently switching on or off along all routes linking the major clusters. We therefore considered this state transition graph as a possible representation of developmental expression state changes based on single-gene switches, and next asked whether this could be used for regulatory network reconstruction. Notably, analysis of real and simulated popu-

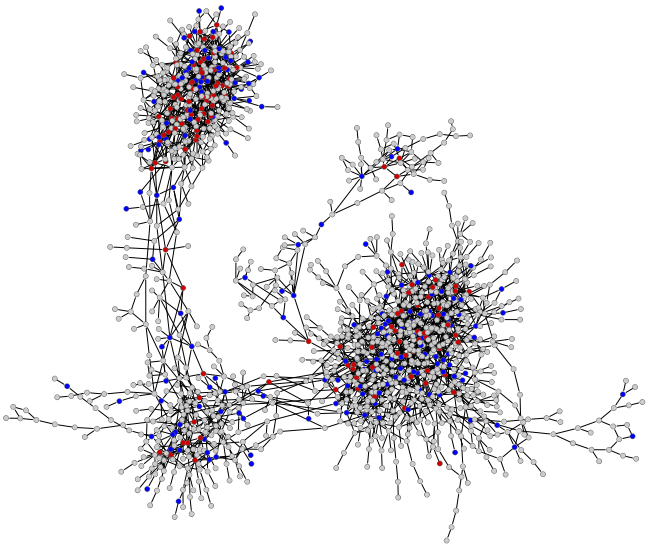


Figure 4.5 Some states occur in multiple cells. State transition graph coloured by the number of times each binary state occurs in the 3934 expression profiles. Grey, occurs once; blue, occurs twice; red, occurs more than twice.

lations of 20 cells showed that pools for the same stage clustered closely together, which masked variation and therefore would not have provided the number of transcriptional states required for network synthesis.

The direction of movement between two states in the state transition graph is initially not defined. Our method assigns a direction to each connection based on overall movement from the early PS to the later 4SG states, and then finds Boolean update functions for each gene that are consistent with its expression changes across the entire transition graph. Unlike previous analyses of single-cell gene expression data, which have largely relied on statistical properties of the data viewed as a whole, our method can recover mechanistic logic and determine the direction of interactions. When the method was applied to our data set, we obtained a core network of 20 transcription factors with endothelial and blood-associated gene modules centered on *Sox7*, endothelial and blood-associated gene modules centered on *Sox7*, *Hoxb4* and *Erg*, and on *Gata1* and *PU.1* (also known as *Spi1*), respectively. For some genes, there were multiple possible consistent update functions. For example, there are two solutions for *Erg*, both of which include activation by *Hoxb4* and *Sox17*. In total there were 39 possible functions, an average of two per gene. This led to 46,656 possible models from the different combinations of the 39 update rules (Figure 4.4c and Appendix A). Repeating the network synthesis with bootstrapping and a different discretization threshold

demonstrated the robustness of our protocol (Appendices B and C).

4.6 Network synthesis predicts direct regulation of Erg

We next asked whether links in our single-cell expression-derived network models can reveal direct regulatory interactions. To provide support for our model, we identified high-confidence gene regulatory regions in the gene loci of the 20 transcription factors in our network by interrogating a compendium of transcription factor ChIP-seq data from hematopoietic cell types (Sánchez-Castillo et al. (2015)), followed by identification of binding sites for the 20 transcription factors within these regions. 27 of the 39 Boolean rules (70%) are supported by the presence of evolutionarily highly conserved motifs for the upstream regulators in the target gene locus (Appendix A), with support for at least one Boolean rule for 16/20 transcription factors. This finding suggested that many of the regulatory interactions proposed in our model may be direct upstream regulator/downstream target gene relationships. To provide further validation, we focused on *Erg*, which our models predicted is activated by *Sox17*, or by *Hoxb4* in combination with *Lyl1* or *Scl* (*Tal1*). By analyzing a Hoxb4 ChIP-seq data set (Fan et al. (2012)), we showed that Hoxb4 can bind to the *Erg+85kb* enhancer, which was previously showed to be active in blood stem and progenitor cells (Thoms et al. (2011); Wilson et al. (2009)). Moreover, comparative sequence analysis revealed that the *Erg+85kb* contains highly conserved Hox and Sox binding sites (Figure 4.6a).

To investigate regulation of *Erg* by Hox and Sox factors, we took advantage of a recently described embryonic stem cell–based reporter system in which single-copy enhancer transgenes linked to the *Hsp68/Venus* reporter are targeted to the *Hprt* locus (Wilkinson et al. (2013)), allowing robust comparisons of wild-type and mutant enhancer activity during *in-vitro* differentiation. We tracked enhancer activity during embryoid body differentiation, where cells transit from a Flk1+ CD41 mesoderm/hemangioblast state, through a Flk1+ CD41+ intermediate, to a Flk1 CD41+ hematopoietic state (Kabrun et al. (1997); Mikkola et al. (2003); Mitjavila-Garcia et al. (2002); Wilkinson et al. (2013)). Flow cytometric analysis revealed a dynamic pattern of YFP expression for the wild-type enhancer, peaking at days 4–5 and highest in the Flk1+ CD41+ population (Figure 4.6b). Similar expression was seen in the Sox mutant, whereas mutation of the Hox motifs caused a reduction of YFP+ cells, and the combined Hox and Sox mutant reduced the proportion of YFP+ cells further still. We also saw similar patterns of expression in the other populations, which constitute a larger proportion of the embryoid body cells but have a lower percentage of YFP+ cells (Figure 4.6b). This suggests that Hox and Sox factors activate and maintain Erg ex-

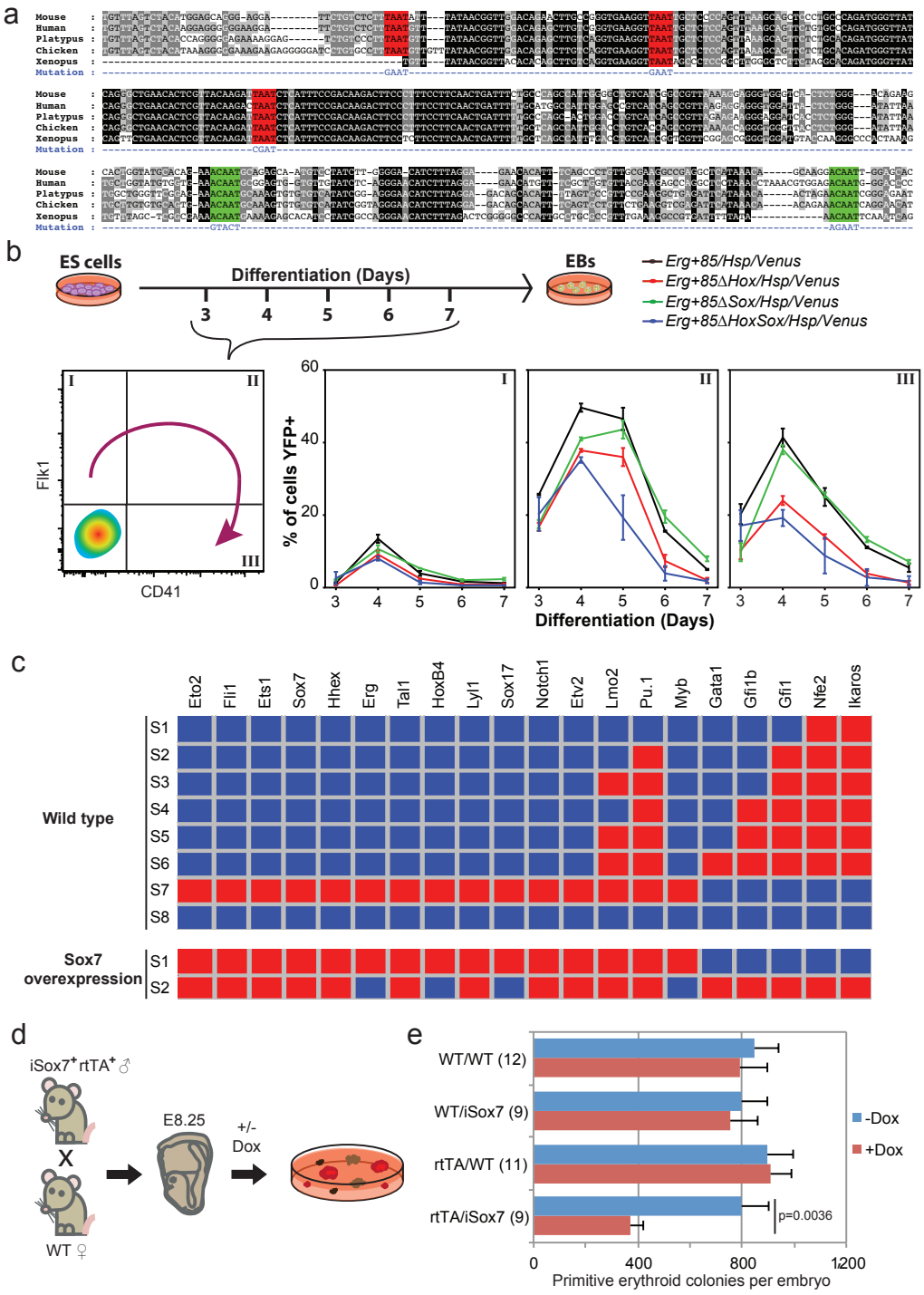


Figure 4.6 Network analysis predicts transcriptional interactions. (a) Alignment of mammalian *Erg+85kb* enhancer. Hox sites, red. Sox sites, light blue. (b) Percentage of Flk1+ CD41, Flk1+ CD41+ and Flk1 CD41+ cells on days 3–7 of differentiation expressing YFP. Data are mean and s.e.m. of triplicate differentiations of two to three clones per construct. (c) Network stable states for wild-type and Sox7 overexpression. Red indicates expressed; blue indicates not expressed. (d) Colony assays with or without doxycycline from genotyped E8.25 embryos from iSox7+ rtTA+ mice crossed with wild types. (e) Quantification of primitive erythroid colonies after 4 days (mean and s.e.m. for the number of embryos indicated). P-value was determined using the student's t-test for the number of embryos indicated.

pression largely independently and additively. When abstracted to the Boolean level, this result is therefore more consistent with the OR logic in our network than with the alternative AND logic, where single mutations would result in an effect as strong as the combined mutant.

4.7 Model execution reveals key switches during development

Next, we assessed whether our network models faithfully recapitulate blood and cardiovascular development, in which endothelial and primitive blood cells emerge from a common mesodermal progenitor. To do this, we determined the stable states of the network model that correspond to those expression patterns for the 20 transcription factors that satisfy all the Boolean network rules, and therefore can remain stable. We found that only eight stable states are reachable in total across all possible models, including "endothelial-like" (WT-S7) and "blood-like" expression states (WT-S2 to S6) (Figure 4.6c). Of note, 432 models had both the endothelial-like state and at least one of the blood-like states (WT-6) as stable states, thus capturing the functionality of bipotential Flk1+ precursors.

Finally, we explored the consequences of *in-silico* perturbation. Overexpression and knockout experiments were simulated for each transcription factor and the ability of the network to reach wild-type or new stable states was assessed. For a number of factors, stable states 6 or 7 were no longer reachable. Among these, enforced expression of Sox7, a factor normally downregulated when cells transit toward the 4SG state (Figure 4.4b), resulted in the stabilization of the endothelial module and an inability to reach any of the blood-like states (Figure 4.6c). Only two stable states were possible, among the lowest for all factors, and furthermore, Sox7 is predicted to regulate more targets than any other transcription factor, suggesting that perturbing its expression could have important downstream consequences. To validate this prediction, we crossed the previously reported iSox7+ rtTA+ male mice 37 with wild-type females, collected embryos at E8.25 and performed colony-forming assays s (Figure 4.6d). Embryos carrying both transgenes showed a 50% reduction of primitive erythroid colony formation and simultaneous appearance of undifferentiated hemangioblast-like colonies following doxycycline-induced *Sox7* expression compared to controls (Figure 4.6e)). This suggests, in agreement with modeling data and gene expression patterns, that downregulation of *Sox7* is important for the specification of primitive erythroid cells.

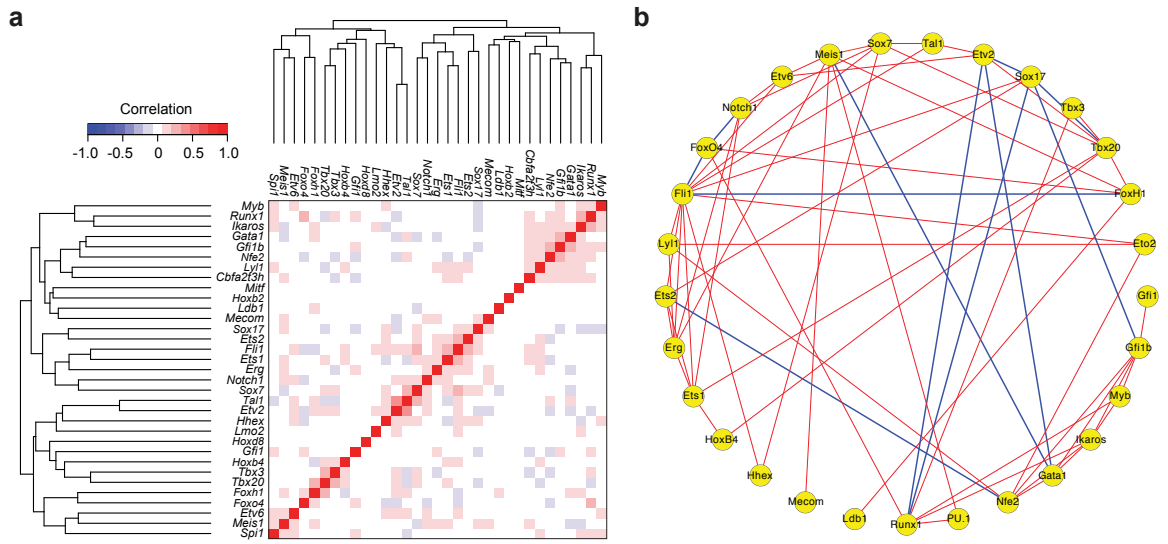


Figure 4.7 Partial correlation analysis. (a) Hierarchical clustering of partial correlation coefficients between pairs of transcription factors for all 3934 expression profiles. Pairwise coefficients were calculated while controlling for the effect of all other transcription factors. *Erg* does not correlate with Sox or Hox factors. (b) Network diagram showing putative undirected activating (red) and inhibiting (blue) relationships suggested by significant correlations (p-value < 1e-10).

4.8 Conclusions

Determining the structure and function of transcriptional regulatory networks is crucial to advancing our understanding of developmental and disease processes and is therefore a key aim of stem and developmental biology. However, studies to date have mainly used population-based data for network construction or have focused on statistical properties of populations of single cells for network inference.

Bayesian network methods provide a very computationally efficient approach to inferring causal relationships among a set of variables and have previously been applied to infer cellular signaling networks from single-cell data (Sachs et al. (2005)). However, these approaches infer a directed acyclic graph where there is no feedback between nodes, a limitation not shared by our approach. In addition, the inference of edges is reliant on network interventions in which many different cell populations are generated by experimentally perturbing genes, and the differences between these populations are used to infer causality. Generating such intervention data is very time consuming and cannot be done when studying wild-type *in-vivo* development. Instead, researchers typically look at the pairwise correla-

tion of genes across single-cell measurements (Guo et al. (2013); Moignard et al. (2013a)). For example, partial correlation analysis measures the degree of association between two genes while controlling for potential effects of all other genes (S. J. Welham, S. A. Gezan, S. J. Clark (1995)). We performed this analysis (Figure 4.7), and found agreement with many of the edges in our synthesized network; however, this analysis failed to predict the Sox/Hox regulation of *Erg*, which we validated experimentally. Moreover, connections do not specify which gene is the upstream regulator and which is the downstream target, and therefore do not reveal mechanistic logic.

To our knowledge no previous study has analyzed the development of an entire mammalian organ at single-cell resolution. Here we have studied the earliest stages of blood development from mesoderm through to the emergence of primitive erythroid cells, and demonstrate that single-cell expression profiling, coupled with computational approaches for network synthesis, can reveal molecular control mechanisms of mammalian organogenesis. Analysis of 46 genes in blood precursors across 1.25 days of post-implantation mouse embryonic development showed that cellular maturation may be asynchronous, with individual cells maturing at different speeds and a large proportion expressing both *Flk1* and *Runx1*, indicating that they are committing to hemogenic endothelial development. The graded changes in expression for key regulators of other mesodermal fates seen in the cell pools analyzed by RNA-seq are also consistent with cells expressing the gene emerging over the time-course analysed, although alternative explanations such as changes in the level of expression cannot be excluded. Furthermore, the diffusion map methodology highlighted the hierarchical nature of organ development, with waves of transcription factor and marker expression and a bifurcation at the four-somite stage. The presence of embryonic globin and erythroid transcription factor *Gata1* in one branch and endothelial markers such as *Pecam1* and *Cdh5* in the other suggests that this bifurcation represents the separation of blood and endothelial fates (Costa et al. (2012); Moignard et al. (2013b)). Trapnell et al. (Trapnell et al. (2014)) recently reported an exciting method related to our diffusion map approach for the analysis of single-cell, RNA-seq, time-course data, where construction of a minimum spanning tree ordered differentiating cells in developmental pseudotime. Although the authors suggested that this methodology could be used to map regulatory networks, such results were not included in their paper. Moreover, cells were sampled from cells differentiating *in-vitro* rather than directly from embryos.

Here we achieved reconstruction of regulatory network models by deriving expression-state graphs from high-throughput, single-cell, gene expression profiling data and using the expression-state graphs to determine gene regulatory rules. First, gene expression profiles

are discretized to binary expression states, where 1 represents a gene that is expressed and 0 represents no measurable expression. Then, pairs of states are connected if they differ in the expression state of exactly one gene, resulting in a state graph. Finally, Boolean rules are found for each gene, which allow a walk from early states to late states by means of a series of single-gene transitions. The result is a set of Boolean rules matching the experimental data that can be combined into a network model. This method requires no prior knowledge of regulatory interactions but instead derives its logic directly from the gene expression data.

We followed this method of network synthesis with steady state and *in-silico* perturbation analyses that identified blood and endothelial-like expression patterns and implicated Sox7 in the regulation of erythroid fate, which we subsequently validated using transgenic mouse assays. Network synthesis also identified several previously known transcription factor interactions, including close linkage of *Etv2*, *Fli1* and *Tal1*, where the latter two are known to function downstream of *Etv2* in the hemangioblast (Kataoka et al. (2011); Pimanda et al. (2007)). To test whether our network model reveals additional direct interactions, we focused on Erg, an essential transcription factor for definitive hematopoiesis and adult HSC function (Loughran et al. (2008); Taoudi et al. (2011)). Our network predicted that *Erg* expression can be activated either by Sox17 or Hoxb4. The *Erg+85kb* enhancer was previously shown to be controlled by Ets and Gata factors and to be active during hematopoietic development (Wilson et al. (2009)) and in HSCs (Thoms et al. (2011)) However, neither Hox or Sox transcription factors had been implicated in *Erg+85kb* activity.

Sox7 and Sox17 belong to the SoxF family of transcription factors, which have recently been shown to confer arterial identity in combination with RBPJ/Notch (Sacilotto et al. (2013)). Arterial identity is linked with the blood-forming potential of hemogenic endothelial cells in the embryo. Moreover, Hoxb4 expression is also known to enhance blood potential (Kyba et al. (2002)), yet there is very little knowledge about how SoxF factors or Hoxb4 integrates into the wider network regulating blood development. Our integrated approach of single-cell expression profiling coupled with network synthesis and subsequent experimental validation identifies *Erg* as a downstream target of Sox and Hox factors during early blood specification. Coupled with our observations here that downregulation of Sox7 is a key event in the development of primitive erythroid cells, our study demonstrates how network modeling from single cells can help to reveal the transcriptional hierarchies that control mammalian development. Rapid technological advances in our ability to perform single-cell profiling (Tang et al. (2011); Tischler and Surani (2013)) suggest that this approach will be widely applicable to other organ systems and may also inform the development of improved cellular programming strategies.

4.9 Materials and Methods

4.9.1 Single-cell qRT-PCR

Single-cell qRT-PCR was carried out using the Single-cell qRT-PCR. Single-cell qRT-PCR was carried out using the Fluidigm BioMark platform as described in Moignard et al. (2013a), with a limit of detection (LOD) of Ct 25. The LOD was determined according to Stahlberg et al. (2011) and manufacturer's instructions. Briefly, standard curves were run on the BioMark with six repeats of each dilution. For each gene, the LOD was the average Ct value for the last dilution at which all six replicates had positive amplification. The overall LOD for the gene set was the median Ct value across all genes. Gene expression was subtracted from the limit of detection and normalised on a cell-wise basis to the mean expression of the four house-keeping genes (*Eif2b1*, *Mrpl19*, *Polr2a* and *Ubc*) in each cell. Cells that did not express all four housekeeping genes were excluded from subsequent analysis, as were cells for which the mean of the four housekeepers was $\pm$ 3 s.d. from the mean of all cells. A dCt value of -14 was then assigned where a gene was not detected. 85–90% of sorted cells were retained for further analysis. *Gata2* did not amplify correctly and *HoxB3* was not expressed in any cells, so these factors have been excluded from the analysis. Further analyses were done on the dCt values for all transcription factors and marker genes, but not housekeeping genes.

4.9.2 Synthesis bootstrapping

To assess the robustness of the predictions of network synthesis, we performed bootstrapping. A random sample of 75% of the 3934 gene expression profiles was retained, and a new state transition graph was built from this reduced data set. This state transition graph was then used as the basis to synthesize new Boolean rules, using the same parameters as the original analysis. The results of repeating this process five times are shown in Appendix A. Bold entries indicate a rule is identical to a rule synthesized from the original data set. Underlined entries indicate that a rule is contained within a larger rule from the original synthesis. We see that in most cases the original rule or a closely related, underlined rule is synthesized. In general, the number of possible solutions for a gene's update function grows as the amount of data used is decreased, and including the full data set narrows these possibilities.

4.9.3 Assessing sensitivity of synthesised rules to binary discretisation threshold

In order to construct a state transition graph and apply our synthesis method, experimental data must first be discretised to binary values that indicate whether a gene is expressed or not expressed. The details of how we determine this threshold are covered in the section entitled "Single-cell q-RT-PCR", above.

To assess sensitivity of results to the choice of threshold, we repeated our analysis with a more stringent cut off, increasing it by two cycles. This resulted in a state transition graph of 1249 nodes (199 fewer nodes than the original state transition graph), which was then used as the basis to synthesize new Boolean rules, using the same parameters as the original analysis. The results are shown in Appendix B. Bold entries indicate a rule is identical to a rule synthesized from the original data set. Underlined entries indicate that a rule is contained within a larger rule from the original synthesis. We see that in most cases the original rule or a closely related, underlined rule is synthesised. In general, the number of possible solutions for a gene's update function grows as the amount of data used is decreased, and including the full data set narrows these possibilities.

Chapter 5

Graphical User Interface

5.1 Introduction

In chapter 3 we introduced an algorithm for synthesising asynchronous Boolean networks from single-cell gene expression data sets. Then, in chapter 4, we applied this method to understand the earliest stage of blood development in the mouse embryo. In this chapter we introduce the Single Cell Network Synthesis (SCNS) toolkit, a general-purpose graphical tool for the synthesis and analysis of models from single-cell data, which can be applied to new data sets as they become available (Figure 5.1).

In systems biology there is a need for general-purpose, user-friendly and efficient tools that can be readily used by biologists who do not have specialist computer science knowledge. SCNS allows biologists to reconstruct asynchronous Boolean network models from single-cell gene expression data sets covering developmental or differentiation time courses. The SCNS toolkit works on both qPCR and RNA-seq data, and can integrate data from both sources. It supports easy deployment to the cloud using the Microsoft Azure platform to increase performance, and control through a web-based graphical interface. Once models have been synthesised, SCNS can compute stable states and perform perturbation analysis, all within a single tool.

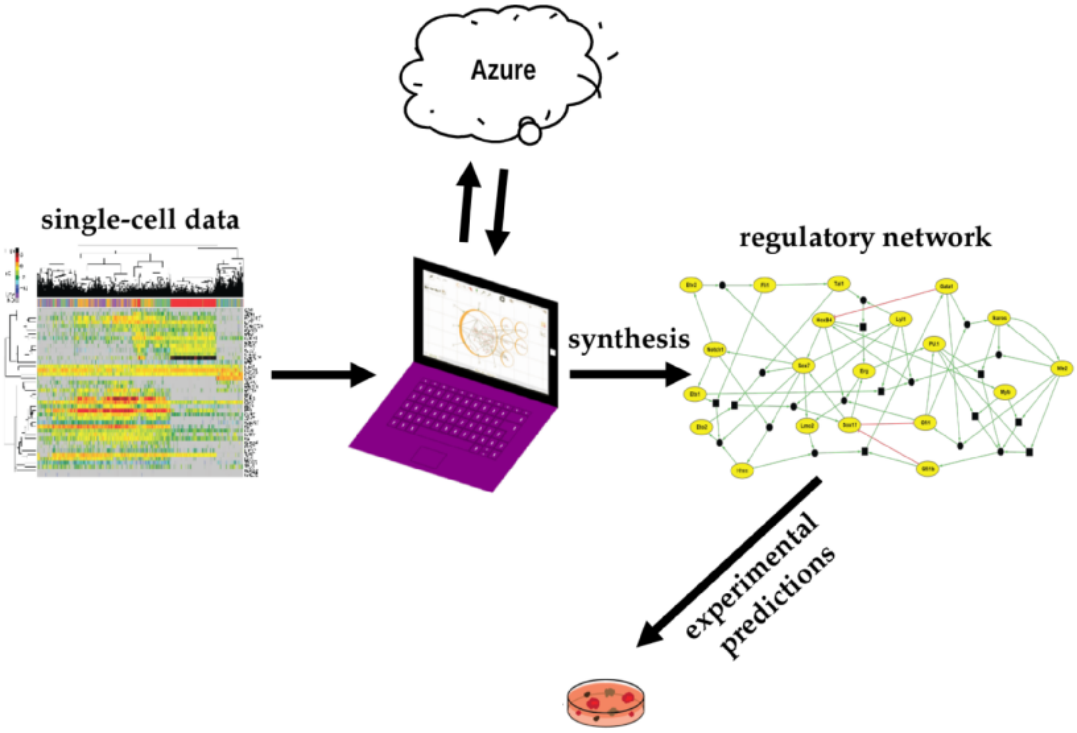


Figure 5.1 The Single Cell Network Synthesis Toolkit.

5.2 SCNS is controlled via a web-based graphical interface

When SCNS is first started, the user is presented with the 'Load Data' page, asking them to upload a .CSV file containing their single-cell gene expression data (Figure 5.2). This file should have rows corresponding to cells and columns corresponding to genes. Each entry should be a 1 or a 0, indicating whether the cell expresses the given gene or not. In addition, the first column should give the class of the cell. This indicates the cell type or day of measurement, and is used to indicate which cells states should be considered initial states and which target states during synthesis.

The browser then automatically switches to the 'STG' page, where a state transition graph automatically constructed from the uploaded data is displayed (Figure 5.3). On this page the user can use two text controls to select initial and target cell classes. For example, for the embryo data set from chapter 4, we would select 'PS' cells as initial states and '4SG' cells as target states (Figure 5.4).

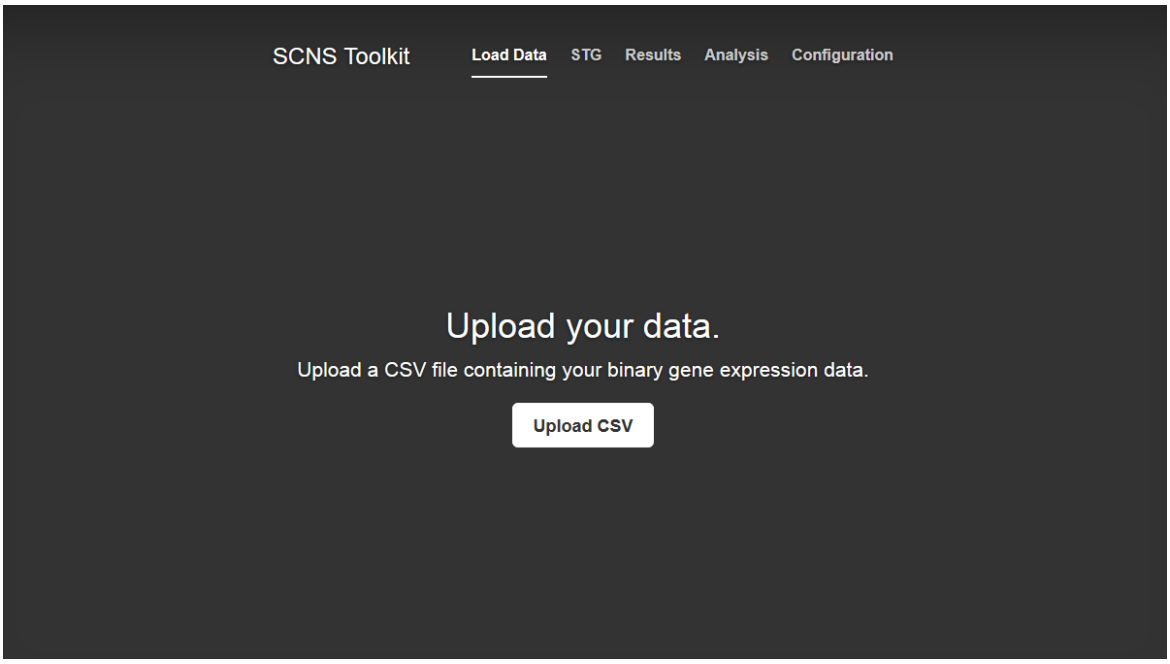


Figure 5.2 The upload data page, which the user is first presented with.

Below these text boxes are controls allowing the configuration of update function parameters (maximum size and threshold, see Chapter 3). The 'Synthesise' button can then be pressed to begin synthesising Boolean network rules. The browser switches to the 'Results' page and Boolean update functions are displayed in a table as they become available (Figure 5.5).

5.3 SCNS finds stable states and performs model perturbations

Once a model, or set of models, has been found, the user can navigate to the 'Analysis' tab to view the computed stable states (Figure 5.6). They can then use two text box controls to select any single or combined overexpression or knockout perturbation. The stable states will be dynamically recomputed with the chosen perturbation and re-displayed on the page (Figure 5.7).

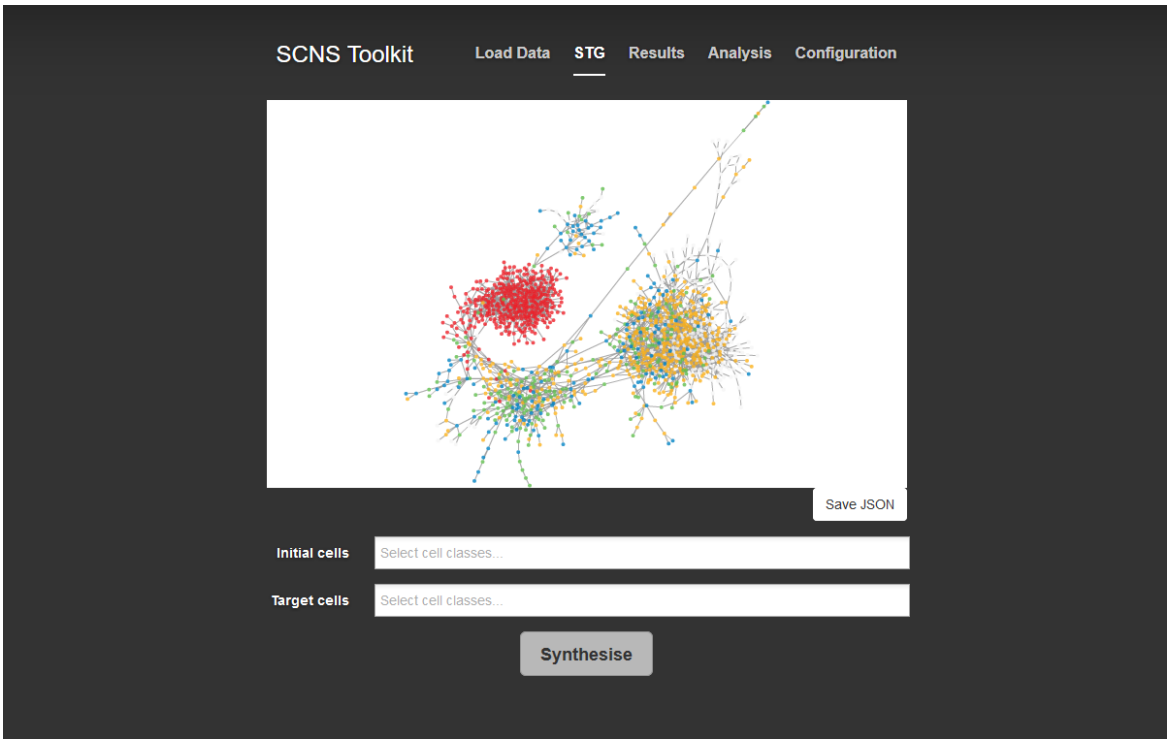


Figure 5.3 The state transition graph page, which allows visualisation of the data, selection of parameters, and running of synthesis.

5.4 SCNS can dispatch computations to the cloud

SCNS can perform all computations on the user's local workstation, or can deploy computations to the cloud and parallelise synthesis across nodes in order to speed up the search through the large number of possible model solutions. The compositional algorithm introduced in Chapter 3 factorises the search for a Boolean network in a way that treats each gene independently of others. SCNS takes advantage of this in order to parallelise computation. On a data set containing 100 genes, computation can in principle be parallelised over 100 compute nodes.

5.5 Tool architecture

SCNS is organised into two components. The frontend is a web-based graphical interface that the user interacts with in order to control the tool. The frontend is written using the AngularJS Javascript web application framework (http://angularjs.org). The backend, writ-

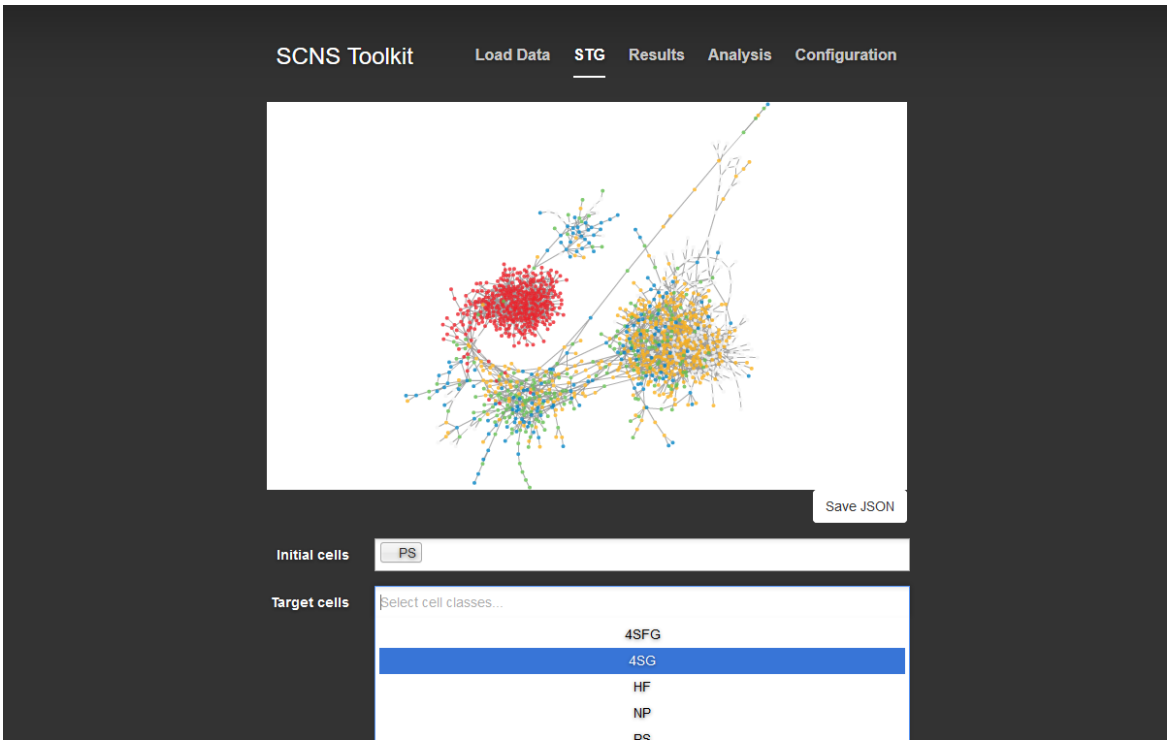


Figure 5.4 Selecting initial and target cell classes.

ten in F#, carries out all state transition graph, synthesis and stable state computations. The backend exposes a lightweight webserver using the Suave library (http://suave.io/, which runs locally and through which the frontend and backend communicate.

The mbrace (Dzik et al. (2013), http://mbrace.io/) cloud programming library is used to distribute computation to the cloud. Currently, this library supports the Microsoft Azure platform (http://azure.microsoft.com/). Amazon Web Services (http://aws.amazon.com/) support is currently under development. Once this is implemented, SCNS can be updated to support both systems.

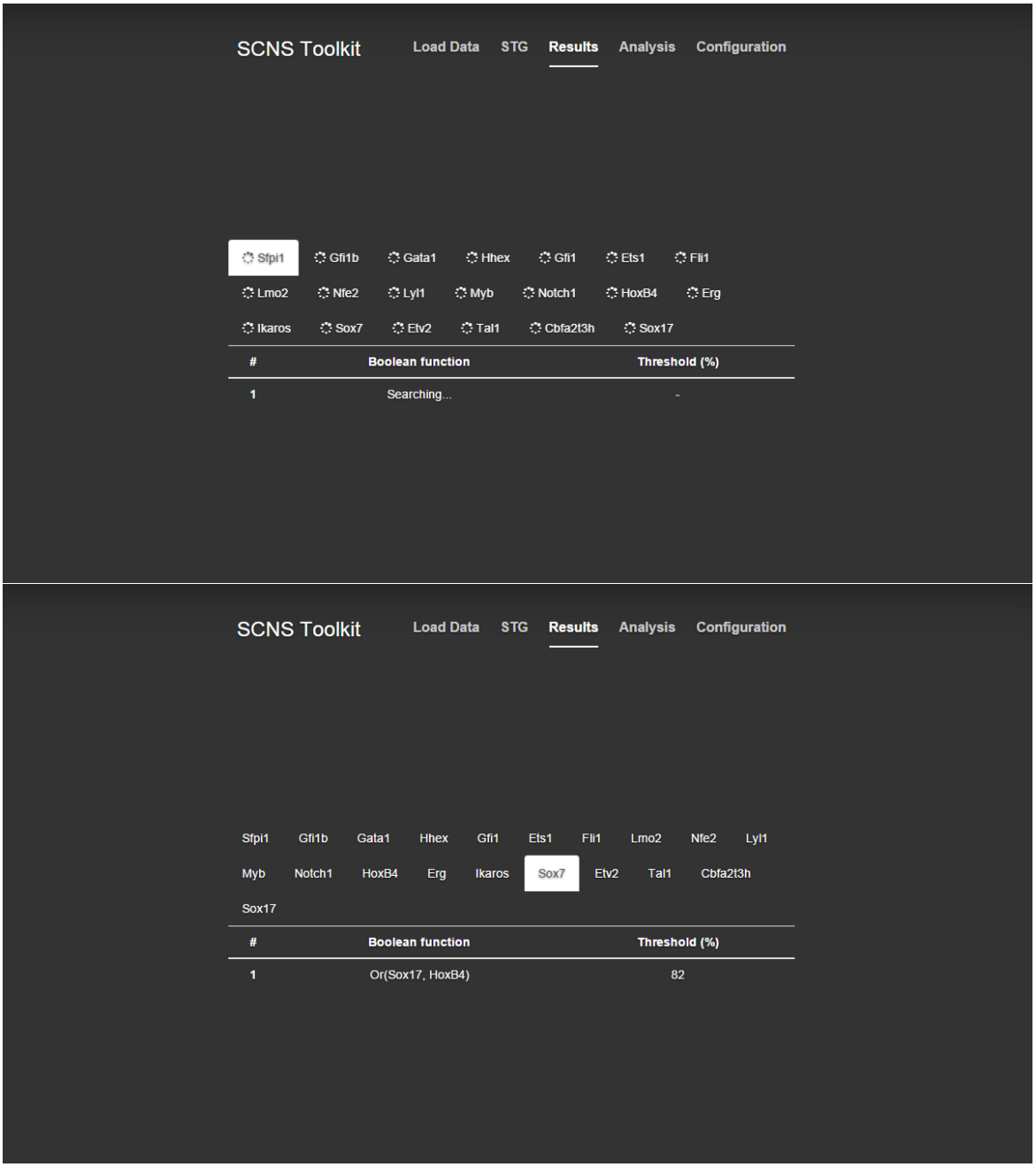


Figure 5.5 Results page. Matching Boolean functions are displayed as they become available. Spinning icons indicate that synthesis has not yet finished.

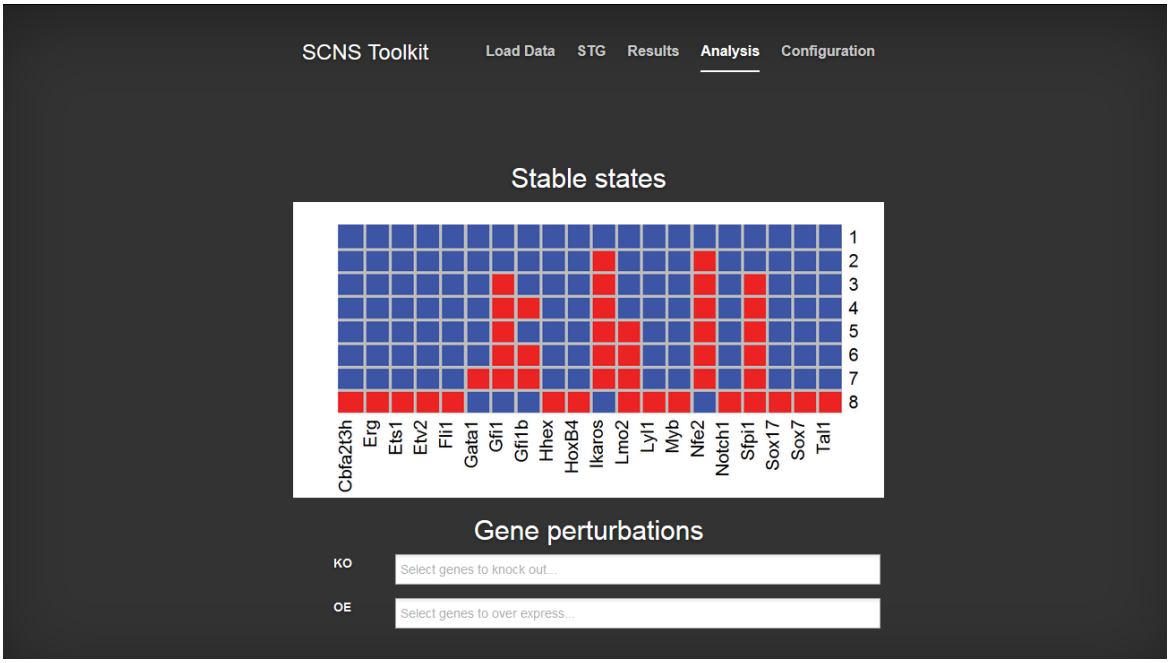


Figure 5.6 Analysis page. Computed stable states are shown.

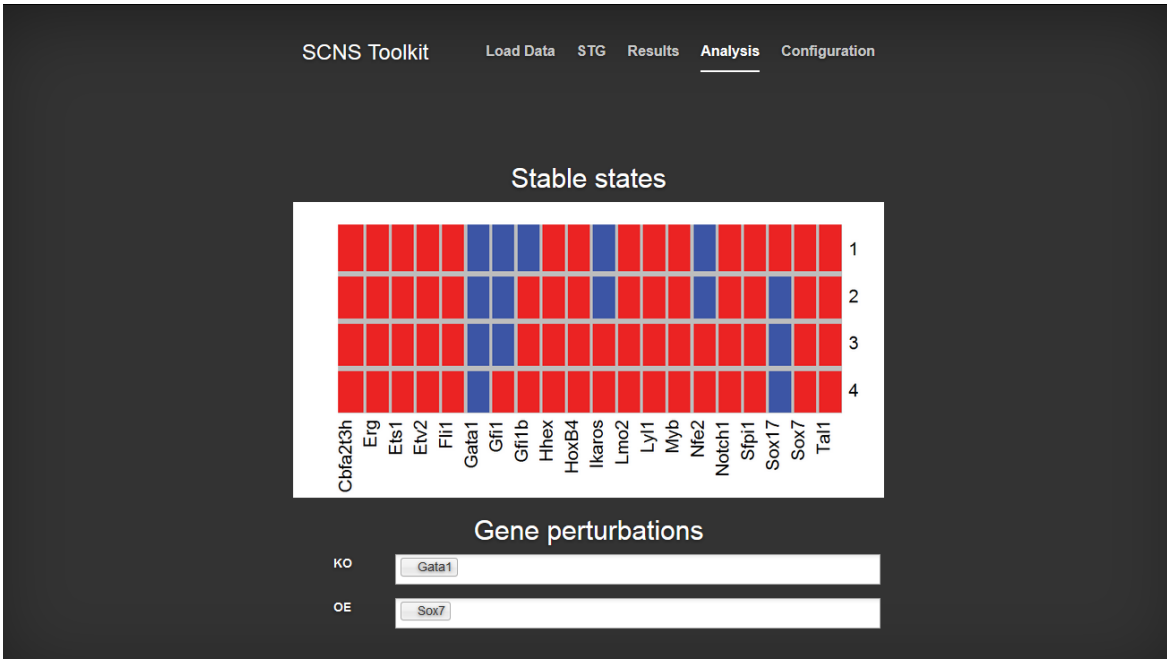


Figure 5.7 Recomputed stable states under a selected combined overexpression/knockout perturbation.

Chapter 6

Discussion

6.1 Gene regulatory network reconstruction from single-cell data

Uncovering and understanding the gene regulatory networks (GRNs) which underlie development and homeostasis is a central issue in molecular cell biology. New single-cell resolution gene expression measurement technology provides snapshots of the gene expression states of the cells that make up a biological tissue, a level of detail which has not been available before. The aim of this PhD was to investigate the possibility of using this new high resolution data to reconstruct mechanistic computational models of gene regulatory networks, which could then be tested experimentally, and used to make useful predictions. This aim led to several objectives:

1. To develop and implement an algorithm for the reconstruction of executable models of gene regulatory networks from single-cell gene expression data.
2. To apply this algorithm to a new data set covering 3934 single cells measured during early embryonic blood development, in order to reconstruct a predictive model of primitive haematopoiesis and generate new biological insights.
3. To develop a user-friendly and efficient graphical tool which can be used by biologists to reconstruct gene regulatory network models from new single-cell gene expression data sets as they become available.

In chapter 3, a synthesis algorithm was introduced for the reconstruction of asynchronous Boolean networks. Taking advantage of the single-cell resolution of the data, I treated ex-

pression profiles as states of an asynchronous Boolean network and framed model inference as the problem of reconstructing a Boolean network from its state space. I then introduced a scalable algorithm to solve this synthesis problem. In order to achieve scalability, this algorithm works in a modular way, treating different aspects of a graph data structure separately before encoding the search for logical rules as Boolean satisfiability problems to be dispatched to a SAT solver.

In chapter 4, this synthesis algorithm was applied to the experimental data set. This resulted in a 20-node asynchronous Boolean network for early blood development, which was consistent with known biology. By applying standard techniques for the analysis of Boolean networks, I found the stable state attractors and performed computational perturbations. The synthesised network, along with the subsequent computational analysis led to a set of novel predictions which were then tested experimentally. I found that these results were robust when performing bootstrapping, removing a third of the data at random and rerunning the synthesis algorithm.

Experimental collaborators were able to validate key predictions made by the analysis. The update function for one of the genes at the core of this network, Erg, which directly activates many other genes, was tested experimentally in an embryonic stem cell model of early blood development. Evidence was found that the activators specified in the gene's synthesised update function (HoxB4 and Sox17) do indeed activate expression of the gene, and furthermore in a fashion consistent with the Boolean "OR" logic of the synthesised update function. This could be regarded as a "local" validation of our model, testing two of the directed edges in the network.

Computational perturbations to another gene at the core of the network, Sox7, indicated that when Sox7 was overexpressed, stable states corresponding to primitive erythroid progenitors no longer exist. Cell-forming assays from an embryos of a mouse with an inducible Sox7 transgene confirmed this prediction experimentally, finding that Sox7 induction results in a significant decrease in erythroid colony formation. This can be thought of as a "global" validation of our model, as it is a prediction about the behaviour of the whole network under a certain perturbation.

Finally, in chapter 5, a general purpose tool was developed to be used by biologists to construct models from new data sets. This tool has a web-based graphical user interface and a backend which can dispatch computations to the cloud in order to scale to larger data.

Previous analyses of single-cell gene expression data had mostly been based on statistical properties of the data viewed as a whole, such as the correlation in the level of expression of

pairs of genes. Such analysis cannot recover mechanistic Boolean logic, does not infer the direction of interactions and cannot easily distinguish direct from indirect influence. During the course of this PhD, two methods for reconstructing mechanistic gene regulatory network models from single-cell gene expression data have appeared. Chen et al. introduced an approach for inferring asynchronous Boolean networks from a directed cell-lineage tree that describes relationships between cell types, together with single-cell measurements from each of these cell types (Chen et al. (2014a)). The method uses a genetic algorithm, a local search method, to optimise network structures. Networks which permit n-step transitions from early cell types to late cell types and minimise transitions from late cell types to early cell types are given high scores. The method was applied to a data set of approximately 500 cells taken during preimplantation mouse development from 16- to 64-cell stage embryos (Guo et al. (2010)). The reconstructed model was found to be in good agreement with a benchmark GRN constructed manually from experiments (Oron and Ivanova (2012)).

This method is conceptually similar to my approach. The first key difference is that the very large number of cells measured in our experimental case study allowed me to construct a much more fine-grained and detailed lineage tree where each node corresponds to an individual cell and each edge to a change in a single gene. The second key difference is that I obtained exact/optimal solutions to my reconstruction problem by using synthesis techniques, rather than employing local search methods.

Ocone et al. introduced a method for deriving ordinary differential equation (ODE) models from single-cell gene expression data (Ocone et al. (2015b)). In this approach, a static relevance network is obtained using GENIE3 (another approach, such as partial correlation or mutual information could also be used) and cells are ordered in "pseudotime" using a diffusion map and the Wanderlust algorithm (Bendall et al. (2014)). This reconstructed time course is then used to optimise the parameters of different candidate ordinary differential equation (ODE) models that fit the static topology of the relevance network. These ODEs can incorporate mechanistic logic such as AND and OR gates. An advantage of this approach is that continuous gene expression levels are taken into account, while SCNS only deals with binary expression. However, the structure of the network is derived using correlation based approaches rather than directly from the time course, and so is subject to the limitations of correlation. A weakness is that only one route through the data is considered, while the SCNS toolkit integrates information from many paths and multiple starting and ending points in the state transition graph. Scalability may also be an issue. Since there a super exponential number of potential ODE models to train and compare, it seems unlikely this approach can scale to very large data sets.

A similar approach should work for the learning of Dynamic Bayesian Networks (DBNs). A DBN is a Bayesian network which incorporates a notion of time, updating the value of variables based upon values at the previous time step. DBNs avoid the two serious limitations of Bayesian networks that were discussed in the introduction: the restriction to an acyclic graph and the need for perturbation data. The training on sequential time course data often allows the direction of edges to be determined without recourse to experimental perturbations, and while the structure of the DBN remains an acyclic graph and there is no feedback within a single time step, the value of downstream nodes from the current time step are allowed to influence upstream nodes at the next time step. It would be interesting to see how such an approach would compare to both the SCNS toolkit and the method of Ocone.

6.2 Synthesis in biology

Synthesis has recently been applied in the context of biology, during the course of this PhD. Koksal et. al. show how to synthesise state-machine-like models from gene mutation experiments using a novel counterexample-guided inductive synthesis (CEGIS) algorithm (Koksal et al. (2013)). Previous work introduced a state-machine based model that describes how signalling between a set of cells leads to a specific, invariant, cell fate pattern in the development of the *C. elegans* vulva. Koksal et. al. show how this model can be reconstructed from a network of known gene interactions and a specification of the effect of genetic perturbations on cell fates, given in a genotype-phenotype table. The approach introduces "holes" in the update functions for specific genes, which are then automatically filled in by the synthesiser. Both the data and the type of model considered in this work were different from those dealt with in the current thesis, which called for a different approach.

Recently, there have been several applications of synthesis to Boolean networks. Sharan and Karp (Sharan and Karp (2013)) synthesise Boolean networks given an existing, static topological network and a set of perturbation experiments, by reduction to the *NP*-hard integer linear programming (ILP) problem (Jünger et al. (2010)). The topological network given as input must be a directed acyclic graph, and it may be signed to indicate whether a relation is activating or repressing. The perturbation experiments fix the value of a subset of genes and report the observed value of another subset of genes. From these two sets of inputs, a network is learnt that optimally fits the observed data. This problem was previously shown to be *NP*-complete (Karlebach and Shamir (2012)). The ILP encoding of this problem introduces 2^n binary variables to represent the truth table of a Boolean function and integer

linear constraints to minimise the number of disagreements with experimental data. The CPLEX solver is then used to obtain an exact solution.

Sharan and Karp apply their method to the well-studied EGFR and IL-1 signalling networks. Large Boolean models for these two systems already exist, and their agreement with experimental perturbation data has been assessed (Ryll et al. (2011); Samaga et al. (2009)). The ILP encoding was used to pinpoint modifications to these models which improve their fit to the experimental data. The hand-curated EGFR network has 112 nodes, and 34 experimental perturbation measurements were made. The network has a 76% agreement with these experimental observations. The ILP solution was used to suggest 4 minimal modifications to the model which increase the agreement to an optimal value of 90%. However, this model is not unique, and searching for other optimal models revealed alternate networks which were markedly different, suggesting that even with an existing topological network much more experimental perturbations are needed to narrow the search space of matching Boolean networks. Together the requirement for comprehensive perturbation data and an existing topological network represent rather extensive prior knowledge of the gene regulatory network, and limit the application of this approach. The limitation to a directed acyclic graph is subject to the same criticisms as for the Bayesian network methods discussed in the introduction of this thesis. The use of ILP solving rather than SAT/SMT solving, however, is a potentially very interesting aspect of this approach. A similar method has been introduced by Guziolowski and Videla et al., which uses an Answer Set Programming (ASP) formulation instead of ILP (Guziolowski et al. (2013)). ASP is a search method based on logic programming (Gebser et al. (2007a, 2012)). It would be very interesting to assess the relative advantages of ILP, ASP and SAT/SMT approaches.

Dunn et. al. and Xu et. al. show how to fit an existing static network for embryonic stem cells to gene expression data in order to obtain an executable Boolean network, under the assumption that experimentally measured data represent stable states of the system (Dunn et al. (2014); Xu et al. (2014)). Dunn et. al. firstly obtain a relevance network by correlating the level of expression of 17 transcription factors, measured in population qPCR data in a range of culture conditions and in time-course experiments. They then define a set of 23 desired stable states which must be reachable from defined initial states. They searched, using an SMT solver, for Boolean networks that have a topology which is a subset of the input relevance network and which have the desired stable states. The resulting models were then used to make 28 new predictions about the effect of genetic perturbations on the stable states of the system, 17 of which were experimentally validated. The use of correlation to obtain the initial topology of the network is subject to the criticisms of correlation-based

approaches to infer networks.

Xu et al. obtained a 30-node signed and directed topological network from previously published ChIP-seq and knockdown studies. The expression of these 30 genes was then profiled by single-cell PCR in 96 cells across two different culture conditions. These single-cell measurements were then treated as stable states, and used to find a Boolean network. The update function for each gene was searched for separately, by exhaustively enumerating all functions that match the given topology and selecting the function that minimises transitions out of the experimentally measured states. The resulting networks were used to make individual, single, and triple *in-silico* knockdowns and found to be in good agreement with experimental data. The assumption that experimentally measured cell states represent state states of the model may be appropriate for cell lines maintained in culture, but it does not transfer to developmental processes such as ours, where cells are transiting through intermediate states in order to develop into a particular lineage.

Paoletti et al. synthesise a related class of models which extend Boolean networks with timing and spatial information (Paoletti et al. (2014)). These are the same class of synchronous Boolean models used by Eric Davidson's lab to model the development of the sea urchin. Their SMT-based approach uses the theory of uninterpreted functions to synthesise model logic from specifications of the synchronous time course of the system in wild type and perturbed conditions. This is therefore the most similar method to ours, different in that it uses the linear path of a synchronous time course rather than a branching asynchronous statespace as a specification. Paoletti et al. apply their method to reconstruct the first model of sea urchin development that can explain all (rather than most) previous experimental data.

These recent results demonstrate the promise of synthesis in biology. New methods are able to automatically reconstruct mechanistic models that satisfy all existing experimental specifications, and which can subsequently be used to make new predictions to be validated in the lab. These models would previously have to be built by hand from experimentally determined regulatory logic, and tested to ensure they behave as expected.

Synthesis approaches generally lead to combinatorial rather than statistical problems, which are then exactly solved using algorithms that leverage highly optimised specialist solvers. Synthesis yields a globally optimal model that satisfies the specification given by the data completely, or otherwise informs the user that no such model exists. Unlike local learning approaches, there is no issue of getting stuck in locally optimal solutions. Synthesis can also be used to find multiple, or all, models that satisfy a given set of model specifications, and design experiments to distinguish between the different possibilities.

The different synthesis approaches discussed here reconstruct models from a range of different experimental data. The approach introduced in this thesis is the first to synthesise gene regulatory network models directly from raw gene expression data, without the need of either genetic perturbation data or a-priori information about the topology of the network. The major disadvantage of my approach is the need for a very large number of measured single cells (thousands rather than hundreds) in order to construct a connected state-transition graph. Development of methods which can incorporate data from multiple experimental sources will improve the quality of reconstructed models.

6.3 Applicability of SCNS to new data

While I have successfully the SCNS toolkit to reconstruct a mechanistic model of early blood development which is consistent with known biology, and with which we were able to validate new predictions, this is ultimately only one test case. To fully evaluate the method it needs to be applied to more data sets.

The field of single-cell genomics is still relatively young, and there is a sparsity of high-quality data sets with a large number of cells. This is set to change as adoption of these protocols becomes more widespread. New single-cell studies will give us insights into organ development and human disease. A larger number of data sets will allow comprehensive evaluation of techniques introduced to reconstruct gene regulatory networks, and allow assessment of the advantages and disadvantages of different approaches and improvement of the methods.

The move towards whole-transcriptome RNA-sequencing data removes the selection bias of qPCR data and allows analysis of the full genetic programme of the cell but presents its own challenges (Macaulay and Voet (2014); Tang et al. (2009)). In chapter 5, I found that I could combine data obtained by single-cell RNAseq with data from qPCR, after removing genes not in the qPCR data set. Cells measured by RNAseq clustered together with cells of the same type measured by qPCR. There does not seem to be a problem in principle, therefore, to applying SCNS to RNAseq data. However, methods for identifying the most important genes and automatically reducing the gene set to a manageable size will be required.

Another interesting project would be to investigate the application of my synthesis tool to single-cell proteomics data sets, such as those measured by single-cell mass cytometry (Zunder et al. (2015)). Protein information is potentially more interesting than mRNA data, as it is the protein that is functional. Ideally, it would be possible to use both sources of informa-

tion. The analysis of proteomics data will present challenges. Single-cell mass cytometry data sets have different characteristics to single-cell qPCR data sets, and often consist of many more cells. A new pipeline for the processing of this new data type would need to be introduced, and then an assessment of whether my tool can successfully reconstruct executable models from it.

6.4 Improvements to the algorithm

There are several modifications which could be made to the synthesis algorithm in an attempt to improve efficiency or to extend it to different classes of model. Firstly, different SAT encodings could be used for the search for Boolean functions. Two different, equally natural SAT encodings for the same problem can lead to very different performance behaviours (Hertel et al. (2007)). For this reason, different encodings should be experimented with and their effect on performance examined. The best encoding for a particular problem for a CDCL solver is not necessarily the one with the smallest number of variables and clauses. One important consideration is the ammendability of the encoding to unit propogation: if a fact or conflict can be derived by unit propagation then backtracking search does not need to be invoked. Brain et al. have recently introduced an approach for automatically deriving "optimally propagating" SAT encodings (Brain et al. (2016)). Propagation-friendly SAT encodings of cardinally constraints $x_1 + \cdots + x_n \leq k$ have been known for some time (Abío et al. (2013); Ansótegui and Manya (2005); Bailleux and Boufkhad (2003); Ben-Haim et al. (2012); Biere et al. (2014); Chen (2010); Eén and Sorensson (2006); Klieber and Kwon (2007); Prestwich (2007); Sinz (2005)).

It is possible that encoding the search for Boolean update functions in a more expressive logic using Satisfiability Modulo Theories solvers could give a more concise representation and lead to a more efficient solution (Abío et al. (2013); Bayless et al. (2015); Sebastiani (2007)). Since my synthesis problem is essentially a discrete optimisation problem, it is also possible that the application of integer linear programming solvers (such as CPLEX or Gurobi) as in the work of Sharan and Karp discussed above (Sharan and Karp (2013)) rather than satisfiability solvers would result in a faster implementation (Li et al. (2004)). The work on satisfiability of linear integer constraints and optimisation in the context of SMT solvers should also be investigated (Bjø rner et al. (2015); Bobot et al. (2012); Bromberger et al. (2015); Dillig et al. (2009); Hendrix and Jones; Jovanović and De Moura (2011); Li et al. (2014); Sebastiani and Tomassi (2012)), as well as answer set programming (Baral (2003); Gebser et al. (2007a,b, 2014)).

It would be interesting to adapt the algorithm to the synthesis of qualitative, rather than Boolean, networks. Qualitative networks extend variables to take values from a specified finite range, say $\{0,1,2,3\}$ to represent zero, low, medium and high levels of expression (Schaub et al. (2007)). There are no obvious barriers to doing this in principle. A method would need to be chosen to discretise single-cell gene expression measurements to the discrete values, and then a state transition graph could be constructed where edges represent the increase or decrease in the level of expression of a single gene. The SAT encoding could be adapted so that update functions increase and decrease expression levels. One problem that this modification would introduce is that the number of possible discrete cell states would increase, along with the number of cell measurements required to successfully construct a connected state transition graph.

6.5 Concluding comments

There are two questions at the heart of modern cell biology research. The first is how to destroy diseased and cancerous cells in the body without harming healthy tissue, and the second is how to produce clinically-relevant cell types in the lab for the purposes of regenerative medicine. There have been breakthroughs in both of these areas in recent years, most strikingly in the immunotherapies and stem cell treatments that are currently going through clinical trials. However, most cancers remain incurable and there is still much more to be done in regenerative medicine research. For example, despite decades of research on the development of blood in the embryo it is still not known how to produce blood stem cells by directed differentiation of embryonic stem cells in the lab, a breakthrough that would have huge clinical applications as a replacement for bone marrow transplantation.

Reconstructing mechanistic models of the gene regulatory networks underlying developmental and disease processes represents an important step towards tackling these problems. The combination of new single-cell data with new computational techniques that take full advantage of this data promises to massively advance the nascent field of systems biology. The development of general-purpose tools that can be used by biologists as new data becomes available, like the SCNS toolkit developed as part of this PhD, is crucial to ensure the widespread adoption and application of these new techniques.

References

Ignasi Abío, Robert Nieuwenhuis, Albert Oliveras, Enric Rodríguez-Carbonell, and Peter J Stuckey. To encode or to propagate? The best choice for each constraint in SAT. In *Principles and Practice of Constraint Programming*, pages 97–106. Springer, 2013.

Karen Adelman and John T Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics*, 13(10):720–31, 2012. ISSN 1471-0064. doi: 10.1038/nrg3293. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3552498&tool=pmcentrez&rendertype=abstract.

By Manindra Agrawal, Neeraj Kayal, Nitin Saxena, Manindra Agrawal, Neeraj Kayal, and Nitin Saxena. PRIMES is in P. *Annals of Mathematics*, 160(August 2009):781–793, 2004. ISSN 0003486X. doi: 10.4007/annals.2004.160.781.

T Akutsu, S Miyano, and S Kuhara. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*, 28: 17–28, 1999. ISSN 2335-6936.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Chromosomal DNA and Its Packaging in the Chromatin Fiber. *Molecular biology of the cell, 4th edition*, (New York: Garland Science), 2002.

Eric Allender, Michael Bauland, Neil Immerman, Henning Schnoor, and Heribert Vollmer. The complexity of satisfiability problems: Refining schaefer's theorem. In *International Symposium on Mathematical Foundations of Computer Science*, pages 71–82. Springer, 2005.

Uri Alon. Network motifs: theory and experimental approaches. *Nature reviews. Genetics*, 8(6):450–61, 2007. ISSN 1471-0056. doi: 10.1038/nrg2102. URL http://www.ncbi.nlm.nih.gov/pubmed/17510665.

Amnon Amir, Oren Kobiler, Assaf Rokney, Amos B Oppenheim, and Joel Stavans. Noise in timing and precision of gene activities in a genetic cascade. *Molecular systems biology*, 3(71):71, 2007. ISSN 1744-4292. doi: 10.1038/msb4100113. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1828745&tool=pmcentrez&rendertype=abstract.

El-ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana

Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 31(6):545–52, 2013. ISSN 1546-1696. doi: 10.1038/nbt.2594. URL http://www.nature.com.ezp-prod1.hul.harvard.edu/nbt/journal/v31/n6/full/nbt.2594.html.

S. Anders, P. T. Pyl, and W. Huber. HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2014. ISSN 1367-4811. doi: 10.1101/002824. URL http://biorxiv.org/content/early/2014/08/19/002824.abstract.

Carlos Ansótegui and Felip Manya. Mapping problems with finite-domain variables to problems with boolean variables. In *Theory and Applications of Satisfiability Testing*, pages 1–15. Springer, 2005.

Timos Antonopoulos, Nikos Gorogiannis, Christoph Haase, Max Kanovich, and Joël Ouaknine. Foundations for decision problems in separation logic with general inductive predicates. In *Foundations of Software Science and Computation Structures*, pages 411–425. Springer, 2014.

Y Arinobu, S Mizuno, Y Chong, H Shigematsu, T Iino, H Iwasaki, T Graf, R Mayfield, S Chan, P Kastner, and K Akashi. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell*, 1(4):416–427, 2007. ISSN 19345909. doi: S1934-5909(07)00070-7[pii]$\backslash$r10.1016/j.stem.2007.07.004. URL http://www.ncbi.nlm.nih.gov/pubmed/18371378.

Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*, volume 6. 2009. ISBN 0521424267. doi: 10.1088/1742-6596/1/1/035. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.7207&rep=rep1&type=pdf$\delimiter"026E30F$nhttp://www.math.sc.edu/~cooper/math778C/abct.pdf.

Gilles Audemard and Laurent Simon. Predicting learnt clauses quality in modern SAT solver.pdf. *Ijcai*, pages 399–404, 2009. ISSN 10450823. doi: 10.1.1.150.1911. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.1911.

Gilles Audemard and Laurent Simon. Refining restarts strategies for SAT and UNSAT. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7514 LNCS, pages 118–126, 2012. ISBN 9783642335570. doi: 10.1007/978-3-642-33558-7_11. URL http://link.springer.com/chapter/10.1007/978-3-642-33558-7_11.

Shirin Azizidoost, Mehrnoosh Shanaki Bavarsad, Mahsa Shanaki Bavarsad, Saeid Shahrabi, Kaveh Jaseb, Fakher Rahim, Mohammad Shahjahani, Fakhredin Saba, Mahdi Ghorbani, and Najmaldin Saki. The role of notch signaling in bone marrow niche. *Hematology*, 20(2):93–104, 2015. ISSN 1607-8454. doi: 10.1179/1607845414Y.0000000167. URL http://www.maneyonline.com/doi/abs/10.1179/1607845414Y.0000000167?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub=pubmed&.

Olivier Bailleux and Yacine Boufkhad. Efficient CNF encoding of boolean cardinality constraints. In *Principles and Practice of Constraint Programming–CP 2003*, pages 108–122. Springer, 2003.

Philip Ball. Portrait of a molecule. *Nature*, 421(6921):421–422, 2003. ISSN 0028-0836. doi: 10.1038/nature01404.

Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011. ISSN 1748-7838. doi: 10.1038/cr.2011.22. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3193420&tool=pmcentrez&rendertype=abstract.

Chitta Baral. *Knowledge representation, reasoning and declarative problem solving*. Cambridge university press, 2003.

Yoseph Barash, John a Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J Blencowe, and Brendan J Frey. Deciphering the splicing code. *Nature*, 465 (7294):53–59, 2010. ISSN 0028-0836. doi: 10.1038/nature09000. URL http://dx.doi.org/10.1038/nature09000.

Margaret H Baron, Andrei Vacaru, and Johnathan Nieves. Erythroid development in the mammalian embryo. *Blood cells, molecules & diseases*, 51(4):213–219, 2013. ISSN 1096-0961 (Electronic). doi: 10.1016/j.bcmd.2013.07.006.

Clark Barrett, Roberto Sebastiani, and Seshia Cesare Tinelli. Handbook of Satisfiability. *New York*, 185(3):980, 2009. ISSN 09226389. doi: 10.3233/978-1-58603-929-5-457. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/1586039296.

Sam Bayless, Noah Bayless, and Alan J Hu. SAT Modulo Monotonic Theories. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3702–3709, 2015.

Attila Becskei, Benjamin B Kaufmann, and Alexander van Oudenaarden. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature genetics*, 37(9):937–944, 2005. ISSN 1061-4036. doi: 10.1038/ng1616.

Thomas Bee, Emma L K Ashley, Sorrel R B Bickley, Andrew Jarratt, Pik-shan Li, Jackie Sloane-Stanley, Berthold Göttgens, and Marella F T R de Bruijn. The mouse Runx1 +23 hematopoietic stem cell enhancer confers hematopoietic specificity to both Runx1 promoters. *Blood*, 113(21):5121–4, 2009. ISSN 1528-0020. doi: 10.1182/blood-2008-12-193003. URL http://www.ncbi.nlm.nih.gov/pubmed/19321859.

C Glenn Begley and Anthony R Green. The SCL gene: from case report to critical hematopoietic regulator. *Blood*, 93(9):2760–2770, 1999.

Yael Ben-Haim, Alexander Ivrii, Oded Margalit, and Arie Matsliah. Perfect hashing and cnf encodings of cardinality constraints. In *Theory and Applications of Satisfiability Testing–SAT 2012*, pages 397–409. Springer, 2012.

Sean C. Bendall, Kara L. Davis, El-ad David Amir, Michelle D. Tadmor, Erin F. Simonds, Tiffany J. Chen, Daniel K. Shenfeld, Garry P. Nolan, and Dana Pe'er. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell*, 157(3):714–725, 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.04.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867414004711.

Josh Berdine, Cristiano Calcagno, and Peter O'Hearn. A Decidable Fragment of Separation Logic. In *In FSTTCS*, pages 97–109, 2004. ISBN 3-540-24058-6, 978-3-540-24058-7. doi: 10.1007/978-3-540-30538-5_9.

Bradley E. Bernstein, Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K. Bailey, Dana J. Huebert, Scott McMahon, Elinor K. Karlsson, Edward J. Kulbokas, Thomas R. Gingeras, Stuart L. Schreiber, and Eric S. Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181, 2005. ISSN 00928674. doi: 10.1016/j.cell.2005.01.001.

Tewodros A Beyene and Andrey Rybalchenko. A Constraint-Based Approach to Solving Games on Infinite Graphs. 2014.

Tewodros A Beyene, Corneliu Popeea, and Andrey Rybalchenko. Solving existentially quantified horn clauses. In *Computer Aided Verification*, pages 869–882. Springer, 2013.

Armin Biere. Yet another local search solver and Lingeling and friends entering the SAT Competition 2014. *SAT Competition*, 2014:2, 2014.

Armin Biere, Alessandro Cimatti, Edmund M. Clarke, Masahiro Fujita, and Yunshan Zhu. Symbolic model checking using SAT procedures instead of BDDs. In *Proceedings of the 36th annual ACM/IEEE Design Automation Conference (DAC´99)*, pages 317–320, 1999. ISBN 1-58113-092-9. doi: 10.1109/DAC.1999.781333. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=781333.

Armin Biere, Daniel Le Berre, Emmanuel Lonca, and Norbert Manthey. Detecting cardinality constraints in CNF. In *Theory and Applications of Satisfiability Testing–SAT 2014*, pages 285–301. Springer, 2014.

Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. 2006. ISBN 9780387310732. doi: 10.1117/1.2819119. URL http://www.library.wisc.edu/selectedtocs/bg0137.pdf.

Nikolaj Bjø rner, Ken McMillan, and Andrey Rybalchenko. On solving universally quantified horn clauses. In *Static Analysis*, pages 105–125. Springer, 2013.

Nikolaj Bjø rner, Anh-Dung Phan, and Lars Fleckenstein. νZ-An Optimizing SMT Solver. In *Tools and Algorithms for the Construction and Analysis of Systems*, pages 194–199. Springer, 2015.

François Bobot, Sylvain Conchon, Evelyne Contejean, Mohamed Iguernelala, Assia Mahboubi, Alain Mebsout, and Guillaume Melquiond. A Simplex-Based Extension of Fourier-Motzkin for Solving Linear Integer Arithmetic. In *Automated Reasoning*, volume 7364, pages 67–81. 2012. doi: 10.1007/978-3-642-31365-3_8. URL http://dx.doi.org/10.1007/978-3-642-31365-3_8.

Charlotta Böiers, Joana Carrelha, Michael Lutteropp, Sidinh Luc, Joanna C.A. Green, Emanuele Azzoni, Petter S. Woll, Adam J. Mead, Anne Hultquist, Gemma Swiers, Elisa Gomez Perdiguero, Iain C. Macaulay, Luca Melchiori, Tiago C. Luis, Shabnam Kharazi, Tiphaine Bouriez-Jones, Qiaolin Deng, Annica Pontén, Deborah Atkinson,

Christina T. Jensen, Ewa Sitnicka, Frederic Geissmann, Isabelle Godin, Rickard Sandberg, Marella F.T.R. de Bruijn, and Sten Eirik W. Jacobsen. Lymphomyeloid Contribution of an Immune-Restricted Progenitor Emerging Prior to Definitive Hematopoietic Stem Cells. *Cell Stem Cell*, 13(5):535–548, 2013. ISSN 19345909. doi: 10.1016/j.stem.2013.08.012. URL http://linkinghub.elsevier.com/retrieve/pii/S1934590913003755.

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

Ingwer Borg and Patrick J.F. Groenen. *Modern Multidimensional Scaling*. 2005. ISBN 9780387981345. doi: 10.1007/0-387-28981-X. URL http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-98134-5?cm_mmc=AD-_-Enews-_-ECS12245_V1-_-978-0-387-98134-5.

Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, David K Gifford, Douglas A Melton, Rudolf Jaenisch, and Richard A Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.08.020. URL http://www.sciencedirect.com/science/article/pii/S0092867405008251.

Martin Brain, Liana Hadarean, Daniel Kroening, and Ruben Martins. Automatic Generation of Propagation Complete SAT Encodings. In *Verification, Model Checking, and Abstract Interpretation*, pages 536–556. Springer, 2016.

Martin Bromberger, Thomas Sturm, and Christoph Weidenbach. Linear Integer Arithmetic Revisited. *arXiv preprint arXiv:1503.02948*, 2015.

James Brotherston and Max Kanovich. Undecidability of propositional separation logic and its neighbours. *Journal of the ACM (JACM)*, 61(2):14, 2014.

James Brotherston, Carsten Fuhs, Juan A Navarro Pérez, and Nikos Gorogiannis. A decision procedure for satisfiability in separation logic with inductive predicates. In *Proceedings of the Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, page 25. ACM, 2014.

Bryant. Graph-Based Algorithms for Boolean Function Manipulation. *IEEE Transactions on Computers*, C-35(8):677–691, 1986. ISSN 0018-9340. doi: 10.1109/TC.1986.1676819.

David Bryder, Derrick J Rossi, and Irving L Weissman. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *The American journal of pathology*, 169(2):338–46, 2006. ISSN 0002-9440. doi: 10.2353/ajpath.2006.060312. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1698791&tool=pmcentrez&rendertype=abstract.

Yosef Buganim, Dina a. Faddah, Albert W. Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L. Klemm, Alexander Van Oudenaarden, and Rudolf Jaenisch.

Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.08.023. URL http://dx.doi.org/10.1016/j.cell.2012.08.023.

Jennifer E F Butler and James T Kadonaga. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & development*, 16(20):2583–2592, 2002. ISSN 0890-9369. doi: 10.1101/gad.1026202.

Atul J Butte and Isaac S Kohane. Relevance Networks: A First Step Toward Finding Genetic Regulatory Networks Within Microarray Data. In *The Analysis of Gene Expression Data*, pages 428–446. 2003. ISBN 0-387-95577-1. doi: 10.1007/b97411. URL http://www.springerlink.com/content/x654845676m42550/$\delimiter"026E30F$nhttp://www.springerlink.com/index/10.1007/b97411.

Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how, and why? *Molecular cell*, 49(5):825–37, 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2013.01.038. URL http://www.ncbi.nlm.nih.gov/pubmed/23473601.

Maurice A Canham, Alexei A Sharov, Minoru S H Ko, and Joshua M Brickman. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS biology*, 8(5):e1000379, 2010. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000379. URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000379.

Michael Carey. The enhanceosome and transcriptional synergy. *Cell*, 92(1):5–8, 1998.

Howard Cedar and Yehudit Bergman. Epigenetics of haematopoietic cell development. *Nature reviews. Immunology*, 11(7):478–488, 2011. ISSN 1474-1733. doi: 10.1038/nri2991. URL http://dx.doi.org/10.1038/nri2991.

Ian Chambers, Jose Silva, Douglas Colby, Jennifer Nichols, Bianca Nijmeijer, Morag Robertson, Jan Vrana, Ken Jones, Lars Grotewold, and Austin Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234, 2007. ISSN 1476-4687. doi: 10.1038/nature06403. URL http://www.ncbi.nlm.nih.gov/pubmed/18097409.

David Chandler and Jerome K. Percus. *Introduction to Modern Statistical Mechanics*, volume 41. 1988. ISBN 019504276X. doi: 10.1063/1.2811680. URL http://scitation.aip.org/content/aip/magazine/physicstoday/article/41/12/10.1063/1.2811680.

Hannah H Chang, Martin Hemberg, Mauricio Barahona, Donald E Ingber, and Sui Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–7, May 2008. ISSN 1476-4687. doi: 10.1038/nature06965. URL http://www.ncbi.nlm.nih.gov/pubmed/18497826.

Chang-Zheng Chen, Ling Li, Harvey F Lodish, and David P Bartel. MicroRNAs modulate hematopoietic lineage differentiation. *Science (New York, N.Y.)*, 303(5654):83–6, 2004. ISSN 1095-9203. doi: 10.1126/science.1091903. URL http://www.ncbi.nlm.nih.gov/pubmed/14657504.

Haifen Chen, Jing Guo, Mishra S Kumar, Paul Robson, Mahesan Niranjan, and Jie Zheng. Single-Cell Transcriptional Analysis to Uncover Regulatory Circuits Driving Cell Fate Decisions in Early Mouse Development. *Bioinformatics (Oxford, England)*, 31(November 2014):1060–1066, 2014a. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu777. URL http://www.ncbi.nlm.nih.gov/pubmed/25416748.

Jiji Chen, Zhengjian Zhang, Li Li, Bi-Chang Chen, Andrey Revyakin, Bassam Hajj, Wesley Legant, Maxime Dahan, Timothée Lionnet, Eric Betzig, Robert Tjian, and Zhe Liu. Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. *Cell*, 156(6):1274–1285, 2014b. ISSN 00928674. doi: 10.1016/j.cell.2014.01.062. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867414001974.

Jingchao Chen. A new SAT encoding of the at-most-one constraint. *Proc. Constraint Modelling and Reformulation*, 2010.

Michael J Chen, Tomomasa Yokomizo, Brandon M Zeigler, Elaine Dzierzak, and Nancy a Speck. Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature*, 457(7231):887–891, 2009. ISSN 0028-0836. doi: 10.1038/nature07619. URL http://dx.doi.org/10.1038/nature07619.

Vijay Chickarmane, Carl Troein, Ulrike a. Nuber, Herbert M. Sauro, and Carsten Peterson. Transcriptional dynamics of the embryonic stem cell switch. *PLoS Computational Biology*, 2(9):1080–1092, 2006. ISSN 1553734X. doi: 10.1371/journal.pcbi.0020123.

Vijay Chickarmane, Tariq Enver, and Carsten Peterson. Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS computational biology*, 5(1):e1000268, 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000268. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000268.

R R Coifman, S Lafon, a B Lee, M Maggioni, B Nadler, F Warner, and S W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7432–7, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0500896102. URL http://www.pnas.org/content/102/21/7432.full.

Stephen a. Cook. The Complexity of Theorem-Proving Procedures. *STOC '71 Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158, 1971. ISSN 08985626. doi: 10.1145/800157.805047.

Guilherme Costa, Valerie Kouskoff, and Georges Lacaud. Origin of blood cells and HSC production in the embryo. *Trends in Immunology*, 33(5):215–223, 2012. ISSN 14714906. doi: 10.1016/j.it.2012.01.012.

Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, Laurie A Boyer, Richard A Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–21936, 2010. ISSN 1091-6490. doi: 10.1073/pnas.1016071107. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3003124&tool=pmcentrez&rendertype=abstract.

Piero Dalerba, Tomer Kalisky, Debashis Sahoo, Pradeep S Rajendran, Michael E Rothenberg, Anne a Leyrat, Sopheak Sim, Jennifer Okamoto, Darius M Johnston, Dalong Qian, Maider Zabala, Janet Bueno, Norma F Neff, Jianbin Wang, Andrew a Shelton, Brendan Visser, Shigeo Hisamori, Yohei Shimono, Marc van de Wetering, Hans Clevers, Michael F Clarke, and Stephen R Quake. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127, 2011. ISSN 1087-0156. doi: 10.1038/nbt.2038. URL http://dx.doi.org/10.1038/nbt.2038.

S. S. Damle and E. H. Davidson. Synthetic in vivo validation of gene network circuitry. *Proceedings of the National Academy of Sciences*, 109:1548–1553, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1119905109.

E. H. Davidson. *The regulatory genome: gene regulatory networks in development and evolution*, volume 310. 2006. ISBN 0120885638. doi: 10.1016/j.ydbio.2007.08.009.

E. H. Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468 (7326):911–920, 2010. ISSN 0028-0836. doi: 10.1038/nature09645. URL http://dx.doi.org/10.1038/nature09645.

Martin Davis, George Logemann, and Donald Loveland. A machine program for theorem-proving. *Commun. ACM*, 5(7):394–397, 1962. ISSN 00010782. doi: 10.1145/368273.368557. URL http://portal.acm.org/citation.cfm?id=368557.

Mark A. Dawson, Tony Kouzarides, and Brian J.P. Huntly. Targeting Epigenetic Readers in Cancer. *New England Journal of Medicine*, 367(7):647–657, 2012. ISSN 0028-4793. doi: 10.1056/NEJMra1112635.

Isil Dillig, Thomas Dillig, and Alex Aiken. Cuts from proofs: A complete and practical technique for solving linear inequalities over integers. In *Computer Aided Verification*, pages 233–247. Springer, 2009.

Guo Ding, Yosuke Tanaka, Misato Hayashi, Shin-ichi Nishikawa, and Hiroshi Kataoka. PDGF receptor alpha+ mesoderm contributes to endothelial and hematopoietic cells in mice. *Developmental dynamics : an official publication of the American Association of Anatomists*, 242(3):254–68, 2013. ISSN 1097-0177. doi: 10.1002/dvdy.23923. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3597973&tool=pmcentrez&rendertype=abstract.

Alexander Dobin, Carrie a Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts635. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3530905&tool=pmcentrez&rendertype=abstract.

Ian J Donaldson, Michael Chapman, Sarah Kinston, Josette Renée Landry, Kathy Knezevic, Sandie Piltz, Noel Buckley, Anthony R Green, and Berthold Göttgens. Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Human molecular genetics*, 14(5):595–601, 2005. ISSN 0964-6906. doi: 10.1093/hmg/ddi056. URL http://www.ncbi.nlm.nih.gov/pubmed/15649946.

L W Donaldson, J M Petersen, B J Graves, and L P McIntosh. Solution structure of the ETS domain from murine Ets-1: a winged helix-turn-helix DNA binding motif. *The EMBO journal*, 15(1):125–134, 1996. ISSN 0261-4189.

Sergei Doulatov, Linda T. Vo, Stephanie S. Chou, Peter G. Kim, Natasha Arora, Hu Li, Brandon K. Hadland, Irwin D. Bernstein, James J. Collins, Leonard I. Zon, and George Q. Daley. Induction of Multipotential Hematopoietic Progenitors from Human Pluripotent Stem Cells via Respecification of Lineage-Restricted Precursors. *Cell Stem Cell*, 13(4):459–470, 2013. ISSN 19345909. doi: 10.1016/j.stem.2013.09.002. URL http://linkinghub.elsevier.com/retrieve/pii/S1934590913004025.

William F Dowling and Jean H Gallier. Linear-time algorithms for testing the satisfiability of propositional Horn formulae. *The Journal of Logic Programming*, 1(3):267–284, 1984.

Sunita L. D'Souza, Andrew G Elefanty, and Gordon Keller. SCL/Tal-1 is essential for hematopoietic commitment of the hemangioblast but not for its development. *Blood*, 105 (10):3862–3870, 2005. ISSN 00064971. doi: 10.1182/blood-2004-09-3611.

Campbell Duff, Kate Smith-Miles, Leo Lopes, and Tianhai Tian. Mathematical modelling of stem cell differentiation: The PU.1-GATA-1 interaction. *Journal of Mathematical Biology*, 64(3):449–468, 2012. ISSN 03036812. doi: 10.1007/s00285-011-0419-3.

S.-J. Dunn, G. Martello, B. Yordanov, S. Emmott, and a. G. Smith. Defining an essential transcription factor program for naive pluripotency. *Science*, 344(6188):1156–1160, 2014. ISSN 0036-8075. doi: 10.1126/science.1248882. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1248882.

Jan Dzik, Nick Palladinos, Konstantinos Rontogiannis, Eirik Tsarpalis, and Nikolaos Vathis. MBrace: cloud computing with monads. In *Proceedings of the Seventh Workshop on Programming Languages and Operating Systems*, page 7. ACM, 2013.

Jack Edmonds. Paths, trees, and flowers, 1965. ISSN 0008-414X.

Niklas Eén and Niklas Sörensson. An Extensible SAT-solver. In *Theory and Applications of Satisfiability Testing*, pages 502–518. 2004. ISBN 978-3-540-20851-8, 978-3-540-24605-3. doi: 10.1007/978-3-540-24605-3_37. URL http://link.springer.com/chapter/10.1007/978-3-540-24605-3_37.

Niklas Eén and Niklas Sorensson. Translating pseudo-boolean constraints into SAT. *Journal on Satisfiability, Boolean Modeling and Computation*, 2:1–26, 2006.

Hanna M Eilken, Shin-Ichi Nishikawa, and Timm Schroeder. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature*, 457(7231):896–900, 2009. ISSN 0028-0836. doi: 10.1038/nature07760. URL http://dx.doi.org/10.1038/nature07760.

Johan Elf, Johan Paulsson, Otto G Berg, and Må ns Ehrenberg. Near-critical phenomena in intracellular metabolite pools. *Biophysical journal*, 84(1):154–170, 2003. ISSN 00063495. doi: 10.1016/S0006-3495(03)74839-5.

M B Elowitz and S Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

Martin Ester, Hans P Kriegel, Jorg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996. ISSN 09758887. doi: 10.1.1.71.1980. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2930.

Andrew G Evans and Laura M Calvi. Notch signaling in the malignant bone marrow microenvironment: implications for a niche-based model of oncogenesis. *Annals of the New York Academy of Sciences*, 1335:63–77, 2015. ISSN 1749-6632. doi: 10.1111/nyas.12562. URL http://www.ncbi.nlm.nih.gov/pubmed/25351294.

Rong Fan, Sabrina Bonde, Peng Gao, Brendan Sotomayor, Changya Chen, Tyler Mouw, Nicholas Zavazava, and Kai Tan. Dynamic HoxB4-regulatory network during embryonic stem cell differentiation to hematopoietic cells. *Blood*, 119(19):e139–47, 2012. ISSN 1528-0020. doi: 10.1182/blood-2011-12-396754. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3362371&tool=pmcentrez&rendertype=abstract.

G Felsenfeld and M Groudiner. Controlling the double helix. *Nature*, 421:444–8, 2003. ISSN 0028-0836. doi: 10.1038/nature01410.

Jasmin Fisher, Ali Sinan Köksal, Nir Piterman, and Steven Woodhouse. Synthesising Executable Gene Regulatory Networks from Single-cell Gene Expression Data. *Computer Aided Verification (CAV), Springer*, 2015.

Samuel D Foster, S Helen Oram, Nicola K Wilson, and Berthold Göttgens. From genes to cells to tissues–modelling the haematopoietic system. *Molecular bioSystems*, 5(12):1413–20, 2009. ISSN 1742-2051. doi: 10.1039/B907225j. URL http://www.ncbi.nlm.nih.gov/pubmed/19763334.

Jenna M. Frame, Kathleen E. McGrath, and James Palis. Erythro-myeloid progenitors: "Definitive" hematopoiesis in the conceptus prior to the emergence of hematopoietic stem cells. *Blood Cells, Molecules, and Diseases*, 51(4):220–225, 2013. ISSN 10799796. doi: 10.1016/j.bcmd.2013.09.006. URL http://dx.doi.org/10.1016/j.bcmd.2013.09.006.

N Friedman, M Linial, I Nachman, and D Pe'er. Using Bayesian networks to analyze expression data. *Journal of computational biology : a journal of computational molecular cell biology*, 7(3-4):601–20, January 2000. ISSN 1066-5277. doi: 10.1089/106652700750050961. URL http://www.ncbi.nlm.nih.gov/pubmed/11108481.

Y Fujiwara, C P Browne, K Cunniff, S C Goff, and S H Orkin. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12355–12358, 1996. ISSN 0027-8424. doi: 10.1073/pnas.93.22.12355.

Jenna L. Galloway, Rebecca a. Wingert, Christine Thisse, Bernard Thisse, and Leonard I. Zon. Loss of Gata1 but not Gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. *Developmental Cell*, 8(1):109–116, 2005. ISSN 15345807. doi: 10.1016/j.devcel.2004.12.001.

J A Garcia-Porrero, I E Godin, and F Dieterlen-Lièvre. Potential intraembryonic hemogenic sites at pre-liver stages in the mouse. *Anatomy and embryology*, 192(5):425–35, 1995. ISSN 0340-2061. doi: 10.1007/BF00240375. URL http://www.ncbi.nlm.nih.gov/pubmed/8546334.

A Garg, A Di Cara, I Xenarios, L Mendoza, and G De Micheli. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics*, 24(17):1917–1925, 2008.

Martin Gebser, Benjamin Kaufmann, André Neumann, and Torsten Schaub. Conflict-driven answer set solving. *IJCAI International Joint Conference on Artificial Intelligence*, 187-188:386–392, 2007a. ISSN 10450823. doi: 10.1016/j.artint.2012.04.001. URL http://dx.doi.org/10.1016/j.artint.2012.04.001.

Martin Gebser, Benjamin Kaufmann, André Neumann, and Torsten Schaub. clasp: A conflict-driven answer set solver. In *Logic Programming and Nonmonotonic Reasoning*, pages 260–265. Springer, 2007b.

Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Answer Set Solving in Practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(3):1–238, 2012. ISSN 1939-4608. doi: 10.2200/S00457ED1V01Y201211AIM019. URL http://www.morganclaypool.com/doi/abs/10.2200/S00457ED1V01Y201211AIM019.

Martin Gebser, Tomi Janhunen, and Jussi Rintanen. Answer set programming as SAT modulo acyclicity. *Front. Artif. Intell. Appl*, 263:351–356, 2014.

D T Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. of Phys. Chemistry*, 81(25):2340–2361, 1977.

Aaron D. Goldberg, C. David Allis, and Emily Bernstein. Epigenetics: A Landscape Takes Shape. *Cell*, 128(4):635–638, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.02.006.

James a Goodrich and Robert Tjian. Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nature reviews. Genetics*, 11(8):549–58, 2010. ISSN 1471-0064. doi: 10.1038/nrg2847. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2965628&tool=pmcentrez&rendertype=abstract.

B Göttgens, L M Barton, J G Gilbert, a J Bench, M J Sanchez, S Bahn, S Mistry, D Grafham, a McMurray, M Vaudin, E Amaya, D R Bentley, a R Green, and a M Sinclair. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nature biotechnology*, 18:181–186, 2000. ISSN 1087-0156. doi: 10.1038/72635.

Berthold Göttgens, Cyril Broccardo, Maria-Jose Sanchez, Sophie Deveaux, George Murphy, Joachim R Göthert, Ekaterini Kotsopoulou, Sarah Kinston, Liz Delaney, Sandie Piltz, Linda M Barton, Kathy Knezevic, Wendy N Erber, C Glenn Begley, Jonathan Frampton, and Anthony R Green. The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. *Molecular and cellular biology*, 24(5):1870–83, 2004. ISSN 0270-7306. doi: 10.1128/MCB.24.5.1870-1883.2004. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=350551&tool=pmcentrez&rendertype=abstract.

Andreas Grönlund, Per Lötstedt, and Johan Elf. Delay-induced anomalous fluctuations in intracellular regulation. *Nature Communications*, 2:419, 2011. ISSN 2041-1723. doi: 10.1038/ncomms1422. URL http://www.nature.com/doifinder/10.1038/ncomms1422.

Andreas Grönlund, Per Lötstedt, and Johan Elf. Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nature Communications*, 4 (May):1864, 2013. ISSN 2041-1723. doi: 10.1038/ncomms2867. URL http://www.nature.com/ncomms/journal/v4/n5/full/ncomms2867.html$\delimiter"026E30F$nhttp://www.nature.com/ncomms/journal/v4/n5/pdf/ncomms2867.pdf.

Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

Guoji Guo, Mikael Huss, Guo Qing Tong, Chaoyang Wang, Li Li Sun, Neil D Clarke, and Paul Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell*, 18(4):675–85, 2010. ISSN 1878-1551. doi: 10.1016/j.devcel.2010.02.012. URL http://www.sciencedirect.com/science/article/pii/S1534580710001103.

Guoji Guo, Sidinh Luc, Eugenio Marco, Ta Wei Lin, Cong Peng, Marc a. Kerenyi, Semir Beyaz, Woojin Kim, Jian Xu, Partha Pratim Das, Tobias Neff, Keyong Zou, Guo Cheng Yuan, and Stuart H. Orkin. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell*, 13(4):492–505, 2013. ISSN 19345909. doi: 10.1016/j.stem.2013.07.017. URL http://dx.doi.org/10.1016/j.stem.2013.07.017.

Carito Guziolowski, Santiago Videla, Federica Eduati, Sven Thiele, Thomas Cokelaer, Anne Siegel, and Julio Saez-Rodriguez. Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming. *Bioinformatics*, 29(18):2320–2326, 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btt393. URL http://bioinformatics.oxfordjournals.org/content/29/18/2320$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/23853063.

L. Haghverdi, F. Buettner, and F. J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, pages btv325–, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv325. URL http://bioinformatics.oxfordjournals.org/content/early/2015/06/27/bioinformatics.btv325.long.

Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673, 2012. ISSN 22111247. doi: 10.1016/j.celrep.2012.08.003. URL http://dx.doi.org/10.1016/j.celrep.2012.08.003.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Number 2. 2009. ISBN 9780387848570. doi: 10.1007/b94608. URL http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf.

Katsuhiko Hayashi, Susana M Chuva de Sousa Lopes, Fuchou Tang, and M Azim Surani. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct

functional and epigenetic states. *Cell Stem Cell*, 3(4):391–401, 2008. ISSN 1875-9777. doi: 10.1016/j.stem.2008.07.027.

David Heckerman. A Tutorial on Learning With Bayesian Networks. *Innovations in Bayesian Networks*, 1995(November):33–82, 1996. ISSN 1860949X. doi: 10.1007/978-3-540-85066-3. URL http://www.springerlink.com/index/62mv333389016034.pdf.

Joe Hendrix and Ben Jones. Bounded Integer Linear Constraint Solving via Lattice Search.

Tracy S P Heng and Michio W Painter. The Immunological Genome Project: networks of gene expression in immune cells. *Nature immunology*, 9(10):1091–4, 2008. ISSN 1529-2916. doi: 10.1038/ni1008-1091. URL http://www.ncbi.nlm.nih.gov/pubmed/18800157.

M. Herold, M. Bartkuhn, and R. Renkawitz. CTCF: insights into insulator function during development. *Development*, 139(6):1045–1057, 2012. ISSN 0950-1991. doi: 10.1242/dev.065268.

Alexander Hertel, Philipp Hertel, and Alasdair Urquhart. Formalizing dangerous SAT encodings. In *Theory and Applications of Satisfiability Testing–SAT 2007*, pages 159–172. Springer, 2007.

Clare Heyworth, Stella Pearson, Gillian May, and Tariq Enver. Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. *EMBO Journal*, 21(14):3770–3781, 2002. ISSN 02614189. doi: 10.1093/emboj/cdf368.

Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12167–12172, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1018832108.

Andreas Hilfinger, Mark Chen, and Johan Paulsson. Using Temporal Correlations and Full Distributions to Separate Intrinsic and Extrinsic Fluctuations in Biological Systems. *Physical Review Letters*, 109(24):248104, 2012. ISSN 0031-9007. doi: 10.1103/PhysRevLett.109.248104. URL http://link.aps.org/doi/10.1103/PhysRevLett.109.248104.

H Hirata, S Yoshiura, T Ohtsuka, Y Bessho, T Harada, K Yoshikawa, and R Kageyama. Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science*, 298(5594):840–843, 2002. ISSN 00368075. doi: 10.1126/science.1074560. URL http://www.ncbi.nlm.nih.gov/pubmed/12399594.

Hanno Hock, Melanie J. Hamblen, Heather M. Rooke, David Traver, Roderick T. Bronson, Scott Cameron, and Stuart H. Orkin. Intrinsic requirement for zinc finger transcription factor Gfi-1 in neutrophil differentiation. *Immunity*, 18(1):109–120, 2003. ISSN 10747613. doi: 10.1016/S1074-7613(02)00501-0.

Sui Huang. Non-genetic heterogeneity of cells in development: more than just noise. *Development (Cambridge, England)*, 136(23):3853–3862, 2009. ISSN 0950-1991. doi: 10.1242/dev.035139.

Sui Huang, G. Eichler, Y. Bar-Yam, and D.E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 94(12):128701, 2005. ISSN 00319007. doi: 10.1103/PhysRevLett.94.128701. URL http://link.aps.org/doi/10.1103/PhysRevLett.94.128701.

Sui Huang, Y.P. Guo, Tariq Enver, and G. May. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, 305(2):695–713, 2007. ISSN 00121606. doi: 10.1016/j.ydbio.2007.02.036. URL http://linkinghub.elsevier.com/retrieve/pii/S0012160607001674.

Jim R Hughes, Nigel Roberts, Simon McGowan, Deborah Hay, Eleni Giannoulatou, Magnus Lynch, Marco De Gobbi, Stephen Taylor, Richard Gibbons, and Douglas R Higgs. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, 46(2):205–212, 2014.

D Huh and J Paulsson. Non-genetic heterogeneity from random partitioning at cell division. *Nature Genetics*, 43(2):95–100, 2011a. doi: 10.1038/ng.729.Non-genetic.

D. Huh and J. Paulsson. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, 108(36):15004–15009, 2011b. ISSN 0027-8424. doi: 10.1073/pnas.1013171108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1013171108.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):1–10, 2010. ISSN 19326203. doi: 10.1371/journal.pone.0012776.

Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167, 2011. ISSN 10889051. doi: 10.1101/gr.110882.110.

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 11(1):163–6, 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2772. URL http://www.ncbi.nlm.nih.gov/pubmed/24363023.

Sorin Istrail and Eric H Davidson. Logic functions of the genomic cis-regulatory code. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (14):4954–4959, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0409624102.

Yoshiaki Ito, Suk-Chul Bae, and Linda Shyue Huey Chuang. The RUNX family: developmental regulators in cancer. *Nature reviews. Cancer*, 15(2):81–95, 2015. ISSN 1474-1768. doi: 10.1038/nrc3877. URL http://dx.doi.org/10.1038/nrc3877.

M. Iwafuchi-Doi and K. S. Zaret. Pioneer transcription factors in cell reprogramming. *Genes & Development*, 28(24):2679–2692, 2014. ISSN 0890-9369. doi: 10.1101/gad.253443.114. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4265672&tool=pmcentrez&rendertype=abstract.

Thierry Jaffredo, Rodolphe Gautier, Anne Eichmann, and Françoise Dieterlen-Lièvre. Intraaortic Hemopoietic Cells are Derived from Endothelial Cells During Ontogeny. *Development*, 125:4575–4583, 1998. ISSN 10501738. doi: 10.1016/j.tcm.2006.02.005.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(February):776–779, 2014. ISSN 1095-9203. doi: 10.1126/science.1247651.

Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome research*, 21(9):1543–51, 2011. ISSN 1549-5469. doi: 10.1101/gr.121095.111. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3166838&tool=pmcentrez&rendertype=abstract.

Dejan Jovanović and Leonardo De Moura. Cutting to the chase solving linear integer arithmetic. In *Automated Deduction–CADE-23*, pages 338–353. Springer, 2011.

Michael Jünger, Denis Naddef, William R. Pulleyblank, Giovanni Rinaldi, Thomas M. Liebling, George L. Nemhauser, Gerhard Reinelt, and Laurence a. Wolsey. 50 years of integer programming 1958-2008: From the early years to the state-of-the-art. *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, pages 1–804, 2010. ISSN 0894069X. doi: 10.1007/978-3-540-68279-0. URL http://link.springer.com/10.1007/978-3-540-68279-0.

Tamar Juven-Gershon and James T. Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology*, 339(2):225–229, 2010. ISSN 00121606. doi: 10.1016/j.ydbio.2009.08.009. URL http://dx.doi.org/10.1016/j.ydbio.2009.08.009.

N Kabrun, H J Bühring, K Choi, a Ullrich, W Risau, and G Keller. Flk-1 expression defines a population of early embryonic hematopoietic precursors. *Development (Cambridge, England)*, 124(10):2039–2048, 1997. ISSN 0950-1991.

James T. Kadonaga. Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology*, 1(1):40–51, 2012. ISSN 17597684. doi: 10.1002/wdev.21.

Michael H Kagey, Jamie J Newman, Steve Bilodeau, Ye Zhan, David A Orlando, Nynke L van Berkum, Christopher C Ebmeier, Jesse Goossens, Peter B Rahl, Stuart S Levine, Dylan J Taatjes, Job Dekker, and Richard A Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010. ISSN 1476-4687. doi: 10.1038/nature09380. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2953795&tool=pmcentrez&rendertype=abstract.

R Kageyama, T Ohtsuka, and T Kobayashi. The Hes gene family: repressors and oscillators that orchestrate embryogenesis. *Development*, 134(7):1243–1251, 2007. ISSN 0950-1991. doi: 10.1242/dev.000786. URL http://www.ncbi.nlm.nih.gov/pubmed/17329370.

Roope Kaivola, Rajnish Ghughal, Naren Narasimhan, Amber Telfer, Jesse Whittemore, Sudhindra Pandav, Anna Slobodová, Christopher Taylor, Vladimir Frolov, Erik Reeber, and Armaghan Naik. Replacing Testing with Formal Verification in Intel^{\scriptsize\circledR} CoreTM i7 Processor Execution Engine Validation. In *Computer Aided Verification*, volume 5643, pages 414–429. 2009. ISBN 978-3-642-02657-7. doi: 10.1007/978-3-642-02658-4_32.

Tibor Kalmar, Chea Lim, Penelope Hayward, Silvia Muñoz Descalzo, Jennifer Nichols, Jordi Garcia-Ojalvo, and Alfonso Martinez Arias. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS biology*, 7(7): e1000149, 2009. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000149. URL http://www.ncbi.nlm.nih.gov/pubmed/19582141.

Guy Karlebach and Ron Shamir. Constructing Logical Models of Gene Regulatory Networks by Integrating Transcription Factor–DNA Interactions with Expression Data: An Entropy-Based Approach. *Journal of Computational Biology*, 19(1):30–41, 2012. ISSN 1066-5277. doi: 10.1089/cmb.2011.0100.

N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984. ISSN 02099683. doi: 10.1007/BF02579150.

P. W. Kasteleyn. Dimer Statistics and Phase Transitions. *Journal of Mathematical Physics*, 4(2):287–293, 1963. ISSN 00222488. doi: doi:10.1063/1.1703953. URL http://jmp.aip.org/resource/1/jmapaq/v4/i2/p287_s1$\delimiter"026E30F$nhttp://link.aip.org/link/?JMAPAQ/4/287/1$\delimiter"026E30F$nhttp://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=JMAPAQ000004000002000287000001&idtype=cvips&doi=10.1063/1.1703953&prog=normal.

Hiroshi Kataoka, Misato Hayashi, Reiko Nakagawa, Yosuke Tanaka, Naoki Izumi, Satomi Shin-ichi Nishikawa, Martin Lars Jakt, Hiroshi Tarui, and Satomi Shin-ichi Nishikawa. Etv2/ER71 induces vascular mesoderm from Flk1+PDGFRα+ primitive mesoderm. *Blood*, 118(26):6975–86, 2011. ISSN 1528-0020. doi: 10.1182/blood-2011-05-352658. URL http://www.ncbi.nlm.nih.gov/pubmed/21911838.

Hiroshi Kataoka, Misato Hayashi, Kumiko Kobayashi, Guo Ding, Yosuke Tanaka, and Shin Ichi Nishikawa. Region-specific Etv2 ablation revealed the critical origin of hemogenic capacity from Hox6-positive caudal-lateral primitive mesoderm. *Experimental Hematology*, 41(6):567–581, 2013. ISSN 0301472X. doi: 10.1016/j.exphem.2013.02.009. URL http://dx.doi.org/10.1016/j.exphem.2013.02.009.

S A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467, 1969.

L Kazemzadeh, M Cvijovic, and D Petranovic. Boolean model of yeast apoptosis as a tool to study yeast and human apoptotic regulations. *Front Physiol*, 3, 2012.

Sepideh Khorasanizadeh. The Nucleosome: From Genomic Organization to Genomic Regulation. *Cell*, 116(2):259–272, 2004. ISSN 00928674. doi: 10.1016/S0092-8674(04)00044-3.

Mark J Kiel and Sean J Morrison. Maintaining hematopoietic stem cells in the vascular niche. *Immunity*, 25(6):862–4, 2006. ISSN 1074-7613. doi: 10.1016/j.immuni.2006.11.005. URL http://www.ncbi.nlm.nih.gov/pubmed/17174928.

Mark J. Kiel, Glenn L. Radice, and Sean J. Morrison. Lack of Evidence that Hematopoietic Stem Cells Depend on N-Cadherin-Mediated Adhesion to Osteoblasts for Their Maintenance. *Cell Stem Cell*, 1(2):204–217, 2007. ISSN 19345909. doi: 10.1016/j.stem.2007.06.001.

Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1778.

Will Klieber and Gihwon Kwon. Efficient CNF encoding for selecting 1 from N objects. In *Proc. International Workshop on Constraints in Formal Verification*, 2007.

Robert J Klose and Adrian P Bird. Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences*, 31(2):89–97, 2006. ISSN 0968-0004. doi: 10.1016/j.tibs.2005.12.008. URL http://www.sciencedirect.com/science/article/pii/S096800040500352X.

Donald E Knuth. The Art of Computer Programming, Volume 4. 2016.

L J Ko and J D Engel. DNA-binding specificities of the GATA transcription factor family. *Molecular and cellular biology*, 13(7):4011–4022, 1993. ISSN 0270-7306. doi: 10.1128/MCB.13.7.4011.Updated.

Taeko Kobayashi and Ryoichiro Kageyama. Hes1 regulates embryonic stem cell differentiation by suppressing Notch signaling. *Genes to Cells*, 15:689–698, 2010. ISSN 13569597. doi: 10.1111/j.1365-2443.2010.01413.x.

Taeko Kobayashi and Ryoichiro Kageyama. Hes1 oscillations contribute to heterogeneous differentiation responses in embryonic stem cells. *Genes*, 2(1):219–28, 2011. ISSN 2073-4425. doi: 10.3390/genes2010219. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3924840&tool=pmcentrez&rendertype=abstract.

Taeko Kobayashi, Hiroaki Mizuno, Itaru Imayoshi, Chikara Furusawa, Katsuhiko Shirahige, and Ryoichiro Kageyama. The cyclic gene Hes1 contributes to diverse differentiation responses of embryonic stem cells. *Genes and Development*, 23(16):1870–1875, 2009. ISSN 08909369. doi: 10.1101/gad.1823109.

Ali Sinan Koksal, Yewen Pu, Saurabh Srivastava, Rastislav Bodik, Jasmin Fisher, and Nir Piterman. Synthesis of biological models from mutation experiments. *Proceedings of the 40th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL '13*, page 469, 2013. ISSN 0362-1340. doi: 10.1145/2429069.2429125. URL http://doi.acm.org/10.1145/2429069.2429125$\delimiter"026E30F$nhttp://dl.acm.org/ft_gateway.cfm?id=2429125&type=pdf$\delimiter"026E30F$nhttp://dl.acm.org/citation.cfm?doid=2429069.2429125.

B Konev and A Lisitsa. A SAT attack on the erdos discrepancy conjecture, 2014. ISSN 16113349. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84904743346&partnerID=40&md5=4263b55c4c7992b003a1ae29bf282787.

Roger D. Kornberg and Yahli Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–294, 1999. ISSN 00928674. doi: 10.1016/S0092-8674(00)81958-3. URL http://www.sciencedirect.com/science/article/pii/S0092867400819583$\delimiter"026E30F$nhttp://ac.els-cdn.com/S0092867400819583/1-s2.0-S0092867400819583-main.pdf?_tid=2cb5db18-c254-11e4-b849-00000aacb35e&acdnat=1425462931_367404aa470e7939c64484c01e2a8d94.

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.02.005. URL http://www.sciencedirect.com/science/article/pii/S0092867407001845.

Smita Krishnaswamy, Matthew H Spitzer, Michael Mingueneau, Sean C Bendall, Oren Litvin, Erica Stone, Dana Pe'er, and Garry P Nolan. Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. *Science (New York, N.Y.)*, 346(6213):1250689, 2014. ISSN 1095-9203. doi: 10.1126/science.1250689. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4334155&tool=pmcentrez&rendertype=abstract.

Jan Krumsiek, Sebastian Poelsterl, Dominik M Wittmann, and Fabian J Theis. Odefy - From discrete to continuous models. *BMC bioinformatics*, 11(1):233, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-233. URL http://www.ncbi.nlm.nih.gov/pubmed/20459647.

Jan Krumsiek, Carsten Marr, Timm Schroeder, and Fabian J. Theis. Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. *PLoS ONE*, 6(8):e22649, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0022649. URL http://dx.plos.org/10.1371/journal.pone.0022649.

Holger Kulessa, Differentiation Programme, and European Molecular. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes and Development*, 9:1250–1262, 1995. ISSN 0890-9369. doi: 10.1101/gad.9.10.1250.

R M Kumar, P Cahan, A K Shalek, R Satija, A J DaleyKeyser, H Li, J Zhang, K Pardee, D Gennert, J J Trombetta, T C Ferrante, A Regev, G Q Daley, and J J Collins. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516(7529):56–61, 2014. ISSN 0028-0836. doi: 10.1038/nature13920. URL http://www.ncbi.nlm.nih.gov/pubmed/25471879.

Michael Kyba, Rita C R Perlingeiro, and George Q. Daley. HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cell and yolk sac hematopoietic progenitors. *Cell*, 109(1):29–37, 2002. ISSN 00928674. doi: 10.1016/S0092-8674(02)00680-3.

Catherine V Laiosa, Matthias Stadtfeld, Huafeng Xie, Luisa de Andres-Aguayo, and Thomas Graf. Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBP alpha and PU.1 transcription factors. *Immunity*, 25(5):731–44,

2006. ISSN 1074-7613. doi: 10.1016/j.immuni.2006.09.011. URL http://www.ncbi.nlm.nih.gov/pubmed/17088084.

Christophe Lancrin, Patrycja Sroczynska, Catherine Stephenson, Terry Allen, Valerie Kouskoff, and Georges Lacaud. The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature*, 457(7231):892–895, 2009. ISSN 0028-0836. doi: 10.1038/nature07679. URL http://dx.doi.org/10.1038/nature07679.

Dirk Landgraf, Burak Okumus, Peter Chien, Tania a Baker, and Johan Paulsson. Segregation of molecules at cell division reveals native protein localization. *Nature Methods*, 9(5): 480–482, 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1955. URL http://dx.doi.org/10.1038/nmeth.1955.

Josette-Renée Landry, Nicolas Bonadies, Sarah Kinston, Kathy Knezevic, Nicola K Wilson, S Helen Oram, Mary Janes, Sandie Piltz, Michelle Hammett, Jacinta Carter, Tina Hamilton, Ian J Donaldson, Georges Lacaud, Jonathan Frampton, George Follows, Valerie Kouskoff, and Berthold Göttgens. Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. *Blood*, 113(23):5783–92, 2009. ISSN 1528-0020. doi: 10.1182/blood-2008-11-187757. URL http://www.ncbi.nlm.nih.gov/pubmed/19171877.

Stefan H Lelieveld, Judith Schütte, Maurits J J Dijkstra, Punto Bawono, Sarah J Kinston, Berthold Göttgens, Jaap Heringa, and Nicola Bonzanni. ConBind: motif-aware cross-species alignment for the identification of functional transcription factor binding sites. *Nucleic acids research*, page gkv1518, 2015.

Ioannis Lestas, Glenn Vinnicombe, and Johan Paulsson. Fundamental limits on the suppression of molecular fluctuations. *Nature*, 467(7312):174–178, 2010. ISSN 0028-0836. doi: 10.1038/nature09333. URL http://dx.doi.org/10.1038/nature09333.

Ruiming Li, Dian Zhou, and Donglei Du. Satisfiability and integer programming as complementary tools. In *Proceedings of the 2004 Asia and South Pacific Design Automation Conference*, pages 879–882. IEEE Press, 2004.

Yi Li, Aws Albarghouthi, Zachary Kincaid, Arie Gurfinkel, and Marsha Chechik. Symbolic optimization with SMT solvers. *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages - POPL '14*, 3(40):607–618, 2014. ISSN 15232867. doi: 10.1145/2535838.2535857. URL http://dl.acm.org/citation.cfm?id=2535857$\delimiter"026E30F$nhttp://dl.acm.org/citation.cfm?doid=2535838.2535857.

H Liang, X Mao, E T Olejniczak, D G Nettesheim, L Yu, R P Meadows, C B Thompson, and S W Fesik. Solution structure of the ets domain of Fli-1 when bound to DNA. *Nature structural biology*, 1(12):871–875, 1994. ISSN 1072-8368. doi: 10.1038/nsb1294-871.

Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–93, 2009. ISSN 1095-9203. doi:

10.1126/science.1181369. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2858594&tool=pmcentrez&rendertype=abstract.

Binbin Lin, Xiaofei He, and Jieping Ye. A geometric viewpoint of manifold learning. In *Applied Informatics*, volume 2, pages 1–12. Springer, 2015.

Kenneth J. Livak, Quin F. Wills, Alex J. Tipping, Krishnalekha Datta, Rowena Mittal, Andrew J. Goldson, Darren W. Sexton, and Chris C. Holmes. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods*, 59 (1):71–79, 2013. ISSN 10462023. doi: 10.1016/j.ymeth.2012.10.004. URL http://dx.doi.org/10.1016/j.ymeth.2012.10.004.

Robert B Lorsbach, Jennifer Moore, Sonny O Ang, Weili Sun, Noel Lenny, and James R Downing. Role of RUNX1 in adult hematopoiesis: analysis of RUNX1-IRES-GFP knock-in mice reveals differential lineage expression. *Blood*, 103(7):2522–9, 2004. ISSN 0006-4971. doi: 10.1182/blood-2003-07-2439. URL http://www.ncbi.nlm.nih.gov/pubmed/14630789.

Stephen J Loughran, Elizabeth a Kruse, Douglas F Hacking, Carolyn a de Graaf, Craig D Hyland, Tracy a Willson, Katya J Henley, Sarah Ellis, Anne K Voss, Donald Metcalf, Douglas J Hilton, Warren S Alexander, and Benjamin T Kile. The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. *Nature immunology*, 9(7):810–819, 2008. ISSN 1529-2908. doi: 10.1038/ni.1617.

K Luger, a W Mäder, R K Richmond, D F Sargent, and T J Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–260, 1997. ISSN 0028-0836. doi: 10.1038/38444.

Christopher T Lux, Momoko Yoshimoto, Kathleen McGrath, Simon J Conway, James Palis, and Mervin C Yoder. All primitive and definitive hematopoietic progenitor cells emerging before E10 in the mouse embryo are products of the yolk sac. *Blood*, 111(7):3435–3438, 2008.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. ISSN 02545330. doi: 10.1007/s10479-011-0841-3.

Ben D. Macarthur and Ihor R. Lemischka. Statistical mechanics of pluripotency. *Cell*, 154(3):484–489, 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.07.024. URL http://dx.doi.org/10.1016/j.cell.2013.07.024.

Ben D. MacArthur, Ana Sevilla, Michel Lenz, Franz-Josef Müller, Berhard M. Schuldt, Andreas a. Schuppert, Sonya J. Ridden, Patrick S. Stumpf, Miguel Fidalgo, Avi Ma'ayan, Jianlong Wang, and Ihor R. Lemischka. Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. *Nature Cell Biology*, 14(11):1139–1147, 2012. ISSN 1465-7392. doi: 10.1038/ncb2603. URL http://dx.doi.org/10.1038/ncb2603.

Iain C Macaulay and Thierry Voet. Single cell genomics: advances and future perspectives. *PLoS genetics*, 10(1):e1004126, 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004126. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3907301&tool=pmcentrez&rendertype=abstract.

J B MacQueen. Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967. URL http://projecteuclid.org/euclid.bsmsp/1200512992.

Bidesh Mahata, Xiuwei Zhang, Aleksandra a. Kolodziejczyk, Valentina Proserpio, Liora Haim-Vilmovsky, Angela E. Taylor, Daniel Hebenstreit, Felix a. Dingler, Victoria Moignard, Berthold Göttgens, Wiebke Arlt, Andrew N J McKenzie, and Sarah a. Teichmann. Single-cell RNA sequencing reveals T helper cells synthesizing steroids De Novo to contribute to immune homeostasis. *Cell Reports*, 7(4):1130–1142, 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.04.011.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Kyle R Allison, Richard Bonneau, Diogo M Camacho, Yukun Chen, James J Collins, Francesca Cordero, James C Costello, Martin Crane, Frank Dondelinger, Mathias Drton, Roberto Esposito, Rina Foygel, Alberto de la Fuente, Jan Gertheiss, Pierre Geurts, Alex Greenfield, Marco Grzegorczyk, Anne-Claire Haury, Benjamin Holmes, Torsten Hothorn, Dirk Husmeier, Vân Anh Huynh-Thu, Alexandre Irrthum, Manolis Kellis, Guy Karlebach, Robert Küffner, Sophie Lèbre, Vincenzo De Leo, Aviv Madar, Subramani Mani, Daniel Marbach, Fantine Mordelet, Harry Ostrer, Zhengyu Ouyang, Ravi Pandya, Tobias Petri, Andrea Pinna, Christopher S Poultney, Robert J Prill, Serena Rezny, Heather J Ruskin, Yvan Saeys, Ron Shamir, Alina Sîrbu, Mingzhou Song, Nicola Soranzo, Alexander Statnikov, Gustavo Stolovitzky, Nicci Vega, Paola Vera-Licona, Jean-Philippe Vert, Alessia Visconti, Haizhou Wang, Louis Wehenkel, Lukas Windhager, Yang Zhang, Ralf Zimmer, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012. ISSN 1548-7091. doi: 10.1038/nmeth.2016.

Joao Marques-Silva, Ines Lynce, and Sharad Malik. Conflict-driven clause learning SAT solvers. *Frontiers in Artificial Intelligence and Applications*, 185(1):131–153, 2009. ISSN 09226389. doi: 10.3233/978-1-58603-929-5-131.

J.P. Marques Silva, K.a. Sakallah, J. P. Marques-Silva, and K.a. Sakallah. GRASP - A New Search Algorithm for Satisfiability. In *Proceedings of International Conference on Computer Aided Design*, pages 220–227, 1996. ISBN 0-8186-7597-7. doi: 10.1109/ICCAD.1996.569607. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=569607.

Jasna Medvedovic, Anja Ebert, Hiromi Tagoh, and Meinrad Busslinger. *Pax5: A Master Regulator of B Cell Development and Leukemogenesis*, volume 111. 2011. ISBN 9780123859914. doi: 10.1016/B978-0-12-385991-4.00005-2. URL http://dx.doi.org/10.1016/B978-0-12-385991-4.00005-2.

Alexander Medvinsky and Elaine Dzierzak. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell*, 86(6):897–906, 1996. ISSN 00928674. doi: 10.1016/S0092-8674(00)80165-8.

Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–70, 2008. ISSN

1476-4687. doi: 10.1038/nature07107. URL http://www.ncbi.nlm.nih.gov/pubmed/18600261.

Menie Merika and Dimitris Thanos. Enhanceosomes. *Current Opinion in Genetics & Development*, 11(2):205–208, 2001. ISSN 0959-437X. doi: 10.1016/S0959-437X(00)00180-5. URL http://www.sciencedirect.com/science/article/pii/S0959437X00001805$\delimiter"026E30F$nhttp://www.sciencedirect.com/science/article/pii/S0959437X00001805/pdfft?md5=3d11f81b4b00c375e3fc7b25ef4582c8&pid=1-s2.0-S0959437X00001805-main.pdf.

David N Messina, Jarret Glasscock, Warren Gish, and Michael Lovett. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome research*, 14(10B):2041–7, 2004. ISSN 1088-9051. doi: 10.1101/gr.2584104. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=528918&tool=pmcentrez&rendertype=abstract.

Hanna K a Mikkola, Yuko Fujiwara, Thorsten M. Schlaeger, David Traver, and Stuart H. Orkin. Expression of CD41 marks the initiation of definitive hematopoiesis in the mouse embryo. *Blood*, 101(2):508–516, 2003. ISSN 00064971. doi: 10.1182/blood-2002-06-1699.

Marcel Mischnik, Stepan Gambaryan, Hariharan Subramanian, Jörg Geiger, Claudia Schütz, Jens Timmer, and Thomas Dandekar. A comparative analysis of the bistability switch for platelet aggregation by logic ODE based dynamical modeling. *Molecular BioSystems*, 10(8):2082–2089, 2014.

Maria Teresa Mitjavila-Garcia, Michel Cailleret, Isabelle Godin, Maria Manuela Nogueira, Karine Cohen-Solal, Valérie Schiavon, Yann Lecluse, Françoise Le Pesteur, Anne Hélène Lagrue, and William Vainchenker. Expression of CD41 on hematopoietic progenitors derived from embryonic hematopoietic cells. *Development (Cambridge, England)*, 129 (8):2003–2013, 2002. ISSN 0950-1991.

Yusuke Miyanari and Maria-Elena Torres-Padilla. Control of ground-state pluripotency by allelic regulation of Nanog. *Nature*, 483(7390):470–473, 2012. ISSN 0028-0836. doi: 10.1038/nature10807. URL http://dx.doi.org/10.1038/nature10807.

Victoria Moignard and Berthold Göttgens. Transcriptional mechanisms of cell fate decisions revealed by single cell expression profiling. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 36(4):419–26, 2014. ISSN 1521-1878. doi: 10.1002/bies.201300102. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3992849&tool=pmcentrez&rendertype=abstract.

Victoria Moignard, Iain C Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J Calero-Nieto, Sarah Kinston, Anagha Joshi, Rebecca Hannah, Fabian J Theis, Sten Eirik Jacobsen, Marella F de Bruijn, and Berthold Göttgens. 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, 2013a. ISSN 1476-4679. doi: 10.1038/ncb2709. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3796878&tool=pmcentrez&rendertype=abstract.

Victoria Moignard, Steven Woodhouse, Jasmin Fisher, and Berthold Göttgens. Transcriptional hierarchies regulating early blood cell development. *Blood Cells, Molecules, and Diseases*, 51(4):239–247, 2013b. ISSN 10799796. doi: 10.1016/j.bcmd.2013.07.007. URL http://dx.doi.org/10.1016/j.bcmd.2013.07.007.

Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, Shin-Ichi Nishikawa, Nir Piterman, Valerie Kouskoff, Fabian J Theis, Jasmin Fisher, and Berthold Göttgens. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33(3): 269–76, 2015. ISSN 1087-0156. doi: 10.1038/nbt.3154. URL http://www.ncbi.nlm.nih.gov/pubmed/25664528.

Rui Monteiro, Claire Pouget, and Roger Patient. The gata1/pu.1 lineage fate paradigm varies between blood populations and is modulated by tif1γ. *The EMBO journal*, 30(6): 1093–1103, 2011. ISSN 0261-4189. doi: 10.1038/emboj.2011.34. URL http://dx.doi.org/10.1038/emboj.2011.34.

M W Moskewicz, C F Madigan, Ying Zhao, Lintao Zhang, and S Malik. Chaff: Engineering an Efficient SAT Solver. In *Design Automation Conference*, pages 530–535, 2001. ISBN 0738-100X VO -. doi: 10.1109/DAC.2001.156196.

N Mossadegh-Keller, S Sarrazin, P K Kandalla, L Espinosa, E R Stanley, S L Nutt, J Moore, and M H Sieweke. M-CSF instructs myeloid lineage fate in single haematopoietic stem cells. *Nature*, 497(7448):239–243, 2013. ISSN 1476-4687. doi: 10.1038/nature12026. URL http://www.ncbi.nlm.nih.gov/pubmed/23575636.

Leonardo De Moura, Bruno Dutertre, and Natarajan Shankar. A Tutorial on Satisfiability Modulo Theories. *Computer Aided Verification*, 4590:20–36, 2007. ISSN 03029743. doi: 10.1007/978-3-540-73368-3_5. URL http://www.springerlink.com/index/10.1007/978-3-540-73368-3$\delimiter"026E30F$nhttp://www.springerlink.com/index/11r20jq883677834.pdf.

James C. Mulloy, Jörg Cammenga, Karen L. MacKenzie, Francisco J. Berguido, Malcolm a S Moore, and Stephen D. Nimer. The AML1-ETO fusion protein promotes the expansion of human hematopoietic stem cells. *Blood*, 99(1):15–23, 2002. ISSN 00064971. doi: 10.1182/blood.V99.1.15.

Boaz Nadler and Meirav Galun. Fundamental limitations of spectral clustering. *Advances in Neural Information Processing Systems 19*, 76100(2):8, 2007. ISSN 1049-5258. doi: 10.1007/BF00668829. URL https://papers.nips.cc/paper/3069-fundamental-limitations-of-spectral-clustering.pdf.

Boaz Nadler, Stephane Lafon, and Ronald Coifman. Diffusion Maps - a Probabilistic Interpretation for Spectral Embedding. *PRINCIPAL MANIFOLDS FOR DATA VISUALIZATION AND DIMENSION REDUCTION*, 10(10):238–260, 2007. ISSN 14397358. doi: 10.1007/978-3-540-73750-6_10. URL http://www.springerlink.com/index/k7131j742231u4q0.pdf.

Jatin Narula, Aileen M. Smith, Berthold Gottgens, and Oleg a. Igoshin. Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate.

PLoS Computational Biology, 6(5):1–16, 2010. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000771.

Felicia S L Ng, Judith Schütte, David Ruau, Evangelia Diamanti, Rebecca Hannah, Sarah J Kinston, and Berthold Göttgens. Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. *Nucleic acids research*, 42(22):13513–13524, 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1254.

T North, T L Gu, T Stacy, Q Wang, L Howard, M Binder, M Marín-Padilla, and N a Speck. Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Development (Cambridge, England)*, 126(11):2563–2575, 1999. ISSN 0950-1991. doi: VL-126.

A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31 (12):i89–i96, 2015a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv257. URL http://bioinformatics.oxfordjournals.org/content/31/12/i89.short.

Andrea Ocone, Laleh Haghverdi, Nikola S. Mueller, and Fabian J. Theis. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96, 2015b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv257. URL http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btv257.

M Ogawa. Differentiation and proliferation of hematopoietic stem cells. *Blood*, 81(11): 2844–53, 1993. ISSN 0006-4971. doi: 10.1182/blood-2001-12-1234. URL http://www.bloodjournal.org/content/81/11/2844.abstract.

Y Okuno, G Huang, F Rosenbauer, E K Evans, H S Radomska, H Iwasaki, K Akashi, F Moreau-Gachelin, Y Li, P Zhang, B Gottgens, and D G Tenen. Potential autoregulation of transcription factor PU.1 by an upstream regulatory element. *Mol Cell Biol*, 25(7):2832–2845, 2005. ISSN 0270-7306. doi: 25/7/2832[pii]$\backslash$r10.1128/MCB.25.7.2832-2845.2005. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15767686$\delimiter"026E30F$nhttp://mcb.asm.org/content/25/7/2832.full.pdf.

Stuart H. Orkin and Leonard I. Zon. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell*, 132(4):631–644, 2008. ISSN 00928674. doi: 10.1016/j.cell.2008.01.025.

Efrat Oron and Natalia Ivanova. Cell fate regulation in early mammalian development. *Physical Biology*, 9:045002, 2012. ISSN 1478-3967. doi: 10.1088/1478-3975/9/4/045002.

Ertugrul M. Ozbudak, Attila Becskei, and Alexander van Oudenaarden. A system of counteracting feedback loops regulates Cdc42p activity during spontaneous cell polarization. *Developmental Cell*, 9(4):565–571, 2005. ISSN 15345807. doi: 10.1016/j.devcel.2005.08.014.

Lior Pachter. A closer look at RNA editing. *Nature Biotechnology*, 30(3):246–247, 2012. ISSN 1087-0156. doi: 10.1038/nbt.2156. URL http://dx.doi.org/10.1038/nbt.2156.

Daniel Panne, Tom Maniatis, and Stephen C Harrison. An atomic model of the interferon-beta enhanceosome. *Cell*, 129(6):1111–23, 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.019. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2020837&tool=pmcentrez&rendertype=abstract.

N Paoletti, B Yordanov, Y Hamadi, C M Wintersteiger, and H Kugler. Analyzing and Synthesizing Genomic Logic Functions. In *Twenty Sixth International Conference on Computer Aided Verification*, 2014.

Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, 344(6190):1396–401, 2014. ISSN 1095-9203. doi: 10.1126/science.1254257. URL http://www.ncbi.nlm.nih.gov/pubmed/24925914.

Rupali P Patwardhan, Choli Lee, Oren Litvin, David L Young, Dana Pe'er, and Jay Shendure. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology*, 27(12):1173–1175, 2009. ISSN 1087-0156. doi: 10.1038/nbt.1589.

Johan Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–8, 2004. ISSN 1476-4687. doi: 10.1038/nature02257. URL http://www.ncbi.nlm.nih.gov/pubmed/14749823.

Johan Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, 2005a. ISSN 15710645. doi: 10.1016/j.plrev.2005.03.003.

Johan Paulsson. Prime movers of noisy gene expression. *Nature genetics*, 37(9):925–926, 2005b. ISSN 1061-4036. doi: 10.1038/ng0905-925.

Juan M Pedraza and Johan Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science (New York, N.Y.)*, 319(5861):339–343, 2008. ISSN 0036-8075. doi: 10.1126/science.1144331.

Carlos-Filipe Pereira, Betty Chang, Jiajing Qiu, Xiaohong Niu, Dmitri Papatsenko, Caroline E. Hendry, Neil R. Clark, Aya Nomura-Kitabayashi, Jason C. Kovacic, Avi Ma'ayan, Christoph Schaniel, Ihor R. Lemischka, and Kateri Moore. Induction of a Hemogenic Program in Mouse Fibroblasts. *Cell Stem Cell*, 13(2):205–218, 2013. ISSN 19345909. doi: 10.1016/j.stem.2013.05.024. URL http://linkinghub.elsevier.com/retrieve/pii/S1934590913002178.

Daniel Peric-Hupkes and Bas van Steensel. Linking Cohesin to Gene Regulation. *Cell*, 132:925–928, 2008. ISSN 00928674. doi: 10.1016/j.cell.2008.03.001.

John M. Perry and Linheng Li. Disrupting the Stem Cell Niche: Good Seeds in Bad Soil. *Cell*, 129(2006):1045–1047, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.05.053.

I S Peter and E H Davidson. A gene regulatory network controlling the embryonic specification of endoderm. *Nature*, 474(7353):635–639, 2011. ISSN 0028-0836. doi: 10.1038/nature10100. URL http://www.ncbi.nlm.nih.gov/pubmed/21623371.

I. S. Peter, E. Faure, and E. H. Davidson. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences*, 109(41):16434–16442, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1207852109.

Jennifer E. Phillips and Victor G. Corces. CTCF: Master Weaver of the Genome, 2009. ISSN 00928674.

Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10:1096–8, 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2639. URL http://www.ncbi.nlm.nih.gov/pubmed/24056875.

J E Pimanda, K Ottersbach, K Knezevic, S Kinston, W Y Chan, N K Wilson, J R Landry, A D Wood, A Kolb-Kokocinski, A R Green, D Tannahill, G Lacaud, V Kouskoff, and B Gottgens. 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, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0707045104. URL http://www.ncbi.nlm.nih.gov/pubmed/17962413.

C Pina, C Fugazza, Aj Tipping, J Brown, S Soneji, J Teles, C Peterson, and T Enver. 17. Inferring rules of lineage commitment in haematopoiesis. *Nature Publishing Group*, 14 (3):287–294, 2012. ISSN 1465-7392. doi: 10.1038/ncb2442. URL http://discovery.ucl.ac.uk/1340480/.

Cristina Pina, José Teles, Cristina Fugazza, Gillian May, Dapeng Wang, Yanping Guo, Shamit Soneji, John Brown, Patrik Edén, Mattias Ohlsson, Carsten Peterson, and Tariq Enver. Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis. *Cell reports*, 11(10):1503–1510, 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.05.016. URL http://www.sciencedirect.com/science/article/pii/S2211124715005288.

Amir Pnueli and Roni Rosner. On the Synthesis of a Reactive Module. In *16th Symposium on Principles of Programming Languages*, pages 179–190. ACM Press, 1989.

Steven Prestwich. Variable dependency in local search: Prevention is better than cure. In *Theory and Applications of Satisfiability Testing–SAT 2007*, pages 107–120. Springer, 2007.

Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, 29(10):886–891, 2011. ISSN 1087-0156. doi: 10.1038/nbt.1991. URL http://dx.doi.org/10.1038/nbt.1991.

Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha a Brugmann, Ryan a Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–83, 2011. ISSN 1476-4687. doi: 10.1038/nature09692. URL http://www.ncbi.nlm.nih.gov/pubmed/21160473.

T Ravasi, H Suzuki, C V Cannistraci, S Katayama, V B Bajic, K Tan, A Akalin, S Schmeier, M Kanamori-Katayama, N Bertin, P Carninci, C O Daub, A R Forrest, J Gough, S Grimmond, J H Han, T Hashimoto, W Hide, O Hofmann, A Kamburov, M Kaur, H Kawaji, A Kubosaki, T Lassmann, E van Nimwegen, C R MacPherson, C Ogawa, A Radovanovic, A Schwartz, R D Teasdale, J Tegner, B Lenhard, S A Teichmann, T Arakawa, N Ninomiya, K Murakami, M Tagami, S Fukuda, K Imamura, C Kai, R Ishihara, Y Kitazume, J Kawai, D A Hume, T Ideker, and Y Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.01.044. URL http://www.ncbi.nlm.nih.gov/pubmed/20211142.

John C. Reynolds. Separation logic: a logic for shared mutable data structures. In *Symposium on Logic in Computer Science*, number 1, pages 55—-74, 2002. ISBN 0-7695-1483-9. doi: 10.1109/LICS.2002.1029817.

Jennifer Rhodes, Andreas Hagen, Karl Hsu, Min Deng, Ting Xi Liu, a Thomas Look, and John P Kanki. Interplay of pu.1 and gata1 determines myelo-erythroid progenitor cell fate in zebrafish. *Developmental cell*, 8(1):97–108, 2005. ISSN 1534-5807. doi: 10.1016/j.devcel.2004.11.014. URL http://www.ncbi.nlm.nih.gov/pubmed/15621533.

Jonah Riddell, Roi Gazit, Brian S. Garrison, Guoji Guo, Assieh Saadatpour, Pankaj K. Mandal, Wataru Ebina, Pavel Volchkov, Guo Cheng Yuan, Stuart H. Orkin, and Derrick J. Rossi. Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. *Cell*, 157(3):549–564, 2014. ISSN 10974172. doi: 10.1016/j.cell.2014.04.006. URL http://dx.doi.org/10.1016/j.cell.2014.04.006.

Michael a Rieger, Philipp S Hoppe, Benjamin M Smejkal, Andrea C Eitelhuber, and Timm Schroeder. Hematopoietic cytokines can instruct lineage choice. *Science (New York, N.Y.)*, 325(5937):217–218, 2009. ISSN 0036-8075. doi: 10.1126/science.1171461.

L Robb, I Lyons, R Li, L Hartley, F Köntgen, R P Harvey, D Metcalf, and C G Begley. Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proceedings of the National Academy of Sciences of the United States of America*, 92 (15):7075–7079, 1995. ISSN 0027-8424. doi: 10.1073/pnas.92.15.7075.

Antony Rodriguez, Elena Vigorito, Simon Clare, Madhuri V Warren, Philippe Couttet, Dalya R Soond, Stijn van Dongen, Russell J Grocock, Partha P Das, Eric A Miska, David Vetrie, Klaus Okkenhaug, Anton J Enright, Gordon Dougan, Martin Turner, and Allan Bradley. Requirement of bic/microRNA-155 for normal immune function. *Science (New York, N.Y.)*, 316(5824):608–11, 2007. ISSN 1095-9203. doi: 10.1126/science.1139253. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2610435&tool=pmcentrez&rendertype=abstract.

Ingo Roeder and Ingmar Glauche. Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. *Journal of theoretical biology*, 241(4):852–65, 2006. ISSN 0022-5193. doi: 10.1016/j.jtbi.2006.01.021. URL http://www.sciencedirect.com/science/article/pii/S0022519306000348.

Philipp Rümmer, Hossein Hojjat, and Viktor Kuncak. Classifying and solving horn clauses for verification. In *Verified Software: Theories, Tools, Experiments*, pages 1–21. Springer, 2014.

Anke Ryll, Regina Samaga, Fred Schaper, Leonidas G. Alexopoulos, and Steffen Klamt. Large-scale network models of IL-1 and IL-6 signalling and their hepatocellular specification. *Molecular BioSystems*, 7(12):3253, 2011. ISSN 1742-206X. doi: 10.1039/c1mb05261f.

A. Mead S. J. Welham, S. A. Gezan, S. J. Clark. *Statistical methods in biology*, volume 104. 1995. ISBN 9780471732877. doi: 10.1002/0471732877.emd318.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas a Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science (New York, N.Y.)*, 308(5721):523–529, 2005. ISSN 0036-8075. doi: 10.1126/science.1105809.

Natalia Sacilotto, Rui Monteiro, Martin Fritzsche, Philipp W Becker, Luis Sanchez-Del-Campo, Ke Liu, Philip Pinheiro, Indrika Ratnayaka, Benjamin Davies, Colin R Goding, Roger Patient, George Bou-Gharios, and Sarah De Val. Analysis of Dll4 regulation reveals a combinatorial role for Sox and Notch in arterial development. *Proceedings of the National Academy of Sciences of the United States of America*, 110 (29):11893–8, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1300805110. URL http://www.pnas.org/content/110/29/11893.full.

Regina Samaga, Julio Saez-Rodriguez, Leonidas G Alexopoulos, Peter K Sorger, and Steffen Klamt. The logic of EGFR/ErbB signaling: theoretical properties and analysis of high-throughput data. *PLoS computational biology*, 5(8):e1000438, 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000438. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000438.

Igor M Samokhvalov, Natalia I Samokhvalova, and Shin-ichi Nishikawa. Cell tracing shows the contribution of the yolk sac to adult haematopoiesis. *Nature*, 446(7139):1056–1061, 2007. ISSN 0028-0836. doi: 10.1038/nature05725.

Manuel Sánchez-Castillo, David Ruau, Adam C Wilkinson, Felicia S L Ng, Rebecca Hannah, Evangelia Diamanti, Patrick Lombard, Nicola K Wilson, and Berthold Gottgens. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic acids research*, 43(Database issue):D1117–23, 2015. ISSN 1362-4962. doi: 10.1093/nar/gku895. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4384009&tool=pmcentrez&rendertype=abstract.

Albin Sandelin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David a Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews. Genetics*, 8(6):424–436, 2007. ISSN 1471-0056. doi: 10.1038/nrg2026.

Sandrine Sarrazin, Noushine Mossadegh-Keller, Taro Fukao, Athar Aziz, Frederic Mourcin, Laurent Vanhille, Louise Kelly Modis, Philippe Kastner, Susan Chan, Estelle Duprez, Claas Otto, and Michael H. Sieweke. MafB Restricts M-CSF-Dependent Myeloid Commitment Divisions of Hematopoietic Stem Cells. *Cell*, 138(2):300–313, 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.04.057. URL http://dx.doi.org/10.1016/j.cell.2009.04.057.

M Schaub, T A Henzinger, and J Fisher. Qualitative Networks: A Symbolic Approach to Analyze Biological Signaling Networks. *{BMC} Systems Biology*, 1(4), 2007.

Koen Schepers, Eric M Pietras, Damien Reynaud, Johanna Flach, Mikhail Binnewies, Trit Garg, Amy J Wagers, Edward C Hsiao, and Emmanuelle Passegué. Myeloproliferative neoplasia remodels the endosteal bone marrow niche into a self-reinforcing leukemic niche. *Cell stem cell*, 13(3):285–99, 2013. ISSN 1875-9777. doi: 10.1016/j.stem.2013.06.009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3769504&tool=pmcentrez&rendertype=abstract.

Thorsten M. Schlaeger, Hanna K a Mikkola, Christos Gekas, Hildur B. Helgadottir, and Stuart H. Orkin. Tie2Cre-mediated gene ablation defines the stem-cell leukemia gene (SCL/tal1)-dependent window during hematopoietic stem-cell development. *Blood*, 105 (10):3871–3874, 2005. ISSN 00064971. doi: 10.1182/blood-2004-11-4467.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998. ISSN 0899-7667. doi: 10.1162/089976698300017467.

Dirk Schübeler, David M. MacAlpine, David Scalzo, Christiane Wirbelauer, Charles Kooperberg, Fred Van Leeuwen, Daniel E. Gottschling, Laura P. O'Neill, Bryan M. Turner, Jeffrey Delrow, Stephen P. Bell, and Mark Groudine. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes and Development*, 18(11):1263–1271, 2004. ISSN 08909369. doi: 10.1101/gad.1198204.

Judith Schütte, Victoria Moignard, and Berthold Göttgens. Establishing the stem cell state: Insights from regulatory network analysis of blood stem cell development. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(3):285–295, 2012. ISSN 19395094. doi: 10.1002/wsbm.1163.

Maria Grazia Scutella. A note on Dowling and Gallier's top-down algorithm for propositional Horn satisfiability. *The Journal of Logic Programming*, 8(3):265–273, 1990.

Roberto Sebastiani. Lazy Satisfiability Modulo Theories. *Journal on Satisfiability, Boolean Modeling and Computation*, 3:141–224, 2007. URL http://jsat.ewi.tudelft.nl/content/volume3/JSAT3_9_Sebastiani_.pdf.

Roberto Sebastiani and Silvia Tomassi. Optimization in SMT with LA(Q) Cost Functions. In *IJCAR 12*, 2012. URL http://arxiv.org/abs/1202.1409.

Vlad C Seitan and Matthias Merkenschlager. Cohesin and chromatin organisation. *Current opinion in genetics & development*, 22(2):93–100, 2012. ISSN 1879-0380. doi: 10.1016/j.gde.2011.11.003. URL http://www.ncbi.nlm.nih.gov/pubmed/22155130.

F Shalaby, J Rossant, T P Yamaguchi, M Gertsenstein, X F Wu, M L Breitman, and a C Schuh. Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice., 1995. ISSN 0028-0836.

F Shalaby, J Ho, W L Stanford, K D Fischer, a C Schuh, L Schwartz, a Bernstein, and J Rossant. A requirement for Flk1 in primitive and definitive hematopoiesis and vasculogenesis. *Cell*, 89(6):981–990, 1997. ISSN 00928674. doi: 10.1016/S0092-8674(00)80283-4.

Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 509(7505): 363–9, 2014. ISSN 1476-4687. doi: 10.1038/nature13437. URL http://www.ncbi.nlm.nih.gov/pubmed/24919153.

Roded Sharan and Richard M Karp. Reconstructing Boolean models of signaling. *Journal of computational biology : a journal of computational molecular cell biology*, 20(3):249–57, 2013. ISSN 1557-8666. doi: 10.1089/cmb.2012.0241. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3590894&tool=pmcentrez&rendertype=abstract.

A. D. Sharrocks. The ETS-domain transcription factor family. *Nature reviews. Molecular cell biology*, 2(11):827–837, 2001. ISSN 1471-0072. doi: 10.1038/35099076.

R A Shivdasani, E L Mayer, and S H Orkin. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*, 373(6513):432–4, 1995. ISSN 0028-0836. doi: 10.1038/373432a0. URL http://www.ncbi.nlm.nih.gov/pubmed/7830794.

Ramesh a. Shivdasani. MicroRNAs: Regulators of gene expression and cell differentiation. *Blood*, 108(12):3646–3653, 2006. ISSN 00064971. doi: 10.1182/blood-2006-01-030015.

Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4):272–86, 2014. ISSN 1471-0064. doi: 10.1038/nrg3682. URL http://www.ncbi.nlm.nih.gov/pubmed/24614317.

Carsten Sinz. Towards an optimal CNF encoding of boolean cardinality constraints. In *Principles and Practice of Constraint Programming-CP 2005*, pages 827–831. Springer, 2005.

Dorothy A Sipkins, Xunbin Wei, Juwell W Wu, Judith M Runnels, Daniel Côté, Terry K Means, Andrew D Luster, David T Scadden, and Charles P Lin. In vivo imaging of specialized bone marrow endothelial microdomains for tumour engraftment. *Nature*, 435 (7044):969–73, 2005. ISSN 1476-4687. doi: 10.1038/nature03703. URL http://www.nature.com/nature/journal/v435/n7044/pdf/nature03703.pdf.

Paul Sjöberg, Per Lötstedt, and Johan Elf. Fokker-Planck approximation of the master equation in molecular biology. *Computing and Visualization in Science*, 12(1):37–50, 2009. ISSN 14329360. doi: 10.1007/s00791-006-0045-6.

Armando Solar-Lezama, Rodric Rabbah, Rastislav Bodík, and Kemal Ebcioğlu. Programming by sketching for bit-streaming programs. *ACM SIGPLAN Notices*, 40:281, 2005. ISSN 03621340. doi: 10.1145/1064978.1065045.

Lingyun Song and Gregory E. Crawford. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 5(2):1–12, 2010. ISSN 15596095. doi: 10.1101/pdb.prot5384.

G J Spangrude, S Heimfeld, and I L Weissman. Purification and characterization of mouse hematopoietic stem cells. *Science (New York, N.Y.)*, 241(4861):58–62, 1988. ISSN 0036-8075. doi: 10.1016/0952-7915(91)90046-4. URL http://www.jstor.org.proxy.library.uu.nl/stable/1701321?seq=1#page_scan_tab_contents.

François Spitz and Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012. ISSN 1471-0056. doi: 10.1038/nrg3207. URL http://dx.doi.org/10.1038/nrg3207.

Matthew H. Spitzer, Pier Federico Gherardini, Gabriela K. Fragiadakis, Nupur Bhattacharya, Robert T. Yuan, Andrew N. Hotson, Rachel Finck, Yaron Carmi, Eli R. Zunder, Wendy J. Fantl, Sean C. Bendall, Edgar G. Engleman, and Garry P. Nolan. An interactive reference framework for modeling a dynamic immune system. *Science (New York, N.Y.)*, 349(6244):1259425, 2015. ISSN 1095-9203. doi: 10.1126/science.1259425. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1259425$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/26160952.

Saurabh Srivastava, Sumit Gulwani, and Jeffrey S. Foster. From program verification to program synthesis. *ACM SIGPLAN Notices*, 45:313, 2010. ISSN 03621340. doi: 10.1145/1707801.1706337.

Saurabh Srivastava, Sumit Gulwani, and JeffreyS. Foster. Template-based program verification and program synthesis. *International Journal on Software Tools for Technology Transfer*, 15(5-6):497–518, 2013. ISSN 1433-2779. doi: 10.1007/s10009-012-0223-4.

Bradford Stadler, Irena Ivanovska, Kshama Mehta, Sunny Song, Angelique Nelson, Yunbing Tan, Julie Mathieu, Christopher Darby, C Anthony Blau, Carol Ware, Garrick Peters, Daniel G Miller, Lanlan Shen, Michele a Cleary, and Hannele Ruohola-Baker. Characterization of microRNAs involved in embryonic stem cell states. *Stem cells and development*, 19(7):935–950, 2010. ISSN 1547-3287. doi: 10.1089/scd.2009.0426.

Anders Stahlberg, Daniel Andersson, Johan Aurelius, Maryam Faiz, Marcela Pekna, Mikael Kubista, and Milos Pekny. Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic acids research*, 39(4):e24, 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1182.

Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics*, 16(January 2014):133–145, 2015. ISSN 1471-0064. doi: 10.1038/nrg3833. URL http://dx.doi.org/10.1038/nrg3833$\delimiter"026E30F$nhttp://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg3833.html#author-information.

Daijiro Sugiyama, Makoto Tanaka, Kenji Kitajima, Jie Zheng, Hilo Yen, Tomotaka Murotani, Atsushi Yamatodani, and Toru Nakano. Differential context-dependent effects of friend of GATA-1 (FOG-1) on mast-cell development and differentiation. *Blood*, 111 (4):1924–1932, 2008. ISSN 00064971. doi: 10.1182/blood-2007-08-104489.

Saulius Sumanas, Gustavo Gomez, Yan Zhao, Changwon Park, Kyunghee Choi, and Shuo Lin. Interplay among Etsrp/ER71, Scl, and Alk8 signaling controls endothelial and myeloid cell formation. *Blood*, 111(9):4500–4510, 2008. ISSN 0006-4971. doi: 10.1182/blood-2007-09-110569.

Miho M Suzuki and Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics*, 9(6):465–76, 2008. ISSN 1471-0064. doi: 10.1038/nrg2341. URL http://www.ncbi.nlm.nih.gov/pubmed/18463664.

Narito Suzuki, Chikara Furusawa, and Kunihiko Kaneko. Oscillatory protein expression dynamics endows stem cells with robust differentiation potential. *PloS one*, 6(11):e27232, January 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0027232. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3207845&tool=pmcentrez&rendertype=abstract.

Gemma Swiers, Claudia Baumann, John O'Rourke, Eleni Giannoulatou, Stephen Taylor, Anagha Joshi, Victoria Moignard, Cristina Pina, Thomas Bee, Konstantinos D Kokkaliaris, Momoko Yoshimoto, Mervin C Yoder, Jon Frampton, Timm Schroeder, Tariq Enver, Berthold Göttgens, and Marella F T R de Bruijn. Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nature communications*, 4:2924, 2013a. ISSN 2041-1723. doi: 10.1038/ncomms3924. URL http://www.ncbi.nlm.nih.gov/pubmed/24326267.

Gemma Swiers, Christina Rode, Emanuele Azzoni, and Marella F.T.R. de Bruijn. A short history of hemogenic endothelium. *Blood Cells, Molecules, and Diseases*, 51(4):206–212, 2013b. ISSN 10799796. doi: 10.1016/j.bcmd.2013.09.005. URL http://linkinghub.elsevier.com/retrieve/pii/S107997961300209X.

Kazutoshi Takahashi and Shinya Yamanaka. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676, 2006. ISSN 00928674. doi: 10.1016/j.cell.2006.07.024.

T Tamura and T Akutsu. An Improved Algorithm for Detecting a Singleton Attractor in a Boolean Network Consisting of AND/OR .... *Lecture Notes in Computer Science*, 2008. URL http://www.springerlink.com/index/l75g6253515637m5.pdf$\delimiter"026E30F$npapers://91dde8cf-0d6e-4036-8d96-0ab5516d506f/Paper/p1304.

Yosuke Tanaka, Misato Hayashi, Yasushi Kubota, Hiroki Nagai, Guojun Sheng, Shin-Ichi Nishikawa, and Igor M Samokhvalov. Early ontogenic origin of the hematopoietic stem cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12):4515–20, 2012. ISSN 1091-6490. doi: 10.1073/pnas.1115828109. URL http://www.pnas.org.proxy.library.uu.nl/content/109/12/4515.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315. URL http://www.ncbi.nlm.nih.gov/pubmed/19349980.

Fuchou Tang, Kaiqin Lao, and M Azim Surani. Development and applications of single-cell transcriptome analysis. *Nature methods*, 8(4 Suppl):S6–S11, 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1557.

Samir Taoudi, Thomas Bee, Adrienne Hilton, Kathy Knezevic, Julie Scott, Tracy a. Willson, Caitlin Collin, Tim Thomas, Anne K. Voss, Benjamin T. Kile, Warren S. Alexander, John E. Pimanda, and Douglas J. Hilton. ERG dependence distinguishes developmental

control of hematopoietic stem cell maintenance from hematopoietic specification. *Genes and Development*, 25:251–262, 2011. ISSN 08909369. doi: 10.1101/gad.2009211.

Daniel P Teufel, Stefan M Freund, Mark Bycroft, and Alan R Fersht. Four domains of p300 each bind tightly to a sequence spanning both transactivation subdomains of p53. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17):7009–14, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0702010104. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1855428&tool=pmcentrez&rendertype=abstract$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17438265.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

To-Ha Thai, Peter a Christiansen, and George C Tsokos. Is there a link between dysregulated miRNA expression and disease? *Discovery medicine*, 10(52):184–94, 2010. ISSN 1944-7930. URL http://www.ncbi.nlm.nih.gov/pubmed/20875339.

Frederic B Thalheimer, Susanne Wingert, Pangrazio De Giacomo, Nadine Haetscher, Maike Rehage, Boris Brill, Fabian J Theis, Lothar Hennighausen, Timm Schroeder, and Michael A Rieger. Cytokine-regulated GADD45G induces differentiation and lineage selection in hematopoietic stem cells. *Stem cell reports*, 3(1):34–43, 2014. ISSN 2213-6711. doi: 10.1016/j.stemcr.2014.05.010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4110750&tool=pmcentrez&rendertype=abstract.

Jean Paul Thiery, Hervé Acloque, Ruby Y J Huang, and M. Angela Nieto. Epithelial-Mesenchymal Transitions in Development and Disease. *Cell*, 139(5):871–890, 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.11.007.

Julie A I Thoms, Yehudit Birger, Sam Foster, Kathy Knezevic, Yael Kirschenbaum, Vashe Chandrakanthan, Georg Jonquieres, Dominik Spensberger, Jason W Wong, S Helen Oram, Sarah J Kinston, Yoram Groner, Richard Lock, Karen L MacKenzie, Berthold Göttgens, Shai Izraeli, and John E Pimanda. ERG promotes T-acute lymphoblastic leukemia and is transcriptionally regulated in leukemic cells by a stem cell enhancer. *Blood*, 117(26):7079–89, 2011. ISSN 1528-0020. doi: 10.1182/blood-2010-12-317990. URL http://www.bloodjournal.org/content/117/26/7079.abstract.

J E Till and E a McCulloch. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation research*, 14(2):213–222, 1961. ISSN 1938-5404. doi: 10.2307/3570892.

Julia Tischler and M Azim Surani. Investigating transcriptional states at single-cell-resolution. *Current opinion in biotechnology*, 24(1):69–78, 2013. ISSN 1879-0429. doi: 10.1016/j.copbio.2012.09.013. URL http://www.ncbi.nlm.nih.gov/pubmed/23084076.

Yayoi Toyooka, Daisuke Shimosato, Kazuhiro Murakami, Kadue Takahashi, and Hitoshi Niwa. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development*, 135(5):909–918, 2008.

Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, 2009. ISSN 1460-2059. doi: 10.1093/bioinformatics/btp120. URL http://www.ncbi.nlm.nih.gov/pubmed/19289445.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–6, 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859. URL http://www.ncbi.nlm.nih.gov/pubmed/24658644.

Leslie G. Valiant. Accidental algorithms. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 509–517, 2006. ISSN 02725428. doi: 10.1109/FOCS.2006.7.

Leslie G Valiant. Holographic algorithms. *SIAM Journal on Computing*, 37(5):1565–1594, 2008. ISSN 01635700. doi: 10.1145/1388240.1388254.

Laurens van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014. ISSN 1532-4435. URL http://jmlr.org/papers/v15/vandermaaten14a.html.

N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. 2007. ISBN 978-0-444-52965-7. doi: 10.1016/B978-044452965-7/50016-7. URL http://www.sciencedirect.com/science/article/pii/B9780444529657500076.

Juan M Vaquerizas, Sarah K Kummerfeld, Sarah a Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–263, 2009. ISSN 1471-0056. doi: 10.1038/nrg2538.

Moshe Vardi. From Verification to Synthesis. *Verified Software: Theories, Tools, Experiments*, page 2, 2008.

Ngan Vo and Richard H Goodman. CREB-binding protein and p300 in transcriptional regulation. *Journal of Biological Chemistry*, 276(17):13505–13508, 2001.

C H Waddington. The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *The strategy of the genes A discussion of some ...*, pages ix +–262 ., 1957. doi: 10.1007/3-540-32786-X_7. URL http://www.cabdirect.org/abstracts/19580101706.html$\delimiter"026E30F$npapers3://publication/uuid/E3C5A230-84D0-4CB4-BA78-4B7DBF839E2C.

Sarah Wareing, Alexia Eliades, Georges Lacaud, and Valerie Kouskoff. ETV2 expression marks blood and endothelium precursors, including hemogenic endothelium, at the onset of blood development. *Developmental Dynamics*, 241(9):1454–1464, 2012. ISSN 10588388. doi: 10.1002/dvdy.23825.

Luigi Warren, David Bryder, Irving L Weissman, and Stephen R Quake. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47):17807–12, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0608512103. URL http://www.ncbi.nlm.nih.gov/pubmed/17098862.

Bartek Wilczynski and E. E M Furlong. Challenges for modeling global gene regulatory networks during development: Insights from Drosophila. *Developmental Biology*, 340 (2):161–169, 2010. ISSN 00121606. doi: 10.1016/j.ydbio.2009.10.032. URL http://dx.doi.org/10.1016/j.ydbio.2009.10.032.

Thomas Wilhelm. The smallest chemical reaction system with bistability. *BMC systems biology*, 3(1):90, 2009. ISSN 1752-0509. doi: 10.1186/1752-0509-3-90. URL http://www.biomedcentral.com/1752-0509/3/90.

Adam C Wilkinson, Debbie K Goode, Yi-Han Cheng, Diane E Dickel, Sam Foster, Tim Sendall, Marloes R Tijssen, Maria-Jose Sanchez, Len a Pennacchio, Aileen M Kirkpatrick, and Berthold Göttgens. Single site-specific integration targeting coupled with embryonic stem cell differentiation provides a high-throughput alternative to in vivo enhancer analyses. *Biology open*, 2(11):1229–38, 2013. ISSN 2046-6390. doi: 10.1242/bio.20136296. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3828770&tool=pmcentrez&rendertype=abstract.

Adam C Wilkinson, Viviane K S Kawata, Judith Schütte, Xuefei Gao, Stella Antoniou, Claudia Baumann, Steven Woodhouse, Rebecca Hannah, Yosuke Tanaka, Gemma Swiers, Victoria Moignard, Jasmin Fisher, Shimauchi Hidetoshi, Marloes R Tijssen, Marella F T R de Bruijn, Pentao Liu, and Berthold Göttgens. Single-cell analyses of regulatory network perturbations using enhancer-targeting TALEs suggest novel roles for PU.1 during haematopoietic specification. *Development (Cambridge, England)*, 141(20):4018–30, 2014. ISSN 1477-9129. doi: 10.1242/dev.115709. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4197694&tool=pmcentrez&rendertype=abstract.

D J Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman and Hall CRC, 2 edition, 2012.

Nicola K Wilson, Diego Miranda-Saavedra, Sarah Kinston, Nicolas Bonadies, Samuel D Foster, Fernando Calero-Nieto, Mark a Dawson, Ian J Donaldson, Stephanie Dumon, Jonathan Frampton, Rekin's Janky, Xiao-Hong Sun, Sarah a Teichmann, Andrew J Bannister, and Berthold Göttgens. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood*, 113 (22):5456–5465, 2009. ISSN 0006-4971. doi: 10.1182/blood-2009-01-200048.

Nicola K. Wilson, Samuel D. Foster, Xiaonan Wang, Kathy Knezevic, Judith Schütte, Polynikis Kaimakis, Paulina M. Chilarska, Sarah Kinston, Willem H. Ouwehand, Elaine Dzierzak, John E. Pimanda, Marella F T R De Bruijn, and Berthold Göttgens. Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, 7(4):532–544, 2010. ISSN 19345909. doi: 10.1016/j.stem.2010.07.016.

Nicola K. Wilson, David G. Kent, Florian Buettner, Mona Shehata, Iain C. Macaulay, Fernando J. Calero-Nieto, Manuel Sánchez Castillo, Caroline a. Oedekoven, Evangelia Diamanti, Reiner Schulte, Chris P. Ponting, Thierry Voet, Carlos Caldas, John Stingl, Anthony R. Green, Fabian J. Theis, and Berthold Göttgens. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell*, 16(6):712–724, 2015. ISSN 19345909. doi: 10.1016/j.stem.2015.04.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1934590915001629.

S. N. Wontakal, X. Guo, C. Smith, T. MacCarthy, E. H. Bresnick, a. Bergman, M. P. Snyder, S. M. Weissman, D. Zheng, and a. I. Skoultchi. A core erythroid transcriptional network is repressed by a master regulator of myelo-lymphoid differentiation. *Proceedings of the National Academy of Sciences*, 109(10):3832–3837, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1121019109.

Steven Woodhouse, Victoria Moignard, Berthold Göttgens, and Jasmin Fisher. Processing, visualising and reconstructing network models from single-cell data. *Immunology and cell biology*, 2015.

Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26(7):873–81, 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq057. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2844994&tool=pmcentrez&rendertype=abstract.

Changchun Xiao, Dinis Pedro Calado, Gunther Galler, To Ha Thai, Heide Christine Patterson, Jing Wang, Nikolaus Rajewsky, Timothy P. Bender, and Klaus Rajewsky. MiR-150 Controls B Cell Differentiation by Targeting the Transcription Factor c-Myb. *Cell*, 131 (1):146–159, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.07.021.

Huafeng Xie, Min Ye, Ru Feng, and Thomas Graf. Stepwise reprogramming of B cells into macrophages. *Cell*, 117(5):663–676, 2004. ISSN 00928674. doi: 10.1016/S0092-8674(04)00419-2.

Huilei Xu, Yen-Sin Ang, Ana Sevilla, Ihor R. Lemischka, and Avi Ma'ayan. Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells. *PLoS Computational Biology*, 10(8):e1003777, 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003777. URL http://dx.plos.org/10.1371/journal.pcbi.1003777.

Z Xue, K Huang, C Cai, L Cai, C Y Jiang, Y Feng, Z Liu, Q Zeng, L Cheng, Y E Sun, J Y Liu, S Horvath, and G Fan. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597, 2013. ISSN 1476-4687. doi: 10.1038/nature12364. URL http://www.ncbi.nlm.nih.gov/pubmed/23892778.

Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, Jin Huang, Ming Li, Xinglong Wu, Lu Wen, Kaiqin Lao, Ruiqiang Li, Jie Qiao, and Fuchou Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–9, 2013. ISSN 1545-9985. doi: 10.1038/nsmb.2660. URL http://www.ncbi.nlm.nih.gov/pubmed/23934149.

Jin Y Yen. Finding the K Shortest Loopless Paths in a Network. *Management Science*, 17 (11):712–716, 1971.

Hiroki Yoshihara, Fumio Arai, Kentaro Hosokawa, Tetsuya Hagiwara, Keiyo Takubo, Yuka Nakamura, Yumiko Gomei, Hiroko Iwasaki, Sahoko Matsuoka, Kana Miyamoto, Hiroshi Miyazaki, Takao Takahashi, and Toshio Suda. Thrombopoietin/MPL Signaling Regulates Hematopoietic Stem Cell Quiescence and Interaction with the Osteoblastic Niche. *Cell Stem Cell*, 1(6):685–697, 2007. ISSN 19345909. doi: 10.1016/j.stem.2007.10.020.

Momoko Yoshimoto, Encarnacion Montecino-Rodriguez, Michael J Ferkowicz, Prashanth Porayette, W Christopher Shelley, Simon J Conway, Kenneth Dorshkind, and Mervin C Yoder. Embryonic day 9 yolk sac and intra-embryonic hemogenic endothelium independently generate a B-1 and marginal zone progenitor lacking B-2 potential. *Proceedings of the National Academy of Sciences of the United States of America*, 108:1468–1473, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1015841108.

Momoko Yoshimoto, Prashanth Porayette, Nicole L. Glosson, Simon J. Conway, Nadia Carlesso, Angelo a. Cardoso, Mark H. Kaplan, and Mervin C. Yoder. Autonomous murine T-cell progenitor production in the extra-embryonic yolk sac before HSC emergence. *Blood*, 119(24):5706–5714, 2012. ISSN 00064971. doi: 10.1182/blood-2011-12-397489.

Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25(21):2227–2241, 2011. ISSN 1549-5477. doi: 10.1101/gad.176826.111. URL http://genesdev.cshlp.org/content/25/21/2227.full.

Amit Zeisel, Ana B Muñoz Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015. ISSN 1095-9203. doi: 10.1126/science.aaa1934. URL http://www.ncbi.nlm.nih.gov/pubmed/25700174$\delimiter"026E30F$nhttp://www.sciencemag.org/cgi/doi/10.1126/science.aaa1934.

Hantao Zhang. SATO: An efficient prepositional prover. In *Automated Deduction—CADE-14*, pages 272–275. Springer, 1997.

Jiwang Zhang, Chao Niu, Ling Ye, Haiyang Huang, Xi He, Wei-Gang Tong, Jason Ross, Jeff Haug, Teri Johnson, Jian Q. Feng, Stephen Harris, Leanne M. Wiedemann, Yuji Mishina, and Li Linheng. Identification of the haematopoietic stem cell niche and control of the niche size. *Nature*, 425:836–841, 2003. ISSN 1476-4687. doi: 10.1038/nature02064.1.

Desheng Zheng, Guowu Yang, Xiaoyu Li, Zhicai Wang, Feng Liu, and Lei He. An Efficient Algorithm for Computing Attractors of Synchronous And Asynchronous Boolean Networks. *PLOS ONE*, 2013.

Tianshou Zhou, Luonan Chen, and Kazuyuki Aihara. Molecular communication through stochastic synchronization induced by extracellular fluctuations. *Physical Review Letters*, 95(17):2–5, 2005. ISSN 00319007. doi: 10.1103/PhysRevLett.95.178103.

Eli R Zunder, Ernesto Lujan, Yury Goltsev, Marius Wernig, and Garry P Nolan. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell stem cell*, 16(3):323–37, 2015. ISSN 1875-9777. doi: 10.1016/j.stem.2015.01.015. URL http://www.sciencedirect.com/science/article/pii/S1934590915000168.

Appendix A

Supporting information for chapter 4 — Synthesised Boolean update rules

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)	Motifs present
Scl	*Fli1*	98	Yes
Etv2	*Notch1*	96	Yes
Fli1	*Etv2*	96	Yes
	Sox7	97	Yes
Lyl1	*Sox7*	92	Yes
Sox7	$Sox17 \vee HoxB4$	82	No (Sox missing)
Erg	$(HoxB4 \wedge Lyl1) \vee Sox17$	84	Yes
	$(HoxB4 \wedge Tal1) \vee Sox17$	83	Yes
Notch1	*Sox7*	94	Yes
Gata1	$Gfi1b \wedge Lmo2$	86	Yes
	$Gfi1b \wedge Hhex$	84	No (Hhex missing)
	$Gfi1b \wedge Ets1$	84	Yes
HoxB4	$(Lyl1 \wedge Ets1) \wedge \neg Gata1$	65	Yes
	$(Lyl1 \vee Nfe2) \wedge \neg Gata1$	65	Yes
	$(Lyl1 \vee Ikaros) \wedge \neg Gata1$	65	No (Ikaros missing)
Sox17	$Lyl1 \wedge \neg Gfi1b$	77	No (Gfi missing)
	$(Eto2 \wedge Sox7) \wedge \neg Gfi1b$	76	No (Gfi missing)
	$(Eto2 \wedge Tal1) \wedge \neg Gfi1b$	75	No (Gfi missing)
Ets1	*Notch1*	96	Yes
Gfi1	$Gata1 \wedge \neg Sox17$	88	Yes
	$Nfe2 \wedge \neg Sox17$	88	Yes
Gfi1b	$Nfe2 \wedge Myb$	87	Yes
	$Pu.1 \wedge Ikaros$	86	No (Ikaros missing)
	$Pu.1 \wedge Nfe2$	86	Yes
	$Pu.1 \wedge Myb$	86	Yes
Eto2	*Sox7*	93	No (Sox missing)
	Hhex	92	No (Hhex missing)
	$Ets1 \wedge Fli1$	94	No (Ets missing)
Hhex	*Sox7*	97	No (Sox missing)
	Notch1	93	No (Rbpj missing)
Ikaros	$Nfe2 \vee Gfi1b$	84	Yes
	$Nfe2 \vee Gata1$	83	Yes
	$Nfe2 \vee Gfi1$	82	Yes
Lmo2	$Sox7 \vee Gfi1$	79	Yes
	$Sox7 \vee Erg$	79	Yes
	$Sox7 \vee HoxB4$	77	Yes
Nfe2	*Ikaros*	78	Yes
Pu.1	$Gfi1 \vee Erg$	67	Yes
Myb	*HoxB4*	64	Yes

Appendix B

Supporting information for chapter 4 — Results of repeating synthesis with a more stringent discretisation threshold

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Scl	*Scl*	100
	Fli1	99
	Sox7	99
Etv2	*Sox7*	98
	Notch1	94
Fli1	*Scl*	99
	Ets1	99
	Sox7	98
	Etv2	98
	Hhex	97
Lyl1	*Fli1*	93
	Eto2	93
	Scl	93
	Ets1	93
	Sox7	91
	Hhex	91
	Etv2	91

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Sox7	*Notch1*	95
	Sox17 $\vee$ ***HoxB4***	96
Erg	*Sox7*	87
	Notch1	84
	HoxB4 $\vee$ *Sox17*	85
Notch1	***Sox7***	96
	Scl	96
	Ets1	96
	Fli1	96
	Etv2	96
Gata1	*Gfi1b*	86
	Gfi1	87
HoxB4	*Lyl1*	74
	Erg	74
Sox17	***Lyl1*** $\wedge \neg$ ***Gfi1b***	75
Ets1	*Sox7*	99
	Notch1	94
Gfi1	*Gata1*	87
	Gata1 $\wedge \neg$ ***Sox17***	87
	Nfe2 $\wedge \neg$ ***Sox17***	87
	HoxB4 $\wedge \neg$ *Notch1*	87
	Gfi1b $\wedge \neg$ *Sox17*	86
Gfi1b	*Gata1*	86
	Nfe2	82
	Gfi1	81
Eto2	*Fli1*	97
	Tal1	96
	Ets1	96
	Sox7	94
	Hhex	94
	Etv2	94
Hhex	*Scl*	99
	Fli1	99

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Ets1	99
	Sox7	98
	Etv2	97
Ikaros	***Nfe2*** $\vee$ ***Gfi1b***	83
	Gata1 $\vee$ *Gfi1b*	80
	Myb $\vee$ *Gfi1b*	80
	Gfi1 $\vee$ *Gfi1b*	80
	Myb $\wedge$ *Eto2*	80
	Myb $\vee$ *Lyl1*	81
Lmo2	*Sox7*	72
Nfe2	*Gata1*	84
	Gfi1b	84
	Gfi1	80
	Ikaros	73
Pu.1	*Ikaros*	75
	Gfi1b	74
Myb	*Ikaros*	79
	Gfi1b	75

Appendix C

Supporting information for chapter 4 — Results of repeating synthesis with multiple rounds of bootstrapping (A-E)

A

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Scl	*Hhex*	96
	Sox7	98
	Etv2	98
	Fli1	99
	Ets1	100
	Scl	100
Etv2	No solution	
Fli1	***Etv2***	98
	Sox7	98
	Notch1	98
Lyl1	*Etv2*	90
	Notch1	91
	Sox7	92
Sox7	***Sox17*** $\vee$ ***HoxB4***	85
Erg	*Sox7* $\vee$ *HoxB4*	90

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Sox7 ∨ *Notch1*	89
	Sox7	89
	Notch1	89
	Sox7 ∧ *Notch1*	88
	Sox17 ∨ *HoxB4*	84
Notch1	No solution	
Gata1	*Gfi1*	89
HoxB4	*Lyl1*	76
Sox17	***Lyl1*** ∧ ¬***Gfi1b***	78
Ets1	*Sox7*	98
	Notch1	99
Gfi1	*Gfi1b* ∧ ¬*Sox17*	88
	Nfe2 ∧ ¬*Sox17*	90
	Gata1	90
	Gata1 ∧ ¬***Sox17***	91
Gfi1b	*Gata1*	87
	Nfe2 ∧ ***Myb***	87
	Nfe2 ∧ *Ikaros*	86
	Pu.1 ∧ ***Nfe2***	86
	Sox7 ∧ *Gata1*	86
	Pu.1 ∧ *Myb*	85
Eto2	*Ets1*	95
	Sox7	94
	Hhex	93
	Notch1	93
	Etv2	93
Hhex	***Sox7***	98
	Notch1	98
Ikaros	***Nfe2*** ∨ ***Gfi1b***	81
Lmo2	*Notch1*	79
	Sox7	79
Nfe2	***Ikaros***	75
	Gfi1	81

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Gfi1b	83
	Gata1	84
Pu.1	*Erg*	71
	Sox7 $\wedge$ *Eto2*	71
	Lyl1 $\wedge$ *Erg*	71
	Notch1 $\wedge$ *Eto2*	71
	Gfi1 $\vee$ ***Erg***	72
	HoxB4 $\vee$ *Erg*	72
	Erg $\wedge$ *Eto2*	73
	Ikaros $\wedge$ *Erg*	73
	Notch1 $\wedge$ *Ikaros*	75
	Sox7 $\wedge$ *Ikaros*	75
Myb	***HoxB4***	60
	Gfi1	67

B

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Scl	*Hhex*	96
	Notch1	96
	Etv2	96
	Ets1	97
	Sox7	98
	Fli1	98
Etv2	No solution	
Fli1	*Notch1*	96
	Etv2	96
	Sox7	98
Lyl1	*Erg*	84
	Notch1	87
	Etv2	87
	Sox7	91

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Sox7	*Sox17* ∨ *Erg*	90
	HoxB4 ∨ *Erg*	88
	Gfi1 ∨ *Erg*	85
	Sox17 ∨ ***HoxB4***	84
	Scl ∧ *Erg*	82
	Fli1 ∧ *Erg*	82
	Erg	82
Erg	*Notch1* ∧ *Etv2*	81
	Sox17 ∨ ***HoxB4***	83
	Sox7 ∧ *Notch1*	85
Notch1	No solution	
Gata1	*Gfi1*	84
	Gfi1b ∧ *Ets1*	85
	Lmo2 ∧ ***Gfi1b***	87
HoxB4	*Lyl1*	75
Sox17	***Lyl1*** ∧ ¬***Gfi1b***	78
Ets1	*Sox7*	95
	Notch1	97
Gfi1	*Gata1*	86
	Nfe2 ∧ ¬***Sox17***	88
	Gfi1b ∧ ¬*Sox17*	88
	Gata1 ∧ ¬***Sox17***	88
	Gfi1b ∧ ¬*Erg*	87
	Gata1 ∧ ¬*Erg*	86
	Nfe2 ∧ ¬*Hhex*	86
	Gata1 ∧ ¬*Hhex*	86
	Ikaros ∧ ¬*Hhex*	86
	Nfe2 ∧ ¬*Erg*	86
Gfi1b	*Gfi1*	81
	Pu.1 ∧ ***Nfe2***	88
	Nfe2 ∧ ***Myb***	89
Eto2	***Sox7***	94
	Hhex	93

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Ets1	92
	Etv2	91
	Notch1	91
Hhex	***Notch1***	94
	Sox7	97
Ikaros	*Myb* $\wedge$ *Eto2*	80
	Nfe2 $\vee$ ***Gata1***	81
	Nfe2 $\vee$ ***Gfi1***	82
	Nfe2 $\vee$ ***Gfi1b***	83
Lmo2	*Sox7*	78
Nfe2	***Ikaros***	77
	Gfi1	80
	Gata1	83
	Gfi1b	86
Pu.1	*Nfe2* $\wedge$ *Erg*	67
	HoxB4 $\wedge$ *Gfi1b*	67
	Myb $\wedge$ *Erg*	67
	Sox7 $\wedge$ *Myb*	67
	Erg $\wedge$ *Eto2*	68
	Gfi1 $\vee$ ***Erg***	68
	Lyl1 $\wedge$ *Erg*	68
	Notch1 $\wedge$ *Nfe2*	68
	Tbx20 $\wedge$ *Gfi1b*	68
	Sox7 $\wedge$ *Nfe2*	69
	Gfi1b $\wedge$ *Erg*	72
	Ikaros $\wedge$ *Erg*	73
	Notch1 $\wedge$ *Gfi1b*	74
	Sox7 $\wedge$ *Gfi1b*	75
	Notch1 $\wedge$ *Ikaros*	75
	Sox7 $\wedge$ *Ikaros*	75
Myb	***HoxB4***	65
	Gfi1	66

C

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Scl	*Hhex*	97
	Notch1	97
	Sox7	98
	Etv2	98
	Fli1	99
	Ets1	100
	Scl	100
Etv2	No solution	
Fli1	*Notch1*	97
	Etv2	97
	Sox7	98
Lyl1	*Notch1*	90
	Sox7	92
Sox7	***Sox17*** $\vee$ ***HoxB4***	85
Erg	*Sox7*	88
	Notch1	89
	Sox17 $\vee$ *HoxB4*	85
Notch1	No solution	
HoxB4	*Lyl1*	76
Sox17	***Lyl1*** $\wedge \neg$ ***Gfi1b***	75
	Erg $\wedge \neg$*Gfi1b*	74
	Fli1 $\wedge \neg$*Gfi1b*	74
	Eto2 $\wedge \neg$*Gfi1b*	74
	Sox7 $\wedge \neg$*Gfi1b*	74
	Lyl1 $\wedge \neg$*Gata1*	73
Ets1	***Notch1***	98
	Sox7	98
Gfi1	*Gfi1b* $\wedge \neg$*Sox17*	86
	Gata1	87
	Nfe2 $\wedge \neg$ ***Sox17***	87
	Gata1 $\wedge \neg$ ***Sox17***	89
Gfi1b	*Gfi1* $\wedge$ *Eto2*	80

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Ikaros $\wedge$ *Gfi1*	80
	Nfe2 $\wedge$ *Ets1*	82
	Notch1 $\wedge$ *Gata1*	82
	Myb $\wedge$ *Ikaros*	83
	Gata1 $\wedge$ *Etv2*	83
	Pu.1 $\wedge$ ***Nfe2***	83
	Sox7 $\wedge$ *Gata1*	83
	Pu.1 $\wedge$ *Gata1*	84
	Pu.1 $\wedge$ ***Ikaros***	84
	Pu.1 $\wedge$ ***Myb***	84
	Gata1 $\wedge$ *Ets1*	85
	Myb $\wedge$ *Gata1*	85
	Nfe2 $\wedge$ ***Myb***	85
Eto2	*Ets1*	96
	Sox7	95
	Hhex	94
	Etv2	94
	Notch1	93
Hhex	***Notch1***	96
	Sox7	97
Ikaros	*Myb*	80
	Myb $\wedge$ *Ets1*	80
	Myb $\wedge$ *Fli1*	80
	Scl $\wedge$ *Myb*	80
	Myb $\wedge$ *Lyl1*	80
	Myb $\vee$ *Gfi1*	80
	Myb $\vee$ *Gata1*	81
	Myb $\vee$ *Gfi1b*	81
	Nfe2 $\vee$ *Myb*	81
	Myb $\wedge$ *Eto2*	82
	Nfe2 $\vee$ ***Gata1***	82
	Nfe2 $\vee$ ***Gfi1***	82
	Nfe2 $\vee$ ***Gfi1b***	83

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Lmo2	*Sox7*	79
	Notch1	78
Nfe2	***Ikaros***	77
	Gfi1	79
	Gfi1b	83
	Gata1	84
Pu.1	*Erg*	67
	Scl $\wedge$ *Erg*	67
	Ets1 $\wedge$ *Erg*	67
	Fli1 $\wedge$ *Erg*	67
	Notch1 $\wedge$ *Eto2*	68
	HoxB4 $\vee$ *Erg*	68
	Sox7 $\wedge$ *Eto2*	68
	Erg $\wedge$ *Eto2*	69
	Gfi1b $\wedge$ *Erg*	69
	Notch1 $\wedge$ *Lyl1*	69
	Gfi1 $\vee$ ***Erg***	70
	Ikaros $\wedge$ *Erg*	70
	Lyl1 $\wedge$ *Erg*	70
	Notch1 $\wedge$ *Gfi1b*	70
	Sox7 $\wedge$ *Lyl1*	70
	Sox7 $\wedge$ *Gfi1b*	71
	Notch1 $\wedge$ *Ikaros*	72
	Sox7 $\wedge$ *Ikaros*	73
Myb	*Gfi1*	64
	HoxB4	62
	Erg	62

D

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Scl	*Etv2*	96

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Notch1	96
	Ets1	96
	Hhex	96
	Fli1	97
	Sox7	97
Etv2	No solution	
Fli1	*Notch1*	96
	Etv2	96
	Sox7	98
Lyl1	*Notch1*	86
	Erg	88
	Sox7	91
Sox7	$Lmo2 \vee HoxB4$	89
	$Sox17 \vee Erg$	89
	$HoxB4 \vee Erg$	87
	$Gfi1 \vee Erg$	85
	$Sox17 \vee Lmo2$	85
	$Scl \wedge Erg$	83
	$Fli1 \wedge Erg$	84
	Erg	83
	$\boldsymbol{Sox17 \vee HoxB4}$	83
Erg	*Sox7*	86
	Notch1	82
	$Sox7 \vee Gfi1$	86
	$Sox17 \vee HoxB4$	87
	$Sox7 \wedge Fli1$	87
	$Notch1 \wedge Eto2$	88
	$Sox7 \wedge Eto2$	89
	$Notch1 \wedge Lyl1$	90
	$Sox7 \wedge Lyl1$	90
Notch1	No solution	
Gata1	*Gfi1b*	85
	Gfi1	88

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
HoxB4	*Lyl1* ∧ ¬*Gfi1*	64
Sox17	***Lyl1*** ∧ ¬***Gfi1b***	78
	Erg ∧ ¬*Gfi1b*	76
	Lyl1 ∧ ¬*Gata1*	75
	Erg ∧ ¬*Gata1*	74
	Eto2 ∧ ¬*Gfi1b*	73
	Lyl1 ∧ ¬*Myb*	72
	Hhex ∧ ¬*Gfi1b*	72
	Sox7 ∧ ¬*Gfi1b*	72
	Erg ∧ ¬*Gfi1*	71
	Lyl1 ∧ ¬*Gfi1*	71
Ets1	*Sox7*	94
	Notch1	97
Gfi1	*Gata1*	88
	Gata1 ∧ ¬***Sox17***	89
Gfi1b	*Gata1*	85
	Nfe2	84
	Nfe2 ∨ *Gfi1*	85
	Nfe2 ∨ *Gata1*	86
	Pu.1 ∧ ***Nfe2***	86
	Myb ∧ *Gata1*	86
	Pu.1 ∧ ***Ikaros***	86
	Nfe2 ∧ *Ikaros*	87
	Pu.1 ∧ ***Myb***	87
	Nfe2 ∧ ***Myb***	88
	Gfi1 ∨ *Gata1*	88
Eto2	*Fli1*	93
	Sox7	92
	Scl	92
	Hhex	92
	Lyl1	92
	Ets1	90
Hhex	***Notch1***	94

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Sox7	97
Ikaros	*Myb* ∨ *Gfi1b*	80
	Myb ∧ *Lyl1*	81
	Nfe2 ∨ ***Gata1***	81
	Myb ∧ *Eto2*	82
	Nfe2 ∨ ***Gfi1b***	83
Lmo2	*Sox7*	76
	Erg	72
Nfe2	*Myb*	70
	Ikaros	76
	Gfi1	81
	Gata1	85
	Gfi1b	86
Pu.1	*Erg*	69
	Gfi1b ∧ *Erg*	74
	Notch1 ∧ *Gfi1b*	75
	Sox7 ∧ *Gfi1b*	75
Myb	*Erg*	62
	Gfi1	64
	HoxB4	66

E

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
Scl	*Lyl1*	96
	Hhex	97
	Eto2	97
	Notch1	97
	Etv2	98
	Sox7	98
	Ets1	100
	Fli1	100

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Scl	100
Etv2	***Notch1***	96
	Sox7	98
Fli1	*Meis1*	90
	Notch1	97
	Etv2	98
	Sox7	98
Lyl1	*Hhex*	95
	Notch1	95
	Sox7	97
Sox7	***Sox17*** $\vee$ ***HoxB4***	88
Erg	*Notch1*	91
	Sox7	90
Notch1	No solution	
Gata1	*Gfi1*	85
HoxB4	*Sox17* $\vee$ *Lmo2*	67
	Eto2 $\wedge \neg$ *Gfi1*	59
	Lyl1 $\wedge \neg$ *Gfi1*	59
	Fli1 $\wedge \neg$ *Gfi1*	59
	Scl $\wedge \neg$ *Gfi1*	59
	Lmo2	57
Sox17	*Fli1* $\wedge \neg$ *Gfi1b*	79
	Sox7 $\wedge \neg$ *Gfi1b*	79
	Scl $\wedge \neg$ *Gfi1b*	79
	Ets1 $\wedge \neg$ *Gfi1b*	79
	Etv2 $\wedge \neg$ *Gfi1b*	79
	Notch1 $\wedge \neg$ *Gfi1b*	79
	Hhex $\wedge \neg$ *Gfi1b*	78
	Eto2 $\wedge \neg$ *Gfi1b*	77
	Lyl1 $\wedge \neg$ ***Gfi1b***	75
Ets1	***Notch1***	97
	Sox7	98
Gfi1	*Gata1*	87

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Nfe2 $\wedge \neg$ ***Sox17***	87
	Gata1 $\wedge \neg$ ***Sox17***	88
Gfi1b	*Notch1* $\wedge$ *Gata1*	81
	Nfe2 $\wedge$ *Etv2*	81
	Sox7 $\wedge$ *Nfe2*	81
	Gata1 $\wedge$ *Etv2*	82
	Sox7 $\wedge$ *Gata1*	82
	Pu.1 $\wedge$ *Gata1*	82
	Nfe2 $\wedge$ *Ets1*	83
	Gata1 $\wedge$ *Ets1*	84
	Myb $\wedge$ *Ikaros*	84
	Myb $\wedge$ *Gata1*	84
	Pu.1 $\wedge$ ***Ikaros***	84
	Pu.1 $\wedge$ ***Myb***	85
	Pu.1 $\wedge$ ***Nfe2***	85
	Nfe2 $\wedge$ ***Myb***	86
Eto2	*Ets1*	99
	Sox7	97
	Etv2	97
	Notch1	96
	Hhex	96
Hhex	***Notch1***	96
	Sox7	98
Ikaros	*Myb* $\wedge$ *Lyl1*	80
	Myb $\vee$ *Gata1*	80
	Myb $\vee$ *Gfi1*	80
	Myb $\vee$ *Mitf*	80
	Myb $\vee$ *Gfi1b*	81
	Nfe2 $\vee$ *Myb*	81
	Nfe2 $\vee$ ***Gfi1b***	82
Lmo2	*Sox7*	78
	Notch1	78
	Notch1 $\vee$ *Erg*	79

Continued on next page

continued from previous page

Gene	Synthesised update functions	% Non-observed transitions disallowed (N_i)
	Sox7 $\vee$ ***Gfi1***	79
	Sox7 $\vee$ ***HoxB4***	79
	Sox7 $\vee$ ***Erg***	79
	Sox7 $\vee$ *Notch1*	79
Nfe2	*Myb*	71
	Ikaros	74
	Gfi1	80
	Gata1	82
	Gfi1b	84
Pu.1	*Erg*	68
	Sox7	67
Myb	*Erg*	62
	Gfi1	63
	HoxB4	65