

Boolean Modeling of Developmental Gene Networks in T Cell Progenitor Differentiation

by

Matthew Langley

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science
Institute of Biomaterials and Biomedical Engineering
University of Toronto

© Copyright by Matthew Langley 2018

Boolean Modeling of Developmental Gene Networks in T Cell Progenitor Differentiation

Matthew Langley

Master of Applied Science

Institute of Biomaterials and Biomedical Engineering
University of Toronto

2018

Abstract

T cell lineage differentiation of hematopoietic progenitors is controlled via gene regulatory networks. We apply computational Boolean network (BN) modeling to simulate systems-level dynamics of this developmental program. Asynchronous Boolean simulations mapped the transcriptional space that is accessible to T cell progenitors under combinations of Notch, interleukin-7 and pre-T cell receptor signaling. Simulations also predict steady states that correspond to known T cell progenitor types and multiple distinct trajectories that can lead to these steady states. Heterogeneous transcriptional dynamics and trajectories were explored further by single-cell transcriptomics to elucidate differences between *in vivo* thymopoiesis and novel *in vitro* differentiation platforms. Finally, our BN modeling framework was integrated into a systems biology software platform to facilitate future extensions to the model. Overall, BN modeling presents a powerful advancement over existing static models for predicting heterogeneous transcriptional responses of T cell progenitors to extrinsic signals during development and *in vitro* differentiation.

Acknowledgments

I would like to thank my supervisor, Prof. Peter Zandstra, both for his expert guidance and for sharing his contagious enthusiasm for research. His scientific and academic mentorship has been invaluable, and his leadership consistently inspires me to become a better scientist.

I would also like to thank my committee members—Profs. Gary Bader, Sid Goyal, and Michele Anderson—for their feedback and encouragement throughout my studies.

I consider myself very fortunate to have had outstanding mentors within the Zandstra Lab. I would especially like to thank Dr. Shreya Shukla for sharing her deep knowledge of T cell biology, her perennial positivity, and her wisdom about finding balance and meaning as a scientist. I am also especially grateful to Dr. Ayako Yachie-Kinoshita, who has been an exceptional mentor in computational biology and was instrumental in making my research exchange in Tokyo a reality.

I would also like to extend special thanks to:

- Ting Yin and Dr. Cynthia Fisher; for their technical support and ensuring I had the materials I needed for my experiments
- Mike Hughes; for his assistance with fetal mouse dissections
- Neil Winegarden, Gurbaksh Basi, Dan Trcka, Ryan Vander Werff, and Tara Stach; for their assistance with single-cell experiments
- Andy Johnson and the staff of the UHN-SickKids Flow Cytometry Facility; for their cell sorting expertise
- Prof. Ellen Rothenberg; for valuable discussions and advice on my Boolean network model during conferences and for sharing her group's pioneering work toward identifying the key genetic elements of the T cell development program
- All members of the Zandstra Lab; for their technical support, helpful comments, and friendship

I would not have made it this far without the steadfast love of my best friend and partner Alice Zhang, my parents Mark and Dinah Langley, and my brother Ryan Langley. Your unwavering support makes all that I do possible, and I am a stronger scientist and person because of you. Thank you for helping me navigate the challenging process of graduate school—I look forward to seeing what new adventures the future has in store!

I have been personally supported by a Canada Graduate Scholarship (CGS-M) from the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was also supported by funding from the Canadian Institutes of Health Research (CIHR), the Stem Cell Network (SCN), Medicine by Design (MbD), and the Institute of Biomaterials and Biomedical Engineering (IBBME) at the University of Toronto.

Table of Contents

Acknowledgments.....	iii
Table of Contents.....	v
List of Figures.....	ix
List of Tables.....	xi
List of Abbreviations.....	xii
1 Introduction.....	1
1.1 Using computation to understand cell fate decisions.....	2
1.2 Gene regulatory networks (GRNs).....	3
1.2.1 GRN inference methods.....	4
1.2.2 Computational simulation of GRNs.....	7
1.2.3 Boolean networks.....	9
1.3 T cell lineage.....	11
1.3.1 Stages of <i>in vivo</i> thymopoiesis.....	12
1.3.2 Environmental signals in the thymic niche.....	15
1.3.3 <i>In vitro</i> methods for T cell lineage differentiation.....	17
1.3.4 Gene regulatory networks for T cell development.....	19
1.3.5 Plasticity and potential shortcuts in T cell development program.....	21
1.3.6 Alternative strategies for computational modeling of T cell development GRNs.....	22
1.3.7 Open questions in T cell development field.....	23
2 Objectives & Aims.....	25
2.1 Hypothesis.....	26
2.2 Specific Aims.....	26

3	Construction of a Boolean network model of T cell development	28
3.1	Introduction.....	29
3.2	Methods.....	30
3.2.1	Literature curation process.....	30
3.2.2	Partial correlation analysis	31
3.2.3	Iterative refinement of Boolean update functions.....	31
3.2.4	Asynchronous Boolean simulation	32
3.2.5	Identification of steady states and strongly connected components	34
3.2.6	<i>In silico</i> genetic knockouts and forced expression	34
3.2.7	Comparison of BN simulations and binarized experimental data	34
3.2.8	Trajectory identification and clustering from asynchronous Boolean simulations	35
3.3	Results.....	36
3.3.1	Construction of BN model of mouse T cell development	36
3.3.2	Steady states of BN correspond to known T cell progenitor states	39
3.3.3	BN simulation captures control of developmental progression by environmental signals	41
3.3.4	BN modeling predicts developmental effect of knockouts.....	43
3.3.5	BN modeling predicts multiple transcriptional trajectories toward T cell commitments	44
3.4	Discussion.....	46
4	Community software for Boolean network modeling.....	50
4.1	Introduction.....	51
4.2	Software	52
4.3	Applications	54
5	Differentiation context-dependent comparison of T cell progenitor transcriptional patterns	57

5.1	Introduction.....	58
5.2	Methods.....	59
5.2.1	Primary tissue dissection.....	59
5.2.2	<i>In vitro</i> differentiation of fetal liver HSPCs toward T cell lineage.....	60
5.2.3	Live imaging of differentiating T cell progenitors	61
5.2.4	Bulk quantitative real-time PCR.....	61
5.2.5	Single-cell qRT-PCR	62
5.2.6	Single-cell RNA-sequencing	62
5.2.7	Flow cytometry	63
5.3	Results.....	64
5.3.1	Characterization of T cell development gene expression dynamics and cell motility during DL4+VCAM differentiation	64
5.3.2	Surface marker-defined stages obscure transcriptional heterogeneity among primary thymocytes.....	72
5.3.3	Single-cell RNA sequencing reveals transcriptional differences between primary and <i>in vitro</i> differentiated T cell progenitors	76
5.4	Discussion.....	83
6	Future Work	87
6.1	Single-cell transcriptomics analysis of T cell progenitor differentiation.....	88
6.1.1	Trajectory inference from single-cell transcriptomics data	88
6.1.2	Boolean network refinement using single-cell transcriptomics.....	92
6.2	Transcriptional memory in T cell progenitor-derived induced pluripotent stem cells	97
6.2.1	T cell progenitor-derived iPSCs exhibit molecular and functional pluripotency	98
6.2.2	T cell development genes are expressed atypically in T cell progenitor-derived iPSCs.....	99

6.2.3	Toward integration of BN models of T cell development and pluripotent fate transitions	102
6.3	Toward multi-scale models of T cell development.....	105
7	Conclusions	108
7.1	Thesis novelty and impact	109
	References.....	110
	Copyright Acknowledgments	123
	Appendix A: Software Resources	124
	Appendix B: Supplementary Tables	131

List of Figures

Figure 1: Overview of mouse T cell development program	13
Figure 2: Notch-driven coherent feedforward loop forms a network motif within T cell development GRN	20
Figure 3: Overview of dynamic Boolean network modeling approach	33
Figure 4: Construction of BN model of mouse T cell development.....	37
Figure 5: Known T cell progenitor cell types are captured by BN modeling.....	40
Figure 6: Density plots of predicted transcriptional state space that is accessible to CLPs when stimulated with various combinations of environmental signals	42
Figure 7: <i>In silico</i> knockout of <i>Bcl11b</i> recapitulates experimentally observed developmental arrest at DN2 stage	43
Figure 8: BN modeling predicts multiple transcriptional trajectories toward T cell lineage commitment	45
Figure 9: A schematic of the defined PSC gene/signal regulatory network model	55
Figure 10: Transcriptional state spaces of mouse ESC network resulting from random asynchronous Boolean simulation	56
Figure 11: DL4+VCAM yields robust mouse T cell progenitor differentiation with accelerated kinetics	67
Figure 12: DL4+VCAM differentiated mouse T cell progenitors express Notch target genes at higher levels than DL4-only	69
Figure 13: DL4+VCAM differentiated human T cell progenitors express Notch target genes at higher levels than DL4-only	70

Figure 14: T cell progenitors exhibit greater motility on DL4+VCAM vs. DL4 only	71
Figure 15: Single-cell qRT-PCR analysis of transcriptional heterogeneity in ETP, DN2A, and DN2B T cell progenitors	75
Figure 16: Summary of single-cell RNA-sequencing experiment design	79
Figure 17: DL4+VCAM differentiated FL HSPCs and primary fetal thymocytes occupy distinct transcriptional spaces	82
Figure 18: Donor cell memory in induced pluripotent stem cells and proposed role of GRN feedback.....	97
Figure 19: T cell progenitor-derived iPSCs exhibit molecular and functional pluripotency	100
Figure 20: iPSCs derived from different T cell progenitor stages express T cell development genes at atypical levels and are transcriptionally distinguishable.....	101
Figure 21: Summary of reported interactions between T cell development and pluripotency GRNs	103
Figure 22: "Discretize" gadget for <i>Garuda</i> pipeline.....	125
Figure 23: "Boolean Simulation" gadget for <i>Garuda</i> pipeline	127
Figure 24: "Boolean SCC Analysis" gadget for <i>Garuda</i> pipeline	129

List of Tables

Table 1: Comparison of GRN simulation methods.....	8
Table 2: Description of T cell progenitor stages and mature T cell types	14
Table 3: Key regulatory inputs to the T cell development program	19
Table 4: Boolean logic functions for BN model of T cell development program	38
Table 5: Summary of <i>Garuda</i> gadgets for Boolean network analysis.....	53
Table 6: Single-cell RNA-sequencing sample metrics	78
Table 7: Methods for trajectory inference from single-cell transcriptomics data.....	90
Supplementary Table 1: Literature evidence for edges in Boolean network (BN) model of mouse T cell development	131
Supplementary Table 2: Microarray datasets used for partial correlation analysis	141
Supplementary Table 3: Primer sequences used for qRT-PCR	156

List of Abbreviations

7-AAD	7-aminoactinomycin D
ABM	agent-based model
BN	Boolean network
CD	cluster of differentiation
ChIP	chromatin immunoprecipitation
ChIP-seq	ChIP followed by DNA sequencing
CRISPR	clustered regularly interspaced short palindromic repeats
DMEM	Dulbecco's Modified Eagle Medium
DMSO	dimethyl sulfoxide
ESC	embryonic stem cell
FBS	fetal bovine serum
GRN	gene regulatory network
HSPCs	hematopoietic stem and progenitor cells
IL	interleukin
IMDM	Iscove's Modified DMEM
mRNA	messenger ribonucleic acid
qRT-PCR	quantitative real-time PCR
R-ABS	random asynchronous Boolean simulation
RNA	ribonucleic acid
RNA-seq	RNA sequencing
PCR	polymerase chain reaction
scRNA-seq	single-cell RNA sequencing
SMT	satisfiability modulo theory
TCR	T cell receptor
TF	transcription factor

1 Introduction

1.1 Using computation to understand cell fate decisions

Biological complexity and specialization are achieved through the coordinated activity of many individual biological components. At the level of individual cell types, specialized transcriptional states and phenotypic functions are achieved by large sets of genes and the proteins they encode. These components interact with each other to form gene regulatory networks (GRNs). Cell fate decisions—in which a cell transitions from one cell fate to another—require a transition between different GRN states. In these cases, the activity of any one gene within the network provides insufficient insight into the cell fate decision process; rather, it is crucial to consider the interactions between all members of the complete network.

Our ability to identify the GRNs underlying cell fate decisions has greatly increased in recent years with the advent of new experimental technologies. For example, biologists can now apply microarrays, RNA sequencing, single-cell transcriptomics, chromatin profiling, knockout models, RNA interference, and CRISPR to elucidate GRNs.

However, despite this vast amount of data, understanding cell fate decision making through systems-level analysis of GRNs continues to be challenging due to the complexity of these networks. Given that GRNs are typically large and frequently comprise interdependent and non-additive regulatory interactions, it can be difficult to develop an intuitive sense of how any given GRN functions from only a static description of its component interactions. Computational modeling is uniquely well-positioned to enable biologists to simultaneously consider the actions of all genes within a particular network and map how specialized cell types and functions emerge from their combined activity.

Herein we examine the challenges of modeling cell fate decisions in the context of one complex yet important biological system: mouse T cell development. We identify open questions pertaining to T cell lineage specification that are amenable to computational investigation and outline how GRN simulations—in particular, Boolean network modeling—can address these questions.

1.2 Gene regulatory networks (GRNs)

Cells make fate decisions by interpreting signals from their environment and activating specific transcriptional programs in response. Gene regulatory networks (GRNs) act as the molecular circuitry that enables this computation inside single cells. Therefore, understanding how cell fate decisions are made first requires knowledge of the genes involved in these circuits, how the genes within these circuits regulate one another, and how input signals from the environment are connected to these circuits.

GRNs are collections of genes and their products which interact with each other to regulate gene expression and, by extension, control cell fate and function. GRNs operate within single cells and define the set of potential transcriptional states that the cell can access. Indeed, by steering gene expression toward a small subset of states within the potential global transcriptional state space, GRNs naturally give rise to distinct cell types that correspond to attractors in this space (Dealy et al., 2005; Huang et al., 2005).

Specific environmental and developmental contexts provide additional layers of regulatory control such as biochemical signals, biomechanical forces, and epigenetic modifications. These factors converge onto the GRN to further constrain the subset of transcriptional states that are realized by the cell. The state of a cell's GRN dictates its cellular identity and determines its phenotypic behaviour by driving the expression or repression of function-associated genes and proteins. For example, the GRN state of a stem cell determines whether it will differentiate or not; the GRN state of a developing lymphocyte determines whether it will begin gene rearrangement of its T cell receptor; and the GRN state of a mature immune cell determines whether it will expand to combat an immune challenge.

By studying cell fate decisions from the perspective of GRNs, we can uncover systems-level properties of these decisions that are not apparent at the level of individual genes. This knowledge can then be harnessed to serve many research purposes. For example, GRN models can be used to create a causal map of molecular interactions and predict the phenotypic effect of signaling changes and genetic perturbations (Emmert-Streib et al., 2014; Xiao, 2009). They can also be used to identify biomarkers of cell types and

diseases (Marbach et al., 2016; Ng et al., 2016). GRN models can also be harnessed to improve *in vitro* differentiation and reprogramming protocols, thereby enabling us to produce therapeutically-relevant cell types at large scales and accelerate efforts in regenerative medicine (Cahan et al., 2014; McNamara et al., 2015). Moreover, since disruption of normal GRN function can lead to disease states such as cancer, a deeper understanding of these regulatory mechanisms can enable new therapeutic interventions to rescue perturbed networks *in vivo* to a healthy state (Karlebach and Shamir, 2008; Mohanty et al., 2014). In particular, GRN control of stem cell fate has been extensively studied in recent years (Dunn et al., 2014; Morris et al., 2014; Yachie-Kinoshita et al., 2018). However, our collective understanding of how GRNs integrate multiple dynamic inputs to make cell fate decisions remains incomplete.

1.2.1 GRN inference methods

Mathematically, a GRN can be conceptualized as a directed graph in which nodes represent individual genes and edges represent the regulatory interactions between genes. It is possible to infer such graphical models of GRNs from gene expression data, direct evidence of molecular binding, knockout and overexpression experiments, and many other experimental modalities (Emmert-Streib et al., 2014).

In some cases, it is possible to construct an accurate GRN model solely through a systematic search of previously published literature for evidence of relevant gene regulatory interactions. Literature evidence may be in the form of binding evidence, such as chromatin immunoprecipitation sequencing (ChIP-seq) analysis of cis-regulatory elements; perturbation evidence, such as genetic knockout models followed by transcriptional profiling; or more likely, a combination of both. Literature-based GRN models have previously been reported in the contexts of myeloid differentiation (Krumisiek et al., 2011), hematopoietic stem and progenitor cell heterogeneity (Bonzanni et al., 2013), myeloid versus lymphoid fate choice (Collombet et al., 2017), heart field specification (Herrmann et al., 2012), and many others.

Literature-based GRN models perform well when the vast majority of important regulatory interactions are believed to have already been reported, which may well be the case for mouse T cell development (Kueh and Rothenberg, 2012; Longabaugh et al., 2017; Rothenberg et al., 2016). However, there is a risk of biasing the resulting computational model toward subnetworks of the complete GRN that happen to be well-studied. In such cases where a literature-based GRN model fails to recapitulate some experimental observations (and assuming the underlying assumptions of the simulation framework are valid), it is often possible to identify genes whose regulatory inputs were underspecified through simulation, and even predict which transcription factors may act as inputs but which remain unreported. Thus, even an incomplete GRN model constructed from available literature evidence may be useful for generating new hypotheses about missing regulatory interactions and guiding further experiments to characterize the true network.

A variety of bioinformatics approaches can be used to either supplement literature-based GRN models or infer GRN models without any direct literature input. Indeed, a burgeoning number of computational algorithms for inference of GRNs from experimental data have been developed in parallel to the increase in available experimental technologies for generating data about GRNs. Efforts such as the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project have benchmarked the performance of many published algorithms against synthetic and high-confidence control GRNs (Marbach et al., 2010). However, it is widely accepted that no single algorithm or paradigm consistently performs best across all test datasets; rather, certain algorithm types perform better on certain classes of datasets and experimental systems (Emmert-Streib et al., 2014). Furthermore, there is increasing evidence that GRN inference accuracy can be improved by applying multiple algorithm types and finding consensus (Marbach et al., 2012).

GRN inference algorithms can typically be classified into a small set of paradigms (Huynh-Thu and Sanguinetti, 2018), which are summarized below. In correlation-based GRN inference methods, regulatory edges are ranked by variants of correlation. In regression-based GRN inference methods, regulatory edges are typically selected by

target gene-specific linear regression, often accompanied by some form of data resampling. In mutual information-based GRN inference methods, regulatory edges are ranked by variants of mutual information and subsequently filtered for causal relationships. Bayesian network methods have also been developed that infer GRNs by heuristically optimizing posterior probabilities of regulatory interactions. Finally, machine learning-enabled GRN inference approaches have also been used; for example, the GENIE3 algorithm employs random forests to predict target gene expression and selects transcription factors as nodes if they reduce the variance of the target gene (Huynh-Thu et al., 2010).

One interesting alternative to the aforementioned statistical GRN inference approaches is satisfiability modulo theory (SMT). An SMT problem involves determining if a logical formula can be satisfied given a set of prior constraints. In the case of biological GRNs, one can formulate an SMT problem as follows: given a set of experimental constraints and a set of possible gene regulatory interactions, what is the full set of logical networks that are able to satisfy the experimental constraints? In this manner, SMT-based approaches are unique in that they do not seek to produce a single logical model of a GRN that best fits an experimental dataset; rather, they identify sets of possible logical models that (mathematically) satisfy all of the provided experimental observations. Two SMT-based methods that employ the Microsoft Z3 solver have been published: Reasoning Engine for Interaction Networks (RE:IN), which integrates user-defined experimental constraints and was used to build a Boolean network (BN) model of naïve mouse embryonic stem cells (Dunn et al., 2014; Yordanov et al., 2016); and Single Cell Network Synthesis (SCNS), which treats single-cell transcriptomics datasets as state transition graphs that must be satisfied and was used to build a BN model of embryonic hematopoiesis (Moignard et al., 2015). However, RE:IN and SCNS are computationally expensive for large networks and can perform poorly if the supplied experimental constraints are insufficient (i.e. if only a few experimental conditions are tested or if not enough single cells are captured to produce a fully connected state transition graph, respectively).

1.2.2 Computational simulation of GRNs

Static network topologies provide only limited insight into the dynamic behaviour of GRNs and their response to various biological conditions. It is possible to gain deeper insights to these aspects of GRN function by converting these topologies into a computable form; that is, by assigning mathematical functions to each edge in the network such that its dynamic behaviour can be simulated computationally.

Computational modeling can extend the utility of GRNs over static descriptions by (1) enabling artificial simulation of network behaviour, (2) predicting unknown regulatory relationships between genes, (3) predicting novel phenotypic states or transitions between states, and (4) identifying methods to manipulate the network's behaviour (Xiao, 2009). Since computation is relatively fast and inexpensive compared to cell culture or *in vivo* experiments, simulations can more efficiently screen large combinatorial spaces of perturbations and experimental conditions to narrow down the set of experiments that need to be performed. Computational simulations have been used to identify mechanisms that direct stem cells toward specific differentiated lineages, explore genetic conditions that lead to abnormal development, and discover new ways of accessing developmentally-important and clinically-relevant cell states (Collombet et al., 2017; Herrmann et al., 2012; Krumsiek et al., 2011; Yachie-Kinoshita et al., 2018). Finally, in cases where a computational model of a GRN fails to recapitulate observed biological behaviour despite including as complete a representation of existing literature and data as possible, they can allow us to systematically identify gaps in current knowledge or faulty modeling assumptions (Bonzanni et al., 2013; Collombet et al., 2017).

Many methods have been used for computational simulation of GRNs (summarized in Table 1) (Karlebach and Shamir, 2008; Le Novère, 2015; Schlitt and Brazma, 2007). Despite the nuances of individual methods, these can be broadly characterized into two types: logical and quantitative. Logical models typically represent gene expression values using discrete levels and describe regulatory interactions between genes using logic functions. Conversely, quantitative models represent gene expression values using continuous variables and describe regulatory interactions between genes using real-

valued functions, most commonly differential equations. In general, logical models are more amenable to inference from experimental data, avoid the need for estimation of biochemical rate parameters, and facilitate the simulation of larger networks due to their relative computational simplicity.

Table 1: Comparison of GRN simulation methods

Method	Description	Strengths	Weaknesses
Logical models			
Boolean networks	Expression values are restricted to 0 or 1, and interactions between genes are represented using Boolean logic operators (AND, OR, NOT). Binary state of network is updated at discrete time steps (Markov process)	Does not require fitting of biochemical rate parameters Computational simplicity permits simulation of large networks Relatively easy to infer from literature evidence and experimental data by satisfiability methods Identifies steady states and models network robustness without continuous values	Poorly models negative autoregulation Binary expression assumption not valid for some systems (ex. multiple thresholds) Discrete time steps have unclear physical interpretation
Multi-level logic models (generalized logical networks)	Generalized form of Boolean network in which variables can take on more than two discrete levels	Allows for recapitulation of dose-dependent gene regulation Can demonstrate non-linear regulatory interactions	Not always clear <i>a priori</i> how many levels should be assigned for each gene
Probabilistic Boolean networks	Considers multiple candidate regulatory functions for model components	Explicitly captures uncertainty about regulatory logic	Can be unclear how to best weight probabilities Sensitivity to probability values
Dynamic Bayesian networks	Probabilistic model that represents dependencies among genes across a time step	Permits efficient inference and learning from data	Requires a large amount of time series data to learn probability dependencies
Petri nets	Non-deterministic models consisting of ‘places’ (genes) and ‘tokens’ (units of expression). Places are connected by transition functions, which adjust number of tokens at input and output place when fired	Does not require fitting of biochemical parameters Permit automatic checking for boundedness and deadlocks Computational simplicity enables simulation of large networks	Can be difficult to determine appropriate transition functions for the model Underlying resource sharing concept is more suited to metabolic networks than gene regulation
Quantitative models			
Continuous ordinary differential equations of chemical kinetics	Describes instantaneous changes in each gene as a function of the concentration of its regulatory inputs, for example using Michaelis-	Uses real-valued parameters with clear physical interpretation Produces continuous output trajectories that can be easily compared against experimentally-observed	Requires adequate experimental data to accurately estimate parameters, which is not available for many biological systems

	Menten kinetics or Hill functions	timecourse expression dynamics.	Large networks are difficult to fit and computationally expensive Mass action kinetics do not account for stochastic effects present in gene regulatory systems
Piecewise linear differential equations	Uses Heaviside step functions (limit of Hill function) leading to discrete gene response to input regulators (i.e. 0 below threshold, maximal above threshold)	Enables efficient numeric computation of network behaviour while still using real- and continuously-valued variables	Sharp threshold boundaries are not appropriate in cases where genes respond to a gradient of input signal
Stochastic simulation algorithms	Simulates individual reactions given an initial number of molecules per molecular species and relevant reaction-probability rates	Accounts for stochastic effects that are prevalent when number of regulatory molecules is low	Does not account for diffusive or transportation effects Computationally expensive to simulate every reaction individually, therefore only tractable for small networks

1.2.3 Boolean networks

The most basic commonly-employed GRN simulation approach is Boolean networks (BNs), which makes two key simplifying assumptions:

1. Each gene in the network can take on only binary values, i.e. 1=ON or 0=OFF.
2. Regulatory interactions between genes are abstracted using Boolean logic, i.e. AND, OR, NOT.

Each gene in the network is assigned a Boolean logic function over its inputs, and this function is used to update the gene's binary state in the next discrete simulation step. In a *synchronous* simulation, all genes in the network are updated at each simulation step. Conversely, in an *asynchronous* simulation, only a random subset of genes in the network are updated at each simulation step. Thus, asynchronous simulations allow for each network state to lead to multiple possible next states. Asynchronous simulations are also more representative of the variable timing of gene regulation in living systems, where there is no universal "clock" to synchronize changes in transcription (Garg et al., 2008). Together, these Boolean update functions define a state transition graph, in which each state corresponds to a possible transcriptional profile and edges correspond to transitions

between states that are permitted by the GRN. In principle, such state transition graphs can be considered analogous to a single-cell level transcriptional space (Lim et al., 2016; Moignard et al., 2015). Attractors (steady states or strongly connected components) within the state transition graph can be compared to experimentally-observed phenotypes. Boolean representations are particularly appealing for GRN modeling because they can recapitulate the behaviour of biological systems without requiring biochemical rate parameters, which are difficult and time-consuming to experimentally validate.

BN modeling has been employed to accurately recapitulate segment polarity patterning in *Drosophila* (Sánchez and Thieffry, 2003), the cell cycle sequence in yeast (Davidich and Bornholdt, 2008), heterogeneity in murine hematopoiesis (Bonzanni et al., 2013), hierarchical differentiation of murine myeloid progenitors (Krumisiek et al., 2011), myeloid versus lymphoid lineage choice (Collombet et al., 2017), embryonic stem cell (ESC) states (Dunn et al., 2014; Xu et al., 2014), and many other biological gene regulation systems. In our own lab, this approach was successfully used to develop a BN model of the endogenous mouse pluripotency network consisting of 30 nodes and 7 signaling pathway inputs (BMP4, Activin, FGF, ERK, Wnt, LIF, PI3K) (Yachie-Kinoshita et al., 2018). The model accurately predicts the extent of transcriptional heterogeneity in various ESC culture conditions, captures the population-level effects of small molecule inhibitors and genetic perturbations, and identifies cellular transitions between distinct pluripotent states (Yachie-Kinoshita et al., 2018).

Although Boolean networks have a demonstrated ability to produce simulation results that are very similar to experimental observations, BN modeling also presents some limitations. One limitation is the need to discretize input data to the model into binary levels. Typically, biological experiments produce real-valued (as opposed to discrete-valued) results, and thus there is some loss of information when these results are discretized into binary ON and OFF values. However, recent single-cell transcriptomics studies suggest that gene expression levels are clearly bimodal at the single-cell level in at least some biological systems such as embryonic blood development (Moignard et al., 2015). These results suggest that binarization can be applied without significant information loss in such systems.

1.3 T cell lineage

In this section, we present T cell development as an interesting model system that would benefit from systematic computational exploration using a Boolean network (BN) approach.

T cells are a type of lymphocyte that develop in the thymus. They play a key role in adaptive immunity by specifically recognizing infected or malignant cells via their T cell receptor (TCR). The T cell lineage is a subject of intense research and clinical interest. Clinically, mature T cells can potentially serve as immunotherapy agents; for example, they can be programmed to recognize and kill specific cancer cell types using chimeric antigen receptors (Fesnak et al., 2016). There is also interest in transplanting T cell progenitors to help reconstitute the thymus in cases of immune deficiency arising from diseases such as primary immunodeficiency or as a result of radiation therapy (Brauer et al., 2016; Dolens and Taghon, 2017). However, at a more basic level, our understanding of the developmental program that drives uncommitted hematopoietic progenitors toward the T cell fate remains incomplete. Applying a BN modeling approach to T cell development would allow us to systematically explore the mechanisms underlying T lineage specification and commitment, predict the dynamic response of T cell progenitors to various environmental contexts and GRN perturbations, and eventually harness the findings to support robust scale-up of T cell-based clinical therapies.

The mouse T cell development program encompasses multiple cellular decision events coordinated by complex interplays between environmental signals and downstream GRNs (Rothenberg et al., 2016; Yui and Rothenberg, 2014). It is commonly described as occurring in 3 phases:

1. *Specification*, in which uncommitted blood progenitors undergo Notch-dependent proliferation and begin to upregulate T cell lineage-associated genes
2. *Post commitment*, in which progenitors commit to the T cell lineage, proliferation slows, and TCR β -chain gene rearrangement commences

3. *Post β -selection*, in which cells switch from Notch-dependence to pre-TCR signal dependence and are either proliferate following selection for functional TCR β -chain rearrangements or die

In vivo, T cells develop both during fetal development and throughout adulthood from blood progenitors that home to the thymus. Efforts to recapitulate aspects of the thymic niche (such as Notch signaling) *in vitro* have yielded multiple methods to effectively differentiate T cells from hematopoietic stem and progenitor cells (Schmitt et al., 2002; Shukla et al., 2017). This section explores these *in vivo* and *in vitro* systems with special emphasis on the GRNs that underlie T cell emergence in each context.

1.3.1 Stages of *in vivo* thymopoiesis

In vivo, T cells differentiate from lymphoid-primed progenitors that settle in the thymus. Upon thymic entry, these progenitors are exposed to multiple niche signals. Chief among these is Notch signaling, mediated by Delta-like ligand (DL4) presentation by the thymic stroma, which is essential for thymic specification of the T cell lineage. Other signals such as interleukin-7 (IL-7) additionally support cell proliferation and survival, though their role in lineage specification, if any, is unclear. Early T cell progenitors (ETPs) in mice proliferate over approximately 10 cell cycles and begin to downregulate “legacy” gene networks associated with the stem cell fate and alternate blood lineages. Sustained exposure to Notch drives progenitors to the CD4/CD8 double negative (DN)2 stages, marked by upregulation of T cell lineage-specific genes (*Tcf7*, *Gata3*, *Bcl11b*, etc.). Whereas B cell lineage potential is lost earlier on, potential for other blood fates such as the myeloid and natural killer (NK) lineages are lost at this stage. Now committed to the T cell fate, DN3 T cell progenitors begin rearranging the β -chain of the T cell receptor (TCR) are selected for functional pre-TCR rearrangements. Selected cells become double positive (DP) for the mature T cell markers CD4 and CD8, and must undergo additional rounds of selection to become either single-positive CD4⁺ helper T cells or CD8⁺ cytotoxic T cells (beyond the scope of this thesis). Additional details on the various stages and types of T cells are listed in Table 2.

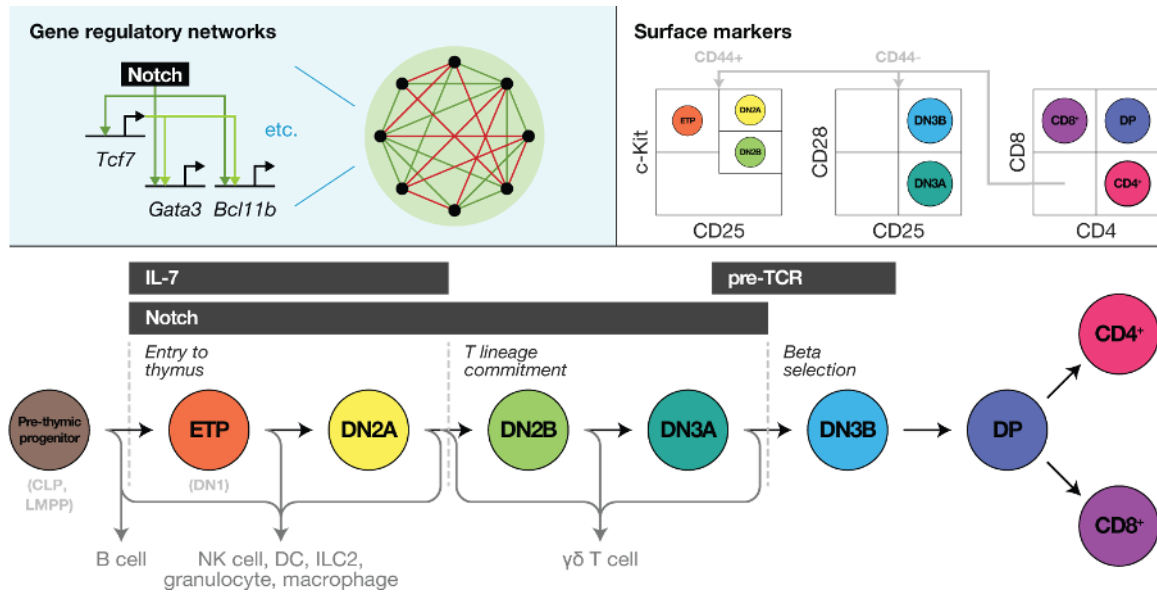


Figure 1: Overview of mouse T cell development program

Top left: Schematic of gene regulatory networks that underlie T cell lineage fate determination, including transcription factors (*Tcf7*, *Gata3*, *Bcl11b*, etc.) downstream of environmental signals (Notch signaling, etc.). *Top right:* Surface markers used to define various T cell progenitor stages and isolate progenitor populations via flow cytometry.

Bottom: Current understanding of mouse thymopoiesis, in which blood progenitors progress through a linear series of surface marker-defined stages in response to environmental signals, thereby losing their potential for other blood lineages and eventually committing to the T cell fate. CD = cluster of differentiation, CLP = common lymphoid progenitor, LMPP = lymphoid-primed multipotent progenitors, ETP = early T cell progenitor, DN = CD4⁻ CD8⁻ double negative, DP = CD4⁺ CD8⁺ double positive, IL-7 = interleukin-7, pre-TCR = pre-T cell receptor

Table 2: Description of T cell progenitor stages and mature T cell types

Stage/Type	Surface Markers	Description
ETP	CD4- CD8- KIT-hi CD44+ CD25-	Early T cell progenitor. Retains potential for non-T blood lineages, as well as expression of some hematopoietic stem cell-associated genes
DN1	CD4- CD8- CD44+ CD25-	Superset of ETP. Heterogeneous subset of blood progenitors with diverse T cell lineage potential (Porritt et al., 2004)
DN2A	CD4- CD8- KIT-hi CD44+ CD25+	Start to express T cell lineage gene markers
DN2B	CD4- CD8- KIT+ CD44+ CD25+	Committed to the T cell lineage
DN3A	CD4- CD8- KIT- CD44- CD25+ CD28-	TCR β gene rearrangement begins
DN3B	CD4- CD8- KIT- CD44- CD25+ CD28+	β -selection, cells with successful TCR β rearrangements receive pre-TCR signaling that rescues survival
ISP	CD4- CD8+ TCR-/lo CD24hi	Intermediate single positive transition state between DN and DP stages
DP	CD4+ CD8+ TCR-/lo	TCR α gene rearrangement occurs, followed by expression of a fully assembled $\alpha\beta$ -TCR. Positive selection occurs, in which cells with functional ligand specificities are selected for
CD4+ helper (T _h)	CD4+ CD8- TCRhi	Release T cell cytokines to regulate immune response, including B cell class switching, cytotoxic T cell activation, and enhancement of macrophage activity

Stage/Type	Surface Markers	Description
CD8+ cytotoxic (T _c)	CD4- CD8+ TCRhi	Kills cancer cells, virus-infected cells, or other damaged body cells via cytotoxin release
$\gamma\delta$ -T	TCRd+	Low abundance T cells that express a T cell receptor (TCR) comprised of one gamma chain and one delta chain; involved in immune response initiation and propagation
Natural killer T (NKT)	CD1d+, mix of other markers	Heterogeneous class of CD1d-restricted T cells that express some NK cells markers and semi-invariant TCRs; respond to activation by rapidly releasing a large amount of cytokines
Regulatory T (Treg)	CD3+ CD4+ CD25hi FOXP3+ CD127lo	Immunosuppressive T cells that maintain self-tolerance and prevent autoimmune disease

1.3.2 Environmental signals in the thymic niche

The thymus provides a uniquely supportive niche for T cell development, largely due to the presence of many crucial environmental signals. Foremost among these signals is **Notch**, mediated primarily by presentation of Delta-like ligand 4 (DL4) by thymic stromal cells and expression of the NOTCH1 receptor by blood progenitors (Love and Bhandoola, 2011). When NOTCH1 binds to DL4 molecules on the thymic stroma, a mechanical force is generated that permits proteolytic release of the intracellular component of the NOTCH1 molecule, which in turn localizes to the nucleus and functions as a transcription factor in complex with RBPJ (recombining binding protein suppressor of hairless). Force generation via a tethered ligand is essential for Notch signaling activation; indeed, our lab has demonstrated that soluble DL4 inhibits both

Notch pathway activation and T cell differentiation during *in vitro* culture (Shukla et al., 2017). Of potential interest, multiple labs have demonstrated that the extracellular and intracellular domains of Notch receptors can be swapped for other protein subunits to synthetically reprogram the receptors' signal-receiving or downstream effector function. The modularity and juxtacrine nature of the Notch pathway makes it an attractive tool for synthetic biology applications (Morsut et al., 2016; Nandagopal et al., 2018; Roybal et al., 2016).

In the specific context of T cell development, Notch signaling promotes growth and survival of thymocytes up to the DN3B stage, at which point their survival is primarily determined by selective pre-TCR signaling. Importantly, Notch signaling also activates a set of downstream transcription factors that are critical for T lineage differentiation, such as *Gata3*, *Tcf7*, and *Bcl11b*. Notch signaling is also primarily responsible for antagonizing the differentiation potential of T cell progenitors toward other blood lineages, especially the B cell lineage (Rothenberg et al., 2016).

Another key element of the thymic signaling niche is **interleukin-7 (IL-7)**. IL-7 supports early thymocyte survival; in particular, cells at the DN2A stage strongly upregulate IL-7 receptor (IL-7R) and are especially sensitive to extracellular IL-7 levels (Wang et al., 2006). Mice that lack IL-7 or IL-7R exhibit strongly impaired T cell development, and IL-7R deficient progenitors cannot fully seed the thymic niche (Prockop and Petrie, 2004; Zięta et al., 2015). Although it is known that IL-7 signaling plays a critical role in $\gamma\delta$ -T cell development and CD8 lineage choice via activation of *Runx3*, the extent of its role in $\alpha\beta$ -T lineage progression remains somewhat unclear. There is evidence that T cell progenitors can sense relative changes in IL-7 levels and that a drop in IL-7 signaling is required to proceed to the DN2B stage and commit to the T lineage (Ikawa et al., 2010).

Kit (stem cell factor) is also present in the thymic niche, where it supports survival and proliferation of ETP cells. However, there is currently no evidence indicating a role for Kit-mediated signaling in T lineage progression.

Although **Flt-3 ligand** is not found in the thymic niche, it is important for maintaining blood progenitors pre-thymically and is a common media component for *in vitro* culture

of hematopoietic stem and progenitor cells (HSPCs). The Flt-3 receptor itself is expressed on thymic seeding progenitors, but becomes downregulated during the ETP stage.

1.3.3 *In vitro* methods for T cell lineage differentiation

The ability to differentiate HPSCs toward the T cell lineage *in vitro* has been of prime importance to the T cell field, both in terms of informing our understanding of T cell developmental biology as well as opening new opportunities for cell therapy. However, recapitulating the complex *in vivo* signaling niche provided by the thymus in an *in vitro* setting is particularly challenging compared to other blood cell differentiation protocols. One of the critical challenges is presentation of Notch ligands in a manner that permits mechanical force generation and subsequent cleavage of the Notch receptor to trigger downstream signaling. In fact, it has been shown that soluble presentation of Notch ligands (which does not allow for mechanical force generation) inhibits T lineage differentiation rather than enhancing it (Shukla et al., 2017). Yet despite this and other challenges, multiple methods have been developed that enable successful T lineage differentiation outside the body (Brauer et al., 2016).

The first of these methods is fetal thymic organ culture (FTOC), which originally came into widespread use during the early 1990s (Jenkinson and Anderson, 1994). This method involves dissection of thymic lobes from E14-15 fetal mice and treating them with 2-deoxyguanosine to remove the resident thymocytes. Treated lobes are then reconstituted together with HSPCs by the hanging drop method and subsequently cultured in the presence of supportive cytokines. Because FTOCs are derived directly from dissected thymi, they provide a relatively faithful 3-dimensional recapitulation of the *in vivo* thymic niche. However, preparation of FTOCs is time consuming and limited by availability of fetal mice, and thymocytes must be extracted from the FTOC to perform endpoint analysis, thus limiting the experimental accessibility of this system.

The OP9-DL1 (and later OP9-DL4) co-culture system was first developed in the early 2000s in the lab of Juan-Carlos Zúñiga-Pflücker, and has proven particularly instrumental during the intervening years for enhancing our understanding of T cell development (Mohtashami et al., 2010; Motte-Mohs et al., 2005; Schmitt et al., 2002). In this platform, OP9 bone marrow stromal cells are transduced to ectopically express either the Notch ligand Delta-like ligand 1 (DL1) or DL4 and are co-cultured with HSPCs in the presence of serum and supportive cytokines. Because the OP9-DL system provides an accessible 2-dimensional culture platform, it facilitated novel studies into the T lineage potential of various progenitors that were impossible using FTOCs. With this platform, HSPCs could be isolated from donors, sorted by flow cytometry to isolate specific populations of interest, seeded on OP9-DL culture, and either harvested for timepoint analysis or observed *in situ* using immunofluorescence and live imaging techniques. However, OP9-DL1 and OP9-DL4 necessitate the use of stromal feeder cells and serum-containing media which limit technical reproducibility and present a significant obstacle to clinical translation in the human system.

Because of these limitations, there has been significant recent research interest in developing stromal feeder-free and serum-free platforms for *in vitro* T cell differentiation. Recently, our lab has developed a novel serum-free and stroma-free *in vitro* platform for mouse and human T cell differentiation that utilizes adsorbed DL4 and vascular cell adhesion molecule-1 (VCAM) (Shukla et al., 2017). After optimization of media conditions, seeding density, and extracellular matrix components, DL4+VCAM was able to produce late-stage T cell progenitors capable of *in vivo* thymus engraftment and immune function at comparable efficiency to OP9-DL4, while also demonstrating faster differentiation kinetics. Furthermore, because the DL4+VCAM platform omits uncharacterized serum and xenogeneic co-culture, it is more readily translatable to clinical applications. These advancements are especially timely given recently increased interest in engineered chimeric antigen receptor T (CAR-T) cells for cancer immunotherapy (Dai et al., 2016).

1.3.4 Gene regulatory networks for T cell development

Specification and commitment to the T cell lineage are driven by the combined activity of many transcription factors (TFs) that together form a gene regulatory network (GRN).

There is no single “master regulator” for the T cell lineage; rather, a set of at least 10 distinct regulatory inputs appear essential for T cell development, as listed in Table 3 and reviewed in Rothenberg et al, 2016. Notably, most of these regulatory inputs are not unique to the T cell lineage, but are shared developmentally by at least one other blood lineage (David-Fung et al., 2006). As a result, tight control over these inputs is especially critical for proper T cell development to occur.

Table 3: Key regulatory inputs to the T cell development program

Regulatory Input	Specific Genes	Description
E proteins	<i>Tcf3</i> (E2A) <i>Tcf12</i> (HEB)	Basic helix-loop-helix (bHLH) TFs
Runx family	<i>Runx1</i> <i>Runx3</i>	TFs
GATA-3	<i>Gata3</i>	TF
c-MYB	<i>Myb</i>	TF
Ikaros-type zinc fingers	<i>Ikzf1</i> (Ikaros) Aiolos Helios	Zinc finger TF
TCF-1/LEF-1	<i>Tcf7</i> (TCF-1) <i>Lef1</i>	HMG box TF
PU.1	<i>Spi1</i>	TF; early “hit-and-run” role
GFI-1	<i>Gfi1</i>	Zinc finger repressor TF
BCL11B	<i>Bcl11b</i>	TF
Notch signaling	<i>Notch1</i> <i>Dll4</i> (DL4, expressed by thymic stroma)	Signaling pathway

These regulatory inputs interact with one another to form network motifs. Network motifs are sub-circuits of gene regulatory networks that are overrepresented in nature compared to random chance and perform characteristic modes of information processing

(Alon, 2007). Feedforward loops (coherent or incoherent) and feedforward loops (positive or negative) are common examples of network motifs and feature prominently within the T cell development GRN. For example, Notch signaling and its target genes *Tcf7*, *Gata3*, and *Bcl11b* form a coherent feedforward loop, in which Notch directly upregulates *Tcf7*, which in turn cooperates with Notch signaling to subsequently activate *Gata3* and *Bcl11b* (Figure 2). This motif is hypothesized to serve as persistence detector for Notch signaling, such that T lineage commitment (marked by upregulation of *Bcl11b* activity) only occurs when Notch signaling levels remain elevated long enough for its primary transcription factor targets to become active (Kueh and Rothenberg, 2012). Other previously studied network motifs within the T cell development GRN include incoherent feedforward loops downstream of IL-7 signaling that are hypothesized to enable T cell progenitors to sense and respond to changes in IL-7 levels (Ikawa et al., 2010); and mutual repression between the T lineage antagonist transcription factor *Spi1* (PU.1) and T lineage promoting transcription factors *Tcf7* (TCF-1), *Gata3*, and *Bcl11b* which reinforces the exclusion of alternate myeloid cell fates during T lineage specification (Del Real and Rothenberg, 2013).

Network motifs

ex. Notch-driven coherent feedforward loop

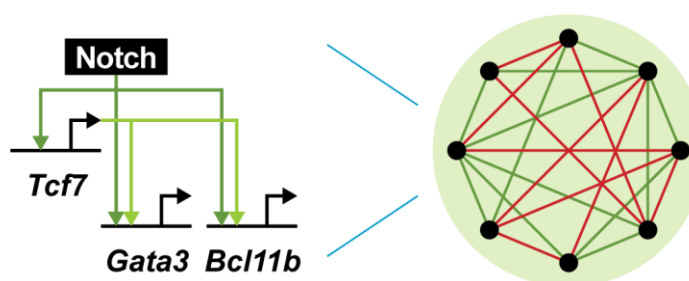


Figure 2: Notch-driven coherent feedforward loop forms a network motif within T cell development GRN

1.3.5 Plasticity and potential shortcuts in T cell development program

Despite the need for tight control over its regulatory inputs, previous evidence suggests there is significant intrinsic plasticity to the T cell development program. For example, DN2 stage progenitors are much more abundant in OP9-DL1 *in vitro* co-culture than they are in the thymus (Balciunaite et al., 2005; Kawazu et al., 2005). Moreover, it remains unclear whether DN2 cells cultured on OP9-DL1 are fully transcriptionally equivalent to primary DN2 thymocytes, underscoring the notion that classic surface marking for defining T cell development stages (CD25, CD44, c-KIT, etc.) may mask developmentally important transcriptional heterogeneity.

Furthermore, many instances of shortcuts or detours from the “canonical” series of events in T cell development have been reported in various contexts. For example, multiple groups have reported that the dwell time spent by T cell progenitors in the DN2 and DN3 stages is greatly reduced in certain contexts. In one instance, the thymi of *Foxn1* splice mutant mice were reported to contain DP cells despite a near complete lack of DN3 cells, suggesting the existence of a shortcut that bypasses the DN3 state (Su et al., 2003). In another case, certain Kit^{lo} HSA⁺ subsets of the DN1 compartment were reported to give rise to DP cells at low efficiencies but without generating any detectable DN2 or DN3-like intermediates (Porritt et al., 2004). Finally, mouse and human models of *Gata3* overexpression contain a fraction of cells that successfully transit to the DP stage despite significant depletion of DN3 intermediates, suggesting it may be possible for DN1 or DN2 stage progenitors to proceed immediately to β -selection (Anderson et al., 2002; Taghon et al., 2001). Taken together, these observations suggest that the classic view of T lineage differentiation as a cleanly linear stage-wise progression may omit alternative routes to the T cell fate which are more readily observed *in vitro* or under perturbed *in vivo* conditions (David-Fung et al., 2006). However, it remains to be understood what physiological role this potential variation might serve.

Another set of particularly striking examples of plasticity in the T cell development program comes through comparing the kinetics and genetic requirements of fetal thymopoiesis versus adult thymopoiesis. In mouse development, hematopoietic

progenitors settle the thymus at approximately E13.5, and the first DP thymocytes emerge at approximately E16.5 (David-Fung et al., 2006), a total period of 3 days. Conversely, in adult thymi it can take up to 10 days for progenitors to reach the DN2 stage and 2 weeks for them to reach the DP stage (David-Fung et al., 2006). In terms of genetic requirements, Ikaros-null mutant mice produce no fetal thymocytes, but exhibit normal T cell development and a fully-populated thymus into adulthood (Wang et al., 1996). Similarly, severely hypomorphic PU.1 mutants exhibit completely arrested fetal thymopoiesis, yet begin to produce T cells following birth for as long the mice can survive (Back et al., 2004; Dakic et al., 2005; Scott et al., 1994). There is evidence to suggest fetal thymopoiesis is also more robust to decreased IL-7 signaling and *Tcf7* (TCF-1) expression levels than adult thymopoiesis, possibly due to a compensatory effect mediated by related cytokines and transcription factors (Crompton et al., 1998; Okamura et al., 1998; Schilham et al., 1998; Verbeek et al., 1995). Finally, many important T cell development genes exhibit highly disparate stage-wise dynamics in fetal versus adult thymopoiesis, including *Id1*, *Id2*, *Runx1*, *Runx3*, *Spi1* (PU.1), *Spib*, *Cd3e*, and *Deltex* (David-Fung et al., 2006). Thus, although both the adult and fetal thymus support T cell production, the pathways employed in each context differ in important ways. In more general terms, it appears likely that while some aspects of the T cell development program must be conserved across all contexts, other aspects permit greater flexibility (David-Fung et al., 2006).

1.3.6 Alternative strategies for computational modeling of T cell development GRNs

In this study, we present Boolean networks (BNs) as an informative framework for modeling the mouse T cell development program. Although Boolean networks are commonly used to model other developmental systems, other groups have employed alternative strategies to model the GRN dynamics of related systems.

For example, **continuous modeling with ordinary differential equations** avoids the need to discretize data, instead producing real-valued outputs that are more reflective of

experimental data. Initial efforts have been made recently by other groups to create differential equation-based models of T cell specification dynamics (Manesso et al., 2016). However, these models are limited to sub-network motifs rather than the complete T cell development GRN and do not generate new predictions about T cell differentiation. As such, there remains a clear research need for network-scale computational models of the T cell development program.

Another class of computable GRN models are **multi-level logical models**. Although regulatory interactions are abstracted as logical functions in these models (similar to BN modeling), multi-level models permit more than 2 discrete activity levels for each variable. This approach was recently employed in the myeloid versus lymphoid lineage choice model published by Collombet et al, in which most genes are constrained to binary values (0 or 1), except for *Spi1* which the authors chose to model with 3 discrete levels (0, 1, or 2) (Collombet et al., 2017).

1.3.7 Open questions in T cell development field

Overall, despite many advances, the field still lacks a complete understanding of the transcriptional dynamics of the T cell development program, and specifically how GRNs encode these dynamics. It is unclear whether transcriptional heterogeneity among progenitor stages is functionally important for T cell development (as observed in pluripotency, for example) (Porritt et al., 2004). Relatedly, it is also unknown whether there are multiple transcriptional trajectories which can lead toward T cell lineage commitment, or only one (Bhandoola and Sambandam, 2006), and whether certain trajectories are more commonly employed in certain differentiation contexts (ex. fetal versus adult thymopoiesis, or *in vivo* thymopoiesis versus *in vitro* differentiation). Finally, the field would benefit from additional insight into key control points within GRNs for T cell differentiation. Characterization of these control points would enable targeted improvements to *in vitro* differentiation protocols; for instance, by enriching for the T cell lineage over alternate fates or reducing dependence on DL4-mediated Notch signals – a key limiting factor for large-scale T cell manufacturing toward therapies.

Boolean network (BN) modeling presents an as-of-yet unexplored opportunity to harness the field's existing knowledge of the T cell development GRN into a computable framework that examines these questions in a relatively rapid and unrestricted manner. Therefore, in this thesis we construct a BN model of the T cell development program and demonstrate its utility over existing static GRN models for predicting heterogeneity and transcriptional dynamics among T cell progenitors in response to extrinsic differentiation signals.

2 Objectives & Aims

2.1 Hypothesis

The gene regulatory networks (GRNs) that comprise the mouse T cell development program support heterogeneous and context-dependent transcriptional responses to extrinsic signals, and these transcriptional responses can be accurately predicted by Boolean network modeling.

2.2 Specific Aims

Aim 1: Construct a computational model of the mouse T cell development program that accurately recapitulates experimental observations

Toward this aim, I developed the first-reported Boolean network (BN) model of the mouse T cell development program based on published high-confidence regulatory interactions and supplemented with a small number of additional interactions derived through partial correlation analysis and iterative refinement through simulation.

Asynchronous simulations of the BN model produce steady states that closely resemble known T cell progenitor cell types. Furthermore, simulations also accurately recapitulated the response of T cell progenitors to various combinations of exogenous signals and genetic perturbations. Finally, the model predicts the T cell development GRN supports multiple possible transcriptional trajectories leading toward T cell lineage commitment. While genes that antagonize the T cell lineage are predicted to be downregulated quickly regardless of trajectory, genes that promote T cell lineage choice are predicted to exhibit more flexibility between trajectories.

Aim 2: Support application and extension of Boolean network modeling by the broader biological community

Toward this aim, I developed a set of software ‘gadgets’ for the *Garuda* systems biology platform. These gadgets enable the broader research community to easily perform gene expression discretization, Boolean network simulation, and state transition graph analysis (including steady states and strongly connected components) without any computer

programming requirements. The gadgets interface with other systems biology software tools via *Garuda* to enable larger data analysis pipelines. Finally, the gadgets were used to simulate a BN model of mouse embryonic stem cells developed in our lab and were publicly released to encourage further expansion of said model by the community (Yachie-Kinoshita et al., 2018).

Aim 3: Compare transcriptional dynamics and heterogeneity of T cell progenitors across various differentiation contexts

Toward this aim, I first assisted in characterizing a novel serum- and stroma-free platform for mouse and human T cell differentiation that utilizes adsorbed Delta-like ligand 4 (DL4) and vascular cell adhesion molecule 1 (VCAM) to create an artificial thymus-like niche. Specifically, it was demonstrated that mouse fetal liver and human umbilical cord blood hematopoietic progenitors upregulate T cell lineage-specific Notch target genes higher when seeded on DL4+VCAM versus DL4 alone. I also demonstrated that T cell progenitors exhibit higher motility during DL4+VCAM culture versus DL4 alone, providing a potential physical explanation for increased activation of the Notch pathway in this novel platform (Shukla et al., 2017).

Separately, single-cell qRT-PCR analysis of sorted primary adult thymocytes was used to demonstrate that surface marker-defined stages of early T cell development mask comprise overlapping and transcriptionally heterogeneous cell populations. Finally, single-cell RNA sequencing (scRNA-seq) was used to examine transcriptional differences and trajectories of differentiating T cell progenitors in fetal thymopoiesis and the aforementioned DL4+VCAM *in vitro* differentiation platform.

3 Construction of a Boolean network model of T cell development

3.1 Introduction

The T cell lineage is a subject of great clinical and research interest. Clinically, T cells hold great potential as cellular therapy agents, especially given recent successes in chimeric antigen receptor (CAR) technology and adoptive cancer immunotherapies (Dai et al., 2016). To support these efforts, more efficient *in vitro* differentiation protocols for the T cell lineage must be developed by harnessing our understanding of how T cells normally develop *in vivo*. Recent studies have identified the genetic players and regulatory interactions that are critical for T cell development from mouse hematopoietic progenitors (detailed in Chapter 1.3). However, integrating these findings into a comprehensive and predictive computable model of T cell lineage specification remains an open challenge.

Boolean network (BN) modeling is a commonly-employed approach for functionalizing qualitative observations about gene regulatory networks (GRNs) into a computable and predictive form (detailed in Chapter 1.2.3). In a BN model, gene expression levels are described with binary variables. Boolean logic functions are used to calculate the binary state of each gene based on the complete network state, thereby enabling simulation of the network's dynamics over many discrete steps. Attractors within the resulting state transition graph often recapitulate experimentally observable cell types (Huang et al., 2005). When applied to biological networks responsible for cell fate decisions, BN models can accurately recapitulate observable cell states. In general, BN modeling and simulation can be used to make informative predictions about the dynamic response of GRNs to different input conditions and genetic perturbations (Albert and Thakar, 2014).

Herein we report the construction and simulation of a Boolean network (BN) model that describes the transcriptional control of single progenitor cells by the T cell development GRN. The model encompasses 34 genes and 3 external signaling pathway inputs that are implicated in mouse T cell development. Asynchronous simulation of the BN model accurately captures known T cell phenotypes under normal and abnormal development conditions, as well as the combinatorial effect of environmental signals on access to transcriptional space. The BN model also predicts that there are multiple possible

transcriptional trajectories leading toward T lineage commitment. Overall, the BN model developed here establishes both a promising framework for exploring the T cell development program *in silico* and a mechanistic basis for designing future improvements to T cell differentiation protocols.

3.2 Methods

3.2.1 Literature curation process

The majority of edges in our Boolean network (BN) model were curated from previous literature. We focused on genes downstream of Notch, interleukin-7 (IL-7), and pre-T cell receptor (pre-TCR) signaling since these inputs are reported to drive progression through the T cell lineage, as opposed to only providing survival support (see Chapter 1.3.2). First, known genes and interactions of interest downstream of Notch, IL-7, and pre-TCR signaling were collected from previous static descriptions of the T cell development GRN (Georgescu et al., 2008; Kueh and Rothenberg, 2012; Longabaugh et al., 2017). Unlike continuous mathematical models that employ ordinary differential equations, BN models do not require quantitative information about binding rates, transcriptional rates, or other biochemical processes. Indeed, such biochemical information remains unknown for most of the studied interactions comprising the mouse T cell development program.

In a BN model, a regulatory interaction is sufficiently well-described if it is known that (1) a gene product X targets another gene Y, and (2) whether X promotes or represses expression of gene Y. Therefore, evidence supporting these two claims were the primary focus of our literature search. Literature evidence for the existence, direction, and sign (activation or repression) of regulatory edges comprised: genetic overexpression or knockout followed by mRNA expression analysis, changes in media-supplemented ligand levels followed by mRNA expression analysis, or overexpression of the intracellular domain of Notch. Evidence of direct binding of transcription factors to downstream

enhancer targets was taken from published chromatin immunoprecipitation (ChIP) studies. A full list of the literature evidence used to inform edges in our BN model is provided in Supplementary Table 1.

3.2.2 Partial correlation analysis

To infer potential gene regulatory interactions that may remain unreported, we employed a partial correlation-based approach to identify pairs of highly-correlated genes. A similar approach had previously been employed in our lab to infer regulatory edges comprising a GRN for mouse embryonic stem cells (Yachie-Kinoshita et al., 2018). A set of 138 bulk microarray profiles of flow-sorted T cell progenitors was downloaded from Gene Expression Omnibus and the Immunological Genome Consortium (Mingueneau et al., 2013). A full list of the microarray datasets employed in this analysis is provided in Supplementary Table 2. Data from all probe sets were quantile normalized by RMA using the ‘oligo’ Bioconductor package for R. Probe sets were then collapsed into unique genes by taking the mean values of probes annotated to the same gene. Batch effect correction was performed using COMBAT. Only those genes that exhibited a range of >2 between the maximum and minimum across all datasets were used for downstream correlation analysis. For each of 10,000 iterations, 100 genes were randomly sampled for pairwise partial correlation analysis with the ‘ppcor’ package in R. Each gene pair was assigned a score, defined as the highest Pearson partial correlation for the pair over all iterations where $p < 0.05$.

3.2.3 Iterative refinement of Boolean update functions

Experimental evidence (overexpression, knockdowns, ChIP-seq, etc.) for each interaction was verified against previous reports. Then, candidate Boolean update functions describing the regulatory inputs to each gene in the network were chosen; for example: *Bcl11b* = *Notch signaling and Gata3 and Tcf7 and Runx1*. In cases where evidence of specific cis-regulatory logic was previously reported in literature, this logic was used to

define the gene's update function. For example, the cis-regulatory element of *Bcl11b* is exceptionally well-characterized and only accurately represented by AND logic (Kueh et al., 2016). For genes with unknown or underspecified cis-regulatory logic, update functions were refined iteratively by asynchronously simulating the BN and comparing the resulting steady states and strongly connected components (SCCs) against biological reference states gathered from microarray data. Specifically, in cases where a variable's state consistently disagreed with experimental expectations, the corresponding update function was modified such that the truth table yielded the expected result, and the modified model was simulated to test for improved accuracy.

3.2.4 Asynchronous Boolean simulation

Under the model's Boolean assumptions, each variable has two possible states: 0 ("OFF") and 1 ("ON"). Biologically, we interpret these values as the minimum and maximum expression values, respectively, observed experimentally across all reference cell types for each gene. In asynchronous simulation of a BN consisting of N variables, n variables ($1 \leq n \leq N$) are selected at each simulation step and updated according to their associated Boolean update function. The number of variables updated with each time step is random, and each variable has an equal chance of being selected. This asynchronous behaviour mimics the stochastic nature of gene expression frequency in organisms (Garg et al., 2008).

A BN model with N variables has 2^N possible states, where each state is a N -length vector of "0" or "1" values representing the ON/OFF profile of all model variables (Figure 3). Since the described T cell development GRN model has 37 model variables, there are 1.37×10^{11} possible expression states.

Furthermore, under an asynchronous update schema, each state has 2^N possible successor states. Therefore, the state transition graph for our reported T cell development GRN model is a directed weighted graph that can contain a maximum of 1.89×10^{22} edges. However, the successors of most states are limited by the model's update rules and

resulting stable limit cycles or steady states. Because of this, outer states are generally negligible in terms of frequency of appearance. Asynchronous Boolean simulation from a limited set of random initial states can be performed to make simulations more computationally tractable. For all simulations, >2000 runs were performed from either a randomly-initialized set of states or a single pre-defined initial state, and each run was simulated for >250 consecutive simulation steps. Random simulations under these conditions produced repeatable outputs. Edge weights were assigned based on how frequently each edge was traversed over all simulation runs. Random asynchronous Boolean simulation was performed in Python 2.7 using the *BooleanNet* package (version 1.2.8).

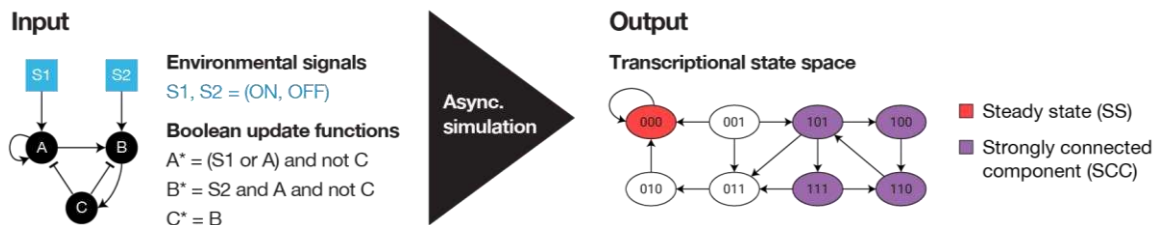


Figure 3: Overview of dynamic Boolean network modeling approach

Interaction networks defined by logic functions are simulated by asynchronously updating individual genes to produce a transcriptional state space (directed weighted graph) containing two types of attractors—steady states (SSs) and strongly connected components (SCCs)

3.2.5 Identification of steady states and strongly connected components

In our BN modeling framework, we investigated two types of attractors: steady states and strongly connected components (Figure 3). Steady states (SSs) are single states at dead ends in the state transition graph, and we assume these reflect relatively homogeneous cell populations, such as differentiated cell states. Conversely, a strongly connected component (SCC) is defined as a set of states wherein each state is reachable from every other state in the set. This is reminiscent of observations of dynamic heterogeneity within pluripotent stem cell (PSC) populations, in which individual cells can transition among numerous states that are high or low in their expression of specific pluripotency genes (Filipczyk et al., 2015; Singer et al., 2014). Note that SCCs are not necessarily closed systems, and SCCs with outgoing transitions were also considered in our analysis. SCCs in a state transition graph were identified by Tarjan's algorithm, which operates by iteratively removing disconnected states and steady state attractors that lack either an incoming or outgoing transition edge. State transition graph analysis and attractor state identification was performed in Python 2.7 using the *networkx* package (version 1.2.6).

3.2.6 *In silico* genetic knockouts and forced expression

In silico loss- and gain-of-function assays were performed by fixing each gene in the GRN to either 0 or 1 and simulating the dynamic behaviour of the perturbed network as described above.

3.2.7 Comparison of BN simulations and binarized experimental data

Agreement between steady states of the BN model and known T cell development stages was quantified using the Hamming distance metric, which measures the number of bits that are different between two binary vectors of length L . Briefly, stage-specific transcriptional profiles for all model genes as measured by bulk microarray were

downloaded from the Immunological Genome Consortium (Mingueneau et al., 2013) and binarized by k -means ($k=2$). The percent agreement score was calculated as:

$$\% \text{ Agreement} = 1 - \frac{\sum_{i=0}^L [x_i \neq y_i]}{L}$$

where x is the binary state vector for the computed steady state, y is the binarized experimentally-observed transcriptional profile, and each index i corresponds to a gene included in the BN model.

The state space of each simulation was visualized by projecting each state along the principal components of a reference set of k -means binarized bulk microarray data from the Immunological Genome Consortium (ImmGen) for each stage of T cell development and alternative blood lineages (including natural killer, natural killer T, macrophages, granulocytes, pre-pro B, and dendritic cells).

3.2.8 Trajectory identification and clustering from asynchronous Boolean simulations

In graph theory, a path from one node to another in a directed graph without repeated states is called a “trajectory”. We used the *networkx* package for Python to identify trajectories from early to late T cell progenitor states via depth-first search. To determine common patterns among these trajectories, we used UPGMA to cluster the trajectories in principal component space based on the Fréchet distance metric. Formally, the Fréchet distance between two curves A and B is defined as:

$$F(A, B) = \inf_{\alpha, \beta} \max_{t \in [0, 1]} \{d(A(\alpha(t)), B(\beta(t)))\}$$

where d denotes Euclidean distance. Common transcriptional patterns were analyzed among the top 6 clusters of trajectories by taking the average expression value at each time step for all trajectories in the cluster.

3.3 Results

3.3.1 Construction of BN model of mouse T cell development

We defined the scope of our Boolean network (BN) model as the differentiation of pre-thymic lymphoid progenitors (such as common lymphoid progenitors, CLPs) into CD4⁺ CD8⁺ double positive (DP) T cells. Intermediates in this process include the surface marker-defined CD4⁻ CD8⁻ double negative (DN)-1, DN2, and DN3 stages, as defined in Chapter 1.3.1. More specifically, we set out to create a Boolean logic representation of the gene regulatory networks (GRNs) that drive progression through these stages of the T cell lineage.

To construct the model (Figure 4a), known genes and interactions of interest downstream of Notch, IL-7, and pre-TCR signaling were collected from a previous static description of the T cell development GRN (Kueh and Rothenberg, 2012; Longabaugh et al., 2017) and recent T cell literature (Supplementary Table 1). Furthermore, we attempted a partial correlation-based approach to predict additional interactions of interest from 138 published bulk microarray profiles of flow-sorted T cell progenitors (Supplementary Table 2). However, this approach yielded a high false negative rate, with 37.9% of reported high-confidence regulatory interactions being identified as uncorrelated (Figure 4b).

Experimental evidence (including overexpression, knockdowns, ChIP, etc.) for each interaction was verified against previous reports. Then, candidate Boolean update functions of the regulatory inputs to each gene in the network were chosen; for example: *Bcl11b*' = *Notch signaling and Gata3 and Tcf7 and Runx1*. These update functions were iteratively refined by comparing the steady states and strongly connected components (SCCs) resulting from asynchronous simulation against biological reference states gathered from microarray data. The finalized BN model of the T cell development GRN (Figure 4c) consists of 37 nodes (3 signaling inputs + 34 genes) and 103 edges. The Boolean update functions for all nodes are provided in Table 4.

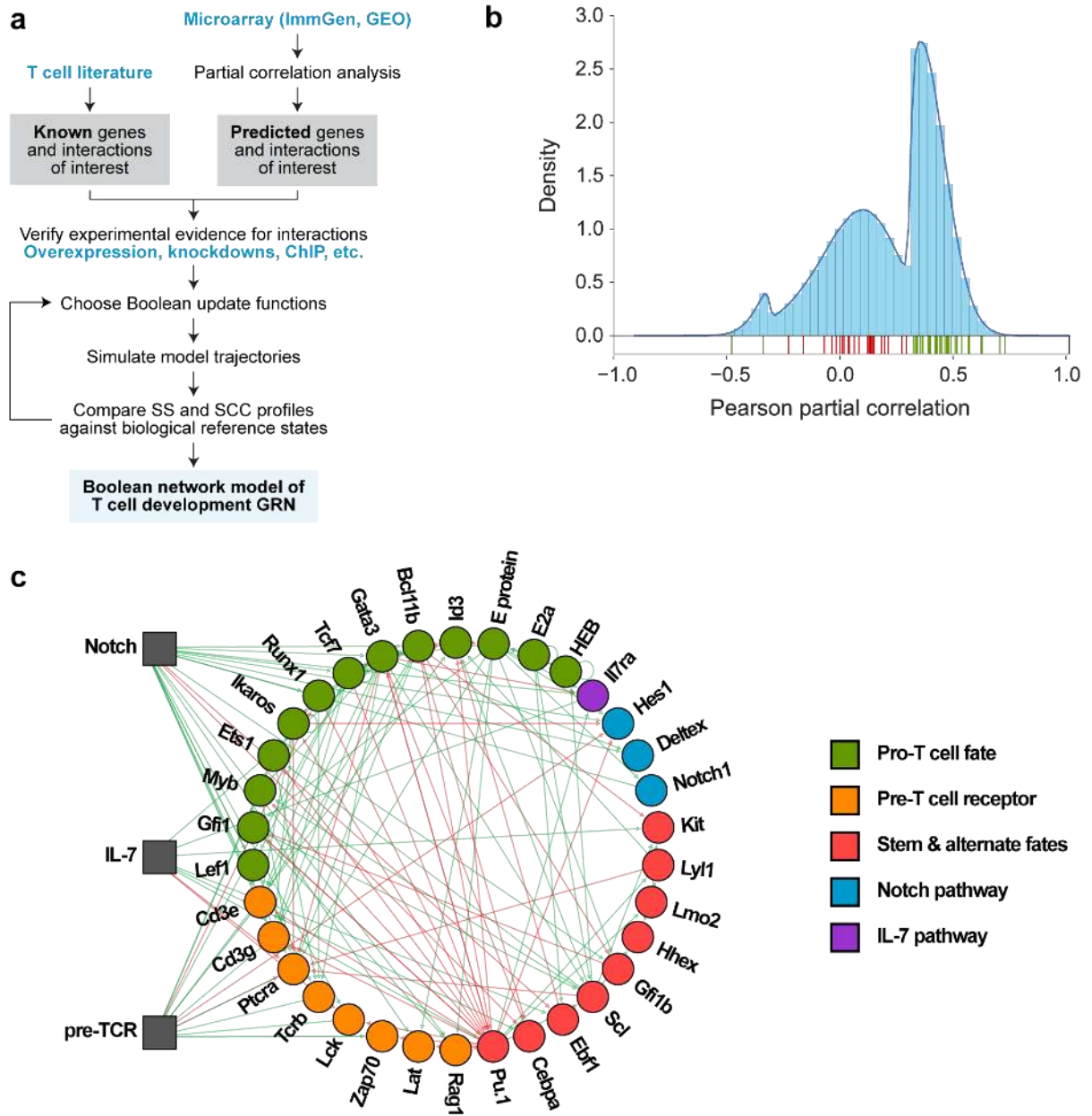


Figure 4: Construction of BN model of mouse T cell development

(a) Iterative method used to construct the BN model of T cell development

(b) Partial correlation analysis of T cell progenitor microarray datasets yields a trimodal distribution with high false negative rate (37.9%). Rugs indicate gene pairs with reported high-confidence regulatory interactions; green = true positive, red = false negative

(c) BN model of T cell development GRN. Green edge = activation, red edge = repression. Node colours indicate groups of functionally-related genes. AND/OR logic not shown.

Table 4: Boolean logic functions for BN model of T cell development program

Node	Update function
NOTCH_SIGNALING	INPUT_DL4 and Notch1
IL7_SIGNALING	INPUT_IL7 and Il7ra
TCR_SIGNALING	INPUT_TCR and Cd3e and Cd3g and Ptcra and Tcrb and Lck and Zap70
E_PROTEIN	E2a and HEB and not Id3
Bcl11b	NOTCH_SIGNALING and Gata3 and Tcf7 and Runx1
Cd3e	(NOTCH_SIGNALING and Bcl11b and (Gata3 or Ikaros or Tcf7)) and not Pu1
Cd3g	(NOTCH_SIGNALING and Bcl11b and (Gata3 or Ikaros or Tcf7)) and not Pu1
Cebpa	IL7_SIGNALING and not NOTCH_SIGNALING and not Bcl11b and not Gata3 and not Hes1
Deltex	NOTCH_SIGNALING and Gata3
E2a	E2a
Ebf1	(IL7_SIGNALING and E2a) and not NOTCH_SIGNALING
Ets1	(TCR_SIGNALING or (Runx1 and Scl)) and not Pu1
Gata3	((NOTCH_SIGNALING or IL7_SIGNALING) and E_PROTEIN and Myb) or Tcf7 or Runx1) and not (Pu1 and not (NOTCH_SIGNALING or Myb))
Gfi1	E_PROTEIN and Lyl1 and (not Pu1 or not Gfi1b)
Gfi1b	(NOTCH_SIGNALING or E_PROTEIN) and not Bcl11b
HEB	HEB or Scl
Hes1	(NOTCH_SIGNALING or (E_PROTEIN and not Pu1)) and not (Ikaros and TCR_SIGNALING)
Hhex	Scl and Lmo2
Id3	TCR_SIGNALING or (Scl and Lyl1 and not NOTCH_SIGNALING and not Bcl11b and not Gata3 and not Ebf1 and not Pu1)
Ikaros	Ikaros or not Pu1
Il7ra	NOTCH_SIGNALING or E_PROTEIN or (Pu1 and not Gata3)
Kit	(IL7_SIGNALING and (Scl or Lmo2)) and not Bcl11b
Lat	(NOTCH_SIGNALING and E_PROTEIN) and not Pu1
Lck	NOTCH_SIGNALING and (not IL7_SIGNALING or not Pu1)
Lef1	NOTCH_SIGNALING and Tcf7
Lmo2	Pu1
Lyl1	Lmo2 and Pu1
Myb	Scl or not Pu1 or TCR_SIGNALING
Notch1	E_PROTEIN
Ptcra	(E_PROTEIN and NOTCH_SIGNALING and Bcl11b and Myb) and (not IL7_SIGNALING or not (Scl and Lyl1) or not Gata3) and not (Ikaros and TCR_SIGNALING)
Pu1	IL7_SIGNALING and not Gata3 and not Gfi1 and not Runx1 and not Tcf7 and not Bcl11b
Rag1	(E_PROTEIN and NOTCH_SIGNALING and Gfi1) and not Pu1
Runx1	NOTCH_SIGNALING and not (Ikaros and TCR_SIGNALING)
Scl	E2a or Gata3 or Gfi1 or Pu1
Tcf7	NOTCH_SIGNALING or (Tcf7 and not Gata3) or TCR_SIGNALING
Tcrb	Ets1 and Gata3 and Runx1 and NOTCH_SIGNALING
Zap70	TCR_SIGNALING or not Pu1

3.3.2 Steady states of BN correspond to known T cell progenitor states

Previous applications of BN modeling to biological network simulation commonly find that the steady states of the model bear high similarity to experimentally-observed cell types (Bonzanni et al., 2013; Collombet et al., 2017; Krumsiek et al., 2011). To determine whether this property was also true of our constructed BN model of T cell development, asynchronous simulations were performed for all combinations of Notch, IL-7, and pre-TCR signaling inputs. As an example, under +Notch +IL-7 input conditions, a state transition map with 10 526 states (unique single-cell expression profiles) and 12 885 possible transitions between these states was generated (Figure 5a). The state transition map contains five steady states and one SCC. Generally, many steady states of the BN model exhibited high agreement with bulk microarray profiles for known T cell progenitor stages. For example, one steady state found in the absence of input signals exhibits 78% agreement (by Hamming distance) with the binarized microarray profile of a CLP, and another steady state found under +Notch +IL-7 conditions exhibits 94% agreement with the T lineage-committed DN3 stage (Figure 5b). Thus, the constructed BN model can recapitulate the transcriptional phenotypes of known T cell progenitor cell types.

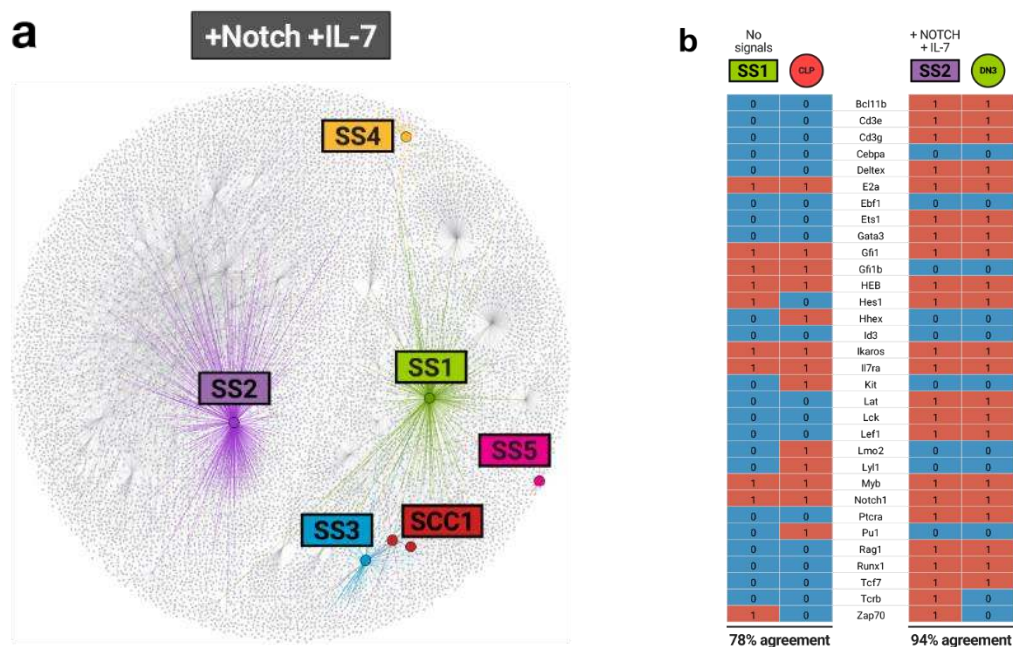


Figure 5: Known T cell progenitor cell types are captured by BN modeling

(a) State transition map produced through asynchronous simulation of T cell development BN model with Notch and IL-7 signaling inputs present. Data represents 4000 random initial conditions and 250 simulation steps per condition. Colours indicate attractors.

(b) Computationally-predicted binary transcriptional profiles for two example steady states agree well with binarized experimentally-measured microarray profiles for common lymphoid progenitors (CLPs) and T lineage-committed DN3 primary cells. Microarray data from ImmGen (Mingueneau et al., 2013) normalized between 0 and 1 across all development stages and alternate fates by k -means clustering. Agreement scored using Hamming distance metric.

3.3.3 BN simulation captures control of developmental progression by environmental signals

To determine whether our BN model captures environmental signal-mediated control of progression through the T cell development program, we assessed the transcriptional space that is accessible from a CLP-like initial state given different combinations of Notch, IL-7, and pre-TCR inputs (Fig. 8). Simulated CLP-like cells were unable to access states beyond DN2, consistent with experimental observations that IL-7 alone is insufficient to specify the T cell lineage (Schmitt et al., 2004). Conversely, *in silico* inclusion of both Notch and IL-7 inputs enabled transcriptional access up to DN3, recapitulating the critical role of Notch signaling in T lineage specification and commitment (Schmitt et al., 2004; Yui and Rothenberg, 2014). Addition of pre-TCR signaling inputs is needed for simulated cells to progress past DN3 to later stages of T cell development. This observation is again consistent with expectations since T cell progenitors must undergo TCR- β selection at the DN3 stage (Yui and Rothenberg, 2014).

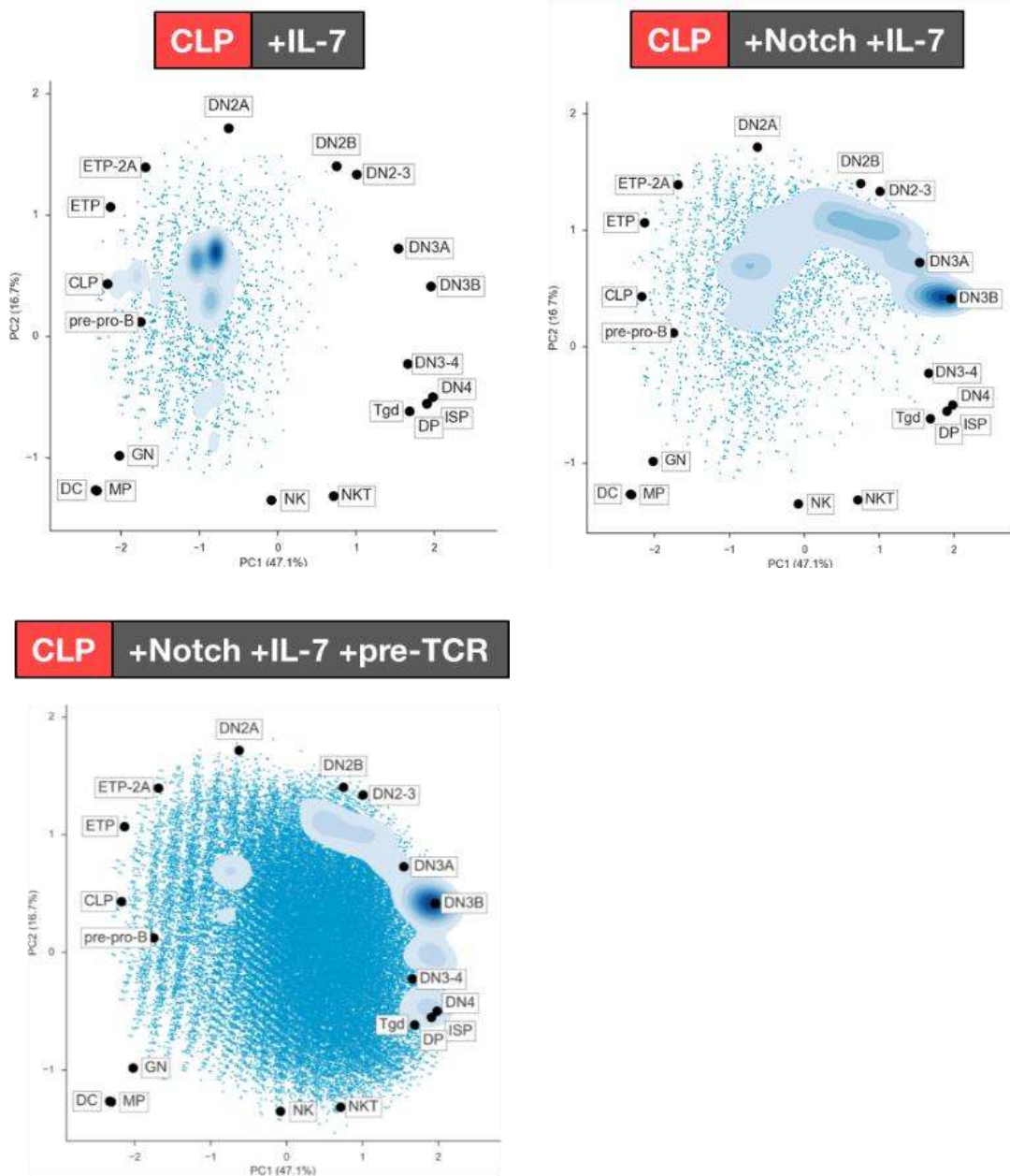


Figure 6: Density plots of predicted transcriptional state space that is accessible to CLPs when stimulated with various combinations of environmental signals
 Produced through asynchronous simulation over 2000 initial conditions and 500 simulation steps. Density contours are weighted by the observed frequency of each state over all simulations. Black dots correspond to reference profiles for T cell progenitor stages and alternate lineages, determined as above. Simulation results are projected onto principal components of reference profiles.

3.3.4 BN modeling predicts developmental effect of knockouts

In addition to simulating different combinations of Notch, IL-7, and pre-TCR signaling inputs, the BN model of T cell development can simulate the knockout or forced expression of specific genes. In this manner, we can predict genes that are necessary for T cell commitment or whose enforced expression leads to arrested T cell development. For example, simulation results for the knockout of *Bcl11b* (*Bcl11b*^{-/-}) given +Notch +IL-7 signaling inputs agree with previous reports of developmental arrest prior to T lineage commitment in *Bcl11b*^{-/-} thymocytes (Ikawa et al., 2010). In particular, asynchronous simulations predict that *Bcl11b*^{-/-} CLPs converge toward an early steady state and cannot progress past the DN2b stage (Figure 7). By contrast, simulated wildtype CLPs are able to reach a DN3b-like steady state.

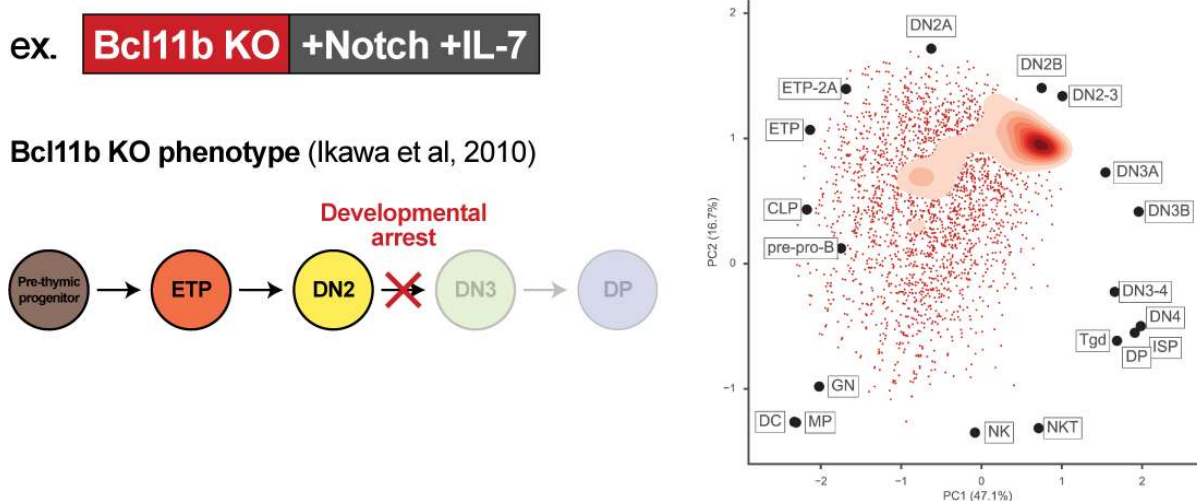


Figure 7: *In silico* knockout of *Bcl11b* recapitulates experimentally observed developmental arrest at DN2 stage

(Left) Genetic knockout of *Bcl11b* has been shown to arrest mouse T cell development at the DN2 stage. (Right) Predicted steady state trajectories from *Bcl11b*^{-/-} CLPs given +Notch +IL-7. Determined by forcing *Bcl11b* = 0 throughout BN simulation over 2000 runs and 500 steps per run. As shown via projection onto principal components, *Bcl11b*^{-/-} CLPs are predicted to arrest prior to the DN2B stage, consistent with previous *in vitro* and *in vivo* observations.

3.3.5 BN modeling predicts multiple transcriptional trajectories toward T cell commitments

In addition to analyzing the steady states and developmental extent of these transcriptional state spaces, we investigated the different paths through these spaces that were observed in individual simulation runs. In graph theory, a path from one node to another in a directed graph without repeated states is a “trajectory”. Using our BN model, we identified 1063 unique trajectories leading from the same CLP-like initial state to the same DN3b-like steady state within the simulated +Notch +IL-7 state transition graph (Figure 8a). This result suggests that there are many transcriptional trajectories that can lead uncommitted blood progenitors toward T cell commitment.

To determine common patterns among these trajectories, we clustered the trajectories in principal component space based on the Fréchet distance metric. Common transcriptional patterns over the simulation were analyzed among the top 6 clusters of trajectories (Figure 8b). In all clusters, genes associated with alternate or stem cell fates (*Pu.1*, *Hhex*, *Lmo2*, etc.) were downregulated early along each trajectory. However, expression of important T cell genes (*Gata3*, *Deltex*, *Cd3e*, etc.) was delayed in some clusters of predicted trajectories but not others. Furthermore, other genes (*Bcl11b*, *Tcf7*, etc.) appeared to fluctuate in their expression before stabilizing to an “ON” level. These patterns suggest that differences among possible trajectories for T lineage specification primarily involve T lineage-promoting genes, whereas rapid silencing of T lineage-antagonizing genes may be a conserved feature of all T cell differentiation trajectories.

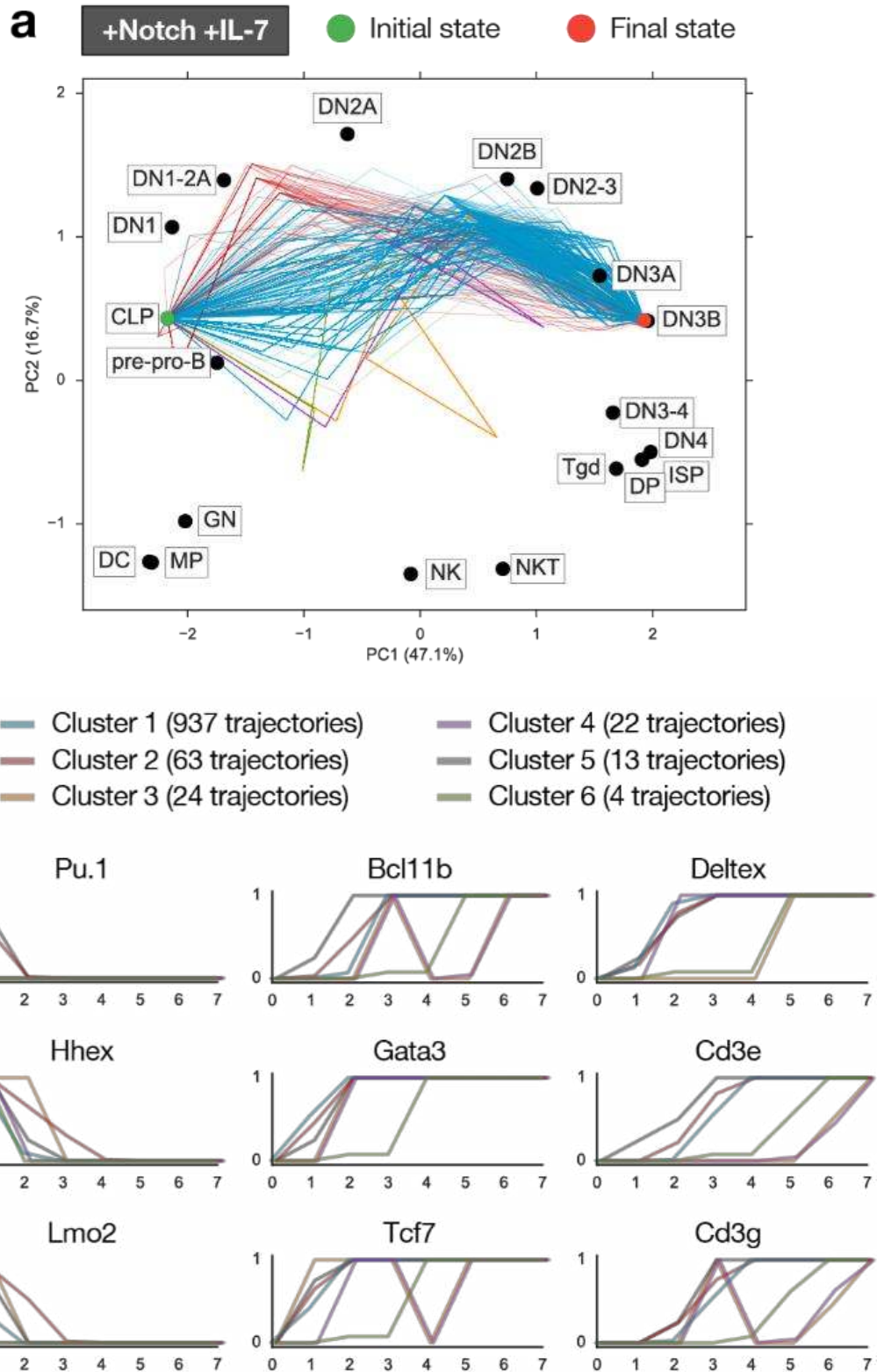


Figure 8: BN modeling predicts multiple transcriptional trajectories toward T cell lineage commitment

(a) Predicted trajectories from CLP to DN3b-like steady state given +Notch +IL-7 signaling, determined using depth-first search on the simulated state transition map. Colours correspond to top-6 clusters of similar trajectories, determined by Fréchet distance and average linkage. (b) Average expression state for a subset of genes in each cluster of similar trajectories shown in (a) over multiple simulation steps.

3.4 Discussion

In this chapter we present a Boolean network (BN) model of the gene regulatory network (GRN) responsible for specifying the T cell lineage in mice. Previous work in the field has focused on static descriptions of the T cell development GRN (Kueh and Rothenberg, 2012) or simulation of small decision-making modules using continuous ordinary differential equations (ODEs) (Manesso et al., 2016). However, the BN model described herein represents the first computable model of the regulatory logic underlying T cell development program in its entirety.

In practice, the assumptions of BN modeling are what facilitated computational modeling of the full T cell development GRN rather than only sub-motifs of the network. Although the source, target, and sign (activation or repression) are known for edges of the T cell development GRN, biochemical rates of binding or transcription remain uncharacterized for most of these edges. The BN modeling framework is well-suited to this level of prior knowledge, since it does not require specification or estimation of rate parameters like a continuous ODE model would. Furthermore, the logical computations inherent to BN simulation are less computationally expensive than the numerical computations demanded by ODE models; thus, simulation of our 37-node BN model is significantly more tractable than an equivalently-sized ODE model would be.

The BN modeling framework facilitates qualitative assessment of the behaviour of complex GRNs consisting of dozens of players and unknown kinetic parameters. This system-level simulation approach represents a key advantage over the static descriptions of GRNs that currently pervade the T cell development field. Specifically, our BN

simulation demonstrates that observed T cell progenitor cell types such as DN3 do in fact correspond to steady state attractors of the underlying GRN (Huang et al., 2005). BN simulation also permitted explicit mapping of the transcriptional space that is accessible to T lineage progenitors given different combinations of IL-7, Notch, and pre-TCR signaling.

A longstanding question within the T cell development field is whether progenitors differentiate from each surface-marker defined stage of T cell development to the next via homogeneous or heterogeneous transcriptional trajectories (Bhandoola and Sambandam, 2006; Yui and Rothenberg, 2014). There has been speculation for over a decade that there may be different or atypical pathways for T cell lineage specification (Bhandoola and Sambandam, 2006). However, it is only recently that computational tools such as BN modeling and experimental techniques such as single-cell transcriptomics have permitted comprehensive investigation of this question. Others within the field have speculated that atypical pathways for T lineage specification arose due to different pre-thymic progenitor cell types that seed the thymus and have dissimilar lineage restriction properties (Petrie and Kincaid, 2005). However, our BN model predicts that this heterogeneity extends to the single-cell level, such that even a single progenitor cell state can potentially follow multiple distinct transcriptional trajectories during T lineage specification.

Our prediction of multiple trajectories leading from the same initial progenitor state to the same final cell type has precedence in other cellular differentiation systems. Theoretical biologists commonly conceptualize differentiation as a process that takes a biological system from one high-dimensional transcriptional attractor to another and predict that many transient pathways leading from the source attractor to the target must be possible (Dealy et al., 2005). Prior to the availability single-cell transcriptomics, Huang et al used timecourse microarray analysis to demonstrate that two chemical signals (all-trans retinoic acid [aTRA] and dimethyl sulfoxide [DMSO]) cause human promyelocytic cells to differentiate via initially divergent transcriptional trajectories that would eventually converge toward the same stable neutrophil attractor state (Huang et al., 2005). Research interest in similar diverging and converging transcriptional trajectories has been

reinvigorated in recent years as single-cell RNA sequencing (scRNA-seq) techniques have matured. For example, Briggs et al employed droplet-based scRNA-seq to demonstrate that mouse embryonic stem cells differentiate into motor neurons via different transcriptional routes in a direct programming protocol (overexpression of 3 TFs; *Ngn2*, *Isl1*, *Lhx3*) than they do in a developmentally-motivated differentiation protocol (timed addition of fibroblast growth factors, retinoic acid, and Sonic hedgehog) (Briggs et al., 2017). Indeed, the fact that certain developmental intermediates can be short-circuited during direct programming protocols is reminiscent of observations of atypical T cell development; *Foxn1* deficient mutants, *Gata3* overexpressing mutants, and a $\text{Kit}^{\text{lo}} \text{HSA}^+$ subset of DN1 progenitors have each been reported to seemingly bypass the DN3 stage of T cell development while still giving rise to DP T cells at reduced efficiencies (Anderson et al., 2002; Porritt et al., 2004; Su et al., 2003; Taghon et al., 2001).

Despite the power of our BN model to predict heterogeneous transcriptional trajectories and accurately recapitulate the transcriptional response of T cell progenitors to extrinsic signals and genetic perturbations, there are limitations to the Boolean network approach when modeling T cell progenitors that merit attention. First, the assumption that all genes can only be fully “ON” or “OFF” poses some challenges when applied to the T cell progenitor system. Gradients of signals (such as IL-7) and gene expression levels are reported to be important for *in vivo* T cell development (Ikawa et al., 2010). These gradual changes in expression levels are also common in qRT-PCR and microarray analysis of T cell development genes in different stages of T cell progenitors, with some genes such as *Tcf7* constantly increasing through to T cell maturation (Mingueneau et al., 2013; Yui et al., 2010). Thus, unlike some earlier differentiation decision points which can be accurately described using binary values, binarization of T cell development genes may be too coarse to permit perfect computational versus experimental correlation.

Second, *in vivo* T cell development also relies on checkpoints external to the T cell development GRN for full commitment and maturation to occur. For example, two T cell receptor-related checkpoints (positive and negative selection) must be passed by DP cells if they are to become mature T cells. Since our model only comprises the transcriptional-

level intracellular decision machinery and excludes physical interactions with the thymic epithelium, this behaviour cannot be accurately captured. In future work, alterations to the Boolean modeling approach or hybrid modeling techniques could be employed to recapitulate these facets of biological T cell development.

Alternatively, one could envision integrating the BN model described here into an agent-based framework, in which each agent corresponds to a cell and incorporates the BN model within it to determine the cell's transcriptional response. By simulating these cellular agents within a spatially-defined thymic niche in which stromal cells provide external cues to the agents depending on their position, one would be able to capture the physical interactions with the thymic niche that mediate selection checkpoints. For example, it would be particularly interesting to computationally examine the dwell time of thymic progenitors in different developmental stages (DN1, DN2, DN3, etc.) for different niche signaling combinations and stroma compositions using such an agent-based modeling approach. We anticipate that multi-scale, multi-framework hybrid models which span all levels of cell fate decision making will facilitate the testing of more complex biological hypotheses. The BN model presented here is a first step towards empowering such efforts.

4 Community software for Boolean network modeling

4.1 Introduction

Previous chapters have outlined how increased experimental throughput and greater availability of computational tools for network inference have expanded our ability to construct predictive systems-level models of gene regulatory networks (GRNs). However, these tools are too often inaccessible to the large fraction of the biological community that lacks computer programming expertise. Additionally, since systems biology research questions typically span multiple experimental modalities and data types, data analysis and modeling must be performed using multiple specialized software packages. Unfortunately, integration of these software packages into complete systems biology pipelines presents a significant user challenge due to inconsistent data interfaces or incompatible computing environment requirements. Altogether, these complications create a barrier to adoption of new systems biology software tools and ultimately slow the progress of biological discovery.

To maximize their utility to the broad research community, new biological datasets and software should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). The Systems Biology Institute (SBI) in Tokyo, Japan, is leading development of a systems biology software platform based on these principles called *Garuda* (Ghosh et al., 2011). *Garuda* (<http://garuda-alliance.org>) enables easy linkage of many computation biology ‘gadgets’ (modules) into novel data analysis pipelines through a shared interface schema. *Garuda* gadgets can be coded in a large variety of programming languages and are accessible through a common repository—the *Garuda Gateway* (<http://gateway.garuda-alliance.org>). Importantly, *Garuda* is designed for use by the broad community of biological researchers, regardless of software programming capability. The advantages of *Garuda* to systems biologists are:




- Ease of discovery of appropriate software tools for their research questions
- Ease of software setup with minimal overhead and no programming requirements
- Ease of data input-output flow from one software tool to the next

As part of my Master's studies, I undertook a 2-month research internship at SBI under the supervision of Dr. Hiroaki Kitano and Dr. Ayako Yachie-Kinoshita. The goal of this internship was to develop a pipeline of *Garuda* gadgets to facilitate simulation and analysis of Boolean network (BN) GRN models (akin to the analyses discussed in previous chapters) and allow their incorporation into larger data analysis pipelines. This chapter describes the *Garuda* gadgets that were developed, as well as how they were applied as part of a computational study of mouse embryonic stem cell (mESC) heterogeneity and fate response.

4.2 Software

Toward our goal of developing a Boolean network modeling and analysis pipeline within *Garuda*, I developed 3 gadgets: “Discretize”, “Boolean Simulation”, and “Boolean SCC Analysis”. To make our pipeline accessible to non-coders, we used the AlgoBuilder graphical user interface framework developed by SBI to wrap our Python scripts into a simple point-and-click interface. These gadgets and associated help documentation were made publicly available for download from the *Garuda Gateway* via the links below. The source code is also available at <https://gitlab.com/stemcellbioengineering/garuda-boolean>. A summary of the completed gadgets in the pipeline is provided in Table 5, and a full description of their features and usage is provided in Appendix A.

Table 5: Summary of *Garuda* gadgets for Boolean network analysis

Gadget	Description	Input	Output
 Discretize	<p>Converts matrices of continuous gene expression data into discrete levels (ex. binary ON/OFF) using <i>k</i>-means</p>	<ol style="list-style-type: none"> 1. Continuous-valued gene expression matrix 	<ol style="list-style-type: none"> 1. Discrete-valued gene expression matrix
 Boolean Simulation	<p>Simulates Boolean network models of GRNs. Starting from user-specified or random initial states, the Boolean logic functions encoded by the model are repeatedly applied (either synchronously or asynchronously) over discrete time steps to produce trajectories of network states.</p>	<ol style="list-style-type: none"> 1. Boolean network rules 2. Initial conditions 3. Configuration parameters 	<ol style="list-style-type: none"> 1. State transition graph 2. State dictionary 3. Edge dictionary
 Boolean SCC Analysis	<p>Analyzes Boolean state transition graphs to identify attractors (including steady states and strongly connected components, SCCs) and their expression profiles. For SCCs, the gadget also calculates a “sustainability” score reflecting the probability of remaining within the SCC over time versus escaping it.</p>	<ol style="list-style-type: none"> 1. State transition graph 2. Configuration parameters 	<ol style="list-style-type: none"> 1. Attractor-annotated state transition graph 2. Expression profiles 3. SCC metrics

4.3 Applications

In addition to being publicly released through *Garuda Gateway*, the aforementioned 3 *Garuda* gadgets were included as part of our recent publication:

Yachie-Kinoshita, A., Onishi, K., Ostblom, J., **Langley, M.A.**, Posfai, E., Rossant, J., and Zandstra, P.W. (2018). Modeling signaling-dependent pluripotent cell states with Boolean logic can predict cell fate transitions. *Molecular Systems Biology* 14, e7952.

This publication reported a Boolean network (BN) model of the mouse embryonic stem cell (mESC) GRN (Figure 9), proposed new metrics for quantifying the stability of heterogeneous attractors of BNs, and facilitated the identification of a novel exogenous signaling combination that robustly generates Cdx2+ Oct4- populations from naïve mESCs. The *Garuda* gadgets described above were used to analyze the state transition graph of this BN model following random asynchronous Boolean simulation of 3 commonly-used mESC culture conditions: LIF+serum (LS), 2i+LIF (2iL), and bFGF+Activin (bF+A) (Figure 10). Releasing our model and simulation analysis software through *Garuda* will enable the stem cell research community to explore and extend our modeling framework without any software programming requirements. Extensibility of our proposed BN model of mESC populations is especially valuable given the large amount of interest in applying logical modeling to embryonic stem cell biology (Dunn et al., 2014; Okawa and del Sol, 2015; Xu et al., 2014).

Importantly, both the mESC model and the mouse T cell development model presented in Chapter 3 are formulated and simulated using the same BN modeling framework. The compatibility of these two models enables a possible future research direction in which the models are combined via *Garuda* to create a BN model of cross-GRN interactions during reprogramming of mouse T cell progenitors to induced pluripotent stem cells (iPSCs). The experimental motivation for modeling T cell progenitor-derived iPSCs and a preliminary framework for integrating the mESC and T cell development BN models are discussed in Chapter 6.2.

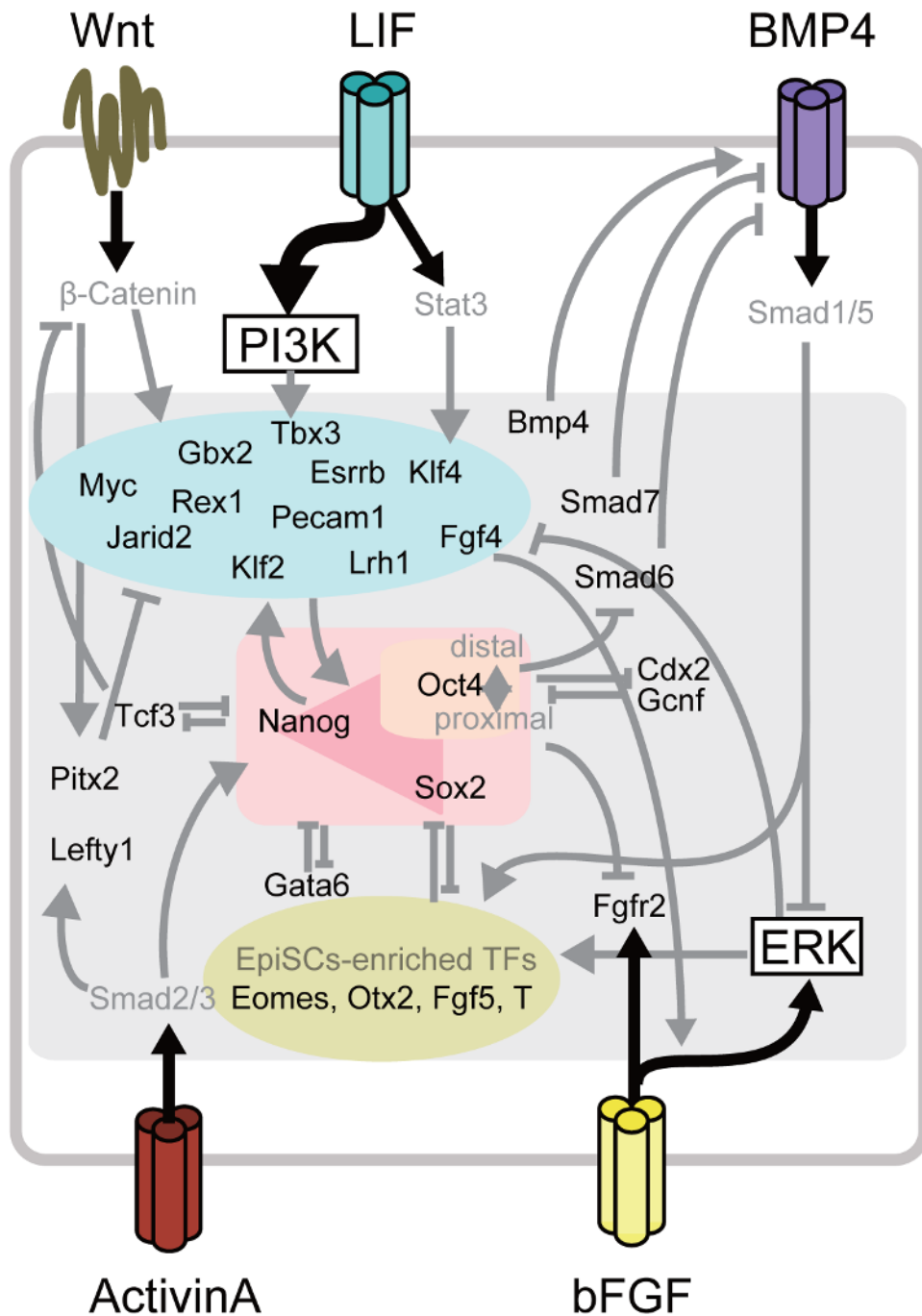


Figure 9: A schematic of the defined PSC gene/signal regulatory network model
 Reprinted from Yachie-Kinoshita, A., Onishi, K., Ostblom, J., **Langley, M.A.**, Posfai, E.,
 Rossant, J., and Zandstra, P.W. (2018). Modeling signaling-dependent pluripotent cell
 states with Boolean logic can predict cell fate transitions. *Molecular Systems Biology* 14,
 e7952.

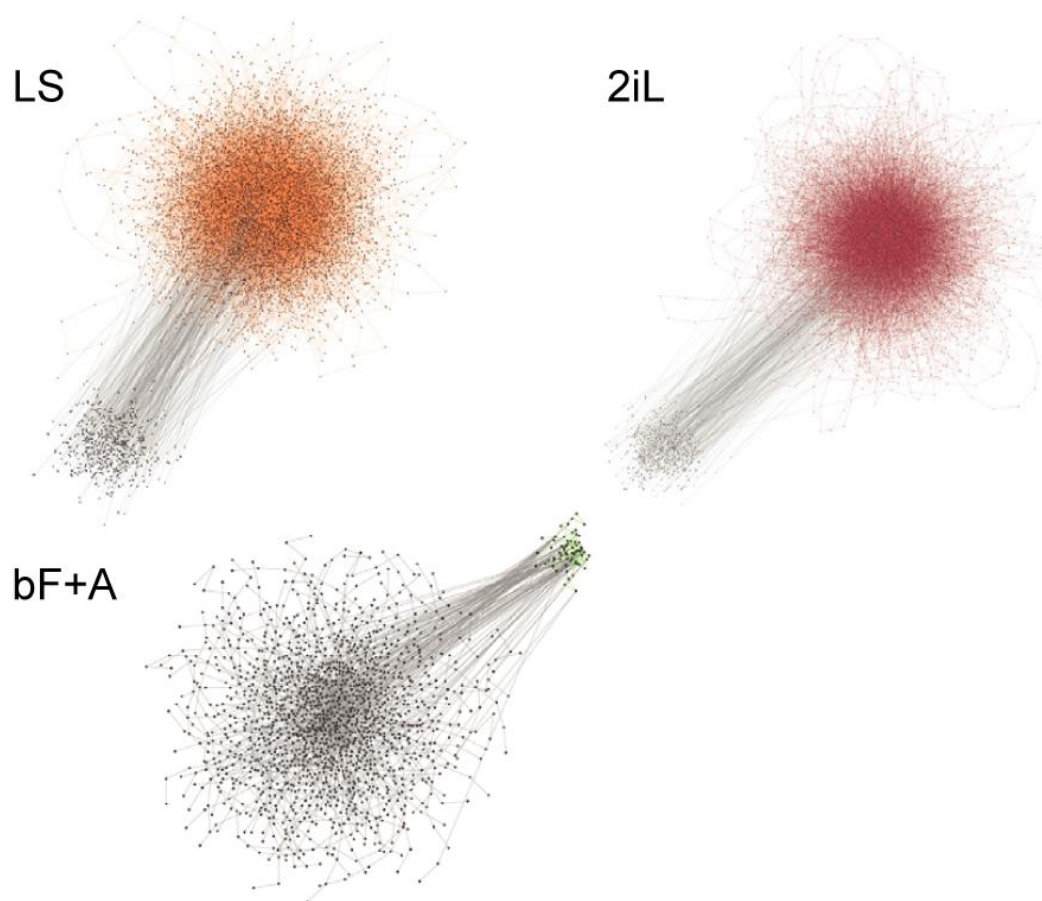


Figure 10: Transcriptional state spaces of mouse ESC network resulting from random asynchronous Boolean simulation

Condition-dependent pluripotent cell populations correspond to strongly connected components (SCCs) in the state transition graphs of asynchronously updated Boolean models. Gray dots represent unique profiles, and edges represent state transitions among the profiles. Coloured edges indicate the transitions within population-specific SCCs. The number of simulations and the number of steps in each simulation were 300-100, 300-100, 300-300 for LS, 2iL, and bF+A condition, respectively. State transition graphs for each condition were calculated using the “Boolean Simulation” gadget, and SCCs were identified using the “Boolean SCC Analysis” gadget.

Reprinted from Yachie-Kinoshita, A., Onishi, K., Ostblom, J., **Langley, M.A.**, Posfai, E., Rossant, J., and Zandstra, P.W. (2018). Modeling signaling-dependent pluripotent cell states with Boolean logic can predict cell fate transitions. *Molecular Systems Biology* 14, e7952.

5 Differentiation context-dependent comparison of T cell progenitor transcriptional patterns

5.1 Introduction

T cells develop in the thymus throughout an organism's life. In mice, thymopoiesis begins at approximately E13.5 and continues through to adulthood (David-Fung et al., 2006). However, there is amounting evidence that T cell development exhibits different stage-wise kinetics and genetic requirements in fetal mice compared to adults (as discussed in Chapter 1.3.4). Of particular interest are reports of shortcuts and detours from the canonical series of T cell progenitor stages (Anderson, 2006; Porritt et al., 2004; Su et al., 2003; Taghon et al., 2001). These reports suggest that certain aspects of the T cell development program are inherently plastic and permit multiple pathways for T lineage specification. However, these differences across developmental contexts have not yet been examined at the level of single-cell gene expression.

Meanwhile, characterization of the thymic niche has informed development of new *in vitro* platforms for T cell differentiation from hematopoietic stem and progenitor cells (HSPCs). The OP9-DL4 (or OP9-DL1) system has had a large impact on enabling in-depth observation of hematopoietic progenitors as they proceed toward the T cell fate (Brauer et al., 2016). However, OP9-DL4 requires the use of both serum and stromal cells, and thus it is difficult to reproducibly control the environmental signals the cells observe and define the specific cues that enable robust T lineage progression. A serum- and stroma-free platform for T lineage differentiation would provide a powerful advance for controlled characterization of the effect of various niche molecules and supplemented cytokines on T cell differentiation. A platform that meets these criteria would enable future engineering studies of the isolated effects of individual cytokines, small molecules, and matrix components on the T lineage differentiation, and specifically how these factors lead to increased or partial activation of the T cell development GRN. Furthermore, in line with previous studies of T cell developmental plasticity across different *in vivo* contexts, it would be interesting to investigate whether T cell progenitors grown *in vitro* follow the same series of transcriptional and developmental events of *in vivo* thymopoiesis, or instead follow pathways unique to the *in vitro* setting.

In this chapter, I outline my contributions toward development of the DL4+VCAM differentiation platform, which facilitates mouse and human T cell progenitor differentiation without serum or stromal feeder cells. Next, I highlight the importance of studying T cell differentiation pathways at a single-cell transcriptional level by demonstrating that sorted surface marker-defined stages of early T cell progenitors that are widely approached as homogeneous populations actually comprise heterogeneous and overlapping transcriptional states as measured by single-cell qRT-PCR. Finally, I compare single-cell transcriptional patterns of developing T cell progenitors across the contexts of fetal thymopoiesis, adult thymopoiesis, and DL4+VCAM *in vitro* differentiation using single-cell RNA sequencing. These results provide preliminary support of the BN model prediction of multiple distinct transcriptional trajectories for T lineage specification, and will enable further refinements to the BN model itself to better capture these heterogeneous and context-dependent observations. Overall, the DL4+VCAM platform and single-cell transcriptomics represent potentially powerful and complementary tools for investigating additional factors that influence T cell differentiation and increasing the resolution at which we understand the T cell development program.

5.2 Methods

5.2.1 Primary tissue dissection

Primary adult thymocytes were obtained from 8-week old adult male CD1 mice. Following CO₂ asphyxiation, thymi were removed and placed in Hank's Balanced Salt Solution (HBSS; Invitrogen) containing 2% fetal bovine serum (FBS; Invitrogen) (abbreviated as 'HF'). Extracted thymi were then pushed through a 40 µm filter to obtain single-cell thymocyte suspensions. To reduce the frequency of erythrocytes and mature T cell populations, cells were subjected to depletion for TER-119, CD4, and CD8 using the EasySep magnetic depletion system (STEMCELL Technologies) and biotin-conjugated antibodies (Biolegend) according to the manufacturer's instructions.

Fetal livers and fetal thymi were isolated from decapitated E13.5 CD1 mouse embryos using surgical forceps and placed in HF. Fetal thymi were processed using a 40 μm filter and magnetically depleted for TER-119, CD4, and CD8 and described above. Fetal livers were disrupted into a single cell suspension by gentle passing through a 21-gauge needle. Subsequently, fetal liver cells were subjected to TER-119 depletion by EasySep magnetic sorting (STEMCELL Technologies) according to the manufacturer's instructions and subsequently cryopreserved in 50% Iscove's Modified DMEM (IMDM; Invitrogen), 40% fetal bovine serum (FBS), and 10% dimethyl sulfoxide (DMSO).

5.2.2 *In vitro* differentiation of fetal liver HSPCs toward T cell lineage

Cryopreserved TER-119-depleted fetal liver cells were thawed and stained for HSPC sorting in HF at 1×10^7 cells/mL. Cells were blocked against non-specific binding with 1% anti-Fc receptor antibody (Fc-block, BD Biosciences) and stained with PE anti-mouse Sca-1 and APC anti-mouse c-Kit antibodies (BD Biosciences) for 20 minutes on ice. Dead cells were excluded using 7-aminoactinomycin D (7-AAD; Life Technologies). Cells were sorted using the BD Influx cell sorter.

Sorted Sca-1⁺ cKit⁺ murine HSPCs were cultured at 3.1×10^3 HSPCs/cm² (corresponding to 1000 cells/well) in DL4 (10 $\mu\text{g}/\text{mL}$) and VCAM-1 (2.32 $\mu\text{g}/\text{mL}$) coated 96-well plates in serum-free Iscove modified Dulbecco medium (Gibco) with 20% bovine serum albumin, insulin, and transferrin serum substitute (BIT; STEMCELL Technologies), 1% Glutamax (Gibco), and 1 $\mu\text{g}/\text{mL}$ low-density lipoproteins (Calbiochem). IMDM+BIT serum-free medium was supplemented with 50 ng/mL Stem Cell Factor (SCF; R&D Systems), 10 ng/mL FMS-like Tyrosine Kinase 3 Ligand (Flt3L; R&D Systems) and 10 ng/mL Interleukin-7 (IL-7; R&D Systems).

5.2.3 Live imaging of differentiating T cell progenitors

Sorted Sca-1+cKit+ HSPCs were seeded at low density (200 cells/well) into triplicate wells of 96-well plates coated with different substrates. After 6 days of culture, cells were stained with conjugated antibodies for CD25-APC and CD44-PE (1:500 dilution) at 37°C for 1 hour. Live cell imaging was then performed without washing on the AxioObserver Z1 (Zeiss) platform in 5% CO₂ and 37°C controlled conditions. Brightfield images were captured at 5-minute intervals over 24 hours using a 10x 0.3 NA air objective. To minimize phototoxicity and photobleaching, images in the fluorescent APC and PE channels were acquired at longer 30-minute (or 60-minute) intervals. Image acquisition and processing was performed using ZEN 2012 blue edition software (Zeiss). Manual tracking was performed using Image-J software. Cells were tracked within 3 unique DL4 only wells and 3 unique DL4+VCAM wells. Manual tracking was performed on 43 cells in the DL4 only condition (15, 10 and 18 cells per well) and 69 cells in DL4+VCAM condition (30, 14 and 25 cells per well).

5.2.4 Bulk quantitative real-time PCR

Sorted Sca-1⁺cKit⁺ murine HSPCs were seeded on no coating, 10 µg/mL DL4, 2.32 µg/mL VCAM-1, and DL4+VCAM-1 at 20,000 cells/well in 96-well plates and were collected at 24 and 48 hours of culture using multiple PBS rinses. CD34⁺ human umbilical blood cells were seeded in the same conditions but using 10 µg/mL DL4 and 2.32 µg/mL VCAM-1. Human cells were collected at 24, 48, and 96 hours of culture. Cells were lysed and RNA was isolated using the PureLink RNA Micro Kit (Invitrogen) according to the manufacturer's protocol. RNA was converted to cDNA using SuperScript III Reverse Transcriptase (Invitrogen) according to the manufacturer's protocol, and amplified together with respective primers in FastStart SYBR Green Master Mix (Roche). Thermocycling and quantification was performed using the QuantStudio 6 Flex (Applied Biosystems). Relative expression of individual genes was calculated by the delta cycle threshold (Δ -Ct) method with the expression of β -actin as an internal reference. PCR primer sequences are available in Supplementary Table 3. Nonparametric

Kruskal–Wallis tests with post hoc Dunn’s analysis were performed in R (version 3.2.5) to determine significant differences between multiple groups.

5.2.5 Single-cell qRT-PCR

Primary ETP, DN2A, and DN2B thymocytes from 6- to 8-week-old CD1 mice were analyzed using single-cell qRT-PCR. Thymocytes were pooled and magnetically depleted for Lin markers (CD4, CD8a, TCR β , TCR $\gamma\delta$, CD11b, CD11c, CD19, NK1.1, GR-1, TER-119). Cells were then stained for CD25 (PE-Cy7), CD44 (PE), c-KIT (APC), and the aforementioned Lin markers (APC-Cy7) and sorted by FACS. Following sorting, cells from each population were suspended in HF and C1 Loading Reagent according to the manufacturer’s instructions, then captured on a Fluidigm C1 Single-Cell Auto Prep IFC for Preamp (5-10 μ m). We were able to achieve a high capture efficiency, such that 75-95% of the capture sites on each C1 IFC contained a single live cell. Following lysis, reverse transcription, and pre-amplification with pooled primers on the C1 IFC, cDNA was harvested in 7 μ L total volume to maximize concentration and qRT-PCR sensitivity. qRT-PCR was then performed on the Fluidigm Biomark for 30 cycles in technical duplicate. 48 primer pairs were used (Supplementary Table 3), targeting genes included in our BN model as well as additional T lineage genes, alternate blood lineage genes, surface marker genes, and housekeeping genes.

5.2.6 Single-cell RNA-sequencing

Single-cell cDNA libraries were prepared using the 10X Chromium controller (10X Genomics) and Chromium Single Cell 3’ reagents according to the manufacturer’s instructions. Input cells for each sample were pooled from at least 3 independent biological replicates (adult thymocytes—3 adult 8-week-old male CD1 mice, fetal thymocytes—E13.5 embryos from 3 pregnant CD1 mothers, DL4+VCAM—input fetal liver HSPCs from E13.5 embryos from 3 different mother mice and cultured separately). Immediately prior to microfluidic capture on the 10X Chromium, cells were sorted by

FACS for CD45+ 7-AAD- live blood cells and suspended in HF at a concentration of 2.37×10^5 cells/mL ($\sim 8.3 \times 10^3$ cells in 35 μ L volume) per manufacturer's guidelines for our intended single-cell capture rate. Following single-cell cDNA library generation, samples were 3' sequenced together on an Illumina Nextseq. Samples were combined in multiplex on each sequencing run to mitigate potential batch effects. Raw sequence data was processed to form gene-barcode expression matrices using the CellRanger pipeline (10X Genomics). Expression matrix processing was performed using the "Seurat" package for R Bioconductor (version 2.3.2) (Butler et al., 2018), including manual filtering of low-quality cells by UMI count and mitochondrial gene presence, log-normalization and scaling of the raw count values, and correction for confounding effects of UMI count differences via regression. Dimensionality reduction was performed using diffusion maps, as implemented in the "destiny" package for R Bioconductor (version 2.10.2) (Angerer et al., 2016).

5.2.7 Flow cytometry

Surface marker staining for mouse experiments was performed with conjugated rat anti-mouse antibodies (BD Biosciences) at 1:400 dilution. All samples were analyzed on FACS LSR Fortessa flow cytometer (BD Biosciences). Cells were washed twice with HF prior to analysis and dead cells were excluded using 7-AAD (Life Technologies) at 1:1000 dilution.

5.3 Results

5.3.1 Characterization of T cell development gene expression dynamics and cell motility during DL4+VCAM differentiation

Future validation of the transcriptional trajectories predicted by the BN model and identification of the key niche factors that promote specific trajectories would benefit from a fully-defined minimal platform for differentiating T cell progenitors. In support of this goal, I assisted in characterizing a novel serum- and stromal cell-free *in vitro* platform for differentiation of mouse and human T cell progenitors that was developed by our lab, referred to here as DL4+VCAM (Shukla et al., 2017) (Figure 11a). The platform represents a minimal engineered thymic niche, including:

- Provision of plate-bound DL4-Fc protein in place of DL4 expression by the thymic epithelium (or by OP9 bone marrow stromal cells, as previously reported)
- Addition of vascular cell adhesion molecule-1 (VCAM), which is abundant in the thymic niche and has been implicated as a stromal matrix for thymic migration of T cell progenitors (Petrie and Zúñiga-Pflücker, 2007)
- Media supplementation of interleukin-7 (IL-7), stem cell factor (SCF, Kit ligand), and FMS-like tyrosine kinase 3 (Flt-3) which play critical roles in maintaining thymocyte viability and expansion (Hosokawa and Rothenberg, 2018)

The DL4+VCAM platform enabled high-yield production of mouse T cell progenitors from mouse fetal liver-derived Sca1+ Kit+ HSPCs (~60% CD25+ CD90+ proT cell frequency after 7 days of differentiation, versus ~40% on DL4 alone) (Figure 11b). T lineage differentiation occurred exceptionally quickly using DL4+VCAM, with ~5% of cells reaching the CD25+ CD44- DN3 stage after just 48 h of culture (versus ~2% on DL4 alone) (Figure 11c-e). The platform also enabled high-yield production of human T cell progenitors from CD34+ human umbilical cord blood (~25% CD7+ CD34- cell frequency after 14 days of differentiation, versus ~5% on DL4 alone). Importantly, these differentiated human T cell progenitors successfully engraft the thymi of humanized mice and mature into cytokine-producing CD3+ T cells. Since the DL4+VCAM platform is a

fully-defined system without serum or stromal cells, it is more amenable to clinical application than other existing T cell differentiation protocols. Overall, the DL4+VCAM platform represents a novel context for T cell progenitor differentiation and can help support translation of T cell-based therapies to the clinic (Shukla et al., 2017).

My specific contributions to this publication were to characterize the mechanism of action of DL4+VCAM, and particularly how VCAM might synergize with DL4 to enhance T cell progenitor yields. Examination of key nodes in the T cell-development gene regulatory network (GRN) in sorted HSPCs within the first 48 h of interaction with DL4 and VCAM revealed rapid upregulation of downstream targets (including *Hes1*, *Deltex*, *Gata3*, and *Tcf7*) of the activated Notch1 intracellular domain (NICD) in the presence of DL4 and VCAM, compared with DL4 alone (Figure 12). In contrast, the myeloid transcription factor PU.1 (encoded by *Spi1*) was more rapidly downregulated within 48 h on DL4+VCAM than on DL4 alone, while the HSPC-associated gene *E2a* (or *Tcf3*) remained unaffected (Figure 12). Thus, VCAM synergistically interacted with DL4 to increase T cell progenitor yields and purity by inducing stronger activation of downstream Notch pathway genes associated with the T cell development GRN.

We next sought to confirm that DL4+VCAM had similar effects on the expression of key nodes in the T cell development GRN in human HSPCs, as we had seen with mouse HSPCs. DL4 and VCAM synergistically enhanced Notch target gene expression compared with DL4 alone (Figure 13). Interestingly, however, the upregulation dynamics observed in human cells were different from those observed in mouse cells. *DELTEX* and *GATA3* were rapidly upregulated within 48 h and showed sustained increases up to 96 h (Figure 13). In contrast, *BCL11B* required 96 h of stimulation before significant enhancement relative to DL4 alone was observed (Figure 13). Thus, synergistic interactions of VCAM with DL4 increased Notch signaling activity in human HSPCs and led to stronger transcriptional activation of downstream T cell development genes, similar to our observations in the mouse system.

We next sought to identify potential biophysical causes that led to increased Notch signaling activity and T cell GRN activation. Toward this, I demonstrated that VCAM

affects DN T cell motility using live-cell imaging. Manual tracking of single cells from day 6 to day 7 with concomitant discrimination of DN1, DN2, and DN3 phenotypes by fluorescent staining revealed that VCAM significantly increased the velocity of DN1 ($P = 0.00001$) and DN3 cells ($P = 0.0006$) compared with DL4 alone (Figure 14a). DN2 single-cell velocities could not be quantified as they typically grew as small aggregates in the engineered thymic-like niche (Figure 14b). We propose that this increased motility may lead to greater Notch signaling activity, either due to enhanced generation of the mechanical force needed to trigger catalytic release of the intracellular domain of Notch, or by increasing the amount of Notch ligand that cells are exposed to per unit time (Figure 14c).

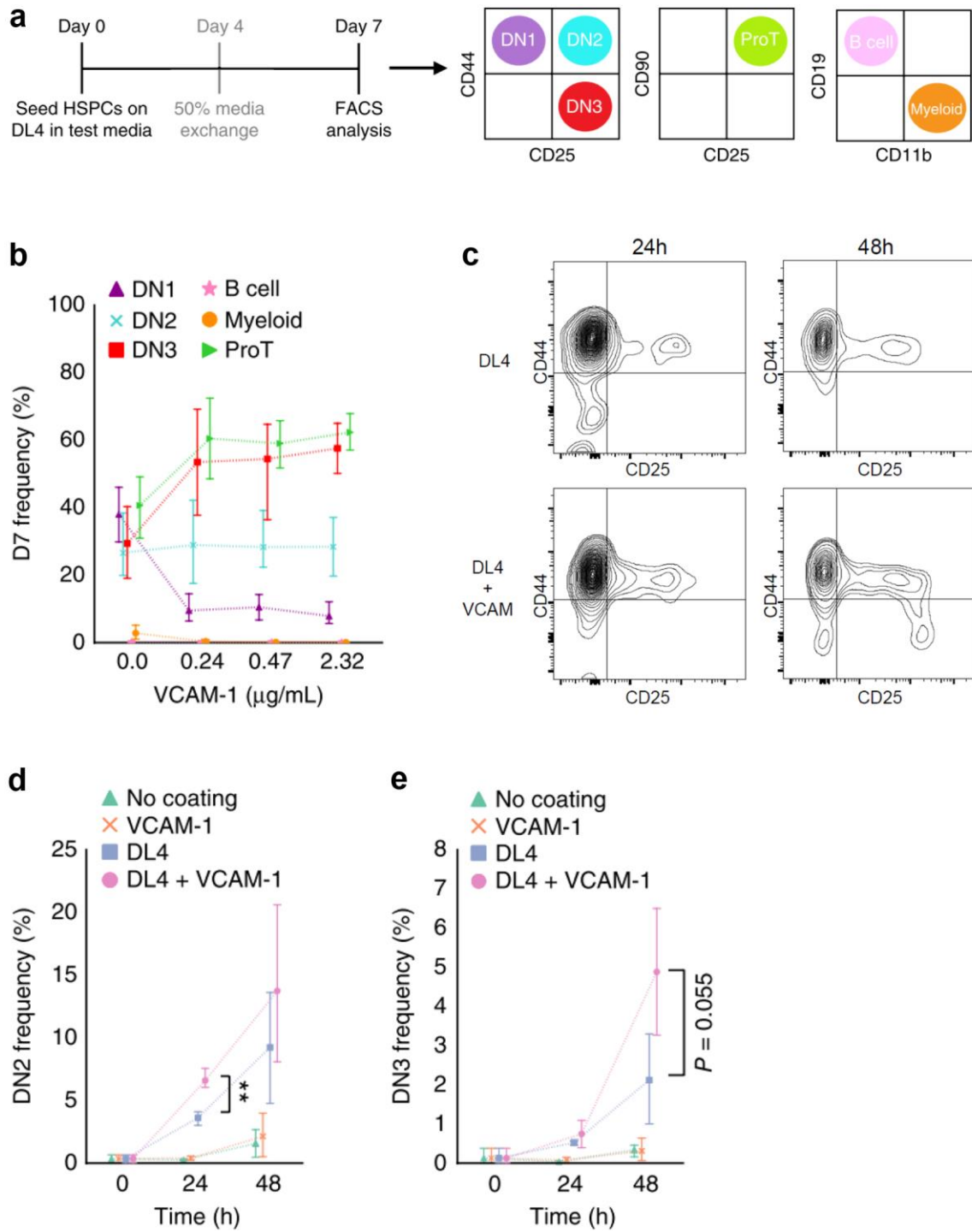


Figure 11: DL4+VCAM yields robust mouse T cell progenitor differentiation with accelerated kinetics

(a) Schematic for 2D coated DL4 (+VCAM) assay. Re-feeding by 50% media exchange was eliminated through assay optimization to enable 7 days of uninterrupted culture

(b) Frequencies of the T cell progenitor types, B cells, and myeloid cells after 7 days of culture of sorted Sca-1+cKit+ mouse HSPCs on DL4-Fc alone or DL4-Fc with increasing concentrations of VCAM (n = 3).

(c) Rapid DN2 and DN3 proT cell differentiation within first two days of culture on DL4+VCAM than DL4 alone. Flow plots are representative of an HSPC sample after 24 and 48 hours after culture on DL4 vs. DL4+VCAM (n = 4).

(d, e) Frequency of DN2 (d) and DN3 (e) cells over 24 h and 48 h of culture time after sorted Sca-1+cKit+ HSPCs were seeded on no coating, 2.32 $\mu\text{g/mL}$ VCAM, 10 $\mu\text{g/mL}$ DL4, or DL4 + VCAM (n = 4).

Data generated and analyzed by S. Shukla. Reprinted from: Shukla, S., **Langley, M.A.**, Singh, J., Edgar, J.M., Mohtashami, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W.

(2017). Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like 4 and VCAM-1. *Nature Methods* 14, 531–538.

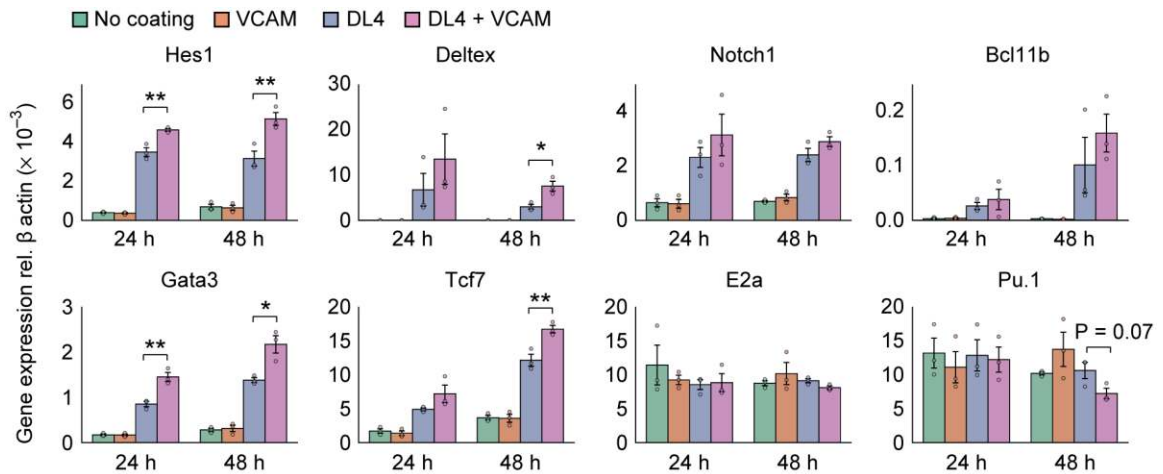


Figure 12: DL4+VCAM differentiated mouse T cell progenitors express Notch target genes at higher levels than DL4-only

qRT-PCR gene expression analysis of downstream Notch pathway genes, an HSPC gene (*E2a*), and a myeloid lineage gene (*Pu.1*) after sorted mouse fetal liver HSPCs were cultured for 24 h or 48 h on each of 4 conditions: no coating, 2.32 μ g/mL VCAM, 10 μ g/mL DL4, and DL4+VCAM. Data represent mean \pm s.e.m. *, $P < 0.05$; **, $P < 0.01$.

Reprinted from: Shukla, S., **Langley, M.A.**, Singh, J., Edgar, J.M., Mohtashami, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W. (2017). Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like 4 and VCAM-1. *Nature Methods* 14, 531–538.

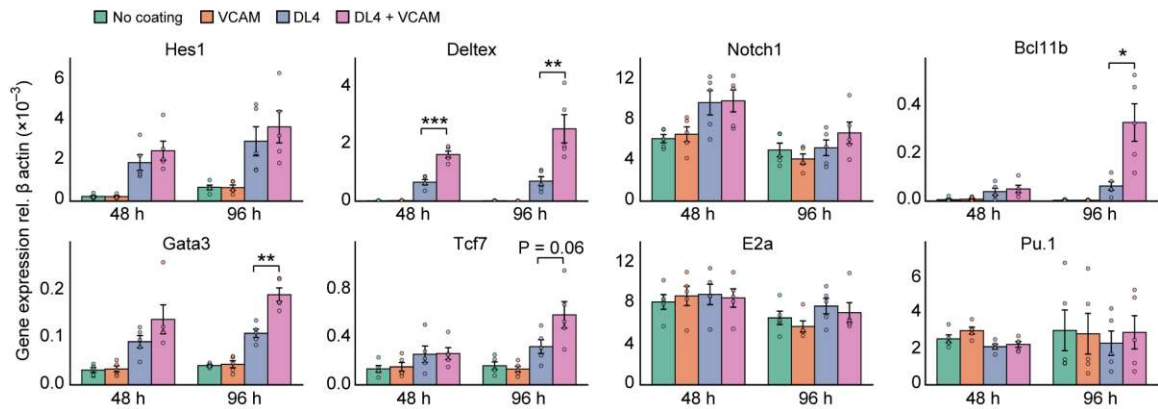


Figure 13: DL4+VCAM differentiated human T cell progenitors express Notch target genes at higher levels than DL4-only

qRT-PCR gene expression analysis of downstream Notch pathway genes, an HSPC gene (*E2A*), and a myeloid lineage gene (*PU.1*) after sorted CD34+ human cord blood cells were cultured for 48 h or 96 h on each of 4 conditions: no coating, 2.32 μ g/mL VCAM, 10 μ g/mL DL4, and DL4+VCAM. Data represent mean \pm s.e.m. *, $P < 0.05$; **, $P < 0.01$. Reprinted from: Shukla, S., Langley, M.A., Singh, J., Edgar, J.M., Mohtashami, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W. (2017). Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like 4 and VCAM-1. *Nature Methods* 14, 531–538.

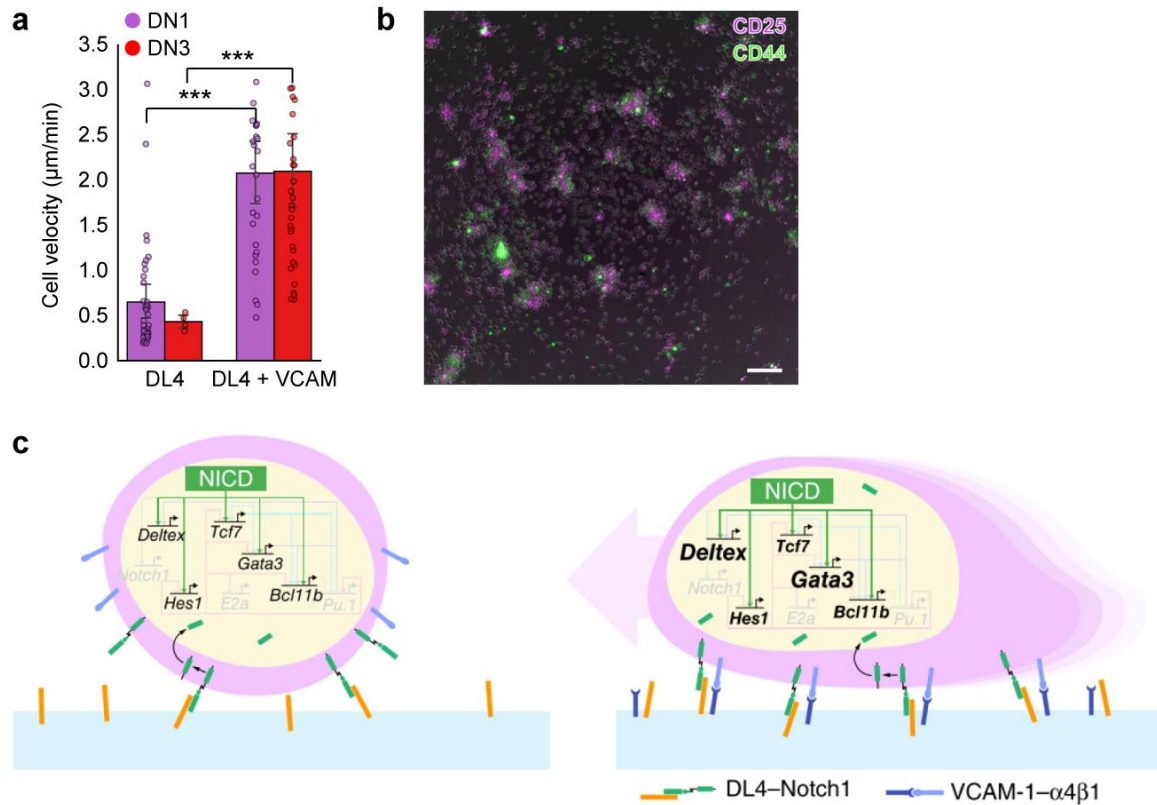


Figure 14: T cell progenitors exhibit greater motility on DL4+VCAM vs. DL4 only

(a) Averaged cell velocity of DN1 and DN3 cells on DL4 alone or DL4+VCAM from days 6 to 7 of culture ($n = 3$).

(b) Representative still image from live imaging of differentiating day 7 progenitor T cells in the DL4+VCAM engineered thymic niche. Cells were stained for CD25 (magenta) and CD44 (green) and merged with bright-field. Scale bar, 100µm.

(c) Schematic of proposed mechanism. DL4 (orange) activates Notch1 receptor (green) on HSPCs, causing translocation of Notch intracellular domain (NICD) to the nucleus and activation of the Notch GRN (left). When DL4 is co-presented with VCAM (right), α 4 integrin receptors (light blue) expressed on HSPCs engage with VCAM (dark blue), leading to higher activation of downstream Notch target genes, increased motility, and accelerated commitment to the T cell fate.

Panel (c) illustrated by J. Ma. Reprinted from: Shukla, S., Langley, M.A., Singh, J., Edgar, J.M., Mohtashami, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W. (2017).

Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like 4 and VCAM-1. *Nature Methods* 14, 531–538.

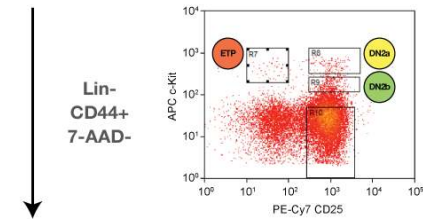
5.3.2 Surface marker-defined stages obscure transcriptional heterogeneity among primary thymocytes

Previous studies have demonstrated that certain surface marker-defined stages of the mouse T cell development program (such as DN1) comprise a heterogeneous mix of subpopulations that exhibit functional differences in their differentiation potential (Porritt et al., 2004). These subpopulations have typically been discriminated by flow cytometry. More recently, however, single-cell qRT-PCR has emerged as a powerful tool for identifying heterogeneity among known cell types (Hamey et al., 2016). We applied single-cell qRT-PCR to test for transcriptional heterogeneity among the earliest surface-marker defined stages of the mouse T cell development program: ETP, DN2A, and DN2B (Figure 15a). Primary thymocytes from adult mice were sorted by FACS and subsequently captured as single cells using the Fluidigm C1. Single-cell qRT-PCR was performed for a panel of 48 genes, including those present in our BN model of T cell development, genes that encode stage-specific surface markers, additional genes corresponding to alternative blood lineages, and housekeeping genes (Supplementary Table 3).

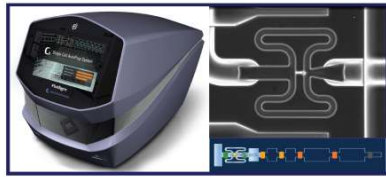
The expression of each gene relative to *Bactin* was calculated using the delta-Ct method for each single cell sample (Figure 15b). The expression patterns of many genes agree with expectations from previous bulk transcriptional experiments. For example, *Ii2ra* (codes for CD25) expression levels are significantly increased in the DN2A and DN2B populations, consistent with the surface marker definitions of these populations. Similarly, *Kit* decreases significantly from DN2A to DN2B, mimicking the drop in c-KIT surface protein levels used to separate these populations in FACS. Important T cell lineage transcription factors including *Gata3*, *Tcf7*, and *Bcl11b*, as well as the T cell co-receptor genes *Cd3e/Cd3g* are detected in the DN2A and DN2B populations, but not in ETP. This is consistent with increased specification to the T cell lineage at the DN2 stage. *Notch1* levels begin to decrease in the DN2B population, which coincides with reports that Notch signal dependence is lost as cells progress to the DN3 stage (Hosokawa and Rothenberg, 2018). Finally, expression levels of alternate blood lineage

genes such as *Pu.1* and *Hhex* are detected at lower frequency in the DN2B population, supporting loss of potential for other blood lineages at this stage.

Plotting the single cell gene expression data on principal components (Figure 15c) produces a pattern analogous to that of bulk microarray profiles. The ETP and DN2B populations are clearly separable by principal components, supporting the notion that these are highly distinct T cell progenitor states. Interestingly, however, the DN2A population interleaves with both ETP and DN2B regions (Figure 15c), even though these populations do not overlap in terms of their FACS gates (Figure 15a). This provides evidence of significant transcriptional heterogeneity within the DN2A surface marker-defined population.

a Primary thymocytes**FACS isolation****Fluidigm C1**

Capture single live cells
Lyse, RT, and pre-amplify

**Fluidigm Biomark**

qRT-PCR for 48 genes

Assess transcriptional heterogeneity
and population overlap

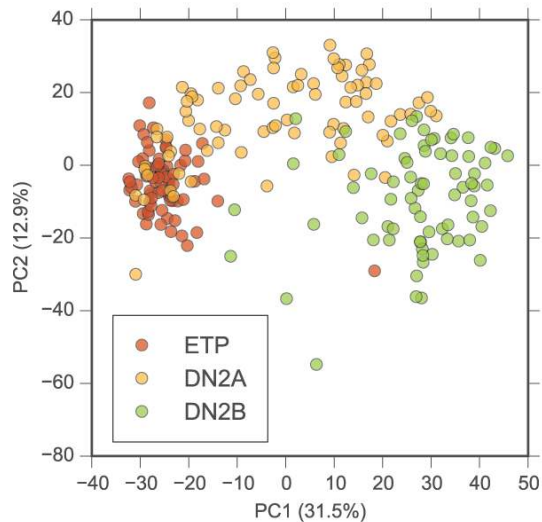
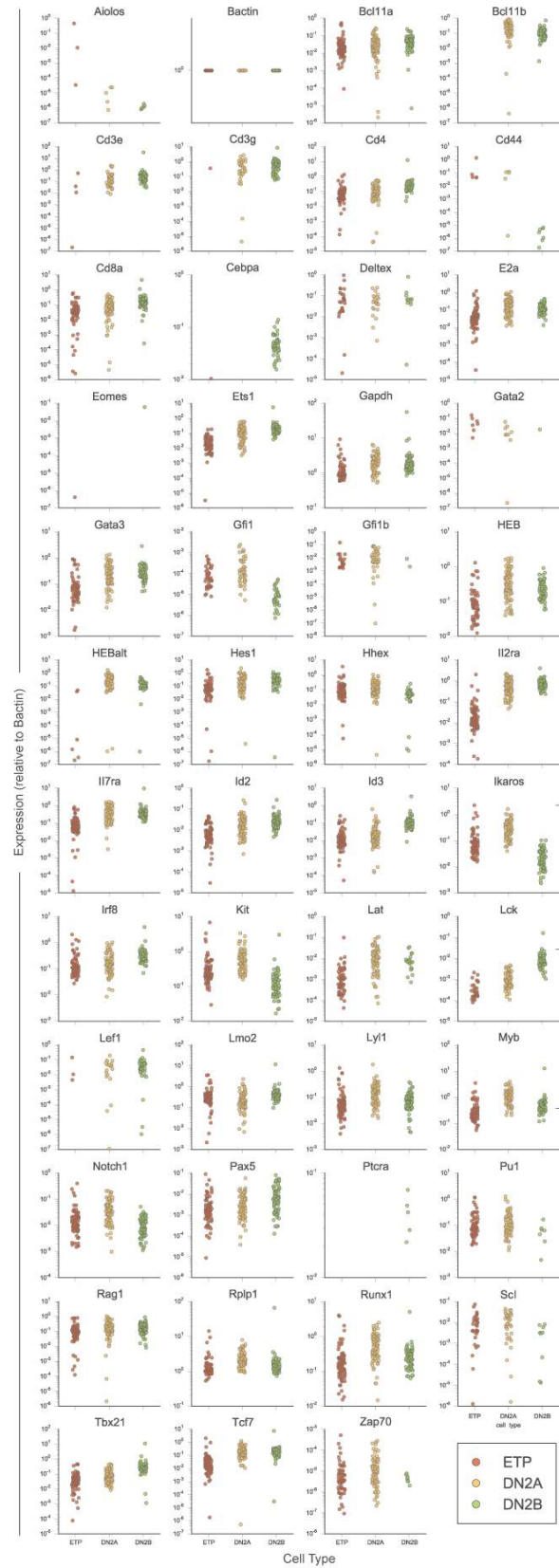
c Principal component analysis**b**

Figure 15: Single-cell qRT-PCR analysis of transcriptional heterogeneity in ETP, DN2A, and DN2B T cell progenitors

(a) Summary of experimental plan for single-cell qRT-PCR gene expression analysis (Fluidigm C1 and Biomark)

(b) Single cell gene expression levels (relative to Bactin) for sorted primary thymocytes

(c) Principal component plot of single cell gene expression profiles for sorted primary thymocytes. The DN2A population interleaves the ETP and DN2B populations.

Red=ETP, yellow=DN2A, green=DN2B.

5.3.3 Single-cell RNA sequencing reveals transcriptional differences between primary and *in vitro* differentiated T cell progenitors

As discussed in Chapter 3.3.5, the BN model of T cell development we developed predicts that there are multiple distinct transcriptional trajectories that uncommitted hematopoietic progenitors could follow toward T lineage commitment. This prediction raised the following questions:

1. Can all of the predicted trajectory patterns be observed experimentally, or are some false artefacts of our BN modeling approach?
2. If multiple trajectories toward T cell lineage commitment are possible, are certain subsets of trajectories favoured in different *in vivo* or *in vitro* contexts?
3. Do different sets of trajectories correspond to the differences in differentiation kinetics, gene expression dynamics, or genetic requirements that have been reported in various experimental contexts?

To answer these questions and gain further insights into differences between *in vivo* T cell development and *in vitro* T cell differentiation that manifest at the single-cell level, single-cell RNA sequencing (scRNA-seq) was performed. These experiments compared mouse T cell progenitor populations from three distinct contexts: primary adult thymocytes, primary E13.5 fetal thymocytes, and E13.5 fetal liver (FL) hematopoietic stem and progenitor cells (HSPCs) differentiated *in vitro* using the DL4+VCAM platform (Figure 16a). Immediately following dissection, primary adult and fetal thymocytes were magnetically depleted to reduce erythrocyte contamination, sorted for live CD45+ cells, and captured as single cells in droplets. In parallel, FL HSPCs were seeded on DL4+VCAM coated plates and cultured for 4 or 7 days prior to analysis, or immediately sorted and captured for library preparation. Pooling cells from multiple differentiation timepoints enabled sampling of cells from the entire T cell lineage progression, rather than just endpoint transcriptional states. Single cells from each context were captured and cDNA libraries were generated using the 10X Chromium microfluidic controller. The libraries were then sequenced together using an Illumina NextSeq. Two separate samples were prepared for each of the three experimental conditions, for a total of 6 single-cell cDNA library samples.

Although roughly equivalent cell concentrations (approximately 8.3×10^3 cells in 35 μL volume) of each sample were provided as input to the 10X Chromium controller, downstream library sequencing revealed that different numbers of cells were captured in each condition, with Fetal Thymocytes Rep. 1 (FTh1) and DL4+VCAM Differentiated FL001 (DVFL1) sample libraries containing roughly half the number of cells compared to the other conditions (Figure 16b, Table 6). We also noted that both Adult Thymocyte samples (ATh1 and ATh2) had much fewer median genes detected per cell and median UMI counts per cell when compared to other conditions (Figure 16cd, Table 6). Because of their low unique molecular identifier (UMI) and detected genes count, the adult thymocyte populations consistently clustered apart from the other four samples despite normalizing for numbers of UMIs. Thus, we decided to exclude ATh1 and ATh2 from our preliminary analysis pending deeper sequencing of these libraries.

After filtering for high-quality cells and normalizing expression values, single-cell transcriptional states for primary fetal thymocytes and DL4+VCAM differentiated FL HSPCs were visualized in reduced dimensional space using diffusion maps (Figure 17a). Diffusion maps employ a non-linear diffusion-based distance metric that emphasizes discovery of the underlying structure, or ‘manifold’, of the dataset (Haghverdi et al., 2015). Although the first diffusion component separates early and late stages of mouse T cell progenitors (marked by genes such as *Spil* and *Bcl11b*, respectively) (Figure 17b), the second diffusion component separates fetal thymocytes from DL4+VCAM differentiated FL HSPCs. Interestingly, the primary branch of the diffusion map projection that connects early-stage fetal thymocytes to later-stage fetal thymocytes appears completely separate from the primary branch connecting undifferentiated DL4+VCAM FL cells to late-stage DL4+VCAM differentiated T lineage cells (Figure 17a, arrows). This suggests that, in addition to occupying distinct regions of transcriptional space, fetal thymocytes and DL4+VCAM differentiated FL HSPCs may indeed follow different trajectories during T lineage specification. However, pseudotime trajectory inference must be performed on the scRNA-seq data to confirm whether the observed trajectories in fetal thymocytes or DL4+VCAM differentiated FL HSPCs align with the trajectories predicted by our BN model of the T cell development program.

Table 6: Single-cell RNA-sequencing sample metrics

Sample	Source	Replicate	Num. Cells	Median Genes per Cell	Median UMI Counts per Cell	Sequencing Saturation
Adult thymocytes	8 wks. CD1, 3 males	Rep 1	3,694	1,316	2,794	70%
Adult thymocytes	8 wks. CD1, 3 males	Rep 2	3,177	1,435	2,874	74%
Fetal thymocytes	E13.5 CD1, 2 litters	Rep 1	1,416	5,078	30,618	61%
Fetal thymocytes	E13.5 CD1, 2 litters	Rep 2	3,236	3,973	16,511	59%
DL4+VCAM differentiated FL HSPCs	E13.5 CD1, 1 litter, 3 wells, Sca1+ Kit+	FL001	1,821	4,011	16,397	64%
DL4+VCAM differentiated FL HSPCs	E13.5 CD1, 1 litter, 3 wells, Sca1+ Kit+	FL002	4,169	3,767	14,775	59%

FL = fetal liver, HSPCs = hematopoietic stem and progenitor cells

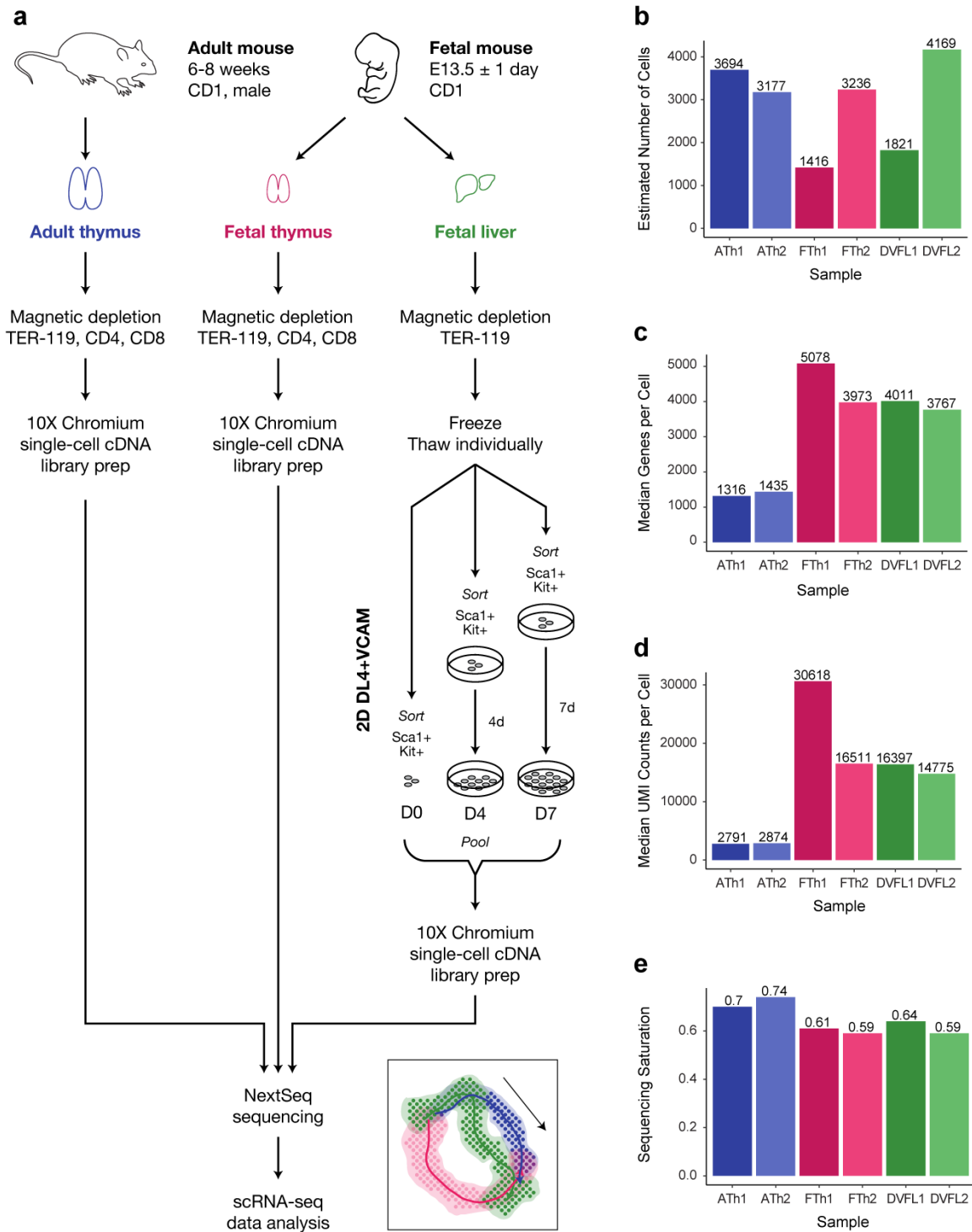


Figure 16: Summary of single-cell RNA-sequencing experiment design

(a) Primary thymocytes were isolated from adult and fetal mice and magnetically depleted to reduce the frequency of erythrocytes and mature T cell populations. Sca1+ Kit+ fetal liver HSPCs were also isolated and differentiated for 4 and 7 days on

DL4+VCAM. Single-cell cDNA libraries were subsequently prepared from each group of cells and sequenced to enable comparison of the transcriptional trajectories that are followed in each of the three experimental contexts.

- (b) Estimated number of cells per sample
- (c) Median genes detected per cell by sample
- (d) Median UMI counts per cell by sample
- (e) Sequencing saturation per sample

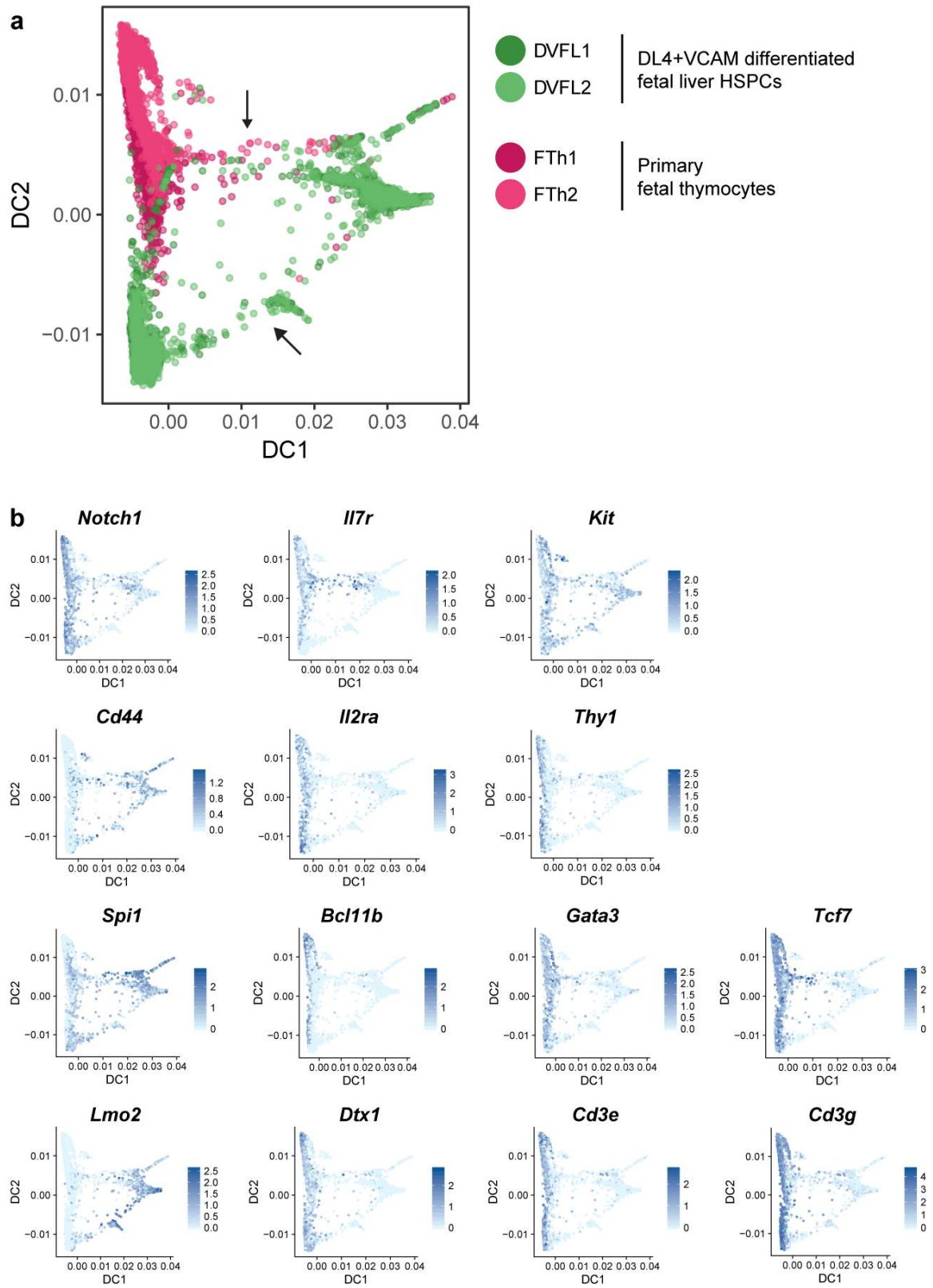


Figure 17: DL4+VCAM differentiated FL HSPCs and primary fetal thymocytes occupy distinct transcriptional spaces

(a) Diffusion map of scRNA-seq measured-transcriptional states in DL4+VCAM differentiated E13.5 FL HSPCs and E13.5 primary fetal thymocytes

(b) Gene expression overlays for select signaling receptors (*Notch1*, *Il7r*, *Kit*), cell surface markers (*Cd44* [CD44], *Il2ra* [CD25], *Thy1* [CD90]), T lineage-antagonizing TFs (*Spi1* [PU.1], *Lmo2*), T lineage-promoting TFs (*Bcl11b*, *Gata3*, *Tcf7* (TCF-1), *Dtx1*), and pre-TCR components (*Cd3e*, *Cd3g*). Early-stage T cell progenitors localize to higher values of DC1, and later-stage T cell progenitors localize to lower values of DC1. DC2 separates DL4+VCAM differentiated FL HSPCs from primary fetal thymocytes, although some overlap between these conditions is observed in reduced dimensional space.

5.4 Discussion

In this chapter, we present multiple complementary experimental platforms that are used to explore heterogeneity and transcriptional response patterns of T cell progenitors, and which can ultimately be harnessed to further refine the BN model presented in Chapter 3.

Overall, the experimental data presented in this chapter supports the hypothesis that the T cell development GRN facilitates heterogeneous and context-dependent transcriptional responses among T cell progenitors. Cellular heterogeneity is observed at the single-cell transcriptional level by qRT-PCR, even within surface marker-defined stages of early T cell development that were previously thought to be largely homogeneous. Context-dependent differences are observed in term of both differentiation kinetics and T cell GRN activation when compared between DL4+VCAM *in vitro* culture and DL4 alone, and are even more discernible between primary fetal thymocytes and DL4+VCAM differentiated FL HSPCs at the level of single-cell transcriptomes.

The proposed effect of VCAM-enhanced cell motility on Notch signaling activity and downstream T cell GRN activation raises an interesting future modeling direction. As currently implemented, the BN model we have developed cannot explicitly account for the effect of biophysical inputs to the GRN. However, it is well-evidenced through this study and others that biophysical cues can serve as an additional layer of regulatory control that converges onto the cellular GRN to influence cell fate decisions (Discher et al., 2009). The cell motility effects we observed on DL4+VCAM suggest that modeling the biophysical factors that influence T lineage differentiation in our engineered thymic-like niche may be as important as investigating direct biochemical inputs to the GRN, such as cytokines. To computationally support such investigations, one could consider extending the BN model presented here into a multi-scale agent-based model that explicitly considers spatial, physical, and intercellular layers of regulatory control in addition to the GRN itself. This future direction is explored in greater detail in Chapter 6.3.

The single-cell RNA sequencing (scRNA-seq) data presented in this chapter form a highly complementary dataset to the single-cell level predictions of our BN model, and therefore open up a number of new opportunities for further analysis. First, pseudotime trajectory inference algorithms can be applied to extract the transcriptional dynamics of individual cells as they differentiate toward the T cell lineage in either fetal thymopoiesis or on DL4+VCAM. Numerous pseudotime trajectory inference algorithms have been developed over recent years (Cannoodt et al., 2016), and are summarized in Chapter 6.1.1. However, none of the well-established trajectory inference algorithms support reconstruction of branched-yet-converging trajectories such as those predicted by the BN model. CellRouter is a recently published trajectory inference algorithm that theoretically supports inference of convergent trajectories (Lummertz et al., 2018); however, we are not able to implement this algorithm presently due to technical issues with the software. Once implemented, pseudotime trajectory inference will allow us to directly compare the trajectories taken by differentiating T cell progenitors *in vivo* and *in vitro* to the trajectories predicted by the BN model. This analysis can also be used to confirm a qualitative prediction of the BN model; specifically, that T cell progenitors rapidly downregulate T lineage antagonist transcription factors regardless of the trajectory taken, whereas T lineage promoting genes are expressed in a trajectory-dependent manner.

Moreover, the scRNA-seq data we have assembled provides an opportunity to further refine the topology and logic functions of our proposed BN model. The literature evidence and microarray data used to construct our present BN model of T cell development were derived from bulk population experiments. By comparison, single-cell transcriptomic data offers greater resolution to examine rare or heterogeneous features of the T cell development GRN, as well as more statistical power due to the increased number of data points (one full transcriptome per single cell) (Fiers et al., 2018). Standard co-expression metrics such as partial correlation could be employed to identify new genes that are significantly correlated to core genes already in the model, such as *Tcf7*, *Gata3*, *Bcl11b*, and *Spil*. This approach would allow us to mitigate potential bias toward well-studied transcription factor interactions and discover additional genes that influence the T lineage decision. Furthermore, single-cell GRN inference tools that directly infer BN models from scRNA-seq datasets have recently been developed,

including BoolTraineR (Lim et al., 2016) and Single Cell Network Synthesis (Moignard et al., 2015). These algorithms can be applied to our dataset to either retrain or fully recreate the BN model using single-cell data, thereby improving agreement with our experimental datasets. By virtue of inferring interactions from single-cell data, it would also be interesting to examine whether these algorithms reveal GRN interactions that are only present in rare T cell progenitor subsets or specific T cell differentiation contexts. New edges or regulatory logic in the revised T cell development BN model could subsequently be validated by ChIP-seq or genetic knockouts. Single-cell enabled refinement of our BN model is discussed in further detail in Chapter 6.1.2.

Overall, the experimental platforms and data presented in this section shed new light on heterogeneity and context-dependent transcriptional patterns within the mouse T cell development program. Looking forward, these platforms and datasets also provide an excellent foundation for making improvements to our BN model (using scRNA-seq) and testing any new predictions that arise from the refined model using our fully-defined engineered thymic niche (DL4+VCAM).

6 Future Work

6.1 Single-cell transcriptomics analysis of T cell progenitor differentiation

6.1.1 Trajectory inference from single-cell transcriptomics data

The BN model of mouse T cell development developed through this project predicts that multiple distinct transcriptional trajectories leading to T lineage commitment are available to hematopoietic progenitors. We hypothesize that certain classes of trajectories are more frequently chosen in different T cell differentiation contexts as a result of external constraints on the GRN, including epigenetic state and environmental signals. Context-dependent enrichment for different transcriptional trajectories could provide a mechanistic explanation for why T lineage differentiation proceeds with such widely varying kinetics and genetic requirements when comparing fetal thymopoiesis, adult thymopoiesis, and various *in vitro* differentiation protocols.

Testing this hypothesis experimentally necessitates the ability to follow transcriptional changes within single cells as they progress down the T lineage. However, timecourse studies of T cell development have previously been limited by multiple factors. Bulk population experiments measure only the average response of cells, even though individual cells may either be following different transcriptional trajectories or progressing along the same trajectory asynchronously. In the T cell field, different stages of T cell progenitors can be isolated by cell surface markers (such as CD25, CD44, and c-KIT) and assessed independently to more finely resolve transcriptional dynamics; however, both we and other groups have demonstrated that these surface marker-defined stages still comprise transcriptionally and functionally heterogeneous subpopulations (Porritt et al., 2004). Furthermore, at least in some developmental contexts, it appears that hematopoietic progenitors can reach the latter stages of T cell development without ever expressing surface markers profiles associated with intermediate states (like DN3). Therefore, bulk analysis of sorted T cell progenitor populations is likely insufficient to experimentally resolve differences in transcriptional trajectories between various T cell differentiation contexts.

Single-cell experimental techniques such as single-cell RNA sequencing (scRNA-seq) have matured over recent years and represent a new opportunity to probe transcriptional dynamics and heterogeneity during T lineage differentiation with single-cell resolution (Fiers et al., 2018). Furthermore, multiple statistical methods have recently been developed to order scRNA-seq profiles in “pseudotime” and place cells along one or more trajectories that approximate the underlying differentiation process (Cannoodt et al., 2016). In this context, pseudotime is an inferred metric that approximates the extent to which a cell has proceeded through a dynamic biological process. Given the potential benefits of these algorithms, we pursued scRNA-seq with the intent to perform computational inference of the pseudotime trajectories followed by differentiating T cell progenitors in various contexts.

Although there are multiple methods available for pseudotime trajectory inference from single-cell transcriptomics data (summarized in Table 7), none of the well-established methods currently support reconstructing trajectories that feature both diverging and converging branches. This is an important current limitation for our application, since the BN model predicts that a single progenitor cell state can follow multiple possible trajectory branches (diverging) before eventually reaching a steady state shared by all trajectories (converging). New trajectory inference methods such as CellRouter (Lummertz et al., 2018) claim to support inference of convergent trajectories; however, we encountered issues getting CellRouter to run without errors in our existing computational environment. Additional troubleshooting and collaboration will be required to pursue this branch of single-cell analysis.

Table 7: Methods for trajectory inference from single-cell transcriptomics data

Method	Trajectory structure	Dimensionality reduction	Trajectory modeling	Other features	Language	Reference
Wanderlust*	Linear	N/A	kNN Euclidean or cosine distance Shortest path	Ensemble method, bootstrapping	MATLAB	(Bendall et al., 2014)
Wishbone	Single bifurcation	PCA, diffusion maps	kNN Similarity weighting Shortest path	Bootstrapping	Python	(Setty et al., 2016)
SLICER	Branching	Locally linear embedding	kNN Extreme cell detection Shortest path	Geodesic entropy used to find branch points	R	(Welch et al., 2016)
Monocle	Branching	ICA (1.0) PCA, t-SNE (2.0)	Reverse graph embedding MST Longest connected path	Monocle 2 uses DDRTree for non-reconstruction with less sensitivity to outliers and cell quality	R	(Trapnell et al., 2014)
Waterfall	Linear	PCA k-means cell clustering	MST (between cluster centers)	Cluster with lowest PC1 value selected as start node	R	(Shin et al., 2015)
SCUBA	Branching	PCA k-means cell clustering	Map cells to clusters in previous time point Bifurcation detection by k-means + gap statistic	Infers from time series or principal curve-based pseudotime ordering	MATLAB	(Marco et al., 2014)
SCOUP**	Branching	PCA	MST Shortest path	Expectation-maximization to refine cell ordering with respect to expression	C++	(Arsenio et al., 2014)
Mpath	Branching	Hierarchical clustering of cells	Waypoint finding MST	Nodes of graph represent “waypoints”, and cells are assigned to edges between their closest waypoints	R	(Chen et al., 2016)

Method	Trajectory structure	Dimensionality reduction	Trajectory modeling	Other features	Language	Reference
TSCAN	Linear	PCA Cell clustering by Gaussian mixture model	MST (between cluster centers) Longest connected path	Number of cell clusters automatically determined by Bayesian information criteria	R	(Ji and Ji, 2016)
CellRouter	Branching / Converging	t-SNE, PCA, diffusion maps, etc.	kNN Edge weighting by Jaccard similarity Community detection by Louvain method Minimum cost flow network	User chooses their own reduced dimension space Trajectories are identified between any two user-specified clusters	R ***	(Lummertz et al., 2018)

PCA = principal component analysis, ICA = independent component analysis, MST = minimal spanning tree, kNN = k-nearest neighbours

- * Wanderlust was originally developed for mass cytometry data (~10 to 50 dimensions), thus dimensionality reduction methods were unnecessary in this context
- ** Due to the computational complexity of its expectation-maximization algorithm, SCOUP is limited to relatively few cells and genes (i.e. single-cell qRT-PCR data)
- *** CellRouter includes an embedded Java library in addition to R pipeline. Limited OS and Java version support

6.1.2 Boolean network refinement using single-cell transcriptomics

The mouse T cell development program serves as an excellent case study in the strengths and limitations of literature-based computational models of gene regulatory networks (GRNs). GRNs such as those involved in mouse T cell development have well-defined biophysical interpretations: transcription factors bind to the cis-regulatory elements of their target gene (often in a combinatorial, cooperative, or competitive manner) to produce either an active transcriptional complex or a repressive one. Therefore, an idealized computational model of a GRN would be fully evidenced by physical characterization of binding events at the cis-regulatory element (using chromatin immunoprecipitation, for example), protein-protein interactions between TF complex members and RNA polymerase, knockout experiments coupled with transcriptional profiling that confirm the effect of losing any of these regulatory inputs, and epigenetic assays to assess the open chromatin status of the gene and its enhancer elements. Obtaining this level of detail for each element of a GRN remains intractable in all but the most simple of model systems (such as sea urchin development) (Peter and Davidson, 2010). Furthermore, most of the associated experimental methods require large numbers of cells to achieve sufficient signal-to-noise ratios. Since GRNs inferred from population-level data rely on average measurements across thousands to millions of single cells, they are prone to underperform when applied to heterogeneous cellular systems. Given that transcriptional heterogeneity is critically important for many stem cell systems, computational GRN inference methods that account for this heterogeneity are needed.

Single-cell transcriptomics presents a promising alternative for GRN inference at high resolution and in a manner that captures transcriptional heterogeneity. Single-cell transcriptomics enables analysis of rare cell populations and can be useful for dissecting heterogeneity within seemingly homogenous populations. Additionally, single-cell transcriptomics studies yield large datasets, with up to tens of thousands of single cells captured in a single run and transcript counts for tens of thousands of genes. As with bulk population data, genes that are important to the underlying cellular process of interest can also be inferred from single-cell transcriptomics data through co-expression statistics. However, the statistical power of methods such as correlation, regression, covariance,

and mutual information are increased when applied to single-cell datasets since the number of samples is greater and since heterogeneous features are preserved in the dataset rather than averaged out.

Furthermore, whereas most network inference algorithms rely on static (single timepoint) measurements of gene interactions, single-cell expression profiles can be ordered in pseudotime (most often using neighbour similarity metrics, like k -nearest neighbour networks). Because of these advantages, the regulatory genomics field is shifting toward single-cell methods (Fiers et al., 2018).

An important caveat to single-cell transcriptomics analyses is that they are noisier than conventional bulk transcriptomic analyses. Some noise is expected as a result of true biological variation within cells due to stochastic gene expression or transcriptional bursting. However, this is often confounded with technical noise arises to low amount of input mRNA. Technical noise is further compounded through PCR amplification bias and dropout effects (false negatives where a gene is called as not-expressed due to poor mRNA capture efficiency). The effects of noise and dropouts must be considered when constructing dynamic GRN models from single-cell data (Fiers et al., 2018).

Boolean network (BN) models are particularly robust to dropout effects since gene expression values are binarized. Therefore, it is perhaps unsurprising that multiple tools for inferring BN models from single-cell transcriptomics data have been developed in recent years. Two of these tools are Single Cell Network Synthesis (SCNS) and BoolTraineR (BTR), and both are potentially suitable options for harnessing the scRNA-seq data generated in this project to improve the BN model of T cell development reported here (Lim et al., 2016; Moignard et al., 2015).

SCNS is a satisfiability modulo theory (SMT)-based tool for inferring BN models *de novo* from single-cell qRT-PCR or single-cell RNA-seq data (Moignard et al., 2015). It approaches the experimentally-measured single-cell transcriptional space as if it were the output of an asynchronous simulation of a BN model—that is, a state transition graph. Because raw single-cell transcriptomics data consists of nodes representing single-cell profiles, but not edges that connect temporally-related single-cell profiles to each other,

these edges must be constructed computationally. SCNS accomplishes this by drawing edges between single-cell states that differ in the expression of only one gene. An important caveat of this approach is that the experimentally-derived graph must be fully connected. This can be difficult to achieve for networks including many genes, since an exponentially greater number of cell states are needed to yield a fully connected graph over more variables. Alternatively, if a k -Nearest Neighbour approach were to be used, one would need to balance increased k values (leading to more densely connected graphs) against increased false positive rates, since a transition between two single-cell states that differ by multiple genes is likely to correspond to a single gene regulatory event. Once a state transition graph has been created from the experimental dataset, SCNS searches for a set of Boolean logic functions that are able to satisfy all transitions within the graph. The Boolean logic functions identified through this process constitute the inferred BN model and, given sufficient data to yield a fully connected graph, are guaranteed to satisfy the experimental observations because of the SMT approach.

BTR differs from SCNS in that it can be used to improve upon an existing BN model using newly-acquired single-cell transcriptomics data rather than constructing a new model *ab initio* (Lim et al., 2016). BTR infers both network topology and Boolean update functions without prior information on cell state trajectories. It accomplishes this through a swarming hill climbing optimization process that iteratively reduces the distance between the model state space (obtained by asynchronous simulation of BN) and the data state space (obtained by single-cell transcriptomics). The output of this iterative optimization is an asynchronous BN model that best represents the input single-cell expression dataset. However, even with sufficient data, BTR does not guarantee perfect agreement between the trained BN model and the single-cell expression dataset.

Importantly, both SCNS and BTR emphasize cellular transitions through intermediate expression states, which is particularly relevant to studies of developmental processes such as T cell development. This constitutes a significant advantage over existing GRN inference tools that assume that the experimentally-measured data must represent stable attractors of the system, akin to a cell population after extended culture in maintained conditions (Yordanov et al., 2016).

Both of these single-cell data-driven approaches are potentially complementary to the network inference method employed in this project. In our approach, we produced a GRN based on literature evidence and bulk population measurements, then asynchronously simulated this network to predict the set of transcriptional trajectories that are possible given the topology and logic of that GRN. This approach is well-suited for cell fate decision systems in which the key players and their input cis-regulatory logic have already been well-characterized through previous reports and are supported by multiple modalities of experimental evidence (i.e. both chromatin immunoprecipitation evidence of binding and genetic knockout evidence of a cause-effect relationship). However, this approach risks biasing toward elements of the underlying GRN that have received greater focus from the research community and thus have more literature evidence. Relatedly, the approach risks underrepresenting or misrepresenting those genes whose roles have not yet been well characterized. Furthermore, GRN models produced through a literature-focused approach may produce artefactual cell trajectories that, while theoretically permissible given the logic and topology of the network, are either impractical or not observed biologically due to constraining factors such as epigenetic state.

Conversely, the approach put forth by SCNS and BTR—and enabled by the availability of single-cell transcriptomics data—starts with a set of observed trajectories and reverse engineers a GRN (in this case, a BN) model that is capable of explaining those trajectories. GRNs inferred via this second approach are thus fully descriptive of the experimental observations when supplied with adequate data. However, it remains unclear whether such models are truly reflective of the underlying biophysical mechanisms and cis-regulatory logic of the actual GRN at work in the cells since no physical binding data or perturbation experiments are considered during the inference process. Furthermore, given that the GRN models are identified or trained to maximally satisfy only the given set of single-cell experimental constraints, it remains unclear whether the models are vulnerable to overfitting. For example, could such a model accurately predict how T cell progenitors would respond transcriptionally to a new environmental signaling condition or genetic perturbation that was not included in the original training (constraint) set?

Nevertheless, we anticipate that using the single-cell RNA-seq data gathered through this project to refine the original BN topology and logic functions would ultimately improve its accuracy when compared to experimental data and potentially capture new aspects of the T cell development program which have not yet been well-studied in our field. If executed carefully, this combined approach would potentially combine the benefits of both literature-focused and single-cell data-driven inference methods.

6.2 Transcriptional memory in T cell progenitor-derived induced pluripotent stem cells

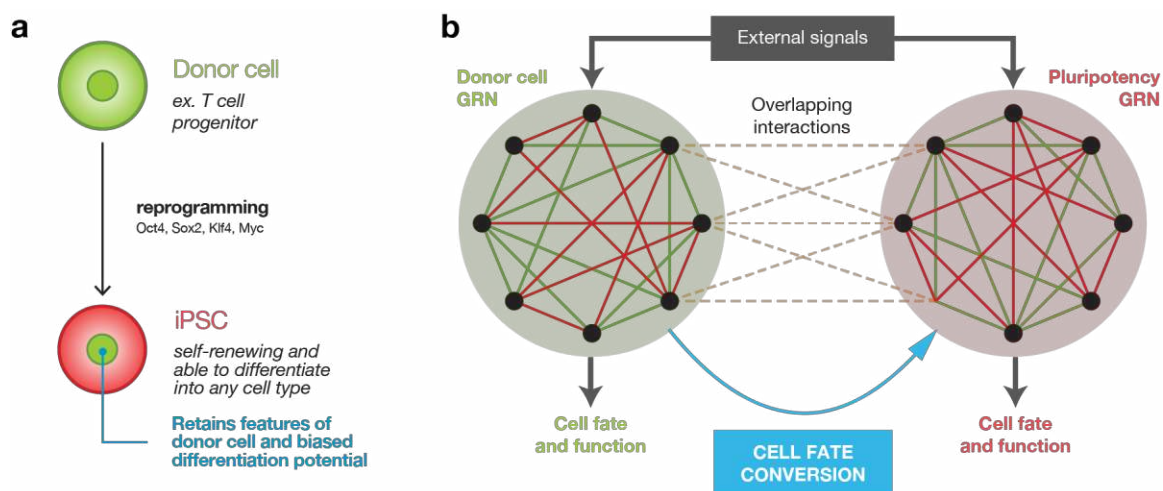


Figure 18: Donor cell memory in induced pluripotent stem cells and proposed role of GRN feedback

In addition to differentiation of T cells, there is also interest in reprogramming T cells into induced pluripotent stem cells (iPSCs). For example, by reprogramming antigen-specific T cells from human immunodeficiency virus (HIV)-positive patients into iPSCs and re-differentiating the cells into CD8⁺ cytotoxic T cells, “rejuvenated” T cells with elongated telomeres but preserved antigen specificity can be created (Nishimura et al., 2013). In another case, significant inhibition of tumor growth was achieved in a xenograft model using T cells that were captured from peripheral blood, reprogrammed into iPSCs, genetically engineered to express a CD19-specific CAR, and re-differentiated into T cells (Themeli et al., 2013).

In our lab, we have observed that iPSC lines derived from different stages of mouse T cell progenitors exhibit biased differentiation potential toward the mesoderm lineage (Shukla *et al*, unpublished data). This is consistent with previous reports of “donor cell memory” – a property of iPSCs to retain genetic and epigenetic features of their cell type of origin and a consequent bias in differentiation potential over early passages (Kim et al., 2010; Polo et al., 2010). Because reprogramming demands that cells transition from their original GRN state to a pluripotent GRN state, one might expect that feedback

between these GRNs could modulate gene expression levels and potentially establish these memory effects. Yet although donor cell memory has been observed in many iPSC systems, no study has investigated the role of somatic GRNs in donor cell memory. The T cell progenitor-derived iPSC system developed in our lab presents a unique opportunity to investigate this through the lens of the well-studied T cell development GRN. Such investigation would enable identification of a potential role for developmental GRNs in induced pluripotency and guide efforts to stabilize donor cell memory states that enhance differentiation potential toward target lineages, such as T cells.

6.2.1 T cell progenitor-derived iPSCs exhibit molecular and functional pluripotency

T cell progenitors were previously isolated in our lab from secondary chimeric adult 1B mice containing a doxycycline (dox)-inducible reverse tetracycline transactivator (rtTA) (*Rosa26 rtTA-IRES-GFP* knock-in) and reprogrammed via addition of 1 mg/mL doxycycline (dox) to culture media (Fluri et al., 2012). Our lab has previously demonstrated that these T cell progenitor-derived iPSC lines exhibit increased differentiation potential and kinetics toward the mesoderm lineage compared to genetically-matched embryonic stem cells (ESCs). We hypothesized that this differentiation potential was accompanied, and perhaps caused, by partially-retained expression of T lineage genes.

To ensure that any gene expression anomalies characterized in the T cell progenitor-derived iPSCs were not an artefact of incomplete reprogramming (Figure 19a), we assessed all iPSC clones for pluripotency at the molecular and functional level. All T cell progenitor-derived iPSC clonal lines expressed the core pluripotency genes *Oct4*, *Sox2*, and *Nanog* at levels comparable to ESCs (Figure 19b) and presented the canonical pluripotency surface markers *OCT-4*, *NANOG*, *SSEA-1*, and *TBX-3* (Figure 19c). Furthermore, DN1-derived iPSCs were able to successfully form chimeric mice and contribute to all three germ layers (Figure 19d). Together, these results suggest that all

clonal lines of T cell progenitor-derived iPSCs were successfully reprogrammed to pluripotency, as defined molecularly and functionally.

6.2.2 T cell development genes are expressed atypically in T cell progenitor-derived iPSCs

Although all T cell progenitor-derived iPSCs achieved pluripotency, qRT-PCR analysis shows that genes associated with T cell development were frequently expressed in medium passage (p7) iPSCs at levels atypical of a “memoryless” pluripotent stem cells, such as ESCs. Some genes (*Bcl2*, *Myb*, *Gata3*, etc.) were consistently expressed at higher-than-ESC levels, whereas others (*Lat*, *Bcl11b*, *Cd3e*, etc.) were consistently expressed at lower-than ESC levels (Figure 20a). Furthermore, these genes exhibited different memory patterns depending on their donor cell stage, such that at least DN1-iPSCs and DN3-iPSCs were fully distinguishable by clustering.

Atypically expressed genes corresponded to a variety of different functions in T cells, including signaling receptors, pre-T cell receptor (TCR) components, and promoters of the T cell and alternative lineages (Figure 20b). These differences were particularly pronounced for *Il7ra* and *Rag1*, which both serve critical roles in developing T cells as a receptor for IL-7 signaling and an activator of T cell receptor chain recombination, respectively. Both *Il7ra* and *Rag1* were expressed up to 100-fold higher in some iPSC clones than in ESCs. Interestingly, genes that share a common function or timecourse progression in primary T cell progenitors do not all exhibit the same donor cell memory trends.

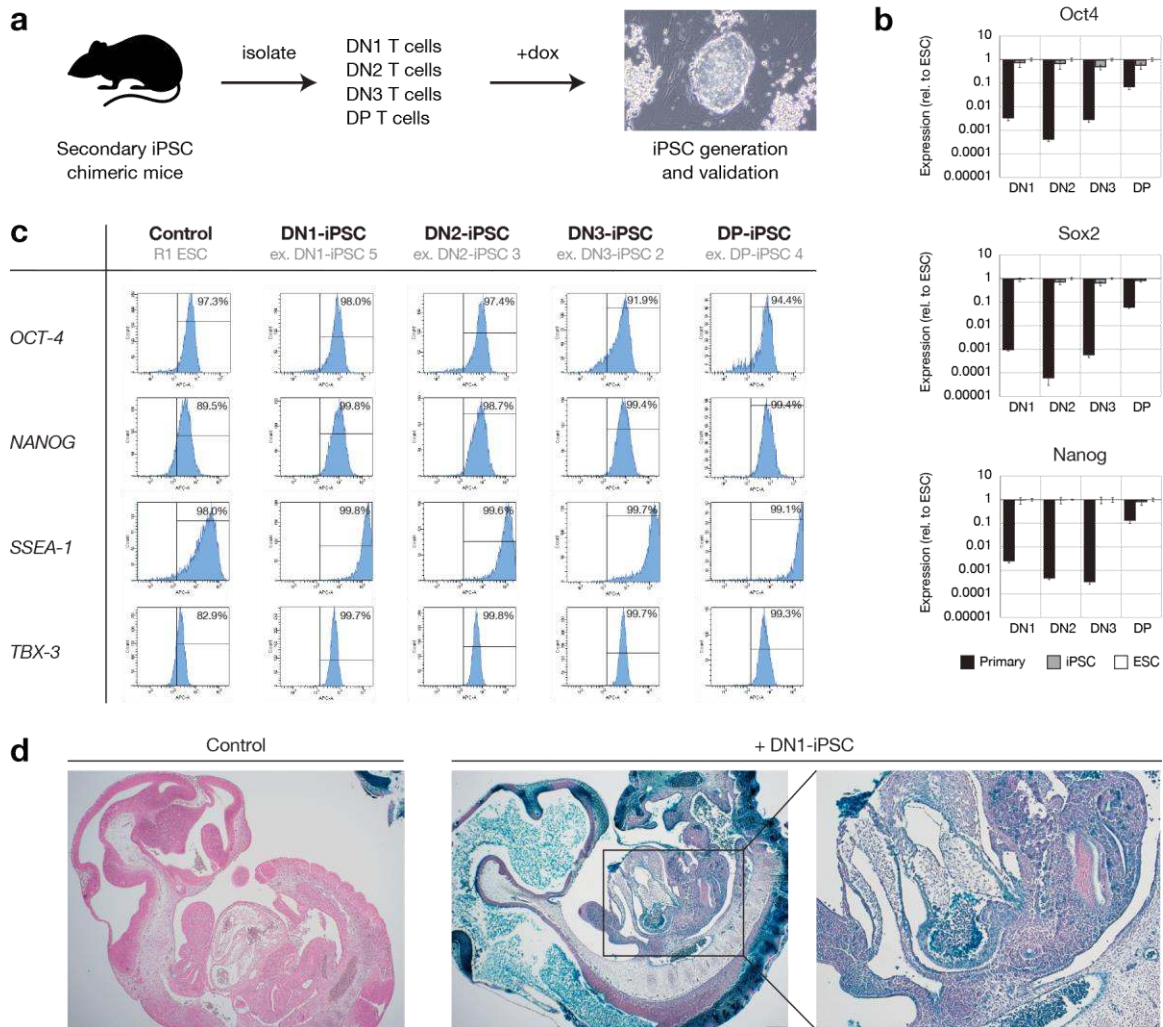


Figure 19: T cell progenitor-derived iPSCs exhibit molecular and functional pluripotency

- (a) Overview of secondary dox-inducible mouse iPSC system and T cell progenitor isolation.
- (b) qRT-PCR-assessed expression levels of core pluripotency genes in all iPSC groups (mean \pm SD of 3 independent clonal lines) are similar to levels in R1 ESC controls.
- (c) Representative flow cytometry plots demonstrate the pluripotency protein markers *OCT-4*, *NANOG*, *SSEA-1*, and *TBX-3* are present in each iPSC line at levels similar to R1 ESC controls.
- (d) E10.5 chimeric mouse embryos produced from aggregated clonal DN1-iPSCs, showing iPSC contribution to endoderm, mesoderm, and ectoderm germ layers (middle,

focus on midsection at right) versus control embryo (left). All tissue stained with hematoxylin and eosin, iPSC contribution visualized by LacZ-mediated staining. *Chimera experiment performed by P. Tonge.*

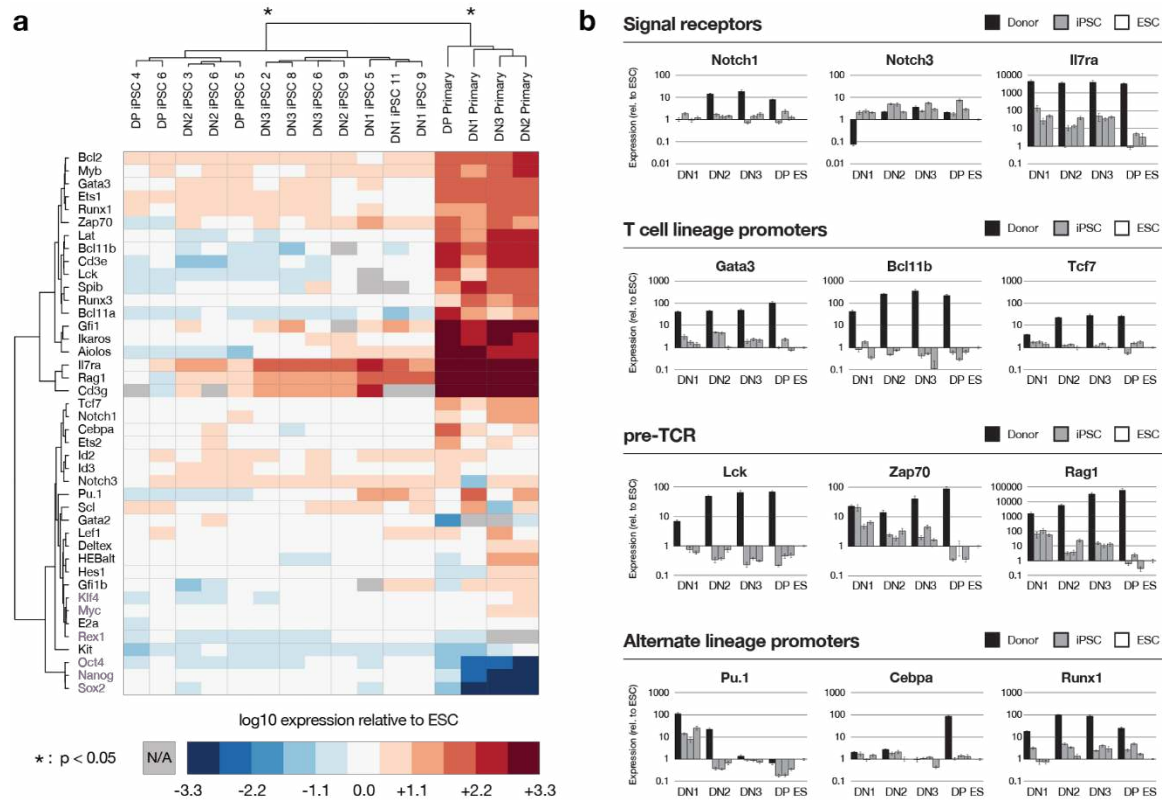


Figure 20: iPSCs derived from different T cell progenitor stages express T cell development genes at atypical levels and are transcriptionally distinguishable

(a) Heatmap of qRT-PCR-assessed gene expression for each T cell progenitor-derived iPSC clonal line and their donor cell types, relative to ESC control. DN1-iPSC and DN3-iPSC lines each clustered fully together (Complete clustering and Euclidean distance algorithms). Purple indicates pluripotency genes.

(b) qRT-PCR-assessed expression levels are shown relative to ESC control for different functional groups of T cell development genes, including signal receptors, T cell lineage promoters, pre-TCR, and alternate lineage promoters. Transcriptional patterns within each group of genes are dissimilar, suggesting T cell development genes with similar functions do not exhibit similar donor cell memory patterns.

6.2.3 Toward integration of BN models of T cell development and pluripotent fate transitions

As discussed in Chapter 4.3, our lab has previously developed and validated a BN model of the endogenous mouse pluripotency network that predicts how cells transition between distinct pluripotent states (Yachie-Kinoshita et al., 2018). By integrating our BN model of the T cell development GRN with this BN of pluripotent stem cell fate transitions, we may potentially capture the reprogramming process from a T cell progenitor to an iPSC. Using Metacore (a curated database of transcription factor binding, receptor-ligand interactions, kinase activity, and metabolic processes; <https://portal.genego.com/>), we queried all known interactions between genes included in the T cell development BN model and those included in the pluripotency BN model. 80 directed interactions from PSC-related gene sources to T cell-related targets and 73 directed interactions from T cell-related gene sources to PSC-related targets were reported (Figure 21). These 153 interactions could serve as the initial basis for new Boolean update functions that link the two BN models. A satisfiability modulo theory (SMT) tool such as RE:IN can be used to generate candidate BNs, with topologies of the T cell development and pluripotency BN models set as “definite” and our aforementioned qRT-PCR results as constraints. Once the integrated BN model is constructed, reprogramming could be simulated by forcing expression of the dox-inducible factors in our secondary reprogramming system (*Oct4*, *Sox2*, *Klf4*, *c-Myc*) to “ON”. By simulating all possible combinations of the input signals, the model could predict methods to control donor cell memory by identifying combinations that result in strongly connected components (SCCs) with previously-unreported expression patterns or higher sustainability scores (Yachie-Kinoshita et al., 2018).

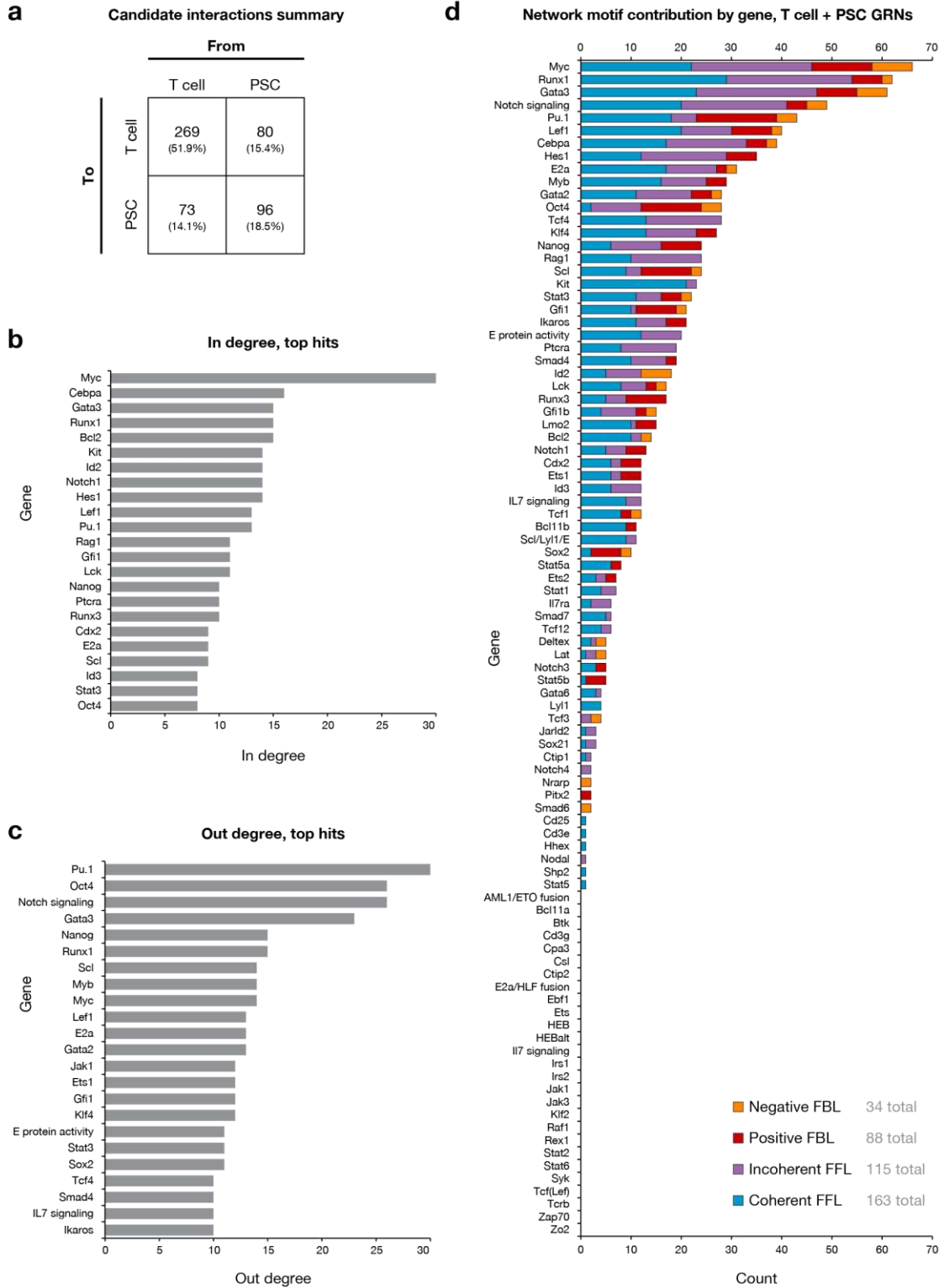


Figure 21: Summary of reported interactions between T cell development and pluripotency GRNs

(a) Reported regulatory interactions within and between each GRN. Top interaction hits in the consolidated dataset can be identified by (b) in degree, (c) out-degree, and (d) network motif contribution and serve as promising starting points for an integrated Boolean model of the T cell-to-iPSC reprogramming process.

6.3 Toward multi-scale models of T cell development

Although the BN model of mouse T cell development presented here constitutes an important advancement toward dynamic and computable models of T lineage specification, its scope comprises the T cell fate decision at only one scale—gene regulatory networks (GRNs) and associated transcriptional events. However, it is well-understood that biological decisions and fate specification processes span multiple scales and types of factors (Discher et al., 2009; Yu and Bagheri, 2016). At the intracellular level, epigenetic regulation of chromatin accessibility, metabolic state, and noise arising through signal transduction exert critical effects on many cell fate decisions.

Intracellularly, cells receive biochemical signals from their environment through receptor-ligand interactions and secrete their own ligands that exert feedback effects on neighbouring cells and other cell types (Qiao et al., 2014). Mechanically, physical forces such as shear stress and substrate stiffness are also known to affect cell fate choice (Discher et al., 2009). Spatially, cells migrate between different niches under the control of various developmental signals and chemokines; for example, hematopoietic stem cells home from the aorta-gonad-mesonephros (AGM) to the fetal liver and eventually to the bone marrow during embryonic development, and lymphoid progenitors continuously migrate from the bone marrow to the thymus to initiate T cell development throughout an organism's life. Spatial positioning and self-organization are also hallmark features of many developing tissues, both during *in vivo* organismal development (such as gastrulation) and well as in various *in vitro* organoid and micropattern confinement platforms (Rahman et al., 2017; Simunovic and Brivanlou, 2017; Tewary et al., 2017). In the context of development, the influence of each of these factors ultimately leads to the emergence of distinct and specialized cell types, tissues, and organs.

Eventually, all the aforementioned effectors of cell fate must converge upon a GRN that resolves the inputs into transcriptional changes, lineage choices, and phenotypic behaviours at the single-cell level. It is for this reason that we focused our initial efforts toward modeling T lineage fate choice on GRNs. However, one could anticipate that a comprehensive model spanning multiple scales of biological regulation would enable

more accurate recapitulation of observed biological behaviour and facilitate testing of more complex biological hypotheses *in silico*.

Agent-based modeling provides one potential framework for achieving comprehensive multi-scale models of cell fate decisions (Kaul and Ventikos, 2015). An agent-based model (ABM) consists of many individual Turing-complete finite-state machines (termed “agents”) which interact with each other based on pre-defined rule sets to simulate emergent properties of a system. Because each agent acts independently based on its own rule set and its perceived environment, ABMs are typically able to capture the behaviour of heterogeneous systems better than continuum mathematical models can. Furthermore, the internal computations performed by each agent are not inherently restricted to any particular class of algorithm, and thus ABMs are amenable to integration with other low-level modeling frameworks, including Boolean networks (BNs). ABMs have been successfully applied to model a variety of cellular systems (Kaul and Ventikos, 2015), and there is specific interest in using ABMs to model self-organization and specialization within stem cell systems.

As an example of one potential future direction within the T cell development field, we can consider an agent-based model of cells within the thymic niche. Agents within the model would represent single cells and could be assigned to different cell type classes, such as hematopoietic progenitors or thymic epithelial cells. Agents would occupy a 3-dimensional physical space that is initialized to mimic the cortical-medullary architecture of the thymus. Agents would also be free to move in 3D space either randomly or in response to chemotactic cues. Each agent could perceive physical and biochemical cues from its environment as a function of its position in space and its neighbours. Agents could also present membrane-bound ligands (such as DL4) or receptors (such as NOTCH1 or CCR7) and secrete cytokines (such as IL-7) or chemokines (such as CCL21) to the extracellular space. The interactions of cellular agents with the extracellular environment would constitute a reaction-diffusion system that can be modeled mathematically using differential equations (Tewary et al., 2017). Finally, at the intracellular level, the BN model presented here could be embedded within each agent to process the environmental signals seen by the agent and decide its transcriptional

response through logical simulation. The effects of this transcriptional response would then be used to adjust the agent's properties, such as its levels of surface-bound molecules and its cytokine secretion profile.

An agent-based model as proposed would enable computational study of interactions between thymocytes and the thymic epithelium, which has previously been difficult to accomplish experimentally due to tissue opacity and imaging limitations. Furthermore, the agent-based model would allow for computational study of anatomical diseases such as DiGeorge syndrome (in which the thymus is smaller than normal) and the effect of radiation therapies employed in cancer treatment and HSC transplant scenarios (which are known to disrupt the architecture and cellular composition of the thymus) (Awong et al., 2013). In the context of normal development, the model could be used to investigate the dwell time of T cells in different regions of the thymus and explore how thymic DN1 cells may function as a thymocyte stem cell-like population. Finally, extending the agent-based model to synthetic *in vitro* T cell development niches such as DL4+VCAM would enable us to explore the mechanism by which increased cell motility in the presence of VCAM gives rise to greater Notch signaling activation and greater T cell yields (Shukla et al., 2017). In this case, we could also proactively screen additional matrix components, cytokines, or small molecules *in silico* for their effect on T cell yields as a means to guide further improvements to the DL4+VCAM T cell differentiation platform. Thus, a multi-scale ABM approach would enable us to investigate new classes of questions regarding the T cell development program and help increase our understanding of the complex regulatory controls that underlie cell fate decision making.

7 Conclusions

7.1 Thesis novelty and impact

In this thesis, I demonstrate the value of a Boolean network (BN) approach to exploring the mouse T cell development program. The dynamic BN model of the T cell development gene regulatory network (GRN) that was constructed represents a significant advancement beyond previous static network topologies and smaller continuous models of sub-motifs of the T cell development program. Simulations of the BN model accurately recapitulate the transcriptional profiles of known T cell progenitor types as well as the response of T cell progenitors to various combinations of environmental signals and genetic perturbations. The BN model also makes the testable prediction that there are multiple possible transcriptional trajectories for T lineage specification, which suggests a potential explanation for the wide variance in differentiation efficiency, kinetics, and genetic requirements that have been observed in different T cell progenitor differentiation contexts.

With respect to experimental novelties, we developed a serum- and stromal cell-free platform for T lineage differentiation that facilitates the study of additional factors that influence the T cell fate within a fully-defined thymic-like niche. We also demonstrated transcriptional heterogeneity within seemingly homogeneous surface marker-defined stages of T cell development using single-cell qRT-PCR. To our knowledge, we generated the first single-cell RNA sequencing dataset that enables comparison of the transcriptional patterns and trajectories taken by differentiating T cell progenitors during *in vivo* thymopoiesis and *in vitro* differentiation. These platforms and datasets form a strong foundation for refining future computational models of the T cell development program.

Finally, through the development of 3 software ‘gadgets’ for the *Garuda* systems biology software platform, our BN simulation and analysis framework are now easily accessible and implementable by the broad biology research community. This effort will enable the integration of BN approaches into larger data analysis pipelines and ultimately extend the impact of our simulation framework beyond the T cell development community.

References

- Albert, R., and Thakar, J. (2014). Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *Wiley Interdiscip. Rev. Syst. Biol. Med.* *6*, 353–369.
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* *8*, 450–461.
- Anderson, M.K. (2006). At the crossroads: diverse roles of early thymocyte transcriptional regulators. *Immunol. Rev.* *209*, 191–211.
- Anderson, M.K., Hernandez-Hoyos, G., Dionne, C.J., Arias, A.M., Chen, D., and Rothenberg, E. V. (2002). Definition of regulatory network elements for T cell development by perturbation analysis with PU.1 and GATA-3. *Dev. Biol.* *246*, 103–121.
- Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* *32*, 1241–1243.
- Arsenio, J., Kakaradov, B., Metz, P.J., Kim, S.H., Yeo, G.W., and Chang, J.T. (2014). Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. *Nat. Immunol.* *15*, 365–372.
- Awong, G., Singh, J., Mohtashami, M., Malm, M., La Motte-Mohs, R.N., Benveniste, P.M., Serra, P., Herer, E., van den Brink, M.R., and Zúñiga-Pflücker, J.C. (2013). Human proT-cells generated in vitro facilitate hematopoietic stem cell-derived T-lymphopoiesis in vivo and restore thymic architecture. *Blood* *122*, 4210–4219.
- Back, J., Dierich, A., Bronn, C., Kastner, P., and Chan, S. (2004). PU.1 determines the self-renewal capacity of erythroid progenitor cells. *Blood* *103*, 3615–3623.
- Balciunaite, G., Ceredig, R., Fehling, H.J., Zúñiga-Pflücker, J.C., and Rolink, A.G.

(2005). The role of Notch and IL-7 signaling in early thymocyte proliferation and differentiation. *Eur. J. Immunol.* *35*, 1292–1300.

Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* *157*, 714–725.

Bhandoola, A., and Sambandam, A. (2006). From stem cell to T cell: one route or many? *Nat. Rev. Immunol.* *6*, 117–126.

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

Brauer, P.M., Singh, J., Xhiku, S., and Zuniga-Pflucker, J.C. (2016). T cell genesis: in vitro veritas est? *Trends Immunol.* *37*, 889–901.

Briggs, A.J.A., Li, V.C., Lee, S., Woolf, C.J., Klein, A.M., and Marc, W. (2017). Mouse embryonic stem cells can differentiate via multiple paths to the same state. *Elife* 1–31.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* *36*, 411–420.

Cahan, P., Li, H., Morris, S.A., Lummertz, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* *158*, 903–915.

Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* *46*, 2496–2506.

Chen, J., Schlitzer, A., Chakarov, S., Ginhoux, F., and Poidinger, M. (2016). Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat. Commun.* *7*, 11988.

- Collombet, S., van Oevelen, C., Sardina Ortega, J.L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., Graf, T., and Thieffry, D. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc. Natl. Acad. Sci.* 201610622.
- Crompton, T., Outram, S. V., Buckland, J., and Owen, M.J. (1998). Distinct roles of the interleukin-7 receptor α chain in fetal and adult thymocyte development revealed by analysis of interleukin-7 receptor α -deficient mice. *Eur. J. Immunol.* 28, 1859–1866.
- Dai, H., Wang, Y., Lu, X., and Han, W. (2016). Chimeric antigen receptors modified T-cells for cancer therapy. *J. Natl. Cancer Inst.* 108.
- Dakic, A., Metcalf, D., Di Rago, L., Mifsud, S., Wu, L., and Nutt, S.L. (2005). PU.1 regulates the commitment of adult hematopoietic progenitors and restricts granulopoiesis. *J. Exp. Med.* 201, 1487–1502.
- David-Fung, E.-S., Yui, M.A., Morales, M., Wang, H., Taghon, T., Diamond, R.A., and Rothenberg, E. V. (2006). Progression of regulatory gene expression states in fetal and adult pro-T-cell development. *Immunol. Rev.* 209, 212–236.
- Davidich, M.I., and Bornholdt, S. (2008). Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One* 3, e1672.
- Dealy, S., Kauffman, S., and Socolar, J. (2005). Modeling pathways of differentiation in genetic regulatory networks with Boolean networks. *Complexity* 11, 52–60.
- Discher, D.E., Mooney, D.J., and Zandstra, P.W. (2009). Growth factors, matrices, and forces combine and control stem cells. *Science* (80-.). 324, 1673–1677.
- Dolens, A.-C., and Taghon, T. (2017). Human T cell development notched up a level. *Nat. Methods* 14, 477–478.
- Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S., and Smith, A.G. (2014). Defining an essential transcription factor program for naïve pluripotency. *Science* 344, 1156–1160.
- Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014). Gene regulatory networks

and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* 2, 38.

Fesnak, A.D., June, C.H., and Levine, B.L. (2016). Engineered T cells: the promise and challenges of cancer immunotherapy. *Nat. Rev. Cancer* 16, 566–581.

Fiers, M.W.E.J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* 1–9.

Filipczyk, A., Marr, C., Hastreiter, S., Feigelman, J., Schwarzfischer, M., Hoppe, P.S., Loeffler, D., Kokkaliaris, K.D., Ende, M., Schauburger, B., et al. (2015). Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat. Cell Biol.* 17, 1235–1246.

Fluri, D.A., Tonge, P.D., Song, H., Baptista, R.P., Shakiba, N., Shukla, S., Clarke, G., Nagy, A., and Zandstra, P.W. (2012). Derivation, expansion and differentiation of induced pluripotent stem cells in continuous suspension cultures. *Nat. Methods* 9, 509–516.

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

Georgescu, C., Longabaugh, W.J.R., Scripture-Adams, D.D., David-Fung, E.-S., Yui, M.A., Zarnegar, M.A., Bolouri, H., and Rothenberg, E. V (2008). A gene regulatory network armature for T lymphocyte specification. *Proc. Natl. Acad. Sci. U. S. A.* 105, 20100–20105.

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y., and Kitano, H. (2011). Software for systems biology: from tools to integrated platforms. *Nat. Rev. Genet.* 12, 821–832.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31, 2989–2998.

Hamey, F.K., Nestorowa, S., Wilson, N.K., and Göttgens, B. (2016). Advancing haematopoietic stem and progenitor cell biology through single-cell profiling. *FEBS Lett.* *590*, 4052–4067.

Herrmann, F., Groß, A., Zhou, D., Kestler, H.A., and Köhl, M. (2012). A Boolean model of the cardiac gene regulatory network determining first and second heart field identity. *PLoS One* *7*, e46798.

Hosokawa, H., and Rothenberg, E. V (2018). Cytokines, transcription factors, and the initiation of T-cell development. *Cold Spring Harb. Perspect. Biol.* *10*, a028621.

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D.E. (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* *94*, 1–4.

Huynh-Thu, V.A., and Sanguinetti, G. (2018). Gene regulatory network inference: an introductory survey. *BioRxiv* 1–23.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* *5*, e12776.

Ikawa, T., Hirose, S., Masuda, K., Kakugawa, K., Satoh, R., Shibano-Satoh, A., Kominami, R., Katsura, Y., and Kawamoto, H. (2010). An essential developmental checkpoint for production of the T cell lineage. *Science* *329*, 93–96.

Jenkinson, E.J., and Anderson, G. (1994). Fetal thymic organ cultures. *Curr. Opin. Immunol.* *6*, 293–297.

Ji, Z., and Ji, H. (2016). TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* *44*, e117–e117.

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* *9*, 770–780.

Kaul, H., and Ventikos, Y. (2015). Investigating biocomplexity through the agent-based

paradigm. *Brief. Bioinform.* *16*, 137–152.

Kawazu, M., Asai, T., Ichikawa, M., Yamamoto, G., Saito, T., Goyama, S., Mitani, K., Miyazono, K., Chiba, S., Ogawa, S., et al. (2005). Functional domains of Runx1 are differentially required for CD4 repression, TCRbeta expression, and CD4/8 double-negative to CD4/8 double-positive transition in thymocyte development. *J. Immunol.* *174*, 3526–3533.

Kim, K., Doi, a, Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I.R., et al. (2010). Epigenetic memory in induced pluripotent stem cells. *Nature* *467*, 285–290.

Krumsiek, J., Marr, C., Schroeder, T., and Theis, F.J. (2011). Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS One* *6*, e22649.

Kueh, H.Y., and Rothenberg, E. V. (2012). Regulatory gene network circuits underlying T cell development from multipotent progenitors. *Wiley Interdiscip. Rev. Syst. Biol. Med.* *4*, 79–102.

Kueh, H.Y., Yui, M.A., Ng, K.K.H., Pease, S.S., Zhang, J.A., Damle, S.S., Freedman, G., Siu, S., Bernstein, I.D., Elowitz, M.B., et al. (2016). Asynchronous combinatorial action of four regulatory factors activates *Bcl11b* for T cell commitment. *Nat. Immunol.* *17*, 956–965.

Lim, C.Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., and Göttgens, B. (2016). BTR: training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics* *17*, 355.

Longabaugh, W.J.R., Zeng, W., Zhang, J.A., Hosokawa, H., Jansen, C.S., Li, L., Romero-Wolf, M., Liu, P., Kueh, H.Y., Mortazavi, A., et al. (2017). *Bcl11b* and combinatorial resolution of cell fate in the T-cell gene regulatory network. *PNAS* *114*, 5800–5807.

Love, P.E., and Bhandoola, A. (2011). Signal integration and crosstalk during thymocyte

migration and emigration. *Nat. Rev. Immunol.* *11*, 469–477.

Lummertz, E., Rowe, R.G., Lundin, V., Malleshaiah, M., Jha, D.K., Rambo, C.R., Li, H., North, T.E., Collins, J.J., and Daley, G.Q. (2018). Reconstruction of complex single-cell trajectories using CellRouter. 1–13.

Manesso, E., Kueh, H.Y., Freedman, G., Rothenberg, E. V., and Peterson, C. (2016). Irreversibility of T-cell specification: insights from computational modelling of a minimal network architecture. *PLoS One* *11*, e0161260.

Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U. S. A.* *107*, 6286–6291.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Consortium, T.D., Kellis, M., Collins, J.J., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* *9*, 796–804.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases.

Marco, E., Karp, R.L., Guo, G., Robson, P., Hart, A.H., Trippa, L., and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci.* *111*, E5643–E5650.

McNamara, L.E., Turner, L.-A., and Burgess, K. V (2015). Systems biology approaches applied to regenerative medicine. *Curr. Pathobiol. Rep.* *3*, 37–45.

Mingueneau, M., Kreslavsky, T., Gray, D., Heng, T., Cruse, R., Ericson, J., Bendall, S., Spitzer, M.H., Nolan, G.P., Kobayashi, K., et al. (2013). The transcriptional landscape of $\alpha\beta$ T cell differentiation. *Nat. Immunol.* *14*, 619–632.

Mohanty, A.K., Datta, A., and Venkatraj, V. (2014). A model for cancer tissue heterogeneity. *IEEE Trans. Biomed. Eng.* *61*, 966–974.

Mohtashami, M., Shah, D.K., Nakase, H., Kianizad, K., Petrie, H.T., and Zuniga-Pflucker, J.C. (2010). Direct comparison of Dll1- and Dll4-mediated Notch activation levels shows differential lymphomyeloid lineage commitment outcomes. *J. Immunol.* *185*, 867–876.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E., et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* *33*, 269–276.

Morris, S.A., Cahan, P., Li, H., Zhao, A.M., San Roman, A.K., Shivdasani, R.A., Collins, J.J., and Daley, G.Q. (2014). Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* *158*, 889–902.

Morsut, L., Roybal, K.T., Xiong, X., Gordley, R.M., Coyle, S.M., Thomson, M., and Lim, W.A. (2016). Engineering customized cell sensing and response behaviors using synthetic Notch receptors. *Cell* *164*, 780–791.

Motte-Mohs, R.N. La, Herer, E., and Zu, J.C. (2005). Induction of T-cell development from human cord blood hematopoietic stem cells by Delta-like 1 in vitro. *Blood* *105*, 1431–1440.

Nandagopal, N., Santat, L., Lebon, L., Sprinzak, D., Bronner, M., and Elowitz, M. (2018). Dynamic ligand discrimination in the Notch pathway. *Cell* *172*, 869–880.

Ng, S.W.K., Mitchell, A., Kennedy, J.A., Chen, W.C., McLeod, J., Ibrahimova, N., Arruda, A., Popescu, A., Gupta, V., Schimmer, A.D., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* *540*, 433–437.

Nishimura, T., Kaneko, S., Kawana-Tachikawa, A., Tajima, Y., Goto, H., Zhu, D., Nakayama-Hosoya, K., Iriguchi, S., Uemura, Y., Shimizu, T., et al. (2013). Generation of rejuvenated antigen-specific T cells by reprogramming to pluripotency and redifferentiation. *Cell Stem Cell* *12*, 114–126.

Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks.

Nat. Rev. Genet. *16*, 146–158.

Okamura, R.M., Sigvardsson, M., Galceran, J., Verbeek, S., Clevers, H., and Grosschedl, R. (1998). Redundant regulation of T cell differentiation and TCRalpha gene expression by the transcription factors LEF-1 and TCF-1. *Immunity* *8*, 11–20.

Okawa, S., and del Sol, A. (2015). A computational strategy for predicting lineage specifiers in stem cell subpopulations. *Stem Cell Res.* *15*, 427–434.

Peter, I.S., and Davidson, E.H. (2010). The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage. *Dev. Biol.* *340*, 188–199.

Petrie, H.T., and Kincade, P.W. (2005). Many roads, one destination for T cell progenitors. *J. Exp. Med.* *202*, 11–13.

Petrie, H.T., and Zúñiga-Pflücker, J.C. (2007). Zoned out: functional mapping of stromal signaling microenvironments in the thymus. *Annu. Rev. Immunol.* *25*, 649–679.

Polo, J.M., Liu, S., Figueroa, M.E., Kulalert, W., Eminli, S., Tan, K.Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., et al. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat. Biotechnol.* *28*, 848–855.

Porritt, H.E., Rumfelt, L.L., Tabrizifard, S., Schmitt, T.M., Zuniga-Pflucker, J.C., and Petrie, H.T. (2004). Heterogeneity among DN1 prothymocytes reveals multiple progenitors with different capacities to generate T cell and non-T cell lineages. *Immunity* *20*, 735–745.

Prockop, S.E., and Petrie, H.T. (2004). Regulation of thymus size by competition for stromal niches among early T cell progenitors. *J. Immunol.* *173*, 1604–1611.

Qiao, W., Wang, W., Laurenti, E., Turinsky, A.L., Wodak, S.J., Bader, G.D., Dick, J.E., and Zandstra, P.W. (2014). Intercellular network structure and regulatory motifs in the human hematopoietic system. *Mol. Syst. Biol.* *10*, 741.

Rahman, N., Brauer, P.M., Ho, L., Usenko, T., Tewary, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W. (2017). Engineering the haemogenic niche mitigates endogenous inhibitory signals and controls pluripotent stem cell-derived blood emergence. *Nat. Commun.* 8, 15380.

Del Real, M.M., and Rothenberg, E. V (2013). Architecture of a lymphomyeloid developmental switch controlled by PU.1, Notch and Gata3. *Development* 140, 1207–1219.

Rothenberg, E. V., Kueh, H.Y., Yui, M.A., and Zhang, J.A. (2016). Hematopoiesis and T-cell specification as a model developmental system. *Immunol. Rev.* 271, 72–97.

Roybal, K.T., Rupp, L.J., Morsut, L., Walker, W.J., McNally, K.A., Park, J.S., and Lim, W.A. (2016). Precision Tumor Recognition by T Cells With Combinatorial Antigen-Sensing Circuits. *Cell* 164, 770–779.

Sánchez, L., and Thieffry, D. (2003). Segmenting the fly embryo: a logical analysis of the pair-rule cross-regulatory module. *J. Theor. Biol.* 224, 517–537.

Schilham, M.W., Wilson, A., Moerer, P., Benaissa-Trouw, B.J., Cumano, A., and Clevers, H.C. (1998). Critical involvement of Tcf-1 in expansion of thymocytes. *J. Immunol.* 161, 3984–3991.

Schlitt, T., and Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics* 8, S9.

Schmitt, T.M., Zú, J.C., and Cker, I.-P. (2002). Induction of T cell development from hematopoietic progenitor cells by Delta-like-1 in vitro. *Immunity* 17, 749–756.

Schmitt, T.M., Ciofani, M., Petrie, H.T., and Zúñiga-Pflücker, J.C. (2004). Maintenance of T cell specification and differentiation requires recurrent notch receptor-ligand interactions. *J. Exp. Med.* 200, 469–479.

Scott, E.W., Simon, M.C., Anastasi, J., and Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science*

265, 1573–1577.

Setty, M., Tadmor, M.D., Reich-zeliger, S., Angel, O., Salame, T.M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 1–14.

Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G., et al. (2015). Single-cell RNA-Seq with Waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17, 360–372.

Shukla, S., Langley, M.A., Singh, J., Edgar, J.M., Mohtashami, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W. (2017). Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like 4 and VCAM-1. *Nat. Methods* 14, 531–538.

Simunovic, M., and Brivanlou, A.H. (2017). Embryoids, organoids and gastruloids: new approaches to understanding embryogenesis. *Development* 144, 976–985.

Singer, Z.S., Yong, J., Tischler, J., Hackett, J.A., Altinok, A., Surani, M.A., Cai, L., and Elowitz, M.B. (2014). Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* 55, 319–331.

Su, D., Navarre, S., Oh, W., Condie, B.G., and Manley, N.R. (2003). A domain of Foxn1 required for crosstalk-dependent thymic epithelial cell differentiation. *Nat. Immunol.* 4, 1128–1135.

Taghon, T., De Smedt, M., Stolz, F., Cnockaert, M., Plum, J., and Leclercq, G. (2001). Enforced expression of GATA-3 severely reduces human thymic cellularity. *J. Immunol.* 167, 4468–4475.

Tewary, M., Ostblom, J., Prochazka, L., Zulueta-Coarasa, T., Shakiba, N., Fernandez-Gonzalez, R., and Zandstra, P.W. (2017). A stepwise model of Reaction-Diffusion and Positional-Information governs self-organized human peri-gastrulation-like patterning. *Development* 144, 4298–4312.

- Themeli, M., Kloss, C.C., Ciriello, G., Fedorov, V.D., Perna, F., Gonen, M., and Sadelain, M. (2013). Generation of tumor-targeted human T lymphocytes from induced pluripotent stem cells for cancer therapy. *Nat. Biotechnol.* *31*, 928–933.
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* *32*, 381–386.
- Verbeek, S., Izon, D., Hofhuis, F., Robanus-Maandag, E., te Riele, H., Watering, M. van de, Oosterwegel, M., Wilson, A., Robson MacDonald, H., and Clevers, H. (1995). An HMG-box-containing T-cell factor required for thymocyte differentiation. *Nature* *374*, 70–74.
- Wang, H., Pierce, L.J., and Spangrude, G.J. (2006). Distinct roles of IL-7 and stem cell factor in the OP9-DL1 T-cell differentiation culture system. *Exp. Hematol.* *34*, 1730–1740.
- Wang, J.H., Nichogiannopoulou, A., Wu, L., Sun, L., Sharpe, A.H., Bigby, M., and Georgopoulos, K. (1996). Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. *Immunity* *5*, 537–549.
- Welch, J.D., Hartemink, A.J., and Prins, J.F. (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* *17*, 106.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* *3*, 160018.
- Xiao, Y. (2009). A tutorial on analysis and simulation of Boolean gene regulatory network models. *10*, 511–525.
- Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I.R., and Ma'ayan, A. (2014). Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput. Biol.* *10*, e1003777.

Yachie-Kinoshita, A., Onishi, K., Ostblom, J., Langley, M.A., Posfai, E., Rossant, J., and Zandstra, P.W. (2018). Modeling signaling-dependent pluripotent cell states with Boolean logic can predict cell fate transitions. *Mol. Syst. Biol.* *14*, e7952.

Yordanov, B., Dunn, S.-J., Kugler, H., Smith, A., Martello, G., Emmott, S., Noverre, N., Le, Jong, H. de, Morris, M.K., Saez-Rodriguez, J., et al. (2016). A method to identify and analyze biological programs through automated reasoning. *Npj Syst. Biol. Appl.* *2*, 16010.

Yu, J.S., and Bagheri, N. (2016). Multi-class and multi-scale models of complex biological phenomena. *Curr. Opin. Biotechnol.* *39*, 167–173.

Yui, M.A., and Rothenberg, E. V (2014). Developmental gene networks: a triathlon on the course to T cell identity. *Nat. Rev. Immunol.* *14*, 529–545.

Yui, M.A., Feng, N., and Rothenberg, E. V (2010). Fine-scale staging of T cell lineage commitment in adult mouse thymus. *J. Immunol.* *185*, 284–293.

Ziętara, N., Łyszkiewicz, M., Puchałka, J., Witzlau, K., Reinhardt, A., Förster, R., Pabst, O., Prinz, I., and Krueger, A. (2015). Multicongenic fate mapping quantification of dynamics of thymus colonization. *J. Exp. Med.* *212*, 1589–1601.

Copyright Acknowledgments

Chapter 4

Shukla, S., **Langley, M.A.**, Singh, J., Edgar, J.M., Mohtashami, M., Zúñiga-Pflücker, J.C., and Zandstra, P.W. (2017). Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like 4 and VCAM-1. *Nature Methods* 14, 531–538.

Chapter 5

Yachie-Kinoshita, A., Onishi, K., Ostblom, J., **Langley, M.A.**, Posfai, E., Rossant, J., and Zandstra, P.W. (2018). Modeling signaling-dependent pluripotent cell states with Boolean logic can predict cell fate transitions. *Molecular Systems Biology* 14, e7952.

Appendix A: Software Resources

Python code developed for Boolean network simulation and state transition graph analysis are available at: <https://gitlab.com/stemcellbioengineering/garuda-boolean>

Garuda gadgets are available at the following links:

Discretize, <http://gateway.garuda-alliance.org/node/86>

Boolean Simulation, <http://gateway.garuda-alliance.org/node/88>

Boolean SCC Analysis, <http://gateway.garuda-alliance.org/node/87>

The following section describes the features and usage pattern for the three publicly-available *Garuda* gadgets that were developed for Boolean network analysis: Discretize, Boolean Simulation, and Boolean SCC Analysis.

Gadget 1: Discretize



Discretize

Input

Gene	Bcl11a	Bcl11b	Bcl2	Cd5e	Cd5g	Cebp
proB_CLP_BM	1498.81	83.86	129.72	79.45	40.32	128.24
1607.79	73.29	77.46	65.6	32.32	216.84	
1774.16	93.18	53.74	100.23	119.59	84.1	
1763.55	170.43	79.25	101.88	109.57	66.72	
733.83	709.64	83.82	186.54	1462.89	68.24	
843.32	544.6		104.8	378.76	58.31	
preT_DN2A_Th	256.45	1297.29	88.1			
preT_DN2B_Th	302.1	1338.32	161			
preT_DN2-3_Th	253.6	2023.82	111			
preT_DN3A_Th	171.54	970.28	75			
preT_DN3B_Th	118.69	640.37	58.71	1702.09	5916.98	74.76
preT_DN3-4_Th	82.03	933.52	75.72	1748.7	6250.51	59.06
T_DN4_Th	73.48	1072.11	46.44	1266.16	4526.46	56.23
T_ISP_Th	132.71	1953.73	73.5	2103.26	5366.85	111.4
T_DP_Th	80.93	1860.74	61.19	1801.03	5080.31	65.25
T_DPhl_Th	103.84	1720.44	65.63	2465.53	6741.44	76.21
T_DPsm_Th						

Output

Gene	Bcl11a	Bcl11b	Bcl2	Cd5e	Cd5g	Cebp
proB_CLP_BM	1	0	1	0	0	0
proB_CLP_FL	1	0	0	0	0	0
preT_ETP_Th	1	0	0	0	0	0
preT_ETP-2A_Th	1	0	0	0	0	0
preT_DN2_Th	0	0	0	0	0	0
preT_DN2A_Th	0	0	0	0	0	0
preT_DN2B_Th	0	1	0	0	1	0
preT_DN2-3_Th	0	1	1	1	1	0
preT_DN3A_Th	0					
preT_DN3B_Th	0					
preT_DN3-4_Th	0					
T_DN4_Th	0					
T_ISP_Th	0	1	0	1	1	0
T_DP_Th	0	1	0	1	1	0
T_DPhl_Th	0	1	0	1	1	0
T_DPsm_Th	0	1	0	1	1	0
T_DPsm_Th						

Figure 22: "Discretize" gadget for Garuda pipeline

<http://gateway.garuda-alliance.org/node/86>

This gadget converts matrices of continuous gene expression data into discrete levels (ex. binary ON/OFF). Discretization is performed using the k -means algorithm, where the parameter k specifies how many levels the samples will be grouped into.

Input

- Gene Expression Data
csv format; one sample per row, one gene per column

Output

- Discretized Expression Profiles
csv format; same layout as input file, but values now set to discrete levels (ex. 0 or 1), where smaller values correspond to lower expression levels

Options

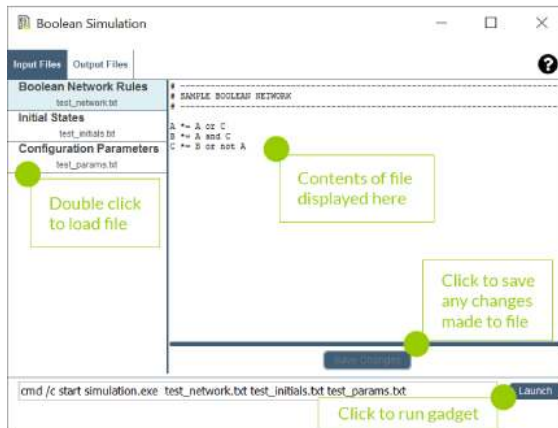
invertOrder	If true, treat large values as lowly-expressed and small values as highly-expressed. This can be useful for working with raw Ct or delta-Ct values from qRT-PCR experiments.
perGene	If true (default), the gadget considers each gene in isolation, such that unique discretization thresholds are chosen for each gene in the dataset. If false, genes are discretized together using a single common threshold.
k	The default value of k is 2 (i.e. binarization, “on” / “off”). However, the gadget can be used to discretize values to more than 2 levels by manually setting this value. For example, “--k 3” will discretize to the levels 0 (low), 1 (medium), and 2 (high).

Gadget 2: Boolean Simulation



Boolean Simulation

Input



Output

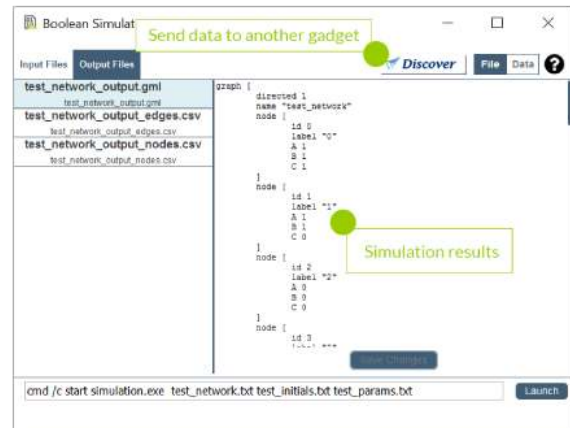


Figure 23: "Boolean Simulation" gadget for *Garuda* pipeline

<http://gateway.garuda-alliance.org/node/88>

This gadget simulates Boolean network models of gene regulatory networks. As input, the simulation takes the initial state of each element (i.e. gene) of the network and a list of Boolean logic functions describing the regulatory control of each element by the rest of the network. The simulation repeatedly applies these functions (either synchronously or asynchronously) over discrete time steps to produce a trajectory of network states. This process is repeated multiple times to generate many trajectories, which are collected into a state transition graph and dictionary files.

Input

1. Boolean Network Rules

txt format; lines beginning with “#” treated as comments

Each line defines a Boolean update function for one element (gene) in the

network; ex. `I17ra *= NOTCH_SIGNALING or E_PROTEIN or (Pu1 and not Gata3)`

2. Initial Conditions

txt format; lines beginning with “#” treated as comments

Each line specifies the initial state of an element (gene); ex. `I17ra = False`
 Can be initialized to “True”, “False”, or “Random”

3. Configuration Parameters

txt format; lines beginning with “#” treated as comments

Parameters are specified as “<parameter> = <value>”; ex. `runs = 250`

Output

1. State Transition Graph

Graph Markup Language (gml) format; defines a directed graph where each node corresponds to a network state observed in the simulation (annotated by its gene expression), and each edge corresponds to an observed transition from one network state to another. Edge weights equal the frequency at which the edge was traversed over all simulation runs.

2. State Dictionary

csv format; one network state per row, first column contains state IDs, other columns contain expression level of all elements (genes) in that state

3. Edge Dictionary

csv format; one edge per row. First column = source state. Second column = target state. Third column = number of times that edge was traversed in all simulations (weight).

Options

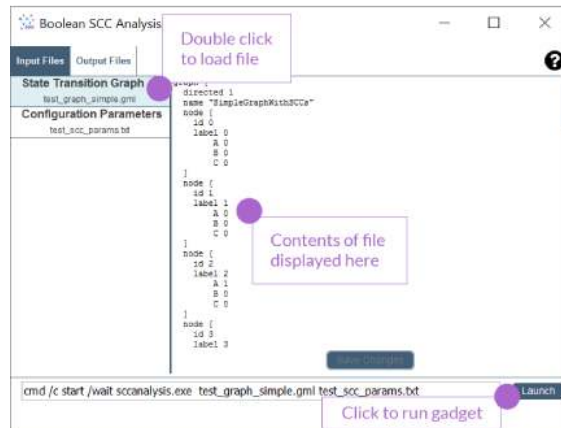
runs	Any whole number (default = 1000)	Number of independent simulations runs to perform
steps	Any whole number (default = 100)	Number of updates (iterations) per simulation trajectory
mode	async (default), sync	<i>Synchronous</i> : Update all elements at each iteration, such that each network state has only one possible following state <i>Asynchronous</i> : Update a random subset of elements at each iteration, such that each network state can lead to many possible following states

Gadget 3: Boolean SCC Analysis



Boolean SCC Analysis

Input



Output

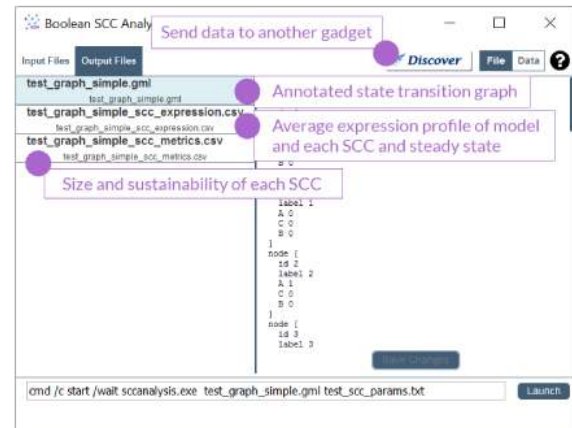


Figure 24: "Boolean SCC Analysis" gadget for *Garuda* pipeline

<http://gateway.garuda-alliance.org/node/87>

This gadget analyzes state transition graphs from Boolean network simulations to identify steady states and strongly connected components. A steady state is a terminal state within the graph that loops back on to itself. Conversely, a strongly connected component (SCC) is a set of mutually reachable states, which can be considered as a “dynamic steady state” of the network. In the case of SCCs, the internal edge weights (frequencies) of the SCC subgraph are used to define a Markov chain. The average gene expression profile for the SCC is calculated as average expression profile is calculated by multiplying the gene expression profile of each state in the SCC against the stationary distribution this Markov chain. A “sustainability” score is also calculated for each SCC based on the probability of any outgoing edges. If no steady states or strongly connected components are identified within the state transition graph, the unmodified state transition graph file and empty csv files with the text “No SCCs or steady states matching criteria were found” will be returned.

Input

1. State Transition Graph

Graph Markup Language (gml) format; defines a directed graph where each node corresponds to a network state observed in the simulation (annotated by its gene expression), and each edge corresponds to an observed transition from one network state to another. Edge weights equal the frequency at which the edge was traversed over all simulation runs.

2. Configuration Parameters

txt format; lines beginning with “#” treated as comments

Parameters are specified as “<parameter> = <value>”; ex. `minSize = 10`

Output

1. State Transition Graph (annotated)

Same format as input, but states which belong to steady states or SCCs are annotated

2. Expression Profiles

csv format; summarizes the probability-weighted average expression profile of each identified steady state and SCC, as well as the weighted average over all SCCs of the model

3. SCC Metrics

csv format; summarizes the number of profiles, number of edges, and sustainability score of each identified SCC

Options

minSize	Any whole number (default = 0)	Minimum size of SCCs to be considered, in terms of numbers of profiles. All returned SCCs will have size > minSize
minSustainability	Any whole number (default = 0.0)	Minimum sustainability score of SCCs to be considered. All returned SCCs will have sustainability > minSustainability
writeSubgraphs	False (default), True	Write GML representations of identified SCCs and SSs to file
annotateGraph	False (default), True	Annotate nodes in input graph with SCC/SS membership

Appendix B: Supplementary Tables

Supplementary Table 1: Literature evidence for edges in Boolean network (BN) model of mouse T cell development

Source	Sign	Target	Reference	Experiment	Cell Type
Bcl11b	Positive	Cd3e	Ikawa (Science, 2010)	Bcl11b overexpression + in vitro T cell culture + mRNA analysis	FL Lin-Sca1+Kit+ (LSK)
Bcl11b	Positive	Cd3e	Longabaugh (PNAS, 2017)	Bcl11b KO + RNA-seq + ChIP with Bcl11b antibody	Bcl11b KO DN2 thymocytes (OP9-DL4 from E13.5 FL)
Bcl11b	Positive	Cd3g	Longabaugh (PNAS, 2017)	Bcl11b KO + RNA-seq + ChIP with Bcl11b antibody	Bcl11b KO DN2 thymocytes (OP9-DL4 from E13.5 FL)
Bcl11b	Negative	Cebpa	Ikawa (Science, 2010)	Bcl11b overexpression + in vitro T cell culture + mRNA analysis	Bcl11b ^{-/-} FL Lin-Sca1+Kit+ (LSK)
Bcl11b	Negative	Id3	Longabaugh (PNAS, 2017)	Bcl11b KO + RNA-seq	Bcl11b KO DN2 thymocytes (OP9-DL4 from E13.5 FL)
Bcl11b	Negative	Kit	Ikawa (Science, 2010)	Bcl11b overexpression + in vitro T cell culture + surface expression analysis using FACS	Bcl11b ^{-/-} FL Lin-Sca1+Kit+ (LSK)
Bcl11b	Negative	Kit	Longabaugh (PNAS, 2017)	Bcl11b KO + RNA-seq + ChIP with Bcl11b antibody	Bcl11b KO DN2 thymocytes (OP9-DL4 from E13.5 FL)
Bcl11b	Positive	Ptcr	Ikawa (Science, 2010)	Bcl11b overexpression + in vitro T cell culture + mRNA analysis	Bcl11b ^{-/-} FL Lin-Sca1+Kit+ (LSK)
Bcl11b	Negative	Pu1	Ikawa (Science, 2010)	Bcl11b overexpression + in vitro T cell culture + mRNA analysis	Bcl11b ^{-/-} FL Lin-Sca1+Kit+ (LSK)
Cd3e	Positive	TCR_SIGNALING	By definition (receptor complex)	-	-
Cd3g	Positive	TCR_SIGNALING	By definition (receptor complex)	-	-
E_PROTEIN	Positive	Gata3	Gregoire (J Biol Chem, 1999)	Analysis of Gata3 cis-regulatory element (wild-type/mutated E-protein binding site)	Jurkat T cell line
E_PROTEIN	Positive	Gata3	Jones-Mason (Immunity, 2012)	Tcf12/Tcfe2a KO or Id2/Id3 KO + intracellular staining	Tcf12 ^{fl/fl} TcfE2 ^{fl/fl} Cd4cre ⁺ and Id2 ^{fl/fl} Id3 ^{fl/fl} Cd4cre ⁺ thymocytes
E_PROTEIN	Positive	Gfi1	Schwartz (PNAS, 2006)	E47 overexpression + mRNA expression analysis	1F9 E2A ^{-/-} T cell lymphoma line
E_PROTEIN	Positive	Gfi1	Xu (Blood, 2007)	Ectopic E2A expression + mRNA expression analysis + ChIP with E47 antibody	0531 and 1F9 E2A ^{-/-} T cell lymphoma line

Source	Sign	Target	Reference	Experiment	Cell Type
E_PROTEIN	Positive	Hes1	Ikawa (J Exp Med, 2006)	E47 overexpression + mRNA expression analysis	E2A ^{-/-} Lin- BM cells (cultured)
E_PROTEIN	Positive	Il7ra	Dias (Immunity, 2008)	mRNA expression profiling	E2A ^{-/-} Lin-Sca1+Kit+ Flt3-hi (LMPP)
E_PROTEIN	Positive	Il7ra	Ikawa (J Exp Med, 2006)	E47 overexpression + mRNA expression analysis	E2A ^{-/-} Lin- BM cells (cultured)
E_PROTEIN	Positive	Il7ra	Welinder (PNAS, 2011)	mRNA expression analysis of HEB KO, E2A KO and WT	E2A ^{-/-} , HEB ^{βTie2Cre} , and WT LY6D- CLPs
E_PROTEIN	Positive	Lat	Ikawa (J Exp Med, 2006)	E47 overexpression + mRNA expression analysis	E2A ^{-/-} Lin- BM cells (cultured)
E_PROTEIN	Positive	Notch1	Dias (Immunity, 2008)	mRNA expression profiling	E2A ^{-/-} Lin-Sca1+Kit+ Flt3-hi (LMPP)
E_PROTEIN	Positive	Notch1	Yashiro-Ohtani (Genes Dev, 2009)	ChIP analysis using E2A antibody E47 overexpression + Notch1 promoter activity measurements	Rag2 ^{-/-} DN3 thymocytes NIH-3T3 cell line
E_PROTEIN	Positive	Notch1	Del Real (Development, 2013)	Ectopic ID2 overexpression + mRNA expression analysis	Scid.adh.2C2 cells
E_PROTEIN	Positive	Ptcr	Ikawa (J Exp Med, 2006)	E47 overexpression + mRNA expression analysis	E2A ^{-/-} Lin- BM cells (cultured)
E_PROTEIN	Positive	Rag1	Dias (Immunity, 2008)	mRNA expression profiling	E2A ^{-/-} Lin-Sca1+Kit+ Flt3-hi (LMPP)
E_PROTEIN	Positive	Rag1	Schwartz (PNAS, 2006)	E47 overexpression + mRNA expression analysis	1F9 E2A ^{-/-} T cell lymphoma line
E2a	Positive	E_PROTEIN	By definition	-	-
E2a	Positive	E2a	Absence of other reported positive inputs	-	-
E2a	Positive	Ebfl	Ikawa (Immunity, 2004)	Ectopic E47 expression + mRNA expression analysis	E2A-deficient BM progenitors
E2a	Positive	Ebfl	Greenbaum (PNAS, 2002)	E2AFH + ChIP with anti-FLAG antibody	Abelson pre-B E2AFH line
Ebfl	Negative	Id3	Thal (PNAS, 2009)	EMSA of Ebfl binding to Id3 promoter + mRNA expression + id3-luciferase reporter following Ebfl expression plasmid	Sorted primary pre-pro-B cells
Ets1	Positive	Tcrb	Kim (EMBO J, 1999)	Ets1 overexpression + analysis of TCRβenhancer activity	p19 cell line
Gata3	Positive	Bcl11b	Kueh (Nat Immunol, 2016)	Gata3 KO + Bcl11b-YFP reporter analysis	Bcl11b-YFP ETP thymocytes

Source	Sign	Target	Reference	Experiment	Cell Type
Gata3	Positive	Bcl11b	Garcia-Ojeda (Blood, 2013)	Gata3 KO + mRNA expression analysis	Gata3 ^{-/-} DN2 thymocytes
Gata3	Positive	Bcl11b	Scripture-Adams (J Immunol, 2014)	Gata3 shRNA-knockdown or flox-knockout + mRNA expression analysis	WT and Gata3 ^{fl/fl} DN thymocytes
Gata3	Positive	Bcl11b	Zhang (Cell, 2012)	ChIP analysis using GATA-3 antibody	Fetal liver derived thymocytes
Gata3	Positive	Cd3e	Zhang (Cell, 2012)	ChIP analysis using GATA-3 antibody	Fetal liver derived thymocytes
Gata3	Positive	Cd3g	Zhang (Cell, 2012)	ChIP analysis using GATA-3 antibody	Fetal liver derived thymocytes
Gata3	Negative	Cebpa	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Positive	Deltex	Wang (Mol Cell Biol, 2009)	Gata3 overexpression + mRNA expression analysis	Adult DN1 thymocytes
Gata3	Negative	Id3	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Negative	Il7ra	Anderson (Devel Bio, 2002)	Gata3 ectopic expression + mRNA expression analysis	Fetal liver derived thymocytes
Gata3	Negative	Il7ra	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Negative	Ptcra	Anderson (Devel Bio, 2002)	Gata3 ectopic expression + mRNA expression analysis	Fetal liver derived thymocytes
Gata3	Negative	Ptcra	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Negative	Pu1	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Negative	Pu1	Anderson (Devel Bio, 2002)	Gata3 ectopic expression + mRNA expression analysis	Fetal liver derived thymocytes
Gata3	Positive	Scl	Anderson (Devel Bio, 2002)	Gata3 ectopic expression + mRNA expression analysis	Fetal liver derived thymocytes
Gata3	Positive	Scl	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Negative	Tcf7	Taghon (Nat Immunol, 2007)	Gata3 overexpression + mRNA expression analysis	Bcl2-transgenic thymocytes
Gata3	Positive	Tcrb	Yang (Blood, 2003)	Measurement of TCR β enhancer activity (wild-type/mutated GATA binding site)	p5424 T cell line

Source	Sign	Target	Reference	Experiment	Cell Type
Gfi1	Negative	Pu1	Spooner (Immunity, 2009)	mRNA expression analysis; ChIP analysis using Gfi1 antibody	Gfi1 ^{-/-} Lin-Sca1+Kit ⁺ (LSK) cells
Gfi1	Negative	Pu1	Wei (Cell Res, 2008)	Gfi1 overexpression/ knockdown + in situ RNA hybridization	Zebrafish embryos
Gfi1	Positive	Rag1	Wei (Cell Res, 2008)	Gfi1 knockdown + in situ RNA hybridization	Zebrafish embryos
Gfi1	Positive	Scl	Wei (Cell Res, 2008)	Gfi1 knockdown + in situ RNA hybridization	Zebrafish embryos
Gfi1b	Negative	Gfi1	Doan (Nucleic Acids Res, 2004)	mRNA expression analysis	Gfi1b transgenic thymocytes
Gfi1b	Negative	Gfi1	Xu (Blood, 2007)	Gfi1b overexpression + mRNA expression analysis	E2A ^{-/-} T cell lymphoma
HEB	Positive	E_PROTEIN	By definition	-	-
HEB	Positive	E_PROTEIN	Welinder (PNAS, 2011)	mRNA expression analysis of HEB KO, E2A KO and WT	E2A ^{-/-} , HEB ^{β^{fl}} Tie2Cre, and WT LY6D- CLPs
HEB	Positive	HEB	Absence of other reported positive inputs	-	-
Id3	Negative	E_PROTEIN	By definition	-	-
Ikaros	Negative	Hes1	Kathrein (J Biol Chem, 2008)	Retroviral reintroduction of Ikaros + mRNA expression analysis	Ikaros ^{-/-} JE131 cell line
Ikaros	Negative	Runx1	Chari (J Immunol, 2008)	mRNA expression analysis	Ikaros ^{-/-} adult DN thymocytes
IL7_SIGNALING	Positive	Cebpa	Ikawa (Science, 2010)	In vitro T cell culture + IL7 drop + mRNA expression analysis	FL Lin-Sca1+Kit ⁺ (LSK) cells
IL7_SIGNALING	Positive	Ebf1	Kikuchi (J Exp Med, 2005)	Constitutive IL7 signaling activation + mRNA expression analysis	IL-7Rα ^{-/-} pre-pro B-cells
IL7_SIGNALING	Positive	Ebf1	Roessler (Mol Cell Biol, 2007)	STAT5 transfection + mRNA expression analysis	Ba/F3 pro-B line
IL7_SIGNALING	Positive	Gata3	Guo (PNAS, 2009)	Flow cytometry with intracellular staining using GATA3 antibody + ChIP with Stat5 antibody	Stat5 ^{-/-} Th2 cells (n.b. cells were stimulated with IL-2 and IL-33)
IL7_SIGNALING	Positive	Kit	Ikawa (Science, 2010)	In vitro T cell culture + IL7 drop + surface expression analysis using FACS	FL Lin-Sca1+Kit ⁺ (LSK) cells
IL7_SIGNALING	Negative	Lck	Ikawa (Science, 2010)	In vitro T cell culture + IL7 drop + plck-GFP reporter + mRNA expression analysis	FL Lin-Sca1+Kit ⁺ (LSK) cells
IL7_SIGNALING	Negative	Ptcr	Ikawa (Science, 2010)	In vitro T cell culture + IL7 drop + mRNA expression analysis	FL Lin-Sca1+Kit ⁺ (LSK) cells

Source	Sign	Target	Reference	Experiment	Cell Type
IL7_SIGNALING	Positive	Pu1	Ikawa (Science, 2010)	In vitro T cell culture + IL7 drop + mRNA expression analysis	FL Lin-Sca1+Kit+ (LSK) cells
Il7ra	Positive	IL7_SIGNALING	By definition (receptor)	-	-
INPUT_DL4	Positive	NOTCH_SIGNALING	By definition (ligand)	-	-
INPUT_IL7	Positive	IL7_SIGNALING	By definition (ligand)	-	-
INPUT_TCR	Positive	TCR_SIGNALING	By definition (ligand)	-	-
Lck	Positive	TCR_SIGNALING	By definition (receptor complex)	-	-
Lmo2	Positive	Hhex	McCormack (Science, 2010)	mRNA expression analysis	Lmo transgenic DN3 leukemic cells
Lmo2	Positive	Hhex	Smith (PLOS One, 2014)	ChIP using LMO2 antibody + LMO2 knockdown + mRNA expression analysis	Human T-ALL line
Lmo2	Positive	Kit	McCormack (Science, 2010)	mRNA expression analysis	Lmo transgenic DN3 leukemic cells
Lmo2	Positive	Lyl1	McCormack (Science, 2010)	mRNA expression analysis	Lmo transgenic DN3 leukemic cells
Lyl1	Positive	Id3	San-Marina (Biochim Biophys Acta, 2008)	Lyl1 overexpression + mRNA expression analysis	Human AML cells
Lyl1	Negative	Ptcr	Herblot (Nat Immunol, 2000)	mRNA expression analysis	Scl-Lmo1 transgenic mice
Lyl1	Negative	Ptcr	Herblot (Nat Immunol, 2000)	Scl overexpression + pT α enhancer activity analysis	AD10.1 immature T cell line
Myb	Positive	Gata3	Del Real (Development, 2013)	Ectopic expression of Myb + intracellular staining of GATA3	Scid.adh.2C2 cells
Myb	Positive	Gata3	Gimferrer (J Immunol, 2011)	Myb KO + intracellular staining of GATA3	c-Myb ^{fl} cd4Cre and WT DP thymocytes
Myb	Positive	Gata3	Maurice (EMBO J, 2007)	Dominant negative Myb overexpression + mRNA expression analysis	E16 cell line
NOTCH_SIGNALING	Positive	Bcl11b	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes
NOTCH_SIGNALING	Positive	Bcl11b	Li (Science, 2010)	ChIP analysis using CSL antibody	Adult thymocytes
NOTCH_SIGNALING	Positive	Bcl11b	Tydell (J Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+CD27+ progenitors
NOTCH_SIGNALING	Positive	Cd3e	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Cd3g	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes

Source	Sign	Target	Reference	Experiment	Cell Type
NOTCH_SIGNALING	Negative	Cebpa	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes
NOTCH_SIGNALING	Positive	Deltex	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes
NOTCH_SIGNALING	Positive	Deltex	Taghon (Genes Dev, 2005)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Deltex	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Negative	Ebf1	Taghon (Genes Dev, 2005)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Gata3	Taghon (Genes Dev, 2005)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Gata3	Van de Walle (Blood, 2009)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Gata3	Weerkamp (Leukemia, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Gata3	Tydell (J Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Hes1	Taghon (Genes Dev, 2005)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+ cells
NOTCH_SIGNALING	Positive	Hes1	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes
NOTCH_SIGNALING	Negative	Id3	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes
NOTCH_SIGNALING	Negative	Id3	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Il7ra	Gonzalez-Garcia (J Exp Med, 2009)	Notch ICN overexpression + analysis of IL-7R α promoter activity	293T and Jurkat cell lines
NOTCH_SIGNALING	Positive	Il7ra	Gonzalez-Garcia (J Exp Med, 2009)	Notch ICN overexpression + in vitro culture + surface expression analysis by flow	DN1 thymocytes
NOTCH_SIGNALING	Positive	Lat	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Lck	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Lef1	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Ptcra	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes

Source	Sign	Target	Reference	Experiment	Cell Type
NOTCH_SIGNALING	Positive	Ptcr	Reizis (Genes Dev, 2002)	Notch ICD overexpression + analysis of pTα enhancer activity	293 cell line
NOTCH_SIGNALING	Positive	Ptcr	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Rag1	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Runx1	Franco (PNAS, 2006)	In vitro culture (+/-DL1) + mRNA expression analysis	Thy1+ fetal thymocytes
NOTCH_SIGNALING	Positive	Runx1	Nakagawa (Blood, 2006)	Notch1 overexpression + mRNA expression analysis	NIH-3T3 cell line
NOTCH_SIGNALING	Positive	Runx1	Taghon (Nat Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	Bcl2-transgenic thymocytes
NOTCH_SIGNALING	Positive	Runx1	Del Real (Development, 2013)	In vitro culture (+/-DL1) + mRNA expression analysis	DN2 and DN3 fetal thymocytes
NOTCH_SIGNALING	Positive	Tcf7	Germar (PNAS, 2011)	ChIP using antibodies for activated Notch-1 and CSL	T6E mouse T cell lymphoma line
NOTCH_SIGNALING	Positive	Tcf7	Tydell (J Immunol, 2007)	In vitro culture (+/-DL1) + mRNA expression analysis	FL Lin-Kit+CD27+ progenitors
NOTCH_SIGNALING	Positive	Tcf7	Weber (Nature, 2011)	In vitro culture (+/-DL1) + mRNA expression analysis	Lin-cKit+Sca1+ (LSK) cells
Notch1	Positive	NOTCH_SIGNALING	By definition (receptor)	-	-
Ptcr	Positive	TCR_SIGNALING	By definition (receptor complex)	-	-
Pu1	Negative	Cd3e	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Cd3g	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Ets1	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Ets1	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes
Pu1	Negative	Ets1	Champhekar (Genes Dev, 2015)	PU.1 or PU.1-Engrailed fusion protein expression + mRNA expression analysis	DN1, DN2A, and DN2B thymocytes
Pu1	Negative	Gata3	Chang (J Immunol, 2009)	ChIP using Gata3 antibody	PU.1-/- CD4+ T cells
Pu1	Negative	Gata3	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes
Pu1	Negative	Gfi1	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes

Source	Sign	Target	Reference	Experiment	Cell Type
Pu1	Negative	Gfi1	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes
Pu1	Negative	Gfi1	Champhekar (Genes Dev, 2015)	PU.1 or PU.1-Engrailed fusion protein expression + mRNA expression analysis	DN1, DN2A, and DN2B thymocytes
Pu1	Negative	Hes1	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Hes1	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes
Pu1	Negative	Id3	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Ikaros	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Positive	Il7ra	DeKoter (Immunity, 2002)	ChIP using PU.1 antibody	FL-derived pro-B cells
Pu1	Positive	Il7ra	DeKoter (Immunity, 2002)	mRNA expression analysis	FL-derived pro-B cells
Pu1	Negative	Lat	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Lck	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Positive	Lmo2	Landry (Blood, 2009)	ChIP using PU.1 antibody	416B myeloid cell line
Pu1	Positive	Lmo2	Landry (Blood, 2009)	PU.1 overexpression + Lmo2 promoter activity measurements	293T cell line
Pu1	Positive	Lmo2	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes
Pu1	Positive	Lmo2	Champhekar (Genes Dev, 2015)	PU.1 or PU.1-Engrailed fusion protein expression + mRNA expression analysis	DN1, DN2A, and DN2B thymocytes
Pu1	Positive	Lyl1	Chan (Blood, 2007)	ChIP using PU.1 antibody	416B myeloid cell line
Pu1	Positive	Lyl1	Chan (Blood, 2007)	Lyl1 promoter activity measurements (wild-type/mutated PU.1 site)	416B myeloid cell line
Pu1	Positive	Lyl1	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes
Pu1	Negative	Myb	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Myb	Del Real (Development, 2013)	Ectopic PU.1 expression + mRNA expression analysis	DN2 and DN3 fetal thymocytes

Source	Sign	Target	Reference	Experiment	Cell Type
Pu1	Negative	Myb	Champhekar (Genes Dev, 2015)	PU.1 or PU.1-Engrailed fusion protein expression + mRNA expression analysis	DN1, DN2A, and DN2B thymocytes
Pu1	Negative	Rag1	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Pu1	Negative	Zap70	Franco (PNAS, 2006)	PU.1 overexpression + mRNA expression analysis	Thy1+ fetal thymocytes
Runx1	Positive	Bcl11b	Kueh (Nat Immunol, 2016)	shRunx1, pan-Runx dominant negative, Runx1 cDNA	Bcl11b-YFP DN2 thymocytes
Runx1	Negative	Pu1	Huang (Nat Genet, 2008)	mRNA expression analysis	Runx1 ^{-/-} DN2, DN3 thymocytes
Runx1	Negative	Pu1	Zarnegar (Mol Cell Biol, 2010)	Runx1/Runx1 dominant negative overexpression + analysis of PU.1 cis-regulatory element	P2C2 immature T cell line and Raw264 myeloid cell line
Runx1	Positive	Tcrb	Kim (EMBO J, 1999)	Runx1 overexpression + analysis of TCR β enhancer activity	p19 cell line
Scl	Positive	Hhex	Donaldson (Hum Mol Genet, 2005)	Analysis of Hhex enhancer containing Scl sites	416B progenitor cell line
Scl	Positive	Hhex	Wilson (Blood, 2009)	Chip-Seq using Scl antibody on putative enhancer	HPC-7 progenitor cell line
Scl	Positive	Id3	San-Marina (Biochim Biophys Acta, 2008)	Lyl1 overexpression + mRNA expression analysis	Human AML cells
Scl	Positive	Kit	Lecuyer (Blood, 2002)	mRNA expression analysis	Immature B cells (B220+) from wild-type/SCL transgenic mice
Scl	Negative	Ptcra	Herblot (Nat Immunol, 2000)	mRNA expression analysis	Scl-Lmo1 transgenic mice
Scl	Negative	Ptcra	Herblot (Nat Immunol, 2000)	Scl overexpression + pT α enhancer activity analysis	AD10.1 immature T cell line
Tcf7	Positive	Bcl11b	Weber (Nature, 2011)	TCF-1 overexpression + mRNA expression analysis + ChIP analysis using TCF-1 antibody	Lin-cKit+Sca1+(LSK) cells
Tcf7	Positive	Cd3e	Germar (PNAS, 2011)	Tcf7 ^{-/-} followed by RNA expression microarray	Tcf7 ^{-/-} thymocytes
Tcf7	Positive	Gata3	Weber (Nature, 2011)	TCF-1 overexpression + mRNA expression analysis	Lin-cKit+Sca1+(LSK) cells
Tcf7	Positive	Gata3	Yu (Nat Immunol, 2009)	ChIP using TCF-1 antibody	TCF ^{-/-} Th2 cells
Tcf7	Positive	Lef1	Li (Mol Cell Biol, 2006)	Wnt pathway activation, ChiP using TCF antibody	DLD1 cancer cell line
Tcf7	Positive	Lef1	Weber (Nature, 2011)	TCF-1 overexpression + mRNA expression analysis	Lin-cKit+Sca1+(LSK) cells

Source	Sign	Target	Reference	Experiment	Cell Type
Tcf7	Negative	Pu1	Rosenbauer (Nat Genet, 2006)	Analysis of PU.1 cis-regulatory element (wild-type/mutated Tcf site)	EL4 T cell line
Tcf7	Positive	Tcf7	Weber (Nature, 2011)	TCF-1 overexpression + mRNA expression analysis + ChIP analysis using TCF-1 antibody	Lin-cKit+Sca1+(LSK) cells
TCR_SIGNALING	Negative	Hes1	Taghon (Immunity, 2006)	TCR β KO + mRNA expression analysis	TCR $\beta^{-/-}$ and WT adult DN thymocytes
TCR_SIGNALING	Positive	Id3	Taghon (Immunity, 2006)	TCR β KO + mRNA expression analysis	TCR $\beta^{-/-}$ and WT adult DN thymocytes
TCR_SIGNALING	Negative	Ptcr	Taghon (Immunity, 2006)	TCR β KO + mRNA expression analysis	TCR $\beta^{-/-}$ and WT adult DN thymocytes
TCR_SIGNALING	Negative	Runx1	Taghon (Immunity, 2006)	TCR β KO + mRNA expression analysis	TCR $\beta^{-/-}$ and WT adult DN thymocytes
Terb	Positive	TCR_SIGNALING	By definition (receptor complex)	-	-
Zap70	Positive	TCR_SIGNALING	By definition (receptor complex)	-	-

Supplementary Table 2: Microarray datasets used for partial correlation analysis

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM1058959	DN2	wt	DN2 replicate 1		Dec 31 2012	Barbara Kee	GPL1261	GSE43224
GSM1058960	DN2	wt	DN2 replicate 2		Dec 31 2012	Barbara Kee	GPL1261	GSE43224
GSM1058961	DN2	wt	DN2 replicate 3		Dec 31 2012	Barbara Kee	GPL1261	GSE43224
GSM1058962	DN2	E2A KO	E2A-deficient DN2 replicate 1		Dec 31 2012	Barbara Kee	GPL1261	GSE43224
GSM1058963	DN2	E2A KO	E2A-deficient DN2 replicate 2		Dec 31 2012	Barbara Kee	GPL1261	GSE43224
GSM1058964	DN2	E2A KO	E2A-deficient DN2 replicate 3		Dec 31 2012	Barbara Kee	GPL1261	GSE43224
GSM1123162	DN3	wt	WT DN3 cells rep 1	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123163	DN3	wt	WT DN3 cells rep 2	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123164	DN4	wt	WT DN4 cells rep 1	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123165	DN4	wt	WT DN4 cells rep 2	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123166	DP	wt	WT DP cells rep 1	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123167	DP	wt	WT DP cells rep 2	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123168	DN3	Ikaros KO	DN3 cells rep 1	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123169	DN3	Ikaros KO	DN3 cells rep 2	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123170	DN4	Ikaros KO	DN4 cells rep 1	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123171	DN4	Ikaros KO	DN4 cells rep 2	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123172	DP	Ikaros KO	DP cells rep 1	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090
GSM1123173	DP	Ikaros KO	DP cells rep 2	sorted cells from mouse thymus	Apr 16 2013	Susan Chan	GPL1261	GSE46090

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM1149208	DN1	Tlx1 OE, Prkdc-Scid	HOXSCID_#19_Thymocytes DN1, biological rep1	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149209	DN2	Tlx1 OE, Prkdc-Scid	HOXSCID_#19_Thymocytes DN2, biological rep1	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149210	DN3	Tlx1 OE, Prkdc-Scid	HOXSCID_#19_Thymocytes DN3, biological rep1	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149211	DN1	Prkdc-Scid	SCID_#687_Thymocytes DN1, biological rep1	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149212	DN2	Prkdc-Scid	SCID_#687_Thymocytes DN2, biological rep1	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149213	DN3	Prkdc-Scid	SCID_#687_Thymocytes DN3, biological rep1	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM1149214	DN1	Prkdc-Scid	SCID_#50_Thymocytes DN1, biological rep2	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149215	DN2	Prkdc-Scid	SCID_#50_Thymocytes DN2, biological rep2	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149216	DN3	Prkdc-Scid	SCID_#50_Thymocytes DN3, biological rep2	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149217	DN1	Tlx1 OE, Prkdc-Scid	HOXSCID_#20_Thymocytes DN1, biological rep2	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149218	DN2	Tlx1 OE, Prkdc-Scid	HOXSCID_#20_Thymocytes DN2, biological rep2	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149219	DN3	Tlx1 OE, Prkdc-Scid	HOXSCID_#20_Thymocytes DN3, biological rep2	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM1149220	DN1	Prkdc-Scid	SCID_#686_Thymocytes DN1, biological rep3	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149221	DN2	Prkdc-Scid	SCID_#686_Thymocytes DN2, biological rep3	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149222	DN3	Prkdc-Scid	SCID_#686_Thymocytes DN3, biological rep3	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149223	DN1	Tlx1 OE, Prkdc-Scid	HOXSCID_#999_Thymocytes DN1, biological rep3	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149224	DN2	Tlx1 OE, Prkdc-Scid	HOXSCID_#999_Thymocytes DN2, biological rep3	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149225	DN3	Tlx1 OE, Prkdc-Scid	HOXSCID_#999_Thymocytes DN3, biological rep3	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM1149226	DN1	Tlx1 OE, Prkdc-Scid	HOXSCID_#535_Thymocytes DN1, biological rep4	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149227	DN2	Tlx1 OE, Prkdc-Scid	HOXSCID_#535_Thymocytes DN2, biological rep4	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149228	DN3	Tlx1 OE, Prkdc-Scid	HOXSCID_#535_Thymocytes DN3, biological rep4	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149229	DN1	Prkdc-Scid	SCID_#525_Thymocytes DN1, biological rep4	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 negative thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149230	DN2	Prkdc-Scid	SCID_#525_Thymocytes DN2, biological rep4	Gene expression from CD4& CD8 double negative; CD44 positive; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421
GSM1149231	DN3	Prkdc-Scid	SCID_#525_Thymocytes DN3, biological rep4	Gene expression from CD4& CD8 double negative; CD44 negative; CD25 positive thymocytes	May 28 2013	Yan Zhen Zheng	GPL1261	GSE47421

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM1162789	HSC	wt	HSC rep1	Gene expression data from hematopoietic stem cells	Jun 13 2013	Maria Alessandra Vigano	GPL6246	GSE47940
GSM1162790	HSC	wt	HSC rep2	Gene expression data from hematopoietic stem cells	Jun 13 2013	Maria Alessandra Vigano	GPL6246	GSE47940
GSM1162791	DN2	wt	ProT rep1	Gene expression data from committed T cells	Jun 13 2013	Maria Alessandra Vigano	GPL6246	GSE47940
GSM1162792	DN2	wt	ProT rep2	Gene expression data from committed T cells	Jun 13 2013	Maria Alessandra Vigano	GPL6246	GSE47940
GSM1162793	DP	wt	DP rep1	Gene expression data from double positive T cells	Jun 13 2013	Maria Alessandra Vigano	GPL6246	GSE47940
GSM1162794	DP	wt	DP rep2	Gene expression data from double positive T cells	Jun 13 2013	Maria Alessandra Vigano	GPL6246	GSE47940
GSM399391	DP	wt	T.DP.Th#1	T.DP.Th#1, Double-Positive, All	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399392	DP	wt	T.DP.Th#2	T.DP.Th#2, Double-Positive, All	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399393	DP	wt	T.DP.Th#3	T.DP.Th#3, Double-Positive, All	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399394	DP69+	wt	T.DP69+.Th#1	T.DP69+.Th#1, Double-Positive, Early Positive Selection	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399395	DP69+	wt	T.DP69+.Th#2	T.DP69+.Th#2, Double-Positive, Early Positive Selection	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399396	DP69+	wt	T.DP69+.Th#3	T.DP69+.Th#3, Double-Positive, Early Positive Selection	Apr 30 2009	Richard Cruse	GPL6246	GSE15907

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM399397	DPbl	wt	T.DPbl.Th#1	T.DPbl.Th#1, Double-Positive, Blasts	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399398	DPbl	wt	T.DPbl.Th#2	T.DPbl.Th#2, Double-Positive, Blasts	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399399	DPbl	wt	T.DPbl.Th#3	T.DPbl.Th#3, Double-Positive, Blasts	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399400	DPsm	wt	T.DPsm.Th#1	T.DPsm.Th#1, Double-Positive, Small Resting	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399401	DPsm	wt	T.DPsm.Th#2	T.DPsm.Th#2, Double-Positive, Small Resting	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399402	DPsm	wt	T.DPsm.Th#3	T.DPsm.Th#3, Double-Positive, Small Resting	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399403	ISP	wt	T.ISP.Th#1	T.ISP.Th#1, Immature Single- Positive	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399404	ISP	wt	T.ISP.Th#2	T.ISP.Th#2, Immature Single- Positive	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM399405	ISP	wt	T.ISP.Th#3	T.ISP.Th#3, Immature Single- Positive	Apr 30 2009	Richard Cruse	GPL6246	GSE15907
GSM594227	DN1	wt	Adult ETP biological rep. 1	Gene expression data from the most immature stage of T- cell differentiation in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594228	DN1	wt	Adult ETP biological rep. 2	Gene expression data from the most immature stage of T- cell differentiation in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM594229	DN1	wt	Adult ETP biological rep. 3	Gene expression data from the most immature stage of T-cell differentiation in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594230	DN2	wt	Adult DN2 biological rep. 1	Gene expression data from an intermediate progenitor stage during T-cell differentiation in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594231	DN2	wt	Adult DN2 biological rep. 2	Gene expression data from an intermediate progenitor stage during T-cell differentiation in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594232	DN2	wt	Adult DN2 biological rep. 3	Gene expression data from an intermediate progenitor stage during T-cell differentiation in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594233	DN3	wt	Adult DN3 biological rep. 1	Gene expression data from irreversibly committed T-cell progenitors in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594234	DN3	wt	Adult DN3 biological rep. 2	Gene expression data from irreversibly committed T-cell progenitors in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM594235	DN3	wt	Adult DN3 biological rep. 3	Gene expression data from irreversibly committed T-cell progenitors in the adult thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594236	DN1	wt	Fetal ETP biological rep. 1	Gene expression data from the most immature stage of T-cell differentiation in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594237	DN1	wt	Fetal ETP biological rep. 2	Gene expression data from the most immature stage of T-cell differentiation in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594238	DN1	wt	Fetal ETP biological rep. 3	Gene expression data from the most immature stage of T-cell differentiation in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594239	DN2	wt	Fetal DN2 biological rep. 1	Gene expression data from an intermediate progenitor stage during T-cell differentiation in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594240	DN2	wt	Fetal DN2 biological rep. 2	Gene expression data from an intermediate progenitor stage during T-cell differentiation in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM594241	DN2	wt	Fetal DN2 biological rep. 3	Gene expression data from an intermediate progenitor stage during T-cell differentiation in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594242	DN3	wt	Fetal DN3 biological rep. 1	Gene expression data from irreversibly committed T-cell progenitors in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594243	DN3	wt	Fetal DN3 biological rep. 2	Gene expression data from irreversibly committed T-cell progenitors in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM594244	DN3	wt	Fetal DN3 biological rep. 3	Gene expression data from irreversibly committed T-cell progenitors in the fetal thymus.	Sep 15 2010	Nikolai Nikolaevich Belyaev	GPL8321	GSE24142
GSM700782	DN3	Miz1 KO	DN3 KO5	Miz1 knockout	Apr 02 2011	Lothar Vassen	GPL1261	GSE28342
GSM700783	DN3	wt	DN3 WT2	wt	Apr 02 2011	Lothar Vassen	GPL1261	GSE28342
GSM769775	DN	wt	DN thymocytes, rep1		Aug 01 2011	Takeshi Egawa	GPL1261	GSE31082
GSM769776	DN	wt	DN thymocytes, rep2		Aug 01 2011	Takeshi Egawa	GPL1261	GSE31082
GSM769777	DN	wt	DN thymocytes, rep3		Aug 01 2011	Takeshi Egawa	GPL1261	GSE31082
GSM769778	DP	wt	DP thymocytes, rep1		Aug 01 2011	Takeshi Egawa	GPL1261	GSE31082
GSM769779	DP	wt	DP thymocytes, rep2		Aug 01 2011	Takeshi Egawa	GPL1261	GSE31082

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM769780	DP	wt	DP thymocytes, rep3		Aug 01 2011	Takeshi Egawa	GPL1261	GSE31082
GSM791134	DN2-3	wt	preT.DN2-3.Th#2	preT.DN2-3.Th#2, DN2-DN3 transitional thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791135	DN2-3	wt	preT.DN2-3.Th#3	preT.DN2-3.Th#3, DN2-DN3 transitional thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791136	DN2	wt	preT.DN2.Th#4	preT.DN2.Th#4, DN2 thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791137	DN2	wt	preT.DN2.Th#5	preT.DN2.Th#5, DN2 thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791138	DN2	wt	preT.DN2.Th#6	preT.DN2.Th#6, DN2 thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791139	DN2A	wt	preT.DN2A.Th#1	preT.DN2A.Th#1, DN2a thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791140	DN2A	wt	preT.DN2A.Th#2	preT.DN2A.Th#2, DN2a thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791141	DN2B	wt	preT.DN2B.Th#1	preT.DN2B.Th#1, DN2b thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791142	DN2B	wt	preT.DN2B.Th#2	preT.DN2B.Th#2, DN2b thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791143	DN3-4	wt	preT.DN3-4.Th#1	preT.DN3-4.Th#1, DN3-DN4 transitional thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791144	DN3-4	wt	preT.DN3-4.Th#2	preT.DN3-4.Th#2, DN3-DN4 transitional thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791145	DN3-4	wt	preT.DN3-4.Th#3	preT.DN3-4.Th#3, DN3-DN4 transitional thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791146	DN3A	wt	preT.DN3A.Th#1	preT.DN3A.Th#1, DN3a thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM791147	DN3A	wt	preT.DN3A.Th#2	preT.DN3A.Th#2, DN3a thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791148	DN3A	wt	preT.DN3A.Th#3	preT.DN3A.Th#3, DN3a thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791149	DN3B	wt	preT.DN3B.Th#1	preT.DN3B.Th#1, DN3b thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791150	DN3B	wt	preT.DN3B.Th#2	preT.DN3B.Th#2, DN3b thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791151	DN3B	wt	preT.DN3B.Th#3	preT.DN3B.Th#3, DN3b thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791152	DN1-2	wt	preT.ETP-2A.Th#3	preT.ETP-2A.Th#3, DN1-DN2 transitional population	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791153	DN1-2	wt	preT.ETP-2A.Th#4	preT.ETP-2A.Th#4, DN1-DN2 transitional population	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791154	DN4	wt	T.DN4.Th#4	T.DN4.Th#4, DN4 thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791155	DN4	wt	T.DN4.Th#5	T.DN4.Th#5, DN4 thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM791156	DN4	wt	T.DN4.Th#6	T.DN4.Th#6, DN4 thymocytes	Sep 06 2011	Richard Cruse	GPL6246	GSE15907
GSM800500	DP	wt	DP wt thymocyte, biological rep1	Gene expression data from mouse DP stage thymocytes.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM800501	DP	wt	DP wt thymocyte, biological rep2	Gene expression data from mouse DP stage thymocytes.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM800502	DP	wt	DP wt thymocyte, biological rep3	Gene expression data from mouse DP stage thymocytes.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM800503	DP	Ikaros KO	Ikaros KO thymocyte stage 1, biological rep1	Gene expression data from mouse DP stage thymocytes.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM800504	DP	Ikaros KO	Ikaros KO thymocyte stage 1, biological rep2	Gene expression data from mouse DP stage thymocytes.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM800505	DP	Ikaros KO	Ikaros KO thymocyte stage 1, biological rep3	Gene expression data from mouse DP stage thymocytes.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM800506	DP	Ikaros dominant negative, pre-leukemic	Ikaros dominant negative thymocyte stage 2 (16 d), biological rep1	Gene expression data from mouse DP stage thymocytes with Ikaros inactivation by dominant negative Ik.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM800507	DP	Ikaros dominant negative, pre-leukemic	Ikaros dominant negative thymocyte stage 2 (16 d), biological rep2	Gene expression data from mouse DP stage thymocytes with Ikaros inactivation by dominant negative Ik.	Sep 22 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM802973	DP	Ikaros dominant negative, pre-leukemic	Ikaros dominant negative thymocyte stage 2 (27 d), biological rep1	Gene expression data from mouse DP stage thymocytes with Ikaros inactivation by dominant negative Ik.	Sep 28 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM802974	DP	Ikaros dominant negative, pre-leukemic	Ikaros dominant negative thymocyte stage 2 (27 d), biological rep2	Gene expression data from mouse DP stage thymocytes with Ikaros inactivation by dominant negative Ik.	Sep 28 2011	Jiangwen Zhang	GPL1261	GSE32311
GSM802975	DP	Ikaros dominant negative, leukemic	Ikaros dominant negative thymocyte stage 3, biological rep1	Gene expression data from mouse DP stage thymocytes with Ikaros inactivation by dominant negative Ik.	Sep 28 2011	Jiangwen Zhang	GPL1261	GSE32311

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM823502	DN3	wt	WT DN3 thymocytes, biological rep 1	Gene expression data from WT DN3 thymocytes	Oct 27 2011	Hai-Hui Xue	GPL6246	GSE33292
GSM823503	DN3	wt	WT DN3 thymocytes, biological rep 2	Gene expression data from WT DN3 thymocytes	Oct 27 2011	Hai-Hui Xue	GPL6246	GSE33292
GSM823504	DN3	wt	WT DN3 thymocytes, biological rep 3	Gene expression data from WT DN3 thymocytes	Oct 27 2011	Hai-Hui Xue	GPL6246	GSE33292
GSM823505	DN3	Tcf7 KO	TCF-1 KO DN3 thymocytes, biological rep 1	Gene expression data from TCF-1-deficient DN3 thymocytes	Oct 27 2011	Hai-Hui Xue	GPL6246	GSE33292
GSM823506	DN3	Tcf7 KO	TCF-1 KO DN3 thymocytes, biological rep 2	Gene expression data from TCF-1-deficient DN3 thymocytes	Oct 27 2011	Hai-Hui Xue	GPL6246	GSE33292
GSM823507	DN3	Tcf7 KO	TCF-1 KO DN3 thymocytes, biological rep 3	Gene expression data from TCF-1-deficient DN3 thymocytes	Oct 27 2011	Hai-Hui Xue	GPL6246	GSE33292
GSM829333	DN1	Tcf7 KO	Tcf7 ^{-/-} ETP thymocyte, biological rep1	Gene expression data from mouse ETP stage thymocytes with ablation of Tcf7.	Nov 07 2011	Jiangwen Zhang	GPL1261	GSE33513
GSM829334	DN1	Tcf7 KO	Tcf7 ^{-/-} ETP thymocyte, biological rep2	Gene expression data from mouse ETP stage thymocytes with ablation of Tcf7.	Nov 07 2011	Jiangwen Zhang	GPL1261	GSE33513
GSM829335	DN1	Tcf7 KO	Tcf7 ^{-/-} ETP thymocyte, biological rep3	Gene expression data from mouse ETP stage thymocytes with ablation of Tcf7.	Nov 07 2011	Jiangwen Zhang	GPL1261	GSE33513

ID	Stage	Perturbation	Title	Description	Date	Submitter	Platform	Series
GSM829336	DN1	wt	Tcf7-/+ ETP thymocyte, biological rep1	Gene expression data from mouse ETP stage Tcf7-/+ thymocytes.	Nov 07 2011	Jiangwen Zhang	GPL1261	GSE33513
GSM829337	DN1	wt	Tcf7-/+ ETP thymocyte, biological rep2	Gene expression data from mouse ETP stage Tcf7-/+ thymocytes.	Nov 07 2011	Jiangwen Zhang	GPL1261	GSE33513
GSM829338	DN1	wt	Tcf7-/+ ETP thymocyte, biological rep3	Gene expression data from mouse ETP stage Tcf7-/+ thymocytes.	Nov 07 2011	Jiangwen Zhang	GPL1261	GSE33513
GSM854335	DN1	wt	preT.ETP.Th#6	preT.ETP.Th#6, Early T Lineage Progenitor	Dec 27 2011	Richard Cruse	GPL6246	GSE15907
GSM854336	DN1	wt	preT.ETP.Th#7	preT.ETP.Th#7, Early T Lineage Progenitor	Dec 27 2011	Richard Cruse	GPL6246	GSE15907
GSM854337	DN1	wt	preT.ETP.Th#8	preT.ETP.Th#8, Early T Lineage Progenitor	Dec 27 2011	Richard Cruse	GPL6246	GSE15907

Supplementary Table 3: Primer sequences used for qRT-PCR

Species	Target	Forward Primer	Reverse Primer
Mouse	Aiolos	CCGAGATGGGAAGTGAGAGA	CGCTTCTCACCGATGAATTT
Mouse	Bactin	GAAATCGTGCGTGACATCAAAG	TGTAGTTTCATGGATGCCACAG
Mouse	Bcl11a	TGGTATCCCTTCAGGACTAGGT	TCCAAGTGATGTCTCGGTGGT
Mouse	Bcl11b	GGGCGATGCCAGAATAGAT	GGTAGCCTCCACATGGTCAG
Mouse	Cd3e	ATGCGGTGGAACACTTTCTGG	GCACGTCAACTCTACACTGGT
Mouse	Cd3g	TGGAGAAGCAAAGAGACTGACA	GCCATCCACTTGTACCAAATTC
Mouse	Cd4	AGGTGATGGGACCTACCTCTC	GGGGCCACCACTTGAACACTAC
Mouse	Cd44	TCTGCCATCTAGACTAAGAGC	GTCTGGGTATTGAAAGGTGTAGC
Mouse	Cd8a	AAGAAAATGGACGCCGAACCTT	AAGCCATATAGACAACGAAGGTG
Mouse	Cebpa	CGGTCATTGTCACTGGTCAACT	GGACAAGAACAGCAACGAGTACC
Mouse	Deltex	GAGGATGTGGTTCGGAGGTA	CCCTCATAGCCAGATGCTGT
Mouse	E2a	TTTGACCCTAGCCGGACATAC	GCATAGGCATTCCGCTCAC
Mouse	Ebfl	TCTACAGCAATGGGATACGGA	GTGTGTGAGCAATACTCGGCA
Mouse	Eomes	GGCCCCTATGGCTCAAATTCC	GAACCACTTCCACGAAAACATTG
Mouse	Ets1	AAAAGTGGATCTCGAGCTTTTCC	CTTTCAAGGCTTGGGACATCA
Mouse	Gapdh	ACTCCACTCACGGCAAATTCA	GCCTCACCCCATTTGATGTT
Mouse	Gata2	ACCACAAGATGAATGGACAGAA	GTCGTCTGACAATTTGCACAAC
Mouse	Gata3	CTCGGCCATTTCGTACATGGAA	GGATACCTCTGCACCGTAGC
Mouse	Gfi1	AGAAGGCGCACAGCTATCAC	GGCTCCATTTTCGACTCGC
Mouse	Gfi1b	CTTCCACCAGAAGTCGGACAT	GAGATTGTGTTGACTCTCACGG
Mouse	HEB	GAGAAGAAGACCGCTCCATGAT	TGGCTTGGGAGATGGGTAAC
Mouse	HEBalt	GTGCTTATCCTGTCCCTGGAATG	TGGCTTGGGAGATGGGTAAC
Mouse	Hes1	TCAACACGACACCGGACAAAC	ATGCCGGGAGCTATCTTTCTT
Mouse	Hhex	CGGACGGTGAACGACTACAC	CGTTGGAGAACCTCACTTGAC
Mouse	Id2	CGACCCGATGAGTCTGCTCTA	GACGATAGTGGGATGCGAGTC
Mouse	Id3	CTGTTCGGAACGTAGCCTGG	GTGGTTCATGTCGTCCAAGAG
Mouse	Ikaros	TCCCAAGTTTCAGGAAAGGA	TCTGCTGTGCTCCAGAGGTA
Mouse	Il2ra	CACTACGAGTGTATTCCGGGA	TCGGTGGTGTCTCTTTTCATCT
Mouse	Il7ra	AGTCCGATCCATTCCCCATAA	ATTCTTGGGTTCTGGAGTTTCG
Mouse	Irf8	AGACGAGGTTACGCTGTGC	TCGGGGACAATTCCGTAAACT
Mouse	Kit	GGCCTCACGAGTTCTATTTACG	GGGGAGAGATTTCCCATCACAC
Mouse	Lat	TTTCTACCCTCTAGTCACTTCC	CCACATTCTTACAGGCTGGCT
Mouse	Lck	TGGAGAACATTGACGTGTGTG	ATCCCTCATAGGTGACCAGTG
Mouse	Lef1	ACCTACAGCGACGAGCACTT	GGGTAGAAGGTGGGGATTTTC
Mouse	Lmo2	GACGATGCGGGTGAAAGACAA	TCACACACTATGTCGGAGTTGA
Mouse	Lyl1	AAAACCTGAGATGGTATGTGCCTC	TGTCCAGGTTTATCACTGGC
Mouse	Myb	GAGCAGAAGAAGTTTCCCGATT	AGCGGGAATCGGATGAATCT
Mouse	Notch1	CCCTTGCTCTGCCTAACGC	GGAGTCCTGGCATCGTTGG
Mouse	Pax5	ACAGCATAGTGTCTACAGGCT	CCCTCTTGC GTTTGTTGGTG
Mouse	Ptcra	GGTGTCAGGCTCTACCATCAG	TGCCTTCATCTACCAGCAGT

Species	Target	Forward Primer	Reverse Primer
Mouse	Pu.1	ATGTTACAGGCGTGCAAATGG	TGATCGCTATGGCTTTCTCCA
Mouse	Rag1	ACCCGATGAAATTCAACACCC	CTGGAAGTACTGGAGACTGTTCT
Mouse	Rplp1	CTCGCTTGCATCTACTCCGC	AGAAAGGTTTCGACGCTGACAC
Mouse	Runx1	GCAGGCAACGATGAAAATACT	GCAACTTGTGGCGGATTTGTA
Mouse	Scl	CTCACTAGGCAGTGGGTTCTTT	GGACCATCAGAAATCTCCATCT
Mouse	Tbx21	AACCGCTTATATGTCCACCCA	CTTGTGTTGGTGAGCTTTAGC
Mouse	Tcf7	AGCTTTCTCCACTCTACGAACA	AATCCAGAGAGATCGGGGGTC
Mouse	Zap70	CTACGTGCTGTCGTTGGTG	GTTACACGGCTTACGCAGGT
Human	Bactin	CATGTACGTTGCTATCCAGGC	CTCCTTAATGTCACGCACGAT
Human	Bcl11b	TCCAGCTACATTTGCACAACA	GCTCCAGGTAGATGCGGAAG
Human	Cebpa	TATAGGCTGGGCTTCCCCTT	AGCTTTCTGGTGTGACTCGG
Human	Deltex	ATCGGAGAAGGCTCTACAGG	CGTCTGGCCTCCTTTCTAACT
Human	E2a	CCGACTCCTACAGTGGGCTA	CGCTGACGTGTTCTCCTCG
Human	Gata3	GTTGGCCTAAGGTGGTTGTG	ACAGGCTGCAGGAATAGGGA
Human	Hes1	CCTGTCATCCCCGTCTACAC	CACATGGAGTCCGCCGTAA
Human	Notch1	GAGGCGTGGCAGACTATGC	CTTGTACTCCGTCAGCGTGA
Human	Pu.1	TGCAATGTCAAGGGAGGGGG	AAACCCTTCCATTTTGCACGC
Human	Tcf7	TGCACATGCAGCTATACCCAG	TGGTGGATTCTTGGTGCTTTTC