Single-Cell Analysis of Normal and Perturbed Early T-Cell Developmental Processes

Thesis by

Wen Zhou

In Partial Fulfillment of the Requirements for

the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2021

(Defended Jan 19th 2021)

To my parents, Lihua Wang and Mao Zhou, for their love and belief in me, and to my grandmother, Hua Xiang, for inspiring me to do science growing up.

To Chen Xu, my husband and the love of my life, and my dog Goli, who walked this journey with me.

Wen Zhou
ORCID: 0000-0003-0357-2744

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Ellen Rothenberg. Your guidance, generous encouragement, kindest care, and unreserved support has been a cornerstone of my PhD studies. Your unwavering enthusiasm, unrelenting dedication, and breadth of knowledge has been an inspiration to me, stimulating me to always think deeper and reach further. You exemplify the highest standard of scientific work, and have taught me to strive for the same. I want to thank you for all the ever-welcoming long discussions on projects, ideas, and science, that we had throughout the years. These discussions nurtured me as a young scientist and person, something indispensable during my PhD career, and I truly believe I will continue to remember and value them in the years to come.

I would like to thank Dr. Long Cai, for the 2.5 years of experience as a research technician in your lab prior to my graduate school. You offered a job to someone fresh out of undergraduate program from across the continent, and brought out the potentials that you saw in me that I would have never recognized. You opened my door and vision to single cell biology, engaged me with cool projects, and inspired me scientifically throughout the time I spent in your lab. While being the nicest supervisor I could ever ask for, you encouraged me to go to graduate school when I was hesitant. I would not be who I am as a scientist today without you.

For the past 5 years, Rothenberg lab felt like a second family to me, which is inseparable from all the mentorship and friendship I received from the members of the Rothenberg group. I would like to thank Dr. Mary Yui, for teaching me, mentoring me, and caring for me scientifically and personally. Mary was largely the reason I joined the Rothenberg lab to begin with, as she was truly the kindest mentor in the world, someone I could talk to about anything, and she always gave me the wisest advice. I want to thank our lab manager Rochelle (Shelley) Diamond, for taking care of the lab, and for being the inspirational figure she is. It was a true pleasure getting to know her. I would also like to thank Maile Romero-Wolf, Xun Wang, Maria Lerica Quiloan, Dr. Boyoung Shin, and Dr. Tom Sidwell for sharing the space, for being amazing colleagues, and for their friendship. I want to especially thank Boyoung for proofreading of my thesis draft and providing feedbacks, and for being a friend I could count on for advice and support, for which I am truly grateful.

I would also like to thank my thesis committee, Dr. Barbara Wold, Dr. Marianne Bronner, and Dr. Long Cai, for the time you spent on me, and the insightful suggestions on my projects over the past many years.

I would also like to acknowledge my collaborations with many great scientists at Caltech: Dr. Brian Williams from Wold lab for teaching me to do my very first scRNA-seq experiment back in 2015, and being extremely kind and supportive throughout my time spent here. Jina Yun from Cai lab for help and friendship. Dr. Fan Gao, director of the bioinformatic resource center, for the help and expertise on bioinformatic pipelines. Jeff Park from the Caltech Single Cell Profiling and Engineering Center for his help and advice. Rochelle (Shelley) Diamond, Diana Perez, Jamie Tijerina, and Patrick Canon from the flow cytometry facility for their advice and expertise on flow cytometry. Dr. Igor Antoshechkin from the Caltech Jacobs Genomics Facility for help on sequencing, but also for his greetings and smiling every time we pass by each other.

I want to thank my close friends: Yusi, Shihan, Haojiang, Hank, and Eileen – You have made my life here at Caltech pleasantly colorful.

Finally, I want to give my sincerest gratitude to my family. My parents: Mao Zhou and Lihua Wang, for their support and belief in me. My grandmother: Hua Xiang, who brought me up during my childhood, and took me to her microbiology lab to watch her do experiments every day, inspiring me to work in biology since I was 4 years old. To my husband, Chen Xu: thank you for your love and support over the past 13 years, and for flying back from Northern California every single weekend to spend time with me for the past few years. We have been fortunate to spend the last couple months together due to the pandemic, but we shall finally truly 'reunite' after I finish my PhD. And to my dog, Goli, thank you for all the love you gave me, tolerating all my late days of work, and still being the sweetest dog.

ABSTRACT

Early T-cell development converts multipotent precursors to committed pro-T cells, silencing progenitor genes while inducing T-cell genes. However, both the underlying steps of developmental progression and the regulations involved have remained obscure. Although some of the expressions of important regulators in early T-cell development have been studied in bulk populations, the nature of heterogeneity in this constantly refreshed developmental continuum makes it difficult to understand the developmental trajectories that the cells have undergone using bulk analysis, both in natural conditions and under gene perturbations.

Combining droplet-based single cell RNA sequencing (scRNA-seq), deep-sequenced whole-transcript scRNA-seq, and seqFISH for key regulatory genes, we established regulatory phenotypes of sequential ETP subsets; confirmed initial co-expression of progenitor- with T-cell specification genes; defined stage-specific relationships between cell-cycle and differentiation; and generated a pseudotime model from ETP to T-lineage commitment, supported by RNA velocity and transcription factor perturbations. This model was validated by developmental kinetics of ETP subsets at population and clonal levels. The results imply that multilineage priming is integral to T-cell specification in natural developing pro-T cells in the thymus.

Moreover, we examined the functional implications of some of the transcription factors (TFs) through bone marrow (BM) derived *ex-vivo* differentiation systems. Using scRNA-seq, Cell Hashing, and a pool-based CRISPR/Cas9 perturbation system, we established the normal and perturbed developmental trajectories before and after the T-lineage commitment stages. Our analysis revealed that, without the essential lineage commitment TF, Bcl11b, the developing early T cells immediately realized the lack of the essential regulator around the proliferating late DN2a stage. But instead of pushing the developmental path backwards to resemble the earlier stage of uncommitted cells, cells lacking *Bcl11b* underwent a diverging route of accumulation of ‘non-T’ genes that are not naturally expressed in earlier stages, potentially leading to the eventual loss of Notch responses. Our results also revealed the

complex regulations by TFs that set up the earliest T-lineage progression and commitment conditions. The SCENIC analysis suggested that *Gata3* and *Tcf7*, despite both being important regulatory factors for T-lineage progression, have very different regulatory roles in controlling proliferation and suppressing myeloid lineages. Furthermore, pseudotime analysis also showed that some of the stem and progenitor genes and 'multilineage' associated genes expressed by early pro-T cells potentially hold back the T-lineage differentiation speed. In summary, our study leveraged both *in vivo* thymic pro-T cells' developmental trajectory obtained through single-cell analysis and *ex-vivo* derived T cells for internal-controlled perturbations, and revealed some profound roles of TFs in regulating early T-cell differentiation processes.

PUBLISHED CONTENT AND CONTRIBUTIONS

Chapter 2 is adapted from the published content:

[1] **Zhou, W.**, Yui, M.A., Williams, B.A., Yun, J., Wold, B.J., Cai, L., and Rothenberg, E.V. (2019). Single-cell analysis reveals regulatory gene expression dynamics leading to lineage commitment in early T-cell development. Cell Systems 9, 321-337.e9. DOI: 10.1016/j.cels.2019.09.008

W.Z. participated in the design of the project, performed most of the experiments and all the data analysis, and wrote the paper.

Chapter 3 is adapted from the manuscript being prepared for publication:

[2] **Zhou, W.**, Gao, F., Romero-Wolf, M., Jo, S., and Rothenberg, E.V.. Single-cell analysis of the transcription factor controlled early T-cell differentiation dynamics and trajectory topology. In preparation.

W.Z. designed the project, performed all of the experiments and data analysis in the current version of the manuscript, and wrote the paper.

Other published content mentioned in or related to this thesis:

[3] *Shah, S., *Takei, Y., ***Zhou, W.**, Lubeck, E., Yun, J., Eng, C.-H.L., Koulena, N., Cronin, C., Karp, C., Liaw, E.J., et al. (2018). Dynamics and spatial genomics of the nascent transcriptome by Intron seqFISH. Cell. 174, 363-376.e16 DOI: 10.1016/j.cell.2018.05.035.

* denotes equal contributions

W.Z. initiated the project, participated in the design of the experiments, performed the experiments, and participated in writing the paper.
I started this project in the Cai lab before graduate school and continued to work on this project through the first 2 years of my PhD program. Although not being extensively discussed in this thesis, this project has a very special place in my heart.

[4] Shah, S., Lubeck, E., **Zhou, W.**, and Cai, L. (2016a). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron 92, 342–357. DOI: 10.1016/j.neuron.2016.10.001

W.Z. performed experiments and edited the paper.

[5] **Zhou, W.**, Rothenberg, EV. (2019). Building a human thymus: A pointillist view, Immunity 51 (5), 788-790

W.Z. and E.V.R. wrote this preview paper together.

[6] Romero-Wolf,M., Shin,B., **Zhou,W.**, Koizumi,M., Rothenberg, E.V., Hosokawa, H. (2020). Notch2 complements Notch1 to mediate inductive signaling that initiates early T cell development. J. Cell Biol. 219, e202005093. DOI: 10.1083/jcb.202005093

W.Z. analyzed the data and edited the paper.

[7] Olariu, V., Yui, M. A., Krupinski, P., **Zhou, W.**, Deichmann, J., Andersson, E., Rothenberg, E. V., and Peterson, C. (2021). Multi-scale dynamical modelling of T-cell development from an early thymic progenitor state to lineage commitment. Cell Reports, in press. DOI: 10.1016/j.celrep.2020.108622.

W.Z. performed and analyzed the FISH experiment, and edited the paper.

[8] Shin, B., Hosokawa, H., Romero-Wolf, M., **Zhou, W.**, Masuhara, K., Tobin, V. R., Levanon, D., Groner, Y., Rothenberg, E. V. (2021). Runx1 and Runx3 drive progenitor to T-lineage transcriptome conversion in mouse T-cell commitment via dynamic genomic site switching. Proc. Natl. Acad. Sci. USA, in press. DOI: 10.1073/pnas.2019655118.

W.Z. performed the experiments and analysis, and edited the paper.

TABLE OF CONTENTS

Acknowledgements……..…..iii
Abstract ……...…..v
Published Content and Contributions…….....vii
Table of Contents…….……..ix

Chapter I: Introduction 1
The Classic Understandings of Hematopoiesis 1
The Current View of Where the 'T-Lineage' Resides on the Hematopoiesis Map..... 4
Essential TFs in the Play of T-lineage Establishment—TCF1, GATA3, and Bcl11b..... 6
The Revolution with Single-Cell Tools..... 7
Single-Cell Technical and Analytical Challenges of the Developmental Continuum and the Regulators Involved 8
A Deeper Dive Using Single-Cell Analysis 11
Bibliography 15

Chapter II: Single-Cell Analysis Reveals Regulatory Gene Expression Dynamics Leading to Lineage Commitment In Early T-Cell Development..... 21
Summary 22
Introduction 23
Results 26
Discussion 42
Main Figures and Legends..... 45
Supplementary Figures and Legends..... 58
Methods..... 72
Supplementary Table Titles and Legends..... 90

Key Resource Table 92
Bibliography 96

Chapter III: Single-Cell Analysis of the Transcription Factor Controlled Early T-Cell Differentiation Dynamics and Trajectory Topology 103
Abstract 104
Introduction 105
Results 108
Discussion 123
Main Figures and Legends 127
Supplementary Figures and Legends 138
Methods 152
Bibliography 163

Chapter IV: Opportunities, Challenges, and Perspectives 167
Does Deeper Sequencing Solve More Problems? 167
How Does a New Dataset Align with the Previous Data? 171
Going Beyond Descriptive Single-Cell Analysis 173
Bibliography 176

Chapter 1

INTRODUCTION

In vertebrates, hematopoietic stem and progenitor cells generate an exceptional diversity of cell types throughout life, and this poses a series of challenges for explanation of developmental dynamics, developmental choice hierarchies, and their underlying mechanisms of regulations. T cells develop in a continuous flux well into adulthood (rev. in Rothenberg, 2019), and they are also of particular interest as one of the 'central players' in the adaptive immune system of mammals. Under the thymic signaling environment, lymphoid-primed multipotent precursors begin their differentiation 'journey' to cells that will irreversibly activate the transcriptional program that confers T-cell identity and excludes other lineage possibilities, this process is termed as 'early T-cell development'. Therefore, early T-cell development is a particularly accessible and functionally relevant system for studying the sequence of regulatory changes through which stem and progenitor cells resolve their multipotency to select a differentiation pathway.

The Classic Understandings of Hematopoiesis

Hematopoietic cells have traditionally been divided into erythroid/megakaryocytic (platelets and erythrocytes), myeloid (i.e. monocytes, macrophages, neutrophils, other granulocytes, mast cells, and dendritic cells), and lymphoid (T cells, B cells, NK cells, nonkiller ILCs) branches. For many years, T cells have been considered a subspecies of lymphoid fate and closely related to B cells, as shown in Figure 1 (Orkin and Zon, 2008). The author will revisit the topic about where exactly T cells should be positioned among the hematopoietic lineages, but it is important to acknowledge that many early lineage decisions in this hematopoiesis map shown in Figure 1 have been extensively studied and validated.

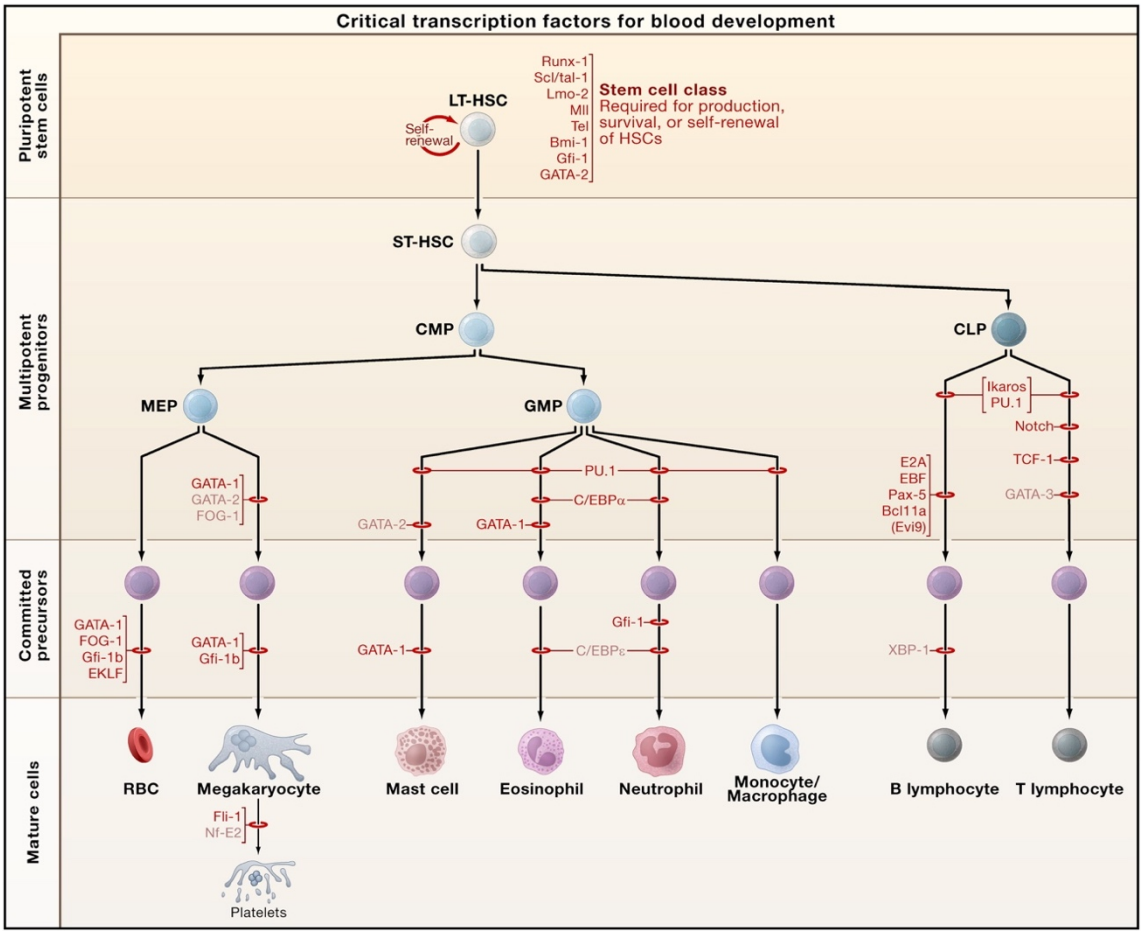


Figure 1. The Classic View of Hematopoiesis Hierarchy and the Regulators (adapted from Orkin and Zon, 2008). The arrows represent the hierarchical relationship between progenitors and different populations. Red bars indicate the stages where hematopoietic development is blocked in absence of a given TF, as determined by conventional gene knockouts. LT-HSC, long-term hematopoietic stem cell; ST-HSC, short-term hematopoietic stem cell; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP, granulocyte/macrophage progenitor; RBC, red blood cell.

On the very top of the hematopoietic hierarchy is the hematopoietic stem cell (HSC). HSCs are defined operationally by their well-known capacity to reconstitute the entire blood system of a recipient. As intrinsic determinants of cellular phenotype, transcription factors (TFs) provide an entry point for resolving how different lineages are related and how lineage-restricted differentiation is programmed (Orkin, 2000; Orkin and Zon, 2008).

HSCs express many non-HSC specific TFs that are shared with different lineages, which include Runx1 (runt-domain protein), Scl/tal1 (the basic helix-loop-helix (bHLH) factors), Lmo2 (LIM domain-containing protein), Mll (SET-domain containing histone methyltransferase), GATA-2 (zinc finger transcription factor), and several others. The lineage specification process involves fundamental changes of the cell's gene expression program regulated and 'coordinated' through essential lineage associated-TFs, some as summarized in Figure 1 (Orkin and Zon, 2008). It is important to note that these TFs' regulatory modules, partially due to the nature of their regulatory requirements, in addition to the involvement of cytokines and various signaling components, survival and 'self-renewal' is often intertwined with the TF-regulated differentiation process. And for reasons like this, hematopoietic cell fate is also intertwined with the origins of leukemias.

It is also important to point out that lineage restricted TFs are more or less limited to their own subtree. For instance, GATA-1 is highly expressed in megakaryocytic/erythroid progenitors (namely MEPs) that give rise to megakaryocyte and red blood cell precursors, whereas a "myeloid factor," such as PU.1 and C/EBPα, is present in GMPs. However, upper in the hierarchy, like HSCs or cells at other transient or stable multipotent stages, can co-express genes associated with multiple lineages, even within single cells, albeit generally at low levels – a phenomenon termed 'multilineage priming' (Hu et al., 1997; Miyamoto et al., 2002; Ng et al., 2009; Olsson et al., 2016; Orkin, 2003). Multilineage priming suggests that the fate of these immature cells is not 'sealed', and that lineage selection is likely a process in which alternative possibilities are eliminated. The coexistence of transcription factors representing different lineages within a common progenitor cell could also offer the potential for immediate "crosstalk" between different

fates at the molecular level, or to simply delay the differentiation choice to allow proper 'additional regulatory apparatus' to be established. In Chapter 2, the author will expand the discussion and significance of multilineage priming in the early T developmental system in depth.

The Current View of Where The 'T-Lineage' Resides on the Hematopoiesis Map

In regards of the T lymphocytes' position on the hematopoiesis map, as seen in Figure 1, T lymphocytes have been often grouped with B lymphocytes and both considered derived from common lymphoid progenitors (CLP). The classic view has been based on the unique antigen receptor generation strategy that T and B cells share, i.e. RAG-mediated recombination, same RAG1/2 enzymes and selection mechanisms, so it was predicted that early separation had occurred between precursors that could generate lymphocytes and precursors that could generate all other hematopoietic cell types. However, a finer picture of cell type identity and of lineage relationships has emerged in the past decade as increasing knowledge has been gained about the newly characterized cell populations, dynamics of lineage-specific developmental processes, and the transcriptional regulatory apparatus that drives them. First, some effector functions of T cells are extensively shared with the NK cells and ILCs, but not with B cells, which do not use the RAG-mediated recombination mechanisms at all (rev. in Rothenberg, 2019). Second, the cell fate decisions do not fall into a strict hierarchy: unlike classic C/EBP(myeloid) vs. GATA1(MEPs) hematopoiesis subtrees, the order in which lymphocyte fates subdivide as compared to the order in which they separate from macrophage, granulocyte, and dendritic cell fates is surprisingly dependent on their cell fate chosen (Rothenberg et al., 2016).

In addition, it was known that in the B-cell differentiation pathway, the myeloid potential is excluded early on, but not the T-lineage potential (Mansson et al., 2010; Welinder et al., 2011; Zandi et al., 2012). In contrast, in the T-differentiation pathway, the B-cell potential is excluded much early on, leaving plenty of residual myeloid potential to be suppressed in later stages both *in vivo* and *in vitro* (Allman et al., 2003; Bell and Bhandoola, 2008; Heinzel et al., 2007; Wada et al., 2008). Thus, it is clearly not a simple hierarchical

relationship between the lymphoid cell fate and myeloid cell fate, or between T cells and B cells, as illustrated in Figure 2 (adapted from Rothenberg, 2019).

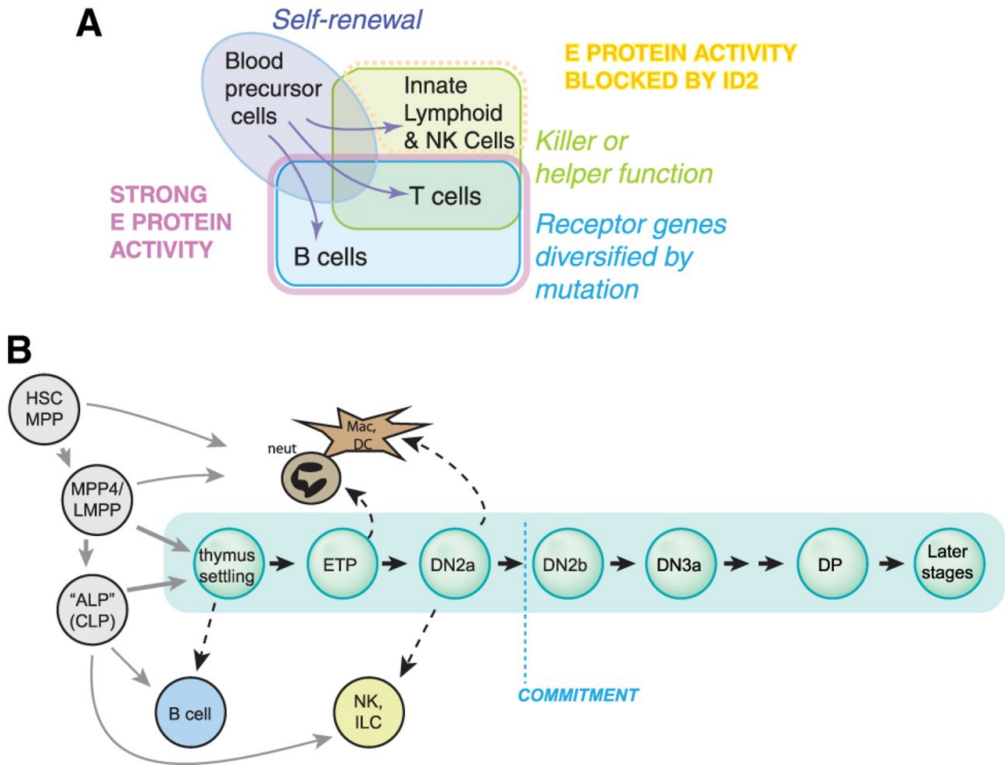


Figure 2. Relationships of the T-cell program to other hematopoietic fates (adapted from Rothenberg, 2019). (*A*) The diagram shows that T cells share a common mechanism for receptor gene diversification with B cells and share similar set of killer and helper functions with NK and ILCs. (*B*) Persistence of alternative lineage potentials in T-cell precursors after entry into the thymus. Dash arrows indicate the last developmental stages at which isolated T-cell precursors can still give rise to the indicated alternative fates, provided that they are removed from the thymic microenvironment. Note that access to the B-cell option is lost a few stages before access to NK and dendritic cell options, unlike the hierarchical structure shown in Figure 1. Mac, Macrophage; DC, dendritic cell; Neut, neutrophilic granulocyte; CLP, common lymphoid progenitor ("ALP" indicates a CLP that is not B-lineage-biased); LMPP, lymphoid-primed multipotent progenitor maintaining myeloid as well as lymphoid potential (similar to "MPP4"); MPP, multipotent precursor.

Early T-cell developmental stages have been intensely studied and are generally well distinguished by combinations of cell-surface markers, which broadly correlate with stereotyped gene expression changes on bulk levels (Yui and Rothenberg, 2014, also discussed in depth in Chapter 2). In mouse systems, the stage markers have been validated by *in vivo* and *in vitro* transfer experiments, differentiation assays under distinct environmental conditions, and targeted genetic perturbation studies. However, individual T-cell precursors in the same thymic cohort can have varied developmental potentials and can go on to divergent fates. What molecular mechanisms control these different developmental outcomes? Also, due to this non-hierarchical position of T-lineage with respect to the hematopoiesis map, and the lack of clear trajectory models for early T-cell development in the primary thymic environment, it remains unclear of how many types of thymic T-cell progenitors there actually are, and the exact steps that they undergo to initiate T-lineage commitment. Are all the precursor cells coming to the thymus capable of giving rise to T cells? What are the steps they need to go through to prepare for lineage commitment? And why do the precursors take a long time to make the T-cell fate choice?

Essential TFs in the Play of T-lineage Establishment—TCF1, GATA3, and Bcl11b

TCF1 (encoded by *Tcf7* gene) and GATA3 are indispensable TFs for early T-cell development, and their expressions are known to be induced in response to Notch signaling in the thymic environment. TCF1 or GATA3 KO result in losses in population size of T cells even in the earliest stage of T cells, e.g. ETP stage (Germar et al., 2011; Hattori et al., 1996; Hosoya et al., 2009; Scripture-Adams et al., 2014; Ting et al., 1996; Weber et al., 2011). TCF1 positively regulates *Gata3*, the DN2 stage marker *Il2ra*, and a commitment marker *Bcl11b*, as well as genes encoding signaling components in early DN cells and a vital TCR complex (Weber et al., 2011). Unlike many other required T-cell factors, an artificial high-level expression of TCF1 from an early stage can instruct T-lineage differentiation, accelerating many T-cell developmental genes' expression, even in pre-thymic precursors without concomitant Notch signaling (Weber et al., 2011). GATA3, similarly to TCF1, is needed for early T populations in fetal as well as adult mice (Hattori et al., 1996; Hosoya et al., 2009; Hozumi et al., 2008; Scripture-Adams et al., 2014; Ting

et al., 1996), although overexpression of GATA3, in contrast to TCF1, is not tolerated by pro-T cells (Taghon et al., 2007; Xu et al., 2013). Later in the DN2 stage, arguably the most critical process in early T-cell development occurs – commitment to T-cell fate, which coincide with the upregulation of the expression of TF, Bcl11b. Bcl11b was discovered in 2010 as a factor required for T-cell commitment by three groups in parallel (Ikawa et al., 2010; Li et al., 2010a, 2010b), and its regulatory functions were further studied with expression profiles with bulk RNA-seq and binding activities with ChIP-seq (Hosokawa et al., 2018; Longabaugh et al., 2017). *In vivo*, Bcl11b is required for the survival of the development of αβ T cells through β selection, and some γδ cells, despite not being strictly required for viability in the way like TCF1 and GATA3.

Moreover, although much previous effort on understanding the roles of important TFs has been performed with bulk RNA-seq and ChIP-seq assays, the kinetics of differentiation, population distributions, and trajectory topologies of the early perturbation outcomes are completely missing. Perturbations of regulatory genes can lead to the emergence of new minor populations of cell type or state, disappearance of some old cell populations, shifting in distributions on the differentiation trajectory, or alternations in gene expressions among the entire population studied. These effects need to be examined in a systematic and internally controlled way with consistent input and experimental setups, and they cannot be observed with bulk assays.

The fine single-cell expression profile of these three TFs together with other regulatory genes are going to be discussed throughout this thesis. And specifically, the perturbation outcomes of these three important TFs and a few more TFs are going to be examined in detail on the single-cell level in Chapter 3.

The Revolution with Single-Cell Tools

The classical understandings of lineage hierarchy and relationships in hematopoiesis have been built on the cell type definition system by cell surface markers analyzed through multicolored fluorescence-activated cell sorting (FACS) and combined with functional assays. However, as mentioned above, because these analyses were conducted on bulk

samples, they can neglect the heterogeneity in the defined population as well as unknown transitional states during the cell fate decision process. Over the past 5-6 years, the rapid development of single-cell tools, mainly single-cell RNA sequencing (scRNA-seq), provided unprecedented opportunities to re-define cell taxonomy, to impute or track differentiation hierarchy, and to uncover transcriptional networks at single-cell resolution for any given isolatable heterogeneous cell population, particularly in the hematopoietic system (Drissen et al., 2016; Giladi et al., 2018; Olsson et al., 2016; Paul et al., 2015).

One of the major advantages of this approach is the potential to bypass the need for a priori markers that define progenitor populations, and the sensitivity to detect rare or even transient transcriptional states *de novo,* given sufficient sample size. For example, recent scRNA-seq have re-defined the transcriptional states of myeloid subtypes and other stem and progenitor populations in the bone marrow (Drissen et al., 2016; Giladi et al., 2018; Nestorowa et al., 2016; Olsson et al., 2016; Paul et al., 2015; Schlitzer et al., 2015; See et al., 2017; Tusi et al., 2018), suggesting that the differentiation from HSCs is actually more complex and less sequential than the classical model, similarly to the non-hierarchical position of the T-cell 'branch' aforementioned. With these single-cell studies, it became more accepted that rather than a stepwise progression of HSCs following a tree-like hierarchy of oligo-, bi-, and unipotent progenitor paths, individual HSCs may gradually acquire lineage biases along multiple directions without necessarily passing through discrete hierarchically organized progenitor populations, forming a so-called 'developmental continuum' (Giladi et al., 2018; Velten et al., 2017). It is fair to conclude that single-cell methods over the past years have revolutionized our understanding of hematopoiesis and the definition of hematopoietic trajectories.

Single-Cell Technical and Analytical Challenges of the Developmental Continuum and the Regulators Involved

In single-cell analysis, there has always been a tradeoff between the number of features (i.e. dimensions) measured and the number of cells measured. The conceptual predecessors of single-cell transcriptome profiling are flow cytometry and mass cytometry, which are

typically restricted to very limited predefined markers, but they can easily profile millions of cells. Single cell RNA profiling techniques like scRNA-seq, in contrast, often do not require prior knowledge of predefined markers, and can measure up to 10^4 genes simultaneously in each cell. However, the platform and method of choice does heavily influence the sensitivity, drop-out rate, and technical noise of genes measured, as well as the throughput of the assays (Svensson et al., 2017). For example, high throughput methods (e.g. droplet-based methods like Drop-Seq, InDrop, and 10X Chromium platforms) often detect 1000-3000 expressed genes per cell (depending on the sequencing saturation), but can easily assay 10^4 cells per experiment (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017); whereas commonly used low-throughput methods (e.g. plate-based methods like Smart-Seq2, C1-Fluidigm system, CelSeq2) are used to profile a few hundred cells per sample, but can detect 5000 genes per cell (Hashimshony et al., 2016; Picelli et al., 2014). Another single-cell RNA profiling method with increasing popularity and can preserve the spatial information, which is independent of enzymatic preparation and subsequent sequencing steps, is single molecule fluorescent in-situ hybridization (smFISH) based quantification, such as seqFISH (Lubeck et al., 2014; Raj et al., 2008; Shah et al., 2016). In the review by Svensson et al. in 2017, the estimated CelSeq sensitivity of mRNA transcript (UMI) measurement was 5-10% compared to the 'gold-standard' smFISH. This review also discussed the sequencing depth needed to detect lowly expressed genes, which is essential for coverage of TFs (Svensson et al., 2017). The most up-to-date detection limit of droplet-based methods is up to about 30% detection efficiency (with 10X V3 Chemistry). In Chapter 2, the author also compared mRNA count measured by seqFISH and measured by scRNA-seq in the same samples, confirming that the detection rate in the 10X V2 method is roughly 10%, and therefore the author took advantage of seqFISH's sensitivity for quantification of important regulators such as TFs.

Why is sensitivity and tradeoff a relevant topic for studying hematopoiesis, developmental continuum, or early T-cell development specifically? And why is developmental continuum analysis particularly challenging? First, developmental systems usually exhibit a fast turnover: as scRNA-seq offers a snapshot of cells and their expression states that we

use to infer the relationship between them, it is important to sample enough cells of interest and capture the genes that can differentiate these states. Second, because of the 'snapshot', the temporal information is missing, therefore we rely on single-cell expression profiles in high-dimensional spaces to generate the transcriptional landscape which encodes information on developmental stage transitions, and enables the ordering of cells along pseudotime, from immature progenitors towards more differentiated states (Bendall et al., 2014; Qiu et al., 2017; Trapnell et al., 2014). This is conceptually exciting but technically difficult, especially for a tightly connected developmental continuum, in a completely unsupervised manner. Questions regarding developmental continuum are fundamentally more challenging than the 'cell type classification problems' in other classic scRNA-seq settings, which can be dealt with clustering and marker identifications, because the genes that differentiate developmentally relevant states are often lowly expressed. In contrast to developmentally relevant genes like TFs, the readily detectable variable and highly expressed genes like cell cycle associated genes can often be confounding factors when one is trying to infer trajectory and pseudotime in reduced dimensional spaces (e.g. PCA, tSNE, UMAP).

It is also important to note that there are assumptions in using single-cell methods to compute developmental trajectories and pseudotime, which should be considered preferably at the design stage of the study. These assumptions include: 1) Coverage of precursor, mature cells, and transitional stages along the differentiation process. If harvested from primary animals, it assumes differentiation happens asynchronously and is a continuous process. Detection of 'jumps' between cell states is difficult. 2) The cells' movement is unidirectional, and additional knowledge is needed to determine the start and finish of the trajectory. 3) Cell state information is complete and accurately represented in the low dimensional spaces. This step may require a fine feature selection or feature 'engineering' step to avoid segregation or dominant spread due to unwanted features, such as cell cycle. 4) Many analytical methods require additional assumptions, such as a tree-like structure of the data, where cells undergo potential bifurcations during differentiation, or absence of oscillations between cell states such as cell cycle, which clearly can be

problematic. In 2018, La Manno et. al. came up with an RNA velocity analysis in scRNA-seq, using the ratio between intron-mapped reads and exon-mapped reads to infer the time derivative of expression states in a static low dimensional representation (La Manno et al., 2018). This provided an additional tool to investigate the potential precursor-product relationship on a trajectory, but the parameters, especially for imputing the sparse intron-mapped read matrix, still need to be closely attended for different datasets.

In later chapters, our studies also calculated trajectory and pseudotime inferences, and in Chapter 2, our study, for the first time in the field, validated the significance of pseudotime by the *ex-vivo* culture of FACS sorted population according to pseudotime, examined both the T-lineage developmental speed and alternative lineage potentials, and mapped the sorted populations' expression profile back onto the pseudotime trajectory.

A Deeper Dive Using Single-Cell Analysis

Various advanced methods that were built upon scRNA-seq opened up more opportunities for further deep dives into understanding mechanisms using single-cell analysis. Cell Hashing enabled pooling of multiple samples into one experiment (Stoeckius et al., 2018), which can hugely improve the experimental design by not only incorporating biological replicates in the same scRNA-seq reactions, but also avoiding the potential confounder effect issues by enabling having both experimental and control samples in the same reaction. Computationally, alignment methods for batch and multi-modal integrations, such as canonical correlation analysis (CCA (Butler et al., 2018) or MultiCCA for more than 2 samples (Stuart et al., 2019)), mutual nearest-neighbor (MNN) correction (Haghverdi et al., 2018), nonnegative matrix factorization (NMF) (Yang and Michailidis, 2015), Harmony (Korsunsky et al., 2019), allowed cross-validation between methods, multi-modal analysis, comparison between organisms, and multiple experimental batch integrations. A good understanding of these technological advancements and their associated assumptions should guide a proper experimental design and usage of single-cell analysis to maximize the yield of insights to the scientific problems of interest. For example, CCA identifies shared aspects of variation between paired datasets, and

multiCCA does integration by iteratively applying CCA; MNN builds MNN graphs between cells from different datasets, where two cells are connected in the graph if they are transcriptionally similar. Most of these methods have the underlying assumption that the datasets being integrated have similar 'variable features' and spread. In other words, all the cell states or clusters should be covered in all the datasets being integrated. This makes it important to utilize integration methods not for comparing the differences between different conditions, but rather comparing the similarity between datasets being integrated, or one needs to have internal controls for proper establishments of 'variable features' or low dimensional spaces within individual datasets prior to integration.

Furthermore, 'Perturbseq' and its derivatives enabled the CRISPR/Cas9 based gene interruptions to be performed together with scRNA-seq, which allows identification of which gene is being perturbed in individual cells as well as the transcriptome information associated within the same cells being 'perturbed' (Dixit et al., 2016). However, due to some technical challenges that 'perturbseq'-based methods faced, such as the viral recombination problem that potentially dis-associated the sgRNA with the barcode being sequenced (Xie et al., 2018), the pool based perturbation studies in scRNA-seq have been technically challenging and still mainly in the technique demonstration land (Datlinger et al., 2017; Gasperini et al., 2019; Replogle et al., 2020). Nevertheless, the 'perturbseq' concept in single-cell analysis has opened up a new dimension of experimental perturbation assays, enabling potentials for dissection of molecular mechanisms, and reaching beyond the 'descriptive analysis'. As discussed earlier, understanding the perturbation outcomes on population distributions of developing T cells will heavily rely on a consistent and internally controlled experimental setup, and unbiased transcriptomic measurements. Therefore, our study has utilized the single-cell perturbation tools extensively, in a pool-based and batch-controlled manner, which will be discussed in Chapter 3.

Gene regulatory network modeling has played a major role in advancing the understanding of developmental systems, as the mechanism of development is based on ordered activations of gene regulatory networks, turning on cascades of regulators and generating

an irreversible path of differentiation (Davidson, 2010; Peter and Davidson, 2015). Classically, GRN inference has been based on analyzing steady-state data corresponding to gene knockout experiments, where one gene is silenced and changes in the steady-state expressions of other genes are observed. However, it can be difficult to know if steady states are achieved in a heterogeneous population. In addition, carrying out knockout experiments on a large number of genes is costly and technically difficult. Gene regulatory network inference has been conducted in numerous bulk gene expression profile studies, using computational tools such as weighted gene co-expression network analysis (WGCNA), or combining transcriptome and epigenome data (Chai et al., 2014; Langfelder and Horvath, 2008; Thompson et al., 2015). Many of the tools are based on the assumption that genes that are highly correlated in expression between different samples should be co-regulated. Therefore, in theory, scRNA-seq data can be simply treated as samples of bulk RNA-seq to infer regulatory structures. However, there are two immediate challenges in using just single-cell expression profiling applications for GRN inferences: 1) Correlation does not infer the direction of regulation. 2) Due to the technical noise in scRNA-seq, network inference needs to be carried out in similar cell types or states, and closely attended. Recently, Aibar et al., 2017 developed the SCENIC method to perform GRN inference based on co-expression of TFs from single cells' expression profiles using an ensemble tree based method (GENIE3, Huynh-Thu et al., 2010) and the TF-binding site search near transcription start sites of all their co-expressed genes. They demonstrated a robust prediction between TFs and target genes using single-cell data (Aibar et al., 2017). In Chapter 3, the author will show our explorations of SCENIC analysis on the scRNA-seq data from our perturbation studies, and will also discuss the limitations and a few newer tools in Chapter 4.

In summary, this thesis will focus on using single-cell analysis to understand the fundamentals of regulations in early T-cell development. The second chapter provides a thorough characterization of *in vivo* thymocytes' single-cell expression profile using complementary single-cell tools, revealing the dynamic expression changes leading to T-lineage commitment. The third chapter focuses on the effects of perturbations of key TFs

in early T-cell development. Combining different types of *ex-vivo* differentiation assays and additional efforts in optimizing single-cell pool-based perturbation strategies, normal and perturbed differentiation trajectories will be presented, as well as the inferred regulatory changes in the different perturbation conditions.

BIBLIOGRAPHY

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: Single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Allman, D., Sambandam, A., Kim, S., Miller, J.P., Pagan, A., Well, D., Meraz, A., and Bhandoola, A. (2003). Thymopoiesis independent of common lymphoid progenitors. Nat. Immunol. *4*, 168.

Bell, J.J., and Bhandoola, A. (2008). The earliest thymic progenitors for T cells possess myeloid lineage potential. Nature *452*, 764–767.

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.

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.

Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. Comput. Biol. Med. *48*, 55–65.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301.

Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. Nature *468*, 911–920.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853-1866.e17.

Drissen, R., Buza-Vidas, N., Woll, P., Thongjuea, S., Gambardella, A., Giustacchini, A., Mancini, E., Zriwil, A., Lutteropp, M., Grover, A., et al. (2016). Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. Nat Immunol *17*, 666–676.

Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell *176*, 377-390.e19.

Germar, K., Dose, M., Konstantinou, T., Zhang, J., Wang, H., Lobry, C., Arnett, K.L., Blacklow, S.C., Aifantis, I., Aster, J.C., et al. (2011). T-cell factor 1 is a gatekeeper for T-cell specification in response to Notch signaling. Proc. Natl. Acad. Sci. U. S. A. *108*, 20060–20065.

Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-Wallscheid, N., Dress, R., Ginhoux, F., et al. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. Nat. Cell Biol. *20*, 836–846.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421–427.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. *17*, 77.

Hattori, N., Kawamoto, H., Fujimoto, S., Kuno, K., and Katsura, Y. (1996). Involvement of transcription factors TCF-1 and GATA-3 in the initiation of the earliest step of T cell development in the thymus. J. Exp. Med. *184*, 1137–1147.

Heinzel, K., Benz, C., Martins, V.C., Haidl, I.D., and Bleul, C.C. (2007). Bone marrow-derived hemopoietic precursors commit to the T cell lineage only after arrival in the thymic microenvironment. J Immunol *178*, 858–868.

Hosokawa, H., Romero-Wolf, M., Yui, M.A., Ungerbäck, J., Quiloan, M.L.G., Matsumoto, M., Nakayama, K.I., Tanaka, T., and Rothenberg, E.V. (2018). Bcl11b sets pro-T cell fate by site-specific cofactor recruitment and by repressing Id2 and Zbtb16. Nat. Immunol. *19*, 1427–1440.

Hosoya, T., Kuroha, T., Moriguchi, T., Cummings, D., Maillard, I., Lim, K.-C., and Engel, J.D. (2009). GATA-3 is required for early T lineage progenitor development. J. Exp. Med. *206*, 2987–3000.

Hozumi, K., Negishi, N., Tsuchiya, I., Abe, N., Hirano, K., Suzuki, D., Yamamoto, M., Engel, J.D., and Habu, S. (2008). Notch signaling is necessary for GATA3 function in the initiation of T cell development. Eur. J. Immunol. *38*, 977–985.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. *11*, 774–785.

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.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Langfelder, P., and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Li, L., Leid, M., and Rothenberg, E.V. (2010a). An early T cell lineage commitment checkpoint dependent on the transcription factor Bcl11b. Science *329*, 89.

Li, P., Burke, S., Wang, J., Chen, X., Ortiz, M., Lee, S.-C., Lu, D., Campos, L., Goulding, D., Ng, B.L., et al. (2010b). Reprogramming of T cells to natural killer-like cells upon Bcl11b deletion. Science *329*, 85–89.

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. Proc. Natl. Acad. Sci. *114*, 5800–5807.

Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. Nat. Methods *11*, 360.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Mansson, R., Zandi, S., Welinder, E., Tsapogas, P., Sakaguchi, N., Bryder, D., and Sigvardsson, M. (2010). Single-cell analysis of the common lymphoid progenitor compartment reveals functional and molecular heterogeneity. Blood *115*, 2601–2609.

Miyamoto, T., Iwasaki, H., Reizis, B., Ye, M., Graf, T., Weissman, I.L., and Akashi, K. (2002). Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment. Dev. Cell *3*, 137–147.

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

Ng, S.Y.-M., Yoshida, T., Zhang, J., and Georgopoulos, K. (2009). Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. Immunity *30*, 493–507.

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

Orkin, S.H. (2000). Diversification of haematopoietic stem cells to specific lineages. Nat. Rev. Genet. *1*, 57–64.

Orkin, S.H. (2003). Priming the hematopoietic pump. Immunity *19*, 633–634.

Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: An evolving paradigm for stem cell biology. Cell *132*, 631–644.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell *163*, 1663–1677.

Peter, I.S., and Davidson, E.H. (2015). Genomic control process: Development and evolution (London, UK ; San Diego, CA, USA: Academic Press is an imprint of Elsevier).

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol.

Rothenberg, E.V. (2019). Programming for T-lymphocyte fates: Modularity and mechanisms. Genes Dev. *33*, 1117–1135.

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.

Schlitzer, A., Sivakamasundari, V., Chen, J., Sumatoh, H.R.B., Schreuder, J., Lum, J., Malleret, B., Zhang, S., Larbi, A., Zolezzi, F., et al. (2015). Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. Nat. Immunol. *16*, 718–728.

Scripture-Adams, D.D., Damle, S.S., Li, L., Elihu, K.J., Qin, S., Arias, A.M., Butler, R.R., Champhekar, A., Zhang, J.A., and Rothenberg, E.V. (2014). GATA-3 dose-dependent checkpoints in early T cell commitment. J. Immunol. Baltim. Md 1950 *193*, 3470–3491.

See, P., Dutertre, C.-A., Chen, J., Günther, P., McGovern, N., Irac, S.E., Gunawan, M., Beyer, M., Händler, K., Duan, K., et al. (2017). Mapping the human DC lineage through the integration of high-dimensional techniques. Science *356*, eaag3009.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron *92*, 342–357.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21.

Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat. Methods *14*, 381–387.

Taghon, T., Yui, M.A., and Rothenberg, E.V. (2007). Mast cell lineage diversion of T lineage precursors by the essential T-cell transcription factor GATA-3. Nat Immunol *8*, 845–855.

Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: From network reconstruction to evolution. Annu. Rev. Cell Dev. Biol. *31*, 399–428.

Ting, C.N., Olson, M.C., Barton, K.P., and Leiden, J.M. (1996). Transcription factor GATA-3 is required for development of the T-cell lineage. Nature *384*, 474–478.

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.

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

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.

Wada, H., Masuda, K., Satoh, R., Kakugawa, K., Ikawa, T., Katsura, Y., and Kawamoto, H. (2008). Adult T-cell progenitors retain myeloid potential. Nature *452*, 768–772.

Weber, B.N., Chi, A.W.-S., Chavez, A., Yashiro-Ohtani, Y., Yang, Q., Shestova, O., and Bhandoola, A. (2011). A critical role for TCF-1 in T-lineage specification and differentiation. Nature *476*, 63–68.

Welinder, E., Åhsberg, J., and Sigvardsson, M. (2011). B-lymphocyte commitment: Identifying the point of no return. Semin. Immunol. *23*, 335–340.

Xie, S., Cooley, A., Armendariz, D., Zhou, P., and Hon, G.C. (2018). Frequent sgRNA-barcode recombination in single-cell perturbation assays. PLOS ONE *13*, e0198635.

Xu, W., Carr, T., Ramirez, K., McGregor, S., Sigvardsson, M., and Kee, B.L. (2013). E2A transcription factors limit expression of Gata3 to facilitate T lymphocyte lineage commitment. Blood *121*, 1534–1542.

Yang, Z., and Michailidis, G. (2015). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics btv544.

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.

Zandi, S., Åhsberg, J., Tsapogas, P., Stjernberg, J., Qian, H., and Sigvardsson, M. (2012). Single-cell analysis of early B-lymphocyte development suggests independent regulation of lineage specification and commitment in vivo. Proc Natl Acad Sci U A *109*, 15871–15876.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Chapter 2

SINGLE-CELL ANALYSIS REVEALS REGULATORY GENE EXPRESSION DYNAMICS LEADING TO LINEAGE COMMITMENT IN EARLY T CELL DEVELOPMENT

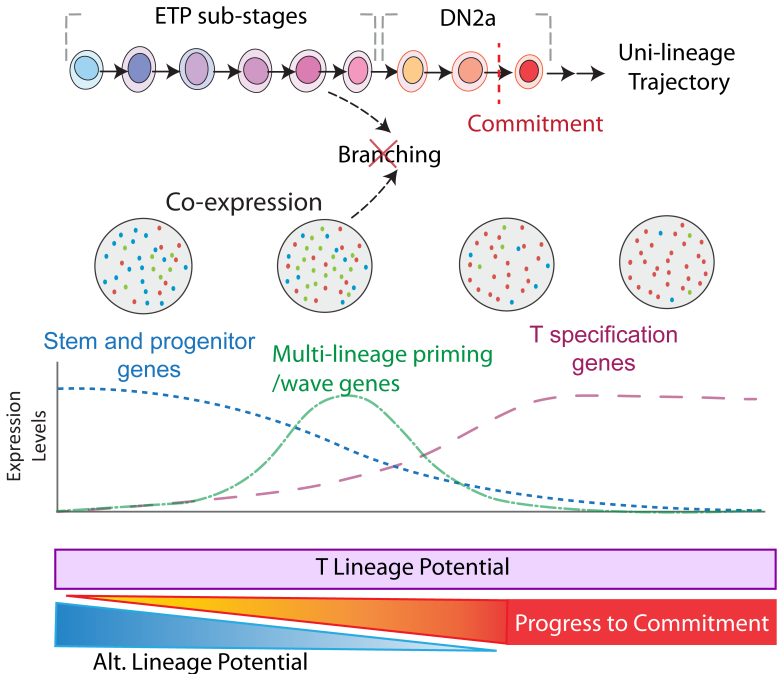
This chapter is adapted from the published article:

Zhou, W., Yui, M.A., Williams, B.A., Yun, J., Wold, B.J., Cai, L., and Rothenberg, E.V. (2019). Single-cell analysis reveals regulatory gene expression dynamics leading to lineage commitment in early T-cell development. Cell Systems 9, 321-337.e9. DOI: 10.1016/j.cels.2019.09.008

W.Z. performed most of the experiments, analyzed the data, and wrote the paper.

SUMMARY

Intrathymic T-cell development converts multipotent precursors to committed pro-T cells, silencing progenitor genes while inducing T-cell genes, but the underlying steps have remained obscure. Single-cell profiling was used to define the order of regulatory changes, employing single-cell RNA-seq for full transcriptome analysis, plus multiplex single-molecule fluorescent in situ hybridization (seqFISH) to quantitate functionally important transcripts in intrathymic precursors. Single-cell cloning verified high T-cell precursor frequency among the immunophenotypically-defined "early T-cell precursor" (ETP) population; a discrete committed granulocyte precursor subset was also distinguished. We established regulatory phenotypes of sequential ETP subsets; confirmed initial co-expression of progenitor- with T-cell specification genes; defined stage-specific relationships between cell-cycle and differentiation; and generated a pseudotime model from ETP to T-lineage commitment, supported by RNA velocity and transcription factor perturbations. This model was validated by developmental kinetics of ETP subsets at population and clonal levels. The results imply that multilineage priming is integral to T-cell specification.



Graphic Abstract

Introduction

The generation of T cells begins in postnatal mice as multipotent precursor cells enter the thymus from the bone marrow and undergo multiple rounds of proliferation and differentiation events before T-lineage commitment (Porritt et al., 2003; Rothenberg et al., 2008; Taghon et al., 2005; Yui et al., 2010). While many key regulators of T-cell specification and commitment are known (Yui and Rothenberg, 2014), the types of thymic T-cell progenitors and the steps that they undergo to initiate commitment remain unclear.

Early T-cell progenitors (ETPs), cells double-negative (DN) for CD4 and CD8 that are Kit^+ $CD44^+$ $CD25^-$, represent the earliest defined stage in each cohort of mouse thymocytes. After ~1 wk of proliferation and differentiation under the influence of environmental signals, including Notch ligands and cytokines from the thymic stroma, ETPs asynchronously progress into the DN2a stage, marked by upregulation of surface CD25 (*Il2ra*) (Porritt et al., 2003) (Fig. 1a). Commitment follows in a separate step, coinciding with the up-regulation of transcription factor *Bcl11b* and global changes in chromatin landscapes (Hu et al., 2018; Ikawa et al., 2010; Kueh et al., 2016; Li et al., 2010). However, ETPs themselves are poorly characterized before they progress to DN2a stage. While single-cell colony assays show that many ETPs are individually multipotent as well as T-cell competent (Bell and Bhandoola, 2008; Wada et al., 2008), none of the ETP markers are exclusive to T cells, so "ETPs" could also include committed non-T-lineage precursors. In addition, T-cell precursors can migrate to the thymus from different hematopoietic precursor states (CLP and LMPP) (Saran et al., 2010) (Fig. 1a). Thus, in a 'snapshot' of single ETP transcriptomes, there could be heterogeneity due to different input origins, different developmental stages, and/or contamination with cells committed to alternative fates.

The expression of important regulators in early T-cell development has mostly been studied in bulk populations. Notch1 signaling (Besseyrias et al., 2007; Pui et al., 1999; Radtke et al., 1999) and transcription factors GATA3 and TCF1 (encoded by *Tcf7*) play

indispensable roles to establish T-cell identity from the earliest stages (García-Ojeda et al., 2013; Germar et al., 2011; Hosoya et al., 2009; Schilham et al., 1998; Scripture-Adams et al., 2014; Ting et al., 1996; Weber et al., 2011). With *Notch1, Gata3,* and *Tcf7,* other regulators more widely shared (*Myb, Gfi1, Runx1, Tcf3*) are also essential for cells starting the T-cell pathway. Expression of these genes is readily detectable in the ETP population by bulk RNA analysis, but in an unknown fraction (De Obaldia and Bhandoola, 2015; Mingueneau et al., 2013; Yui and Rothenberg, 2014; Yui et al., 2010; Zhang et al., 2012). Further, many legacy "non-T" genes, associated with "stemness" and/or non-T-lineage fates, are also expressed at low levels in early pro-T cell populations, including several with potential gene network interactions with the "T-cell" regulators (Longabaugh et al., 2017; Yui and Rothenberg, 2014). It is unclear if they are an integral part of the T-lineage program or merely expressed in contaminating cells. If the former, the expressions of stem and progenitor "non-T" genes may be indicators of multi-lineage priming and/or important regulatory network relationships between the declining stem cell program and ongoing T-cell specification. The single-cell expression patterns of these genes relative to T-cell genes are essential to elucidate the significance of their expression in T-cell development.

Single cell transcriptional profiling by RNAseq (scRNA-seq) has transformed our understanding of hematopoietic differentiation and heterogeneity (Boudil et al., 2013; Giladi et al., 2018; Ishizuka et al., 2016; Karamitros et al., 2018; Knapp et al., 2018; Olsson et al., 2016; Paul et al., 2015; Pina et al., 2012; Tusi et al., 2018; Velten et al., 2017; Zandi et al., 2012; Zheng et al., 2018), providing nominally unbiased full-transcriptome information and effectively separating distinct cell types within complex populations. However, in most scRNA-seq applications the accuracy and robustness of measurement are biased towards highly expressed genes, which mostly characterize already-diverged developmental end states. Here, the goal is to resolve a continuum of changing transcriptome states within a developmental pathway, and relate them to changes in the controlling regulatory network state. This demands accurate, statistically robust quantitation of regulatory genes encoding key transcription factors (TFs), which

are often expressed at low RNA copy numbers per cell. Therefore, we have taken advantage of recent advances in single-molecule fluorescence in situ hybridization (Raj et al., 2006), which visualizes and counts individual mRNA transcripts directly in individual cells at very high sensitivity. Recently, a version incorporating a temporal barcoding scheme, “seqFISH”, has been developed that uses a limited set of fluorophores but can detect hundreds to thousands of distinct sequences in the same cells (Lubeck et al., 2014; Shah et al., 2016a, 2016b), and another similar strategy, “merFISH”, has also been described (Chen et al., 2015). We have used the highly sensitive seqFISH technique to quantify transcripts of a curated panel of 65 regulatory and developmental state marker genes in pro-T cells.

Thus, combining droplet-based scRNA-seq, deep-sequenced whole-transcript scRNA-seq, and seqFISH for key regulatory genes, together with developmental assays of sorted subsets and clones from sorted founders, we have characterized the sequence of gene expression transitions in early intrathymic mouse T-cell precursors and regulatory gene dynamics of T-cell specification. Our results show an unexpectedly complex, multistep progression through which the cells shed stem cell characteristics and approach T-cell lineage commitment. The results give new insights into the transition from multipotency to commitment and how it is controlled.

RESULTS

Single-cell developmental competence and bulk population phenotype of ETPs

Broad outlines of mouse T-cell development are well-studied, but the initial events upon entry of T-cell precursors into the thymus remain obscure. Most uncertain are events that occur within the ETP population and in transition to DN2a. While later stages are clearly defined as shown in Fig. 1a, ETPs are rare, individually multipotent and poorly separable by flow cytometry from other, irrelevant multipotent precursors. However, vital regulatory events including the exclusion of B-cell potential (Heinzel et al., 2007) and epigenetic priming of the cells for later commitment (Kueh et al., 2016; Ng et al., 2018) occur during the ETP stage(s). Thus, we have investigated whether different precursors contribute functional ETP starting state(s); their precise sequence of regulatory state changes leading to T-lineage commitment; and whether they develop by single or branched pathways. In Fig. S1 a-b, we summarize the logical sequence of questions addressed, the experimental approaches, and the data- handling pipeline.

To characterize the earliest mouse thymic T-cell progenitors through T-lineage commitment, we used fluorescence-activated cell sorting (FACS) of DN cells to isolate the ETP and DN2a subsets (cf.1a). Only a tiny fraction of total thymocytes (<0.01%) at steady state are uncommitted ETP and DN2a cells, distinguished from all others by their expression of growth factor receptor c-Kit. Expression of a *Bcl11b-YFP* knock-in reporter (Kueh et al., 2016) that distinguishes uncommitted (YFP-) from newly committed (YFP+) DN2a cells was used to mark the commitment milestone (Fig. 1b,c; Table S1). Another growth factor receptor, Flt3, has been reported to characterize the least mature ETPs (Ramond et al., 2014; Sambandam et al., 2005), and in many experiments we used it to subdivide ETPs either by FACS or in silico.

To estimate the fraction of "ETPs" that actually possess T-lineage developmental potential, we carried out single-cell clonal culture experiments. Individual ETP cells were plated in microwells and tracked by live imaging in T-cell development culture conditions to determine how many could generate progeny that reach DN2 stage and undergo

commitment (Fig. 1b, top, see Methods). Of 78 founder ETPs, 66 survived and were tracked for 6 days. Almost all clones generated cells expressing CD25 and Bcl11b-YFP by day 6 (Fig. 1b, bottom). Two of the 66 clones only produced small non-T lineage cells resembling granulocytes, consistent with alternative lineage affiliation, as discussed below. Thus, >90% of viable clonogenic ETPs possessed T-lineage precursor activity.

Bulk RNA expression patterns showed that ETP populations were clearly distinct from DN2a populations, with many of the differences reflecting downregulation of ETP-expressed genes in DN2a (Fig. 1c,d). ETP populations expressed many characteristic "non-T" genes, including genes expressed in mature granulocytes, macrophage, dendritic cells, NK cells, and stem cells, but not in mature T cells (www.immgen.org) (Fig. 1c), consistent with previous bulk RNA expression studies (Mingueneau et al., 2013) [rev. by (Rothenberg et al., 2016; Yui and Rothenberg, 2014)]. Both uncommitted and committed DN2a cells expressed lower levels of multipotent progenitor-associated genes *Flt3, Lmo2,* and *Mef2c* than ETPs, although the DN2a cells continued to express another multipotency-associated gene, *Spi1* (encoding transcription factor PU.1) (Fig. 1c, d). In contrast, sorted $Flt3^{+}$ and $Flt3^{-}$ ETP populations appeared similar, and both expressed the essential T-cell regulatory genes *Gata3* and *Tcf7,* implying that at least some ETPs have started the T-lineage specification (Fig. 1d). Such population-level analysis raised the question of how many substates were comprised in ETPs, how homogeneously cells progressed through them, and which states reflected the presence of contaminating cells with no T-cell potential.

To determine the sequence of developmental changes in these earliest pro-T cells, we FACS-purified Kit^{high} thymocytes across the ETP-DN2 developmental continuum, and analyzed their single-cell transcriptomes and also their developmental potentials (Fig. S1c). To anchor the developmental direction, for most analyses we also added a small number of purified committed DN3 cells (Fig. S1c). The transcriptomes of these samples were defined by three methods: seqFISH, whole-transcriptome 3'-end biased sequencing (10X Chromium), and whole-transcriptome full-transcript sequencing (Fluidigm C1-

SmartSeq2). Results from these methods were highly concordant, but highlighted different aspects of the gene expression programs.

Sensitive monitoring of developmentally important regulatory genes in single cells by seqFISH

Expression in ETP populations of the essential T-cell regulatory genes, *Gata3* and *Tcf7,* appeared in accord with their high clonogenic T-cell precursor frequency overall, but single-cell methods were needed to determine which ETP subsets activate these T-cell regulators. First, we sought to determine whether the ETPs expressing characteristic multipotent progenitor-associated regulatory genes included the individual cells entering the T-cell pathway. However, as shown in Table S1 and in previous studies, regulatory genes have bulk RNAseq signals measured at <10 FPKM, below the robust detection limit of common single-cell approaches (also see below). We therefore applied a targeted seqFISH approach, focused on a curated set of regulatory and lineage-informative genes. Most of these 65 genes are known to be functionally significant in early T or multipotent progenitor cells (Hosokawa et al., 2018a; Rothenberg et al., 2016; Yui and Rothenberg, 2014), while others are distinctive markers for stages in T and non-T pathways (genes and criteria for selection shown in Table S2). Probes for 54 genes with low to medium expression level were used in barcoding rounds of seqFISH with hybridization chain reaction (HCR seqFISH), followed by sequential rounds of non-barcoding HCR single molecule FISH (HCR smFISH) to detect the remaining genes, including highly expressed genes, controls, and genes with shorter transcripts, and finally followed by immunofluorescent staining (Fig. 2a; see STAR Methods). Analyses used sorted populations of ETP-DN2a from mice of 3 different ages (4 wk, 2874 cells; 5 wk, 4413 cells; 8 wk, 1736 cells) (Fig. S2a, c), plus similar numbers of DN3s from the same animals imaged in separate lanes of hybridization-cells.

As detailed in Fig. S2a-c, seqFISH measurements were sensitive and reproducible across all three ages tested in independent experiments without batch correction. It faithfully detected critical genes like *Tcf7* and *Notch1* that were hard to detect consistently in ETPs

with 10X Chromium scRNA-seq (Fig. S2b, d). Furthermore, protein and seqFISH RNA expression (c-Kit, PU.1 and TCF-1 protein vs. *Kit, Spi1* and *Tcf7*) correlated in the same cells (Fig. S2e,f).

SeqFISH reveals co-expression of stem/progenitor and T-cell regulators in individual ETPs

SeqFISH confirmed regulatory state differences between Kit-high cells categorized as ETP or DN2 based on expression of *Il2ra* (CD25). DN2s expressed lower levels of multiple ETP-associated genes (*Flt3*, *Cd34, Mpo, Lmo2)* while a subset expressed much higher levels of the commitment-associated gene *Bcl11b* (Fig. 2b). Pairwise coexpression patterns of the seqFISH gene set among all ETP-DN3 cells sampled (Fig. 2c) clearly distinguished a "T-associated" group of genes, including a subset highly coexpressed in DN3s (*Ptcra, Rag1, Cd3e, Cd3g, Spib*, *Tcf12,* and *Lef1*), from at least two other gene groups containing coexpressed 'Stem and Progenitor' genes (*Kit, Spi1, Lyl1, Bcl11a, Runx3, Pim1, Erg, Cd34, Hhex, Lmo2*, and *Cd44*). Each of these stem/progenitor groups also contained genes normally associated with non-T cells (e.g. *Mpo, Irf8, Pdgfrb*) (Fig. 2c). In addition, other separate gene subgroups contained *Gata3* and Ikaros (*Ikzf*) family TFs, plus their interaction partners found in T and innate-lymphoid cells (*Zfpm1*, *Gfi1*, and *Zbtb16*). These "T/ILC" groups of genes showed intermediate correlation both with the stem/progenitor genes and with the T-associated genes.

The seqFISH results enabled the cells to be resolved into 9 clusters (Fig. 2d,e), based on high-dimensional analysis using Smart Local Moving (SLM) clustering (Waltman and van Eck, 2013). Clusters were provisionally ordered by known "endpoint" genes, starting from the earliest ETP cells, identified by *Flt3* and *Cd34* enrichment, to committed DN3 cells, marked by high *Ptcra*, *Cd3e* and *Cd3g*. This initial clustering was broadly consistent with results from previous bulk RNA analysis. However, it revealed that progenitor- or alternative-lineage genes were not all co-expressed, but instead displayed distinct although overlapping patterns. Among the earliest cells, for example, *Lmo2* and *Flt3* were co-expressed in a more restricted developmental pattern (mainly cluster 2), than *Kit* and

Spi1. Cells in DN3 split into 3 clusters, two of which represented DN3a stages with high levels of *Bcl11b, Ets1*, *Ptcra, Cd3g, Cd3e,* and *Rag1* (clusters 6 & 0, mainly distinguished by different levels of *Tcf7*). The third DN3 cluster (cluster 7) could be identified as DN3b cells that had passed the β-selection checkpoint based on the T-cell receptor expression (see Fig. 1a), with enrichment of *Lef1, Id3*, *Tcf7, and Pgk1* but downregulation of DN3a genes. Only one small ETP subpopulation, low in *Tcf7* expression (Fig. 2d,e, cluster 8) and highly coexpressing *Mpo*, *Spi1*, *Cebpa*, *Lmo2,* and *Irf8,* but not other progenitor genes, appeared to be discontinuous from the others. This outgroup population was seen in every analysis we performed, and is identified below. Note that, in each of the seven clusters spanning $Flt3^+$ ETP to DN3a, the expression of key regulatory genes such as *Spi1*, *Tcf7*, and *Bcl11b* was relatively homogeneous; 89-100% of cells expressed >3 copies/cell in relevant clusters (Fig. S2g).

Given the distinctive expression of progenitor-associated genes among ETPs, a central question was whether the cells expressing these genes are representative of the cells entering the T-cell program. We used seqFISH to assess which legacy stem and progenitor genes are coexpressed with *Gata3* and *Tcf7* in individual cells. *Gata3* activation began in ETPs with varying levels of *Tcf7 transcripts,* and became concordant in DN2-DN3 stages (Fig. 2f). As expected (Kueh et al., 2016), the T-lineage commitment gene *Bcl11b* was activated exclusively in cells that express *Tcf7*, and almost completely within the DN2 stage (Fig. 2f).

To ask directly how ETPs expressing Notch-induced *Gata3* and/or *Tcf7* differ from ETPs not expressing these genes, we compared the transcript counts of all other seqFISH genes between ETP cells with and without expression of *Gata3* (>10 transcripts vs. ≤3 transcripts) and/or *Tcf7* (>20 transcripts vs. ≤5 transcripts) (Table S3). The seqFISH results confirmed that ETPs activating *Gata3* and/or *Tcf7* were markedly different from committed, *Bcl11b*-expressing DN2s in their expression levels of >30 genes (p values $<10^{-6}$, two-tailed T test, unequal variances). However, ETPs expressing *Gata3* and/or *Tcf7* differed very little from ETPs lacking expression of both *Gata3* and *Tcf7*. ETPs with and

without *Gata3* and/or *Tcf7* expression were statistically indistinguishable in their expression of *Notch1,* or of stem/progenitor-associated genes *Spi1, Cd34, Mpo, Mef2c,* or *Bcl11a*, which were expressed by the great majority of both (Table S3). Only *Gfi1b* and *Runx3* differed with $p < 10^{-6}$, while *Flt3* and *Lmo2* were slightly lower in expression, and T-promoting genes, including *Hes1* and *Ets1,* were slightly higher in the cells expressing *Gata3* and/or *Tcf7* than in those without *Gata3* or *Tcf7*. Overall, these seqFISH results show that there is continuity between the stem/progenitor gene expression patterns in those individual ETPs starting T-cell development and most other ETPs.

Individual ETPs in fact spanned boundaries of the gene-set co-expression clusters seen in the overall ETP—DN3 population (cf. Fig. 2c). For example, the myeloid-associated gene, *Mpo*, encoding myeloperoxidase, was expressed at higher levels in ETPs than either *Gata3* or *Bcl11b*, but a major fraction of *Mpo*-expressing cells also clearly expressed *Tcf7* (>20 copies/cell) (Fig. 2f). The growth-promoting gene *Pim1,* which marked intermediate clusters (Fig. 2d, clusters 3,5), was activated in both *Tcf7*-low and *Tcf7*-high ETPs and then increased in DN2 cells with varied *Tcf7* expression. These results suggest that although not expressed in mature T cells, *Mpo* as well as *Pim1* were substantially expressed within cells initiating the T-cell program and are not from contaminants.

Deep-sequencing confirms stem/progenitor and "non-T" associated regulatory genes co-expressed with *Gata3* and *Tcf7* in individual ETPs

To extend this inquiry to a sensitive genome-wide analysis of single cells, we carried out whole-transcript Smartseq2 scRNA-seq analysis (from C1 Fluidigm; "C1") of highly purified ETP-DN2a cells (n=193 cells) (Fig. 3). Despite the low cell numbers, semi-supervised clustering of the C1 dataset (based on differentially expressed genes described in Fig.1c and Table S1) yielded high-quality gene expression patterns that supported and extended those seen in seqFISH. DN3 endpoint cells could not be included, but the results again separated ETP-DN2a cells expressing combinations of multipotent progenitor-associated genes from the cells more highly expressing T-lineage associated genes (Fig. 3a-e; Table S4, "C1_supervised_markers"). Again, one small outgroup was found with a

highly divergent program (Fig. 3a, PC2) lacking T-cell gene expression, apparently among cells with a "Flt3$^-$ ETP" phenotype (Fig. 3b-e; cluster 9). Nevertheless, in the rest of the cells, C1-Smartseq data confirmed that multipotency-associated genes *Spi1*, *Flt3, Lmo2*, *Mef2c, Cd7* and *Irf8,* were all frequently co-expressed with *Tcf7* and *Gata3* in individual ETPs, sometimes continuing into DN2s. But whereas *Spi1* could still be co-expressed with the late DN2a gene *Bcl11b*, in contrast *Irf8, Lmo2*, and *Flt3* expression was almost dichotomous with *Bcl11b* (Fig. 3f,g). This supports the interpretation that expression of these stem/progenitor genes selectively characterizes most ETPs as they enter the T-cell developmental program.

10x scRNA-seq shows tightly connected ETP-DN2 cell populations

The seqFISH and C1 results indicated that the regulatory states of most ETP cells are within the continuum of the T-cell specification trajectory. We therefore dissected this trajectory in depth by whole-transcriptome analyses of thousands of enriched ETP-DN2a cells, again with DN3 cells as an internal reference, using 10X Chromium v2 (10X). Samples of 4627 (replicate1) and 7076 (replicate2) ETP-DN2 cells plus 10% DN3 cells yielded 3' end-enriched transcriptome profiles with UMI quantitation. Upon dimensional reduction (tSNE or UMAP), RNA expression phenotypes separated the cells into 2-3 distinct clusters. These corresponded respectively to a large mix of ETP-DN2 cells, DN3 cells, and a small outgroup (Figure 4a,b), judged by expression patterns of genes characterizing different developmental stages or lineages *(e.g., Elane* (granulocytes)*, Mpo* (macrophages)*, Klrd1* (NK cells))(Fig. 4c, highlighted in red). Within the ETP-DN2 continuum, stage-defining genes such as *Kit* (ETP-DN2), *Il2ra* (DN2-DN3), and *Bcl11b* (committed DN2-DN3) were localized to different regions but not well-separated. Again, the small outgroup expressed granulocyte-associated genes, e.g. *Elane* (Fig. 4c) along with some progenitor-associated genes (*Kit, Spi1, Lmo2*), as in the seqFISH (cluster 8 in Fig. 2g, h) and C1 analyses (cluster 9, Fig. 3b-e). Highly concordant results were found in an independent 10X experiment (Fig. S3a-d), and the 10X results overall agreed well with the C1 and seqFISH results after CCA scaling (Fig. S3e).

Fine resolution unsupervised clustering by SLM distinguished 14 sub-clusters of cells across the ETP-DN3 range (Fig. 4a,b,d; Table S4, "10X unsupervised"). *Bcl11b* expression again marked clusters of recently committed cells (Fig. 4d, clusters 5, 2, 9, 11). The spiked-in DN3 cells again included both pre-β-selection DN3a cells (cluster 9: high *Ptcra*, *Cd3g, Cd3d,* and *Cd3e*, non-proliferative) and DN3b cells that had begun β-selection (cluster 11: high levels of *Lef1* and proliferative markers). The *Elane*-expressing outgroup was cluster 13. This left the clusters of greatest interest, representing earlier, pre-commitment pro-T cells (clusters 0, 10, 4, 6, 7, 8, 12, 1), provisionally identified by their expression of progenitor-associated genes such as *Cd34, Lmo2* and *Mef2c*. However, among these earlier clusters, the ordering was ambiguous in unsupervised clustering, and the relationship to cluster 13 was still unclear. This was partly because transcripts of key T-cell genes *Notch1, Gata3,* and *Tcf7* did not change sharply enough to be identified as highly enriched in any particular ETP-DN2a cluster(s). Another source of ordering ambiguity among ETP-DN2a cells was the prominence of multiple states associated with cell cycle, in both biological replicates (Fig. 4d, Fig. S3d). Cells expressing S- or G2+M related genes (e.g. *Birc5*, *Mki67)* were found in clusters apparently representing different stages along the early-to-late developmental continuum.

Distinct T-cell differentiation kinetics and identification of committed granulocyte precursors among 'ETPs'

To confirm which gene expression clusters were associated with T- or non-T- lineage potential and to verify which were more or less advanced in T-lineage progression, we used marker genes that distinguished some of these clusters to fractionate ETPs by FACS, and then directly compared their developmental kinetics and fates under T-cell and non-T cell developmental conditions (Fig. S4a). We also sought to resolve whether the *Elane*-positive cells (Fig. 2e, cl. 8; Fig. 3, cl. 9; Fig. 4d, cl. 13) were part of the T-cell developmental pathway or a separate lineage. These cells uniquely expressed several granulocyte-associated genes, including *Elane, Ms4a3, Ly6c2,* and *Prtn3,* but lacked expression of *Notch1* or Notch-induced genes (*Hes1, Dtx1*), possibly resembling a bone marrow early pre-neutrophil precursor (Evrard et al., 2018). Distinctively, these cells co-

expressed surface receptors, CD63 and Ly6c2, detectable with antibodies that were used to purify them away from other ETP subsets for developmental tests.

We first confirmed that Flt3$^+$ ETPs were indeed more immature than the Flt3$^-$ ETPs. Flt3$^+$ and Flt3$^-$ ETPs (excluding CD63$^+$ Ly6c$^+$ cells) and DN2a (CD25$^+$Bcl11b-YFP$^-$) cells were co-cultured with OP9-DL1 stroma to provide T-cell differentiation conditions (Fig. S4b). Their progression was scored by two T-cell milestones: onset of CD25 expression, denoting transition from ETP to DN2a, and the subsequent expression of Bcl11b-YFP. Then, to test the developmental potential of the Elane$^+$ cells, CD63$^+$ Ly6c$^+$ cells were sorted and compared with CD63$^-$ Ly6c$^-$ ETPs. Unlike other ETPs, CD63$^+$ Ly6c$^+$ cells could not turn on CD25 or Bcl11b-YFP in T-cell culture conditions. Instead, they turned on the granulocyte marker Gr1 after 4-5 days (Fig. S4c-d). These populations were also tested for their ability to generate alternative lineages in non-T conditions, in the absence of Notch signaling and with cytokines supporting myeloid differentiation. Under these conditions, while other subsets of ETPs generated multiple types of non-T cells, CD63$^+$ Ly6c$^+$ cells exclusively gave rise to Gr1$^+$ granulocytes (Fig. S5). Thus, the CD63+Ly6c+ cluster in the thymic 'ETP compartment' is a committed granulocyte precursor, has no T potential, and differentiates independently of Notch signaling. Thus, expression of *Elane and Prtn3* in single-cell and bulk ETP RNA-seq is attributable to a distinct non-T- lineage population rather than to expression by uncommitted T-cell precursors.

Developmental progression shows stage-dependent relationships to cell cycle states

We could now address the gene regulatory states associated with T-cell specification per se, in the 10X data. To gain better resolution of possible component processes by topology on a more complex developmental manifold, we applied a force-directed layout algorithm using SPRING, visualizing long-distance as well as nearest-neighbor relationships of cells across three reduced dimensions (Weinreb et al., 2018) (Fig. S6). The SPRING graph revealed an ordered developmental continuum from ETP (*Il2ra* negative), through DN2a (*Il2ra* positive) and committed DN2 cells (*Bcl11b* positive), and into the separated DN3a

and DN3b cells (offset *Bcl11b* high populations) (Fig. S6a-b), roughly progressing from top right to bottom left (arrow in Fig. S6b). Within the main zones, the early ETP marker *Flt3* highlighted the top right edge, while ETP-DN2 gene *Spi1* lit up a distinctly larger area of the ETP-DN2 cluster, and the T-lineage commitment gene *Bcl11b* was activated only at the edge away from the *Flt3*-enriched zone and continuing into the offset DN3a and DN3b cells (Fig. S6a), consistent with known developmental relationships. However, the cells also varied strongly along an axis orthogonal to the developmental direction (Fig. S6). This second axis was represented by proliferative and cell cycle state markers as annotated in Fig. S6a. The resolution of two biologically meaningful but orthogonal axes of variance suggests that cells transition through multiple cell cycles as they progress through successive differentiation states, rather than confining cell cycling to a single state.

Notably, expression of many functionally important genes was not uniform across each band of cells along the "developmental axis". The G1-associated ETP-DN2 region (upper left) had a concentration of cells expressing *Gata3* and *Tcf7*, yet this region was also most enriched for cells expressing high levels of *Spi1*, *Cd7,* and *Tyrobp*, genes characteristic of non-T cells. Depending on the actual trajectory the cells take, this state could represent a developmental branch point, an alternative entry point for precursors, or a transiently induced upregulation of non-T genes even along the T-cell pathway.

RNA velocity analysis maps the developmental flux from ETP through DN2 and commitment

To elucidate the developmental fluxes between populations in the ETP-DN2 transition, we used RNA velocity analysis (Velocyto)(La Manno et al., 2018)(Fig. S7; Fig. 5a, b). This algorithm uses the ratio of unspliced, presumably nascent, pre-mRNAs to mature mRNAs to estimate the rate of RNA production change, and therefore the direction of regulatory change in low-dimensional transcriptome space, for cells moving through development. Indeed, 17% of reads in the 10X scRNA-seq data mapped to intronic regions of the genome (Fig. S7a, b). Data from the 10X analysis, omitting DN3b and granulocyte precursors, were plotted on a principal component space (PC1 and 2 shown in Fig. 5a,b),

with RNA velocity-based differentiation vectors superimposed on the same axes (Fig. 5b). Similarly to the SPRING layout, expression of known genes showed that cells separated orthogonally with cell cycle differences most evident along PC1 and developmental stage differences more along PC2 (Fig. 5a, b). Notably, though, despite this cell-cycle correlation, differences in cell cycle genes did not drive the velocity vector patterns, for the velocity vectors were nearly identical even when cell cycle genes were excluded from the calculation (Fig. S7c).

The velocity vector map indicated complex, PC1-biased differentiation trends within the ETP compartment distinct from those in DN2, and suggested that transition from ETP to DN2a occurred from a preferential regulatory state (Fig. 5b). While velocity vectors indicated that DN2 cells in all cell cycle states were uniformly progressing toward DN3 (central band of downward pointing arrows), the early ETPs (along the topmost zone, colocalized with *Flt3*) had velocity vectors suggesting two different attractors with distinct cell cycle states. Velocity vectors for the *Birc5*$^{+}$ ETPs (extreme top right, presumably in G2+M) appeared to be pointing to the left, toward another ETP state, where a subset of these *Birc5*$^{+}$ ETPs appeared to be developmentally static (dots or shortest arrows). Of note, these more static ETPs, possibly representing a self-renewing subset, also showed the highest ongoing transcription of *Hoxa9*, a homeobox gene associated with prethymic progenitor specification and leukemia (Gwin et al., 2013)(Fig. S7d). In contrast, ETPs with differentiation velocity vectors pointing toward an *Il2ra*$^{+}$*Bcl11b*$^{-}$ early DN2a state (down) were on the left, among *Birc5*-nonexpressing ETPs. Here, transitions from a *Cd7*-high ETP subset (extreme upper left) were most prominent. The velocity data suggest that the immediate precursors of DN2a cells were among particularly *Spi1*-high G1 phase ETP cells, many also transiently *Cd7* high, in the process of downregulating *Flt3* (Fig. 5a, Fig. S7e).

Supervised analysis of 10X data reveals a developmental trajectory from ETP through T-lineage commitment

The RNA velocity analysis was reinforced by the topology obtained when we used the 10X datasets to construct a developmental gene expression trajectory. The curated list of seqFISH genes (Table S2) was now used for supervised analysis of the whole transcriptome data, with DN3b cells and granulocyte precursors excluded (Fig. S8a,b; clusters in Table S4, "10X_supervised_markers"). DDRtree (Qiu et al., 2017a) was used to obtain a connected developmental trajectory and pseudotime staging of the cells (Fig. 5c-e, Fig. S8). From the independent replicates of the 10X analysis, 763 genes were significantly differentially expressed along the pseudotime axis in both (qval $<10^{-8}$), and these genes were clustered according to their expression patterns in a heat map (Fig 5f; listed in order in Table S5). Fig. 5f also indicates approximate subdivisions and regulatory landmarks; the pattern of expression in pseudotime of the curated genes themselves is shown in Fig. S8c. While the pseudotime model clearly supported the distinction between ETP and DN2a stages (approx. between subdivisions B & C, Fig 5f), additional substages were present, in accord with the seqFISH analysis (cf. Fig. S1), and these were not based on cell cycle gene clusters. Instead of monotonic increases or decreases in gene expression across the trajectory, another group of progenitor-associated genes (e.g. *Spi1*, *Cd7, Mpo,* and *Tyrobp*) was predicted to rise transiently upon down-regulation of *Flt3* within the ETPs (*Il2ra* negative), followed by their own down-regulation at a later DN2 stage. This implication also accorded with the unsupervised RNA velocity analysis. Similarly, in second or third waves during the ETP-DN2 transition and DN2 stages (subdivisions C & D-E, Fig 5f), other groups of genes including *Pim1* were predicted to undergo transient expression changes before the final committed DN3 regulatory state.

These predicted pseudotime trends were generally consistent with known regulatory relationships between landmark TFs, Bcl11b and PU.1 (encoded by *Spi1*) and individual target genes, based on perturbation experiments that defined targets of these factors genome-wide (Hosokawa et al., 2018a, 2018b; Ungerbäck et al., 2018). These perturbation tests defined 326 PU.1-upregulated genes, 237 PU.1-repressed genes, 394 Bcl11b-dependent genes, and 747 Bcl11b-repressed genes. Bcl11b and/or PU.1 targets represented 214 of the 763 pseudotime-indicator genes (Table S5), so we compared the

changes in these genes in pseudotime with changes in expression of *Bcl11b* and *Spi1* themselves. Fig. 5g shows the fractions of genes in individual pseudotime expression clusters that were significantly repressed by or dependent on PU.1 or Bcl11b (pattern details in Table S5). PU.1 indeed positively regulated genes in several distinct early clusters, particularly in the early transient wave (Fig. 5g, orange margin), but negatively regulated genes in late (DN3-associated) clusters. Bcl11b primarily activated genes upregulated late in pseudotime. Bcl11b repression targets were concentrated among early and intermediate pseudotime-expressed genes, especially in the two intermediate expression waves (Fig. 5g, groups with green and orange margins). These genes had been deduced to be Bcl11b repression targets because acute deletion of *Bcl11b* caused their expression to increase even in committed pro-T cells that had already reached DN2b. This supports the interpretation that the genes upregulated in the intermediate wave are expressed within the T-lineage specification pathway, and that their expression is then truncated by Bcl11b.

In vitro culture supports the single-cell trajectory and multilineage priming model

These intermediate expression waves were unpredicted (Mingueneau et al., 2013; Yui and Rothenberg, 2014), and might either reflect a succession of transient regulatory states during T-cell development or be computational artifacts of forcing branched gene expression changes into a single pathway. Specifically, in the DDRtree model, the end stage ETPs exhibited a small branch going off the trajectory, associated with upregulation of *Spi1, Hhex, Cd7*, and *Tyrobp*, genes strongly affiliated with myeloid, NK, or DC alternative fates. In pseudotime, however, these genes were modeled as transiently up-regulated in ETP. In support of the pseudotime model, ETPs expressing high levels of these genes (G1-enriched ETPs) were identified in the velocity analysis within the region most likely to transition to DN2 (Figure 5b). In seqFISH and C1 distribution analysis, we had also confirmed that these genes are expressed by a substantial population of cells (Fig. 2-3). Thus, two hypotheses can explain this early wave or branch pattern (Fig. 6a): 1. lineage branching, where levels of these non-T-cell associated transcripts are accumulated in a subset of cells that have branched off towards alternative fates; or 2. multilineage

priming, in which genes associated with alternative lineages are expressed transiently in early stages, reflecting the intrinsic regulatory network structure and phenotypic plasticity of uncommitted early T-cell stages. If lineage branching were true, then the pseudotime model expression pattern of transiently upregulated genes in late ETP would be inaccurate.

To test the two hypotheses functionally, we used the pseudotime analysis to identify markers that could distinguish between ETP subpopulations. We then FACS-purified ETP subsets based on their expression of these markers, and followed their T-lineage developmental kinetics, as well as their alternative lineage potentials, through *in vitro* culture (Fig. 6b). Whereas Flt3 marks earlier ETPs (Fig. 2, Fig. S4), the cell surface marker HSA (*Cd24a*) was predicted in pseudotime to be gradually up-regulated during late ETP stages, followed by Ly6d up-regulation (Fig. 5f). Unfortunately, CD7 could not be used for subset enrichment due to lack of a specific antibody. We therefore sorted ETPs into 6 sub-populations according to Flt3, HSA, and Ly6d expression (Fig. S8d), and tested them in OP9 co-culture systems with and without Notch ligand to compare their developmental potentials and speeds of T-lineage progression, as measured by upregulation of the Bcl11b-YFP reporter. In this T-lineage developmental assay, after 4 days, these 6 populations showed a clear range of T-lineage developmental speeds (Fig. 6c-d). The most advanced population repeatedly appeared to be $Ly6d^{+}$ $Flt3^{-}$ ETPs (pop. 6, approximately late substage B, Fig. 5f), and the least advanced population, the $Flt3^{+}$ $Ly6d^{-}$ HSA^{-} cells (pop. 1), in good agreement with the single-cell pseudotime trajectory model. In tests of non-T lineage potential using co-culture without Notch ligand, $Flt3^{+}$ cells (pops. 1-3) differentiated readily into dendritic cells (DCs), macrophages, natural killer cells (NKs), and some granulocytes, as expected for uncommitted precursors. However, despite their association with higher expression of myeloid-affiliated genes *Spi1, Hhex, Tyrobp*, and *Mpo*, all the $Flt3^{-}$ subpopulations (pops. 4-6) revealed less potential to give rise to DCs and macrophages than the $Flt3^{+}$ ones, although similar to $Flt3^{+}$ ETPs in their output of NKs (Fig. 6e). This agreed with the different outputs of $Flt3^{+}$ and $Flt3^{-}$ ETP subsets when myeloid potential was promoted with alternative cytokines, omitting Flt3 ligand

(Fig. S5b). Thus, potential towards DC and macrophage development is reduced, not increased, in ETPs when they turn off *Flt3*.

Finally, to determine whether the developmental potentials of individual cells truly match the transcriptome features of the pseudotime model, we repeated this experiment at the clonal level. First, we determined the distribution of developmental states from Bcl11b^{-} DN2a to Bcl11b^{+} DN2a to DN2b, within clones generated by single precursors from sorted ETP subsets 1-6 (Fig. 6f). The results showed that nearly all cells in clones from all subsets of input cells had crossed the ETP-DN2 boundary in five days (Fig. 6f, top). In accord with the sorted bulk population results (Fig. 6d), clones seeded by precursors from subsets 1 and 2 were slower than the rest and those seeded from subsets 5 and 6 were faster than the rest at turning on Bcl11b-YFP and progressing to DN2b (Fig. 6f, middle, bottom). However, despite these differences, >75% of the individual subset 3 and 4 precursors generated clones in which at least 30%-50% of the cells had turned on Bcl11b-YFP by day 5 (Fig. 6f), confirming the T-lineage potential of the founders. To determine how homogeneously the transcriptomes of these sorted subsets were actually distributed in pseudotime, at single-cell level, we used Cell Hashing for scRNA-seq of 5 populations simultaneously, combining barcoded antibodies with 10X analysis (Stoeckius et al., 2018)(Fig.6b, also see Methods). Purified ETP subsets 1, 3, 4, and 6 and a reference ETP-DN3 population were labeled and pooled for 10x single-cell transcriptome analysis. A new DDRTree and a new 'ETP-enriched' pseudotime analysis were calculated from the results (Fig. 6g,h, Fig. S8e), and the distinct subset features were deconvolved from the data by sample cell hashing barcode. The separation and spread of the clonal developmental assay and the transcriptomic pseudotime profiles of precursors from sorted gates were in good agreement. Cells in the Flt3^{-} subsets 4 and 6 resolved to different pseudotime positions particularly well, and both subsets were distinct from subsets 1 and 3 (Fig. 6g,h). Fig. S8e confirms that their enhanced T-lineage differentiation relative to subsets 1 and 3 was indeed correlated with their higher expression of "non-T" genes *Spi1, Hhex, Cd7*, *Mpo,* and *Tyrobp*, as predicted by RNA velocity results.

These results thus confirm that ETPs advance toward T-lineage progression as they turn off *Flt3*, but that strong multipotency regulators and non-T markers are transiently elevated in these cells relative to earlier T-cell precursors. This result favors the multilineage-priming model and indicates that the transient upregulation of these "non-T" genes is an integral feature of the early T-cell developmental program.

DISCUSSION

The T-lineage commitment transition has been much studied, but the events leading up to commitment have been poorly understood until now. Here, we have dissected the gene regulatory changes and associated developmental potentials during this process, encompassing ETP to DN2a stages, at the single-cell level (Fig. S1a), with results summarized in Figure 7. This analysis has provided evidence for an ordered sequence of at least three transient regulatory states leading toward T-lineage commitment. Evidence that these transient states are truly within the T-cell developmental progression and not representing cells of different lineages comes from the high T-lineage precursor frequency in the starting ETP population, from the relative differentiation kinetics of the candidate intermediate populations, and from the robust coexpression of T-lineage specification TFs (*Tcf7, Gata3*) together with genes specific for the intermediate states within individual cells. This study thus provides insight into gene expression dynamics of the earliest T-cell precursors, essential for more accurate modeling of the underlying T-cell specification gene regulatory network.

The results of this study were greatly strengthened by the complementary contributions from three single-cell transcriptome analysis approaches. Genome-wide transcriptome profiles based on 10X Chromium droplet-based sequencing had to be supplemented with highly sensitive seqFISH measurements to obtain accurate relationships between regulatory genes expressed in the same cells, while deep sequencing of a smaller number of cells with C1-SmartSeq2 provided full-transcript corroboration. We validated the biological predictions of the pseudotime trajectory using primary cell culturing assays to test directly the T and other lineage differentiation potentials among sub-populations of ETPs. The pseudotime model of gene expression dynamics in early T-cell differentiation was also consistent with recent empirical knock-out studies of known regulatory factors, PU.1 (*Spi1*) and Bcl11b (Hosokawa et al., 2018a, 2018b; Ungerbäck et al., 2018), which activate and repress target genes that cluster appropriately relative to Bcl11b and PU.1 expression changes.

Transcriptome clustering and RNA velocity analyses indicated that developmental progression could be initially linked with cell cycle control in ETPs, later becoming cell cycle-unrestricted in DN2s. Through RNA velocity and pseudotime analysis, we identified the most likely phenotype of the immediate DN2 precursors within the ETP population. Notably, these cells were particularly enriched for expression of *Spi1* and other genes that are not specific for the T-cell pathway, supporting multilineage priming. This population was distinct from an outgroup of granulocyte-committed precursors found in every population of ETPs analyzed. Finally, primitive populations of ETPs with unusually high *Hoxa9* transcription were detectable by cell cycle and distinctive regulatory gene expression velocity (Fig. S6e), and could represent an ETP subset with augmented self-renewal potential.

Using seqFISH and C1 data, we showed that within the ETP state the majority of individual cells co-express legacy progenitor genes with the critical Notch-induced T-cell regulatory genes, *Gata3* and *Tcf7*. This demonstrates rigorously that intra-thymic Notch signaling does not immediately shut down expression of stem and progenitor genes, even as it turns on T-cell genes, and that the two regulatory networks operate together in the same cells throughout ETP and even into DN2 stages, implying timescales of days (Kueh et al., 2016). This also suggests the possibility of crossover regulatory network connections, which remain to be determined but may help to explain the observed transient regulatory states. Previous studies suggested that hematopoietic stem cells (HSCs) maintain low-level expression of lineage-associated genes to stay poised for multilineage blood production while balancing self-renewal and differentiation, a state termed multilineage priming (Hu et al., 1997; Mercer et al., 2011; Orkin, 2003; van Galen et al., 2014). Seemingly-overlapping patterns of expression of *Spi1, Bcl11a, Cebpa* and T-cell specification genes at the population level have been suggested to explain the persistence of multilineage differentiation potential in ETP-DN2a cells under conditions of Notch withdrawal (Del Real and Rothenberg, 2013; Franco et al., 2006; Kueh et al., 2016; Laiosa et al., 2006; Wang et al., 2014; Yui et al., 2010), but this has previously been a hypothesis. The results shown here are the first to demonstrate this co-expression in individual ETPs.

Furthermore, in ETPs, even some "effector" genes representative of non-T cell lineages, such as *Mpo,* were also robustly co-expressed with *Gata3* and *Tcf7* at the single-cell level, in populations showing a high T-lineage precursor frequency; the seqFISH data ruled out possible doublets. This pattern of coexpression strongly supports multilineage priming in many individual ETP (and even DN2a) cells rather than contamination with cells lacking T-lineage potential.

In summary, we have established a detailed model of single-cell transcriptome dynamics during the transition from multipotentiality to T-cell lineage commitment, with single-cell sequencing tools, bolstered by highly sensitive seqFISH analysis, and supported by *in vitro* differentiation kinetics and the results of acute transcription factor perturbation studies. This study provides new potential regulatory steps to explore and validate. For the first time, the complexity and regulatory substructure within the first phase of T-cell development can be perceived.

ACKNOWLEDGMENTS

We thank Jeff Park, Paul Rivaud and Sisi Chen from the Caltech Single Cell Profiling and Engineering Center for help with the 10X Genomics samples, Andres Collazo and the Biological Imaging facility of Caltech for clonal live imaging support, Sean Upchurch and Diane Trout for C1 bioinformatic support, and members of the Rothenberg, Wold, and Cai labs for advice. We also thank Rochelle Diamond and members of the Caltech Flow Cytometry facility for sorting, Ingrid Soto for mouse care, Igor Antoshechkin and Vijaya Kumar of the Caltech Jacobs Genomics Facility and Xiwei Wu and the Integrative Genomic Core of City of Hope for Smartseq2 and bulk RNA sequencing. Support for this project came from USPHS grants (R01HL119102 and R01HD076915) to E.V.R., The Beckman Institute at Caltech for support of all the Caltech facilities, the Biology and Biological Engineering Division Bowes Leadership Chair Fund, the Louis A. Garfinkle Memorial Laboratory Fund, the Al Sherman Foundation, and the Albert Billings Ruddock Professorship to E.V.R.

MAIN FIGURES

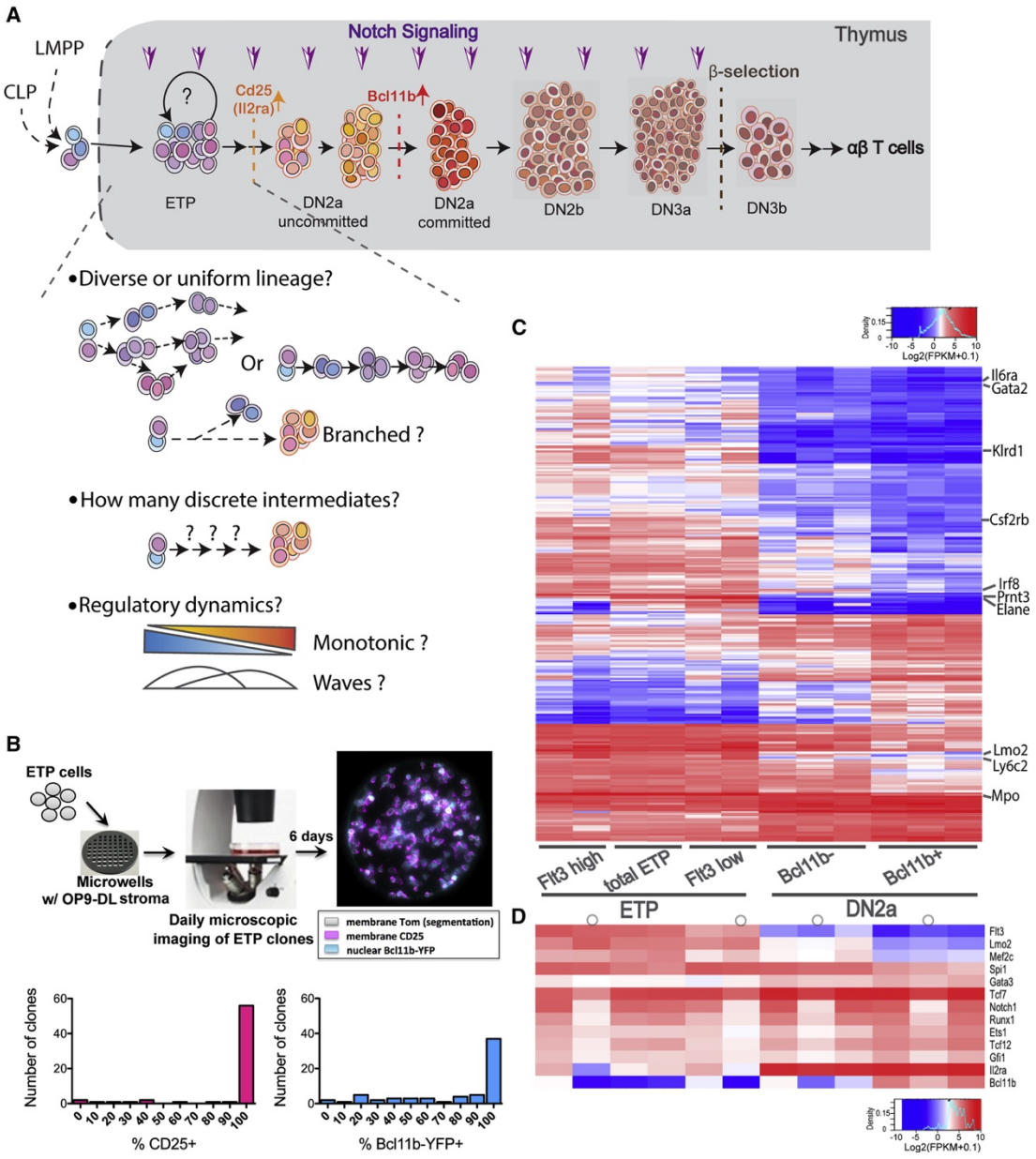


Figure1. High T-cell precursor frequency in ETP cells and bulk population gene expression comparison with DN2a cells. A) Schematics of early T-cell developmental stages, checkpoints, associated key developmental markers, and previously unresolved questions addressed in this study. B) Diagram of clonal culture and imaging methods for following the development of individual sorted ETP cells and a representative false color

image of the progeny of an ETP clone (top). Histogram plots showing the numbers of ETP clones with different percentages of CD25+ (magenta) or Bcl11b+ (cyan) cells on day 6 of culture (n = 66 viable clones) (bottom). c-d) Heatmaps of bulk RNAseq measurements on $Flt3^+$ and $Flt3^-$ ETP and $Bcl11b^-$ (uncommitted) and $Bcl11b^+$ (committed) DN2a sorted populations. Color scales indicate raw expression levels as log(FPKM+0.1), without row normalization. Some samples were sequenced with pre-amplification, indicated (o) (see Methods). C) Clustered expression heatmap of bulk RNAseq measurements for genes differentially expressed between all ETP and committed $Bcl11b^+$ DN2a cells ($n \geq 3$, adj. pval<0.05, fold change ≥ 2 either way, also see Table S1). Representative non-T or stem/progenitor genes are labeled. D) Selected key genes involved in T development, on the same populations as in (c).

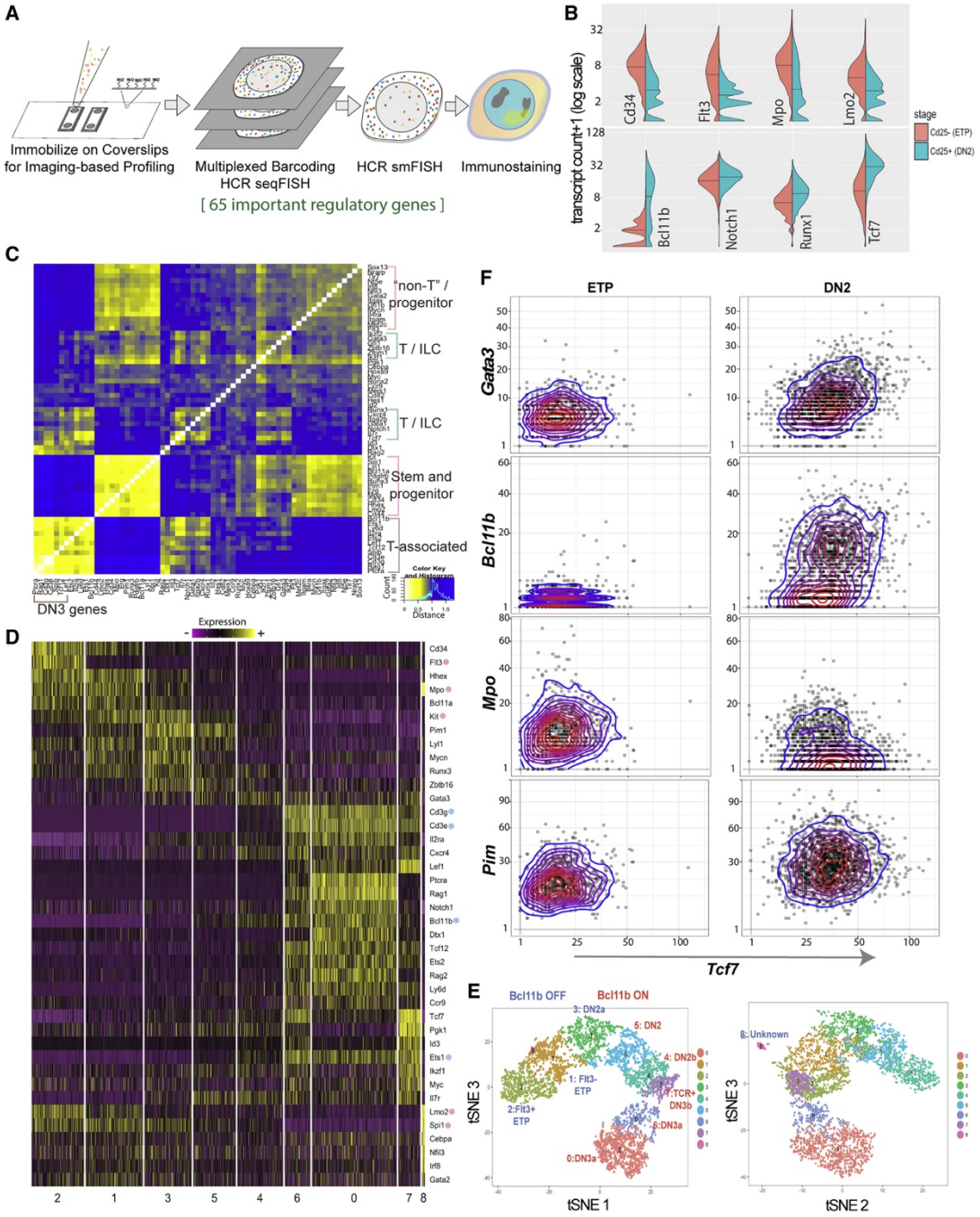


Figure 2 High sensitivity measurement and coexpression of key regulatory genes in single early pro-T cells using seqFISH. A) Experimental design for seqFISH analysis with FACS enriched cells. B) Transcript distributions of genes in thymic ETP (cKit^{high}, *Kit* transcript ≥5, *Il2ra* transcript ≤3, N=890) and DN2 (cKit^{high}, *Kit* transcript ≥5, *Il2ra* transcript >3, n=1984) cells, in cells from 4 week-old-animals as detected by seqFISH. C)

Gene-Gene Pearson distance heatmap of co-expression of genes measured based on 2963 ETP-DN2 cells plus 1587 DN3 cells. D-E) Clustering analysis of seqFISH data for 4550 cells across ETP-DN3 stages. The Smart Local Moving (SLM) algorithm was used based on PC 1-6 of size-normalized data for 65 genes. Heatmap of genes enriched in expression in each sub-cluster, ordered based on connectivity in tSNE and reflecting developmental progression (Wilcoxon rank sum test with threshold of 0.2 and minimum fraction of expressing cells ≥0.2 using Seurat 2). E) Annotated tSNE display generated using PC1-6, colored by clusters. F) Pair-wise scatter plots, overlaid with color-coded density contours, of copy numbers of transcripts for *Tcf7* against those of T-specification genes *Gata3* and *Bcl11b* and of "non-T" gene *Mpo* and growth-control gene *Pim1*. ETP and DN2 cells are defined as in (B), displayed on sqrt+1 scale.

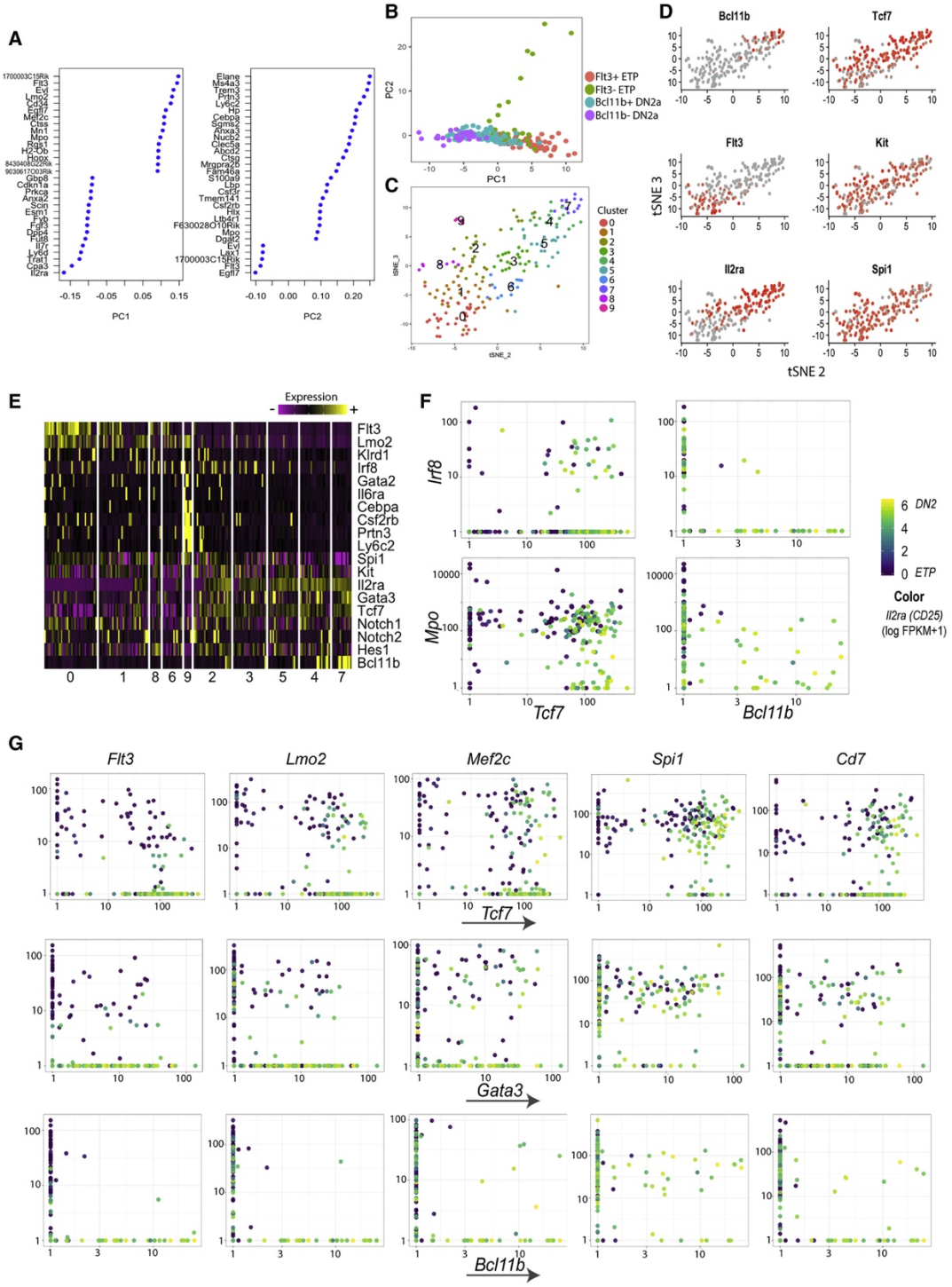


Figure 3. Semi-supervised C1 Fluidigm (C1) analysis of single cells in the ETP-DN2a developmental continuum supports co-expression hierarchy of T-lineage and progenitor-associated genes. A) Principal component (PC) loading of first 2 PCs

of the analysis based on genes that are differentially expressed in bulk RNAseq shown in Table S1. B) PC1-2 display of 193 cells measured by C1, colored by stage categorization of Flt3, Il2ra (ETP vs. DN2a), and Bcl11b positivity. C) tSNE display of C1 data with SLM clusters color projected. Both tSNE and clustering with SLM were performed with PC 1-10. D) tSNE display with expression patterns of specific genes as indicated overlaid in red. E) Heatmap of expression patterns of selected genes ('non-T' genes and 'T-associated' genes). The clusters are ordered by approximate T developmental order, according to C) and D). Also see Table S4 for the list of feature genes that are enriched in individual clusters. F) Bi-plots of expression patterns of two non-T lineage markers *Irf8* and *Mpo*, against T-specification genes *Tcf7* and *Bcl11b*, showing the pattern of overlap of *Mpo* and both T-specification genes. *Irf8*, on the other hand, overlaps with early T-specification gene, *Tcf7*, but minimally with *Bcl11b*, which is expressed at a later stage. The dots are colored by expression of Il2ra (CD25) on a log transformed color scale. G) Co-expression patterns of stem and progenitor genes and T-specification genes *Tcf7, Gata3* and *Bcl11b*. n= 228 total cells measured, n= 193 cells were shown in this figure after filtering for single cells with a minimum of 3600 genes and a mitochondrial gene fraction under 0.11.

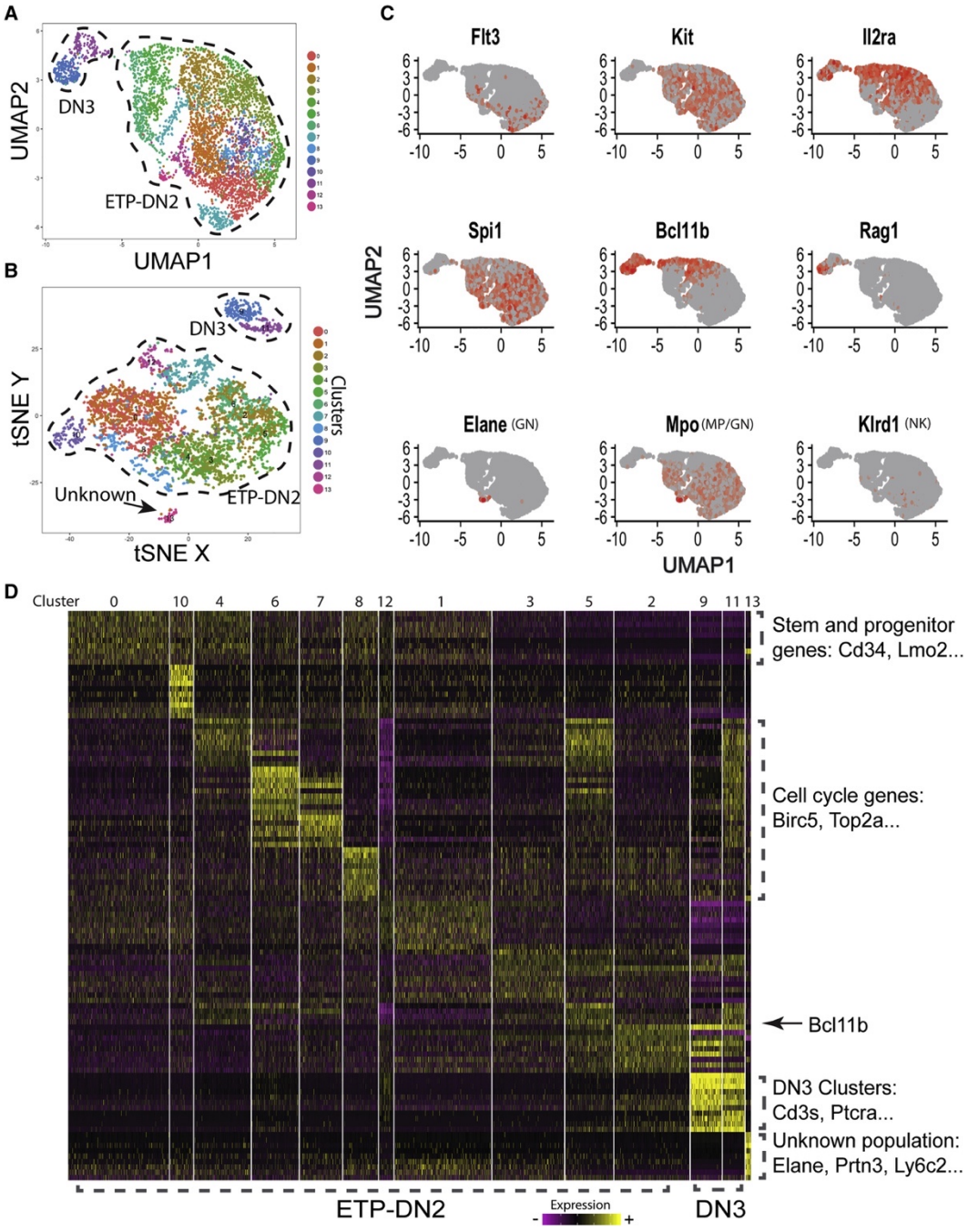


Figure 4. A dense developmental continuum of gene expression in early DN pro-T cells based on 10x Chromium scRNA-seq analysis. A-B) UMAP(A) and tSNE(B) displays of 10X Chromium data, colored by sub-clusters. Clustering performed with SLM algorithm using PC1-10. C) UMAP display with expression patterns of genes that characterize different developmental stages (*Flt3, Kit, Il2ra, Spi1, Bcl11b, Rag1*) or different lineages

[*Elane* (granulocytes, GN), *Mpo* (macrophages, MP), *Klrd1* (NK cells, NK)] overlaid in red. D) Heatmap displaying the top 10 enriched genes in each sub-cluster ordered by approximate developmental progression based on gene expression and connectivity in low dimensional displays. (Seurat 2 pipeline with minimum fraction of expressing cells ≥0.2, Wilcoxon rank sum test with threshold of 0.2; see Table S4). n=4627 cells: ~90% ETP-DN2 and ~10% DN3 cells.

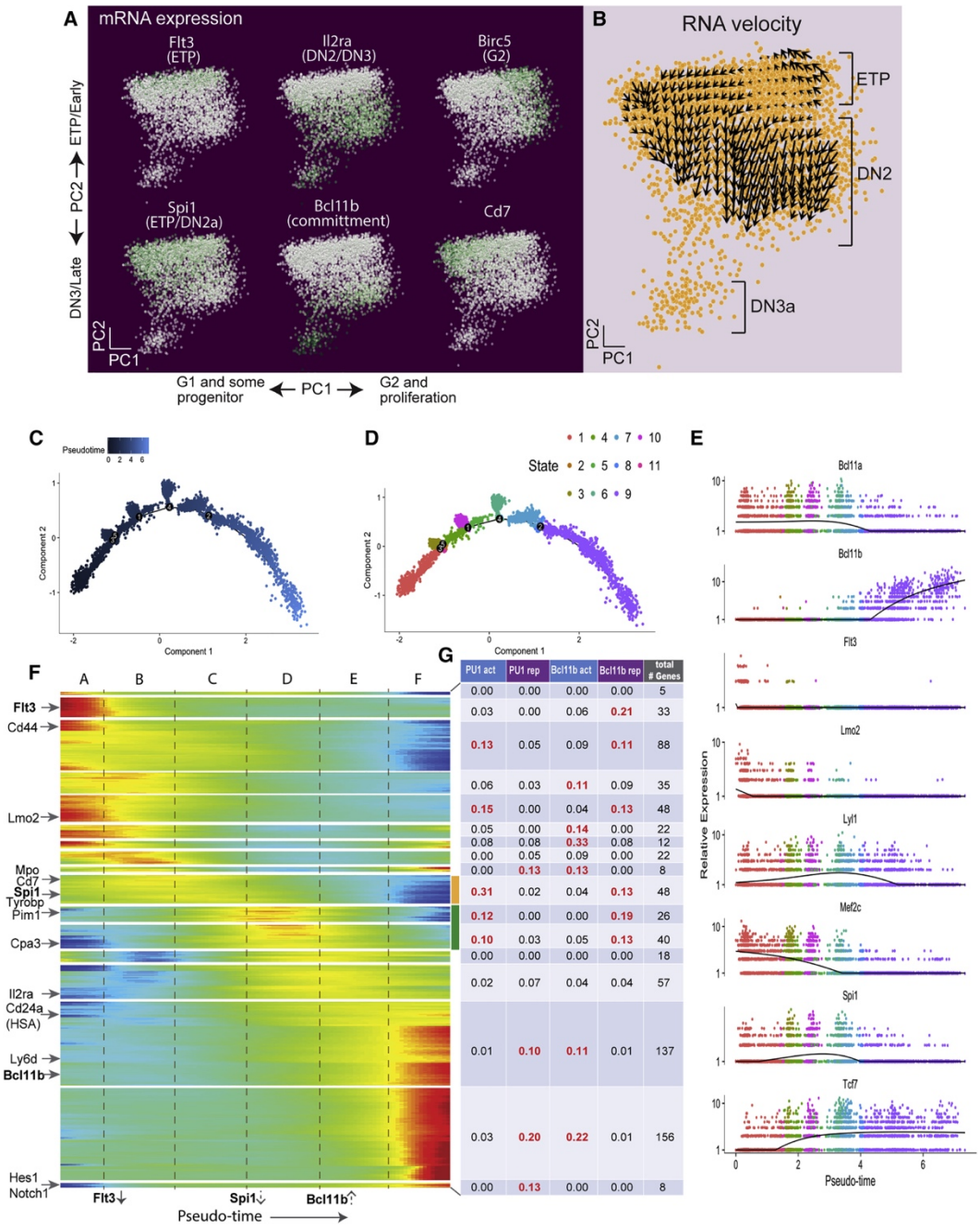


Figure 5. Stage ordering by RNA velocity and pseudotime modeling from supervised analysis of 10X scRNA-seq data: evidence for gene expression waves during early T-cell differentiation. A-B) RNA velocity analysis on trimmed data using Velocyto (excluding granulocyte precursor and DN3b clusters). A) mRNA expression patterns for key genes

on PC1-2: higher expression, darker green. B) Grid arrows indicating relative transition probabilities based on un-spliced/spliced transcript calculations (imputation with k = 90, displayed on PC1-2) using Velocyto. Also see Fig. S7. C-D) DDRtree display analyzed with Monocle 2 and based on the curated instructive gene list (Table S2), overlaid with pseudo-time staging (C), and branching state (D). Granulocyte precursor and DN3b clusters excluded, n=4438 cells. E-F) Gene expression patterns along pseudo-time. E) Relative expression patterns of representative regulatory genes across pseudo-time, colored by DDRtree 'state' (legend in (D)). Also see Fig.S8C. F) Clustered expression heatmap of 763 genes that are differentially expressed along the pseudo-time (Monocle 2, with qval<10^{-8}, in both biological replicates). Red= high expression level, blue = low expression level, on a relative scale normalized to each gene. Dashed vertical lines are positioned for descriptive purposes, hierarchical clustering based on expression using the "complete" method. G) Summary table of fractions of pseudotime-differentially expressed genes in each cluster that overlap with regulatory targets activated (act) or repressed (rep) by key TFs PU.1 and Bcl11b in perturbation assays, and the total number of genes in each cluster. Also see Table S5. Red font highlights fractions above 10% (0.1).

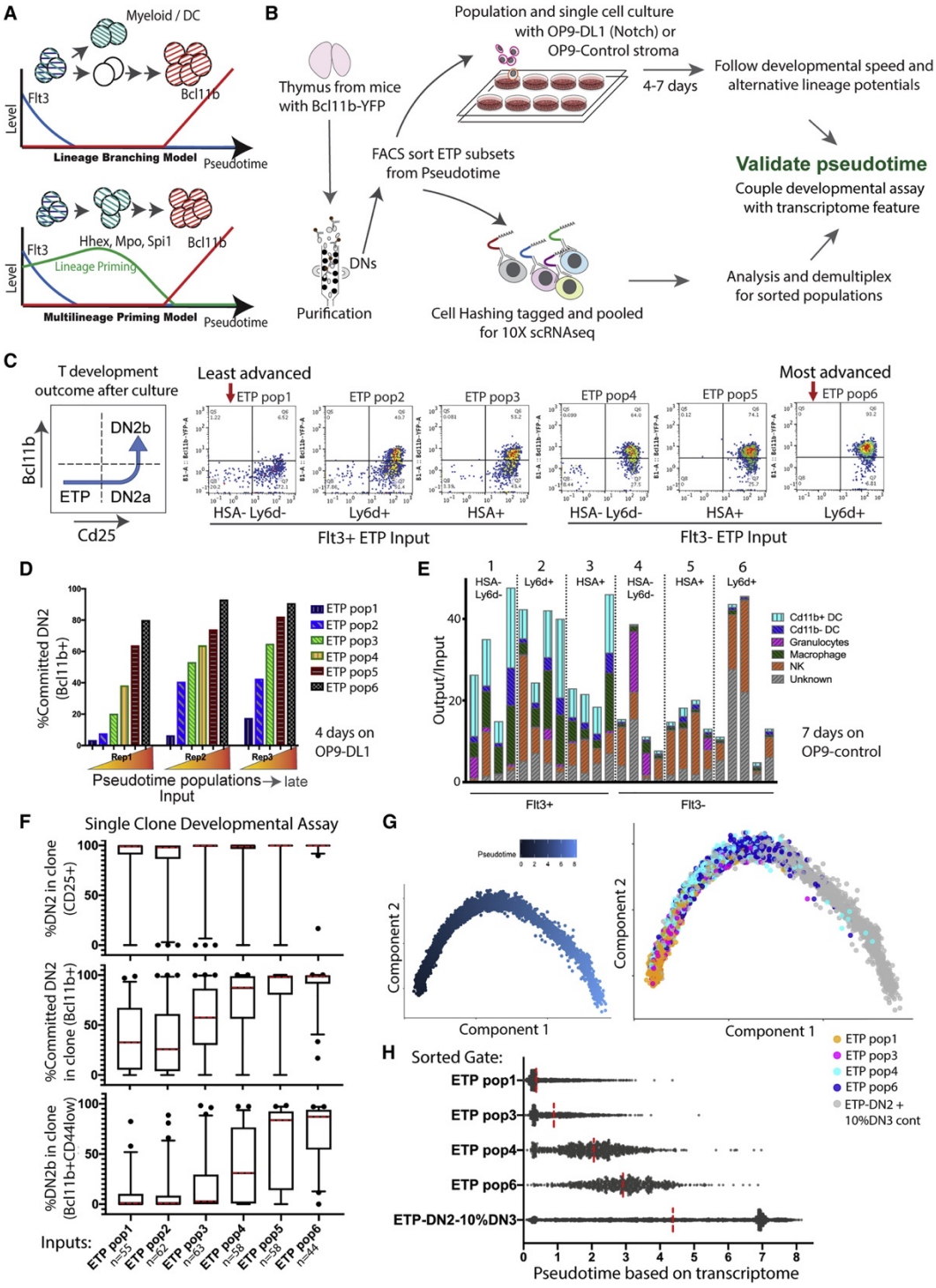


Figure 6. *In vitro* test of ETP developmental staging favors a multilineage-priming model for gene expression waves. A) Diagram of two hypotheses to explain the branch or early wave patterns observed in the DDRtree and pseudotime analyses. B) Diagram of the *in*

vitro developmental culture assays and ETP subset scRNA-seq setups. C-E) ETPs (stages A and B, Fig 6D) were subdivided into 6 populations according to surface markers Flt3, HSA, and Ly6d, and analyzed for their developmental progression after 4-7 days. C) Representative flow cytometry plots of the development of sorted ETP populations after 4 days of culture on OP9-DL1. D) Bar-graphs showing the fraction of committed T cells (measured by Bcl11b-YFP upregulation) after 4 days in OP9-DL1 culture, ordered according to the pseudo-time pattern. (n=3 independent biological replicates, 3rd replicate (Rep3) an average of 2 technical replicates.) E) non-T lineage potential of individual sorted populations after 7d of culture on OP9-Control (no Notch ligand, non-T conditions) with lymphoid supporting cytokines. n≥4. F) Summary plots of percentage of cells passing T-developmental milestones in individual clones from individual FACS sorted precursors (gates same as in (C)-(E)) cultured 5 days on OP9-DL1. Whiskers represent 5-95 percentiles. n=55, 62, 63, 58, 58, 44 live clones in ETP pop1 through 6, respectively. G-H) Reconstructed transcriptome single-cell pseudotime trajectory with 4 ETP subsets (pops 1, 3, 4, 6 from (C)-(F)) and an ETP-DN3 control group tagged with antibody barcodes. G) DDRtree with pseudotime coloring and highlighted ETP subsets. H) Pseudotime distribution of individual cells from the 4 sorted subpopulations. (Analyzed with Monocle 2 and based on the curated instructive gene list). n= 1333, 1144, 1044, 823, 3172 cells in ETP pop1, 3, 4, 6 and control, respectively.

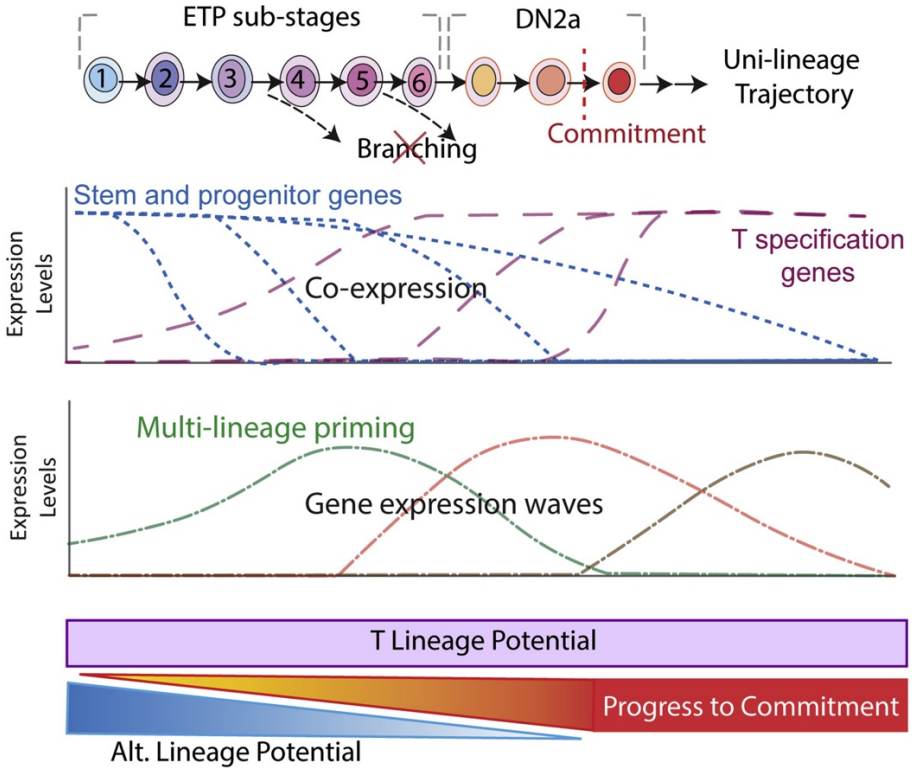
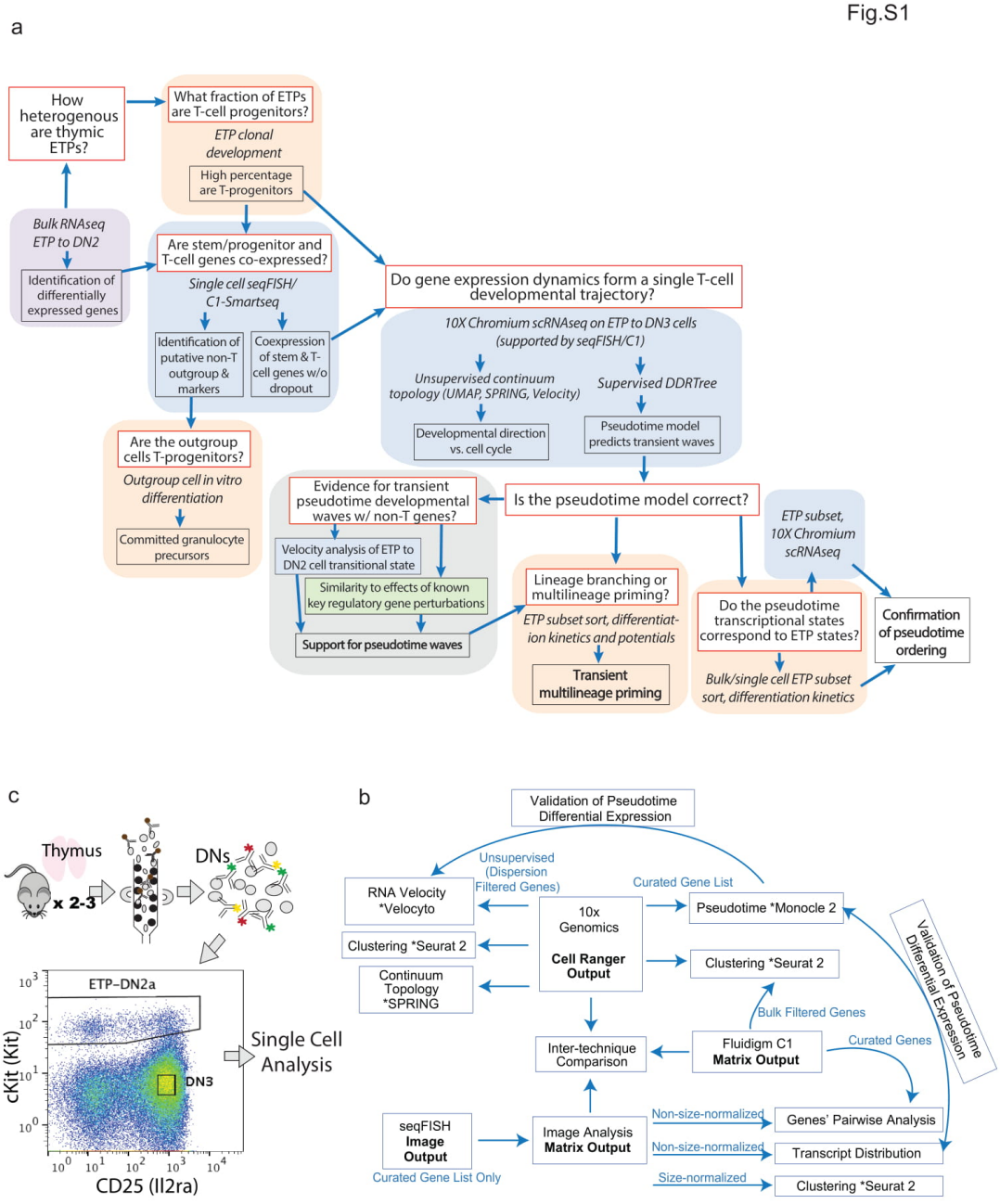


Figure 7. Summary of key findings in this study.

Data imply sequential sub-stages within the ETP compartment before transition to DN2a, not only marked by asynchronous downregulation of progenitor genes but also by transient activation of gene waves as the cells progress toward commitment. The frequency of T-lineage potential is very high in ETPs overall, and although some transiently activated genes are otherwise associated with non-T fates (multilineage priming), alternative lineage potential in pro-T cells decreases monotonically as the cells progress from $Flt3^{+}$ ETP to $Flt3^{-}$ ETP to DN2a to commitment.

SUPPLEMENTARY FIGURES

Fig.S1

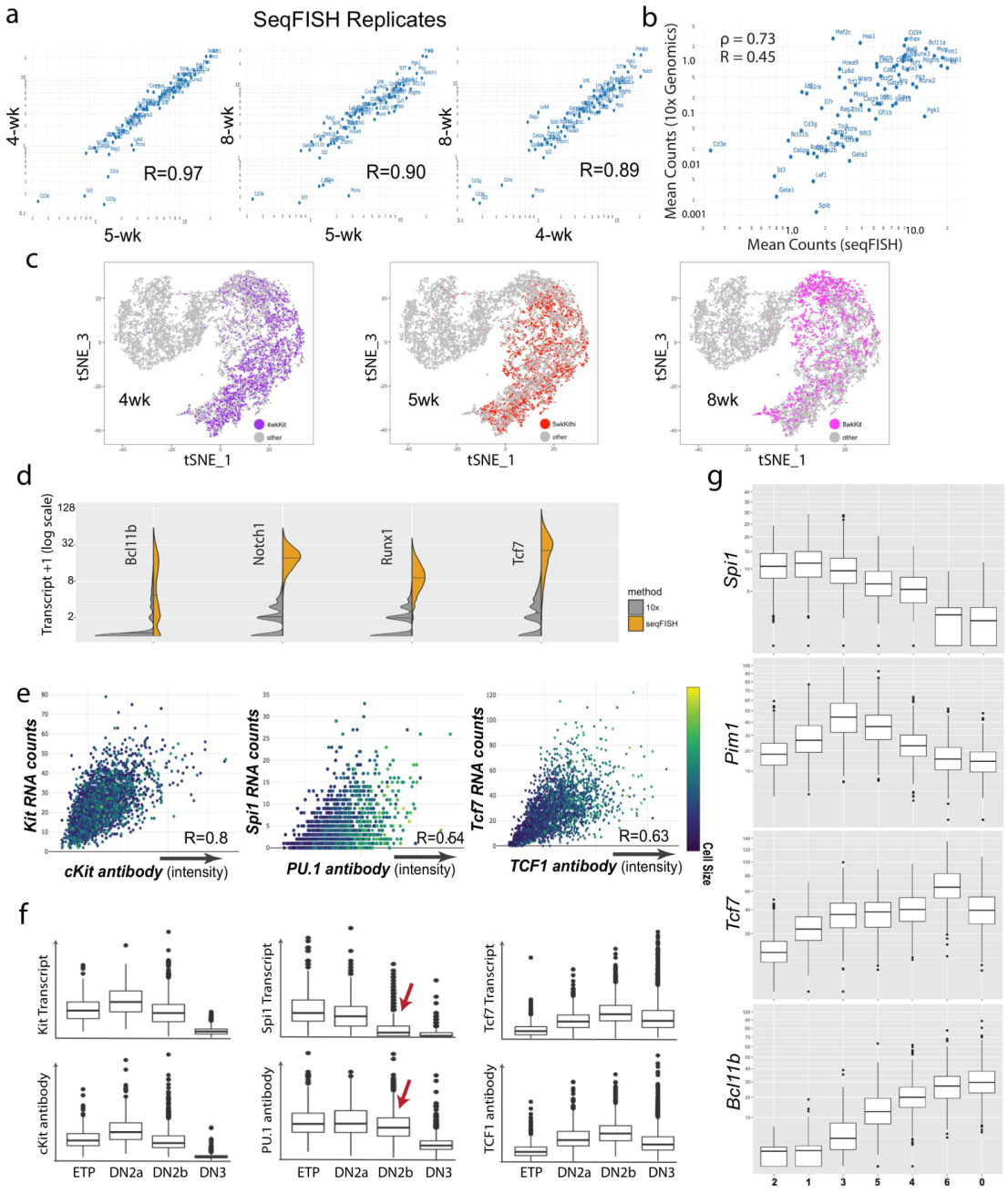


Supplementary Figure S1, related to Figs. 1, 2, 3, 4, 5, 6, and 7. Summary schematics of biological questions addressed, and analysis pipelines used.

a) Summary: logic flow of central biological questions in this study, how each step provides the rationale for the next, and breakdown of specific technologies and

analyses used to address the specific questions. Questions are highlighted in red boxes, and results are shaded in gray boxes. Techniques and analysis used are described in italic text and colored in background with blue shading indicating analysis using single-cell transcription profiling tools, purple shading indicating bulk RNA analysis, and orange shading indicating cell culturing assays. b) Summarizes relationships between methods, gene and cell filters being used, and data analytical pipelines used in this study. c) Sorting gates and logistics for purifying Kit^{hi} ETP-DN2a and DN3 cells for single-cell analyses.

Fig. S2



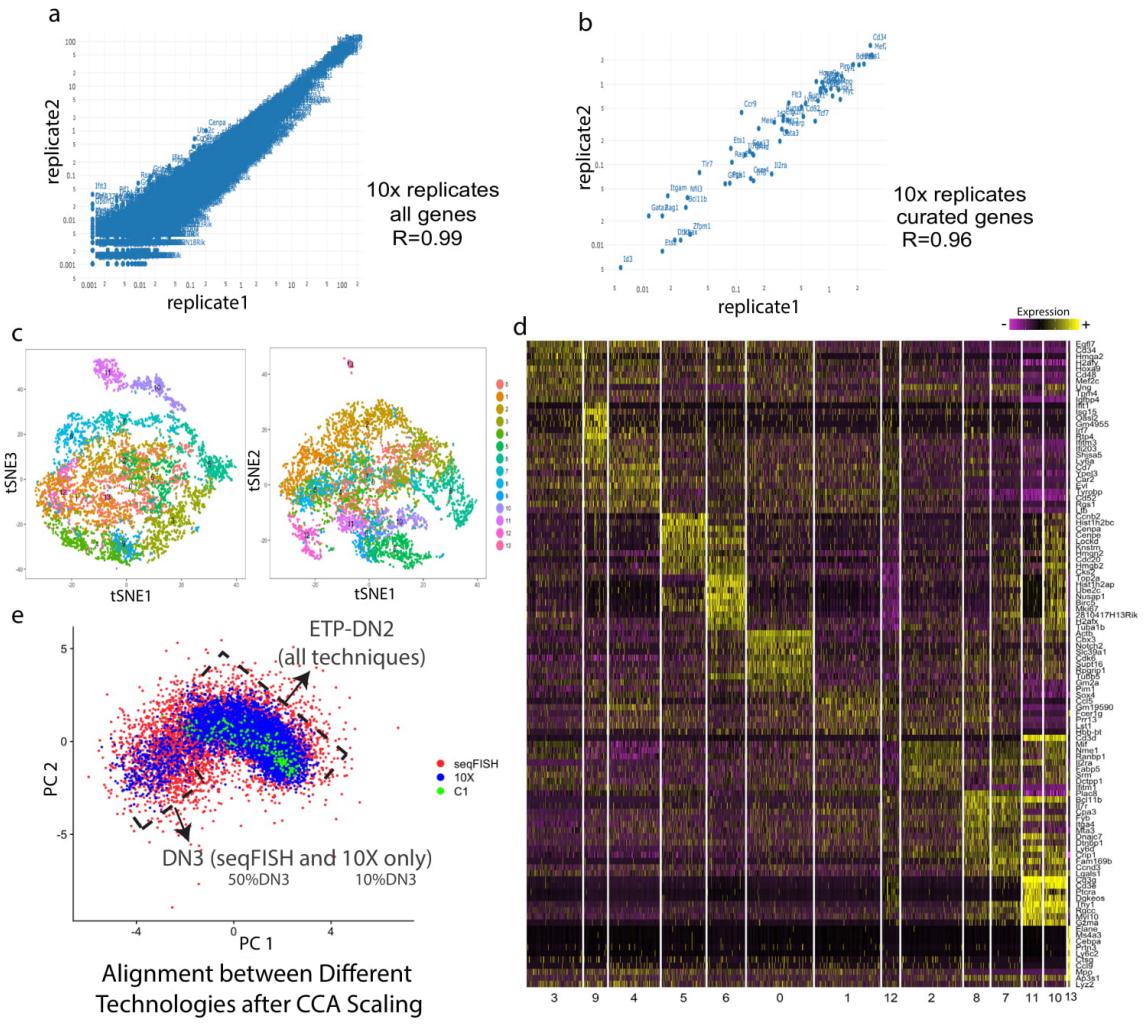
Supplementary Figure S2, related to Fig. 2. Highly sensitive seqFISH provides reproducible and robust RNA transcript quantitation for regulatory genes.

a) Scatterplot comparison between mean values of expression measured in a comparable population (Flt3+ ETP) in different seqFISH experiments with

thymocytes from 4, 5, 8-week old animals. b) Scatterplot comparison between mean values of expression, measured in seqFISH and 10X Chromium, of genes listed in Table S2. Mean values taken from cells in comparable cell populations (Flt3+ ETPs). Patterns were broadly correlated, but seqFISH detected approximately 10 molecules of RNA for each UMI count in the 10X analysis (seqFISH: n=1656, from 3 replicates. 10X: n=863, from replicate1). The result is consistent with the previously described 10% sampling rate of 10X Chromium v2 scRNA-seq, at the sequencing depth being used (Islam et al., 2014; Kolodziejczyk et al., 2015) c) tSNE plots with combined 3 seqFISH replicates, including Kit positive and DN3 populations, colored in each panel to indicate the distribution of Kit positive cells from one of the individual replicates. The samples from different experiments and ages are interspersed without batch corrections. d) Detected transcript distribution comparison between seqFISH and 10X Chromium experiment on key regulatory genes in ETP-DN2 cells. (seqFISH: n=2524, 10X: n=4234.) SeqFISH detected the expression of *Notch1, Tcf7,* and *Runx1* in almost all ETP-DN2 cells, in agreement with their known functional roles whereas 10X had a high false negative rate. e) Scatterplots of antibody staining and RNA transcript count correlations, colored by the cell size estimation (area of image segmentation). f) Transcript and antibody distribution of cells at different stages (binned by *Il2ra, Bcl11b* transcripts). c-Kit and TCF1 agree well with *Kit* and *Tcf7* expression at all stages. Note that arrows indicating the PU.1 protein (encoded by *Spi1*) and *Spi1* RNA disagree at DN2b stage, as *Spi1* RNA drops in expression between DN2a-b stage while PU.1 protein appears to persist longer. This is likely a reflection of the extreme stability of PU.1 protein, as reported previously (Kueh et al., 2013). The antibody signals were plotted in arbitrary units on linear scales, with signal quantitation described in Methods. g) Developmentally ordered clusters of seqFISH transcript distribution. Clusters shown here were as

presented in Figure 2f, excluding DN3b (cluster 7) and the 'outlier' myeloid cluster (cluster 8).

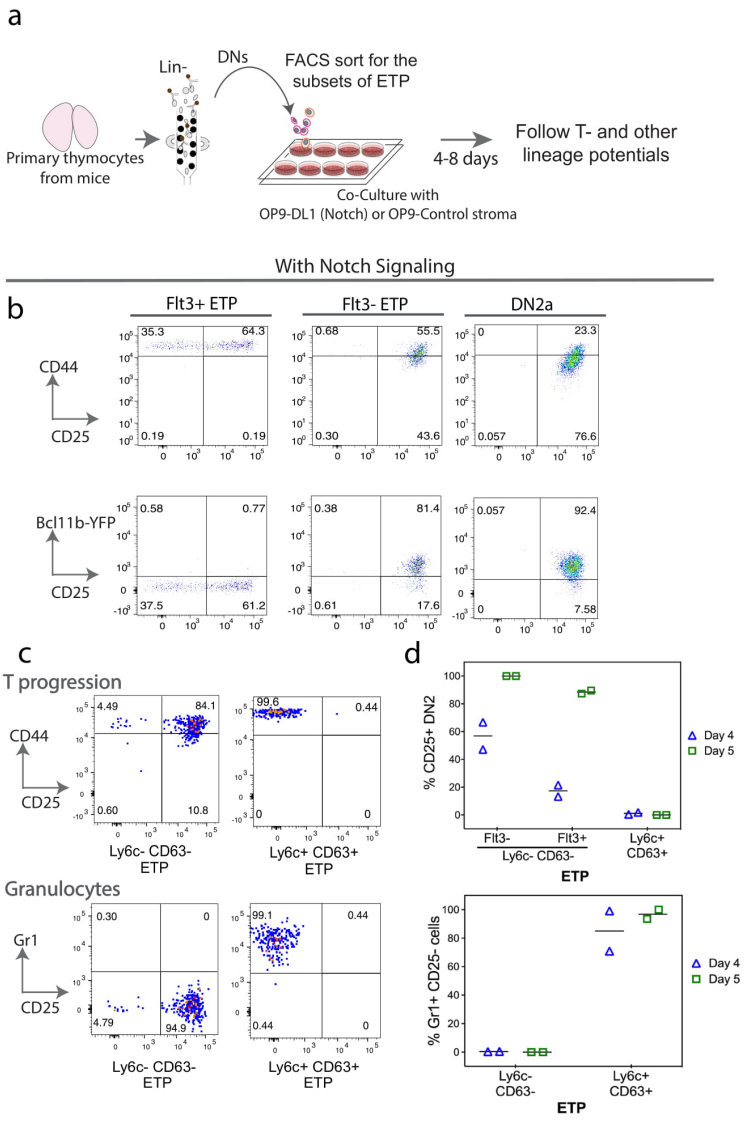
Fig. S3



Supplementary Figure S3, related to Fig. 4. 10X Chromium scRNA-seq replicates confirm the similar continuity and heterogeneity of cell states and lineage progression within the purified early T-cell population.

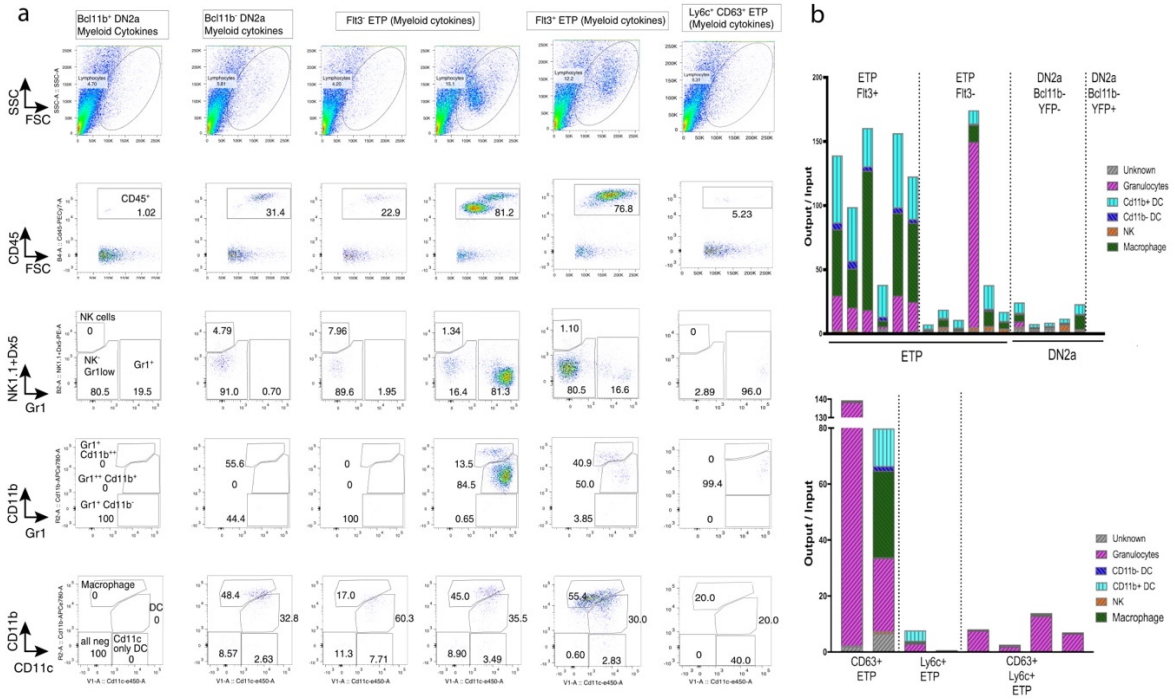
a-b) Scatterplot comparison between mean values of expression measured in comparable population (early Flt3+ ETPs) in two 10X Chromium scRNA-seq replicates (n=863 cells replicate1, n=1442 cells replicate2). c) tSNE display of 10X replicate2 (7076 cells) colored by cluster. Clustering was performed with SLM algorithm, using PC 1 to 10. d) Heatmap of feature genes enriched in each sub-cluster analyzed in 10X replicate 2, ordered by approximate developmental state. Yellow=high expression, purple=low expression. Compare with similar clustering for replicate 1, shown in Fig. 4d. e) Alignment of seqFISH, 10X, and C1 datasets after CCA scaling, shown in principal components 1-2.

Fig S4



Supplementary Figure S4 (previous page), related to Figs. 2-4. Discrete granulocyte precursor subset in the ETP compartment.

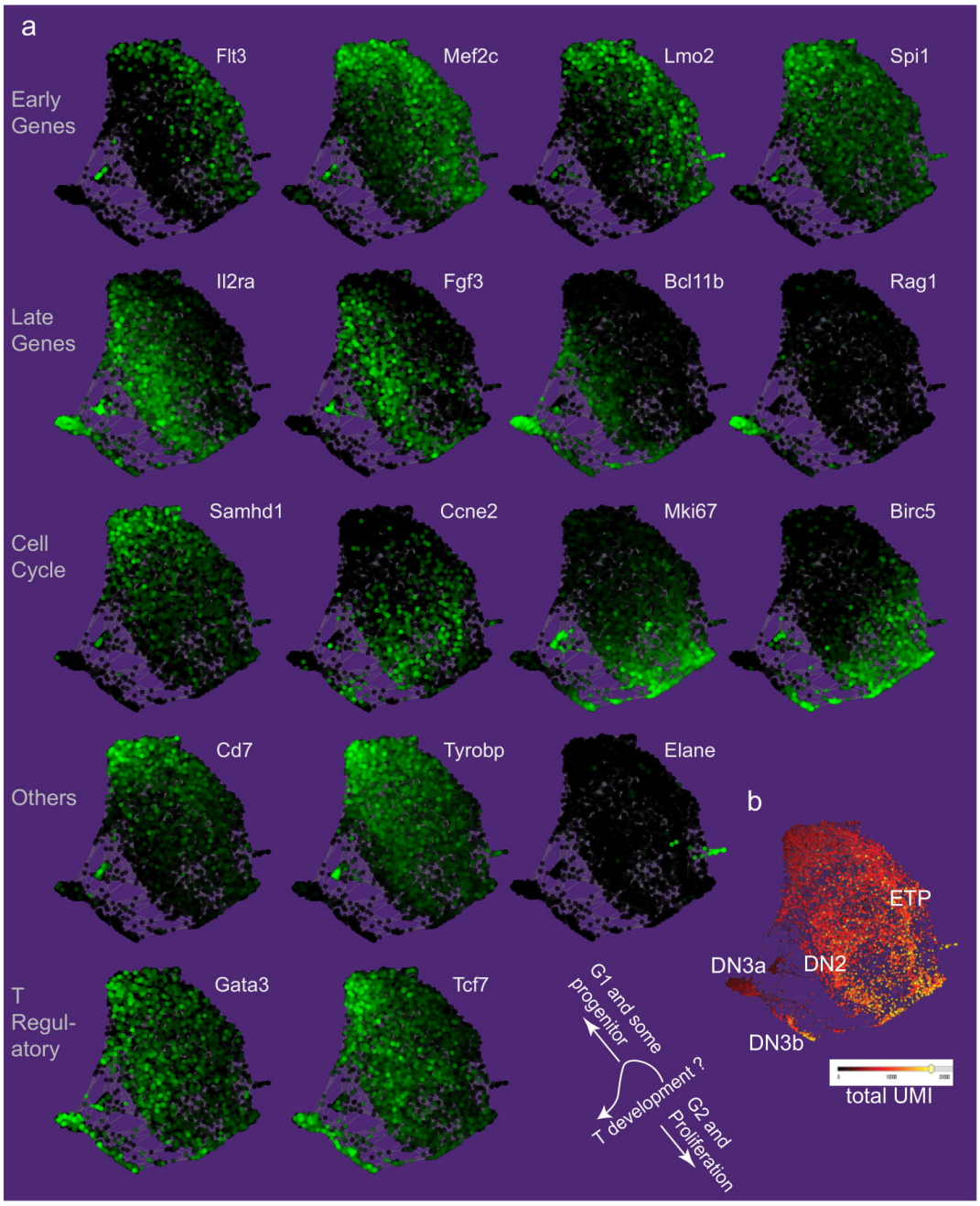
a) Experimental plan to test developmental potential. Purified Lin$^-$DN thymocyte subsets were sorted into wells (25 or 50 cells/well into 96 well plate) for co-culture with pre-plated OP9 stroma, with Notch ligands (OP9-DL1) or without, and then stained and FACs analyzed at indicated timepoints. b) Flt3$^+$, Flt3$^-$ ETPs, and DN2a cells cultured on OP9-DL1 for 4 days, then analyzed for developmental markers, CD44 and CD25, and Bcl11b-YFP. c) CD63+ Ly6c+, and CD63- Ly6c- ETP cells cultured on OP9-DL1 for 4d, then analyzed for markers of T-cell progression, CD25, and granulocytes (Gr1). d) Summary plot of percentages of CD25+ cells and Gr1+ cells after 4-5d culture with OP9-DL1 stromal cells. Thymocytes from Bcl2 transgenic mice were used to enhance cell survival. Also see Fig. S5.



Supplementary Figure S5, related to Figs. 2-4. Commitment assays under conditions lacking Notch signaling to test ETP subsets for alternative lineage potentials.

Subsets of ETPs and DN2a cells were sorted according to surface marker expression patterns indicated and tested for developmental potential under conditions favoring myeloid development. Assays were performed with cells isolated from Bcl2-tg mice to promote survival. 25 or 50 cells/well were plated into 96 well plate with pre-seeded OP9-control stroma cells, cultured for 7 days under myeloid conditions, and analyzed by flow cytometry as shown. One representative culture is shown from each subset except Flt3$^-$ ETPs, which are represented by two different cultures. Here, subsets are shown with data from the most advanced T-lineage precursors, Bcl11b-YFP$^+$ DN2a cells, which are already T-lineage committed. a) Gating strategies and representative flow cytometry analysis for alternative lineage assays at day 7 of culture. Cells were gated on FSC and SSC, 7AAD negative and CD45 positive for live lymphocytes (top two rows), and then separated by anti-NK1.1 + anti-Dx5 and anti-Gr1 for NK cells (NK) and Granulocytes (Gr1$^+$ cells) respectively (third row). The non-NK and non-Granulocyte population (lower left of panels in third row) was further separated using anti-CD11b and anti-CD11c for Macrophage (MP) and Dendritic cells (DC), respectively. The cells that were negative for all alternative-lineage markers in the staining panels were categorized as ‘unknown’. The cell numbers generated from individual categories were divided by the input cell number and displayed in stacked bar graphs in b) and also in Fig. 6e. b) Summary graphs for results of alternative lineage potential assays of ETP subsets distinguished by Flt3, CD63 and Ly6c expression, compared with DN2a cells separated into Bcl11b-YFP$^-$ and Bcl11b-YFP$^+$. Isolated cells were cultured on OP9-Control stroma, under myeloid lineage supporting cytokine conditions for 7 days (see STAR Methods). The stacked bar-graphs represent the developmental potentials of each ETP subset to generate cells of non-T lineages under these permissive conditions. Top panel shows results from Ly6c$^-$CD63$^-$ ETP cells subdivided by Flt3, and DN2a cells subdivided by Bcl11b-YFP. The bottom panel shows results from Ly6c and CD63 single and double positive ETP populations. Under these conditions, CD63$^-$ Ly6c$^-$ ETPs generated multiple types of cells of alternative non-T lineages, including

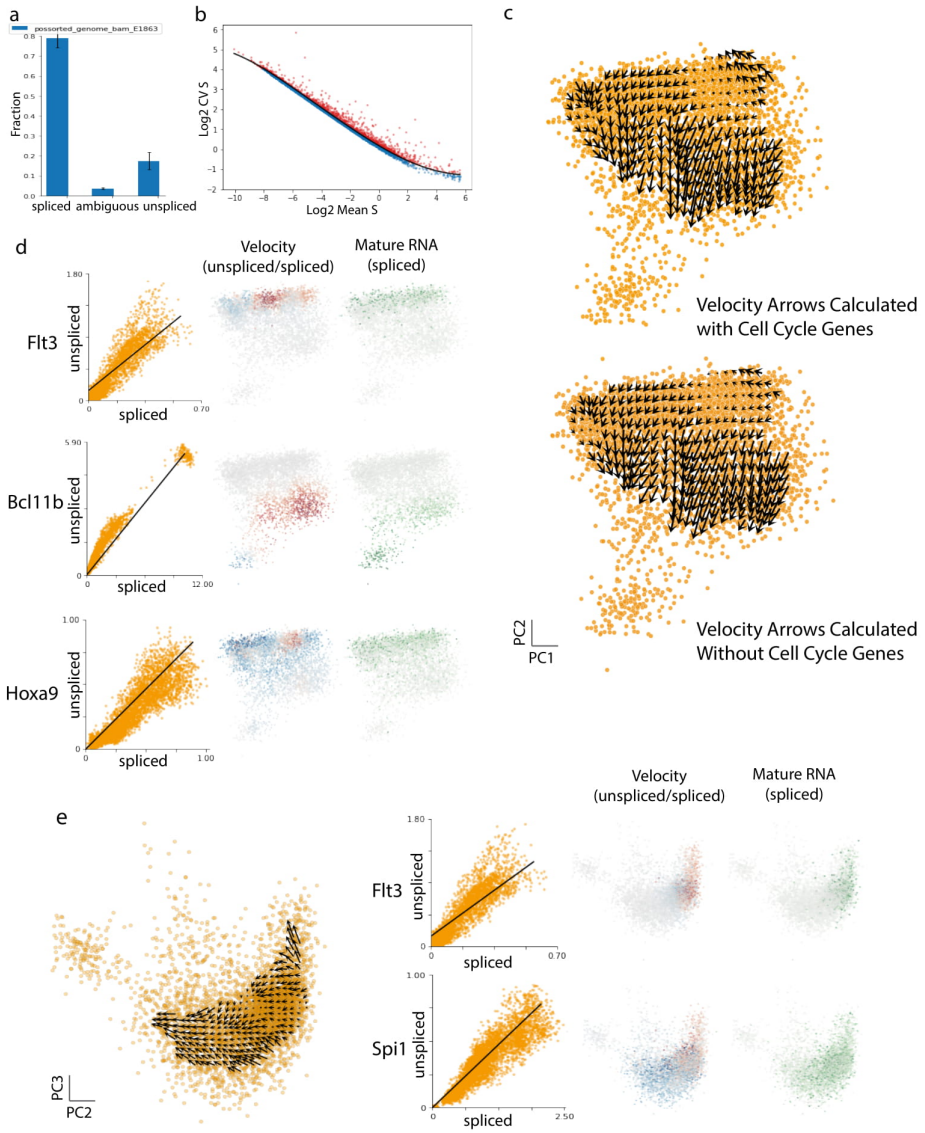
Gr1$^+$ granulocytes (magenta) (top). However, Ly6c$^+$ CD63$^+$ double positive cells gave rise exclusively to Gr1$^+$ granulocytes, while CD63$^+$ or Ly6d$^+$ single positive cells generated Gr1$^+$ cells as well as other lineages (bottom). Individual replicates are presented in separate bars.



Supplementary Figure S6 (previous page), related to Fig. 4. Developmental connectivity coupled with orthogonal spread of cell cycle signatures in SPRING analysis.

a-b) SPRING display of expression topology of ETP-DN2-DN3 cells. (Performed with PC1-20, and k=5 on raw 10x dataset with cells filtered by minimum 2500 UMI counts, but not by mitochondrial content, and genes filtered by 60th percentiles for variability) a) Expression levels of key genes are highlighted in green on relative scales. The key genes were categorized as early developmental genes (*Flt3, Lmo2*, and *Mef2c*, early ETPs; *Spi1*, all ETP and DN2a cells) and later genes (*Il2ra*, ETP to DN2a transition marker; *Fgf3*, DN2a-specific; *Bcl11b*, commitment marker; *Rag1*, upregulated in DN3a) for marking the developmental direction. The second orthogonal axis was represented by proliferative and cell cycle state markers, with G2/M-active genes *Birc5* and *Mki67* (similarly with *Top2a* and *Cenpa*, not shown*)* expressed by cells at the lower right with the highest UMI counts, with G1-to-S phase cyclin *Ccne2* immediately adjacent, and G1-expressed gene *Samhd1* concentrated at the other end). The committed granulocyte precursor population appeared as a spur (upper right) away from the main distribution (*Elane*). b) Developmental stages and axes annotated based on overall marker expression patterns, with total UMI counts displayed on red-yellow scale as shown.

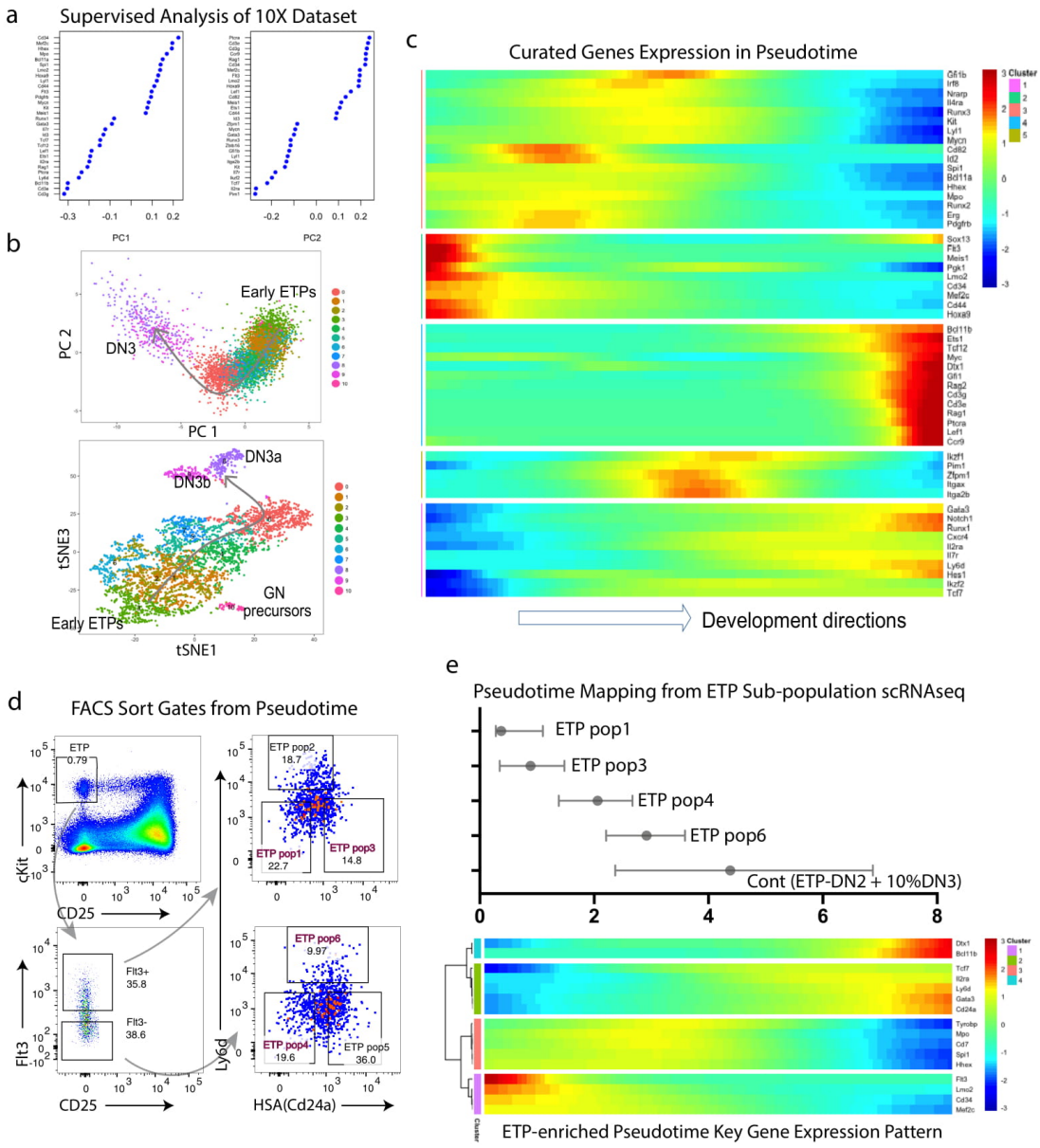
Fig. S7



Supplementary Figure S7, related to Fig. 5. Supporting analysis for RNA velocity using Velocyto.

a) Fraction of 10X Chromium reads mapped to different genomic regions: “spliced” represents exonic reads, “unspliced” represents intronic reads. b) Mean and variable filter of genes that are used in velocity analysis: red dots highlight the gene filter (‘spliced’) for PCA analysis. c) PC1-2 display with arrows indicating the transition probability of cells (imputation and transitioning probability estimation with k=90, quiver scale=0.7, scale type = “relative”.). The vector calculation was

performed with (top) and without (bottom) including cell cycle genes (as defined by gene ontology annotation (using Goatools), used in (La Manno et al., 2018)). d) Scatterplots (left panels) display the un-spliced vs spliced isoform distributions after imputation, and gamma fit of the rates of RNA processing for individual genes. Red-blue heatmaps (center plots) highlight the unspliced fraction of the individual genes, indicating active synthesis of transcripts (red), and apparently decreasing synthesis (blue), on PC1-2 displays as shown in (c). Green (right-hand plots) highlight spliced transcripts on the same axes. e) PC2-3 display with arrows indicating the transition probability of cells as described in the top panel of (c)(and Fig. 5b) plus corresponding gene plots as described in (d).



Supplementary Figure S8, related to Fig. 6.

Supervised analysis of 10X Chromium data: low dimensional representation based on the curated, instructive gene list in Table S2. a) PC loading of first 2 PCs with supervised analysis. b) PC (PC1 vs. PC2) and tSNE (tSNE2 vs. tSNE3) displays of 10X data (replicate 1, 4627 cells) with clusters color projected. tSNE and the SLM clustering algorithm were performed based on PC1-6 (same as seqFISH analysis). c) Clustered expression patterns of members of the curated instructive gene list (Table S2) on pseudo-time (same pseudo-time scale and

calculations as in Fig. 5; displayed are genes that were detected in ≥11 cells). Colored by log transformed and row normalized relative expression level. d) FACS gating strategy for ETP sub-population sorts in Fig. 6, both on population level and single-cell level. e) Precursors in individual gates shown in (d) were profiled for transcriptome expression and pseudotime prediction using Cell Hashing. Top panel labels the mean and interquartile ranges of individual-cell pseudotime predictions from each subpopulation. Bottom heatmap displays the key genes' expression pattern on this recalculated, 'ETP-enriched' pseudotime scale, aligned to the top panel. Note enrichment of *Tyrobp, Mpo, Cd7, Spi1,* and *Hhex* expression corresponding to the sorted ETP subsets 4 and 6.

METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

All sequence data generated in this study have been deposited in Gene Expression Omnibus and all genotypes of mice used in this study were crossed from strains available from Jackson Laboratories, or from strains we reported previously (Kueh et al., 2016), which are available upon reasonable requests. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ellen V. Rothenberg (evroth@its.caltech.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animals

Mice of a variety of genotypes were used exclusively as sources of primary cells to be analyzed *ex vivo* in these studies. B6.*Bcl11b*$^{yfp/yfp}$ reporter (Kueh et al., 2016) mice were used for bulk RNAseq analysis, *in vitro* developmental assays, and ETP subpopulation Cell Hashing 10X scRNA-seq. This nomenclature is used for animals which have a nondisruptive insertion of *IRES-mCitrine* into the 3'-untranslated region of *Bcl11b*, so that they have wildtype *Bcl11b* function despite simultaneously expressing the yellow fluorescent protein. C57BL/6(B6) mice (stock originally from Jackson Laboratories) were used for seqFISH and all other scRNA-seq analysis. B6.ROSA26-*mTom;Bcl11b-YFP* mice were used for clonal imaging analysis. They were generated by crossing and backcrossing B6.129(Cg)-*Gt(ROSA)26Sor*$^{tm4(ACTB-tdTomato,-EGFP)Luo}$/J mice, which express ubiquitous membrane Tomato (Jackson Laboratories), with the B6.*Bcl11b*$^{yfp/yfp}$ reporter mice until both loci were homozygous. Eμ-Bcl-2-25(Bcl2-tg) (Strasser et al., 1991) and B6.*Bcl11b*$^{yfp/yfp}$*;Bcl2-tg* mice were used for specific culturing assays as indicated below. B6.*Bcl11b*$^{yfp/yfp}$*;Bcl2* mice were generated through crossing B6.*Bcl11b*$^{yfp/yfp}$ x Bcl2-tg until the *Bcl11b* locus was homozygous. All adult animals used were mice between 4 and 8 weeks of age, and all samples within experiments were pools from multiple age and sex-matched animals. Animals used for these experiments were bred and maintained at the Animal Facilities at California Institute of Technology under conventional Specific Pathogen-Free conditions, and animal protocols were reviewed and approved by the Institute Animal Care and Use Committee of California Institute of Technology (Protocol

#1445-18G). To maximize both the thymus population sizes and fertility of the mice in the colony, care was taken to protect these animals from stress throughout their lifetimes to the greatest extent possible.

Cell lines

To provide a microenvironment that supports T-lineage differentiation in vitro, we co-cultivated primary cells with the OP9-DL1 stromal cell line (Schmitt and Zúñiga-Pflücker, 2002), which was obtained from Dr. Zúñiga-Pflücker (Sunnybrook Research Institute, University of Toronto) and maintained in our laboratory as described in the original reference. Control OP9 cells not expressing the Notch ligand DL1 were used to establish a microenvironment to support non-T cell developmental pathways of primary cells. The OP9-control cells were also obtained from Dr. Zúñiga-Pflücker. Both OP9-DL1 and OP9-control cell lines were tested and found to be negative for mycoplasma contamination. For live imaging experiments, a derivative of the OP9-DL1 cells was used, OP9-DL1-delGFP1, in which the GFP marker in the cell line had been removed by Cas9-mediated disruption as described elsewhere (Olariu et al., 2021). Details of the differentiation cultures are given below under Method Details.

METHOD DETAILS

Primary Cell Purification

Early stage thymocytes were purified from thymi removed from 4- to 8-week-old animals prior to flow cytometry analysis or fluorescence-activated cell sorting (FACS). Harvested thymi were mechanically dissociated to make single-cell suspensions that were re-suspended in Fc blocking solution with 2.4G2 hybridoma supernatant (prepared in the Rothenberg lab), followed by depletion of mature T and non-T lineage cells using a biotin-streptavidin-magnetic bead removal method. Briefly, thymocyte suspensions were labeled with biotinylated lineage marker antibodies (CD8α, TCRβ, TCRγδ, Ter119, CD19, CD11c, CD11b, NK1.1), incubated with MACS Streptavidin Microbeads (Miltenyi, Biotec) in HBH buffer (HBSS (Gibco), 0.5% BSA (FractionV), 10 mM HEPES, (Gibco)), pre-filtered through nylon mesh, and passed through a magnetic column (Miltenyi Biotec)

on a cell separation magnet (BD Biosciences) to obtain enriched DN cells. Then, the DN cells were stained with conjugated fluorescent cell surface antibodies (See STAR Key Resources Table) to purify the ETP, DN2a, and DN3 populations. ETP: Kit^{high} $CD44^{high}$ $CD25^{neg}$. DN2a: Kit^{high} $CD44^{high}$ $CD25^{+}$. DN2b: $Kit^{intermed}$ $CD44^{high/intermed}$ $CD25^{+}$. DN3: Kit^{low} $CD44^{low}$ $CD25^{+}$. Where the *Bcl11b-YFP* allele is present, the onset of *Bcl11b-YFP* expression distinguishes T-lineage committed DN2a cells from earlier, uncommitted DN2a cells (Kueh et al., 2016).

Flow Cytometry and Cell Sorting

Unless otherwise noted, flow cytometry analysis and FACS of all samples were carried out using the procedures outlined. Briefly, cultured cells on tissue culture plates and primary cells from thymus were prepared as single-cell suspensions, incubated in 2.4G2 Fc blocking solution, stained with respective surface cell markers as indicated (See STAR Key Resources Table), resuspended in HBH, and filtered through a 40 μm nylon mesh. They were then analyzed using a benchtop MacsQuant flow cytometer (Miltenyi Biotec, Auburn, CA) or sorted with a Sony Synergy 3200 cell sorter (Sony Biotechnology, Inc, San Jose, CA) for most of the single-cell transcriptome analyses and seqFISH samples, or with a FACSAria Fusion cell sorter (BD Biosciences) for the culture assays and ETP sub-population Cell Hashing scRNA-seq. All antibodies used in these experiments are standard, commercially available monoclonal reagents widely established to characterize immune cell populations in the mouse; details are given in the STAR Key Resources Table. Acquired flow cytometry data were all analyzed with FlowJo software (Tree Star).

Cell Cultures

Subsets of primary DN thymocytes FACS-purified as described above were cultured on a OP9-DL1 or OP9-control stromal monolayer system (Schmitt and Zúñiga-Pflücker, 2002) at 37°C in 7% CO_2 conditions with standard culture medium [80% αMEM (Gibco), 20% Fetal Bovine Serum (Sigma-Aldrich), Pen-Strep-Glutamine (Gibco), 50 μM β-mercaptoethanol (Sigma)] supplemented with appropriate cytokines (Lymphoid condition:

Flt3L (Pepro Tech Inc.) 10 ng/mL, Human IL7 (Pepro Tech Inc.) 5 ng/mL; Myeloid condition: M-CSF(Pepro Tech Inc.), GM-CSF(Miltenyi Biotec), and IL-6(Pepro Tech Inc.) each at 5 ng/mL, SCF(Pepro Tech Inc.) at 1 ng/mL, and IL-3 (Pepro Tech Inc.) at 0.1 ng/mL.

Bulk RNAseq Analysis

Kit^{hi} $CD44^{hi}$ cells purified from B6.*Bcl11b*$^{yfp/yfp}$ animals were subdivided into $Flt3^{high}CD25^{low}$ ETP, $Flt3^{low}CD25^{low}$ ETP, Bcl11b-$YFP^{neg}CD25^{hi}$ DN2a, and Bcl11b-$YFP^{pos}CD25^{hi}$ DN2a. fractions, followed by RNA purification following the instructions of the RNeasy Micro Kit (Qiagen 74004). cDNA from each sample was prepared with or without pre-amplification as indicated in Fig. 1. Pre-amplified samples were prepared with SMART-Seq v4 Ultra Low Input RNA Kit (Takara 634888) and Nextera XT library preparation kits (FC-131-1096) for Illumina sequencing, column 2, 6, 8,11 in Fig1b-c). Samples without pre-amplification were prepared using NEBNext Ultra RNA Library Prep Kit for Illumina (E7530, NEB). All bulk libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50 nt. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4 and produced approximately 30 million reads per sample.

RNA-seq reads were mapped onto the mouse genome build GRCm38/mm10 using STAR (v2.4.0) and were post-processed with RSEM (v1.2.25; http://deweylab.github.io/RSEM/) according to the settings in the ENCODE long-rna-seq-pipeline (https://github.com/ENCODE-DCC/long-rna-seq-pipeline/blob/master/DAC/STAR_RSEM.sh), with the minor modifications that the setting '–output-genome-bam–sampling-for-bam' was added to rsem-calculate-expression. STAR and RSEM reference libraries were created from genome build GRCm38/mm10 together with the Ensembl gene model file Mus_musculus.GRCm38.gtf. The resulting bam files were used to create HOMER tag directories (makeTagDirectory with –keepAll setting). For analysis of statistical significance among DEGs, the raw gene

counts were derived from each tag directory with 'analyzeRepeats.pl' with the '–noadj -condenseGenes' options, followed by the 'getDiffExpression.pl' command using EdgeR (v3.6.8; http://bioconductor.org/packages/release/bioc/html/edgeR.html). For data visualization, RPKM normalized reads were derived using the 'analyzeRepeats.pl' command with the options '–count exons –condenseGenes –rpkm'; genes with an average of RPKM ≥1 across samples were kept, and their RPKM values were processed by log transformation. The normalized datasets were then hierarchically clustered with R hclust function based on Euclidean distance and 'complete' linkage. The heatmap is visualized with R pheatmap with log2 transformed RPKM data (after adding 0.1 to all values).

Clonal Imaging Assay of Individual ETPs

To follow individual ETP clones by microscopic imaging, Kit^{hi} $CD44^{hi}$ $CD25^{-}$ ETP cells were purified from B6.ROSA26-*mTom;Bcl11b-YFP* mice (generated as described in the Animal sections above). Sorted ETP cells were plated onto OP9-DL1 stromal cells lacking GFP (OP9-DL1-delGFP1) in 24-well glass bottom plates with black 8mm circular poly(dimethyl siloxane) PDMS micromeshes with multiple microwells 250μM wide x 100 μM deep, custom fabricated by Microsurfaces (Australia). Cells were cultured in OP9 culture medium prepared as previously described except for the omission of the pH indicator, phenol red, from the medium, and with the addition of 10mM Hepes buffer to stabilize the pH of the wells during imaging, plus 10 ng/ml Flt3L, 5 ng/ml IL-7, and 0.05 μg/ml CD25-AlexaFluor647 (BioLegend), for detection of CD25 surface expression. Wells were imaged daily for 6 days on a Leica 6000 wide-field fluorescence inverted microscope with Metamorph software and an incubation chamber preset to 37°C, 7% CO_2. Wells found to have exactly one mTomato positive cell on either day 1 or 2 were followed subsequently and scored for CD25 and Bcl11b-YFP fluorescence.

SeqFISH

Experimental Design

Using seqFISH, single transcripts can be robustly detected and localized in 3D in light-scattering tissue or in samples of thousands of cells. The strategy detects each targeted gene with up to 24 probes per gene using Hybridization Chain Reaction (HCR) amplification, in which all the probes against a given gene share the same HCR amplification handle and are detected in repeated sequential rounds of color-coded HCR in which each gene is decoded by a different sequence of colors (Shah et al., 2016a). Signals can be aligned by keeping the sample immobile under the microscope throughout all rounds of processing. This technique enables detection of transcripts even < 1 kb in size, with a fidelity comparable to conventional single molecule FISH (smFISH), and can be sequentially multiplexed (Shah et al., 2016b, 2016a).

T cells have relatively small cytoplasm compared to many cell lines and other cell types, and it was observed that smFISH analysis was relatively hard to perform due to the high relative content of cytoplasmic membrane and nuclear membrane sandwiching the small cytoplasm, yielding relatively dim fluorescent signals. To amplify the signal, therefore, we designed a 5-color-sequential barcoding scheme of HCR-seqFISH, using an error correction scheme that tolerates 1 round of signal dropout or inaccuracy as described before (Shah et al., 2016b). We applied HCR-seqFISH against 54 genes on FACS sorted and immobilized early T cells, followed by additional targeted HCR smFISH analyses and immunostaining on the same samples. Targeted HCR smFISH analyses, of only five genes at a time, were used for functionally important genes with particularly short transcripts which required maximal sensitivity, or for those particularly abundant transcripts which can obstruct detection of other species in the barcoding rounds. Briefly, 14-24 primary probes incorporating designed hairpin initiation sequence handles (hyb1) were hybridized to mRNA transcripts of genes of interest, followed by HCR signal amplification in 5 colors against the “handles”. Targeted mRNAs detected by amplified signals appear to be individual bright dots in microscope images, and were recorded and registered in space. Without moving the slide on the microscope, primary probes and readout hairpins were then digested with DNaseI, leaving mRNAs intact, and the second hybrization round of primary probes, with attached handles permuted (hyb2), were hybridized again. After

HCR amplification, the second round of amplified signals in 5 colors were collected and registered to the previous hybridization. The steps were repeated until the completion of the designed sequential rounds of hybridization. The individual mRNA molecules were represented by the sequence of colors that appeared in the same registered spots. The identities of the mRNAs were encoded in the color sequence (color barcode details in Table S6).

SeqFISH Probe Design and Synthesis

The curated gene set that we selected as targets for seqFISH analysis consisted of regulatory genes that were judged likely to be functionally important in early T and lymphomyeloid development, based on previous genetic perturbation evidence, and lineage-associated genes that would be particularly informative as developmental state indicators (www.immgen.org) (Mingueneau et al., 2013) [reviewed in (Longabaugh et al., 2017; Rothenberg et al., 2016; Yui and Rothenberg, 2014)], as detailed in Table S2. The final list included 65 genes.

Gene-specific primary probes (35 nt long) were designed as previously described (Shah et al., 2016b), where 5 pairs of dye-coupled HCR hairpins (IR800, Alexa 647, Alexa 594, Cy3b, and Alexa 488) were used for signal amplification and readout from primary probes, and the 405nm channel was used for segmentation. Probes to be used in barcoding seqFISH were first subjected to stringent screening to avoid cross-reactivity, using the probe design software previously described (Shah et al., 2016b) with the following settings for this study. First, all candidate probes were BLASTed against the mouse transcriptome, and expected copy numbers of off-target probe hits were calculated using predicted RNA counts in the ENCODE database for murine thymocytes. BLAST hits with a 15-nt match on any sequences other than the target gene were considered off-target hits. For each target gene, any candidate probe that hit an expected cumulative total off-target copy number exceeding a threshold >0.1% of total was dropped, and candidate probes were sequentially dropped until no off-target gene was hit by more than 6 individual

probes from the entire pool. At this stage, all of the "viable" probes for each gene had been identified. For the final probe set, the best possible subset from the viable probes was selected such that the final probes were non-overlapping and at least 2-nt bases apart from each other. The choice between which of two overlapping candidate probes to keep was based on their respective distances from the target GC content (55% in this case). As a final step to minimize cross-hybridization between probe sets, a local BLAST database was constructed from all the viable probe sequences, and all of the probes (including "handle" sequences) were queried against it. All matches of 17 nt or longer between probes were removed by dropping the matched probe from the larger probe set. The final probe set size for barcoding seqFISH was 14-24 probes per gene. For targeted, non-barcoding smHCR, 8-24 probes per gene were used, and genes were analyzed in groups of 5 per HCR round, with groups based on similar probe numbers per gene.

The template oligos were generated from array-synthesized oligopools from Oligoarray or Twist Bioscience, and amplified as described by Chen et al., 2015 and Shah et al., 2016b. To balance the probes' concentrations, each of the template oligos were synthesized 3 times in the oligo pool, and probe pools for individual hybridizations were assigned a validated primer and assembled according to the following template (complete list in Table S6):

5' -[Primer 1] - [KpnI] - ["TAG"] - [primary probe] - [HCR initiator] - ["GAT"] - [EcoRI] - [Primer 2] - 3'

List of amplification primers:

Name	Primer1	Primer2	Pool #
Barcode hyb 1	AATTGAGCAGCTCGGGCCAC	GGCGATGGAAGCCTGCAACT	1
Barcode hyb 2	CCGCACGCCGTCCTTAAATC	CTTTCCGTGCTGCCGGATCT	1

Barcode hyb 3	GACGCACATATGCGGGCAAG	GGCATCTTCGTGACTGCGGA	1
Barcode hyb 4	ATTGAGGGTCTTCGCGTGCC	GTAACCGGCGCTTTGCAACC	1
smHCR hyb 1	TGTGCGCTCCGATTGTCCTC	GCAAATGGGGTCTGTTGGCC	1
smHCR hyb 2	TGCAGCTCCGCGAAATGAAG	CGCTGCCTGTCTGTGCCATT	1
smHCR hyb 3	TCAGGGCACGAGGACATTCG	TCCGGCAAGATTGCTCTCCC	2
smHCR hyb 4	ATGCGCTGCAACTGAGACCG	TTGTGCCAGCCTTGGTCGAG	2

SeqFISH Experimental Procedures and Imaging

The DN cells were purified as described in "Cell Purification" above, the ETP-DN2 population was FACS-sorted as a continuum as shown in Fig. S1c, and an equal number of DN3 cells was sorted separately, each population into tubes containing HBH buffer. Next, the isolated DN cell fractions were crosslinked with 4% Formaldehyde (ThermoScientific 28908) in 1X PBS for 10min. Then, cells were spun onto an amino-silane modified coverslip in hyb-cells (Grace Bio-Labs, RD478685-M). They were then crosslinked again with 4% Formaldehyde (ThermoScientific 28908) in 1× PBS for 10min, and permeabilized in 70% EtOH overnight at 4°C. Samples were imaged first to record the surface antibody signals, followed by briefly bleaching away antibody signals through incubation in 0.1% $NaBH_4$ (Sigma 452882) in 1× PBS for 10min. Then, the samples were washed with PBS and pretreated with DNaseI (Roche Cat. #04716728001) at 1 U/µl for 2 hrs at 37°C, and washed 3 times with 50% Hybridization Buffer (50% HB: 2× SSC (Invitrogen 15557-036), 50% Formamide (v/v) (Ambion AM9344), 10% Dextran Sulfate (Sigma D8906) in Ultrapure water (Invitrogen 10977-015)). Following pre-treatment, samples were (1) hybridized overnight at 37°C with primary intron probes at concentrations of 1 nM each oligo in 50% Hybridization Buffer, then (2) washed in 50% Wash Buffer (2× SSC, 50% Formamide (v/v), 0.1% Triton-X 100 (Sigma X-100)) for 20

minutes, followed by incubation in 2× SSC for 10 minutes. The samples were then (3) incubated with HCR hairpins in Amplification Buffer (2× SSC, 10% Dextran Sulfate in Ultrapure water) for 30 minutes followed by (4) washing in 2× SSC for 5 min, and then in 10% Wash Buffer (2× SSC, 10% Formamide (v/v), 0.1% Triton-X 100 (Sigma X-100)) for 10 minutes. Before imaging, brief DAPI staining was performed for cell background registration and segmentation (DAPI 5μg/mL, 1min, Sigma D8417), then (5) imaged as described below. After image acquisition, (6) the samples were incubated with 1 U/μl DNaseI (Roche) for 3 hours at 37°C, and the remaining enzymes were washed out by 30 min incubation with 50% wash buffer at 37°C. The procedures (3)-(6) constituted one round and were repeated until the completion of all rounds of barcoding and non-barcoding HCR seqFISH.

Post RNA profiling, additional immunostaining with antibodies was performed in some experiments to quantitate transcription factor proteins. Specifically, samples were blocked with 1 × PBS, 1% BSA for 1 hour at room temperature, followed by incubation with anti-PU.1 or anti-TCF1, and anti-CD44 (not shown) (See STAR Key Resources Table) at 1:100 for 2 hours at room temperature, then washed in PBS 3 times, and then imaged. Note that antibodies used for surface staining, e.g. anti-cKit, were imaged before hybridization as described above.

Samples were imaged in an anti-bleaching buffer (20 mM Tris-HCl, 50 mM NaCl, 0.8% glucose, saturated trolox (Calbiochem 648471), pyranose oxidase (OD405 = 0.05) (Sigma P4234), and catalase at a dilution of 1/1000 (Sigma C3155)). Sample port covers were closed with a glass coverslip or a transparent polycarbonate sheet to exclude oxygen. The images were acquired with a microscope (Leica, DMi8) equipped with a confocal scanner unit (Yokogawa CSU-W1), sCMOS camera (Andor Zyla 4.2 PLUS), 40x oil objective lens (Leica NA 1.30), and a motorized stage (ASI MS2000). Lasers from CNI and filter sets from Semrock were used. Snapshots were acquired with 0.5 μm z steps for more than 30 positions per sample.

Image Processing and Analysis

The images were first corrected to remove the uneven illumination profiles in each channel, the effects of chromatic aberration, and registered for shift across all hybridizations as described before (Shah et al., 2016b).

For cell segmentation, the cell background taken in the DAPI channel without staining was first maximum z projected and blurred using a 2D Gaussian blur with a sigma of 1 pixel. The ImageJ-FIJI built in default dark thresholding algorithm was then used to separate out the cell boundary from background. Finally, the thresholded image was run through a watershed algorithm to demarcate individual cells. The obtained individual cell masks were further filtered by size (number of pixels between 600-3000) and circularity (between 0.7 to 1). The subsequent segmentation results were manually curated and corrected to obtain a final accurate segmentation of images.

The potential mRNA signals were then found by LOG filtering the registered images and finding points of local maxima above a specified threshold value. Once all potential points in all channels of all hybridizations were obtained, dots were matched to potential barcode partners in all other channels of all other hybridizations using a 3-pixel search radius to find symmetric nearest neighbors. The number of each barcode was then counted in each of the assigned segmented cells. Signals were decoded using the designed sequences of colors that should uniquely represent each targeted gene (Table S6).

The antibody staining quantification was performed with maximum z-projections for each channel. Average pixel intensities were quantified within individual cell segmentations, subtracted by average background intensity acquired in dummy segmentations (no cells) in the same fields of view, and multiplied by area to estimate the total signal. Because the quantification was performed after subtraction of background intensity, the total signal quantitation is not sensitive to segmentation accuracy or area size.

C1^{TM}-Fluidigm Smartseq2 Single Cell RNA-seq

ETP-DN2a cells were purified as a continuum as described above (Fig. S1c), except that no DN3 cells were pooled in for C1 analysis. The cells were then washed and resuspended to 250,000 cells/mL concentration in HBH buffer; 12 μL of this suspension was added to 8 μL of Fluidigm Cell Suspension Reagent for loading on the Fluidigm IFC (5-10 μm size). Cells were visually inventoried for doublets and empty chambers, and returned to the C1 for lysis, reverse transcription and amplification using the SMART-Seq v4 protocol. All amplified cDNA samples were quantified on Qubit and a subset were selected for BioAnalyzer sizing based on yield and chamber occupancy. The cDNA libraries were then tagmented using the Nextera XT DNA sample prep kit and Nextera XT indices. After tagmentation and amplification, libraries were pooled, cleaned up with Ampure XP beads (0.9× volume), quantified on Qubit and sized on the BioAnalyzer. Following the library preparation, the sequencing was performed with single read sequencing of 50nt on HiSeq2500 with a sequencing depth of $1.5x10^6$ reads per cell. The reads were mapped onto the GRCm38/mm10 mouse genome assembly.

10X Chromium V2 Single Cell RNA-seq

The DN thymocytes were enriched as described above, the ETP-DN2 population was sorted together as a continuum as shown in Fig. S1c, and DN3 cells were sorted separately. A small aliquot of DN3 cells representing ~10% of the total ETP-DN2 cells was added into the ETP-DN2 sample as a developmental endpoint internal reference. The sample was then washed and resuspended to 1 million cells/mL concentration in HBSS supplemented with 10% FBS and 10 mM HEPES, 17,400 cells were loaded into each 10X Chromium v2 lane, and the subsequent preparation was conducted following the instruction manual of 10X Chromium v2. The cDNA library and final library after index preparation were checked with bioanalyzer (High Sensitivity DNA reagents, Agilent Technology #5067-4626; Agilent 2100 Bioanalyzer) for quality control. Following the library preparation, the sequencing was performed with paired-end sequencing of 150nt each end on one lane of HiSeq4000 per sample, by Fulgent Genetics, Inc. (Temple City, CA). The reads were mapped onto the mouse genome Ensembl gene model file

Mus_musculus.GRCm38.gtf using a standard CellRanger pipeline. Cells were sequenced to an average depth of 40,000-50,000 reads per cell (target $4x10^8$ reads per lane).

Cell Hashing with Single Cell RNAseq

DN cells were purified as described above, pooling thymus from eight female B6.*Bcl11b*$^{yfp/yfp}$ mice, 5.5-weeks old. The 4 subsets of ETP cells (pops 1, 3, 4, 6) were sorted 4-way using the gates described in Fig. S8d. The sorted cells (total yield ~2000 per gate) were concentrated and each subset was incubated individually with TotalSeq A (Biolegend) anti-Mouse Hashtag 1, 2, 3, or 4 (1:50), respectively. A sorted reference population of ETP-DN2 continuum plus 10%DN3 cells, as in Fig. S1c, was tagged in parallel with anti-Mouse Hashtag 5. The samples were then washed 3 times with HBSS supplemented with 10% FBS and 10 mM HEPES, and pooled to load onto one lane of a 10X Chromium V3 chip. The cDNA preparation was performed following the instruction manual of 10X Chromium v3, and the hashtag library was prepared following the Biolegend TotalseqA guide. The cDNA, tag library, and final library after index preparation were checked with bioanalyzer (High Sensitivity DNA reagents, Agilent Technology #5067-4626; Agilent 2100 Bioanalyzer) for quality control. The cDNA final library was sequenced on NovaSeq 6000, and the tag library was sequenced on HiSeq4000, by Fulgent Genetics, Inc.. Cells were sequenced to an average depth of ~50,000 reads per cell for cDNA and ~2,500 reads per cell for hashtags.

Single-Cell Expression Profile Data Analysis

Analytical Pipelines

The analysis methods applied and the relationships between different datasets and methods are abbreviated in the schematics in Fig.S1b. Specifically, the software/packages Seurat v2.3.4 and 3.0.1 (Butler et al., 2018; Stuart et al., 2019), Monocle v2 (Qiu et al., 2017a, 2017b), Velocyto v0.17.8 (La Manno et al., 2018), and SPRING (Weinreb et al., 2018) were used in this study, and 10X raw reads were mapped and assigned by Cell

Ranger. Unsupervised analysis of low dimensional representations (tSNE, UMAP, SPRING), RNA velocity, and clustering were performed with gene sets filtered as described below.

Supervised clustering and pseudotime analysis of 10X data were performed based on the curated list of genes in Table S2, using quality control (QC)-trimmed 10X datasets from which the DN3b and granulocyte precursor clusters were computationally removed. For trajectory analysis, this improves developmental connectivity and T-lineage relevance. For seqFISH analysis, data from the cells were QC trimmed as described below, and for high dimensional analysis, the expression was further normalized by RNA content/size, as described below.

Gene and Cell Filtering: Quality Control

In seqFISH analysis, cells with less than 250 barcoded transcripts detected (total from 54 barcoded genes) were omitted. In PCA and clustering analysis, similar to scRNA-seq, the cells were first size-normalized to estimated RNA content. The RNA content in individual cells was estimated by total number of mRNA signals detected in one barcoding hybridization round without decoding. Applying the Quality Control (QC) filter resulted in 4551 cells from 4-week-old animals, 7150 cells from 5-week-old animals, and 2598 cells from 8-week-old animals being presented in this study.

The C1 Fluidigm-Smartseq2 analysis was performed based on data filtered on cells that visually appeared to be single cells observed under the microscope in the Fluidigm chip, with at least 3600 genes expressed, less than 11% mitochondrial content, and with detectable expression of genes that are differentially expressed in bulk analysis described in Fig. 1d. The filter resulted in 193 cells presented in this study.

Unless otherwise specified, both supervised and unsupervised analysis of 10X Chromium V2 scRNA-seq was based on data filtered on cells with at least 1200 genes expressed (transcript count over 1); outliers with more than 4500 genes were also removed (potential doublet), and only genes that were found expressed in at least 3 cells were kept in the

analysis. For clustering, the cells were further cleaned to keep only cells with mitochondrial content of less than 5%, with signals normalized to total number of UMI and mitochondrial content as recommended by Seurat2. The QC filter resulted in 4627 cells in replicate 1 and 7076 cells in replicate 2 being presented in this study. The RNA velocity and pseudo-time analysis with Velocyto and Monocle 2, respectively, were performed on the cells that passed the filtering steps described above, and also with the DN3b cluster and granulocyte precursor cluster removed (cluster 13 in unsupervised analysis, both replicates).

Unsupervised clustering analysis of 10X scRNA-seq data was performed after log normalization and scaling, with 4307 variable genes identified in Seurat2 (average expression between 0.0125 and 3, and minimum dispersion of 0). Note that the dispersion filter was set low to allow capture of subtle features of the developmental continuum.

Inter-technique Comparison

We calculated the average raw gene expression levels in comparable cell input populations between different techniques in their own measurement units. The general expression levels were found to agree, allowing that the target genes mainly encode transcription factors and are expressed at very low levels. Overall, seqFISH was approximately tenfold more sensitive than 10X Chromium v2, in terms of estimated transcript counts per gene (Fig. S2b) and in a greatly reduced dropout rate, as shown for the functionally essential developmental regulatory genes in Fig. S2d. This finding is consistent with the previously described 10% sampling rate of 10X Chromium v2, at the sequencing depth being used (Islam et al., 2014; Kolodziejczyk et al., 2015). The discrepancies between the C1-Smartseq2 and 10X systems (Spearman correlation=0.68, Pearson correlation=0.57) are likely due to the difference in UMI and non-UMI based measurement unit, as amplification steps in Smartseq2 could result in biased readout of some genes. Aside from sensitivity differences, the biggest qualitative differences between sequencing-based (C1-Smartseq2 and 10X) and seqFISH measurements on the selected genes are likely due to

the fact that seqFISH by-passes any poly(A)-based reverse transcription-amplification step and probes directly at the exon regions of mRNAs. This can lead to the following: a) seqFISH can also probe the pre-mRNAs of genes of interest that have not been poly-adenylated; b) when the reverse transcription step in scRNA-seq is inefficient that will lead to dropouts, such that sequencing would more robustly detect genes that are expressed at high levels; and c) miscalling of transcripts in seqFISH can occur due to crowded transcript signals in limited-sized lymphocytes. Indeed, the expression patterns of genes between seqFISH and 10X showed general agreement but were still only moderately correlated, as represented by Spearman correlation of 0.73 and Pearson correlation of 0.45 on the lowly-expressed regulatory gene transcripts (Fig S2b).

PU.1 and Bcl11b Perturbation data

The pseudotime model is compared with recently determined functional targets of PU.1 and Bcl11b, in Table S4 and Fig. 5f. Lists of genes activated or repressed by PU.1 were taken from the overlap of acute perturbation data for PU.1 gain and loss of function in DN2a-DN2b pro-T cells [Table S6B in (Ungerbäck et al., 2018)]. Specifically, the 326 PU.1-activated genes showed both enhanced activation 48h after exogneous PU.1 was introduced into DN2b cells and reduced expression 4d after endogenous *Spi1* was disrupted from DN2a cells (p.adj<0.1). The 237 PU.1-repressed genes showed both downregulation in response to the exogenous PU.1 and upregulation when endogenous *Spi1* was disrupted (p.adj<0.1). The 747 Bcl11b-repressed genes and 394 Bcl11b-dependent genes were defined from the intersection of genes responding significantly (p.adj < 0.05, at least twofold change) in the same direction in at least two different types of loss of function perturbations affecting DN2b-stage cells: in vivo deletion by *Vav1-iCre*, in vivo deletion by *pLck-Cre*, and/or in vitro acute deletion by Cas9 and guide RNA in DN2b cells [Supplementary Table 3 in (Hosokawa et al., 2018a)].

QUANTIFICATION AND STATISTICAL ANALYSIS

Experiments and techniques were carried out independently at least twice. Three independent seqFISH experiments were carried out, two independent 10X analyses were carried out on completely separate biological samples, and cell hashing 10X analysis of ETP subsets was carried out on a third completely independent biological sample. C1 data were pooled from ETP-DN2a cells sorted onto the chips in three separate experiments. While analyses shown in the paper are primarily from one of the three seqFISH replicates (in most cases the 4 wk old mouse sample) or one of the two 10X replicates (mostly replicate 1, which yielded greater sequencing depth per cell), the data were highly consistent between independently generated samples using the same technique, and highly consistent with the C1 analysis, as shown in Figs. S2 and S3. Cell culture experiments were carried out three to four times independently with concordant results as indicated in Fig.6 and Supplementary Figures S4, S5, and S8. Only the single-cell sorted experiments in Fig. 6f-h, which corroborate other data in Figs. 5, 6c-e, and S8, were not repeated as such. Cloning data in Fig. 1 (>60 clones) and Fig. 8f (>300 clones) each came from one experiment.

The statistical tests and specific settings used for each comparison are indicated in the individual figure and table legends.

DATA AND CODE AVAILABILITY

All sequence data generated in this study have been deposited in Gene Expression Omnibus and are available under accession numbers GSE130812 and GSE137165. Sources for code used in this study are indicated in the Key Resources Table.

LIST OF SUPPLEMENTARY MATERIAL

SUPPLEMENTARY TABLES

Supplementary Table S1: related to Fig. 1

Supplementary Table S2: related to Fig. 2 and Fig. 5

Supplementary Table S3: related to Fig. 2

Supplementary Table S4: related to Figs. 3, 4, and S8

Supplementary Table S5: related to Fig. 5

Supplementary Table S6: related to Fig. 2 and STAR Methods

SUPPLEMENTARY FIGURES

Figure S1: related to Figs. 1, 2, 3, 4, 5, 6, and 7

Figure S2: related to Fig. 2

Figure S3: related to Fig. 4

Figure S4: related to Figs. 2, 3, and 4

Figure S5: related to Figs. 2, 3, and 4

Figure S6: related to Fig. 4

Figure S7: related to Fig. 5

Figure S8: related to Fig. 6

SUPPLEMENTARY TABLE TITLES AND LEGENDS

Supplementary Table S1, related to Fig. 1. Bulk RNAseq data of genes differentially expressed between all ETP samples and Bcl11b-YFP^{+} DN2a samples. Genes with the average RPKM larger than 1, expression fold change larger than 2 either way and adjusted pval < 0.05 are shown. Values in RPKM.

Supplementary Table S2, related to Fig. 2 and Fig. 5. Curated regulatory and marker genes used in seqFISH analysis and supervised 10X Chromium analysis. Table indicates gene names and the combinations of criteria used for selecting each of these genes as particularly informative, based on their genetically defined functional importance or use as developmental state indicators (www.immgen.org) (Mingueneau et al., 2013) [reviewed in (Longabaugh et al., 2017; Rothenberg et al., 2016; Yui and Rothenberg, 2014)].

Supplementary Table S3, related to Fig. 2. SeqFISH raw transcript data and analysis of transcript distribution comparison between different stages of pro-T cells. Populations being compared are *Gata3*- (≤3 transcripts) **and** *Tcf7*- (≤5 transcripts) double negative ETPs, *Gata3*+ (>10 transcripts) **or** *Tcf7*+ (>20 transcripts) ETPs, and *Bcl11b*+ (> 5 transcripts) DN2s. Thresholds for binning were drawn to identify clear positives and negatives and avoid ambiguous intermediate levels of expression. Highlighted are p values $<10^{-6}$, two-tailed T test, unequal variances. Analysis performed from the 4-week-animal dataset.

Supplementary Table S4 , related to Fig. 3, Fig. 4, and Fig. S8. C1 and 10X marker genes identified in each sub-cluster in the analyses shown. Clustering based on SLM, markers identified with minimum fraction of 0.2 in the cluster and threshold of 0.2 using Wilcoxon rank sum test in Seurat 2. C1 supervised analysis was performed as shown in Fig.3. 10X unsupervised analysis was performed in Seurat 2 as shown in Fig. 4. 10X

supervised analysis was performed as described in Fig. S8. pct.1, pct.2: weightings in principal components 1, 2 respectively.

Supplementary Table S5, related to Fig. 5. Differentially expressed genes identified by supervised pseudo-time analysis from 10X analysis (intersection of both independent 10X replicates, qval<1E-08). The genes are ordered and clustered based on the pseudotime expression pattern as shown in Fig. 5. The crosses mark the individual genes that overlap with perturbation assays, which were shown to be significantly regulated by PU.1 or Bcl11b, as described in Fig. 5f-g. Lists of genes regulated by PU.1 or by Bcl11b were from published data (Hosokawa et al., 2018a, 2018b; Ungerbäck et al., 2018) as described in STAR Methods.

Supplementary Table S6, related to Fig. 2 and STAR methods. Designed oligo probe template pools and sequential color barcode used for seqFISH experiments.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-human/mouse CD44 PE	eBioscience	Cat#12-0441-83
Anti-mouse CD117 (cKit) APC	eBioscience	Cat#17-1171-82
Anti-mouse CD25 eFluor-450	eBioscience	Cat#48-0251-82
Anti-mouse CD25 APCe780	eBioscience	Cat#47-0251-82
Anti-mouse CD25-Alexa Fluor 647	Biolegend	
Anti-mouse CD45 PECy7	eBioscience	Cat#25-0451-82
Anti-mouse CD11b PE	eBioscience	Cat#12-0112-85
Anti-mouse CD11b AF488	eBioscience	Cat#53-0112-82
Anti-mouse CD11b APCe780	eBioscience	Cat#47-0118-42
Anti-mouse CD11c e450	eBioscience	Cat#48-0114-82
Anti-mouse CD11c APCe780	eBioscience	Cat#47-0114-82
Anti-mouse CD63 PE	Biolegend	Cat#143903
Anti-mouse Ly6c PE	Biolegend	Cat#128008
Anti-mouse Ly6c Alexa Fluor 647	Biolegend	Cat#128010
Anti-mouse CD135 (Flt3) BV421	Biolegend	Cat#135313
Anti-mouse CD24(HSA) APC	Biolegend	Cat#138506
Anti-mouse Ly6d PE	Biolegend	Cat#138603
Anti-mouse Gr1 APC	Biolegend	Cat#108412
Anti-mouse NK1.1 PE	eBioscience	Cat#12-5941-83
Anti-mouse Dx5 PE	eBioscience	Cat#12-5971-83
Anti-mouse NK1.1 Biotin	eBioscience	Cat#13-5941-85
Anti-mouse CD19 Biotin	eBioscience	Cat#13-0193-85
Anti-mouse Ter119 Biotin	eBioscience	Cat#13-5921-85
Anti-mouse CD11b Biotin	eBioscience	Cat#13-0112-86
Anti-mouse CD11c Biotin	eBioscience	Cat#13-0114-85
Anti-mouse CD8a Biotin	eBioscience	Cat#13-0081-86
Anti-mouse TCR$\gamma\delta$ Biotin	eBioscience	Cat#13-5711-85
Anti-mouse TCRβ Biotin	eBioscience	Cat#13-5961-85
Streptavidin PerCP-Cy5.5	eBioscience	Cat#45-4317-82
PU.1 (9G7) Rabbit mAb (Alexa Fluor 647 conjugate)	Cell Signaling	Cat#2240
TCF1/TCF7 (C63D9) Rabbit mAb (Alexa Fluor 647 conjugate)	Cell Signaling	Cat#6709
Totalseq-A0301 anti-mouse Hashtag1	Biolegend	Cat#155801
Totalseq-A0301 anti-mouse Hashtag2	Biolegend	Cat#155803
Totalseq-A0301 anti-mouse Hashtag3	Biolegend	Cat#155805
Totalseq-A0301 anti-mouse Hashtag4	Biolegend	Cat#155807
Totalseq-A0301 anti-mouse Hashtag5	Biolegend	Cat#155809
Biological Samples		
Primary murine thymocytes	This work	
Chemicals, Peptides, and Recombinant Proteins		

MEM Alpha	GIBCO	Cat#12561-056
Fetal Bovine Serum	SigmaAldrich	Cat#F7305
Human IL-7	PeproTech Inc	Cat#200-07
Human FLT-3-Ligand	PeproTech Inc	Cat#300-19
Stem Cell Factor	PeproTech Inc	Cat#250-03
Murine M-CSF	PeproTech Inc	Cat#315-02
Mouse GM-CSF	Miltenyi Biotec	Cat#130-095-739
Murine IL3	PeproTech Inc	Cat#213-13
Murine IL6	PeproTech Inc	Cat#216-16
HBSS	GIBCO	Cat#14175-095
HEPES	GIBCO	Cat#15630-080
Pen Strep Glutamine	GIBCO	Cat#10378-016
MACS Streptavidin Microbeads	Miltenyi Biotec	Cat#130-048-101
37% formaldehyde	ThermoFisher Scientific	Cat#28908
7AAD	eBioscience	Cat#00-6993-50
β-mercaptoethanol	SigmaAldrich	Cat#M6250
$NaBH_4$	SigmaAldrich	Cat#452882
DNaseI recombinant, RNase-free	Roche	Cat#4716728001
20× SSC	Invitrogen	Cat#15557-036
Formamide	Ambion	Cat#AM9344
HCR amplification hairpins	Molecular Instruments	Custom order
Dextran Sulfate	SigmaAldrich	Cat#D8906
Trolox	Calbiochem	Cat#648471
Pyranose oxidase	SigmaAldrich	Cat#P4234
Catalase	SigmaAldrich	Cat#C3155
Critical Commercial Assays		
Illumina Nextera DNA preparation Kit	Illumina	Cat#FC-121-1030
Nextera Index Kit (96 indexes, 384 samples)	Illumina	Cat#FC-121-1012
RNeasy Micro Kit	QIAGEN	Cat#74004
C1™ Single-Cell mRNA Seq IFC, 5–10 µm	Fluidigm	Cat#100-5759
Chromium i7 Multiplex Kit	10X Genomics	Cat#120262
Chromium Single Cell 3′ Library & Gel Bead Kit v2	10X Genomics	Cat#120267
Chromium Single Cell A Chip Kit	10X Genomics	Cat#1000009
High Sensitivity DNA Kit	Agilent Technologies	Cat#5067- 4626
Qubit dsDNA HS Kit	ThermoFisher Scientific	Cat#Q32854
SPRIselect reagent kit	Beckman Coulter	Cat#B23318
Chromium Single Cell 3′ GEM, Library & Gel Bead Kit v3	10X Genomics	Cat#1000092
Chromium Chip B Single Cell Kit	10X Genomics	Cat#1000074

Deposited Data		
Bulk RNA-seq data	This work	Gene Expression Omnibus GSE130812
Two samples, 10X Chromium RNA-seq	This work	Gene Expression Omnibus GSE130812
C1 Smartseq2 RNA-seq, 226 cells	This work	Gene Expression Omnibus GSE130812
10X Chromium RNA-seq cell hashing sample, 5 cell fractions barcoded	This work	Gene Expression Omnibus GSE137165
Experimental Models: Cell Lines		
OP9-DL1	Schmitt et al., 2002	N/A
OP9-DL1 dGFP	Olariu et al., 2021	N/A
OP9-control	Schmitt et al., 2002	N/A
Experimental Models: Organisms/Strains		
Mouse: C57BL/6	Jackson laboratories	Stock NO: 664
Mouse: B6.Cg-Tg(BCL2)25 Wehi/J (Bcl2-tg)	Jackson laboratories	Stock NO: 002320
Mouse: Bcl11b-YFP	Kueh et al., 2016	N/A
Mouse: Bcl11b-YFP x BCL2	This work	N/A
Mouse: B6.ROSA26-*mTom;Bcl11b-YFP*	This work	N/A
Oligonucleotides		
Listed in Table S6		
Software and Algorithms		
Bedtools (v.2.17.0)	Quinlan and Hall, 2010	http://bedtools.readthedocs.io/en/latest/
Bioconductor (v3.4)	N/A	http://bioconductor.org/
DESeq2 (v.1.14.1)	Love et al., 2014	https://www.bioconductor.org/packages/devel/bioc/html/DESeq2.html
EdgeR (v.3.16.5)	Robinson et al., 2010	http://bioconductor.org/packages/release/bioc/html/edgeR.html
FlowJo (v10.0.8)	N/A	https://www.flowjo.com/
Ggplot2 (v.2.2.1)	N/A	http://ggplot2.org/

HOMER (v4.8)	Heinz et al., 2010	http://homer.ucsd.edu/homer/
MATLAB (R2016a)	N/A	https://www.mathworks.com/products/matlab.html
R (v3.4.2)	N/A	https://www.r-project.org/
RSEM (v1.2.25)	Li and Dewey, 2011	http://deweylab.github.io/RSEM/
Rstudio (v1.1.383)	N/A	https://www.rstudio.com/
Samtools (v0.1.19-96b5f2294a)	Li et al., 2009	http://samtools.sourceforge.net/
STAR (v2.4.0; v2.5.2a)	Dobin et al., 2013	https://github.com/alexdobin/STAR/releases
Python(v3.6)	N/A	https://www.python.org
Custom probe design software	Shah et al., 2016b	Long Cai lab
Velocyto.py (v0.17.8)	La Manno et al., 2018	http://velocyto.org/velocyto.py/
Seurat (v2.3.4; v3.0.1)	Butler et al., 2018; Stuart et al. 2019	https://satijalab.org/seurat/
SPRING	Weinreb et al., 2018	https://kleintools.hms.harvard.edu/tools/spring.html
Monocle v2	Qiu et al., 2017a, 2017b	http://cole-trapnell-lab.github.io/monocle-release/
Other		
BD FACS Aria II Cell Sorter	BD Bioscience	N/A
Illumina HiSeq 2500	Illumina	N/A
Illumina HiSeq 4000	Illumina	N/A
iCyt Mission Technology Reflection Cell Sorter	Sony	N/A
BD FACSARIA FUSION Cell Sorter	BD Bioscience	N/A
Miltenyi Biotech MACSQuant 10 Flow Cytometer	Miltenyi Biotec	N/A
hyb-cells	Grace Bio-Labs	RD478685-M
Microscope	Leica	DMi8
Confocal Scanner Unit	Yokogawa	CSU-W1
sCMOS camera	Andor	Zyla 4.2 PLUS
40x Oil Objective Lens NA1.30	Leica	N/A
Motorized stage MS2000	ASI	N/A
Leica wide-field fluorescence inverted microscope	Leica	6000
Black PDMS micromesh inserts	Microsurfaces	MMA-0250-100-08-01

BIBLIOGRAPHY

Bell, J.J., and Bhandoola, A. (2008). The earliest thymic progenitors for T cells possess myeloid lineage potential. Nature *452*, 764–767.

Besseyrias, V., Fiorini, E., Strobl, L.J., Zimber-Strobl, U., Dumortier, A., Koch, U., Arcangeli, M.-L., Ezine, S., MacDonald, H.R., and Radtke, F. (2007). Hierarchy of Notch–Delta interactions promoting T cell lineage commitment and maturation. J. Exp. Med. *204*, 331–343.

Boudil, A., Skhiri, L., Candéias, S., Pasqualetto, V., Legrand, A., Bedora-Faure, M., Gautreau-Rolland, L., Rocha, B., and Ezine, S. (2013). Single-cell analysis of thymocyte differentiation: Identification of transcription factor interactions and a major stochastic component in ab-lineage commitment. PLoS One *8*, e73098.

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.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.

De Obaldia, M.E., and Bhandoola, A. (2015). Transcriptional regulation of innate and adaptive lymphocyte lineages. Annu Rev Immunol *33*, 607–642.

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

Evrard, M., Kwok, I.W.H., Chong, S.Z., Teng, K.W.W., Becht, E., Chen, J., Sieow, J.L., Penny, H.L., Ching, G.C., Devi, S., et al. (2018). Developmental Analysis of Bone Marrow Neutrophils Reveals Populations Specialized in Expansion, Trafficking, and Effector Functions. Immunity *48*, 364-379.e8.

Franco, C.B., Scripture-Adams, D.D., Proekt, I., Taghon, T., Weiss, A.H., Yui, M.A., Adams, S.L., Diamond, R.A., and Rothenberg, E.V. (2006). Notch/Delta signaling constrains reengineering of pro-T cells by PU.1. Proc. Natl. Acad. Sci. U. S. A. *103*, 11993–11998.

García-Ojeda, M.E., Klein Wolterink, R.G.J., Lemaître, F., Richard-Le Goff, O., Hasan, M., Hendriks, R.W., Cumano, A., and Di Santo, J.P. (2013). GATA-3 promotes T-cell specification by repressing B-cell potential in pro-T cells in mice. Blood *121*, 1749–1759.

Germar, K., Dose, M., Konstantinou, T., Zhang, J., Wang, H., Lobry, C., Arnett, K.L., Blacklow, S.C., Aifantis, I., Aster, J.C., et al. (2011). T-cell factor 1 is a gatekeeper for T-cell specification in response to Notch signaling. Proc. Natl. Acad. Sci. U. S. A. *108*, 20060–20065.

Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-Wallscheid, N., Dress, R., Ginhoux, F., et al. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. Nat. Cell Biol. *20*, 836–846.

Gwin, K.A., Shapiro, M.B., Dolence, J.J., Huang, Z.L., and Medina, K.L. (2013). Hoxa9 and Flt3 signaling synergistically regulate an early checkpoint in lymphopoiesis. J. Immunol. Baltim. Md 1950 *191*, 745–754.

Heinzel, K., Benz, C., Martins, V.C., Haidl, I.D., and Bleul, C.C. (2007). Bone marrow-derived hemopoietic precursors commit to the T cell lineage only after arrival in the thymic microenvironment. J Immunol *178*, 858–868.

Hosokawa, H., Romero-Wolf, M., Yui, M.A., Ungerbäck, J., Quiloan, M.L.G., Matsumoto, M., Nakayama, K.I., Tanaka, T., and Rothenberg, E.V. (2018a). Bcl11b sets pro-T cell fate by site-specific cofactor recruitment and by repressing Id2 and Zbtb16. Nat. Immunol. *19*, 1427–1440.

Hosokawa, H., Ungerbäck, J., Wang, X., Matsumoto, M., Nakayama, K.I., Cohen, S.M., Tanaka, T., and Rothenberg, E.V. (2018b). Transcription factor PU.1 represses and activates gene expression in early T cells by redirecting partner transcription factor binding. Immunity *49*, 782.

Hosoya, T., Kuroha, T., Moriguchi, T., Cummings, D., Maillard, I., Lim, K.-C., and Engel, J.D. (2009). GATA-3 is required for early T lineage progenitor development. J. Exp. Med. *206*, 2987–3000.

Hu, G., Cui, K., Fang, D., Hirose, S., Wang, X., Wangsa, D., Jin, W., Ried, T., Liu, P., Zhu, J., et al. (2018). Transformation of accessible chromatin and 3D nucleome underlies lineage commitment of early T cells. Immunity *48*, 227-242.e8.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. *11*, 774–785.

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.

Ishizuka, I.E., Chea, S., Gudjonson, H., Constantinides, M.G., Dinner, A.R., Bendelac, A., and Golub, R. (2016). Single-cell analysis defines the divergence between the innate lymphoid cell lineage and lymphoid tissue-inducer cell lineage. Nat Immunol *17*, 269–276.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J., et al. (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. Nat. Immunol. *19*, 85–97.

Knapp, D.J.H.F., Hammond, C.A., Hui, T., van Loenhout, M.T.J., Wang, F., Aghaeepour, N., Miller, P.H., Moksa, M., Rabu, G.M., Beer, P.A., et al. (2018). Single-cell analysis identifies a CD33+ subset of human cord blood cells with high regenerative potential. Nat. Cell Biol. *20*, 710–720.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell *58*, 610–620.

Kueh, H.Y., Champhekar, A., Champhekhar, A., Nutt, S.L., Elowitz, M.B., and Rothenberg, E.V. (2013). Positive feedback between PU.1 and the cell cycle controls myeloid differentiation. Science *341*, 670–673.

Kueh, H.Y., Yui, M.A., Ng, K.K., 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.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Laiosa, C.V., Stadtfeld, M., Xie, H., de Andres-Aguayo, L., and Graf, T. (2006). Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBP alpha and PU.1 transcription factors. Immunity *25*, 731–744.

Li, L., Leid, M., and Rothenberg, E.V. (2010). An early T cell lineage commitment checkpoint dependent on the transcription factor Bcl11b. Science *329*, 89.

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. Proc. Natl. Acad. Sci. *114*, 5800–5807.

Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. Nat Methods *11*, 360–361.

Mercer, E.M., Lin, Y.C., Benner, C., Jhunjhunwala, S., Dutkowski, J., Flores, M., Sigvardsson, M., Ideker, T., Glass, C.K., and Murre, C. (2011). Multilineage priming of enhancer repertoires precedes commitment to the b and myeloid cell lineages in hematopoietic progenitors. Immunity *35*, 413–425.

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 αβ T cell differentiation. Nat. Immunol. *14*, 619–632.

Ng, K.K., Yui, M.A., Mehta, A., Siu, S., Irwin, B., Pease, S., Hirose, S., Elowitz, M.B., Rothenberg, E.V., and Kueh, H.Y. (2018). A stochastic epigenetic switch controls the dynamics of T-cell lineage commitment. ELife *7*.

Olariu, V., Yui, M.A., Krupinski, P., Zhou, W., Deichmann, J., Andersson, E., Rothenberg, E.V., and Peterson, C. (2021). Multi-scale dynamical modeling of T cell development from an early thymic progenitor state to lineage commitment. Cell Rep. *34*, 108622.

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

Orkin, S.H. (2003). Priming the hematopoietic pump. Immunity *19*, 633–634.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell *163*, 1663–1677.

Pina, C., Fugazza, C., Tipping, A.J., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. Nat Cell Biol *14*, 287–294.

Porritt, H.E., Gordon, K., and Petrie, H.T. (2003). Kinetics of steady-state differentiation and mapping of intrathymic-signaling environments by stem cell transplantation in nonirradiated mice. J. Exp. Med. *198*, 957–962.

Pui, J.C., Allman, D., Xu, L., DeRocco, S., Karnell, F.G., Bakkour, S., Lee, J.Y., Kadesch, T., Hardy, R.R., Aster, J.C., et al. (1999). Notch1 expression in early lymphopoiesis influences B versus T lineage determination. Immunity *11*, 299–308.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017a). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017b). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315.

Radtke, F., Wilson, A., Stark, G., Bauer, M., van Meerwijk, J., MacDonald, H.R., and Aguet, M. (1999). Deficient T cell fate specification in mice with an induced inactivation of Notch1. Immunity *10*, 547–558.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol *4*, e309.

Ramond, C., Berthault, C., Burlen-Defranoux, O., de Sousa, A.P., Guy-Grand, D., Vieira, P., Pereira, P., and Cumano, A. (2014). Two waves of distinct hematopoietic progenitor cells colonize the fetal thymus. Nat Immunol *15*, 27–35.

Rothenberg, E.V., Moore, J.E., and Yui, M.A. (2008). Launching the T-cell-lineage developmental programme. Nat. Rev. Immunol. *8*, 9–21.

Rothenberg, E.V., Ungerbäck, J., and Champhekar, A. (2016). Forging T-Lymphocyte Identity: Intersecting Networks of Transcriptional Control. Adv. Immunol. *129*, 109–174.

Sambandam, A., Maillard, I., Zediak, V.P., Xu, L., Gerstein, R.M., Aster, J.C., Pear, W.S., and Bhandoola, A. (2005). Notch signaling controls the generation and differentiation of early T lineage progenitors. Nat. Immunol. *6*, 663.

Saran, N., Lyszkiewicz, M., Pommerencke, J., Witzlau, K., Vakilzadeh, R., Ballmaier, M., von Boehmer, H., and Krueger, A. (2010). Multiple extrathymic precursors contribute to T-cell development with different kinetics. Blood *115*, 1137–1144.

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. Baltim. Md 1950 *161*, 3984–3991.

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

Scripture-Adams, D.D., Damle, S.S., Li, L., Elihu, K.J., Qin, S., Arias, A.M., Butler, R.R., Champhekar, A., Zhang, J.A., and Rothenberg, E.V. (2014). GATA-3 dose-dependent checkpoints in early T cell commitment. J. Immunol. Baltim. Md 1950 *193*, 3470–3491.

Shah, S., Lubeck, E., Schwarzkopf, M., He, T.F., Greenbaum, A., Sohn, C.H., Lignell, A., Choi, H.M., Gradinaru, V., Pierce, N.A., et al. (2016a). Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. Development *143*, 2862–2867.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016b). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron *92*, 342–357.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*.

Strasser, A., Harris, A.W., and Cory, S. (1991). bcl-2 transgene inhibits T cell death and perturbs thymic self-censorship. Cell *67*, 889–899.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21.

Taghon, T.N., David, E.S., Zúñiga-Pflücker, J.C., and Rothenberg, E.V. (2005). Delayed, asynchronous, and reversible T-lineage specification induced by Notch/Delta signaling. Genes Dev *19*, 965–978.

Ting, C.N., Olson, M.C., Barton, K.P., and Leiden, J.M. (1996). Transcription factor GATA-3 is required for development of the T-cell lineage. Nature *384*, 474–478.

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

Ungerbäck, J., Hosokawa, H., Wang, X., Strid, T., Williams, B.A., Sigvardsson, M., and Rothenberg, E.V. (2018). Pioneering, chromatin remodeling, and epigenetic constraint in early T-cell gene regulation by SPI1 (PU.1). Genome Res. *28*, 1508–1519.

van Galen, P., Kreso, A., Wienholds, E., Laurenti, E., Eppert, K., Lechman, E.R., Mbong, N., Hermans, K., Dobson, S., April, C., et al. (2014). Reduced Lymphoid Lineage Priming Promotes Human Hematopoietic Stem Cell Expansion. Cell Stem Cell *14*, 94–106.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.

Wada, H., Masuda, K., Satoh, R., Kakugawa, K., Ikawa, T., Katsura, Y., and Kawamoto, H. (2008). Adult T-cell progenitors retain myeloid potential. Nature *452*, 768–772.

Waltman, L., and van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B *86*.

Wang, H., Zang, C., Taing, L., Arnett, K.L., Wong, Y.J., Pear, W.S., Blacklow, S.C., Liu, X.S., and Aster, J.C. (2014). NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. Proc. Natl. Acad. Sci. U. S. A. *111*, 705–710.

Weber, B.N., Chi, A.W.-S., Chavez, A., Yashiro-Ohtani, Y., Yang, Q., Shestova, O., and Bhandoola, A. (2011). A critical role for TCF-1 in T-lineage specification and differentiation. Nature *476*, 63–68.

Weinreb, C., Wolock, S., and Klein, A.M. (2018). SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. Bioinforma. Oxf. Engl. *34*, 1246–1248.

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.

Zandi, S., Åhsberg, J., Tsapogas, P., Stjernberg, J., Qian, H., and Sigvardsson, M. (2012). Single-cell analysis of early B-lymphocyte development suggests independent regulation of lineage specification and commitment in vivo. Proc Natl Acad Sci U A *109*, 15871–15876.

Zhang, J.A., Mortazavi, A., Williams, B.A., Wold, B.J., and Rothenberg, E.V. (2012). Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. Cell *149*, 467–482.

Zheng, S., Papalexi, E., Butler, A., Stephenson, W., and Satija, R. (2018). Molecular transitions in early progenitors during human cord blood hematopoiesis. Mol. Syst. Biol. *14*, e8041.

Chapter 3

SINGLE-CELL ANALYSIS OF THE TRANSCRIPTION FACTOR CONTROLLED EARLY T CELL DIFFERENTIATION DYNAMICS AND TRAJECTORY TOPOLOGY

ABSTRACT

The establishment of T-cell identity involves a series of signal-modulated gene network steps. In the previous study (Chapter2), while we established a detailed model of single-cell transcriptome dynamics during the transition from multipotentiality to T-cell lineage commitment, the functional roles of many of these genes remain obscure. In this study, we leveraged the *ex-vivo* differentiation systems, combined scRNA-seq with batch indexing and different perturbation strategies, to unravel the differentiation paths upon perturbations of single transcription factors (TFs) that are involved in setting up the early T-cell identity. Specifically, we examined the choices faced by the cells during commitment, under the control of an important TF Bcl11b. We found that cells without this critical TF quickly 'realized' the abnormality, around the stage where it is first expressed. And rather than a simple developmental block or regression to an earlier stage on the developmental trajectory, the cells took a diverging path to 'exit' the T-lineage. We also examined the more complexed regulatory network of early TFs that may be involved in controlling the alternative lineage suppression and T-lineage progression early on, which set up the population dynamic topology leading up to T-lineage commitment. Our results revealed the diverse and multi-module-spanning regulatory roles of these TFs in controlling the kinetics and differentiation outcomes of the earliest T cells.

INTRODUCTION

The establishment of T-cell identity emerges from multipotent precursors through a series of signal-modulated gene network steps (Porritt et al., 2003; Rothenberg et al., 2008; Taghon et al., 2005; Yui and Rothenberg, 2014; Yui et al., 2010). Commitment to the T-cell lineage begins in the thymus from the precursors trafficked from the bone marrow and is induced by complex spatiotemporal interactions between precursor T cells and thymic epithelial cells, among which Notch signaling from ligands on the thymic stroma is playing critical roles to suppress alternative lineage possibilities, as well as regulating cell proliferation (García-Peydró et al., 2006; Porritt et al., 2003; Romero-Wolf et al., 2020; Taghon et al., 2005). While this process takes place naturally in the thymus, it can also be replicated in *ex-vivo* cell culture systems such as OP9-DL1 monolayer co-culture (Schmitt and Zúñiga-Pflücker, 2002) and recently in a 3D serum-free artificial thymic organoid culture system, M-ATO (with DLL4) (Montel-Hagen et al., 2020).

The key remaining questions in the early T-cell developmental process are how the differentiation kinetics and population distributions are regulated by the gene regulatory network in early T cells. Unlike developing embryos, the differentiation timing from stem cells is not strictly deterministic. *In vivo*, individual hematopoietic stem/progenitor cells can make rather stochastic decisions about when to become activated (Busch et al., 2015; Naik et al., 2013). With the help of *ex-vivo* cell culture systems, we can more conveniently control the time and duration of signaling environment encounter. However, after days of differentiation, we will still yield cells in heterogeneous developmental states, due to the reasons discussed above. This makes bulk profiling techniques inadequate to study regulatory outcomes of population distributions and the topology of developmental trajectories. Therefore we combined scRNA-seq with Cell Hashing for sample and batch indexing (Stoeckius et al., 2018), and different perturbation strategies, to unravel the differentiation paths upon perturbation of single transcription factors that are involved in setting up early T-cell identity.

This paper uses single-cell transcriptome analyses of genetic perturbations to dissect two different phases of pro-T cell programming: (1) the choices faced by the cells during

commitment, under the control of the Bcl11b transcription factor, and (2) the more complex, previously obscure early gene regulatory network that guides hematopoietic precursors into the beginning of the T-cell program, which plays integrated roles in setting up the population dynamic topology leading up to T-lineage commitment.

First, the transcription factor Bcl11b is well known to be required for the development of αβ T cells and most γδ T cells, and its normal expression initiates precisely during commitment to the T-cell lineage in the thymus (Kueh et al., 2016; Liu et al., 2010; Shibata et al., 2014; Wakabayashi et al., 2003). Previous studies have shown that the progression towards the T-lineage is blocked or highly abnormal in cells that lack Bcl11b (Ikawa et al., 2010; Li et al., 2010a, 2010b; Longabaugh et al., 2017), including being prone to differentiate into natural killer (NK) cells in reduced Notch environment, and retaining abnormal expression of 'immature' or 'non-T' genes associated with stem, or other lineages (Hosokawa et al., 2018a; Li et al., 2010a, 2010b; Longabaugh et al., 2017). Although much previous work has identified differential expressed genes (DEGs) through bulk RNA-seq (Hosokawa et al., 2018a; Longabaugh et al., 2017), the pattern is quite abnormal compared to normal pro-T cells and does not appear to represent a simple developmental block. Our study first used the ATO with scRNA-seq to resolve the single-cell normal and defective developmental trajectories, revealing a 'realization' process in which the cells are fully set up to differentiate into T cells, but lack the critical transcription factor Bcl11b. Specially, we wanted to ask if these Bcl11b-absent-cells regress backwards in differentiation trajectory and resemble earlier cells, or do they bifurcate to an abnormal developmental trajectory? When do the 'abnormal genes' come up? Are the 'non-T' genes expressed in all Bcl11b-absent-cells or just a subset of them?

Second, we wanted to understand how the earliest transcription factors set up the cells to support or control differentiation and commitment, both in terms of speed and differentiation outcomes. Our recent study has characterized the fine gene expression pattern of murine pro-T cells prior to lineage commitment (Zhou et al., 2019). Importantly, we found that many 'multilineage' and 'stem' genes are expressed in distinct patterns in early stages, but their functional impacts were obscure. Are the stem and progenitor genes involved in establishing the normal differentiation speed and population distributions?

Meanwhile, TFs such as TCF1 (encoded by *Tcf7*, referred to as "Tcf7" below) and Gata3, two important regulators involved in early T-cell fate decisions [rev. in Yui and Rothenberg, 2014], are upregulated during the early stage (ETP) in normal development. Will the loss of Gata3 and Tcf7 result in an immediate change in earliest lineage decisions? What are the regulatory components involved? Our study optimized a strategy based on CRISPR/Cas9 KO 'perturbseq' system (Dixit et al., 2016), and took advantage of the 10X Chromium V3 chemistry's direct gRNA capture capability (Replogle et al., 2020), to resolve this finest detail of regulatory networks in primary *ex-vivo* derived early pro-T cells.

RESULTS

Fine comparison of time-controlled differentiation kinetics between BM derived *ex-vivo* culture systems and thymic early T cells

To choose a robust and precise time window that would be suitable for investigation of gene regulation during the murine T-cell commitment processes, we compared fine sub-populations of early pro-T cells generated from Bone Marrow (BM) precursors through different *ex-vivo* differentiation methods, as well as their *in vivo* thymocytes (Thy) counterparts (Figure 1A). Specifically, purified Lin- BM from C57BL/6 mice were either seeded on OP9-DLL1/DLL4 monolayer stroma culture with the standard serum condition, or aggregated with MS5-mDLL1/mDLL4 and deployed on culture inserts at the air-liquid interface, forming artificial thymic organoids (ATO) under serum-free conditions (Montel-Hagen et al., 2020). DN1, DN2a populations (sub-divided by Bcl11b-YFP positive and negative, indicated by "maturation steps" at the top of the heatmap) derived from different co-culture systems were sorted and sequenced through bulk RNA-seq. The gene expression profiles of the early pro-T populations generated through *ex-vivo* differentiation assays exhibited strong agreement on the important regulatory genes between different conditions, as well as agreement with their *in vivo* thymocyte counterparts (Figure 1B). The heatmap not only shows the agreement on the 'early-to-late stage' dynamic pattern ranges of expression, but also the absolute level of expression in FPKM values, that *ex-vivo* differentiated T cells in OP9-DLL1, ATO-DLL1, and ATO-DLL4 systems recapitulated accurately the thymic pro-T developmental regulations.

Despite the general agreement on important developmentally curated regulators, the transcriptome-level differentially gene expression analysis did show some differences between thymocytes and *ex-vivo* derived cells (Fig.S1A). Among all the conditions, ATO-DLL4 most closely recapitulated the thymic populations in both DN1 and DN2 stages, shown by the least number of differentially expressed genes compared to thymic counterparts (Fig.S1A-B). The only clear deviation from normal pro-T populations, among all eight tests performed, was the DN1 population derived from our OP9-DLL4 cultures, as it expressed many more myeloid genes than other samples. Also, the cells

derived from OP9-DLL systems generally express more *Id2* than other systems (Figure 1B), which encodes an important transcription factor promoting innate lineages such as NK, NKT, and ILCs.

Single-cell analysis aligns cells differentiated from BM to thymic early pro-T, in a continuum covering a good spectrum of early T differentiation

To determine whether the *ex-vivo* (*in vitro*) cultures gave full coverage of the fine population topology, we also performed a proof-of-concept scRNA-seq analysis with LSK derived pro-T cells and aligned with our previously published thymic T-cell data (Figure 1 C-D). We performed standard canonical correlation analysis (CCA) with 3000 anchor features using Seurat3, resulting in the well-aligned and intermixed *ex-vivo* derived and *in vivo* thymic early T-cell low dimensional representation (Figure 1D). Further cell cycle regression was performed to reveal only the fine developmental-related distribution of *ex-vivo* derived vs in vivo thymic cells (Fig S1C-E). It was striking to find that a rather pure precursor population (LSK) could give rise to a full spectrum of DN1-DN2b pro-T cells within 6 days under the controlled *ex-vivo* culture environment (Fig. S1F), and almost entirely recapitulating the thymic pro-T cell gene expression profiles on single-cell levels (Fig S1G).

The fine RNA-seq profiles together with the experience in previous cell culture assays (W. Z., M. R.-W., data not shown) helped us strategize the use of *ex-vivo* system for perturbation experiments at different stages to optimize cell output. We noticed that ATOs are good at generating consistent results and providing adequate numbers of cells at day 6 and beyond, while the OP9-DL1 system supports better proliferation and cell recovery early, especially under conditions that require viral vector delivery. Hence, we have utilized both ATO-DLL4 and OP9-DLL1 for the two different stages in which we investigated TF gene regulation using scRNA-seq for the rest of the study.

Differentiation kinetics and outcomes without an important T-lineage transcription factor – Bcl11b

We first wanted to use single-cell analysis to examine how the loss of Bcl11b alters the developmental trajectory of pro-T cells during commitment. Bcl11b is not expressed in the cells until they reach the commitment transition, after which it is expressed in all committed T-lineage cells(Kueh et al., 2016; Zhou et al., 2019). From previous bulk studies(Hosokawa et al., 2018a; Longabaugh et al., 2017), Bcl11b KO pro-T cells not only failed to suppress some of the early 'immature' or 'non-T' associated genes during commitment, but also abnormally turned on genes that were not active in normal precursors before Bcl11b expression onset. Therefore, loss of Bcl11b does not appear to represent a simple developmental block, but rather enables the cells to be diverted to an aberrant alternative differentiation pathway or pathways which have been completely uncharacterized until now. The genes upregulated in these Bcl11b KO cells appear to be a mixture of genes associated with various hematopoietic lineages raising the question of whether there is one abnormal trajectory promoted or several. Therefore, bulk studies have been unable to resolve the differentiation trajectory of the normal and abnormal differentiation process. It was also unclear whether the abnormal expression of different immature and alternative lineage-associated genes were occurring in the same cells or in different sub-populations of the Bcl11b KO cells. We utilized the long-term culture capability of the ATO system to compare the WT and Bcl11b knock out (referred as 'Bcl11b KO' or '11b KO' hereafter) differentiation processes at single-cell resolution, with two staggered timepoints, starting from LSK precursors purified from multiple individual animals (Figure 2A). For the Bcl11b KO, we used a conditional knockout strain with *Bcl11b*$^{flx/flx}$ and *Vav1-iCre*, which deletes the main functional coding domains of *Bcl11b* during early hematopoiesis before T-cell development begins. In each experiment, cells derived from different animals were tagged with different antibody-oligonucleotide conjugates using a "cell hashing" technique to give different barcodes to cells from each donor (Stoeckius et al., 2018). They were then pooled for scRNA-seq to serve as biological replicates or for control vs experimental comparisons within a single 10X v3 run.

First, the surface staining phenotype of ATOs consistently recapitulated thymic phenotypes of the Bcl11b KO pro-T cells (Figure 2B, S2A). We sorted Lin- CD45+

CD25+ cells from ATOs derived from LSK precursors from the two different genotypes, and the Bcl11b KO cells clearly exhibited higher cKit surface staining compared to WT, consistent with previous studies (Hosokawa et al., 2018a; Ikawa et al., 2010; Li et al., 2010a, 2010b; Longabaugh et al., 2017). The separation between KO and WT was more dramatic in the D13 culture compared to D10 (Figure 2B). Using the scRNA-seq data, we first compared the datasets in a 2-genotype-2-timepoints 'pseudo-bulk' analysis, testing the genes that were differentially expressed between D10 and D13 in the WT cells and in the Bcl11b KO cells, respectively (Figure 2C). Interestingly, among the genes that were differentially expressed between D10 and D13 in WT (total 1775 genes), 721 genes were not among the differentially expressed genes in Bcl11b KO; whereas among the genes that were differentially expressed between D10 and D13 in Bcl11b KO (total 1995 genes), 941 were not observed differentially expressed in WT (Figure 2C-D). This result indicates that a large fraction of genes that were differentially regulated, representing the differentiation progression happening between D10 and D13, were diverging between the two genotypes. We then looked more closely at the genes that were significantly differentially expressed in both genotypes between the two timepoints, as shown in Figure 2E. Most of the differential expression patterns of these genes agreed, although some were more highly expressed in the WT at both stages (Fig. 2E, cyan) and others always more highly expressed in the KO at both these stages (dark blue). However, a small group of abnormal genes were discordantly regulated from D10 to D13 (Fig. 2E, red). Some were up-regulated in Bcl11b KO but down-regulated in WT, including *Cd63, Tyrobp, Cd244, Cpa3*; and another small group of genes were down-regulated in Bcl11b KO but up-regulated in WT, including *Cd3g* and *Pcna* (Figure 2E). The genes that were completely differentially expressed in one but not the other are discussed in detail below.

Single-cell analysis: shifts in patterns in low dimensional space and the further WT-KO divergence at day 13

To ensure reproducibility, the single-cell analyses were performed through two separate scRNA-seq experiments, each with 6-8 hashtagged samples. To integrate the results, we first performed CCA integration for the two experiments, and subsequent clustering and

demultiplexing of the Cell Hashing index (Figure 2F,G, S2B). Upon demultiplexing, it was clear that the assay provided excellent reproducibility after integration, both on the biological replicate level across experiments, and between different animals of the same genotypes (Fig S2B-D). Although there was still substantial overlap between Bcl11b KO and WT patterns in D10, in UMAP 1 and 2, it was striking to see that the two genotypes took very different paths in D13 (Figure 2G, Fig. S2B). Based on gene expression patterns in these clusters (Fig. S2G, also from Zhou et al 2019(Zhou et al., 2019)), at D10 there were still many cells representing the most immature states present in both cultures (clusters 12, 6, 3), and many cells at intermediate stages showing enrichment for cell cycle genes (clusters 7, 5). By D13, the putative immature clusters were depleted, while WT cells increasingly shifted to more mature stages in clusters 4, 2, 9, and 10 (based on genes shown in Fig. S2G), whereas the Bcl11b KO cells showed a strong enrichment in clusters 0, 1, and 8 instead.

To more accurately measure the low dimensional space patterns and cluster distributions of all individual samples, we calculated the pair-wise Kullback-Leibler (KL) divergence between all the samples based on each sample's cluster distributions, as shown in the heatmap in Figure 2H. First, the KL divergence confirmed the pattern agreements between all the experimental and biological replicates, and the contrast between WT and Bcl11b KO samples. Second, the KL divergence calculation also agreed with the visual pattern that there was further divergence between WT and Bcl11b KO progressing with time. The shifts of cluster distributions with genotype could be investigated sample by sample, in experiment 2, where two timepoints were collected (Fig.S2 E-F). Because KL divergence is calculated based on each sample's cluster distributions, it is important to pinpoint which were the most quantitatively important clusters that describe the genotype differences. Scatterplots of cluster proportions between WT and Bcl11b KO showed that in both D10 and D13, cluster 2 was always enriched in WT and cluster 0 was always enriched in the KO (Figure 2I). Indeed, a finer look at each cluster's expression profile (Fig S2G) and each sample's UMAP 1 and 2 pattern suggested that the clusters 0 and 2 potentially indicated where WT and Bcl11b KO were 'branching' into the 2 different directions. We also performed differential gene expression analysis between clusters 0 and 2, and the

expression pattern of top 20 genes enriched in either cluster are shown in the heatmap (Figure 2J). It is particularly interesting to see that the cluster 2 enriched genes were mostly T-lineage associated genes while the cluster 0 enriched genes were 'immature' or 'innate immune' cell associated genes, and many were known Bcl11b repression targets (Hosokawa et al., 2018a; Longabaugh et al., 2017). These genes were clearly differentially expressed between clusters 0 and 2, but even more differentially expressed further down the 'branches' (clusters 2,9,10 for WT and 0,1,8 for Bcl11b KO, respectively).

The beginning of divergence between genotypes happened in clusters with high cell cycle activities, and the differentially expressed genes have different timing

The potential differentiation trajectories of WT and Bcl11b KO could be sketched manually, based on the connectivity in UMAP 1 and 2 and expression patterns of lineage markers (Figure 3A-C, S2G). Of special interest was cluster 5, which was enriched for proliferation-associated genes (Fig. S2G), had a loop-like topology, and was represented in both genotypes at both timepoints. Interestingly, when we highlighted the cells in cluster 5, we observed that the WT and Bcl11b KO formed parallel but slightly separate loops (Figure 3B). This suggested that the branching of the WT vs KO trajectory may happen earlier, around the seemingly mixed 'proliferation stage', before branching further into cluster 0 and cluster 2, as discussed before (Figure 3A). To test this hypothesis, we performed differential gene expression analysis between WT and Bcl11b KO, only on cells within cluster 5 (Figure 3D). To our surprise, a large number of genes were significantly differentially expressed between WT and Bcl11b KO in cluster 5 only, clearly substantiating the separation observed in Figure 3B. These differentially expressed genes overlapped largely with genes noted to be differentially expressed between cluster 0 and cluster 2 (cf. Fig. 2J). Moreover, we observed that even within the same cluster which was present at both timepoints, namely cluster 5, WT to KO difference at D13 could still be bigger than in D10. For example, expressions of *Ikzf2* (same for *S100a10* and *Itga4*), in D10 Bcl11b KO cells were not very different compared to D10 WTs, while at D13 these genes were dramatically upregulated in Bcl11b KO cells in cluster 5, compared to WT cells in cluster 5. Thus, clearly the 'branching' between WT and Bcl11b KO

already occurred earlier in the trajectory, at cluster 5, despite the close proximity in low dimensional space. The genotype-specific differentially expressed genes in cluster 5 are shown in Fig. 3D. In addition, the added time resolution in the heatmap Figure 3D showed that the accumulation of subtle abnormal expression features can precede substantial movement between discrete clusters in low dimensional space.

The destiny of cells after the realization of their inability to become T cells

Although Bcl11b KO cells have previously been noted to express 'immature' or 'non-T' genes (Hosokawa et al., 2018a; Longabaugh et al., 2017), it was unknown whether the abnormally expressed B-, NK-, other innate-, stem-associated genes were all expressed in the same cells or segregated among different cells that lack Bcl11b. Our data now showed that cells most highly expressing, NK (*Il2rb*) and ILC (*Rora, Zbtb16*) genes were concentrated at the 10 o'clock 'tip' within the UMAP1-UMAP2 space (cluster 8) which seems to be the 'exit' point of the Bcl11b KO developmental progression (Figure 3C). *Myo1e*, a B cell and ILC-associated gene, was more spread across cluster 1 and cluster 8. *Cd163l1*, a gene associated with the TCRγδ cell lineage, highlighted cluster 8 and far left of cluster 1. Fig. S3C-F shows a detailed analysis of the expression patterns of many of these genes within the Bcl11b KO cells specifically. Although the different genes upregulated in Bcl11b KO cells were found more or less spread across the population distribution, nearly all reached their highest levels at the same 'tip'. Therefore, the Bcl11b KO developmental pathway appeared to progress toward a major single endpoint, distinct from the T-cell program, not several diverse alternative states.

To gain more insight into the processes that drive the program for Bcl11b KO cells, we computationally subset only Bcl11b KO cells, and performed co-regulated 'gene module' analysis with Monocle3, extracting modules of genes which help to further sub-define the component states within the Bcl11b KO population (Fig S3A-D). In Bcl11b KO only UMAP 1-3 (Fig S3E), with this fine resolution visualization, we observed that the 'tip of exit' showed sign of loss of Notch response genes, such as *Nrarp, Il2ra*. Cells around the tip further upregulated *Gata3, Id2, Ikzf2*, while downregulating *Rag1*. The orthogonal UMAP 1-2 space also showed a small cluster, shown in Fig. S3F (also seen in Figure 2F,

S2G cluster 13), has a small distinct cell cluster with the expression of interferon-response genes, *Ifit1* and *Ifit3*. Thus, the end stage of the Bcl11b KO pathway may involve the upregulation effort of many 'multi-lineage' genes, but at the end, it was clear that the D13 terminus of the Bcl11b KO program 'exiting point' involved at least partial downregulation of Notch signaling.

These results thus show that pro-T cells reaching the threshold of commitment, when they would normally upregulate *Bcl11b,* respond to the lacking of Bcl11b by gradually accumulating 'early stage' genes that would normally be turned-off around this stage, and progressively upregulate 'abnormal' genes in a uni-lineage path diverging from the normal developmental trajectory. The cells lack Bcl11b do not regress back along the normal developmental trajectory, but rather go on an aberrant branch that would finally lead to a small fraction of cells 'exiting' the 'T-programs' through the downregulation of Notch signaling responses and further upregulation of the innate lineage associated expression features.

TFs with early stage dynamic expression patterns are examined through batch controlled dual gRNA direct capture perturb-seq

Many regulatory network changes seem to precede the T-lineage commitment decision(Zhou et al., 2019), but the basis of their regulation has been poorly understood. This is partly due to the rarity of the cells in these stages and to the overlap between TFs expressed in the earliest T-cell precursors and those used in other hematopoietic precursors. To reveal the underlying network topology, we examined the effects on differentiation speed and outcomes when we knocked out candidate regulators of these early events, specifically candidate TFs for regulatory functions that also exhibit dynamic expression pattern changes during early stages in differentiation (Figure 4A). In particular, *Bcl11a, Spi1, Hoxa9, Meis1*, and *Erg* encode stem-progenitor associated factors, whereas *Tcf7* and *Gata3* encode T-lineage associated factors, and we focused on these for targeted KO studies. Note that several of these genes have previously been shown to be important for viability of early T-cell precursors (Champhekar et al., 2015; Germar et al., 2011; Hosokawa et al., 2018b; Scripture-Adams et al., 2014; Yu et al., 2012), but we ensured

maximum retrieval of viable perturbed cells by including a *Bcl2* transgene in the Cas9 mice. KO effects were examined by introducing 2-3 sgRNAs against the gene of interest, delivered through retrovirus, to Lin^- or LSK prethymic precursor cells from bone marrow that were purified from mice with constitutive expression of Cas9 proteins. After transduction, the cells were then cultured in either ATO-DLL4 or OP9-DLL1 systems to initiate the T-cell program (details in Methods; see Figs S1, S2). Interestingly, in various experimental settings in preliminary studies, KOs of 'stem'-related genes, *Bcl11a* or *Meis1*, promoted a 'faster' developmental phenotype promoting the DN2a to DN2b transition, as indicated by surface marker combinations of CD44, cKit and CD25. *Erg* KO increased the rate of DN1 to DN2 transition (examples shown in Figure 4B). KO of *Spi1*, encoding PU.1, a pioneer factor which is important for myeloid vs lymphoid decision, was previously shown to accelerate DN2 to DN3 progression if deleted after T-cell development was initiated (Champhekar et al., 2015; Hosokawa et al., 2018b; Scripture-Adams et al., 2014). Here, with earlier deletion, it lowered the CD44 level in all CD25+ cells, but retained a population that is $CD25^-CD44^+$. However, we did often observe variations in infection rates, cell number yields, and phenotype inconsistency (between culture systems or experiments during screens of different knockouts). Therefore, to combat variability and non-autonomous effects, we designed a definitive, highly-controlled scRNA-seq assay to determine whole-transcriptome effects of these perturbations, using a pool-synthesized, batch-controlled dual gRNA system (Figure 4C-D).

Briefly, paired gRNAs (each pair designed against the same exon of the same gene to ensure sufficient KO) were synthesized through array-based oligo synthesis, then the oligo pool was PCR amplified and Gibson assembled to generate the paired dual gRNA insert pool, and subsequently incorporated into retroviral backbone and packaged into the final retroviral vector library. The paired gRNAs being expressed were compatible with direct capture of both gRNA sequences in scRNA-seq using 10X Chromium V3 chemistry ('Cap1' and 'Cap2' in Figure 4C, details and description of quality controls and titration given in Methods; Figs. S4A, S4B). Importantly, each packaged retroviral pool was titered on the same type of primary cells to precisely target multiplicity of infection (MOI),

0.5-1 in this study, to ensure single gene effects (Fig S4B, Methods). Here, we used CRISPR/Cas9 in the KO context, similarly to the original perturbseq setup(Dixit et al., 2016). FACS sorted precursor LSK cells were infected with retroviral pools at MOI of 0.5-1, then cultured on OP9-DL1 for 5 days differentiation before FACS sorting to purify the infection positive populations for scRNA-seq (Figure 4D, details in Methods). To further minimize batch variations, we multiplexed multiple batches of biological replicates using antibody-conjugated cell-hashing technique (Stoeckius et al., 2018). After sequencing libraries of cDNA, gRNA, and hashtags yielded from the same experiment, the three respective FASTQ files were aligned with CellRanger3 (for cDNA) and an in-house pipeline (for gRNA and hashtags, Fig S4C-D, details in Methods). We also performed additional validation experiments of dual gRNA in our primary cell differentiation system, showing that two gRNAs in the same vector did improve upon single gRNA KO effect (Fig S4E-F). The multiple replicate infections, multiple sgRNA pairs against the same target, and pooled sequencing of all in the same 10X v3 analysis yielded a high quality, internally controlled resource of data in which to identify specific perturbation effects.

Dramatic changes in topology with single TF perturbations

It is important to note that the 5-day-culture of LSK with Notch signaling is aiming to look at the immediate loss of function effect at the very early stage of differentiation (before any identity establishment), as normally 5 days post infection, the majority of LSK cells are still in DN1 stage by surface markers. However, the scRNA-seq result revealed that some of the individual TF perturbations already resulted in dramatic shifts in low dimensional representation (Figure 4E-G, Fig. S4G). While control cells spread across the UMAP1 and 2 space, forming a sparse differentiation continuum (Figure 4F left), cells expressing sg.Tcf7 or sg.Gata3 stalled at the more un-differentiated stage; cells expressing sg.Spi1 shifted towards more differentiated stage but slightly veering from the control trajectory; and cells expressing sg.Erg formed a distinct cluster, aligned parallel to the ‘normal’ trajectory with some DN2 signature genes enriched (Figure 4F-G, S4G, S5D). In the entire pool of cells, more sg.Erg expressing cells were detected than any other cells,

implying that Erg loss may enhance proliferation (Figure 4H). In contrast, Tcf7 and Meis1 perturbations slightly suppressed proliferation despite the presence of anti-apoptotic Bcl2, as all pairs of gRNAs of these genes showed consistent cell number differences compared to the control vectors (Figure 4H). For simplicity, hereafter, we refer to the integrated effect of all sg.RNA against the same genes as 'KO'.

To describe the cluster distribution relationships of different KOs, pair-wise KL divergence was calculated. This showed that KOs of Tcf7, Erg, and Spi1 exhibited dramatic cluster pattern differences among each other, and they were also very distinct from the control (Figure 4I, individual pairs of gRNA's cluster distribution in Fig S5A). However, except for the clusters specific for the Erg KO cells, most of the KO and controls still shared the same common clusters, although their distributions among these clusters varied greatly, as shown in Figure 4E. In light of the results previously shown for the Bcl11b KO in the later-stage cells (Fig. 3A, B, D), we wondered whether the gene expression profiles of KO and control within the same 'common clusters' were different here. However, here the representative scatter plots (Fig S5B), correlation heatmaps (Fig S5C), and differential gene expression analysis (not shown) between Cont and KOs all showed that for cells within the same 'common' clusters, the gene expression patterns between control and KOs were rather similar, in contrast to the behavior described previously for cluster 5 of the Bcl11b KO cells and WT cells. Therefore, we conclude that the effects of these early KO perturbations were mostly described by the shifts between clusters in low dimensional space, not by subtle changes of individual genes' expression in the same shared clusters.

Differentially expressed genes between different KOs are partially explained by the early to late progression in control cells

The key questions we sought to answer in knocking out TFs were a) what target genes they regulated and b) what their potential functional implications were during this early differentiation process. To first examine the regulated targets, we performed differential expression analysis looking at the top up- and down- regulated genes in response to each of the KOs. All KOs except Hoxa9 and Meis1 showed many significantly differentially

expressed genes in response to the KOs (Figure 5A-B). The Bcl11a KO up-regulated 'late' genes and down-regulated the 'stem' genes, while the Spi1 KO partially overlapped with the Bcl11a KO's differential expression pattern but induced more innate-immune genes such as *S100a6* and *Ifitm2*, and also upregulated transcripts of the T-cell receptor γ locus, *Tcrg-C4*. Erg KO cells upregulated many cytoskeleton- and growth/signaling-related genes, and downregulated *Ctla2a* and *Myl10*. Tcf7 KO cells showed expression enriched for stem-related genes such as *Sox4* and *Mef2c*, as well as *Malat1*, a lncRNA known for involvement of nuclear-speckle and mRNA splicing. While Gata3 KO cells were also enriched for some stem-related genes, they were not enriched for expression of *Sox4*; instead, they noticeably upregulated DC and macrophage-related genes, as well as cell cycle related genes which will be discussed later.

We then asked if the genes up-regulated and down-regulated in each of the KOs and shifts of the KO cells from control or 'normal' differentiation trajectories could explain each other. To resolve this, we compared each knockout with a differential expression analysis of the control cells only, defined by the gene expression changes that occur normally in the stages from early DN1 to late DN2b, i.e. (in only the Cont) between early clusters (7,1,11) and late clusters (0,6,5,8,9,14), as shown in Fig S5F. For example, Bcl11a and Spi1 KOs both pushed the cells away from clusters representing the most un-differentiated states (Figure 4F, S4G); Were all the genes these cells up-regulate essentially 'late' or 'DN2' genes, such as *Cpa3* and *Fgf3*? Tcf7 and Gata3 KO both had stalled differentiation at different stages (Figure 4F, Fig. S4G). Was this why both KO conditions seemed to have enriched expressions of at least some 'stem'-related genes, such as *Mef2c* and *Bcl11a*? To generalize, could most of the differentially expressed genes be simply explained by the shifts of 'early' to 'late' stages that normally happen during differentiation? The results of this comparison showed that while many changes in Tcf7 knockouts seemed to follow predictions of a simple developmental block (Fig. S5E), many of the differentially expressed genes in other KO conditions changed in ways that are not explained purely by the early to late transition expression profile changes in control (Fig.S5E-I, dark brown dots).

KOs resulted in dramatic changes in cell cycles and different lineage program regulations

To further understand the potential changes of regulatory activities induced by deletion of these TFs, we used SCENIC (Aibar et al., 2017) algorithm to identify groups of genes correlated with, and potentially regulated by the same "central" factors. The SCENIC analysis was performed on the computationally separated individual KOs (Fig.S6), and we found that these predicted "central" factors and their potential target genes responded distinctively in the different individual KOs. SCENIC infers probable network connections from the enrichment of the target motif for TFs in genomic regions near TSS's of genes co-expressed with those TFs, and integrate the expressions of the potential target genes as the predicted TF activities, namely 'regulons'. Among the most significant regulon activities identified by SCENIC, it revealed several coherent regulons that responded in markedly different ways in our KOs (Fig. 5E). Interestingly, SCENIC predicted a 'central' TF, Ybx1, governing a prominent regulon in all samples (shown in Fig S6), the function of which has not been extensively studied in hematopoietic systems. However, inferred Ybx1 activity appeared strongly correlated with cell cycle: among the genes co-expressed with Ybx1, links between Ybx1, cytoskeleton genes and G2/S/proliferation markers, such as Birc5 and Hmgb1, were discovered (Figure 5C-E, detailed in Fig S6 and S7C-D). Moreover, the distribution across the cell cycle/proliferative stages changed dramatically upon perturbations of some of the genes, as shown in Figure 5C-E. The Erg KO shifted cells to a highly proliferative stage, and the Gata3 and Bcl11a KOs also seemed to promote proliferation, whereas the Spi1 and Tcf7 KOs seemed to suppress proliferation (Figure 5C). Both Gata3 and Erg KO further induced a 'new' Hmgb1 module which potentially further extends the cell cycle and cytoskeleton regulation. The observation on cell cycle regulation in pair-wise plots and SCENIC analysis substantiated the earlier transcriptome-wide differential expression analysis in Figure 5A, where Gata3 KO upregulated *Mki67* and *Top2a*, two canonical G2/S/proliferative markers, which are profoundly downregulated in Tcf7 KO. Moreover, SCENIC predicted that integrated Myc 'regulon' activities are significantly decreased in Tcf7 and Meis1 KOs, but significantly increased in Erg KO (Figure S7A). Cell cycle

stages, RNA contents, and expression of cytoskeleton genes were further compared between control and all KOs, and the results support these findings on Erg, Gata3, Meis1, and Tcf7 KO effects (Figure S7B-D).

Besides cell cycle, cytoskeleton, and Myc related regulations, we focused on developmental and lineage-associated regulons. Although Gata3 and Tcf7 both showed a 'stalled phenotype' on differentiation space, their up-regulated and down-regulated genes had minimum overlaps. The Gata3 KO seemed to be more encouraging for innate lineages, upregulating *S100a4, Ifitm1,* and *Ifitm3d*, whereas cells without Tcf7 downregulated these genes, and upregulated *Tyrobp* and *Sox4* instead (Figure 5A). Indeed, the SCENIC summarized TF-target regulation prediction showed evidence for differential regulation of Spi1 (PU.1), Irf8, and C/EBP family TF activities upon deletion of *Tcf7, Gata3*, and *Erg* (Summarized in Figure 5E, details in Fig S6). Loss of Gata3 immediately promoted all myeloid programs including upregulating Irf8 (DC) and C/EBP family genes (MF), while supporting proliferation (Olsson et al., 2016). However, loss of *Tcf7* only retained Irf8 and Spi1 expressions, but surprisingly did not promote a MF/GN module involving Spi1 and C/EBP family TF activities (Figure 5E, S7E-H). Moreover, it is not surprising that Tcf7 KO completely abolished the T-lineage regulatory module (Figure 5E). The Erg KO not only suppressed the myeloid regulating modules, but also the stem module (Figure 5E). Therefore, it was clear that KOs of individual TFs could lead to complexed changes in regulatory activities in developing early T cells, including proliferation, Myc activity, alternative lineage programs, and T-lineage programs.

KO of individual TF shifts cells in T differentiation pseudotime trajectory

To better describe the differentiation continuum along T fate lineage, we performed trajectory and pseudotime analysis with Monocle3 in 3D UMAP space, as shown in Figure 5F. The predicted T developmental pseudotime of individual cells from different KOs are represented in Figure 5G. Kruskal-Wallis test has shown the significant pseudotime acceleration shift of Bcl11a, Spi1, and Erg KOs from control, and significant deceleration of Gata3 and Tcf7 KOs from control. The results indicate the expression of Spi1, Bcl11a and Erg in early T cells slows down the differentiation process, while expression of Gata3

and Tcf7 are used (or required) for the advancement T-lineage, which agrees with previous studies (Champhekar et al., 2015; Scripture-Adams et al., 2014; Weber et al., 2011). The unsupervised trajectory and pseudotime analysis based on full transcriptome profiles further substantiated that these TFs are involved in regulating the differentiation speed. Clearly, these genes expressed in the early stage of T cells are functionally relevant in setting up the proper proliferative state and differentiation speed of the cells. Our results revealed the diverse and multi-module-spanning regulatory roles of these TFs in controlling the kinetics of early T-lineage differentiation.

DISCUSSION

The perception of cell fate commitment is usually a discrete and irreversible decision point made by the multipotent cells, resulting in the acute loss of alternative lineage potentials. However, for a stem cell population that is regulated under a non-fully deterministic gene regulatory network, we need to view the process as a developmental or decision continuum, that are potentially coupled by multi-module regulations, in order to understand the requirement for cell fate commitment 'event' and the modulators for the speed of the differentiation process. Indeed, many previous studies have shown that hematopoietic differentiation processes often involve gradual exclusions of alternative lineages. But in T cells, the process does not seem to be regulated by any single 'master regulator' (Longabaugh et al., 2017; Mingueneau et al., 2013; Naito et al., 2011; Thompson and Zúñiga-Pflücker, 2011; Yui and Rothenberg, 2014; Zhang et al., 2012). Previous studies have shown perturbations of some of the regulators either in later stages of development or in slightly different settings, such as fetal liver precursors (Champhekar et al., 2015; García-Ojeda et al., 2013; Germar et al., 2011; Hosokawa et al., 2018a, 2018a, 2018b; Longabaugh et al., 2017; Scripture-Adams et al., 2014; Weber et al., 2011; Yu et al., 2012). Most importantly, the lack of single-cell resolution has previously made it impossible to understand the full developmental continuum and trajectory topology; any interpretation of bulk RNA-seq perturbation data could have been an averaging effect of mixed regulatory outcomes on single-cell levels. In this study, we used scRNA-seq coupled with different carefully designed- and optimized- perturbation strategies to further examine the loss of function outcomes of regulators to this early T-cell differentiation continuum. Specifically, two different stages of the cell fate decision process were focused: with Bcl11b KO – around and right after lineage commitment; and with selected TF-'perturbseq' – prior and leading up to lineage commitment.

Our results showed a clear branching developmental trajectory through the lineage commitment stage of WT and Bcl11b KO precursors. The loss of Bcl11b impacted cells early on after Bcl11b would normally begin to be expressed, in the proliferative DN2 stage, where immediate divergence of normal and defective trajectory occurs. Despite the

clustering algorithm assignment of WT and KO to the same cluster, the changes in the expression pattern preceded their shifts in low dimensional space. Furthermore, the cells without Bcl11b seemed to undergo a major uni-lineage path of accumulation of many 'abnormal' genes in the same cells over time, through which some of the Notch response genes were transiently upregulated, but eventually lost. Notably, upon the final loss of Notch signaling, many NK or ILC-associated genes were turned on together, potentially representing the 'exiting' path to become 'NK-like' or 'innate immune' cells observed in previous studies (Hosokawa et al., 2018a; Li et al., 2010b).

Previous study of *in vivo* thymocytes, using the highly sensitive imaging-based seqFISH method, showed that within the ETP state, the majority of individual cells co-express legacy progenitor genes with the critical Notch-induced T-cell regulatory genes, *Gata3* and *Tcf7* (Zhou et al., 2019). This implies that the stem and Notch-induced regulatory modules operate together to potentially set up lineage progression to DN2 stage, which later leads to lineage commitment. This study examined individual TF perturbations, including *Gata3* and *Tcf7* themselves, around the ETP equivalent stage. First, our result showed that deletion of some stem-related genes, like *Bcl11a* and *Spi1*, shifted cells to a more differentiated state, and Erg KO shifted the cells to an aberrant proliferative DN2 state. Second, SCENIC analysis suggested that some of the TFs are controlling not only the differentiation state but also are involved, directly or indirectly, in controlling proliferation, cytoskeleton and Myc activity modules. Our results showed surprising contrast of myeloid and proliferation modules regulated by Gata3 and Tcf7: Gata3 was involved in suppressing many myeloid lineages and promoting T-lineages; whereas Tcf7 was absolutely required for setting up the T-lineage program at the earliest stage, but did not appear directly to suppress myeloid lineages. Although Gata3 and Tcf7 are both known, essential T-cell regulatory factors, our analysis showed that the presence of Gata3 normally seemed to suppress proliferation while Tcf7 seemed to promote proliferation. Finally, it was clear that perturbation of individual TFs can significantly shift differentiation kinetics, indicated through pseudotime calculations. Bcl11a, Spi1, and Erg KOs accelerated the differentiation process while Gata3 and Tcf7 KOs stalled it. The

results imply that some of the stem related genes naturally hold back the differentiation speed, while Gata3 and Tcf7 at least are promoting the differentiation process.

It is fair to conclude that developmental timing is initially restrained through a network of positively cross-regulating transcription factors that actively maintain stem-like properties, which are expressed in the progenitors. The differentiation progression then results from the tipping of the balance between the differentiation-promoting gene network regulators and the stem-like gene network regulators. Therefore, some of the early stage transcription factors are likely controlling the balance, hence differentiation kinetics.

In summary, we presented a detailed single-cell study of the population distributions, trajectory topology, and differentiation kinetics upon knocking out important regulators during or before the T-cell identity establishment.

ACKNOWLEDGMENTS

We thank Jeff Park and Sisi Chen from the Caltech Single Cell Profiling and Engineering Center for providing supports for 10X Chromium experiments, Rochelle Diamond and members of the Caltech Flow Cytometry facility for sorting, Ingrid Soto for mouse care, Maria Quiloan for mouse genotyping, Igor Antoshechkin and Vijaya Kumar of the Caltech Jacobs Genomics Facility for bulk RNA sequencing. Support for this project came from USPHS grants (R01HL119102 and R01HD076915) to E.V.R., the Beckman Institute at Caltech for support of all the Caltech facilities, the Biology and Biological Engineering Division Bowes Leadership Chair Fund, the Louis A. Garfinkle Memorial Laboratory Fund, the Al Sherman Foundation, and the Albert Billings Ruddock Professorship to E.V.R.

AUTHORSHIP STATEMENT: W.Z. designed the project, carried out the experiments, analyzed the data, and wrote the paper. F.G. wrote the in-house bioinformatic pipeline for perturbseq and hashtag alignment and assignment. M.R.W. and S.J. performed

preliminary experiments. E.V.R. supervised research, guided the design of the project, and wrote the paper.

MAIN FIGURES

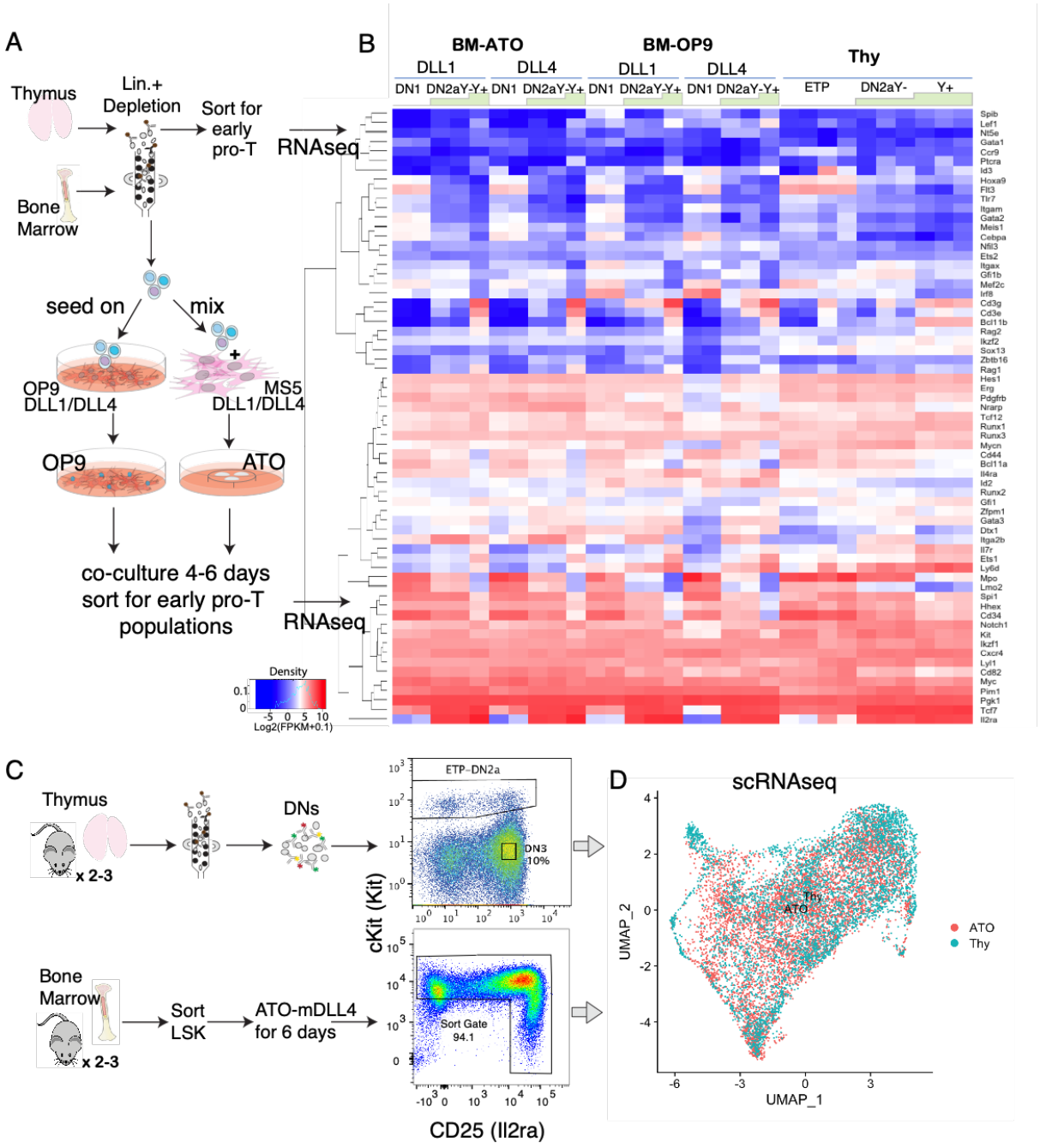


Figure 1. BM derived early T cells' expression profiles recapitulated *in vivo* thymic early T cells', on bulk and single-cell levels.

A) Illustration of early T cells harvested from thymus or derived from bone marrow (BM) precursors. The *ex-vivo* culture systems include both the OP9-DL co-culture system with

OP9-DLL1 or OP9-DLL4 stroma cells, and the 3D artificial thymic organoid (ATO) system with MS5-mDLL1 or MS5-mDLL4 stroma cells, as detailed in Methods. *B)* Clustered expression heatmap of bulk RNA-seq measurements comparing early T cells harvested *in vivo* and early T cells derived from BM as illustrated in A). All genes plotted are from a list of curated important regulatory gene list described in the previous study (Zhou et al., 2019). Color scales indicate raw expression levels as log(FPKM+0.1), without row normalization. *C)* Illustration of sample purification procedures and FACS sorting strategies for the scRNA-seq experiments, comparing *in vivo* and *ex-vivo* derived early T cells' single-cell expression profiles. *D)* Aligned *in vivo* and *ex-vivo* derived scRNA-seq profiles after CCA scaling, shown in UMAP1-2. More detailed analysis of the aligned scRNA-seq profile and comparisons are shown in Figure S1.

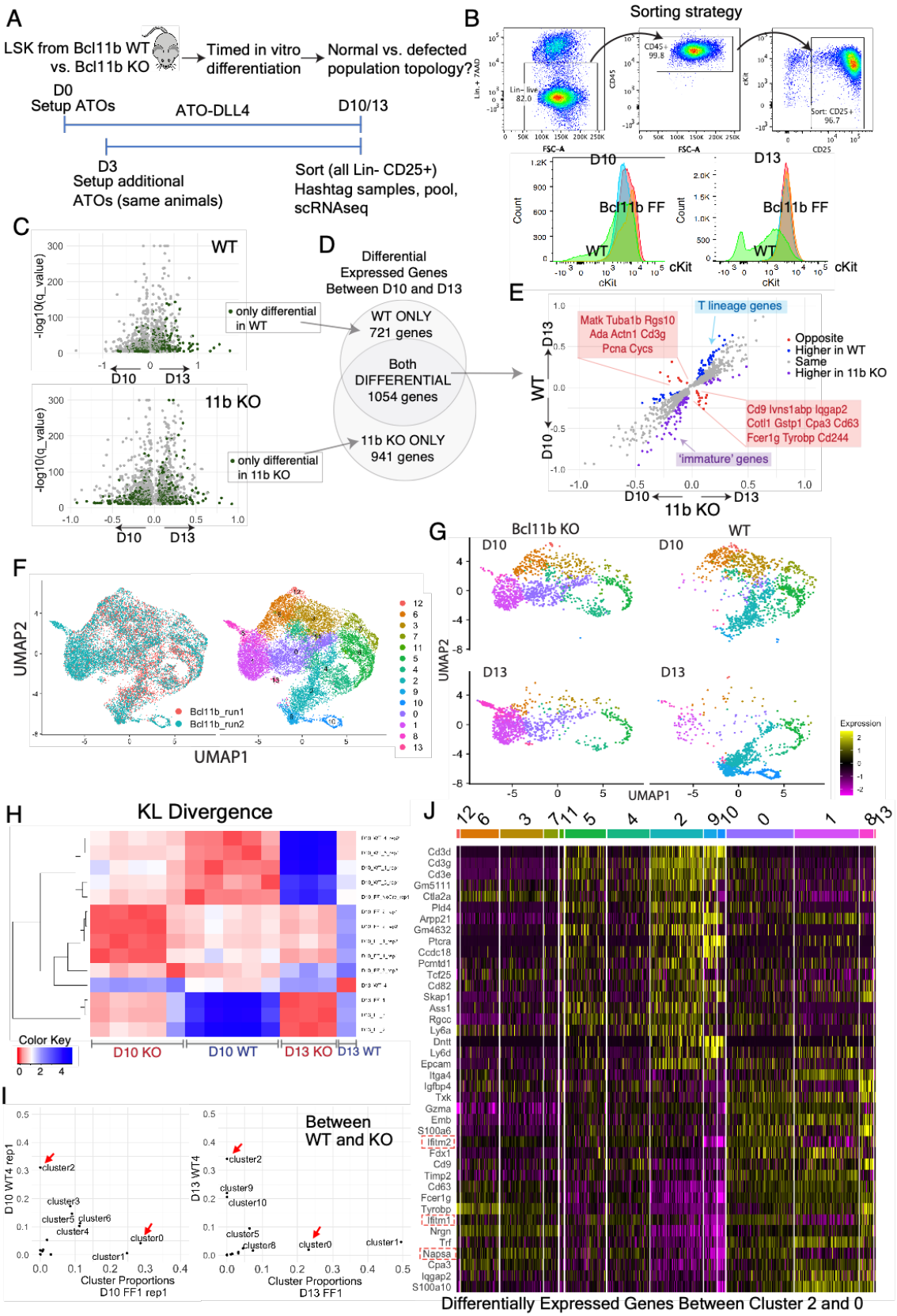

A
LSK from Bcl11b WT vs. Bcl11b KO
Timed in vitro differentiation
Normal vs. defected population topology?
D0
Setup ATOs
ATO-DLL4
D10/13
D3
Setup additional ATOs (same animals)
Sort (all Lin- CD25+) Hashtag samples, pool, scRNAseq
B
Sorting strategy
Lin- live 82.0
CD45+ 99.8
Sort: CD25+ 96.7
D10
D13
Bcl11b FF
WT
cKit
Count
C
WT
11b KO
-log10(q_value)
D10
D13
D
Differential Expressed Genes Between D10 and D13
only differential in WT
only differential in 11b KO
WT ONLY 721 genes
Both DIFFERENTIAL 1054 genes
11b KO ONLY 941 genes
E
T lineage genes
Matk Tuba1b Rgs10 Ada Actn1 Cd3g Pcna Cycs
Opposite
Higher in WT
Same
Higher in 11b KO
Cd9 Ivns1abp Iqgap2 Cotl1 Gstp1 Cpa3 Cd63 Fcer1g Tyrobp Cd244
'immature' genes
F
UMAP2
UMAP1
Bcl11b_run1
Bcl11b_run2
G
Bcl11b KO
WT
Expression
H
KL Divergence
Color Key
D10 KO
D10 WT
D13 KO
D13 WT
I
Between WT and KO
D10 WT4 rep1
D13 WT4
Cluster Proportions D10 FF1 rep1
Cluster Proportions D13 FF1
J
Differentially Expressed Genes Between Cluster 2 and 0

Figure 2. (Previous page) Single-cell population distributions of Bcl11b knockouts revealed time-progressed abnormality compared to wild type.

A) Schematics showing the experimental design and setup of internal controlled scRNA-seq experiments comparing wildtype (WT) and Bcl11b knockout (11b KO). *B)* Top panels describe the FACS purification strategy for the Bcl11b scRNA-seq experiments. Bottom panels summarize the surface staining phenotype of cKit levels in WT and Bcl11b KO (noted as FF for flx/flx, details in Methods). *C-E)* Differential expression analysis of expression profiles from cells that were harvested on D10 and D13, in WT and Bcl11b KO, separately. *C)* Volcano plots of genes with differential expression between D10 and D13 (x axis represent the 'estimates' from the generalized linear model fit of gene expression with respect to time), and their adjusted p-values (qval, on a $\log_{10}$ scale). The dot color represents whether the identified gene's differential expression is also tested differential in the other genotype. *D)* Venn diagram showing the number of differentially expressed genes' overlaps between the two genotypes. *E)* Scatterplot showing the genes that were significantly differential expressed in both of the genotypes. Comparing the WT and Bcl11b KO's 'estimates', showing whether the directions of differential expression regulations with respect to time agreed in both of the genotypes. Red dots represent the genes that were regulated in opposite directions; Blue dots show the genes expressed higher in WT (≥1.7 fold difference in 'estimates', and the 'estimate' in at least one of the genotype ≥0.1); Purple dots showed the genes expressed higher in Bcl11b KO (≥1.7 fold difference in 'estimates', and the 'estimate' in at least one of the genotype ≥ 0.1). *F)* Left panel shows the aligned two experiments of Bcl11b scRNA-seq profiles after CCA scaling, as 'Bcl11b_run1' and 'Bcl11b_run2', in UMAP1-2. Note that 'Bcl11b_run1' samples were only collected in D10. Right panel shows the Louvain clustering of the integrated samples (details in Methods). *G)* Samples subset according to the hashtag demultiplexed genotype and time of harvest, and displayed in UMAP1-2, colored by the same clustering annotation from Fig.2F. (Also see Fig.S2B). *H)* Heatmap showing the KL divergence of all integrated samples, calculated based on cluster distributions (as shown in Fig.2G and Fig.S2B). *I)* Pair-wise cluster distribution scatterplots comparing WT and Bcl11b KO. Red arrows indicate the most dramatic and consistent difference between the two

genotypes in both time points, cluster 0 and 2. *J)* Heatmap showing the top 20 differentially expressed genes between cluster 0 and cluster 2, in both directions. (Wilcoxon Rank Sum test, filtered by minimally expressed by 25% cells in of one of the clusters, and adjusted p-val < 1e-50).

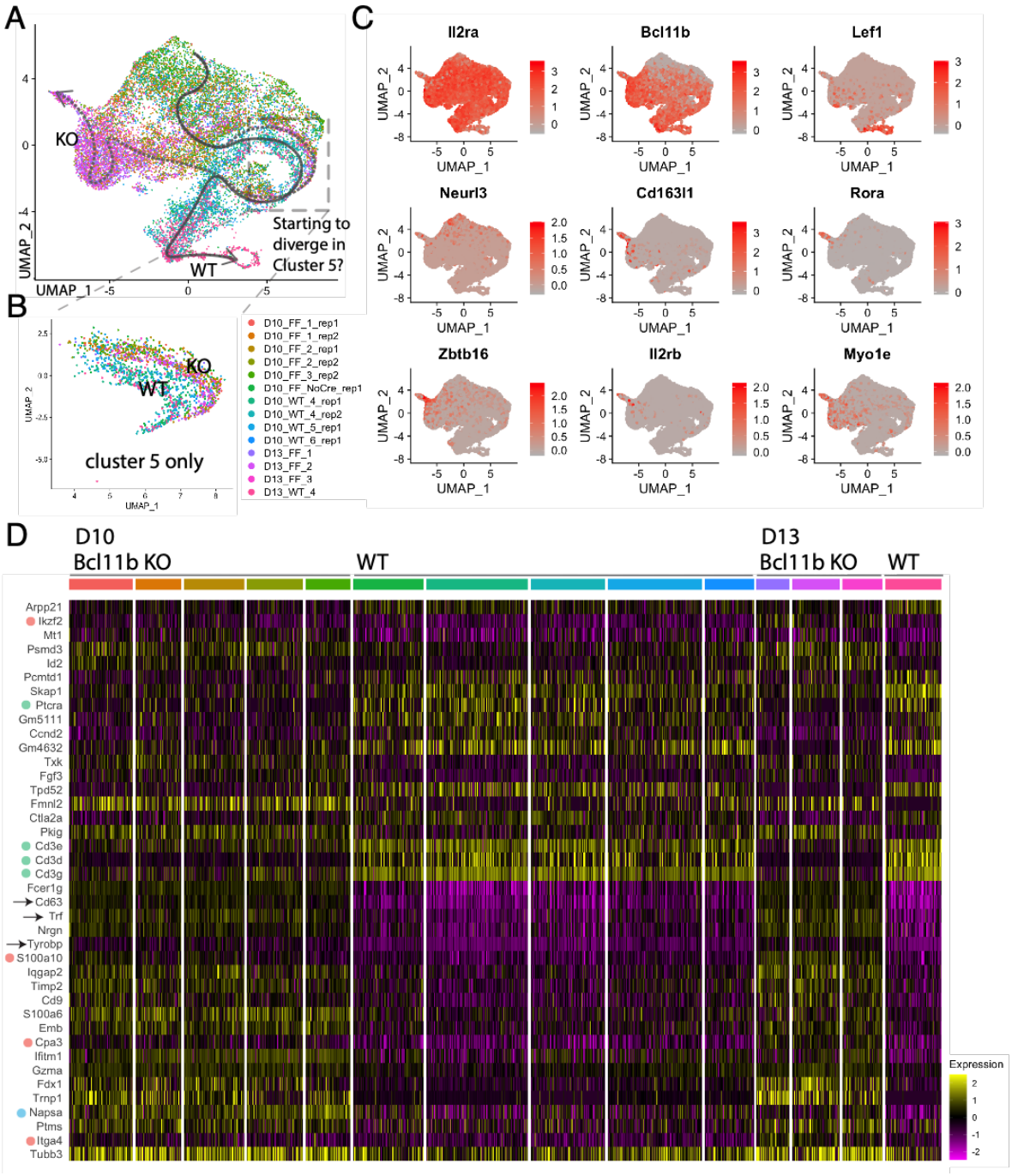


Figure3. Trajectory divergence between WT and Bcl11b knockout occurred immediately after the normal-Bcl11b expressing proliferating late DN2a stage.

A-B) UMAP 1-2 colored by different demultiplexed samples. *A)* The hand-sketched inferred trajectories of differentiation of WT and Bcl11b KO. *B)* The zoom-in view of the

subset of cells only in cluster 5 from Fig.2F, showing a slight separation of genotypes. The same color map shown in B) is used in panel A), B), and D). *C)* Selected genes highlighted on UMAP 1-2, marking T differentiation stages (*Il2ra, Bcl11b, Lef1*), and alternative lineage associated genes (*Neurl3, Cd163l1, Rora, Zbtb16, Il2rb, Myo1e*). *D)* Heatmap of differentially expressed genes between WT and Bcl11b KO in only the cells from cluster 5 (Fig.3B), revealing the gene expression differences that caused the separation shown in fig3B. (Wilcoxon Rank Sum test, filtered by minimally expressed by 25% cells in of one of the clusters, and adjusted p-val < 1E-20, top and bottom 20 genes ranked by average log expression differences ('avg_logFC' in Seurat) are displayed, calculated using Seurat v3.) Red dots label the genes that are enriched in Bcl11b KO compared to WT, but more dramatically in D13 than D10.

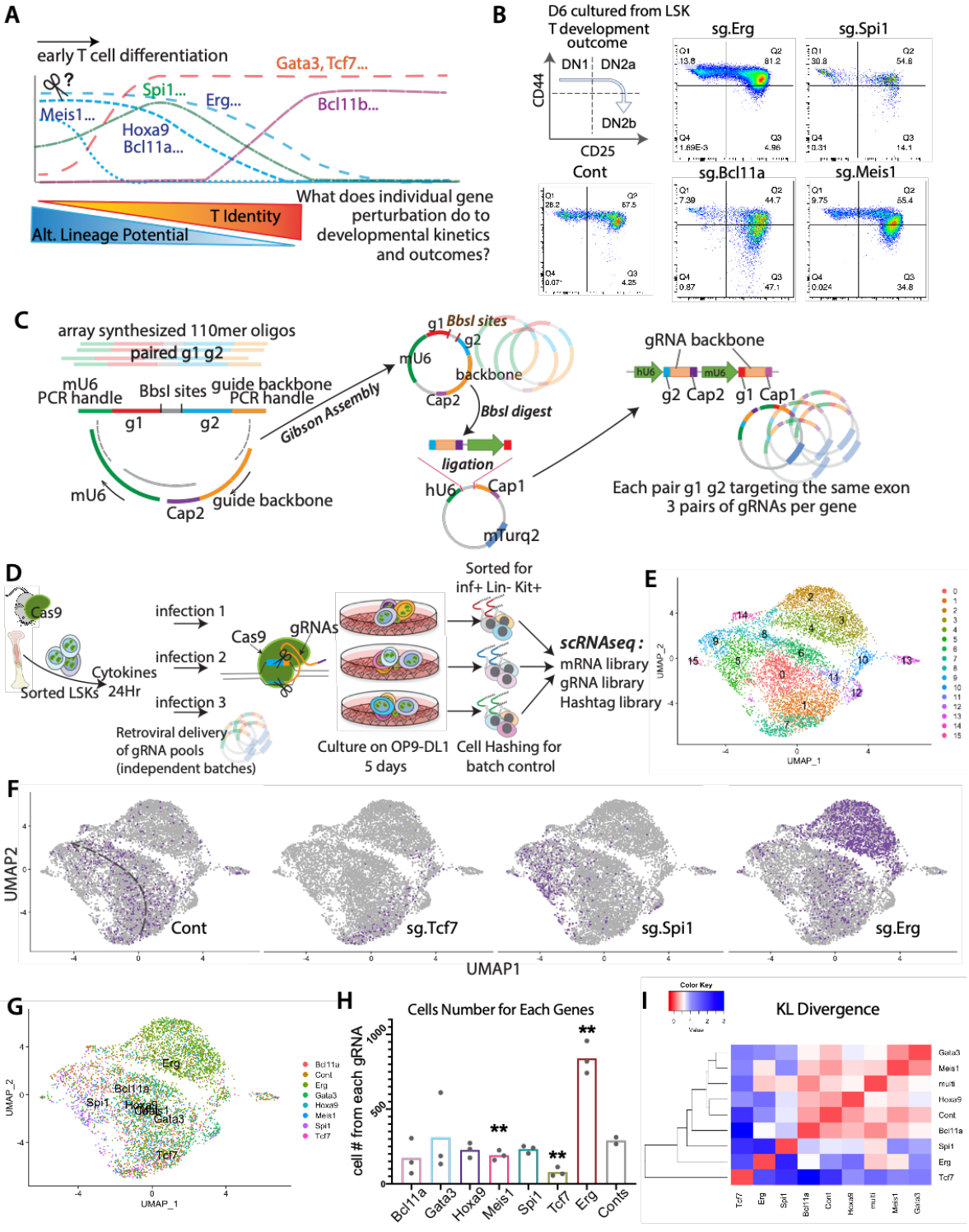


Figure 4. Pool-based batch-controlled TF perturbseq revealed deletions of TF resulted in dramatic changes on population distributions on UMAP.

A) Gene expression dynamics of selected TFs with known or potential regulatory roles in early T-cell development, and illustrations of different lineage potentials as described from the previous study (Zhou et al., 2019). *B)* Representative surface expression phenotypes analyzed by flow cytometry. Different developmental kinetics were resulted from acute TF deletions on precursors prior to Notch signaling encounter (details in Methods), and cultured for 6 days with OP9-DLL1, and developmentally staged by surface expression of CD44 and CD25. *C)* Illustration of pool-based dual gRNA cloning strategy used in the following experiments in this study. *D)* Internal- and batch-controlled cell biology experimental setup for the single-cell perturbation (perturbseq) experiment. *E-G)* UMAP 1-2 on the scRNAseq data based on PC 1-16, the analysis was performed with the expression data after being scaled to UMI counts, mitochondrial content, and cell cycle stages (using Seurat v3(Butler et al., 2018)). *E)* The cells are colored by clustering result using PC 1-16 and Louvain clustering algorithm. *F-G)* The cells are colored by sgRNA assignment. 3 pairs of dual sgRNAs against the same gene were aggregated together. *F)* The purple-colored dots highlight individual gene's KO effect, compared with the Cont. Trajectory was sketched by hand according to marker gene expressions in each cluster, as shown in Fig.S5D. More genes are plotted in Fig.S4G. *G)* Merged representation with labels showing the centroids of different KO distributions on UMAP1-2. *H)* The cell number recovered from the scRNA-seq pool separated by genes being perturbed. Each dot represents cell number recovered from one of paired gRNA vectors. Statistical significance showed t-test analyzed between Control and KOs. (**: p-val<0.01, *: p-val<0.05). *I)* Heatmap showing the KL divergence of all WT and KO samples of each genes.

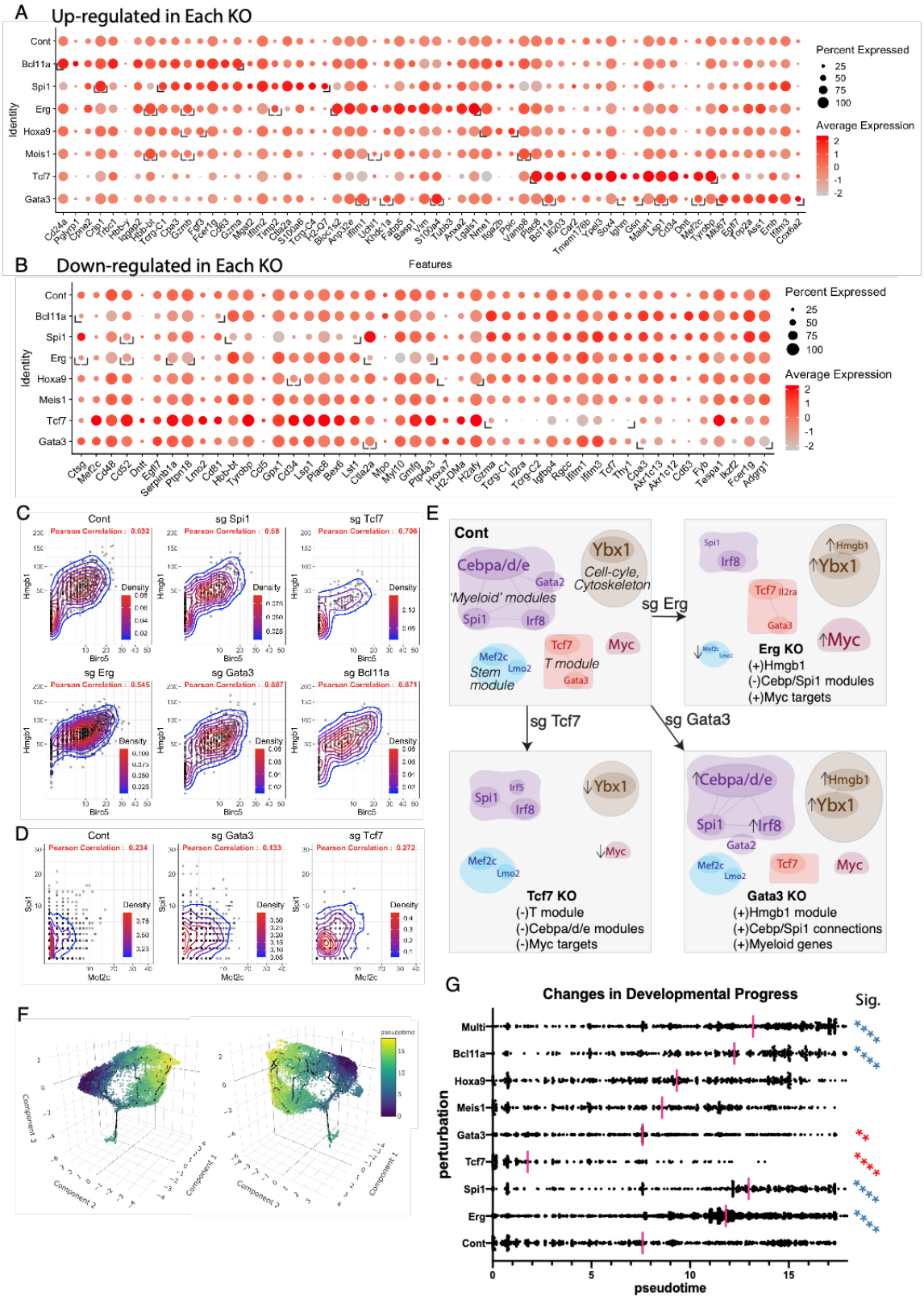
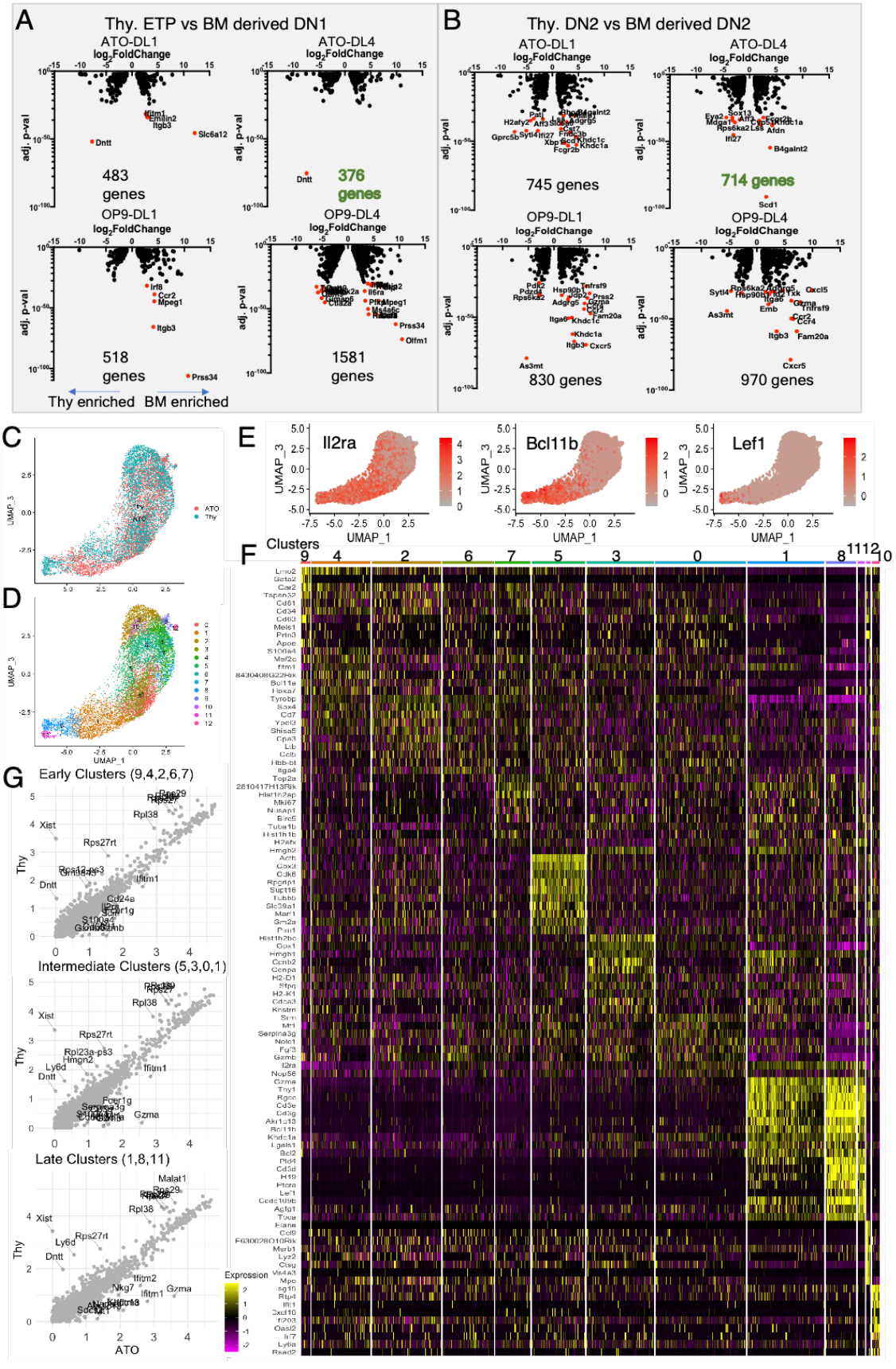

A
Up-regulated in Each KO
Identity
Cont
Bcl11a
Spi1
Erg
Hoxa9
Meis1
Tcf7
Gata3
Percent Expressed
25
50
75
100
Average Expression
2
1
0
-1
-2
B
Down-regulated in Each KO
Features
C
Cont
sg Spi1
sg Tcf7
sg Erg
sg Gata3
sg Bcl11a
Pearson Correlation : 0.632
Pearson Correlation : 0.68
Pearson Correlation : 0.706
Pearson Correlation : 0.545
Pearson Correlation : 0.607
Pearson Correlation : 0.671
Hmgb1
Birc5
Density
D
Cont
sg Gata3
sg Tcf7
Pearson Correlation : 0.234
Pearson Correlation : 0.133
Pearson Correlation : 0.272
Spi1
Mef2c
E
Cont
Cebpa/d/e
Gata2
'Myeloid' modules
Spi1
Irf8
Ybx1
Cell-cyle, Cytoskeleton
Tcf7
Myc
Mef2c
Lmo2
Stem module
T module
Gata3
sg Erg
sg Tcf7
sg Gata3
Spi1
Irf8
Tcf7
Il2ra
Gata3
Hmgb1
↑Ybx1
↑Myc
Erg KO
(+)Hmgb1
(-)Cebp/Spi1 modules
(+)Myc targets
Irf5
↓Ybx1
↓Myc
Tcf7 KO
(-)T module
(-)Cebpa/d/e modules
(-)Myc targets
↑Cebpa/d/e
↑Hmgb1
↑Ybx1
↑Irf8
Gata3 KO
(+)Hmgb1 module
(+)Cebp/Spi1 connections
(+)Myeloid genes
F
pseudotime
Component 1
Component 2
Component 3
G
Changes in Developmental Progress
Sig.
perturbation
Multi
Bcl11a
Hoxa9
Meis1
Gata3
Tcf7
Spi1
Erg
Cont
pseudotime

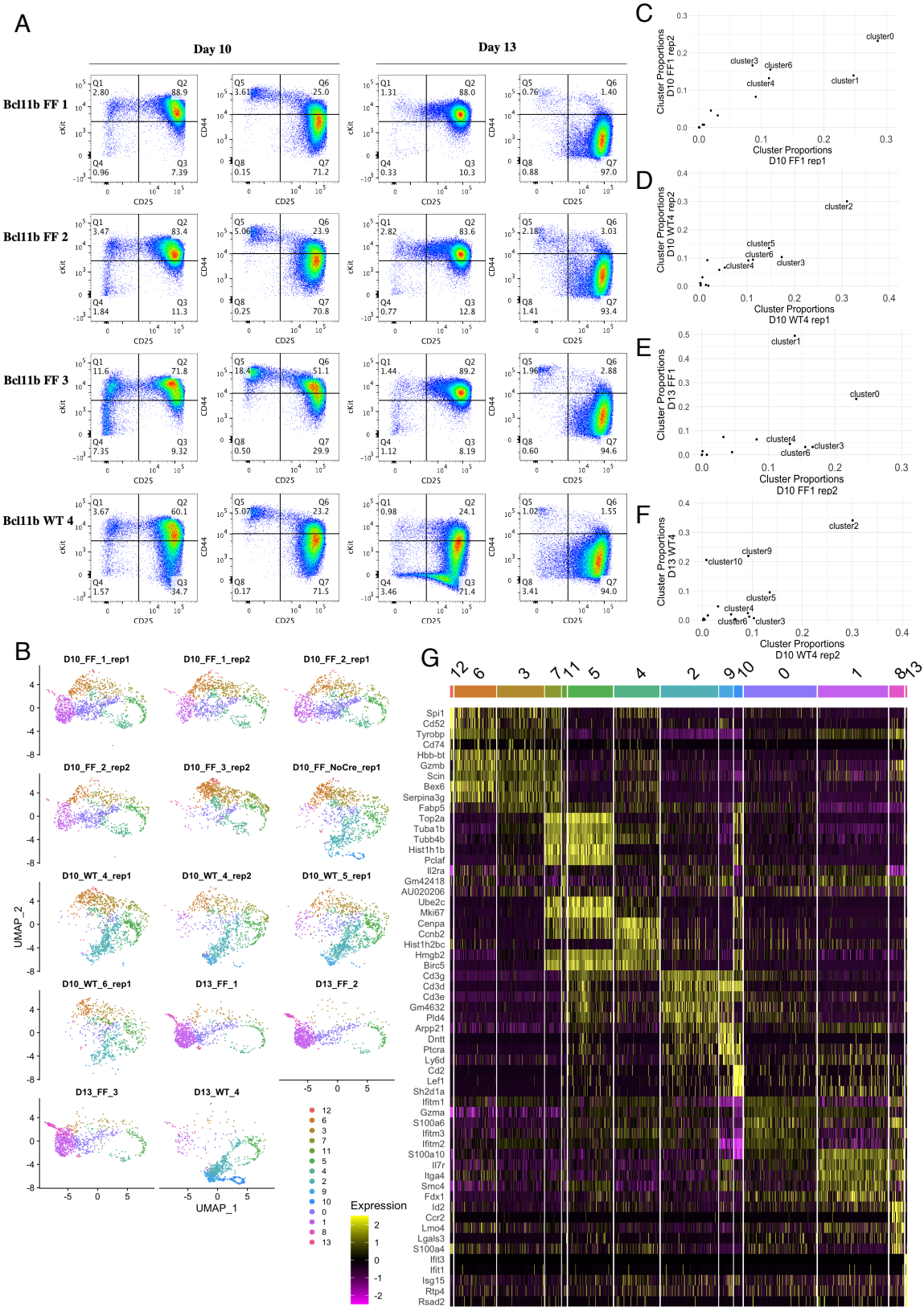
Figure 5. (Previous page) TFs involved in multi-lineage and multi-module gene regulations, affecting cell number, lineage decisions and differentiation kinetics of early T cells. *A)* Top 15 up-regulated genes in each KO (the top 15 was defined by average log expression differences (avg_logFC in Seurat3), test was performed with Wilcoxon Rank Sum test, filtered by minimally expressed by 5% cells in either the control or the KO, minimum average log expression difference cutoff of 0.1, and adjusted p-val < 1E-2). The right-angle brackets indicate the top differential expressed genes. Note some overlapping genes that were differentially expressed in more than one KO conditions were also indicated through the bracket annotations. *B)* Top 10 down-regulated genes in each KO (similarly to Fig.5A with minimum average log expression difference cutoff of 0.1, and adjusted p-val < 1E-4). *C-D)* Pairwise scatterplots of the transcript distributions and correlations, separated by different KOs. *C)* Transcript distributions of *Hmgb1* and *Birc5,* which indicate the cell cycle and proliferative stage of cells. (G2/M or proliferative cells express *Birc5*). *D)* Transcript distributions of *Spi1* and *Mef2c*, showing these stem and progenitor genes' expression level in the different KOs. *E)* Summary cartoon of findings from the 'regulon activity' of SCENIC analysis, as detailed in Fig.S6 and S7 and Methods. *F)* 3D UMAP colored by inferred pseudotime. The pseudotime was calculated with Monocle 3, based on the trajectory inference from 3D UMAP built using the size and cell cycle scaled data as described in Fig.4E, details in Methods. *G)* Pseudotime distributions of cells from different KO, showing that KOs of *Erg, Spi1,* and *Bcl11a* exhibit faster differentiation speed compared to the control; KOs of *Gata3* and *Tcf7*, on the other hand, had slowed or stalled developmental progression according to pseudotime distributions. In order for pseudotime to reflect only the T-lineage progression relevance, the alternative lineage population at the bottom of F) was excluded. For statistical significance, Kruskal-Wallis test of multiple comparisons was performed, comparing each KO to Cont. Level of statistical significance: ** marks adj.p-val<1E-2, **** marks adj.p-val <1E-4, by the Kruskal-Wallis test. The asterisk colors indicate the direction of peudotime change compared to Cont.

SUPPLEMENTARY MATERIALS

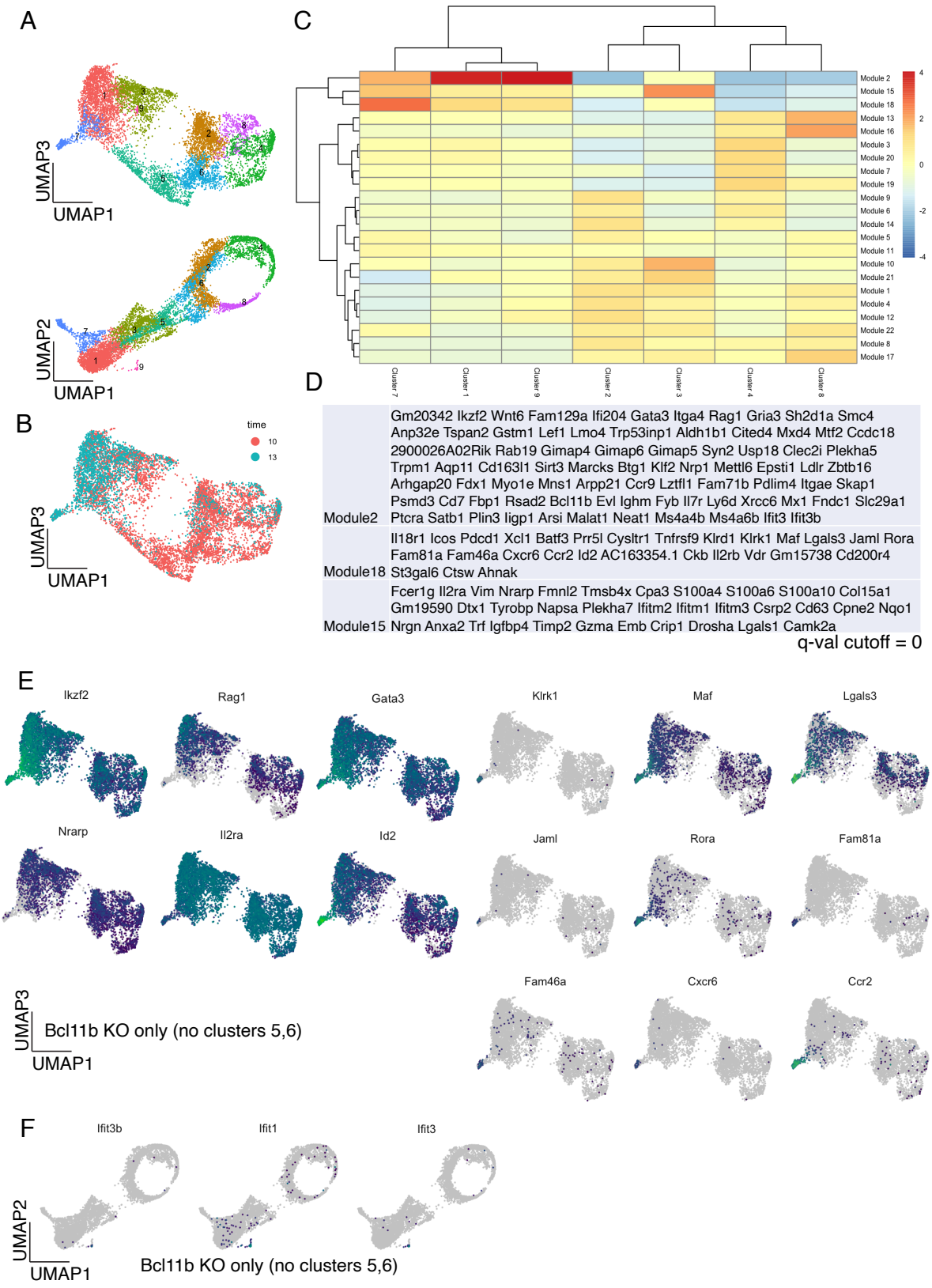
SUPPLEMENTARY FIGURES


A
Thy. ETP vs BM derived DN1
ATO-DL1
ATO-DL4
OP9-DL1
OP9-DL4
log2FoldChange
adj. p-val
483 genes
376 genes
518 genes
1581 genes
Thy enriched
BM enriched
B
Thy. DN2 vs BM derived DN2
745 genes
714 genes
830 genes
970 genes
C
D
E
Il2ra
Bcl11b
Lef1
UMAP_1
UMAP_3
F
Clusters
G
Early Clusters (9,4,2,6,7)
Intermediate Clusters (5,3,0,1)
Late Clusters (1,8,11)
Thy
ATO
Expression

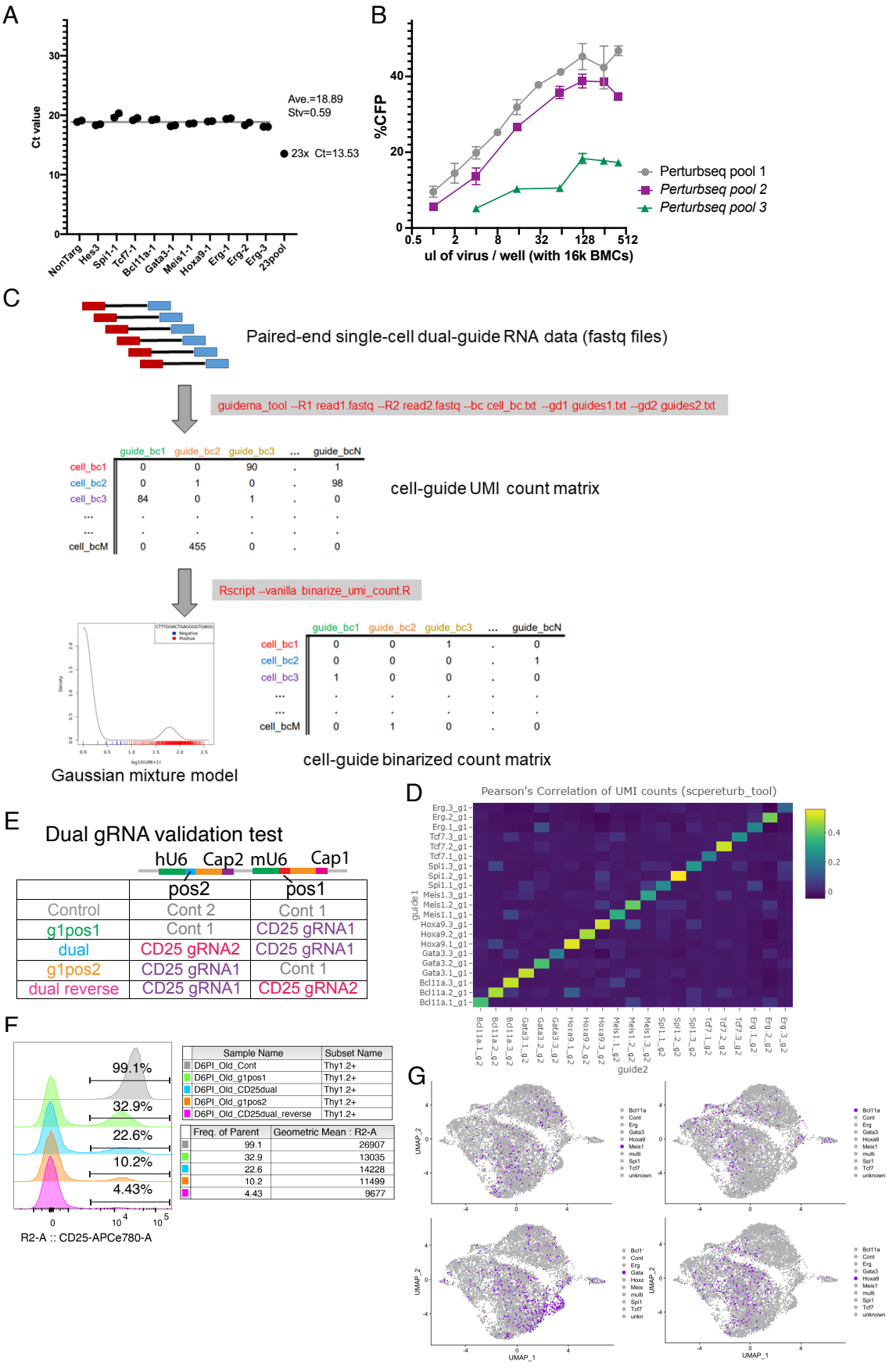
Supplementary Figure S1 (previous page), related to Fig.1. Full bulk and single-cell RNA-seq analysis of *in vivo* and *ex-vivo* derived early T cells. *A-B)* Volcano plots showing the full differential expression analysis of bulk RNA-seq profiles, comparing between *in vivo* thymic early T cells and *ex-vivo* BM derived early T-cell populations. The total number of differentially expressed genes were labeled on each plot, which was analyzed with EdgeR and filtered by $\log_2$ Fold change larger than 1 and adjusted p.val <1E-3. *A)* Differential gene analysis between thymic ETP and *ex-vivo* derived DN1, in the 4 culture conditions. *B)* Differential gene analysis between thymic DN2 and *ex-vivo* derived DN2, in the 4 culture conditions. *C-F)* Single-cell analysis of *in vivo* and *ex-vivo* derived T cells, aligned with CCA scaled data, as shown in Fig.1C. The data is shown on UMAP1-3 for the clear separation on developmental stages. *C)* Cells colored by origin of sample, i.e. 'Thy' for *in vivo* thymocytes and 'ATO' for T cells derived from ATO-DLL4 as discussed in Fig.1C. *D)* Cells colored with clustering assignment on integrated data. *E)* T developmental marker genes expression pattern. *F)* Heatmap displaying the top 10 enriched genes in each sub-cluster ordered by approximate developmental progression based on gene expression and connectivity in low dimensional displays. (Seurat 3 pipeline with minimum fraction of expressing cells ≥ 0.25, Wilcoxon rank sum test with avg_logFC threshold of 0.3). *G)* Within the early, middle and late developmental sub-clusters, the average gene expression level (size and log normalized transcript count data) comparison between thymic T-cell data (Thy) and the *ex-vivo* derived T cells (ATO). Off-diagonal outlier genes were labeled. Note that the 'Thy' data was obtained from female mice whereas the 'ATO' data was derived from LSKs of male mice, hence the Xist expression on 'Thy' data only.


A
Day 10
Day 13
Bcl11b FF 1
Bcl11b FF 2
Bcl11b FF 3
Bcl11b WT 4
cKit
CD44
CD25
B
D10_FF_1_rep1
D10_FF_1_rep2
D10_FF_2_rep1
D10_FF_2_rep2
D10_FF_3_rep2
D10_FF_NoCre_rep1
D10_WT_4_rep1
D10_WT_4_rep2
D10_WT_5_rep1
D10_WT_6_rep1
D13_FF_1
D13_FF_2
D13_FF_3
D13_WT_4
UMAP_1
UMAP_2
Expression
C
Cluster Proportions D10 FF1 rep2
Cluster Proportions D10 FF1 rep1
D
Cluster Proportions D10 WT4 rep2
Cluster Proportions D10 WT4 rep1
E
Cluster Proportions D13 FF1
Cluster Proportions D10 FF1 rep2
F
Cluster Proportions D13 WT4
Cluster Proportions D10 WT4 rep2
G
Spi1
Cd52
Tyrobp
Cd74
Hbb-bt
Gzmb
Scin
Bex6
Serpina3g
Fabp5
Top2a
Tuba1b
Tubb4b
Hist1h1b
Pclaf
Il2ra
Gm42418
AU020206
Ube2c
Mki67
Cenpa
Ccnb2
Hist1h2bc
Hmgb2
Birc5
Cd3g
Cd3d
Cd3e
Gm4632
Pld4
Arpp21
Dntt
Ptcra
Ly6d
Cd2
Lef1
Sh2d1a
Ifitm1
Gzma
S100a6
Ifitm3
Ifitm2
S100a10
Il7r
Itga4
Smc4
Fdx1
Id2
Ccr2
Lmo4
Lgals3
S100a4
Ifit3
Ifit1
Isg15
Rtp4
Rsad2

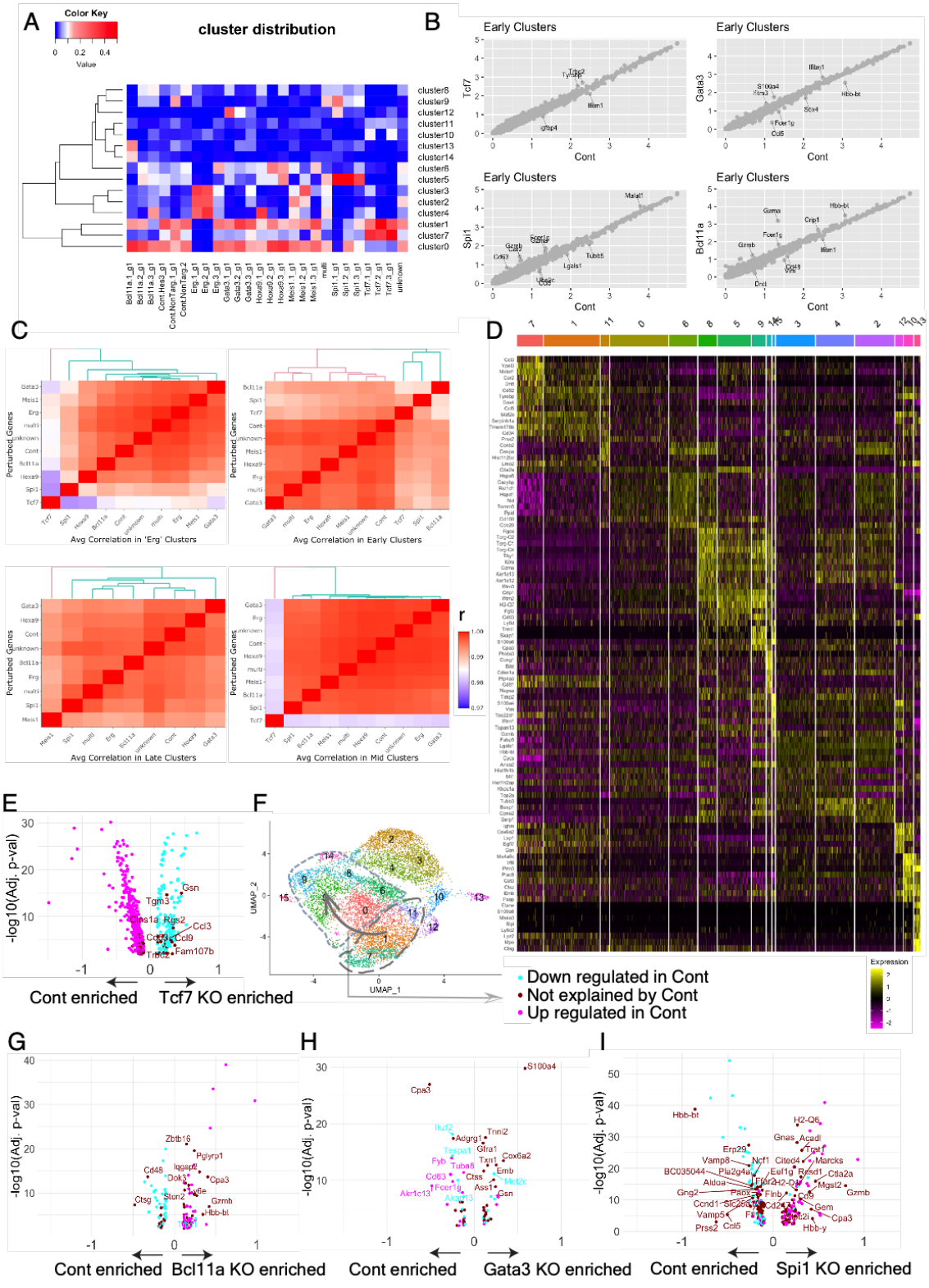
Supplementary Figure S2 (previous page), related to Fig. 2 and 3. Surface and expression profiles of WT and Bcl11b KO (labeled as 'FF' for Bcl11b homozygous flx/flx locus) single-cell samples. *A)* Flow cytometry profiles of WT and Bcl11b KO cells collected from ATO-DLL4 culture system at 2 different time points. Compared to WT control, the cells missing Bcl11b can still similarly turn off CD44, but failed to down regulate cKit expression levels. In D13, the cKit was further up regulated compared to D10 in Bcl11b KO samples. *B)* UMAP 1-2 display of the integrated scRNA-seq data, separated by individual samples. *C-D)* Cluster distributions (shown in proportions), comparing different biological replicates of cells derived from BM of the same animal origin. This shows great replicability of *ex-vivo* derivation and scRNA-seq experimental setups. *E-F)* Cluster distributions comparing samples harvested from different time points of the same animal origins. Cluster assignment same as described in Fig. 2F and Fig. S2G. *G)* Heatmap displaying the top 5 enriched genes in each sub-cluster ordered by approximate developmental trajectories of WT and Bcl11b KO, based on gene expression and connectivity in UMAP displays. (Seurat 3 pipeline with minimum fraction of expressing cells ≥ 0.25, Wilcoxon rank sum test with avg_logFC threshold of 0.3).


A
UMAP3
UMAP1
UMAP2
UMAP1
B
time
10
13
UMAP3
UMAP1
C
Module 2
Module 15
Module 18
Module 13
Module 16
Module 3
Module 20
Module 7
Module 19
Module 9
Module 6
Module 14
Module 5
Module 11
Module 10
Module 21
Module 1
Module 4
Module 12
Module 22
Module 8
Module 17
Cluster 7
Cluster 1
Cluster 9
Cluster 2
Cluster 3
Cluster 4
Cluster 8
D
Module2 Gm20342 Ikzf2 Wnt6 Fam129a Ifi204 Gata3 Itga4 Rag1 Gria3 Sh2d1a Smc4 Anp32e Tspan2 Gstm1 Lef1 Lmo4 Trp53inp1 Aldh1b1 Cited4 Mxd4 Mtf2 Ccdc18 2900026A02Rik Rab19 Gimap4 Gimap6 Gimap5 Syn2 Usp18 Clec2i Plekha5 Trpm1 Aqp11 Cd163l1 Sirt3 Marcks Btg1 Klf2 Nrp1 Mettl6 Epsti1 Ldlr Zbtb16 Arhgap20 Fdx1 Myo1e Mns1 Arpp21 Ccr9 Lztfl1 Fam71b Pdlim4 Itgae Skap1 Psmd3 Cd7 Fbp1 Rsad2 Bcl11b Evl Ighm Fyb Il7r Ly6d Xrcc6 Mx1 Fndc1 Slc29a1 Ptcra Satb1 Plin3 Iigp1 Arsi Malat1 Neat1 Ms4a4b Ms4a6b Ifit3 Ifit3b
Module18 Il18r1 Icos Pdcd1 Xcl1 Batf3 Prr5l Cysltr1 Tnfrsf9 Klrd1 Klrk1 Maf Lgals3 Jaml Rora Fam81a Fam46a Cxcr6 Ccr2 Id2 AC163354.1 Ckb Il2rb Vdr Gm15738 Cd200r4 St3gal6 Ctsw Ahnak
Module15 Fcer1g Il2ra Vim Nrarp Fmnl2 Tmsb4x Cpa3 S100a4 S100a6 S100a10 Col15a1 Gm19590 Dtx1 Tyrobp Napsa Plekha7 Ifitm2 Ifitm1 Ifitm3 Csrp2 Cd63 Cpne2 Nqo1 Nrgn Anxa2 Trf Igfbp4 Timp2 Gzma Emb Crip1 Drosha Lgals1 Camk2a
q-val cutoff = 0
E
Ikzf2
Rag1
Gata3
Klrk1
Maf
Lgals3
Nrarp
Il2ra
Id2
Jaml
Rora
Fam81a
Fam46a
Cxcr6
Ccr2
UMAP3
UMAP1
Bcl11b KO only (no clusters 5,6)
F
Ifit3b
Ifit1
Ifit3
UMAP2
UMAP1
Bcl11b KO only (no clusters 5,6)

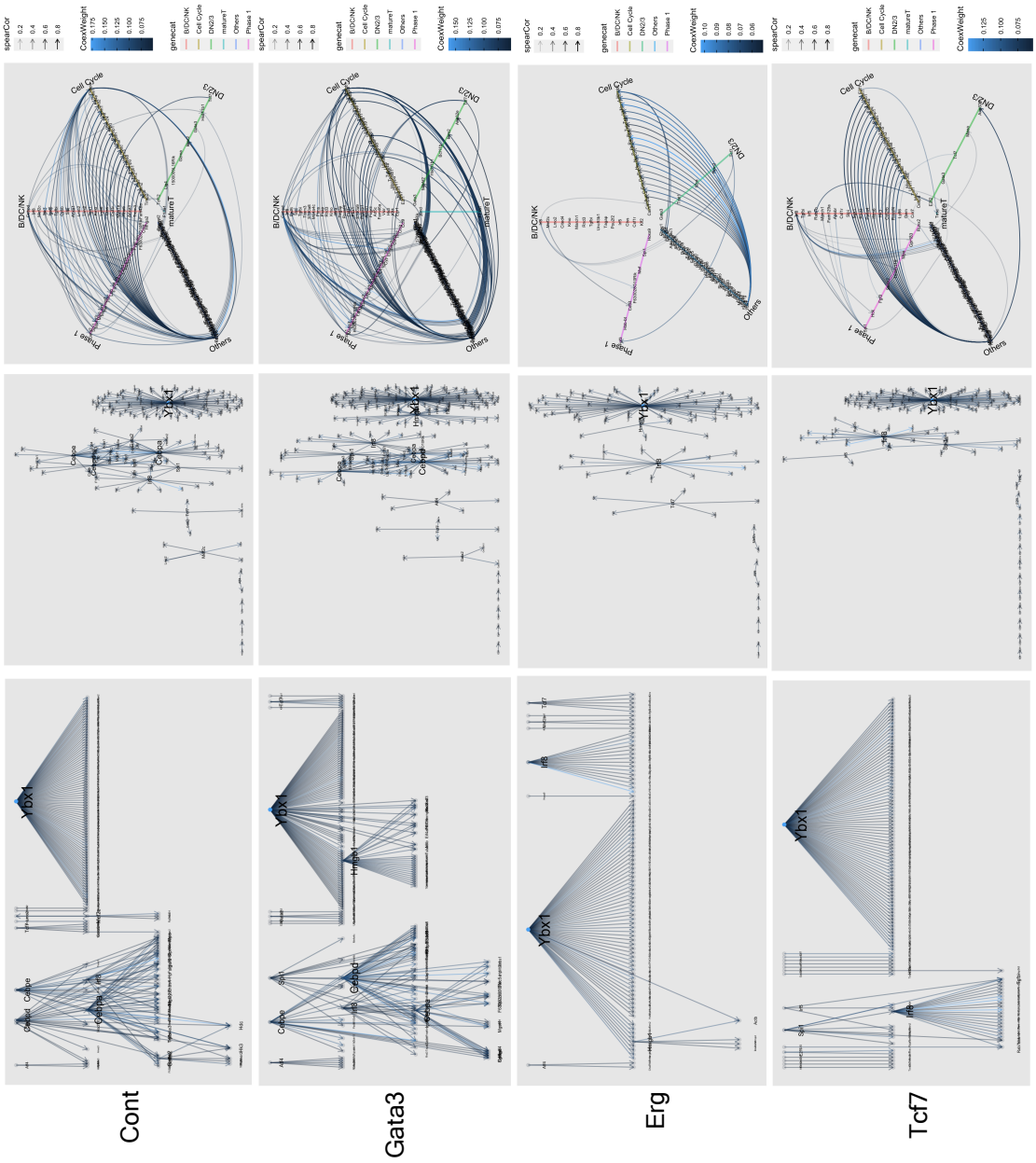
Supplementary Figure S3 (previous page), related to Fig.3. A fine look of only the Bcl11b KO trajectory revealed accumulation of abnormal gene expressions, and the potential 'exiting point' of T-lineage. *A-B)* UMAP1-2 and 1-3 of only the cells derived from Bcl11b KO animals from both experiments (computational subset by the hashtag assignments). *A)* Cells colored by clustering analysis of Bcl11b KO cells only, performed with Monocle3 (Cao et al., 2019) with PhenoGraph algorithm (Levine et al., 2015). *B)* Cells colored by time point of sample collection, showing Bcl11b KO samples' distribution changes with respect to time, on UMAP1-3. Note that the lower clusters 5 and 6, as shown in Fig.S3A and here, are exclusively expressed by D10 samples and represent mostly cells before the normal Bcl11b expression. For display clarity, the cluster 5 and 6 will be removed in the panels below. *C)* Heatmap showing the Monocle3 inferred co-regulated gene modules based on the scRNA-seq data, and the aggregated expression level in the clusters mentioned above in Fig.S3B. *D)* The list of genes in the inferred 'gene module' 2, 18 and 15, which are most enriched in the most differentiated abnormal cells. *E-F)* Selected genes' expression patterns on the 'Bcl11b KO only' UMAP1-2, and 1-3. *E)* The expression of Notch response genes, such as *Nrarp* and *Il2ra*, were down regulated at the bottom left 'tip' of this UMAP 1-3 display. Many genes potentially involved in alternative lineages gradually accumulate as the cells progress toward the bottom left 'tip', where more NK and ILC marker genes started to be up regulated. This implies that the 'non-T' genes slowly accumulate in Bcl11b KO cells in a rather homogeneous fashion, e.g. there are no major bifurcations of the Bcl11b KO trajectory, and the final 'departure' from the T-lineage involves loss of Notch signaling responses. *F)* Some interferon response genes (*Ifit3b*, *Ifit1,* and *Ifit3*) uniquely highlighted the minor cluster of cells at the bottom left on Bcl11b KO only UMAP 1-2.


A
Ct value
Ave.=18.89
Stv=0.59
23x Ct=13.53
NonTarg
Hes3
Spi1-1
Tcf7-1
Bcl11a-1
Gata3-1
Meis1-1
Hoxa9-1
Erg-1
Erg-2
Erg-3
23pool
B
%CFP
ul of virus / well (with 16k BMCs)
Perturbseq pool 1
Perturbseq pool 2
Perturbseq pool 3
C
Paired-end single-cell dual-guide RNA data (fastq files)
guiderna_tool --R1 read1.fastq --R2 read2.fastq --bc cell_bc.txt --gd1 guides1.txt --gd2 guides2.txt
cell-guide UMI count matrix
Rscript --vanilla binarize_umi_count.R
Gaussian mixture model
cell-guide binarized count matrix
D
Pearson's Correlation of UMI counts (scpereturb_tool)
guide1
guide2
E
Dual gRNA validation test
hU6
Cap2
mU6
Cap1
pos2
pos1
Control
Cont 2
Cont 1
g1pos1
Cont 1
CD25 gRNA1
dual
CD25 gRNA2
CD25 gRNA1
g1pos2
CD25 gRNA1
Cont 1
dual reverse
CD25 gRNA1
CD25 gRNA2
F
99.1%
32.9%
22.6%
10.2%
4.43%
R2-A :: CD25-APCe780-A
Sample Name
Subset Name
D6PI_Old_Cont
Thy1.2+
D6PI_Old_g1pos1
Thy1.2+
D6PI_Old_CD25dual
Thy1.2+
D6PI_Old_g1pos2
Thy1.2+
D6PI_Old_CD25dual_reverse
Thy1.2+
Freq. of Parent
Geometric Mean : R2-A
99.1
26907
32.9
13035
22.6
14228
10.2
11499
4.43
9677
G
UMAP_1
UMAP_2

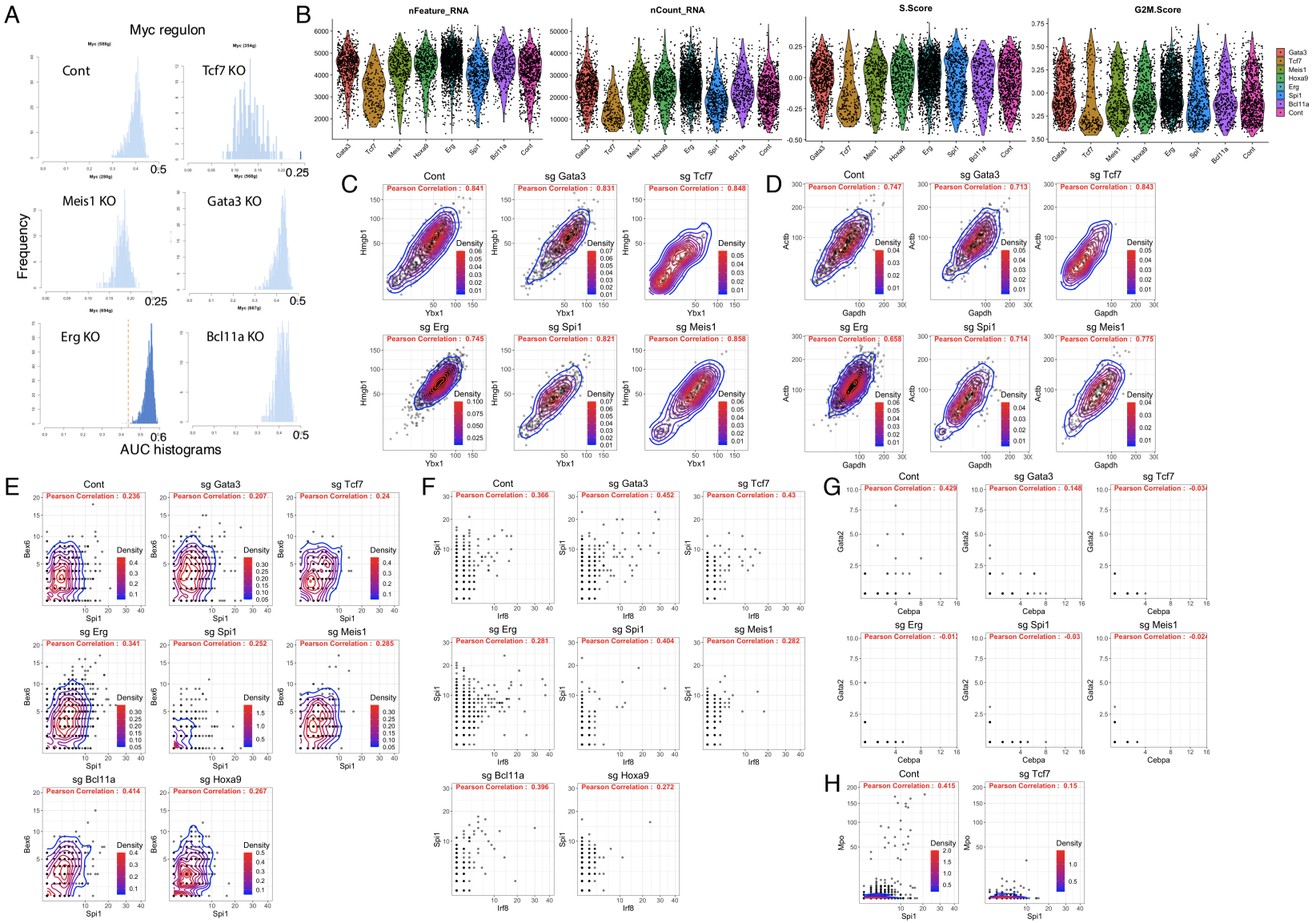
Supplementary Figure S4 (previous page), related to Fig.4. Technical and analytical details of dual gRNA perturbseq. *A)* The qPCR ct values of selected vectors, which sample the evenness of plasmid vector backbones in the cloned dual-gRNA pool. The result showed an evenly synthesized pool of plasmids that was used for viral packaging. *B)* Multiplicity of infection (MOI) titration with the viral pools containing dual gRNA vectors. All the batches of viral titers were tested separately on primary BM cells, to target precisely MOI of 0.5-1 for the scRNA-seq experiment (viral usage ~ 40-64% of the plateau). Because of the inferior infectivity of pool 3, only pool 1 and pool 2 were used in this study. *C)* In house bioinformatic processing pipeline to align the dual gRNA with both Cap1 and Cap2 information (detailed in Methods). *D)* Pearson correlation of UMI counts from guide1 and guide2 assignment. *E-F)* The validation experiment of dual gRNA effectiveness of acute gene perturbation. *E)* Due to the low recovery of guide2-cap2 counts (from pos2 on the illustration), we designed the experiment to validate whether the gRNAs from pos2 are adding effectiveness to the acute perturbation in our system. Here, CD25 (encoded by *Il2ra*) were targeted by two gRNA with switching positions. *F)* Flow cytometry analysis of the dual gRNA validation test. The result surprisingly showed that the same gRNA sequence in pos2 was more effective in perturbation compared to pos1. But most importantly, the dual gRNA perturbation efficiency towards the same gene is consistently better than a single gRNA. *G)* Additional KOs' single-cell distribution patterns highlighted on UMAP 1-2, as discussed in Fig.4F.


A
Color Key
cluster distribution
B
Early Clusters
C
Avg Correlation in 'Erg' Clusters
Avg Correlation in Early Clusters
Avg Correlation in Late Clusters
Avg Correlation in Mid Clusters
D
E
-log10(Adj. p-val)
Cont enriched
Tcf7 KO enriched
F
UMAP_1
UMAP_2
Down regulated in Cont
Not explained by Cont
Up regulated in Cont
Expression
G
Cont enriched
Bcl11a KO enriched
H
Cont enriched
Gata3 KO enriched
I
Cont enriched
Spi1 KO enriched

Supplementary Figure S5 (previous page), related to Fig. 4 and Fig. 5. Low dimensional representations and clustering assignments mostly represent the differences between genotypes (KOs). A) Heatmap of cluster distributions of individual dual gRNA perturbations. The result showed that there was a general agreement of cluster distributions of perturbations against the same genes, with a few exceptions, such as Erg.3 (3rd pair of gRNAs targeting Erg). B) Scatterplots of average gene expressions between KOs and Cont, in the shared early common clusters (7,1,11 in sub-clusters defined in Fig.4E). The result showed surprising similarity of the gene expressions between KOs and Cont in the same 'common clusters'. C) The heatmaps of Pearson correlations of average expressions in the shared 'common clusters' between different perturbations. (The 4 'common clusters' were defined as: 'early clusters': 7, 1, 11; 'mid clusters': 0, 6; 'late clusters': 9, 5, 8; 'Erg clusters': 4, 3, 2 sub-clusters defined in Fig.4E, respectively.) Note that the color scale of heatmap represents Pearson correlation of 0.97 to 1. This implies that within the same 'common clusters', the expression profile between KOs and Cont are very similar. D) Heatmap displaying the top 10 enriched genes in each sub-cluster ordered by approximate developmental trajectories. Note that Erg KO formed a parallel trajectory in the UMAP display from the main trajectory of other cells. (Seurat 3 pipeline with minimum fraction of expressing cells ≥ 0.25, Wilcoxon rank sum test with avg_logFC threshold of 0.3). E-I) Volcano plots showing the differential expressed genes between Cont and KOs, the color of dots represent if the gene was differentially expressed in Cont cells only during the normal developmental progression (transition from 'early clusters' to 'mid' and 'late' clusters, as annotated in F.) Specifically, genes that were upregulated during normal early to late transition in Cont cells are labeled as cyan, and the downregulated genes are labeled as magenta. Genes that were differentially expressed between KOs and Cont but not differentially regulated in this normal development transition are labeled as dark red. The color labels help visualize whether the differential expression between WT and KOs were merely reflecting a developmental acceleration or a stalled progression.



Supplementary Figure S6, related to Fig5. Inferred TF activities and regulatory connections by SCENIC. SCENIC was performed on subsets of individual KOs. The font sizes of the gene labels reflect the number of edges shown in the graph, showing the most prominent TF regulators. The detailed analysis is described in Methods.


A
Myc regulon
Cont
Tcf7 KO
Meis1 KO
Gata3 KO
Erg KO
Bcl11a KO
Frequency
AUC histograms
B
nFeature_RNA
nCount_RNA
S.Score
G2M.Score
Gata3
Tcf7
Meis1
Hoxa9
Erg
Spi1
Bcl11a
Cont
C
Cont
sg Gata3
sg Tcf7
sg Erg
sg Spi1
sg Meis1
Pearson Correlation : 0.841
Pearson Correlation : 0.831
Pearson Correlation : 0.848
Pearson Correlation : 0.745
Pearson Correlation : 0.821
Pearson Correlation : 0.858
Hmgb1
Ybx1
Density
D
Pearson Correlation : 0.747
Pearson Correlation : 0.713
Pearson Correlation : 0.843
Pearson Correlation : 0.658
Pearson Correlation : 0.714
Pearson Correlation : 0.775
Actb
Gapdh
E
Pearson Correlation : 0.236
Pearson Correlation : 0.207
Pearson Correlation : 0.24
Pearson Correlation : 0.341
Pearson Correlation : 0.252
Pearson Correlation : 0.285
Pearson Correlation : 0.414
Pearson Correlation : 0.267
Bex6
Spi1
F
Pearson Correlation : 0.366
Pearson Correlation : 0.452
Pearson Correlation : 0.43
Pearson Correlation : 0.281
Pearson Correlation : 0.404
Pearson Correlation : 0.282
Pearson Correlation : 0.396
Pearson Correlation : 0.272
Irf8
G
Pearson Correlation : 0.429
Pearson Correlation : 0.148
Pearson Correlation : -0.034
Pearson Correlation : -0.011
Pearson Correlation : -0.03
Pearson Correlation : -0.024
Gata2
Cebpa
H
Pearson Correlation : 0.415
Pearson Correlation : 0.15
Mpo

Supplementary Figure S7, related to Fig. 5. Detailed evidence of TF regulated activities on developing T cells. *A)* Myc regulon activity distributions from SCENIC analysis. The result showed an upregulated Myc activity by *Erg* KO cells and downregulated Myc activity by *Meis1* and *Tcf7* KOs. *B)* Distributions of number of genes detected, number of transcripts detected, and inferred cell cycle stages, according to genes perturbed. *C-D)* Scatterplots of transcript distributions of cell cycle and cytoskeleton related genes, separated by genes perturbed. *E)* Transcript distributions of *Spi1* and a known DC marker and Spi1 downstream target, *Bex6*. The clear downshift of the density in *Spi1* KO, compared to all other KOs and Cont, substantiated the effectiveness of perturbation in this experiment. *F-H)* Transcript distributions of some stemness and myeloid program markers.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animals

C57BL/6J, B6.Cg-Tg(BCL2)25Wehi/J(Bcl2-tg), *Vav1*-iCre mice (B6N.Cg-*Commd10*$^{Tg(Vav1–icre)A2Kio}$/J) and B6.Gt(ROSA)26Sortm1.1(CAG-cas9*,- EGFP)Fezh/J (Cas9) mice were purchased from the Jackson Laboratory. B6.*Bcl11b*$^{yfp/yfp}$ reporter mice(Kueh et al., 2016) were used for bulk RNAseq analysis, and B6.*Bcl11b*$^{fl/fl}$ mice(Hosokawa et al., 2018a; Longabaugh et al., 2017) were both reported previously. All mice were maintained on the B6 background. For CRISPR/Cas9 experiments, BCL2 transgenic mice and Cas9 mice were crossed to generate B6-Cas9/+; +/Bcl2 heterozygotes for each experiment. For Bcl11b experiments, B6.Bcl11b$^{fl/+}$ *Vav1*-iCre heterozygous mice were bred to obtain *Bcl11b*$^{+/+}$ and *Bcl11b*$^{fl/fl}$ *ROSA26R-YFP* mice with *Vav1*-iCre, as previously described(Hosokawa et al., 2018a), annotated as WT and Bcl11b KO. Animals used for these experiments were bred and maintained at the Animal Facilities at California Institute of Technology under conventional Specific Pathogen-Free conditions, and animal protocols were reviewed and approved by the Institute Animal Care and Use Committee of California Institute of Technology (Protocol #1445-18G).

Cell lines

To provide a microenvironment that supports T-lineage differentiation *in vitro*, we co-cultivated purified BM cells with the OP9-DL1, OP9-DL4 stromal cell line (Schmitt and Zúñiga-Pflücker, 2002), which were obtained from Dr. Zúñiga-Pflücker (Sunnybrook Research Institute, University of Toronto), or MS5-mDLL1 or MS5-mDLL4 (Montel-Hagen et al., 2020), which were obtained from Dr. Gay Crook (UCLA) and maintained in our laboratory as described in the original reference. Details of the differentiation cultures are given below under Method Details.

METHOD DETAILS

Primary Cell Purification

For *in vitro (ex-vivo)* differentiation of pro-T cells, bone marrow hematopoietic progenitors were used for input. Bone marrow (BM) was removed from the femurs and tibiae of 10-12 week-old mice. Suspensions of BM cells were prepared and stained for lineage markers using biotin-conjugated lineage antibodies: CD3ε (eBioscience, clone 145-2C11), CD19 (eBioscience, clone 1D3), B220 (eBioscience, clone RA3-6B2), NK1.1 (eBioscience, clone PK136), CD11b (eBioscience, clone M1/70). CD11c (eBioscience, clone N418), Gr1 (eBioscience, clone RB6-8C5), and Ter119 (eBioscience, clone TER-119), then incubated with streptavidin-coated magnetic beads (Miltenyi Biotec), and passed through a magnetic column (Miltenyi Biotec), denoted as 'Lin^{-} BM'. For all scRNA-seq experiments, the Lin^{-} BM cells were immediately further FACS sorted for live (7AAD^{negative}), CD45^{positive}LSK (Lin^{negative}Scal^{high}cKit^{high}), detailed as below. All the BM precursors (Lin- or LSK) were frozen down in liquid nitrogen for storage in freeze down medium containing 10% DMSO, 40% FCS, and 50% OP9 medium, before further differentiation assays.

Flow Cytometry and Cell Sorting

Unless otherwise noted, flow cytometry analysis and FACS of all samples were carried out using the procedures outlined. Briefly, cultured cells on tissue culture plates and primary cells from thymus were prepared as single-cell suspensions, incubated in 2.4G2 Fc blocking solution, stained with respective surface cell markers as indicated, resuspended in HBH, and filtered through a 40 μm nylon mesh. They were then analyzed using a benchtop MacsQuant flow cytometer (Miltenyi Biotec, Auburn, CA) or sorted with a Sony Synergy 3200 cell sorter (Sony Biotechnology, Inc, San Jose, CA) or with a FACSAria Fusion cell sorter (BD Biosciences). All antibodies used in these experiments are standard, commercially available monoclonal reagents widely established to characterize immune cell populations in the mouse. Acquired flow cytometry data were all analyzed with FlowJo software (Tree Star).

BM Cell Differentiation

Upon usage, the hematopoietic progenitors were thawed and either cultured on OP9-DLL1 or OP9-DLL4 monolayers using OP9 medium (α-MEM, 20% FBS, 50 μM β-mercaptoethanol, Pen-Step-Glutamine) supplemented with 10 ng/ml of IL-7 (Pepro Tech Inc) and 10 ng/ml of Flt3L (Pepro Tech Inc); or aggregated to artificial thymic organoids with ATO-mDLL1 or ATO-mDLL4, seated at the air-medium interface on a culture insert (Millipore Sigma) in serum-free ATO medium (DMEM-F12, 2X B27, 30 μM Ascorbic acid, Pen-Step-Glutamine) supplemented with 5 ng/ml of IL-7 (Pepro Tech Inc) and 5 ng/ml of Flt3L (Pepro Tech Inc). If required viral delivery of gRNA, thawed BM precursors were incubated for 20-24 hours in OP9 medium supplemented with 10 ng/ml of SCF (Pepro Tech Inc), 10 ng/ml of IL-7 (Pepro Tech Inc) and 10 ng/ml of Flt3L (Pepro Tech Inc), without stroma cells, detailed as below in the 'CRISPR/Cas9-mediated Acute Deletion' section.

Cloning

The retroviral vector backbone used for sgRNA expression cloning was based on previously published E42-dTet(Hosokawa et al., 2018b) with the following modifications: 1) Capture sequence 1 (Cap1) was added to the sgRNA scaffold before the termination signal. 2) One nucleotide 'G' was deleted before the sgRNA protospacer insertion site (two *AarI* restriction enzyme cutting sites) to allow compatibility with dual sgRNA vector cloning. The cloning was achieved through high-fidelity PCR (primers as shown below) and Gibson assembly, the final cloned product, containing the human U6 (hU6) promoter, two *AarI* cutting sites, gRNA backbone with Cap1 sequence and mTurquoise2 fluorescent marker, was as shown in the bottom middle in Fig. 4c.

For dual sgRNA cloning, a 'donor' sequence containing gRNA backbone and mouse U6 (mU6) promoter were obtained from a plasmid modified from Vidigal and Ventura, 2015. Specifically, the capture sequence 2 (Cap2) was added prior to the termination signal of the sgRNA scaffold backbone; we also found that the linker sequence between gRNA backbone and mU6 promoter contained a partial sgRNA backbone sequence that hinders the PCR capability and Gibson assembly accuracy, therefore we cloned to remove the

partial repeat sequence in the linker region. The cloning was performed through sequential high-fidelity PCR (primers as listed below) and blunt end ligation.

The pool-based dual gRNA cloning was performed similarly to the protocol described by Vidigal and Ventura, 2015 with the modified vector and plasmids above (workflow shown in Figure 4C), with minor protocol modifications: 1) the 'Donor sequence' containing gRNA scaffold - Cap2 - modified linker – mU6 were obtained through PCR with the modified plasmid, rather than enzymatic digestion. 2) All gel purification steps were avoided, and purifications were achieved with Ampure XP or SPRIselect beads (Beckman Coulter) instead. 3) Selected gRNAs from the oligopool were qPCR quantified before and after the pool-based vector cloning process (Figure S4A) for quality control, ensuring the amplification evenness of the final plasmid pool. 4) A retroviral vector was used instead of lentiviral vectors.

Primers used in cloning

Primers for E42 modification and addition of Cap1	
backbone_fwd	gctttaaggccggtcctagcaatttttttctcgagtggctc
backbone_rev	tgtgttcacctgcgagcggtgtttcgtcctttccacaag
insert and loop_fwd	accgctcgcaggtgaacacaaca
insert and loop_rev	ttgctaggaccggccttaaagcgcaccgactcggtgccac
Primers for donor mU6 modifications and addition of Cap2	
pD_mU6_rev_cap2_blunt	gctaataggtgagcGCACCGACTCGGTGCCAC
pD_mU6_fwd_cap2_blunt	ggctaaggTTTTTTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTC
rev_del_partial_primer	aaaaaaCCTTAGCCGCTAATAGGTGAG
fwd_del_partial_primer	tttagcgcgtgcgccaattc
Primers for pool-based dual gRNA vector assembly	
Fwd lib amp primer	GTTTTGAGACTATAAATATGCATGCGAGAAAAGCCTTGTT
Rev lib amp primer	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC
FWD pDonor opening	gttttagagctagaaatagcaagtt
REV pDonor opening	caaacaaggcttttctcgca

CRISPR/Cas9-mediated Acute Deletion in Precursor Cell Cultures

To generate input cells, Cas9 mice were first bred to Bcl2-tg mice to generate heterozygotes for both transgenes. LSK or Lin^-BM cells from these Cas9;Bcl2-tg animals were purified and stored as described above. 20-24 hours after thawing and recovery in cytokines, the cells were transduced with retroviral vectors encoding reporters (CFP) and the indicated guide RNAs (sgRNAs) as detailed below, and then seeded to a OP9-DLL1 stromal culture. The methods used to generate the virus supernatant and infecting BM cells were described previously(Hosokawa et al., 2018b). For infecting LSK precursors for scRNA-seq ('perturbseq'), different batches of viruses were tested on primary BM precursors prior to the 'perturbseq' experiments to determine the accurate titers (Figure S4B), and delivered to target a precise multiplicity of infection (MOI) of 0.5-1. For phenotypical assays, cells were analyzed after 2-6 days after culture. For scRNA-seq, retrovirus infected $Lin^-CD45^+c\text{-}Kit^{hi}CFP^+$ cells were sorted on a FACSAria Fusion cell sorter (BD Biosciences).

Bulk RNAseq Analysis

Lin^- BM cells were harvested from B6.*Bcl11b*$^{yfp/yfp}$ animals, and cultured in differentiation conditions as described above. Upon harvesting, cells were subdivided into $CD25^{low}$ for DN1, Bcl11b-$YFP^{neg}CD25^{hi}$ DN2a, and Bcl11b-$YFP^{pos}CD25^{hi}$ DN2a. fractions, followed by RNA purification following the instructions of the RNeasy Micro Kit (Qiagen 74004). cDNA from each sample was prepared using NEBNext Ultra RNA Library Prep Kit for Illumina (E7530, NEB). All bulk libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50 nt. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4 and produced approximately 30 million reads per sample.

RNA-seq reads were mapped onto the mouse genome build GRCm38/mm10 using STAR (v2.4.0) and were post-processed with RSEM (v1.2.25; http://deweylab.github.io/RSEM/) according to the settings in the ENCODE long-rna-seq-pipeline (https://github.com/ENCODE-DCC/long-rna-seq-

pipeline/blob/master/DAC/STAR_RSEM.sh), with the minor modifications that the setting '–output-genome-bam–sampling-for-bam' was added to rsem-calculate-expression. STAR and RSEM reference libraries were created from genome build GRCm38/mm10 together with the Ensembl gene model file Mus_musculus.GRCm38.gtf. The resulting bam files were used to create HOMER tag directories (makeTagDirectory with –keepAll setting). For analysis of statistical significance among DEGs, the raw gene counts were derived from each tag directory with 'analyzeRepeats.pl' with the '–noadj -condenseGenes' options, followed by the 'getDiffExpression.pl' command using EdgeR (v3.6.8; http://bioconductor.org/packages/release/bioc/html/edgeR.html). For data visualization, RPKM normalized reads were derived using the 'analyzeRepeats.pl' command with the options '–count exons –condenseGenes –rpkm'; genes with an average of RPKM ≥1 across samples were kept, and their RPKM values were processed by log transformation. The normalized datasets were then hierarchically clustered with R hclust function based on Euclidean distance and 'complete' linkage. The heatmap is visualized with R pheatmap with log2 transformed RPKM data (after adding 0.1 to all values).

Single Cell RNA-seq (10X Chromium V2)

Note that only the scRNA-seq data from Figures 1 and S1 was obtained through 10X Chromium V2, the rest of the scRNA-seq data were obtained through the V3 chemistry. The early T cells derived in ATO-DLL4 from LSK were sorted as shown in Figure 1C (bottom). The sample was then washed and resuspended to 1 million cells/mL concentration in HBSS supplemented with 10% FBS and 10 mM HEPES, 17,400 cells were loaded into a 10X Chromium v2 lane, and the subsequent preparation was conducted following the instruction manual of 10X Chromium v2. The cDNA library and final library after index preparation were checked with bioanalyzer (High Sensitivity DNA reagents, Agilent Technology #5067-4626; Agilent 2100 Bioanalyzer) for quality control. Following the library preparation, the sequencing was performed with paired-end sequencing of 150nt each end on one lane of HiSeq4000 per sample, by Fulgent Genetics, Inc. (Temple City, CA). The reads were mapped onto the mouse genome Ensembl gene

model file Mus_musculus.GRCm38.gtf using a standard CellRanger pipeline. Cells were sequenced to a targeted depth of 50,000 reads per cell.

Single Cell RNA-seq (10X Chromium V3) on Bcl11b Samples with Cell Hashing

LSK from Bcl11b WT and KO animals were obtained, aliquoted into 6-7k cell/tube, and stored in liquid nitrogen as described above (individual animals were not pooled). To setup the culture, cells were thawed and aggregated with MS5-mDLL4 (800-1000 LSK and 150k MS5-DLL4 cells per ATO), and seeded on culture inserts as described above. The ATO medium was changed every 3-4 days. After culturing for 10-13 days (note experiment 1 had only D10, and experiment 2 had both D10 and D13), the ATO was mechanically disrupted and *ex-vivo* derived T cells were prepared for FACS sorting as described above. Specifically, cells derived from each animal and each time point were stained with a biotin-conjugated lineage cocktail (TCRγδ (eBioscience, clone GL-3), CD19, NK1.1, CD11b, CD11c, and Gr1). Secondary surface staining was performed with fluorescently conjugated streptavidin, CD45, cKit (eBioscience, clone 2B8), CD44 (eBioscience, clone IM7), CD25 (eBioscience or Biolegend, clone PC61.5), and TotalSeq A (Biolegend) anti-Mouse Hashtag 1-8 (1:50, in separate samples). A viability dye 7AAD (eBioscience) was applied to exclude dead cells. The sorted cells ($CD45^{positive}Lin^{low}7AAD^{negative}CD25^{positive}$), washed 2 times with HBSS supplemented with 10% FBS and 10 mM HEPES, were pooled to target an equal cell number from each Hash-tagged sample, and loaded onto one lane of a 10X Chromium V3 chip. The cDNA preparation was performed following the instruction manual of 10X Chromium v3, and the hashtag library was prepared following the Biolegend TotalseqA guide. The cDNA, tag library, and final library after index preparation were checked with the bioanalyzer (High Sensitivity DNA reagents, Agilent Technology #5067-4626; Agilent 2100 Bioanalyzer) for quality control. The cDNA final libraries was sequenced on HiSeq4000 or NovaSeq 6000, and the tag library was sequenced on HiSeq4000, by Fulgent Genetics, Inc. Cells were sequenced to an average depth of 50,000-70,000 reads per cell for cDNA and ~2,500 reads per cell for hashtags.

Direct-capture Perturbation scRNA-seq (with 10X Chromium V3)

LSK were purified, recovered in cytokines and infected with MOI 0.5-1, and cultured with OP9-DL1 as described above. Note that multiple packages of the viral pools were infected in parallel, in separate wells, to serve as biological replicates. The medium was changed on day 3. On day 5, the cells were harvested through scrapping, and filtered and prepared for FACS sorting as described above. Specifically, cells derived from each animal and each time point were stained with a biotin-conjugated lineage cocktail (TCRβ (ebioscience, clone H57-597), TCRγδ (eBioscience, clone GL-3), CD19, NK1.1, CD11b, CD11c, and Gr1). Secondary surface staining was performed with fluorescently conjugated streptavidin, CD45, cKit (eBioscience, clone 2B8), CD44 (eBioscience, clone IM7), CD25 (eBioscience or Biolegend, clone PC61.5), and TotalSeq A (Biolegend) anti-Mouse Hashtag 1-6 (1:50, in separate infected samples). A viability dye 7AAD (eBioscience) was again applied to exclude dead cells. The sorted cells ($CD45^{positive}Lin^{low}7AAD^{negative}CFP^{high}cKit^{positive}$) were washed 2 times with HBSS supplemented with 10% FBS and 10 mM HEPES, pooled to target equal cell number from each Hash-tagged sample, and loaded onto one lane of a 10X Chromium V3 chip. The cDNA preparation was performed following the instruction manual of 10X Chromium v3 for perturbation with minor modifications, and the hashtag library was prepared following the Biolegend TotalseqA guide. The cDNA, gRNA library, Hashtag library, and final libraries after index preparation were checked with bioanalyzer (High Sensitivity DNA reagents, Agilent Technology #5067-4626; Agilent 2100 Bioanalyzer) for quality control. All libraries were sequenced on HiSeq4000, by Fulgent Genetics, Inc. Cells were sequenced to at least medium depth of 50,000 reads per cell for cDNA, 20M reads/sample for hashtags and 20M reads/sample for gRNAs.

Data Analysis

Mapping of scRNA-seq Sequences, Hashtag, and gRNA Identification

Single-cell RNA-seq data were processed using 10X Cellranger 3.0.0 software. Standard cellranger-mm10-3.0.0 reference annotations were loaded to the pipeline for read mapping and gene quantification.

To process single-cell hashtag and guide RNA sequencing data, two ultrafast in-house tools (**hashtag_tool** and **guiderna_tool**) (https://github.com/gaofan83/single_cell_perturb_seq/) were developed to process raw fastq data and generate count tables (Fig.S4C). The results are typically delivered within one minute. Downstream R codes can be used to binarize the count tables for identity assignment using Gaussian Mixed Modeling.

As note, the **guiderna_tool** was specifically developed for our dual-guide system with two guide-RNA sequences (targeting different sequences) in engineered in the viral vector backbones. Based on 10X bead chemistry, **Capture1** (5'-GCTTTAAGGCCGGTCCTAGCAA-3') and **Capture2** (5'-GCTCACCTATTAGCGGCTAAGG-3') sequences recognize expressed **Guide1** and **Guide2** RNA molecules that have reverse complement capture sequences inserted. Specifically, **Capture1** and **Capture2** sequences should pair with **Guide1** and **Guide2**, respectively. From in-house single-cell guideRNA data, UMI counts can be calculated for the **Guide1** list of barcodes and the **Guide2** list of barcodes. As note, the **guiderna_tool** uses both capture sequences in R1 reads and template switching oligo sequence (TSO) in R2 read for read filtering and sorting; then potential **protospacer** sequences in R2 reads (after 5' TSO sequence) are mapped against the corresponding guide library (**Guide1** or **Guide2**) for quantification. In contrast, **Cellranger** finds a constant region after **protospacer** region in R2 first, then **protospacer** abundances in R2 are calculated. Since **guiderna_tool** utilizes both R1 and R2 read information for filtering, it is expected to be more accurate.

Gene and Cell Filtering, Data Alignment, and Clustering Analysis

10X Chromium V2 scRNA-seq (Figure1 and S1) analysis was based on data filtered on cells with at least 1200 genes expressed (transcript count over 1); outliers with more than

4300 genes or 23k unique transcripts were also removed (potential doublet) from the ATO scRNA-seq dataset, and outliers with more than 4400 genes or 27k unique transcripts were also removed from thymocyte dataset (10X V2 run1 from chapter 2), and only genes that were found expressed in at least 3 cells were kept in the analysis. The cells were further filtered to keep only cells with mitochondrial contents of less than 7.5-9%. The QC filter resulted in 6167 cells in the ATO scRNA-seq sample and 4783 cells in the thymocyte sample, which were being presented in Figure 1 and S1. The top 3000 variable features were identified from each of the two datasets and integrated with the CCA algorithm using the 3000 anchor features and 20 dimensions in Seurat v3 (Stuart et al., 2019). The principal component analysis was performed on the integrated dataset, and the UMAP display was analyzed based on PCs 1-20. For clustering, Louvain clustering was performed on the first 20 PCs with the resolution set to be 0.7, and the top 10 enriched genes in each cluster were calculated with Wilcoxon Rank Sum test, shown in the heatmap (Figure S1F).

10X Chromium V3 scRNA-seq (all scRNA-seq dataset except in Figures 1 and S1) analysis was based on data filtered on cells with at least 1300 genes expressed (transcript count over 1). The doublet elimination was guided through Cell Hashing. Specifically, number of features vs. number of unique transcripts detected were plotted, and cells with more than 1 Cell Hashing tags were considered doublet and highlighted on the scatterplot. Both the ‘cell hashing identified doublets’ and outliers with only one hashtag identified but fell in the region of high feature and transcripts content similarly to ‘cell hashing doublets’, were dropped. The subsequent integration and clustering analysis were performed similarly described above.

Unless specified, the trajectory and pseudo-time analysis with Monocle 3 were all performed on the cells that passed the filtering steps described above.

SCENIC Analysis and Visualization Graphics

We performed SCENIC (Aibar et al., 2017) analysis by starting from the raw counts of the computationally subset 'genotype' of WT and individual TF KOs (described above), and following the proposed workflow using the default parameters in SCENIC R setup. The co-expression network was generated using GENIE3(Huynh-Thu et al., 2010), and potential direct-binding targets (regulons) were based on DNA-motif analysis. AUC, which identifies and scores gene regulatory networks or regulons in single cells, was calculated using AUCell as previously desribed(Aibar et al., 2017). The motif bindings were inferred based on publicly available motif binding databases provided by the Aerts lab. The regulon output where the co-expression weight attributed to each predicted TF-target interaction, was used to filter the graphic display, retaining interactions with a co-expression weight above 0.05 and with 'high confidence annotations'. The retained interaction edgelist was used to generate graphs using the igraph R library, which was in turn visualized as plots using the ggraph library ('sugiyama' or 'stress' layout, Figure S6, left and middle panels). To further examine TF-TF interactions, predicted interactions between TFs with a co-expression weight above 0.01 and with 'high confidence annotations' were visualized with hive plots with ggraph in R. The axis on hive plots represent the categories curated with genes enriched in different cell types or states, according to RNA-seq datasets on the Immgen website (The Immunological Genome Project Consortium et al., 2008).

Software Details

The analyses were performed mainly in R (version 4.0.2) with the following packages: ggplot2(v3.3.2), dplyr(v1.0.2), cowplot(1.1.0), Seurat(v3.2.2), AUCell(v1.10.0), RcisTarget(v1.8.0), GENIE3(v1.10.0) SCENIC(v1.2.2), monocle3(v0.2.3.0), ggraph(v2.0.4), igraph(v1.2.6).

BIBLIOGRAPHY

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: Single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S.M., Reth, M., Höfer, T., and Rodewald, H.-R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature 518, 542–546.

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.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502.

Champhekar, A., Damle, S.S., Freedman, G., Carotta, S., Nutt, S.L., and Rothenberg, E.V. (2015). Regulation of early T-lineage gene expression and developmental progression by the progenitor cell transcription factor PU.1. Genes Dev. 29, 832–848.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 167, 1853-1866.e17.

García-Ojeda, M.E., Klein Wolterink, R.G.J., Lemaître, F., Richard-Le Goff, O., Hasan, M., Hendriks, R.W., Cumano, A., and Di Santo, J.P. (2013). GATA-3 promotes T-cell specification by repressing B-cell potential in pro-T cells in mice. Blood 121, 1749–1759.

García-Peydró, M., de Yébenes, V.G., and Toribio, M.L. (2006). Notch1 and IL-7 receptor interplay maintains proliferation of human thymic progenitors while suppressing non-T cell fates. J. Immunol. Baltim. Md 1950 177, 3711–3720.

Germar, K., Dose, M., Konstantinou, T., Zhang, J., Wang, H., Lobry, C., Arnett, K.L., Blacklow, S.C., Aifantis, I., Aster, J.C., et al. (2011). T-cell factor 1 is a gatekeeper for T-cell specification in response to Notch signaling. Proc. Natl. Acad. Sci. U. S. A. 108, 20060–20065.

Hosokawa, H., Romero-Wolf, M., Yui, M.A., Ungerbäck, J., Quiloan, M.L.G., Matsumoto, M., Nakayama, K.I., Tanaka, T., and Rothenberg, E.V. (2018a). Bcl11b sets pro-T cell fate by site-specific cofactor recruitment and by repressing Id2 and Zbtb16. Nat. Immunol. 19, 1427–1440.

Hosokawa, H., Ungerbäck, J., Wang, X., Matsumoto, M., Nakayama, K.I., Cohen, S.M., Tanaka, T., and Rothenberg, E.V. (2018b). Transcription factor PU.1 represses and activates

gene expression in early T cells by redirecting partner transcription factor binding. Immunity 49, 782.

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.

Kueh, H.Y., Yui, M.A., Ng, K.K., 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.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 162, 184–197.

Li, L., Leid, M., and Rothenberg, E.V. (2010a). An early T cell lineage commitment checkpoint dependent on the transcription factor Bcl11b. Science 329, 89.

Li, P., Burke, S., Wang, J., Chen, X., Ortiz, M., Lee, S.-C., Lu, D., Campos, L., Goulding, D., Ng, B.L., et al. (2010b). Reprogramming of T cells to natural killer-like cells upon Bcl11b deletion. Science 329, 85–89.

Liu, P., Li, P., and Burke, S. (2010). Critical roles of Bcl11b in T-cell development and maintenance of T-cell identity: Bcl11b has essential functions in T cells. Immunol. Rev. 238, 138–149.

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. Proc. Natl. Acad. Sci. 114, 5800–5807.

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 αβ T cell differentiation. Nat. Immunol. 14, 619–632.

Montel-Hagen, A., Sun, V., Casero, D., Tsai, S., Zampieri, A., Jackson, N., Li, S., Lopez, S., Zhu, Y., Chick, B., et al. (2020). In vitro recapitulation of murine thymopoiesis from single hematopoietic stem cells. Cell Rep. 33, 108320.

Naik, S.H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., and Schumacher, T.N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature 496, 229–232.

Naito, T., Tanaka, H., Naoe, Y., and Taniuchi, I. (2011). Transcriptional control of T-cell development. Int. Immunol. 23, 661–668.

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

Porritt, H.E., Gordon, K., and Petrie, H.T. (2003). Kinetics of steady-state differentiation and mapping of intrathymic-signaling environments by stem cell transplantation in nonirradiated mice. J. Exp. Med. 198, 957–962.

Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol.

Romero-Wolf, M., Shin, B., Zhou, W., Koizumi, M., Rothenberg, E.V., and Hosokawa, H. (2020). Notch2 complements Notch1 to mediate inductive signaling that initiates early T cell development. J. Cell Biol. 219, e202005093.

Rothenberg, E.V., Moore, J.E., and Yui, M.A. (2008). Launching the T-cell-lineage developmental programme. Nat. Rev. Immunol. 8, 9–21.

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

Scripture-Adams, D.D., Damle, S.S., Li, L., Elihu, K.J., Qin, S., Arias, A.M., Butler, R.R., Champhekar, A., Zhang, J.A., and Rothenberg, E.V. (2014). GATA-3 dose-dependent checkpoints in early T cell commitment. J. Immunol. Baltim. Md 1950 193, 3470–3491.

Shibata, K., Yamada, H., Nakamura, M., Hatano, S., Katsuragi, Y., Kominami, R., and Yoshikai, Y. (2014). IFN-γ–producing and IL-17–producing γδ T cells differentiate at distinct developmental stages in murine fetal thymus. J. Immunol. 192, 2210–2218.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 19.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21.

Taghon, T.N., David, E.S., Zúñiga-Pflücker, J.C., and Rothenberg, E.V. (2005). Delayed, asynchronous, and reversible T-lineage specification induced by Notch/Delta signaling. Genes Dev 19, 965–978.

The Immunological Genome Project Consortium, Heng, T.S.P., Painter, M.W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S.J., Koller, D., Kim, F.S., Wagers, A.J., et al. (2008). The Immunological Genome Project: Networks of gene expression in immune cells. Nat. Immunol. 9, 1091–1094.

Thompson, P.K., and Zúñiga-Pflücker, J.C. (2011). On becoming a T cell, a convergence of factors kick it up a Notch along the way. Semin. Immunol. 23, 350–359.

Vidigal, J.A., and Ventura, A. (2015). Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. Nat. Commun. 6, 8083.

Wakabayashi, Y., Watanabe, H., Inoue, J., Takeda, N., Sakata, J., Mishima, Y., Hitomi, J., Yamamoto, T., Utsuyama, M., Niwa, O., et al. (2003). Bcl11b is required for differentiation and survival of αβ T lymphocytes. Nat. Immunol. 4, 533–539.

Weber, B.N., Chi, A.W.-S., Chavez, A., Yashiro-Ohtani, Y., Yang, Q., Shestova, O., and Bhandoola, A. (2011). A critical role for TCF-1 in T-lineage specification and differentiation. Nature 476, 63–68.

Yu, Y., Wang, J., Khaled, W., Burke, S., Li, P., Chen, X., Yang, W., Jenkins, N.A., Copeland, N.G., Zhang, S., et al. (2012). Bcl11a is essential for lymphoid development and negatively regulates p53. J. Exp. Med. 209, 2467–2483.

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.

Zhang, J.A., Mortazavi, A., Williams, B.A., Wold, B.J., and Rothenberg, E.V. (2012). Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. Cell 149, 467–482.

Zhou, W., Yui, M.A., Williams, B.A., Yun, J., Wold, B.J., Cai, L., and Rothenberg, E.V. (2019). Single-cell analysis reveals regulatory gene expression dynamics leading to lineage commitment in early T cell development. Cell Syst. 9, 321-337.e9.

Chapter 4

OPPORTUNITIES, CHALLENGES, AND PERSPECTIVES

In this thesis, we first established a detailed model of single-cell transcriptome dynamics during the transition from multipotentiality to T-cell lineage commitment, with single-cell sequencing tools, bolstered by highly sensitive seqFISH analysis, and supported by *in vitro* differentiation kinetics. To further understand the functional relevance of TFs that exhibit these dynamic gene expression patterns, we optimized a few *ex-vivo* culture systems to derive cohorts of early T cells from bone marrow precursors, with and without perturbations, and examined the outcomes of population distributions and trajectory topologies upon perturbations using single-cell analysis. For the first time, we revealed the complexed roles of TFs in regulating the topology of early T-cell differentiation trajectory, cell cycle state, alternative lineage potentials, and differentiation kinetics. In this chapter, the author will discuss some additional technical opportunities and challenges in using single-cell analysis for future understanding of regulatory mechanisms with respect to our T-cell development system, or to developmental processes in general.

Does Deeper Sequencing Solve More Problems?

In earlier chapters, we have briefly mentioned the underlying problem of high dropout rate for droplet-based scRNA-seq techniques, e.g. 10X Genomics. Yet, the ease of large-scale sample preparation, the compatibility of Cell Hashing and direct capture 'perturbseq' make the usage of the 10X platform still more accessible among other single-cell tools. In fact, the author has shown that even with this presumably 'zero-inflated' transcript count data matrix, we could still resolve many interesting biological questions in the early T-cell developmental continuum. One underlying question that affects all scRNA-seq strategy and budget allocation, which the author has not touched upon, is sequencing depth. Should deeper sequencing of fewer cells or shallower sequencing of more cells be favored: a practical limit in the total number of reads that can be sequenced per experiment. There is no consensus or general rules for sequencing depth requirement, and in reality, this is likely

highly question dependent and system dependent. While the 10X V3 recommends a minimum coverage of 20k reads/cell, the most recently published data vary hugely – single-cell datasets for relevant questions in developmental biology or hematopoiesis only, from <10k reads/cell (Bulaeva et al., 2020; Hawkins et al., 2020; Naganuma et al., 2021) to, less commonly, 100k reads/cell (Holloway et al., 2020). Interestingly, the conclusions made by various groups regarding the sequencing depths, using computational tools and datasets from different platforms, are rather controversial. In 2014, Jaitin et al. suggested that 20k reads per cell with the plate-UMI-based MARS-seq tool, only unambiguously define 200 to 1500 distinct RNA molecules, could accurately represent cell types. Other computational groups also supported the idea of shallower sequencing of more cells, for example, Zhang et al. suggested optimal sequencing coverage of '1 read per gene per cell', and Svensson et al. showed potentially marginal return of deeper sequencing beyond 15k reads/cell. However, as shown in Chapter 2, at the sequencing depth of 50k/cell using 10X V2, we knew that we were detecting ~10% of the lowly expressed molecule, comparing to seqFISH, and clearly agreeing with many published studies (Islam et al., 2014; Kolodziejczyk et al., 2015; Svensson et al., 2017; Torre et al., 2018). Also, as shown in Figure1 (adapted from Svensson et al. 2017 and Mereu et al., 2020, and from data generated in house), it is clear that 20k reads/cell does not saturate the detection limit. In fact, 20k reads/cell is even further from saturation in data generated from 10X Chromium V3 chemistry (Figure 1d). For all of the experiments the author performed and discussed in the previous chapters, the author had recovered a minimum of 50-60k reads/cell (median). The questions are: Would the low dimensional representations change if the author had sequenced less? And would it be worth it to sequence less cells in exchange for deeper sequencing (twice as deep at 120k reads/cell or four times at 200k+ reads/cell)?

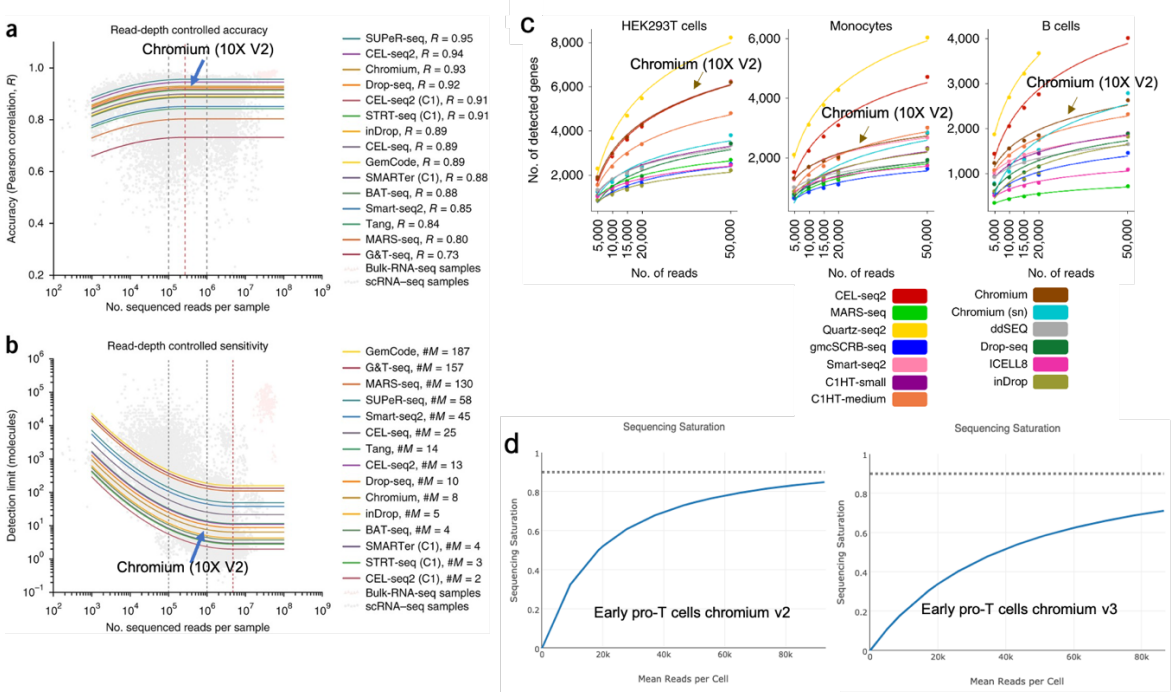


Figure 1. Sequencing Depth's Influence of Genes Detected and Sequencing Saturation in Previous Studies and in The Developing Early T cells. (Panel a and b were modified from Svensson et al. 2017, panel c was modified from Mereu et al., 2020, and panel d was generated in house.) *(A)* Accuracy is marginally dependent on sequencing depth beyond 100k reads/cell. Saturation occurs at 270,000 reads per cell in the model (dashed red line). Protocol names are ordered by performance on the basis of predicted correlation (R) at 1 million reads. *(B)* Depending on the techniques, sensitivity (i.e. detection limit for lowly expressed transcripts) can be critically dependent on sequencing depth. Saturation occurs at 4.6 million reads per cell (dashed red line). The gain from 1 to 4 million reads per sample is marginal, whereas moving from 100,000 reads to 1 million reads corresponds to an order-of-magnitude gain in sensitivity (dashed black lines). Protocols are ordered by performance on the basis of predicted detection limit (#M, number of molecules at 1 million reads). *(C)* Number of detected genes at down-sampled sequencing depths. Points represent the average number of detected genes as a fraction of all cells of the corresponding cell type at the corresponding sequencing depth. *(D)* Saturation curves generated with CellRanger, representing detection of additional transcripts with down-sampled sequencing depth. The

left and right panels represent thymic early pro-T cell scRNA-seq data generated with 10X Chromium V2 and 10X Chromium V3, respectively. While with 10X V2 chemistry, the saturation is close to be achieved at 80k reads/cell, the saturation is less than 70% using 10X V3 chemistry at the same sequencing depth, inferring that a better sensitivity can certainly be achieved through deeper sequencing.

Generally, the low dimensional representations, as shown by Svensson et al. (2019), are unlikely to change if the sequencing depth were slightly lower. This is partially due to the fact that dimension reductions are usually performed based on well-expressed and highly variable genes only. Also, the clustering and trajectory formed through complete unsupervised analysis are usually less sensitive to lowly expressed genes that are around the detection limit. This usually means if the sample actually contains multiple discrete cell types, or very distinct highly expressed features across the trajectory, a slightly shallower sequencing would give the same results. However, we found that the complete unsupervised clustering and trajectory with highly expressed variable genes do not accurately represent developmental trajectories in our developing early T cells. A more accurate developmental trajectory can be obtained through the usage of a curated list of genes for building the connected graph in low dimensional space and for ordering the cells in pseudotime. Some of the genes in the curated list, however, are lowly expressed. Therefore, the sequencing depth was likely needed for trajectory analysis in populations with subtle expression changes. The latter question regarding decreasing cell number in exchange for further increase of depth, also depends on the situation. For heterogeneous population that may contain rare cell population of interest, or pool-based ‘perturbseq’ assays, a decrease in cell number coverage will reduce the statistical confidence, therefore it certainly will not be preferred. In addition, from past experiences, most sequencing reads are mapped to genes encoding cell cycle related processes and ribosomal proteins. The increased read depth will again mostly map against these already highly expressed genes that is of less interest, resulting in a huge waste of resource. However, there are potential workarounds for using the new targeted sequencing technology with the 10X platform (Replogle et al., 2020) ; or using seqFISH for even better accuracy of detection for TFs (Zhou et al., 2019, or chapter

2). The only limitations are that the list needs to be curated before the acquisition, and potential data analytical challenges to integrate datasets with different sets of genes measured in different experiments.

Moreover, imputation methods for dropouts have been developed by many computational groups over the past 3-4 years. For example, MAGIC (van Dijk et al., 2018) imputes missing expression values by sharing information across similar cells, resulting in an essentially smoothening effect. This smoothening concept is also heavily used in RNA velocity (La Manno et al., 2018) calculation, as intron-mapped reads are not intentionally captured through single end (3' or 5' end of the mRNA transcripts) scRNA-seq methods, therefore they are very sparse. While imputation can efficiently bring up the real dropout, it can often lead to the 'over-smoothing' problem and loss of biologically relevant information, such as stochasticity or other variability. Filling in the dropout should not be a computational effort alone. Further improvement of the chemistry for better capture efficiency, combinations of targeted gene panel (Replogle et al., 2020), potentially targeted splicing variant or intron panel in the future, or using imaging-based methods such as intron seqFISH (Shah et al., 2018) and seqFISH+ (Eng et al., 2019), will not only decrease the need for imputation, but also preserve more biologically meaningful variances. In short, if one knows the genes of interest (or intronic regions of interest), targeted panel single-cell sequencing or imaging-based tools offer great opportunity to improve the sequencing depth vs. cell number problem, and will potentially improve the quality of developmental trajectory inference, GRN inference, and RNA velocity analysis.

How Does a New Dataset Align with the Previous Data?

Another challenge in the field of single-cell analysis is the so-called 'unified analysis'. Samples collected across different methods, platforms, experimental setups, animals /patients, and batches can be extremely challenging to compare with. Obviously, these 'batch effects' can lead to false discovery, and also the identification of shared cell types or states can be very complicated. As discussed and demonstrated in earlier chapters, ideally, careful experimental design using multiplexed scRNA-seq to pool cells into a same batch

of sequencing is desired. However, this may not be practical due to logistical limitations concerning sample preparation, time constraints, etc. As many single-cell atlas projects are in progress including a most recent organoid atlas project (Regev et al., 2017; Rozenblatt-Rosen et al., 2017; the Human Cell Atlas 'Biological Network' Organoids et al., 2020), how can new researchers take proper usage of atlas single-cell data as references?

In Chapter 1, the author has briefly mentioned some commonly used computational alignment methods for different datasets and some of their underlying assumptions (e.g. multiCCA, MNN, Harmony). In fact, Seurat (a well-known and widely used software package for scRNA-seq analysis) V4 was just released around the time the author was drafting this thesis (Hao et al., 2020), using a 'weighted-nearest neighbor' (WNN) approach to integrate atlas style scRNA-seq data. In order to map a new dataset to the atlas reference, a 'supervised PCA analysis' can be performed to identify a projection of the transcriptome dataset that maximally captures the structure defined in the atlas WNN graph. This method can potentially 'supervise' the analysis of gene expression data to ignore the variables that are irrelevant to the WNN graph of interest, improving cell type identification and robust positioning on developmental trajectories for new datasets.

Around the same time, a few approaches based on deep learning methods started to emerge, leveraging the advances in wet lab capability of generating large scale data and the increase of computing power. For example, DCA (Eraslan et al., 2019) utilizes the autoencoder concept to denoise the data, returning the scaled gene by cell matrix, which is the exact same size as the input. scVI (Lopez et al., 2018) uses deep generative modeling based on a hierarchical Bayesian model (with the assumption of a zero-inflated negative binomial transcript count distribution obtained by scRNA-seq methods), removing unwanted factors (e.g. batch effects), and returning latent space vectors that can serve as input to downstream analysis. In contrast, SAUCIE (Amodio et al., 2019) uses a deep neural network in which some of the layers are designed to perform cluster annotation and 2D visualization, returning the output of cluster annotation and low dimensional visualization directly. However, because some of the output latent space obtained from deep neural network methods are not

as interpretable as methods like PCA or NMF, they can be susceptible to overfitting and other technical issues. Lately, more tools were developed using generative adversarial networks (GANs) for scRNA-seq imputation, *in silico* data generation and augmentation (Marouf et al., 2020; Xu et al., 2020), aiming to boost the robustness of detecting biologically interesting variable features against technical noises and batch effects.

Around the time when we published our early T-lineage single-cell study in mouse systems (Zhou et al. 2019, presented in chapter 2), several other groups also published similar single-cell profiles of mouse and human, with the focuses on different stages and cell populations. In mouse systems, a comprehensive, dynamic single-cell analysis of hematopoietic and stromal cells during thymic organogenesis in the mouse fetus was published by Kernfeld et al. in 2018, and this was complemented by single-cell dissections of thymic stromal cell types by Bornstein et al. in 2018. In human systems, Zeng et al., 2019; Lavaert et al., 2020; and Le et al., 2020 also published valuable single-cell analyses on human early T lymphopoiesis and thymic stromal cell development. There are unprecedented opportunities to discover new scientific insights by comparing these single-cell datasets through integration methods mentioned above.

In summary, with the increasing availability of public datasets and advancement of computational tools, one should not overlook the power of prior knowledge from these old datasets. Perhaps, in the future, the easiest and most straightforward usage of 'prior knowledge' is pre-experimental data analysis with publicly available single-cell data of similar cell types, regardless of the platform of data acquisition. Not only will future experimental design processes benefit from the datamining process, but also the collections of data may increase the statistical power and the confidence for the conclusions being drawn on the new datasets.

Going Beyond Descriptive Single-Cell Analysis

There are a few commonly accepted stages of data analytic maturity, often used in areas such as statistics and business analytics. Although the exact terminology may differ, the analytical stages often include these 3 levels in a progressing manner: "descriptive analysis",

"predictive analysis", and "prescriptive analysis". While descriptive analysis is often used to quantify relationships between different features of the samples or samples themselves, predictive analysis leverages predictive models to analyze a specific performance in a sample and one or more measured features in the same sample. The objective for the latter model is to assess the likelihood that similar features in a different sample will exhibit the same performance. Prescriptive analytics suggest the decision options to take advantage of the results of the descriptive and predictive analysis. This concept is easily applicable to single-cell analysis. The major use of single-cell analysis up to this point has focused on descriptive analytics, e.g. identifying cell types or states, differential expression analysis, integrating datasets, denoting new markers in clusters, etc. Much effort in the single-cell analytical field has been put to extend the single analysis to tackle some non-trivial problems, and to explore the usage of predictive analytics. In previous chapters, we also went beyond the usage of classical single-cell descriptive analysis through trajectory inference which we later experimentally validated, and the GRN inferences based on the internal-controlled perturbation scRNA-seq data using SCENIC. However, some of the newer tools promise a further integration of multi-modal single-cell data (e.g. scATAC-seq) or previous knowledge on GRN topology, thereby leveraging a better predictive power of single-cell analysis. CellOracle is a machine learning-based tool to infer GRNs via the integration of different single-cell data modalities (i.e. transcriptome and chromatin accessibility profiles), and can also potentially integrate prior knowledge via regulatory sequence analysis to infer TF-target gene interactions (Kamimoto et al., 2020). Note that the major difference between SCENIC and CellOracle is that SCENIC calculates the potential target of TF through a motif search near transcription start sites (up to 10kb) of co-expressed genes; CellOracle builds on the SCENIC strategy, but expands motif search to co-accessible regions of the chromatin of the transcription start sites (from scATAC-seq data) (Kamimoto et al., 2020). Note that regulatory regions of the genes can easily be megabases away from the transcription start sites, therefore CellOracle's expansion of the regulatory region search may be very important to better identify the TF targets. Moreover, some TFs may lack the known binding motifs, therefore it may be helpful in the future to incorporate custom TFs' ChIP-seq profiles to further assist GRN inferences from scRNA-

seq data, rather than having the target gene lists determined completely based on motif search. Moreover, CellOracle can also leverage the inferred GRN to simulate gene expression changes in response to TF perturbations, *in silico*. *In silico* perturbation is a truly exciting concept and a promising usage for predictive analytics of single-cell field, especially on the arguably most important governors of developmental biology – TF regulated processes. In fact, a few other groups have also demonstrated the predictive potentials with existing single-cell datasets (Dibaeinia and Sinha, 2020; Sun et al., 2020; Tian et al., 2020; Zhang et al., 2019). SERGIO (Dibaeinia and Sinha, 2020) simulates stochastic gene expressions in steady-state or differentiating cells according to a user-provided GRN. It is worth noting that Dibaeinia and Sinha, 2020 demonstrated an *in silico* perturbations of some of the interesting TFs leveraging our single-cell seqFISH data presented in Chapter 2 and Zhou *et al*. 2019, and using GENIE3 (Huynh-Thu et al., 2010) predicted GRN as well as the GRN information from Longabaugh et al. 2017. First, they demonstrated using either GRN calculated through GENIE3 or published GRN annonation (Longabaugh et al., 2017), SERGIO could very nicely simulate the profiles that resembled the seqFISH data. Interestingly, they also generated "*in silico* knockout" of *Tcf7, Runx1, Hes1, Bcl11b, Spi1, Lmo2, Gata3,* and *Gfi1b*, many of which have recently been experimentally persued by other members of our lab (Romero-Wolf et al., 2020, Shin et al. in press) or the author herself (presented in Chapter 3). Among the genes being *in silico* knocked out, *Tcf7* KO seemed to exhibit an agreement of our experimental results discussed in Chapter 3. However, although they showed that perturbation resulted in alterations of developmental trajectories, some of the other KOs did not seem to reflect our experimental observations, such as *Gata3*, *Bcl11b*, *Hes1* (Romero-Wolf et al., 2020), and *Runx1*(Shin et al. in press). This could be largely due to the incomplete knowlegde of the underlying GRNs and limited number of genes included in the seqFISH panel. However, *in silico* perturbations with SERGIO, CellOracle, or other co-expression focused predictive methods alike (Sun et al., 2020; Tian et al., 2020), may help refine a shorter list for potential perturbation targets of interest for experimental validations, assisting the 'next-generation' single-cell experimental design, offering a glance of the 'prescriptive' power of single-cell analysis.

BIBLIOGRAPHY

Amodio, M., van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., et al. (2019). Exploring single-cell data with deep multitasking neural networks. Nat. Methods *16*, 1139–1145.

Bornstein, C., Nevo, S., Giladi, A., Kadouri, N., Pouzolles, M., Gerbe, F., David, E., Machado, A., Chuprin, A., Tóth, B., et al. (2018). Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells. Nature *559*, 622–626.

Bulaeva, E., Pellacani, D., Nakamichi, N., Hammond, C.A., Beer, P.A., Lorzadeh, A., Moksa, M., Carles, A., Bilenky, M., Lefort, S., et al. (2020). MYC-induced human acute myeloid leukemia requires a continuing IL-3/GM-CSF costimulus. Blood *136*, 2764–2773.

Dibaeinia, P., and Sinha, S. (2020). SERGIO: A single-cell expression simulator guided by gene regulatory networks. Cell Syst. *11*, 252-271.e11.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716-729.e27.

Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature *568*, 235–239.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zagar, M., et al. (2020). Integrated analysis of multimodal single-cell data (Genomics).

Hawkins, F.J., Suzuki, S., Beermann, M.L., Barillà, C., Wang, R., Villacorta-Martin, C., Berical, A., Jean, J.C., Le Suer, J., Matte, T., et al. (2020). Derivation of airway basal stem cells from human pluripotent stem cells. Cell Stem Cell S1934590920304926.

Holloway, E.M., Czerwinski, M., Tsai, Y.-H., Wu, J.H., Wu, A., Childs, C.J., Walton, K.D., Sweet, C.W., Yu, Q., Glass, I., et al. (2020). Mapping development of the human intestinal niche at single-cell resolution. Cell Stem Cell S1934590920305464.

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.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Kamimoto, K., Hoffmann, C.M., and Morris, S.A. (2020). CellOracle: Dissecting cell identity via network inference and in silico gene perturbation (Genomics).

Kernfeld, E.M., Genga, R.M.J., Neherin, K., Magaletta, M.E., Xu, P., and Maehr, R. (2018). A single-cell transcriptomic atlas of thymus organogenesis resolves cell types and developmental maturation. Immunity *48*, 1258-1270.e6.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell *58*, 610–620.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Lavaert, M., Liang, K.L., Vandamme, N., Park, J.-E., Roels, J., Kowalczyk, M.S., Li, B., Ashenberg, O., Tabaka, M., Dionne, D., et al. (2020). Single cell profiling of immature human postnatal thymocytes resolves the complexity of intra-thymic lineage differentiation and thymus seeding precursors (Immunology).

Le, J., Park, J.E., Ha, V.L., Luong, A., Branciamore, S., Rodin, A.S., Gogoshin, G., Li, F., Loh, Y.-H.E., Camacho, V., et al. (2020). Single-cell RNA-seq mapping of human thymopoiesis reveals lineage specification trajectories and a commitment spectrum in T cell development. Immunity *52*, 1105-1118.e9.

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. Proc. Natl. Acad. Sci. *114*, 5800–5807.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D.S., Krebs, C.F., and Bonn, S. (2020). Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. Nat. Commun. *11*, 166.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J.K., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol. *38*, 747–755.

Naganuma, H., Miike, K., Ohmori, T., Tanigawa, S., Ichikawa, T., Yamane, M., Eto, M., Niwa, H., Kobayashi, A., and Nishinakamura, R. (2021). Molecular detection of maturation stages in the developing kidney. Dev. Biol. *470*, 62–73.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. ELife *6*, e27041.

Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol.

Romero-Wolf, M., Shin, B., Zhou, W., Koizumi, M., Rothenberg, E.V., and Hosokawa, H. (2020). Notch2 complements Notch1 to mediate inductive signaling that initiates early T cell development. J. Cell Biol. *219*, e202005093.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017). The Human Cell Atlas: From vision to reality. Nature *550*, 451–453.

Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C.-H.L., Koulena, N., Cronin, C., Karp, C., Liaw, E.J., et al. (2018). Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. Cell *174*, 363-376.e16.

Sun, T., Song, D., Li, W.V., and Li, J.J. (2020). scDesign2: An interpretable simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured (Bioinformatics).

Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat. Methods *14*, 381–387.

Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2019). Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq (Genomics).

the Human Cell Atlas 'Biological Network' Organoids, Bock, C., Boutros, M., Camp, J.G., Clarke, L., Clevers, H., Knoblich, J.A., Liberali, P., Regev, A., Rios, A.C., et al. (2020). The organoid cell atlas. Nat. Biotechnol.

Tian, J., Wang, J., and Roeder, K. (2020). ESCO: Single cell expression simulation incorporating gene co-expression (Genetics).

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. Cell Syst. *6*, 171-179.e5.

Xu, Y., Zhang, Z., You, L., Liu, J., Fan, Z., and Zhou, X. (2020). scIGANs: Single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Res. *48*, e85–e85.

Zeng, Y., Liu, C., Gong, Y., Bai, Z., Hou, S., He, J., Bian, Z., Li, Z., Ni, Y., Yan, J., et al. (2019). Single-cell RNA sequencing resolves spatiotemporal development of pre-thymic lymphoid progenitors and thymus organogenesis in human embryos. Immunity *51*, 930-948.e6.

Zhang, M.J., Ntranos, V., and Tse, D. (2020). Determining sequencing depth in a single-cell RNA-seq experiment. Nat. Commun. *11*, 774.

Zhang, X., Xu, C., and Yosef, N. (2019). Simulating multiple faceted variability in single cell RNA sequencing. Nat. Commun. *10*, 2611.

Zhou, W., Yui, M.A., Williams, B.A., Yun, J., Wold, B.J., Cai, L., and Rothenberg, E.V. (2019). Single-cell analysis reveals regulatory gene expression dynamics leading to lineage commitment in early T cell development. Cell Syst. *9*, 321-337.e9.