Simultaneous Inference of Cell Types, Lineage Trees, and Regulatory Genes From Gene Expression Data

Simultaneous Inference of Cell Types, Lineage Trees, and Regulatory Genes from Gene Expression Data

A dissertation presented

by

Leon Adam Furchtgott

to

The Committee on Higher Degrees in Biophysics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biophysics

Harvard University

Cambridge, Massachusetts

May 2016

Dissertation Advisor: Prof. Sharad Ramanathan Leon Adam Furchtgott

Simultaneous Inference of Cell Types, Lineage Trees, and Regulatory Genes from Gene Expression Data

Abstract

Important goals of developmental biology include identifying cell types, understanding the sequence of lineage choices made by multipotent cells and unconvering the molecular networks controlling these decisions. Achieving these goals through computational analysis of gene expression data has been difficult. In this dissertation supervised by Sharad Ramanathan, I develop a probabilistic framework to identify cell types, infer lineage relationships and discover core gene networks controlling lineage decisions. Working with Sandeep Choubey and Sumin Jang, we infer the gene expression dynamics of early differentiation of mouse embryonic stem cells, revealing discrete state transitions across nine cell states. Using a probabilistic model of the gene regulatory networks, we predict that these states are further defined by distinct responses to perturbations and experimentally verify three such examples of state-dependent behavior. Working with Vilas Menon and Sam Melton, we infer a lineage tree for early neural development and putative regulatory transcription factors from single-cell transcriptomic profiles. The lineage tree shows a prominent bifurcation between cortical and mid/hindbrain cell types, and the inferred lineage relationships were confirmed by clonal analysis experiments. In summary, this study provides a framework to infer predictive models of the gene regulatory networks that drive cell fate decisions.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First I would like to thank my advisor, Sharad Ramanathan, for his guidance over the past six years. I owe a lot of my thinking about science to our conversations over the past years. Sharad has helped me learn how to take inchoate musings and instincts and to flesh them out into a theoretical and computational language.

I also want to thank my thesis committee members Andrew Murray, Leonid Mirny, Ariel Amir, and David Scadden. They have provided valuable advice and encouragement in my efforts. This dissertation, and my experience as a graduate student, would be impoverished without their help.

My scientific experience and daily life at Harvard would be very different without the wonderful current and former members in the Ramanathan lab: Askin Kocabas, Jeffrey Lee, Sumin Jang, Dann Huh, Ching-Han (Hannah) Shen, Zhechung (Lance) Zhang, Sandeep Choubey, Josselin Milloz, Adele Doyle, Ling-Nan Zou, Abdullah Yonar, Ethan Loew, Jim Valcourt, Steven Zwick, and Sam Melton. Hannah cheerfully initiated me to *C. elegans* experiments. Sandeep and Sumin were valuable teammates in the writing of Chapter 3 of this dissertation. It has been a real pleasure working with Sam in the preparation of Chapter 2 over the past few months.

The Harvard Biophysics Program has been a wonderful source of support, and I am particularly grateful to Jim Hogle and Michele Jakoulov for their help in removing bureaucratic hurdles, giving me the freedom to explore, and being there when I needed them the most.

I am also grateful to the following people for their friendship, mentorship and support during my PhD: Danny Ben-Zvi, James Briggs, Jeremy England, Genya Frenkel,

KC Huang, Gerhard Hummer, Christof Koch, Sigiswald Kuijken, Erel Levine, L. Mahadevan, Catherine McKenna, Vilas Menon, Jimmy Mitrani, Toshihiko Oki, Jacob Oppenheim, Vipul Periwal, David Salzman, Fred Sommers, and Ned Wingreen. I am sure that I am forgetting many more.

My parents and siblings have been with me through my entire education. My wife and daughter have been a constant source of encouragement and support throughout my PhD, and I dedicate this dissertation to them.

To Deborah and Alice

Fallax gratia, et vana est pulchritudo; mulier timens Dominum, ipsa laudabitur.

Chapter 1. Introduction

1.1. The problem: understanding development

During development, organisms generate a plethora of specialized cell types, each with unique defining characteristics and functions. Remarkably, this process starts from a single cell, which successively divides and differentiates, leading to the generation of all of the cell and tissue types in an adult organism. Many fascinating questions remain about how the developmental process self-organizes in both space and time. How do millions of individual cells specialize and coordinate in order to build complex patterned organs or powerful networks of neurons?

A step on the way to addressing these questions of developmental patterning and of interactions between cells is the question of how an individual cell makes decisions during development. More concretely, major goals in the field of developmental biology include (a) categorizing and characterizing what constitute the distinct cell types that an individual cell might belong to, (b) understanding the sequence of transitions between these different cell types that an individual cell might make, and (c) characterizing the molecular network that underlies this sequence of transitions. But underlying each cell's behavior is an incredibly complex molecular network composed of thousands of different types of proteins (which are themselves translated from RNA molecules transcribed from DNA), external signaling molecules, epigenetic factors, etc. Moreover, as the cell changes cell state or is in different environments, this molecular network also changes. The number of potential interactions between the different factors comprising a cell's molecular network is astronomical. Understanding and making predictions about such a complicated system

is truly challenging, because of the large number of parameters that are impossible to characterize fully. How then can we proceed?

1.2. Two approaches: traditional biology and large-scale data analysis

Traditional developmental biology and large-scale gene profiling and analysis represent two diverging paradigms for studying development. Traditional developmental biology has made great strides in understanding differentiation in certain key systems (such as mammalian hematopoiesis and Drosophila development) as the result of decades of careful experiments (Gilbert, 2014; Kondo et al., 1997; Orkin and Zon, 2008; Till and McCulloch, 1961). However, in many systems (including human neural development), little is known about the number of cell types, their lineage relationships, or the genes guiding development. The advent of large-scale gene expression profiling techniques (including microarrays (Heng and Painter, 2008) and, more recently, single-cell RNA-Seq (Jaitin et al., 2014; Klein et al., 2015; Zeisel et al., 2015)) offers the promise that large gene expression datasets can allow us to make testable predictions about development.

When we contrast the respective approaches and paradigms used by traditional developmental biologists and by computational biologists, the differences are striking. Take for example studies of differentiation in the hematopoietic system. Hematopoiesis is the development of all of the different cell types that exist in the blood, from red blood cells to white blood cells such as T- and B-cells. All of these cell types have a common ancestor – hematopoietic stem cells – a fact that was established in the early 1960s (Till and McCulloch, 1961). Since then, biologists have attempted to piece together the exact "family tree" of hematopoietic differentiation by isolating and defining one intermediate progenitor at a time. Biologists have also been able to identify many of the key genes

involved in hematopoiesis by carefully discovering one gene at a time using the tools of classical genetics.

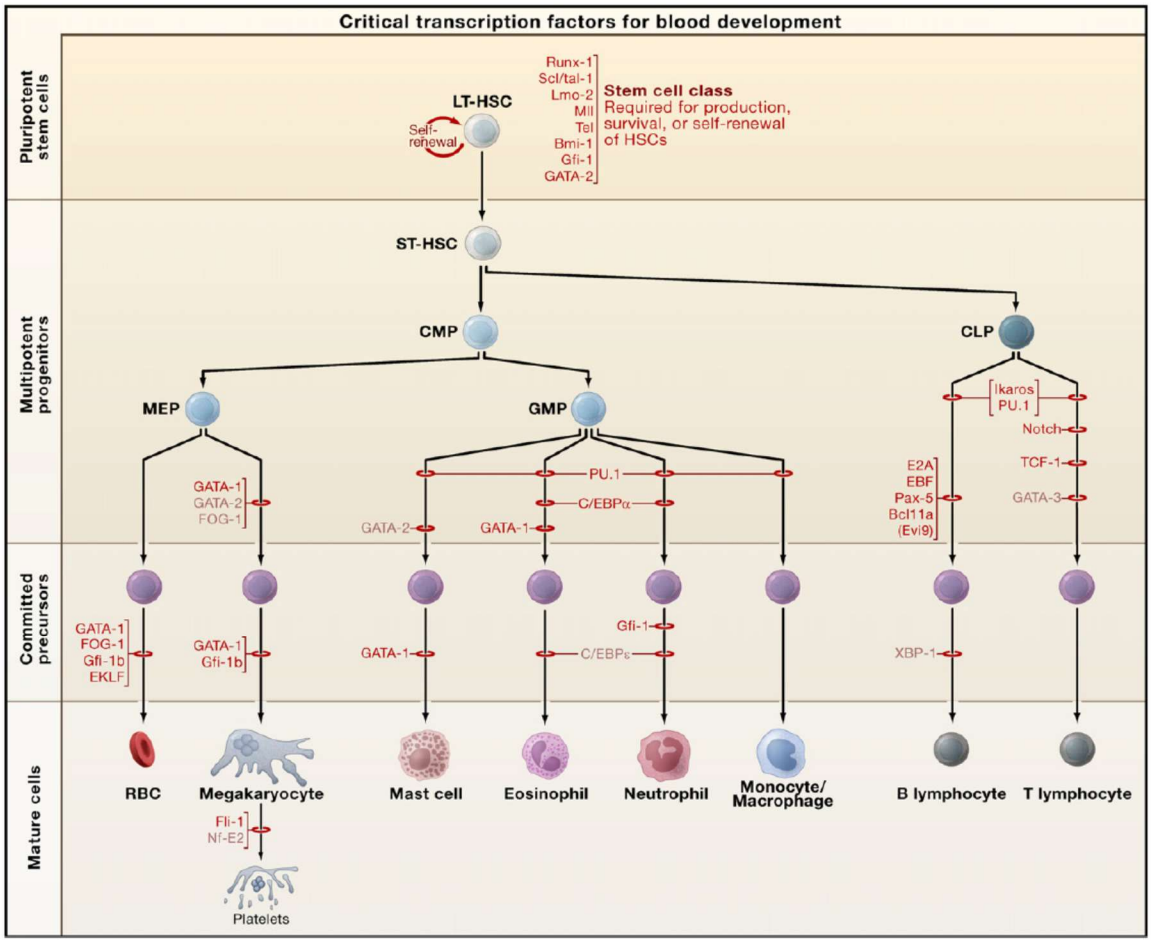


Figure 1.1: Traditional picture of hematopoiesis. Figure reproduced from (Orkin and Zon, 2008). Traditional picture of hematopoiesis, with different mature cell types coming about through a series of cell-fate decisions mediated by key transcription factors. Key genes, as determined through conventional gene knockouts, are indicated by red bars; those associated with oncogenesis in black. Abbreviations: 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; RBCs, red blood cells.

As Figure 1 illustrates, the picture that emerges from traditional biology is one in which hematopoietic stem cells give rise to the different cells in the blood through a series

of binary cell-fate decisions, collectively describing a tree. Each cell-fate decision is the result of a small number of key "master" genes, which in turn control other genes and unleash a process that gives each cell type a defined identity. This picture suggests that it is possible to extract key genes from the cell's complex set of molecular interactions, and these key genes are determinants of cell fate.

A very different picture is painted by approaches that take comprehensive sets of gene expression data and use statistical methods to infer the gene regulatory networks underlying hematopoiesis (Basso et al., 2005; Jojic et al., 2013; Laurenti et al., 2013; Novershtern et al., 2011). As Figure 2 shows, instead of a hierarchy of decision-making master genes, these models show "strong evidence for the role of complex interconnected circuits in hematopoiesis" (Novershtern et al., 2011). Other recent papers relying on single-cell transcriptomics conclude that one of the cell types in Figure 1 (the common myeloid progenitor) in fact should be seen as a composite of multiple distinct subpopulations (Guo et al., 2013; Paul et al., 2015). The explosion of biological measurement has not only provided larger amounts of high-quality data. It has caused us to reevaluate the organization of molecular networks and how they give rise to distinct cell fates.

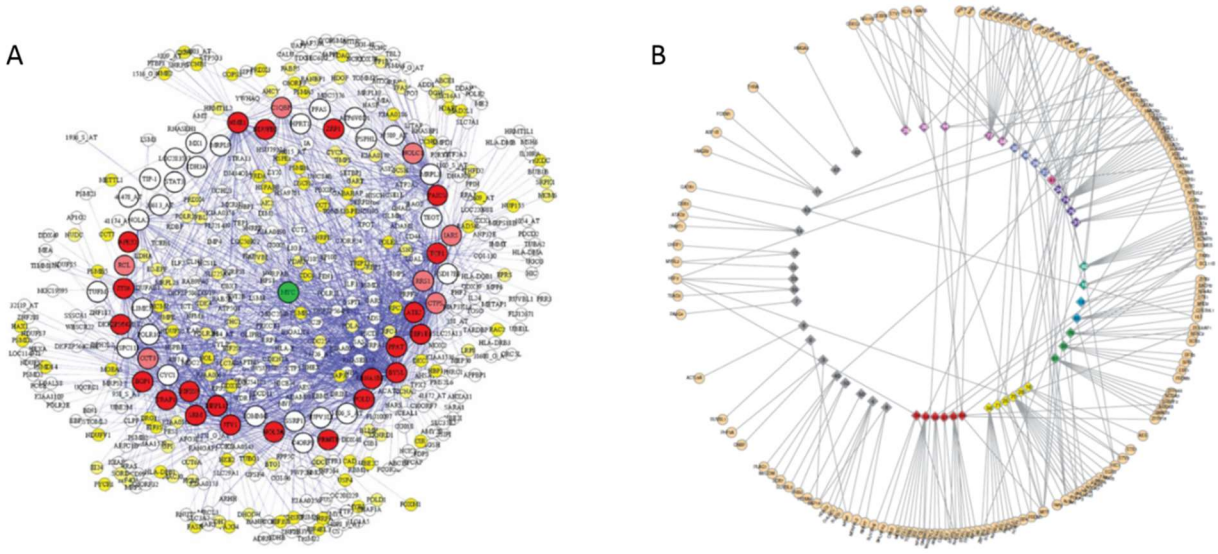


Figure 1.2: Gene regulatory networks inferred from computational models. (A). Figure reproduced from (Basso et al., 2005). Subnetwork of the B-cell developmental network surrounding transcription factor Myc (green). Circles represent genes. (B). Figure reproduced from (Jojic et al., 2013). Diamonds represent coexpressed gene modules; circles represent putative module regulators.

From the "complex systems" point of view of the modern computational biologists of Figure 2, the traditional "master gene" view of Figure 1 is hopelessly naïve. Rather than being determined by a small number of master genes, cell fate is a property of the molecular network as a whole. Yet the traditional master-gene vision has certain advantages, and complex-systems approaches have yet to account for them properly.

First, although hematopoietic development undoubtedly involves the coordination of a very large number of molecular factors, the macroscopic phenomenon that results – namely the daily production of a hundred billion blood cells by hematopoietic progenitors – is one that is robust, reliable and consists of a series of cell-fate decisions. The traditional picture gives a mechanistic model for how cells make decisions, which allows for tracking cells, generating further hypotheses, and modeling the system. In contrast, it is unclear how

a complex network such as the ones shown in Figure 2 can give rise to a series of simple cell-fate decisions.

Furthermore, recent reprogramming and transdifferentiation experiments demonstrate that a small number of key transcription factors (fewer than 4) are sufficient to induce changes in cell state (Graf and Enver, 2009; Iwasaki and Akashi, 2007; Takahashi and Yamanaka, 2006). If cell state is determined by a complex network, reprogramming and transdifferentiation should be almost impossible. Reprogramming could not have been predicted from the complex networks of Figure 2. The fact that reprogramming and transdifferentiation have been observed in a large variety of contexts suggests that some kind of simple structure exists under all of the complex details of the regulatory networks.

1.3. A toy model

A third potential reason to worry about large-scale computational approaches to understanding development has to do with the assumptions behind techniques used for network inference.

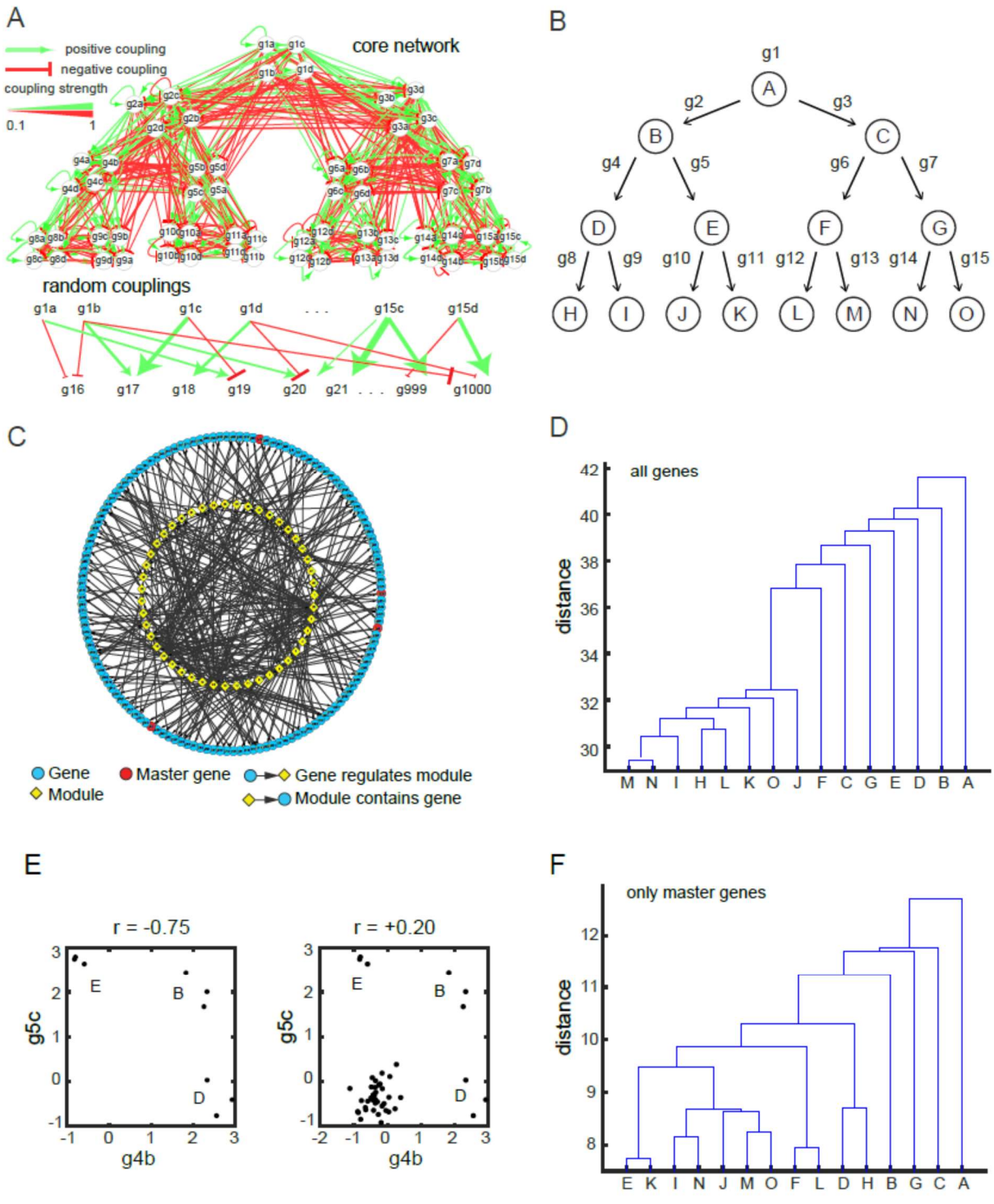


Figure 1.3: Generation and analysis of a synthetic gene network (A) Gene network used to generate the data. Top: the core network is composed of pairs of mutually-inhibitory master gene modules (g1-g15) which sequentially drive subsequent gene pairs.

(Figure 1.3, continued) Bottom: 1000 additional genes are driven by the master genes with couplings that can be either positive or negative (pointed or flat arrows), and with strengths taken from a Gaussian distribution (represented by the different arrow thicknesses). (B) Lineage tree generated by the network, containing 15 cell types denoted A-O. The master gene module associated with each decision is shown above the corresponding arrow. (C) Interaction network between the 48 modules (yellow diamonds) and 141 regulator genes (circles) identified by Module Networks for gene expression data generated by the network shown in (A). Arrows from regulator genes to modules indicate regulation of the module by the gene; arrows from modules to regulator genes indicate inclusion of the regulator gene within the modules. The red circles indicate true master genes; blue circles indicate non-master genes that are identified as regulator genes. (D), (F) Average-linkage hierarchical clustering dendrograms of cell types A-O from the computationally-generated gene expression data. The distance metrics used are the Euclidean distance of the gene expression data for all genes (D) and for only the master genes (F). (E) Scatter plots of the expression of genes g4b and g5c in the replicates of cell types B, D and E (left) and all cell types (right). The value of Pearson correlation (r), shown above each scatter plot, is negative when calculated over cell types B, D and E but positive when calculated over all cell types.

To better understand the computational challenges in lineage and regulatory network inference, we built a mathematical model of a network that gives rise to a series of lineage decisions. Inspired by the traditional picture from Figure 1, and numerous examples of mutually-inhibitory pairs of factors determining cell fate (Graf and Enver, 2009; Qi et al., 2013; Thomson et al., 2011; Zhang et al., 1999), our model gene network contains 7 mutually-inhibitory pairs of 'master' gene modules, with global activatory and inhibitory interactions between modules as well as activatory and inhibitory interactions within modules. In addition, the network includes one thousand other genes randomly coupled to the master genes (Figure 2A). This simulated network gives rise to 15 progenitor and terminally-differentiated cell types along a lineage tree (Figure 2B). We generated triplicate gene expression data for the different cell types by adding noise to our underlying network to reflect measurement noise as well as biological variability (Figure 2C). We hid the original network and the lineage relationships from ourselves, and sought, using only

this gene expression data, to infer both the lineage relationships between the cell types and the core master-gene network.

None of the established computational methods we applied to our data were successful in recovering either the original lineage tree or the key features of the underlying differentiation network. Classification of the genes using Module Networks (Segal et al., 2003), a widely-used network inference algorithm, yielded 48 distinct gene modules, only 11 of which contained the master genes that generated the lineage tree. Additionally, the Module Networks analysis categorized 141 genes as being regulators of modules, producing a network of modules and regulators (Figure 2C). It was not possible to recover the original master genes from this complex network, nor was it clear how the network gave rise to the observed lineage tree. Further, distances between cell types in hierarchical clustering and principal component analysis (PCA) showed significant differences from the actual lineage (Figure 2D). Thus, inferring the correct lineage relationships from gene expression data alone was also not possible.

We next asked why computational methods failed at inferring the correct network. Key genes involved in lineage decisions showed changing patterns of relative expression during differentiation and thus changing values of correlation between each other. For example, pairs of master genes which have mutually inhibitory relationships in the original network (such as g4b and g5c in our model, Figure 2A) show a strongly negative correlation when evaluated over the progenitor and its descendant cell types ($r=-0.75$ in B, D, E), and a weak positive correlation ($r=+0.20$) when evaluated using all the cell types (Figure 2G). Reuse of factors in different lineage decisions in real developmental networks can only exacerbate such changes in correlation. In the absence of any knowledge of the

lineage relationship, changing patterns of correlation are hard to interpret. Thus, network inference methods that depend on correlation patterns evaluated over all cell types miss these changing patterns of expression in key genes and fail at inferring the correct network.

To understand why inferring lineage relationships between cell types using gene expression is challenging, we questioned whether the distance metrics between cell types used by these methods accurately reflected lineage relationships. When we performed PCA and hierarchical clustering of the cell types using only the original master genes, relative distances corresponded better to lineage relationships (Figure 2F). Clustering methods and PCA measure distances between cell types in gene expression space by using either Euclidean distance (square root of the sum of the squared differences in expression of each gene following Pythagoras) or Hamming distance (the total number of genes that are differentially expressed in the two cell types). Since only a subset of genes reflected the lineage relationships between cell types, we hypothesized that the inclusion of all the other genes corrupted distance metrics.

In this example, measuring more genes does not give more accuracy when measuring similarities between different cell types – only the original master genes serve this purpose. Furthermore, the network inference algorithm was incapable of recovering the original master-gene network from the overall complex network, suggesting that the algorithm's assumptions are naturally biased towards complex-looking networks.

1.4. Bridging the gap

A goal of this dissertation is to develop computational methods to take large-scale gene expression data and to find key variables in order to get to a picture closer to the traditional picture in Figure 1 than the complex networks of Figure 2. This is not because

I am taking sides and saying that the traditional picture is completely correct or that I dismiss the inherent complexity of developmental networks. Rather, fitting data to a traditional lineage tree with key control variables is a useful approximation for the important features of the differentiation network. It gives an interpretable, experimentally testable model of differentiation, and it allows for computational modeling and prediction using a small number of key variables.

This dissertation describes a computational framework to infer cell states, cell state transitions and key genes from gene expression data. The computational framework developed fits the gene expression data to a small, sparse, subset of variables. By projecting the gene expression into a correctly-chosen subspace, we can also infer lineage relationships between cell types and cluster individual cells more accurately. This framework bridges a gap between computation and experiment because it looks for interpretable and measureable variables, probabilistically ranking potential functional genes for each cell-fate decision.

I have applied this framework to gene expression data from mouse hematopoiesis, mouse germ layer differentiation, and human cortical development; these efforts are described respectively in Chapters 2, 3 and 4 in this dissertation.

1.5. Open questions: reaction coordinates in biological processes

I am interested in exploring the role of reaction coordinates in building coarse-grained models of high-dimensional complex systems. There is an extensive literature in the field of computational molecular simulation interested in finding optimal "reaction coordinates" or order parameters for protein folding trajectories (Best and Hummer, 2005, 2011; Du et al., 1998; Krivov, 2013; McGibbon and Pande, 2016). This reduction in

dimensionality still captures and allows for modeling of the slowest-timescale dynamics of the system.

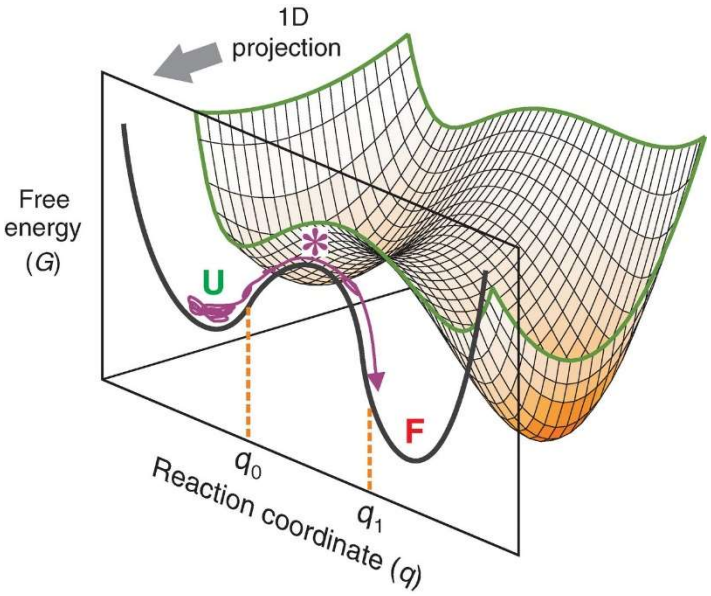


Figure 1.4: Protein Folding Reaction Coordinates Figure reproduced from (Chung et al., 2015). Can we do the same for cellular differentiation?

This work is somewhat analogous to this body of research in that it is looking for reaction coordinates for cell state transitions. I am interested in exploring the theoretical implications of this work: what role do reaction coordinates have in an era of abundant multi-dimensional next-generation sequencing? What constitute good reaction coordinates for biology? And how can we go from a "hairball" complex-network picture of biological networks to a "reaction-coordinate" picture?

1.6. Open questions: definition of cell type

Recent advances in single-cell genomics and transcriptomics have allowed for comprehensive probing of gene expression in thousands of cells in a single experiment. An important goal from these experiments is to identify rare cell types or subpopulations of known cell types that could not be previously recognized due to limitations in bulk

population measurements, which average the expression levels of thousands of cells (Paul et al., 2015; Trapnell, 2015).

Traditionally, cell types have been identified according to certain functional assays, surface markers, and developmental potential (Orkin and Zon, 2008; Pereira et al., 2007). In contrast, given single-cell gene expression measurements for thousands of cells, cell types can be identified in an unbiased way by clustering the cells according to their gene exprression using various statistical techniques (Macosko et al., 2015a; Satija et al., 2015).

Ultimately, however, every individual cell is its own microstate and is slightly different from every other cell. Clustering techniques can classify cell types with arbitrary resolution and with different outcomes depending on assumptions about the nature of the heterogeneity. In order to have useful definitions of cell types, some kind of functional information is necessary. This is currently missing from most single-cell transcriptomics, but it will be important in order to make progress.

1.7. Open questions: data complexity and data analysis.

The framework described in this dissertation looks for one-dimensional patterns in individual genes in order to infer cell types, lineage relationships and key genes. Are one-dimensional patterns in fact sufficient to reconstruct the lineage tree and network? In Chapter 2, we verify that, at least for known lineage relationships in B- and T-cell development, one-dimensional patterns are correlated with and predictive of lineage relationships. In addition, some of the experimental predictions from the framework in the context of germ layer differentiation and neural development have been experimentally tested, and others are in accordance with previous literature. The fact that the simple model

can make experimental predictions suggests the developmental networks we have studied might have a low-dimensional structure.

We could certainly imagine other systems in which one-dimensional patterns are not nearly enough to rebuild the network or to make experimental predictions. Such systems would require recognizing patterns of two genes, three genes, or more in order to reconstruct the network, suggesting that cell identity would be determined by combinatorially complex gene interactions (e.g. cell type A goes to cell type B only if gene 1 and gene 2 are upregulated and gene 3 is downregulated, but cell type A goes to cell type C only if gene 1 and gene 3 are upregulated and gene 2 is downregulated). How to build methods for data analysis that take into account the combinatorial complexity of the underlying data, and how to independently measure this combinatorial complexity, is an open question.

1.8. References

Abrikosov, A.A. (2004). Nobel Lecture: Type-II superconductors and the vortex lattice. Rev. Mod. Phys. *76*, 975–979.

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C.T., Bryder, D., Yang, L., Borge, O.-J., Thoren, L.A.M., et al. (2005). Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell *121*, 295–306.

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature *404*, 193–197.

Anderson, P.W. (1978). Local moments and localized states. Rev. Mod. Phys. *50*, 191–201.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. Mol. Syst. Biol. *3*, 78.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. Nat. Genet. *37*, 382–390.

Best, R.B., and Hummer, G. (2005). Reaction coordinates and rates from transition paths. Proc. Natl. Acad. Sci. U. S. A. *102*, 6732–6737.

Best, R.B., and Hummer, G. (2011). Diffusion models of protein folding. Phys. Chem. Chem. Phys. *13*, 16902–16911.

Buckingham, M.E., and Meilhac, S.M. (2011). Tracing cells for tracking cell lineage and clonal behavior. Dev. Cell *21*, 394–409.

Busslinger, M. (2004). Transcriptional control of early B cell development. Annu. Rev. Immunol. *22*, 55–79.

Chung, H.S., Piana-Agostinetti, S., Shaw, D.E., and Eaton, W.A. (2015). Structural origin of slow diffusion in protein folding. Science *349*, 1504–1510.

Corson, F., and Siggia, E.D. (2012). Geometry, epistasis, and developmental patterning. Proc. Natl. Acad. Sci. U. S. A. *109*, 5568–5575.

Crispino, J.D. (2005). GATA1 in normal and malignant hematopoiesis. Semin. Cell Dev. Biol. *16*, 137–147.

Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T., and Shakhnovich, E.S. (1998). On the transition coordinate for protein folding. J. Chem. Phys. *108*, 334.

Ficara, F., Crisafulli, L., Lin, C., Iwasaki, M., Smith, K.S., Zammataro, L., and Cleary, M.L. (2013). Pbx1 restrains myeloid maturation while preserving lymphoid potential in hematopoietic progenitors. J. Cell Sci. *126*, 3181–3191.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G., and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. Cancer Res. *68*, 5924–5931.

Gazit, R., Garrison, B.S., Rao, T.N., Shay, T., Costello, J., Ericson, J., Kim, F., Collins, J.J., Regev, A., Wagers, A.J., et al. (2013). Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. Stem Cell Reports *1*, 266–280.

Gilbert, S.F. (2014). Developmental Biology (Sinauer).

Goossens, S., Janzen, V., Bartunkova, S., Yokomizo, T., Drogat, B., Crisan, M., Haigh, K., Seuntjens, E., Umans, L., Riedt, T., et al. (2011). The EMT regulator Zeb2/Sip1 is essential for murine embryonic hematopoietic stem/progenitor cell differentiation and mobilization. Blood *117*, 5620–5630.

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

Guo, G., Luc, S., Marco, E., Lin, T.-W., Peng, C., Kerenyi, M.A., Beyaz, S., Kim, W., Xu, J., Das, P.P., et al. (2013). Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. Cell Stem Cell *13*, 492–505.

Heng, T.S.P., and Painter, M.W. (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. *9*, 1091–1094.

Iwasaki, H., and Akashi, K. (2007). Myeloid lineage commitment from the hematopoietic stem cell. Immunity *26*, 726–740.

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.

Jojic, V., Shay, T., Sylvia, K., Zuk, O., Sun, X., Kang, J., Regev, A., Koller, D., Best, A.J., Knell, J., et al. (2013). Identification of transcriptional regulators in the mouse immune system. Nat. Immunol. *14*, 633–643.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M. V, Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

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.

Kondo, M., Weissman, I.L., and Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell *91*, 661–672.

Koschmieder, S., Rosenbauer, F., Steidl, U., Owens, B.M., and Tenen, D.G. (2005). Role of transcription factors C/EBPalpha and PU.1 in normal hematopoiesis and leukemia. Int. J. Hematol. *81*, 368–377.

Krivov, S. V (2013). On Reaction Coordinate Optimality. J. Chem. Theory Comput. *9*, 135–146.

Kurotaki, D., Osato, N., Nishiyama, A., Yamamoto, M., Ban, T., Sato, H., Nakabayashi, J., Umehara, M., Miyake, N., Matsumoto, N., et al. (2013). Essential role of the IRF8-KLF4 transcription factor cascade in murine monocyte differentiation. Blood *121*, 1839–1849.

Landau, L.D., and Lifshitz, E.M. (1951). Statistical Physics, Volume 5 (Elsevier).

Laurenti, E., Varnum-Finney, B., Wilson, A., Ferrero, I., Blanco-Bose, W.E., Ehninger, A., Knoepfler, P.S., Cheng, P.-F., MacDonald, H.R., Eisenman, R.N., et al. (2008). Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. Cell Stem Cell *3*, 611–624.

Laurenti, E., Doulatov, S., Zandi, S., Plumb, I., Chen, J., April, C., Fan, J.-B., and Dick, J.E. (2013). The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. Nat. Immunol. *14*, 756–763.

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc. Natl. Acad. Sci. U. S. A. *102*, 4936–4942.

Machta, B.B., Chachra, R., Transtrum, M.K., and Sethna, J.P. (2013). Parameter space compression underlies emergent theories and predictive models. Science *342*, 604–607.

Macian, F. (2005). NFAT proteins: key regulators of T-cell development and function. Nat. Rev. Immunol. *5*, 472–484.

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. (2015a). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

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. (2015b). Highly Parallel Genome-

wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Margolin, A.A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., and Califano, A. (2006). Reverse engineering cellular networks. Nat. Protoc. *1*, 662–671.

McGibbon, R.T., and Pande, V.S. (2016). Identification of simple reaction coordinates from complex dynamics. 16.

Van der Meer, L.T., Jansen, J.H., and van der Reijden, B.A. (2010). Gfi1 and Gfi1b: key regulators of hematopoiesis. Leukemia *24*, 1834–1843.

Min, I.M., Pietramaggiori, G., Kim, F.S., Passegué, E., Stevenson, K.E., and Wagers, A.J. (2008). The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. Cell Stem Cell *2*, 380–391.

Miyawaki, K., Arinobu, Y., Iwasaki, H., Kohno, K., Tsuzuki, H., Iino, T., Shima, T., Kikushige, Y., Takenaka, K., Miyamoto, T., et al. (2015). CD41 marks the initial myelo-erythroid lineage specification in adult mouse hematopoiesis: redefinition of murine common myeloid progenitor. Stem Cells *33*, 976–987.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

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.

Pereira, C., Clarke, E., and Damen, J. (2007). Hematopoietic colony-forming cell assays. Methods Mol. Biol. *407*, 177–208.

Perrimon, N., Pitsouli, C., and Shilo, B.-Z. (2012). Signaling mechanisms controlling cell fate and embryonic patterning. Cold Spring Harb. Perspect. Biol. *4*, a005975.

Pina, C., May, G., Soneji, S., Hong, D., and Enver, T. (2008). MLLT3 regulates early human erythroid and megakaryocytic cell fate. Cell Stem Cell *2*, 264–273.

Pongubala, J.M.R., Northrup, D.L., Lancki, D.W., Medina, K.L., Treiber, T., Bertolino, E., Thomas, M., Grosschedl, R., Allman, D., and Singh, H. (2008). Transcription factor EBF restricts alternative lineage options and promotes B cell fate commitment independently of Pax5. Nat. Immunol. *9*, 203–215.

Qi, X., Hong, J., Chaves, L., Zhuang, Y., Chen, Y., Wang, D., Chabon, J., Graham, B., Ohmori, K., Li, Y., et al. (2013). Antagonistic regulation by the transcription factors C/EBPα and MITF specifies basophil and mast cell fates. Immunity *39*, 97–110.

Radomska, H.S., Huettner, C.S., Zhang, P., Cheng, T., Scadden, D.T., and Tenen, D.G. (1998). CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. Mol. Cell. Biol. *18*, 4301–4314.

Ragu, C., Boukour, S., Elain, G., Wagner-Ballon, O., Raslova, H., Debili, N., Olson, E.N., Daegelen, D., Vainchenker, W., Bernard, O.A., et al. (2010). The serum response factor (SRF)/megakaryocytic acute leukemia (MAL) network participates in megakaryocyte development. Leukemia *24*, 1227–1230.

Rebollo, A., and Schmitt, C. (2003). Ikaros, Aiolos and Helios: transcription regulators and lymphoid malignancies. Immunol. Cell Biol. *81*, 171–175.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Riddell, J., Gazit, R., Garrison, B.S., Guo, G., Saadatpour, A., Mandal, P.K., Ebina, W., Volchkov, P., Yuan, G.-C., Orkin, S.H., et al. (2014). Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. Cell *157*, 549–564.

Robert-Moreno, A., Espinosa, L., de la Pompa, J.L., and Bigas, A. (2005). RBPjkappa-dependent Notch function regulates Gata2 and is essential for the formation of intra-embryonic hematopoietic cells. Development *132*, 1117–1126.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Satoh, Y., Yokota, T., Sudo, T., Kondo, M., Lai, A., Kincade, P.W., Kouro, T., Iida, R., Kokame, K., Miyata, T., et al. (2013). The Satb1 protein directs hematopoietic stem cell differentiation toward lymphoid lineages. Immunity *38*, 1105–1115.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Solar, G.P., Kerr, W.G., Zeigler, F.C., Hess, D., Donahue, C., de Sauvage, F.J., and Eaton, D.L. (1998). Role of c-mpl in early hematopoiesis. Blood *92*, 4–10.

Stier, S., Cheng, T., Dombkowski, D., Carlesso, N., and Scadden, D.T. (2002). Notch1 activation increases hematopoietic stem cell self-renewal in vivo and favors lymphoid over myeloid lineage outcome. Blood *99*, 2369–2378.

Sugawara, T., Oguro, H., Negishi, M., Morita, Y., Ichikawa, H., Iseki, T., Yokosuka, O., Nakauchi, H., and Iwama, A. (2010). FET family proto-oncogene Fus contributes to self-renewal of hematopoietic stem cells. Exp. Hematol. *38*, 696–706.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.

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

Tamura, T., Nagamura-Inoue, T., Shmeltzer, Z., Kuwata, T., and Ozato, K. (2000). ICSBP directs bipotential myeloid progenitor cells to differentiate into mature macrophages. Immunity *13*, 155–165.

Thomson, M., Liu, S.J., Zou, L.-N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875–889.

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

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

Vaillant, F., Blyth, K., Andrew, L., Neil, J.C., and Cameron, E.R. (2002). Enforced expression of Runx2 perturbs T cell development at a stage coincident with beta-selection. J. Immunol. *169*, 2866–2874.

Wang, H., Lee, C.H., Qi, C., Tailor, P., Feng, J., Abbasi, S., Atsumi, T., and Morse, H.C. (2008). IRF8 regulates B-cell lineage specification, commitment, and differentiation. Blood *112*, 4028–4038.

Zeisel, A., Manchado, A.B.M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-.). *347*, 1138–1142.

Zhang, P., Behre, G., Pan, J., Iwama, A., Wara-Aswapati, N., Radomska, H.S., Auron, P.E., Tenen, D.G., and Sun, Z. (1999). Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. Proc. Natl. Acad. Sci. U. S. A. *96*, 8705–8710.

Chapter 2. Simultaneous inference of lineage trees and regulatory genes from gene expression data

[A large part of this chapter is in preparation for submission as Leon Furchtgott, Samuel Melton, Ling-Nan Zou, Sharad Ramanathan, "Simultaneous inference of lineage trees and regulatory genes from gene expression data." LF and SR designed the study. LF and SM performed computational analyses. LNZ and SR developed the gene regulatory network model and linear programming method. LF and SR wrote the manuscript with input from LNZ and SM.]

Abstract

Two goals of developmental biology are to determine the sequence of lineage choices made by multipotent cells and to understand the molecular networks controlling these decisions. Achieving these goals through computational analysis of gene expression data has been difficult. Here we show that challenges in inferring lineage relationships and gene networks are intrinsically related. We develop a probabilistic framework to simultaneously infer lineage relationships and discover core gene networks controlling lineage decisions. The only free parameter of this framework is our expectation for the number of master regulators for each lineage decision. By applying our methods to analyze gene expression data from the hematopoietic system, we discover both the sequence of lineage decisions and the core network controlling each one. Our study provides a conceptual approach for discovering and quantitatively modelling gene regulatory networks controlling lineage decisions and for making predictions about reprogramming.

2.1. Introduction

During development, pluripotent cells make a series of lineage decisions to give rise to the different cell types of the body. These lineage decisions are controlled by an intra-cellular molecular network that includes transcription factors and signaling molecules. There are two fundamental questions associated with understanding the differentiation of individual cells. The first is to identify lineage relationships: how cells and their progeny move from pluripotent through intermediate to terminally differentiated cell states. The second is to identify the architecture and dynamics of the gene networks that allow cells to make fate decisions along their developmental trajectory.

The problems of reconstructing cell lineages and computationally inferring gene networks have typically been approached separately. Determining lineage relationships between cells, a problem studied since the 19th century, has involved tracking cells and their progeny over time (Buckingham and Meilhac, 2011). This principle has been used repeatedly in different biological contexts. In the nematode *Caenorhabditis elegans*, for example, progenitor cells have been followed visually to determine the lineage of each cell in the entire organism (Sulston et al., 1983). The hierarchy of hematopoietic progenitors has been studied extensively through experiments involving transplantation of prospectively isolated hematopoietic progenitors into lethally irradiated recipients and tracking of the cell progeny (Orkin and Zon, 2008).

Determining lineage relationships between cell types using gene expression alone is challenging because, unlike a tracking dye injected in a progenitor or heritable mutations in DNA, gene expression changes are dynamic and transient. While cell types can be characterized by the gene expression levels of thousands of genes, it is unclear how to

measure distances between cell types in this high-dimensional gene expression space. Figure 2.1B shows the principal component analysis (PCA) of gene expression data for 1,459 transcription factors from 11 early hematopoietic progenitors (Figure 2.1A) (Heng and Painter, 2008). Distances between cell types in this PCA projection and in hierarchical clustering of the cell types (Figure 2.1C) show notable discrepancies with lineage distances known from experimental models of the early hematopoietic cellular hierarchy (Figure 2.1D). Because of the challenges associated with determining lineage from gene expression data, DNA sequencing has been used to determine lineage relationships, for example by tracking mutations in hyper-mutating regions of the genome in mouse tumors (Frumkin et al., 2008).

A

Abbreviation in text	Full name	Phenotype
LT	Long-Term Reconstituting Stem Cell	CD34- Flk2- Lin- ckit+ Sca1+
ST	Short-Term Reconstituting Stem Cell	CD34+ Flk2- Lin- ckit+ Sca1+
MPP	Multipotent Progenitor	CD34+ Flk2+ Lin- ckit+ Sca1+
MLP	Multilineage Progenitor	Lin- AA4+ Kit++ IL7Ra- B220-
CMP	Common Myeloid Progenitor	Lin- IL7R- Sca1- ckit+ FcgRlo CD34+
MEP	Megakaryocyte-Erythroid Progenitor	Lin- IL7R- Sca1- ckit+ FcgRlo CD34-
GMP	Granulocyte-Monocyte Progenitor	Lin- IL7R- Sca1- ckit+ FcgRhi CD34+
CLP	Common Lymphoid Progenitor	Lin- AA4+ Kit+ IL7Ra+ B220-
FrA	Pre-pro-B cell	Lin- AA4+ Kit+ IL7Ra+ B220+
FrBC	Pro-B cell	Lin- AA4+ IgM- CD19+ CD43+ HSA+
ETP	Early T lineage Precursor	4- 8- 11b- 11c- 19- NK1.1- TCR- 44hi 117hi 25-

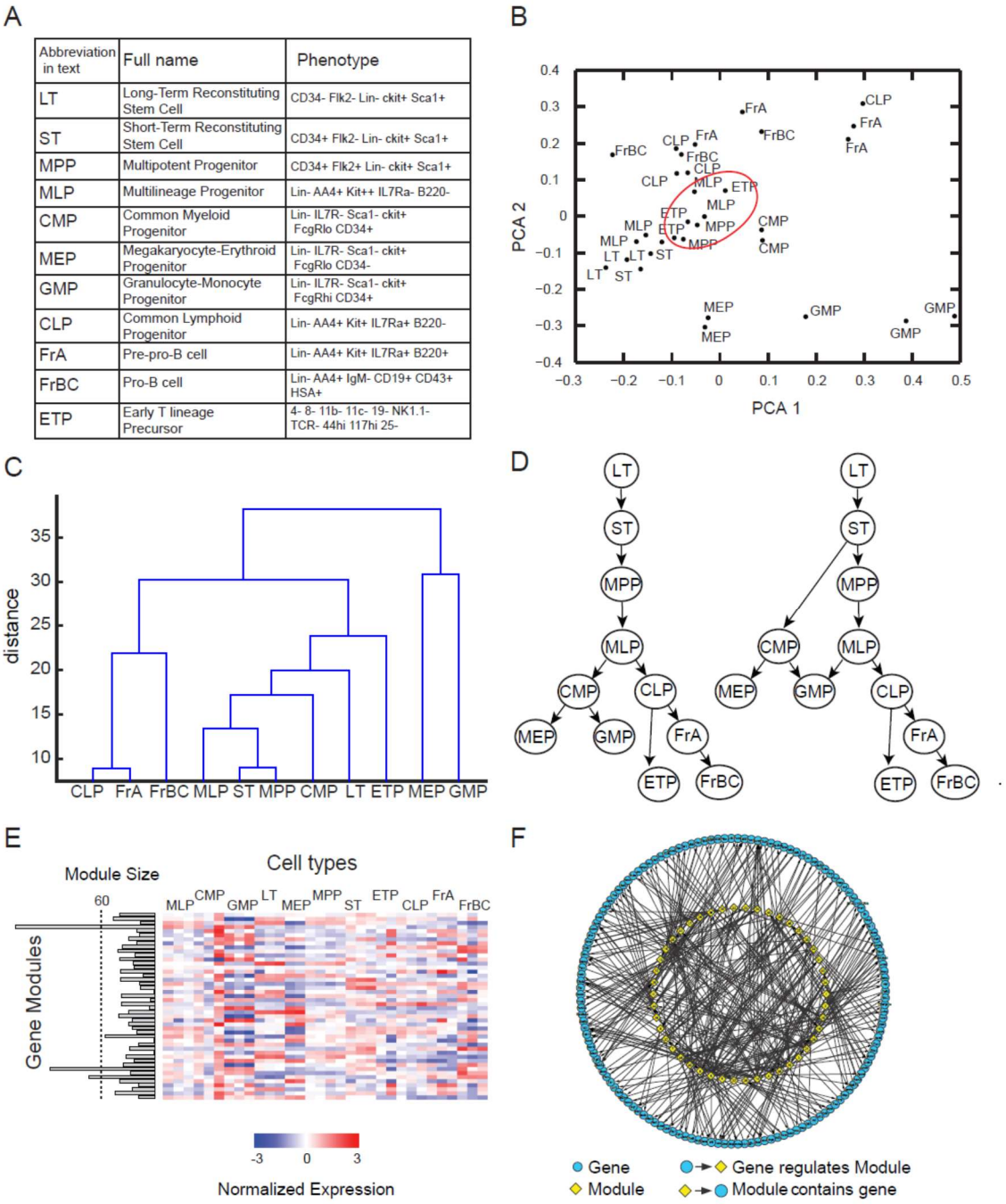


Figure 2.1: Application of common data analysis methods to gene expression data from early hematopoietic progenitors. (A) Hematopoietic cell types considered. Listed for each cell type are the abbreviation used in this paper and its full name and phenotype. (B) Projections of the early hematopoietic progenitors along first two principal components. (PC1: 30%; PC2: 17%). Each point represents a different replicate. Note the proximity between ETP and MPP samples (red circle), which does not reflect either of the lineage models shown in (D).

(Figure 2.1, continued) (C) Average-linkage hierarchical clustering dendrogram of the early hematopoietic progenitors. The distance metric used is the Euclidean distance of the log2-transformed gene expression data. Note that ETP does not cluster with CLP or FrA, which does not reflect either of the lineage models shown in (D). (D) Two models of the hierarchy of early hematopoietic progenitors, both built based on prospective isolation of lineage-restricted progenitors, include (left) the traditional model, in which the first split strictly separates myeloid and lymphoid lineages (Akashi et al., 2000; Kondo et al., 1997; Reya et al., 2001) and (right) an alternative hierarchy proposed by Adolfsson and colleagues, in which lymphoid progenitors subsequent to the first split retain some myeloid potential (Adolfsson et al., 2005). (E) Module Networks analysis of transcription factor microarray data for the 11 early hematopoietic progenitors results in 48 distinct gene modules (rows) with expression levels across all cell type replicates (columns). Bar graph indicates number of genes in each module (left). (F) Interaction network between 143 regulator genes (blue circles) identified in (A) and 48 modules (yellow diamonds). Arrows from regulator genes to modules indicate regulation of the module by the gene; arrows from modules to regulator genes indicate inclusion of the regulator gene within the modules.

Considerable progress has been made in developing algorithms for inferring gene regulatory networks from gene expression data (Bansal et al., 2007; Levine and Davidson, 2005; Margolin et al., 2006; Segal et al., 2003). Such analyses have characterized gene networks as complex and scale-free (Basso et al., 2005) or as complex circuits composed of many interconnected modules of genes (Laurenti et al., 2013; Novershtern et al., 2011). Figure 2.1E and 1F show the complex network of modules and regulators for the same early hematopoietic progenitors, inferred from gene expression data by Module Networks, a widely-used network inference software (Segal et al., 2003). These methods allow us to infer the structure of the underlying networks through the analysis of gene expression patterns across cell types, but understanding how the networks inferred by these methods lead to the sequence of lineage decisions is difficult.

In this chapter, we show that determining lineage relationships between different cell types using only gene expression data allows us to simultaneously infer the networks that control the lineage decisions. We first studied simulated data generated by a

mathematical gene regulatory network model. Despite the simplicity of the network that generated the data, conventional analysis techniques produced incorrect, complex networks with densely interconnected modules and were unable to reconstruct the lineage tree. We next developed a probabilistic approach that infers both lineage relationships and the underlying network simultaneously. We then applied our approach to analyze gene expression data from the hematopoietic system; by computationally reconstructing the lineage tree of the early hematopoietic progenitors we could infer the underlying transcription factor network involved in lineage decisions. Finally, we developed a framework for quantitatively modeling the inferred transcription factor network and made specific predictions about the hematopoietic system, including possibilities of reprogramming. In sum, we show that difficulties inherent in inferring lineage trees and in inferring gene networks can be overcome when considering both problems simultaneously, allowing us to model the underlying network effectively.

2.2. Results

2.2.1. A low-dimensional pattern correlated with lineage transitions

In order to discover patterns of gene expression that are predictive of lineage relationships, we studied known lineage relationships between 41 cell types from B- and T-cell development (Figure 2.2A-B). We identified 150 triplets of cell types with experimentally-verified lineage relationships constituting both cell fate decisions (e.g. cell type A gives rise to cell types B and C) and lineage progressions (cell type A gives rise to cell type B which then gives rise to cell type C). For each triplet, we noted which cell type was the progenitor or intermediate cell type ("intermediate" cell type henceforth) and which cell types were not ("daughter" cell types henceforth). We analyzed transcription

factor microarray gene expression data for these triplets from the Immunological Genome Consortium, including between 2 and 4 biological replicates per cell type (Heng et al 2008).

For each triplet of cell types, we sought to find low-dimensional patterns that were correlated with the lineage relationship. In particular, we evaluated three potential gene-expression patterns across the 150 triplets (Figure 2.2C-E):

- Genes with a "downregulation pattern" for cell type A show strong downregulation of gene expression in cell type A compared to cell types B and C, resulting in a statistically significant *minimum* expression level of the gene in cell type A ($p < 0.005$, two-sample t-test).

- Genes with a "median pattern" for cell type A show significant downregulation of gene expression in B (or C) and strong upregulation of gene expression in C (or B), with a statistically significant *median* expression level of the gene in cell type A ($p < 0.01$, two-sample t-test).

- Genes with an "upregulation pattern" for cell type A show strong upregulation of gene expression in cell type A compared to cell types B and C, resulting in a statistically significant *maximum* expression level of the gene in cell type A ($p < 0.005$, two-sample t-test).

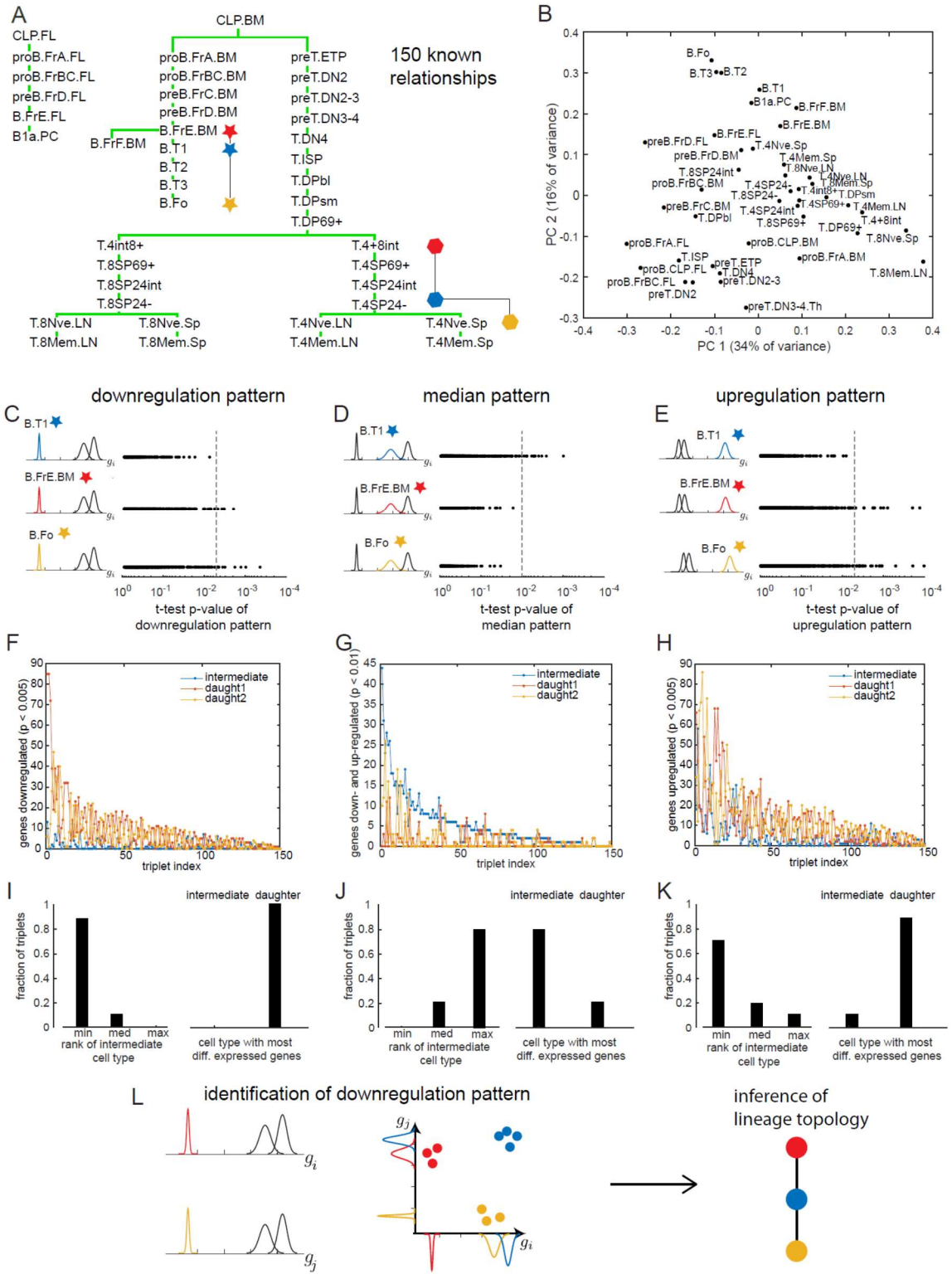


Figure 2.2: Lineage relationships in B- and T-cell development are correlated with 1-dimensional gene expression patterns. (A). Known lineage relationships between 42 cell types in B- and T-cell development.

(Figure 2.2, continued) Two examples of triplets of cell types are shown, with the intermediate cell type indicated in blue. (B) Projection of cell types along first two principal components of their gene expression data (PC1: 34%; PC2: 16%). (C)-(E) Genes showing the downregulation, median, or upregulation patterns in triplet (B.T1, B.FrE.BM, B.Fo). Each gene is represented by a dot. For each pattern (downregulation, median, and upregulation), the gene's p-value for the pattern is plotted on the x-axis, and the cell type featured by the pattern (e.g. gene downregulated in B.T1) is plotted on the y-axis. The p-value threshold is indicated by the gray dotted line.(F)-(H) For each pattern, number of genes showing the pattern in the three cell types of each triplet. The number of genes in the intermediate cell type of each triplet is indicated in blue; the number of genes in the two daughters is represented in yellow and red. For each pattern, the triplets are ordered according to the total number of genes showing the pattern in the triple.(I)-(K) Statistics of the relative number of genes showing the pattern in each triplet, for triplets with 10 or more genes. Left plot: fraction of triplets in which the intermediate cell type has the minimum, median or maximum number of genes showing the pattern, compared to the two daughter cell types. Right plot: fraction of triplets in which the cell type with the most genes showing the pattern is the intermediate or daughter cell type.(L) Inference strategy: identification of genes showing the downregulation pattern leads to a vote against the cell type in which they are downregulated, allowing for inference of lineage relationship topology.

In order to evaluate the predictive power of each pattern, we determined for each triplet the number of genes showing the pattern in each of the cell types (Figure 2.2C-E). The total number of genes showing a downregulation, median, or upregulation pattern in a given triplet varied between 0 and 118. For each of the three patterns, we considered the triplets with more than 10 genes showing the pattern, and we asked whether the relative number of genes showing the pattern in the different cell types of a triplet was correlated with the lineage relationship (Figure 2.2F-H).

Among the 84 triplets with more than 10 genes showing a downregulation pattern, the number of genes downregulated in the intermediate cell type was much smaller than the number of genes downregulated in either of the two daughter cells (mean 1.3 compared to 7.5). In 87% of these triplets, the intermediate cell type had the fewest significantly downregulated genes; in the remaining 13% of triplets, the intermediate cell type had the median number of downregulated genes (Figure 2.2I, left). Importantly, in each triplet, the

cell type with the most downregulated genes was always one of the daughter cell types (Figure 2.2I, right).

The other two patterns showed weaker correlation with lineage relationships. For the median pattern, the intermediate cell type had the highest number of genes that were significantly up- and down-regulated in 78% of triplets, but a daughter cell type had the most such genes in 22% of triplets (Figure 2.2J). For the upregulation pattern, the intermediate cell type had the fewest upregulated genes in 70% of triplets, but also the most such genes in 10% of triplets (Figure 2.2K).

Of the three patterns that we evaluated, we identified the downregulation pattern as having the strongest correlation with lineage relationships: in particular, our results show that in a given triplet, the cell types with the most downregulated genes are most likely not the intermediate cell type. The gene expression pattern that we observed in known lineage relationships suggested a strategy to discover unknown lineage relationships from gene expression data by identifying genes with significant downregulation (Figure 2.2L). We next developed a Bayesian framework to determine significantly downregulated genes and to sum up the contributions of different genes in order to find the most likely cell lineage topology.

2.2.2. Bayesian statistical approach to infer cell states and state transitions

The pattern discovered in the previous section suggests a method for lineage inference resting on two assumptions:

1. Lineage relationships can be inferred from genes that have a clear minimum expression level in one of the three cell types.

2. Each gene that has a minimum expression level in a cell type contributes to the probability that this cell type is not the intermediate state.

We next developed a Bayesian probabilistic framework that reflects our confidence in whether or not these assumptions are satisfied. This framework, given gene expression data for a collection of n cell types, infers lineage relationships between cell types and finds the key sets of marker genes $\{\alpha_i\}$ defining each cluster and transition genes $\{\beta_i\}$ determining transitions between cell types. We assume that lineage relationships between cell types are characterized by "transition" genes $\{\beta_i\}$ showing the downregulation pattern in the daughter cell types as observed in our corpus of triplets; we also assume that certain marker genes $\{\alpha_i\}$ are expressed in one cell type and nowhere else (Figure 2.3A-B). Given this model, the framework that we developed finds the most likely lineage tree and sets of marker and transition genes given the gene expression data.

Since the transition gene pattern that our model is based upon is between 3 cell types, we first determined lineage relationships and key genes for all sets of 3 cell types, and then built a tree using the set of inferred relationships. We note that we seek to infer the *topology* of the relationship between the three cell types: given our lack of any temporal information, we will not distinguish between A being the progenitor of B and C and A being an intermediate cell type through which B makes a transition to C or vice versa. We present an outline of the underlying math below and the details in Chapter 5 Appendix: Mathematical derivation of Bayesian Framework.

Let a set of three cell types be A, B and C, with gene expression data $\{g_i^{A,B,C}\}$ for all genes ($i = 1$ to N) in cell types A, B and C including data from biological and technical replicates for each cell type. The term $g_i^{A,B,C}$ denotes the expression data for just gene i in

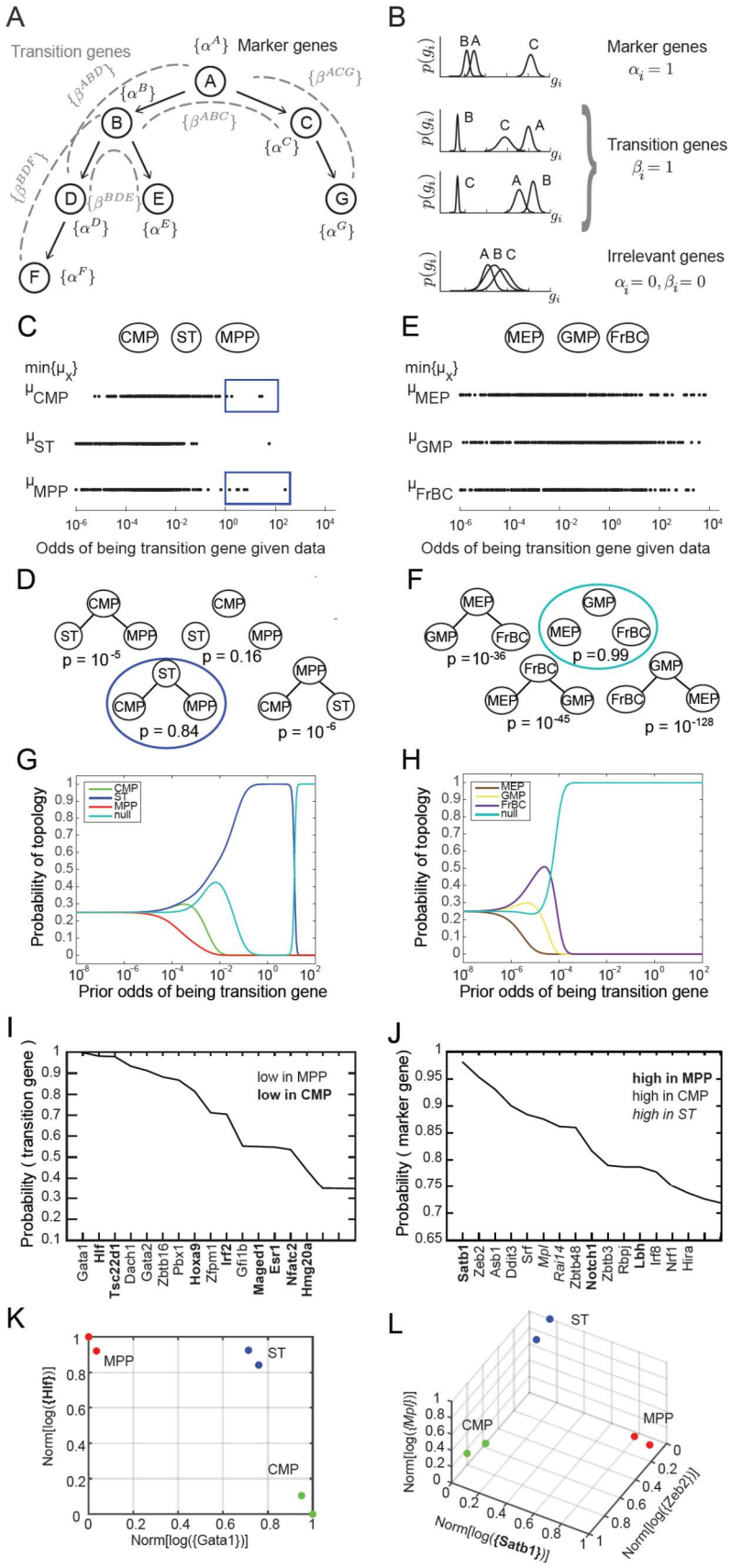
the three cell types. We would like to infer the topology T of the relationships between cell types A, B and C. We note that T can take on four possible values: $T = \mathcal{A}$: cell type A is the intermediate (either A is the progenitor of B and C, or A is an intermediate cell type between B and C); $T = \mathcal{B}$: cell type B is the intermediate; $T = \mathcal{C}$: cell type C is the intermediate; or $T = \emptyset$: we cannot determine the topology. We thus want to compute the probability of each of the possible topologies given the gene expression data, $p(T|\{g_i^{A,B,C}\})$.

For each gene i, we define a variable α_i equal to 1 if the gene is a marker gene and 0 if not, and a variable β_i equal to 1 if the gene is a transition gene and 0 not. Given the replicate measurements of the gene expression of gene i in cell types A, B and C, we can determine probabilistically whether each gene is a transition gene. The probability that a gene is a transition gene $p(\beta_i = 1|g_i^{A,B,C})$ will be closer to one if the gene has a unique minimum expression level in one cell type. For each gene, we can calculate the odds $\mathcal{O}_i$ that its expression level has a unique minimum in one cell type, $\mathcal{O}_i \equiv p(\beta_i = 1|g_i^{A,B,C})/p(\beta_i = 0|g_i^{A,B,C})$.

Each gene with high odds of being a transition gene tells us that the cell type in which its expression level is a minimum cannot be the intermediate state. In our probabilistic framework, each transition gene casts a vote $p(T \mid g_i^{A,B,C}, \beta_i = 1)$ against the topology in which the cell type with the minimum expression level of that gene is intermediate, and thus equally for the two other topologies in which it is not. To determine the probability of a topology, we next must add up the votes from all the genes, weighted by the probability that they are transition genes.

Figure 2.3: Identification of lineage topology and gene sets for 3 cell types. (A) Model of lineage tree underlying inference framework. Lineage relationships between cell types are characterized by "transition" genes $\{\beta_i\}$ showing a downregulation pattern in the daughter cell types as observed in our corpus of triplets; marker genes $\{\alpha_i\}$ are expressed in one cell type and nowhere else. (B) Gene expression patterns of marker genes, transition genes, and irrelevant genes in cell types A, B and C. Marker genes are highly expressed in only one cluster, whereas transition genes are highly expressed in two clusters and downregulated in the third. High probability transition genes alone are used for the determination of set of transitions; both high probability transition and marker genes are used for re-clustering. (C) Dot plot of the cell type that is most likely to have the minimum mean expression of each gene among CMP, ST and MPP as a function of the odds O_i of that gene being a transition gene. Each gene votes against the topology whose central node has the minimum mean among the three cell types, and this vote is weighted by the odds that the gene is a transition gene (Equation 3). Two groups of genes (boxed) are most likely to be transition genes and thus cast a strong vote against CMP or MPP being the intermediate cell type. (E) Computed probability of the topology given gene expression data indicates .84 probability that ST is the intermediate type. (F) Dot plot for triplet MEP/GMP/FrBC of the cell type that is most likely to have the minimum mean expression as a function of the odds O_i of that gene being an asymmetric gene. (G) Computed probability of the topology given gene expression data is the null hypothesis (.99). (H)-(I) Plot of the probabilities of the topologies given the data, $p\left(T\middle|\left\{g_i^{A,B,C}\right\}\right)$ as a function of the prior odds of genes being transition genes, $p(\beta_i = 1)/p(\beta_i = 0)$, for triplets CMP/ST/MPP and MEP/GMP/FrBC. For both triplets, the dominant topology is not affected by the choice of prior odds over a large range of values. (J) – (K) Replicates of cell types CMP, ST and MPP (dots colored based on cluster identity) in the gene expression space defined by transition and marker genes (probability > 0.8) associated with triplet CMP/ST/MPP. Axes represent the normalized log expression values of each class of transition genes (J), and marker genes for the three cell types (K). The most likely gene of each class is represented in curly brackets.

(Figure 2.3, continued)



2.2.2.1 *Summing up votes from different genes*

In order to obtain the probability of the topology given gene expression for all genes, $p\left(T \mid \{g_i^{A,B,C}\}\right)$, we must sum up each gene's vote $p\left(T \mid g_i^{A,B,C}, \beta_i = 1\right)$. Using a Bayesian framework, we derive this sum as (Methods):

$$p\left(T|\{g_i^{A,B,C}\}\right) \propto p(T) \prod_i \left(1 + \frac{1}{p(T)} \mathcal{O}_i \times p\left(T|g_i^{A,B,C}, \beta_i = 1\right)\right), \quad (1)$$

where $p(T)$ is the prior probability of T. Thus each gene's contribution $p\left(T \mid g_i^{A,B,C}, \beta_i = 1\right)$ to the probability of the topology given total gene expression $p\left(T \mid \{g_i^{A,B,C}\}\right)$ is weighted by the odds $\mathcal{O}_i$ that it is a transition gene.

The joint probability of the topology and the marker and transition genes is then:

$$p\left(T, \{\alpha_i\}, \{\beta_i\} \mid \{g_i^{A,B,C}\}\right) = p\left(T|\{g_i^{A,B,C}\}\right) \prod_i p\left(\alpha_i, \beta_i | T, g_i^{A,B,C}\right), \quad (2)$$

where $p\left(\alpha_i, \beta_i | T, g_i^{A,B,C}\right)$ is the joint probability of gene i being a transition gene or marker gene for topology T given the gene expression data.

Equation (1) can be rewritten as:

$$p\left(T|\{g_i^{A,B,C}\}\right) \propto p(T) \prod_i \left(1 + \frac{3}{2} \mathcal{O}_i \left[1 - p\left(\mu_T^i \text{is min} | g_i^{A,B,C}, \beta_i = 1\right)\right]\right), \quad (3)$$

where, for $T = \mathcal{A}$, $p\left(\mu_A^i \text{is min} | g_i^{A,B,C}, \beta_i = 1\right)$ is the probability that the mean μ_A^i of the distribution of gene i in cell type A is less than those of cell types B and C. Every gene casts a vote $-p\left(\mu_T^i \text{is min} | g_i^{A,B,C}, \beta_i = 1\right)$ against cell type T being the progenitor, and this vote is weighted by the odds O_i of the gene i being a transition gene and having a unique minimum. Genes that are in fact expressed in only one of the cell types automatically get a low vote since their confidence to tell which of the three cell types has minimum gene expression is small. Similarly, genes expressed in all three cell types at a comparable level

have a low vote. On the other hand, transition genes get a high vote since these genes can determine with much higher confidence the cell type in which the expression level is the minimum.

2.2.2.2 Computing the terms in equations (1)-(3)

The weights of the votes of each gene and the sum of these votes arise naturally in our Bayesian framework, allowing us to calculate the probability of the topology given all gene expression data. The odds O_i of a gene being a transition gene, the probabilistic vote $p\left(T\middle|g_i^{A,B,C},\beta_i=1\right)$ it casts for different topologies, and its probability $p\left(\alpha_i,\beta_i\middle|T,g_i^{A,B,C}\right)$ of being a marker or transition gene given the topology can all be calculated using this framework. They depend on the likelihood of the data $p\left(g_i^{A,B,C}\middle|T,\alpha_i,\beta_i\right)$ given the topology and whether the gene is a transition or marker gene. Each of these probabilities $p\left(g_i^{A,B,C}\middle|T,\alpha_i,\beta_i\right)$ must be computed numerically by integrating over a prior probability distribution of the means and standard deviations of the distribution functions of gene expression in cell types A, B and C, with the constraints on which cell type has the minimum mean defining the domains of integration (Materials and Methods). The odds O_i of a gene being a transition gene is proportional to the prior odds $p(\beta_i=1)/p(\beta_i=0)$ in the absence of any data of a gene being a transition gene. The prior odds are the one free parameter in our model, and we vary this sparsity parameter to test the robustness of our results.

2.2.3. Application to hematopoietic gene expression data

We used our Bayesian framework to understand the lineage and gene network governing early hematopoietic differentiation. We considered 11 early hematopoietic

progenitors from the ImmGen Consortium microarray data set (Heng and Painter, 2008), and focused our analysis on transcription factor gene expression. Given only the gene expression data for these different subpopulations of cells, we determined the lineage relationships and the key factors associated with each lineage decision. We calculated the probabilities of topology and marker and transition genes for the $\binom{11}{3} = 165$ possible triplets of cell types using equations (1) and (2). To illustrate our method, we first describe the analysis of the expression data from two such triplets of cell types: CMP/ST/MPP and MEP/GMP/FrBC (Figure 2.3). We then assembled the triplets to form an undirected lineage tree (Figure 2.4).

We computed the most likely topology and the identity of the transition genes simultaneously by maximizing $p\left(T, \{\alpha_i\}, \{\beta_i\} \middle| \{g_i^{A,B,C}\}\right)$ determined by Equation (2). Following Equation (3), each gene votes against the topology whose central node has the minimum expression of that gene among the three cell types, and this vote is weighted by the odds that the gene is a transition gene. To illustrate this for the triplet of cell types CMP, MEP and GMP, we plotted the topology each gene voted most against, *i.e.* the topology T for which $p\left(\mu_T^i \text{is min} \middle| g_i^{CMP,ST,MPP}, \beta_i = 1\right)$ is the maximum, versus the odds $\mathcal{O}_i$ of that gene being a transition gene (Figure 2.3C).

We find two groups of genes that are much more likely to be transition genes than any of the other genes, with values of $O_i \sim 10^2$ compared to 10^0 at most for other genes (Figure 2.3C, blue boxes). These two groups of genes have a large value for either $p\left(\mu_{CMP}^i \text{is min} \middle| g_i^{CMP,ST,MPP}, \beta_i = 1\right)$ or $p\left(\mu_{MPP}^i \text{is min} \middle| g_i^{CMP,ST,MPP}, \beta_i = 1\right)$ and thus vote against $T = CMP$: cell type CMP is the intermediate or $T = MPP$: cell type MPP is the

intermediate. Together these genes that have a high odds of being transition genes appear to most support topology $T = ST \equiv CMP - ST - MPP$.

In fact, the intuition in Figure 2.3C is borne out in the calculation of $p\left(T\middle|\left\{g_i^{CMP,ST,MPP}\right\}\right)$. Using Equation (1) above and assuming prior odds $p(\alpha_i = 1)/p(\alpha_i = 0)$ in the absence of any data of a gene being a transition gene to be 0.05, we calculate that there is an 84% chance that the topology is ST (Figure 2.3D).

In contrast to the case of the triplet of cell types CMP/ST/MPP, for triplet MEP/GMP/FrBC, there are no genes that have a much higher likelihood of being transition genes than other genes, and the distributions of genes supporting different topologies are similar (Figure 2.3E). Thus the most likely topology calculated using equation (1) is the null hypothesis (99%), which is that transition genes, if they exist, do not have patterns that depend on the cellular topology (Figure 2.3F), in which case there is insufficient evidence to classify the triplet according to a particular non-null topology.

2.2.3.1 *Choice of prior odds does not affect the most likely topology*

The only free parameter in our calculation above is a sparsity parameter: the prior odds of gene *i* being a transition gene, $p(\beta_i = 1)/p(\beta_i = 0)$ required in equations (1)-(3). At one extreme, if $p(\beta_i = 1)/p(\beta_i = 0) \to 0$, then $p\left(T\middle|\left\{g_i^{A,B,C}\right\}\right) \to p(T)$: if we assume that none of the genes are transition genes, then knowing gene expression does not give us any new knowledge of the topology T, since only transition genes are informative about T. At the other extreme, if $p(\beta_i = 1)/p(\beta_i = 0) \to \infty$ then the null hypothesis dominates: if all genes are transition genes, then there will be negative votes against all topologies. We

computed the behavior of $p(T|\{g_i^{A,B,C}\})$ between these two limits to determine the sensitivity of our answer to $p(\beta_i = 1)/p(\beta_i = 0)$.

Figure 2.3G and Figure 2.3H shows the dependences of the probabilities $p(T|\{g_i^{A,B,C}\})$ on the prior odds for triplets CMP/ST/MPP and CLP/MEP/GMP for values of $p(\beta_i = 1)/p(\beta_i = 0)$ between 10^{-8} and 10^{2}. For triplet CMP/ST/MPP the topology $ST \equiv CMP - ST - MPP$ dominates for $p(\beta_i = 1)/p(\beta_i = 0)$ between 10^{-2} and 10, whereas for triplet MEP/GMP/FrBC there is no value of the prior odds that strongly favors a non-null topology. For most triplets, the most likely topology does not depend on the choice of prior odds; we ignore those triplets where different choices of prior odds lead to different most-likely topologies.

2.2.3.2 Transition and marker genes for the CMP/ST/MPP triplet

For each triplet, we evaluated each gene's probability of being a transition or marker gene. Figure 2.3I shows the names and associated probabilities of the 15 genes most likely to be transition genes for the triplet CMP – ST – MPP. The transition genes fall into two groups, corresponding to the two boxes in Figure 2.3C. One group, which includes genes *Gata1*, *Dach1*, and *Gata2*, has higher expression in CMP than in MPP; the other group, which includes *Hlf*, *Tsc22d1*, and *Hoxa9*, has higher expression in MPP. Although the values of the probabilities of the genes being transition genes vary with the value of the sparsity parameter, the relative order of different genes does not change. The genes identified include many genes previously identified as being important for lineage specification (Crispino, 2005; Gazit et al., 2013; Miyawaki et al., 2015). The transition genes we discovered thus not only have gene expression patterns that reflect the lineage decision but also include functionally important genes.

In addition to the transition genes, we identified marker genes present only in ST (including *Mpl* and *Rai14*, consistent with (Solar et al., 1998)) and then *symmetrically* downregulated in both CMP and MPP (Figure 2.3J, L). Marker genes for CMP include Srf, Zeb2, Rbpj and Irf8 (consistent with (Goossens et al., 2011; Kurotaki et al., 2013; Ragu et al., 2010; Robert-Moreno et al., 2005; Tamura et al., 2000)); marker genes for MPP include Satb1, consistent with (Satoh et al., 2013). Although these genes were not used to determine the topology, they are good markers for cell types ST and LT.

Plotting the cell types using the mean expression levels of the two transition gene classes captures the fork in the gene expression space associated with the cell-fate decision (Figure 2.3K). In contrast with the PCA analysis of the cell types (Figure 2.1A), in which MPP appears to be an intermediate between the hematopoietic stem cell types (LT and ST) and CMP, the projection of the cell types onto the transition-gene subspace cleanly shows that ST splits into CMP and MPP.

2.2.4. A lineage tree for early hematopoiesis

We next constructed both the lineage tree for the cell types and a gene regulatory network. Out of the 165 possible triplets of hematopoietic progenitors, 144 showed one single non-null topology with probability greater than 0.6 over a large range of values of the prior odds.

To illustrate the construction of the tree from individual triplets, consider first the gene expression data for four cell-types: long-term hematopoietic stem cell (LT), short-term hematopoietic stem cell (ST), multipotent progenitor (MPP) and common myeloid progenitor (CMP) (Heng and Painter, 2008). We first infer the topology of the $\binom{4}{3} = 4$

triplets involving those cell types. For each triplet, we found a non-null topology that had high probability over a large range of prior odds (Equation (1), Figure 2.4A-D). We assembled these triplet topologies into a tree that recapitulates each of the local topologies. Although triplet CMP/LT/MPP has topology CMP – **LT** – MPP (Figure 2.4D), we could determine that LT cannot be the *direct* progenitor of CMP or MPP, because ST is an intermediate between LT and both cell types (Figure 2.4A, C). We can thus "prune" this triplet to infer the local lineage relationships between the four cell types (Figure 2.4E). The inferred tree between the four cell types is undirected; however, it is interesting to note that triplet CMP – **LT** – MPP, although not necessary for the construction of the tree, gives a hint of directionality, suggesting that LT is an ancestor of CMP and MPP, not an intermediate.

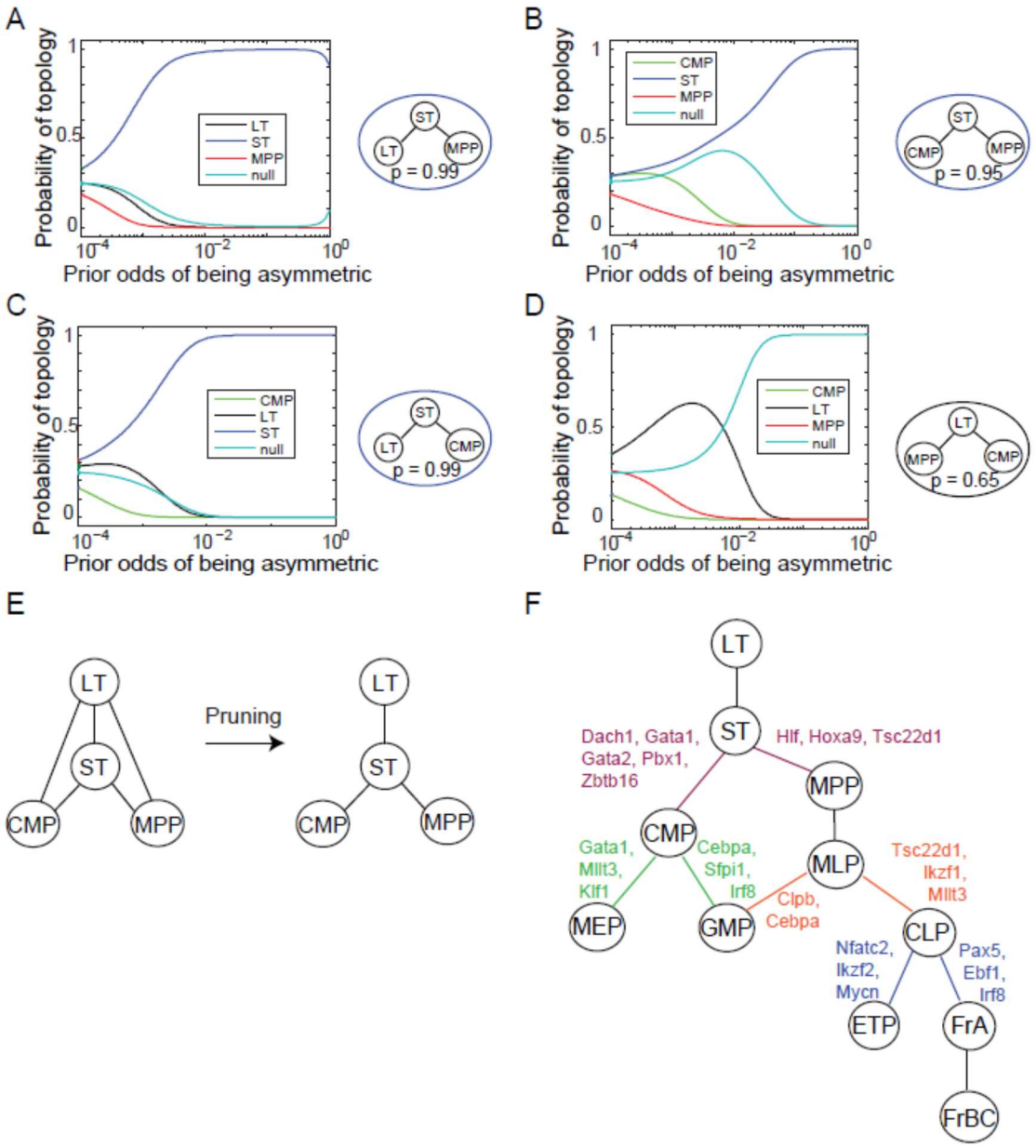


Figure 2.4: Determination of lineage tree and transition and marker genes given gene expression for early hematopoiesis. (A)-(D) Determination of the topologies of the four triplets involving cell types LT, ST, MPP and CMP: LT/ST/MPP (A), CMP/ST/MPP (B), LT/ST/CMP (C), and MPP/LT/CMP (D). Left: Plots of the probabilities of the topologies given the data, $p\left(T\middle|\left\{g_i^{A,B,C}\right\}\right)$ as a function of the prior odds of genes belonging to the asymmetric class, $p(\alpha_i = 1)/p(\alpha_i = 0)$, for the four triplets. Right: For each triplet, there is a non-null topology that has high probability over a large range of prior odds (probabilities indicate the maximum probability of the topology over the range of prior odds).

(Figure 2.4, continued) (E) Schematic of the pruning rule used to assemble the triplet topologies into a tree that recapitulates each of the local topologies. Although triplet CMP/LT/MPP has topology CMP – **LT** – MPP (Figure 2.4D), LT cannot be the *direct* progenitor of CMP or MPP, because ST is an intermediate between LT and both cell types (Figure 2.4A, C). Left: Superposition of all four topologies, before using the pruning rule. Right: Resulting tree after using the pruning rule, eliminating the connections from LT to CMP and MPP. (F) Final lineage tree, with key inferred transition genes indicated along cell fate decisions.

The ST/CMP/MPP triplet immediately allows us to distinguish between two competing models regarding the hierarchy of early hematopoietic progenitors. According to the traditional picture (Iwasaki and Akashi, 2007), MPP is the progenitor of CMP, and ST is the progenitor of MPP – therefore MPP should be an intermediate between ST and CMP and the topology of triplet ST/MPP/CMP should be ST – **MPP** – CMP (Figure 2.1D, left). According to a model suggested by Adolfsson and colleagues (Adolfsson et al., 2005), ST splits into CMP and MPP (Figure 2.1D, right), and the topology should be CMP – **ST** – MPP. We identify both CMP – **ST** – MPP and CMP – **LT** – MPP as the correct topologies, lending support to the Adolfsson model.

We determined the topologies of the remaining triplets of cell types, and we constructed the full lineage tree by assembling and pruning the triplets we identified as connected. For each triplet of cell types, we discovered marker and transition genes genes. The lineage tree that we determined is consistent with the Adolfsson model and contains three additional lineage decisions (Figure 2.4F). First, CMP gives rise to MEP (megakaryocyte/erythroid progenitor) and GMP (granulocyte/macrophage progenitor). Second, MPP gives rise to MLP (multilineage progenitor), which then splits into the GMP and CLP (common lymphoid progenitor) cell types. The final lineage decision, in which CLP gives rise to the ETP (pre-T) and FrA (pre-pro-B) and FrBC (pro-B) cell types.

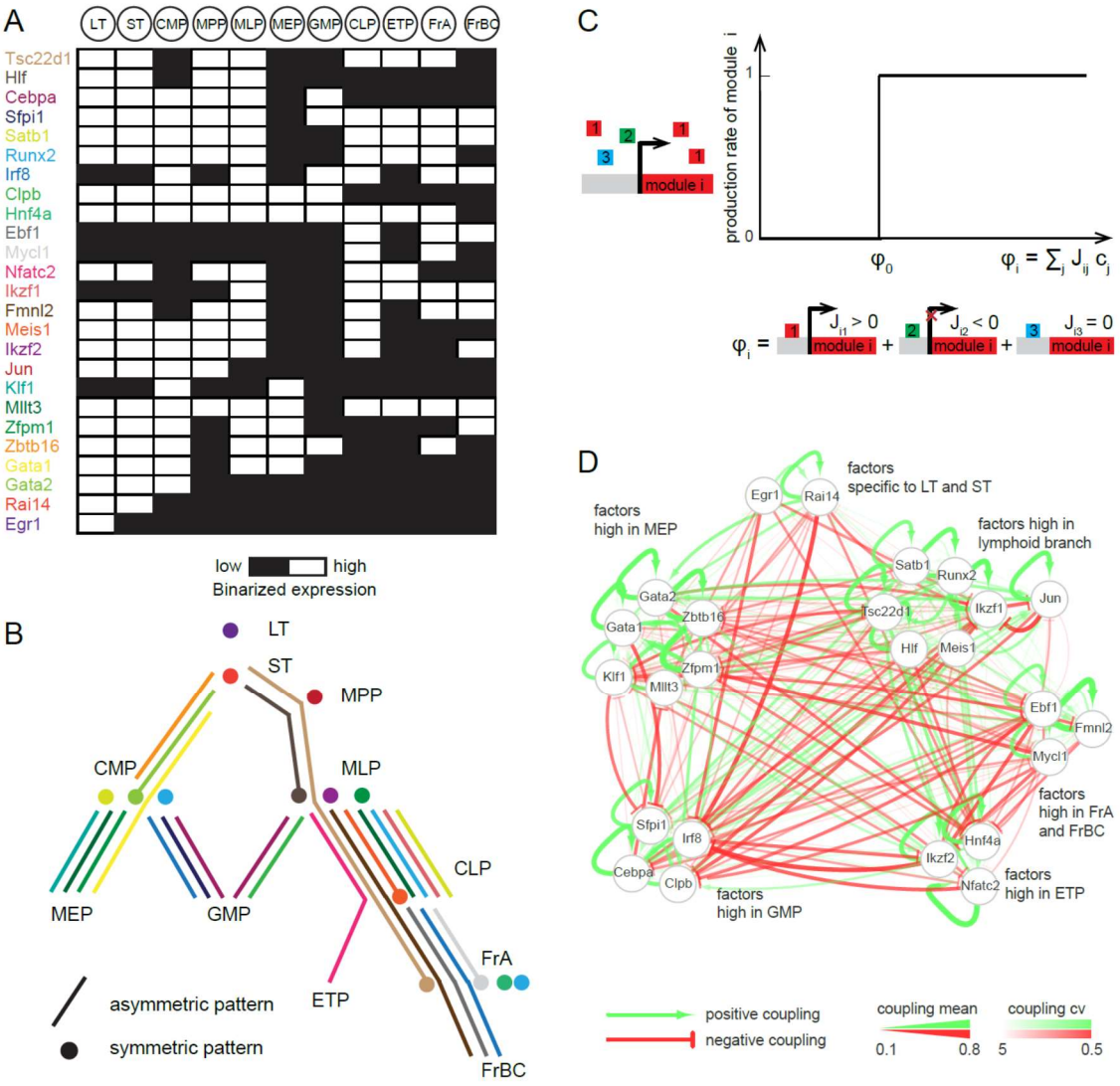
Many genes we discover as belonging with high probability to the transition and marker classes of genes at each lineage decision are known in the literature to be functionally important genes. Marker genes for LT include *Egr1* and *Fus*, consistent with (Sugawara et al., 2010) and (Min et al., 2008)). For the CMP/MEP/GMP decision, markers for CMP include *Pbx1* and *Mycn*, known to be important for renewal of the pluripotent state (Ficara et al., 2013; Laurenti et al., 2008), transition genes maintained in MEP include *Gata1* and *Mllt3*, involved in erythroid and megakaryocytic cell fate (Pina et al., 2008; Zhang et al., 1999), and transition genes maintained in GMP include known myeloid regulators *Sfpi1*, *Cebpa*, and *Gfi1* (Koschmieder et al., 2005; van der Meer et al., 2010; Radomska et al., 1998; Zhang et al., 1999). In the subsequent lineage decisions, we again rediscovered a number of known regulators, including lymphoid regulators *Satb1*, *Ikzf1* and *Notch1* (Rebollo and Schmitt, 2003; Satoh et al., 2013; Stier et al., 2002), and factors important for the B/T cell-fate decision such as *Ebf1, Pax5, Irf8*, *Nfatc2*, and *Runx2* (Busslinger, 2004; Macian, 2005; Vaillant et al., 2002; Wang et al., 2008).

At the different lineage decisions along the tree, we found that several genes are reused and belong to multiple gene classes as differentiation progresses. For example, the transition genes present in LT and ST and expressed CMP but downregulated in MPP exhibit three distinct behaviors in the lineage decision from CMP to MEP and GMP. First, some genes such as *Zbtb16* do not have differential expression in the CMP/MEP/GMP lineage decision. Second, some genes including *Pbx1* and *Gata2* are symmetrically downregulated in both MEP and GMP. Gene *Gata1*, on the other hand, is upregulated from CMP to MEP but downregulated in GMP. Thus, in cell type CMP, the set of transition genes from ST to CMP contains three new subclasses, one that does not have any role in

the next lineage decision, a second that is symmetric and potentially contributes to stabilizing CMP, and a third that serves as a transition gene at the next decision. We find a similar pattern of reuse in the other lineage decisions, with genes acting as transition genes or marker genes in multiple cell-fate decisions.

Our approach allowed us to categorize different genes within a particular cell-fate decision as different classes of transition or marker genes. Based on the behavior of genes in the different cell-fate decisions along the tree, we grouped genes into 25 modules (Methods), with genes in each module showing the same binarized expression pattern along the lineage tree (Figure 2.5A-B). We will denote each module by a representative gene, but most modules contain multiple genes (Figure 2.6A-B). By simultaneously inferring both the lineage tree and the gene regulatory network, we were thus able to discover a simple core gene network that gives rise to the lineage tree.

As we show next, inferring the lineage relationships and the underlying network can help us quantitatively understand the gene regulatory networks that govern differentiation.



E

Start state	End state	Overexpress module	Probability	Experimental validation	Reference
CMP, MPP	LT	Egr1	100%	Yes	Riddell et al., 2014
CMP	MPP	Hlf	48%	Yes	Riddell et al. ,2014
MPP	CMP	Gata2	47%	No	
CMP	CLP	Meis1	46%	No	
GMP, ETP, FrA	MEP	Gata1+Klf1	11%	Yes	Iwasaki and Akashi, 2007
MEP, ETP, FrA	GMP	Cebpa+Irf8+Clpb	13%	Yes	Iwasaki and Akashi, 2007
MEP, GMP, ETP	FrA/BC	Ebf1	45%	Yes	Pongubala et al., 2008
MEP, GMP, FrA	ETP	Nfatc2	33%	No	

Figure 2.5: Quantitative modeling of the core network underlying hematopoiesis. (A) Binary gene expression patterns across all cell types (columns) for the 25 inferred gene modules. Each gene module represents multiple genes that have similar behaviors along the lineage tree. (B) Gene expression patterns of the 25 modules in the 11 hematopoietic cell types (colors according to A).

(Figure 2.5, continued) Straight lines indicate asymmetric regulation favoring the colored branch; dots indicate symmetric downregulation in the subsequent two branches. (C) Plot of the production rate of module i, $r_i(\vec{m})$, as a function of the drive from the other modules $\phi_i(\vec{m}) = \sum_{j=1}^{N} J_{ij} m_j$. The production rate is equal to 1 if the drive is greater than a critical drive ϕ_0 and 0 otherwise. (D) Shared features of the inferred gene regulatory networks between the 25 modules. Green arrows represent couplings whose average values across 5000 sampled solutions are positive; red arrows represent couplings with negative mean values. Line thickness represents the mean strength of the coupling; transparency represents the coefficient of variation (c.v.) of the coupling. The modules are shown in groups that have similar interactions with other modules. Four groups of modules are characteristic of the four terminal cell types (MEP, GMP, FrA/BC and ETP) and have mutually inhibitory interactions. There is also a group of modules that are characteristic of the HSC cell types (LT and ST), and one that is characteristic of the lymphoid cell types (MPP and MLP). The progenitor cell types are stabilized through a combination of positive and negative interactions. (E) Table of reprogramming predictions. Shown are the start and end states of the prediction, the modules to overexpress, the probability of reaching the end state given starting in the starting state and overexpressing the given modules, and any prior experimental validation of the prediction.

2.2.5. Modeling the underlying network

The Bayesian analysis in the last section allowed us to obtain the probabilities of lineage relationships as well as the marker and transition genes for each triplet of cell types. Classifying genes based on their patterns of expression along the inferred lineage tree rather than by gene-gene correlations allowed us to identify 25 modules of genes with similar expression patterns in successive cell-fate decisions (Figure 2.5B). Here we describe a framework, distinct from the preceding method, for building a mathematical model to obtain a quantitative understanding of hematopoietic differentiation.

We assume that the expression of each module is driven by the expression of all other modules. We consider a network that contains only direct interactions, in which each module *j* exerts a drive on module *i* which is equal to an interaction strength J_{ij} (positive or negative) multiplied by the concentration of module *j*. The total drive on module *i* is the sum of the drives from the different modules. We further consider that the total drive on module *i* affects expression in a highly non-linear manner, with high gene expression for

drives that exceed a critical drive ϕ_0, and low gene expression otherwise (Figure 2.5C). Thus the effective dynamics of expression levels, m_i of each module i of genes is given by the non-linear equation:

$$\frac{dm_i}{dt} = H\left(\sum_j J_{ij} m_j - \phi_0\right) - \frac{m_i}{\tau},$$

where H is the Heaviside step function and τ_i is the effective lifetime of module i. The effective lifetime can be absorbed into the other parameters in the model through rescaling, and in steady state m_i will have high and low expression levels that can be rescaled to be 0 and 1 (Methods).

This model has to be constrained by our data. To do so, we make the assumption that experimentally discovered cell types must be stable over a certain period of time. Thus, we asked what set of interactions J_{ij} were consistent with the observed cell types (LT, ST, CMP, etc.), defined by the appropriate expression patterns of the modules in each cell type, being stable states of the network. If state $\vec{m}^\alpha = \{m_1^\alpha, \ldots, m_{25}^\alpha\}$ with expression level m_i^α in module i is a stable state of the network, then the interactions J_{ij} must be such that the total drive on each module that is expressed in $\vec{m}^\alpha$ is *greater* than the critical drive, and the total drive on each module that is not expressed in $\vec{m}^\alpha$ is *less* than the critical drive:

$$m_i^\alpha = 1 \Rightarrow \sum_j J_{ij} m_j^\alpha \geq \phi_0,$$

$$m_i^\alpha = 0 \Rightarrow \sum_j J_{ij} m_j^\alpha < \phi_0. \quad (4)$$

Thus, for each stable state, we have 25 constraints on the possible values of J_{ij}, one for each module. Given 11 cell types, there are 275 inequalities that constrain the values of the 625 different interaction strengths J_{ij}. Clearly, the problem is underdetermined, and

there is an infinite number of solutions that would allow for the observed cell types to be stable states. Faced with this massive parameter uncertainty, we noted recent research (Machta et al., 2013) that shows that when fitting high-dimensional models to data, some parameters or combinations of parameters, referred to a "sloppy" parameters, have little effect on model behavior, whereas a small number of parameter combinations, designated as "stiff", tend to be well-constrained and crucial for the model behavior. Inspired by this work, we sought to determine which interaction strengths in the network were the most important for producing the known cell types, and which interaction strengths could take on a wide variety of values and still produce these cell types.

We used a linear programming method with the constraints set by Equation (4) to generate a representative set of models and study their behaviors (Methods). We sampled 5,000 different solutions J_{ij} for which the known cell types are stable, and we explored any common features of these networks. We found that 79 couplings could be considered near-universally positive or negative (70% or more of the sampled couplings were of one sign or another, corresponding to coefficients of variation less than 2). The remaining 546 couplings could take a wide range of values and still produce the observed stable states. We were thus able to discover a core network between the different modules that is shared by the majority of solutions and is thus necessary to produce the observed cell types (Figure 2.5D).

The inferred functional network (Figure 2.5D) shows certain motifs common across cell fate decisions. First, in all 5000 solutions, the parameters for interaction strengths are tuned such that the cell-fate decisions along the tree are controlled by mutually-inhibitory interactions between modules that serve as transition genes along the decision. For example,

in a majority of models, mutual inhibition between modules Zbtb16 and Tsc22d1 is responsible for the decision from ST to CMP and MPP. Second, activatory interactions stabilize the mutually inhibitory interactions in the progenitor. For example, modules Zbtb16 and Tsc22d1 are present together in cell types LT and ST; their mutual inhibition is stabilized through indirect activation of both factors by modules Egr1 and Rai14 through Gata2 and Hlf respectively.

Similarly, the decision of CMP cells to differentiate to MEP or GMP is controlled by the strong mutual inhibition between modules Gata1, Klf1 and Mllt3 on the one hand and modules Sfpi1 and Cebpa on the other hand, and these modules are stabilized in CMP through activation by Gata2 and Satb1. The transition from MPP to MLP is mediated by the mutual inhibition between module Jun, which is downregulated from MPP to MLP, and modules Ikzf1 and Irf8, which are upregulated from MPP to MLP. The decision from MLP to GMP and CLP is controlled by the mutual inhibition between Cebpa and Clpb, and Ikzf1 and Ebf1; this inhibition is stabilized in MLP by modules Sfpi1 and Ir8. Finally, the FrA/ETP split is controlled by the mutual inhibition between Nfatc2 and Irf8, and the mutual inhibition between Ebf1 and both Hnf4a and Ikzf2 is further responsible for the split between ETP and FrBC. These modules are stabilized in CLP by Ikzf1 and Meis1.

In sum, we find that the network contains 4 groups of modules each of which is characteristic of one of the 4 terminal cell types (MEP, GMP, FrBC and ETP). These four groups have mutually inhibitory interactions, and the progenitor cell types (LT, ST, CMP, MPP, MLP and CLP) are stabilized through positive interactions with symmetrically-downregulated modules.

2.2.5.1 Reprogramming predictions

We next used the sampled solutions to make reprogramming predictions. Given the set of possible coupling constants that satisfy the fixed-point constraints $\{J_{ij}\}$, we modeled the dynamics of moving between the stable states of the network. Given any two stable states $\vec{m}^{\alpha}$ and $\vec{m}^{\beta}$, we calculated the probability of moving from $\vec{m}^{\alpha}$ to $\vec{m}^{\beta}$ when overexpressing some set of modules (Figure 2.5E, Methods). Our model identified modules whose overexpression could achieve transdifferentiation between the terminal cell types of the tree. Thus reprogramming to MEP can be achieved by overexpressing members of modules Gata1 and Klf, consistent with reprogramming experiments involving *Gata1* (Iwasaki and Akashi, 2007). Similarly, transdifferentiation to GMP can be achieved through overexpression of modules Cebpa, Irf8 and Clpb, consistent with experimentally observed *Cebpa* or *Sfpi1* overexpression (Iwasaki and Akashi, 2007), and transdifferentiation to ETP and FrA/BC is successful through overexpression of modules Nfatc2 and Ebf1, respectively. Transdifferentiation to B-like cells has been achieved using overexpression of *Ebf1* (Pongubala et al., 2008) but has not been successfully attempted for T-like cells.

Our model also predicts modules for reprogramming to the pluripotent and early multipotent progenitors. For example, reprogramming to LT-HSC, MPP, CMP and CLP from the more terminal cell types can be achieved through overexpression of members of modules Egr1, Hlf and Ikzf2, Gata2 and Meis1 respectively. Several members of these modules have been successfully employed in attempts to convert differentiated cells towards hematopoietic stem cells, including genes *Mycn*, *Hlf*, *Lmo2*, *Meis1* and *Pbx1* (Riddell et al., 2014).

In conclusion, we have shown that we can categorize the relevant genes into different modules, and these modules can be the starting point for successfully modeling the core differentiation network responsible for the cell-fate decisions along the tree and making transdifferentiation and reprogramming predictions.

2.3. Discussion

Gene expression patterns during the course of differentiation give us snapshots of the dynamics of the molecular network that leads progenitor cells to more differentiated states. Not all genes give us equal information about the sequence of lineage decisions pluripotent cells make, and in fact, the expression levels of many genes corrupt this information. We first studied known lineage relationships and found that one-dimensional downregulation patterns in individual genes are correlated with lineage relationships. We built a probabilistic framework based on the lineage-decision expression patterns we observed. This framework allowed us to weigh the data from each gene while employing a sparsity parameter, the prior probability for any gene to be a transition gene. When we fail to detect patterns that support any lineage relationship, our probabilistic approach allows us to declare failure in finding the correct topology and the genes. Using this framework, we identified both lineage relationships and known master factors for the hematopoietic system, including *Gata1*, *Cebpa*, *Sfpi1*, *Ebf1* and *Pax5* (Orkin and Zon, 2008). By using the patterns of gene expression along the inferred lineage tree we build models, sampled 5,000 possible combinations of parameters and found the key interactions of networks that lead to hematopoiesis. We were thus able to uncover simple core networks involving a small number of key transcription factors, in agreement with the traditional paradigm in the study of hematopoiesis (Iwasaki and Akashi, 2007; Orkin and Zon, 2008),

but in contrast with recent computational studies arguing for the role of complex networks in hematopoiesis (Basso et al., 2005; Novershtern et al., 2011). Using our models, we can also predict factors that can reprogram one cell type into another, with many of our findings consistent with recent literature.

There are many examples of genes known to be functionally important for lineage decisions whose expression patterns fit the downregulation pattern that we observed. In the case of patterning involving lateral inhibition, progenitor cells express genes together (for example, *Notch* and *Delta*) which are differentially expressed in the differentiated states (only *Notch* or only *Delta*) (Perrimon et al., 2012). The same pattern is also seen in multiple examples of lineage decisions often involving mutual inhibition, where key genes expressed in the progenitor are differentially regulated in the progeny (Graf and Enver, 2009; Qi et al., 2013; Thomson et al., 2011; Zhang et al., 1999). In these latter examples, the proteins that integrate external signals to decide whether progenitor cell type A differentiates into cell types B or C are (a) present in A before the decision is made and (b) not only favor one of the two differentiated cell fates but also inhibit the progenitor cell from choosing the other fate, so their levels must be downregulated in the differentiated cell type whose fate they inhibit. In each of these examples, the genes show a clear minimum expression level in one of the three cell states, and the cell type in which their expression is a minimum is not the progenitor. Given the prevalence of the downregulation pattern among known functional genes, it is perhaps not surprising that the transition genes that we discovered based on their gene expression pattern are good candidates for having a functional role, and indeed include known regulators of hematopoiesis.

Our observation that downregulation patterns are correlated with lineage relationships in B- and T-cell development suggests that this pattern, seen anecdotally in the context of important functional regulators of cell state, may be one that can be observed more generally in a larger number of transcription factors and wide variety of lineage transitions, including lineage progressions.

The transitions of multipotent cells from one cell state to another are controlled by a network involving a very large number of molecular factors and interactions. One approach to studying complex biological networks is to carefully measure every variable and rate constant in the underlying network and then build mathematical models for the different interactions (Karr et al., 2012). In striking contrast, state transitions of complex physical systems have been studied successfully by extracting, often using symmetry principles, reduced sets of key variables known as order parameters. Such approaches have led to a deep understanding of a diverse set of phenomena including complex metallic alloys transitioning to a superconducting state and the localization of electrons in disordered solids to the acquisition of magnetism in solids (Abrikosov, 2004; Anderson, 1978; Landau and Lifshitz, 1951).

The two classes of transition genes at each lineage decision best describe the cell state transition: in the two-dimensional space of these two classes of transition genes, the progenitor cells must trace a fork as they transition to the two differentiated cell types (Figure 2.2L and Figure 2.3K). The two classes of transition genes thus constitute the relevant order parameters for the cell fate decision. Monitoring the levels of these genes in real time in single cells should capture the dynamics of transitions of individual cells at a lineage decision, thus allowing us to understand how individual cells make developmental

decisions, whether their transitions from one state to another are probabilistic, and how different cells coordinate their decisions. Similar to the dimension-reduction approaches for the analysis of protein-folding trajectories to obtain correct reaction coordinates (Best and Hummer, 2005), our framework leads us to the transition genes as the 'reaction coordinates' for understanding lineage decisions.

Single-cell data being generated for different developmental systems (Klein et al., 2015; Macosko et al., 2015b) gives us glimpses into how single cells make lineage decisions. Our finding that distance measurements using all genes are misleading implies that the analysis of this data might require more than straightforward clustering (Jaitin et al., 2014; Zeisel et al., 2015). A probabilistic approach that simultaneously determines the cluster identity of each cell, the lineage relationships between cell types and the underlying networks, rather than treat each problem independently, will be needed to understand single-cell transcriptomics data. Combining such statistical approaches with phenomenological modeling of developmental signaling networks (Corson and Siggia, 2012) will likely help us understand the networks that control the decisions of individual cells.

2.4. Methods

2.4.1. Gene Expression Data

Hematopoietic gene expression data was downloaded from the Immunological Genome Project (Heng et al., 2008; GEO GSE15907) and log-2 transformed. We restricted the genes considered to 1,459 transcription factors.

Table 2.1 Hematopoietic Cell Types Considered. Listed for each cell type considered in this paper are the Immunological Genome Project descriptor for the cell type, its common name and phenotype, its age and location, and the number of replicates in the data set.

Abbreviation in text	Immgen Descriptor	Long name	Phenotype	Age	Location	# of replicates
LT	SC.LT34F.BM	Long-Term reconstituting Stem Cell (LT-HSC)	CD34- Flk2- Lin- ckit+ Sca1+	8w	Bone marrow	3
ST	SC.ST34F.BM	Multipotent Progenitor (ST-HSC)	CD34+ Flk2- Lin- ckit+ Sca1+	8w	Bone marrow	2
MPP	SC.MPP34F.BM	Multipotent Progenitor (MPP)	CD34+ Flk2+ Lin- ckit+ Sca1+	8w	Bone marrow	2
MLP	MLP.BM	Multilineage Progenitor	Lin- AA4+ Kit++ IL7Ra- B220-	6w	Bone marrow	4
CMP	SC.CMP.BM	Common Myeloid Progenitor	Lin- IL7R- Sca1- ckit+ FcgRloCD34+	8w	Bone marrow	2
MEP	SC.MEP.BM	Megakaryocyte-Erythroid Progenitor	Lin- IL7R- Sca1- ckit+ FcgRloCD34-	8w	Bone marrow	2
GMP	SC.GMP.BM	Granulocyte-Monocyte Progenitor	Lin- IL7R- Sca1- ckit+ FcgRhiCD34+	8w	Bone marrow	3
CLP	proB.CLP.BM	Common Lymphoid Progenitor	Lin- AA4+ Kit+ IL7Ra+ B220-	10w	Bone marrow	4
FrA	proB.FrA.BM	Fr. A (pre-pro-B)	Lin- AA4+ Kit+ IL7Ra+ B220+	10w	Bone marrow	4

(Table 2.1, continued)

FrBC	proB.FrBC.BM	Fr. B/C (pro-B)	Lin- AA4+IgM-CD19+CD43+HSA+	10w	Bone marrow	3
ETP	preT.ETP.Th	Early T lineage Precursor	4- 8- 11b- 11c-19- NK1.1- TCR-44hi 117hi 25-	6w	Thymus	3

2.4.2. Software

Calculations were performed using custom written MATLAB code (The Mathworks) on the Harvard Research Computing Odyssey cluster. Module Networks software (Segal et al., 2003) was downloaded as part of the Genomica suite (http://genomica.weizmann.ac.il/) and was run using default parameters.

2.4.3. Membership in the transition and marker gene classes

Membership in classes G1-G15 was calculated by assuming a prior of 5×10^{-2} for the hematopoietic data. The threshold for membership was $p(\alpha_i|g_i, T) > 0.8$ or $p(\beta_i|g_i, T) > 0.8$.

2.4.4. Determination of gene modules

A total of 265 genes belong to any of the marker and transition gene classes. We partitioned the transcription factors into a total of 41 subclasses, each of which is a unique intersection or set difference of the marker and transition gene classes. We determined binary gene expression profiles by calculating the mean log2 fold-change in expression level for each subclass. Some subclasses had identical binary gene expression profiles: for

example, gene Mpl is present in ST, downregulated symmetrically in CMP and MPP, and then further downregulated asymmetrically from CMP to GMP. Such a gene expression pattern would require 3 distinct levels (high in ST, medium in CMP, and low in GMP); instead, our binary representation of Mpl matches that for Rai14: on in LT and ST and off in all other cell types.

We grouped subclasses with identical binary expression profiles together, leaving us with a total of 25 modules with unique binary gene expression profiles. We denote each module by a representative gene; the genes that belong to each module are shown in Table 2.2. The binary profiles for each module are shown in Figure 2.6A, and the number of genes per module is shown in Figure 2.6B.

Table 2.2: Composition of the 25 Gene Modules.

Module name	Module members								
Tsc22d1	*Hoxa9*	*Tsc22d1*							
Hlf	*Hlf*								
Cebpa	*Cebpa*	*Gfi1*							
Sfpi1	*Bmyc*	*Hcls1*	*Hdac5*	*Lass5*	*Lass6*	*Lbh*	*Sfpi1*	*Zscan2*	
Satb1	*Aff3*	*Arid5b*	*Elk3*	*Erg*	*Esr1*	*Etv6*	*Hhex*	*Mef2c*	*Ncoa3*
	Rel	*Satb1*	*Smad5*	*Sry*	*Stat4*	*Zbtb20*	*Zfp238*		
Runx2	*Runx2*	*Smarca2*							
Irf8	*Irf8*								
Clpb	*Clpb*	*Men1*	*Nfe2*	*Nfkbib*	*Pdlim1*	*Polr2l*	*Psmd9*	*Ruvbl2*	*Rxrb*
	Tcfec	*Tysnd1*							
Hnf4a	*Ctbp2*	*Hnf4a*	*Nfil3*	*Pbx4*	*Prdm5*				
Ebf1	*Ddx54*	*E2f2*	*Ebf1*	*Pax5*					
Mycl1	*Mycl1*								
Nfatc2	*Nfatc2*	*Nr1d2*							
Ikzf1	*672048 9N17Rik*	*Adnp*	*Ankrd6*	*Arhgap17*	*Arid1a*	*Atf2*	*Atf6*	*Atf7ip*	
	Bach2	*Baz2a*	*Bbx*	*Brd8*	*Carf*	*Clock*	*Crebbp*	*Ctcf*	
	Ddx58	*Dido1*	*Elf1*	*Elk4*	*Ep300*	*Ets1*	*Ezh1*	*Foxj3*	
	Foxp1	*Grlf1*	*Helz*	*Hipk2*	*Hivep1*	*Hmbox1*	*Hmg20a*	*Huwe1*	
	Ikzf1	*Irf1*	*Irf2*	*Irf7*	*Klf6*	*Klf7*	*Lcor*	*Mllt10*	
	Myst4	*Nab1*	*Ncoa1*	*Ncoa6*	*Ncor1*	*Nfatc1*	*Nfatc3*	*Notch1*	
	Nsd1	*Otud4*	*Papolg*	*Pbrm1*	*Pdcd11*	*Phf21a*	*Pias2*	*Pogk*	
	Pou2f1	*Ppp1r16b*	*Rara*	*Rbl1*	*Rfc1*	*Rfx3*	*Rnf38*	*Runx1*	

(Table 2.2, continued)

	Runx3	*Sertad2*	*Sin3a*	*Smad4*	*Smad7*	*Smarca5*	*Sox4*	*Sp1*	
	Stag1	*Stat1*	*Suz12*	*Tcerg1*	*Tcf12*	*Tcf20*	*Tcf7l2*	*Tial1*	
	Trim24	*Trim30*	*Trp53bp1*	*Trps1*	*Ubp1*	*Wasl*	*Zdhhc15*	*Zeb2*	
	Zfp113	*Zfp131*	*Zfp182*	*Zfp39*	*Zfp407*	*Zfp606*	*Zfp81*	*Zfp90*	
	Zfx	*Zhx1*	*Zkscan1*	*Zkscan3*	*Zmynd11*				
Fmnl2	*Fmnl2*	*Hdac9*							
Meis1	*Meis1*	*Cdc5l*	*Zfp192*						
Ikzf2	*Atf1*	*Ikzf2*	*Klf10*	*Limd1*	*Lmo2*	*Nfic*	*Nfix*	*Taf1a*	*Zfp295*
Jun	*Jun*	*Atoh1*							
Klf1	*6330416L07Rik*	*Aebp2*	*Ankrd32*	*Ccnt2*	*Fem1b*	*Klf1*	*Mid1*	*Nudt12*	*Pms1*
	Polr3e	*Zfp597*							
Mllt3	*1810007M14Rik*	*Ahctf1*	*Ankhd1*	*Arid4a*	*Ash1l*	*Atm*	*BC003267*	*Bcl11a*	
	Bclaf1	*Brwd1*	*Cebpg*	*Cebpz*	*Creb1*	*Dek*	*Dmtf1*	*Garnl1*	
	Hltf	*Jmy*	*Lcorl*	*Mef2a*	*Mga*	*Mll1*	*Mll3*	*Mll5*	
	Mllt3	*Mtf2*	*Myst3*	*Narg1*	*Ncoa2*	*Nfat5*	*Nr3c1*	*Nrip1*	
	Nsbp1	*Papola*	*Phf14*	*Rb1*	*Rbm39*	*Rybp*	*Shprh*	*Sirt1*	
	Sp4	*Taf1*	*Tcf4*	*Tmem131*	*Uhrf2*	*Wwp1*	*Zbtb1*	*Zeb1*	
	Zfp148	*Zfp281*	*Zfp292*	*Zfp36l1*	*Zfp386*	*Zfp445*	*Zfp451*	*Zfp59*	
	Zfp68	*Zfp748*	*Zmym2*						

(Table 2.2, continued)

Zfpm1	*Zfpm1*							
Zbtb16	*Zbtb16*							
Gata1	*Gata1*							
Gata2	*Dach1*	*Gata2*	*Pbx1*					
Rai14	*Mpl*	*Rai14*						
Egr1	*Atf4*	*Cbx7*	*Cited2*	*Egr1*	*Fli1*	*Foxo1*	*Fus*	*Gata3*
	Id2	*Lass4*	*Mycn*	*Ndn*	*Nfkbiz*	*Sfrs5*	*Tsc22d 3*	

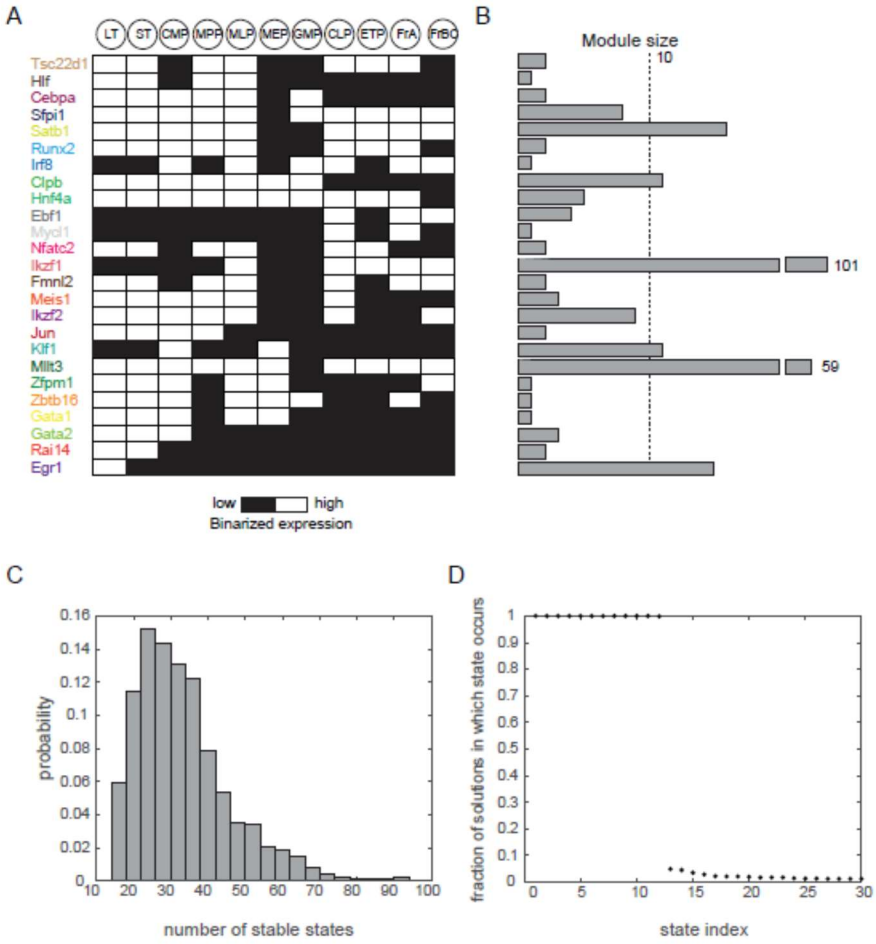


Figure 2.6: Characteristics of Gene Modules and Quantitative Modeling of the Core Network Underlying Hematopoiesis. (A) Binary gene expression patterns across all cell types (columns) for the 25 inferred gene modules. Color code of gene modules as in Figure 2.5. (B) Bar graph indicating the number of genes in each module. (C) Probability distribution of the number of stable states for each sampled network. Distribution estimated using sampling of 5,000 distinct solutions J_{ij}. (D) Fraction of solutions J_{ij} in which each state is stable. Shown are the states which are stable in the greatest fraction of states. Apart from the null state $\vec{m} = \vec{0}$ and the states associated with the 11 cell types, no state was stable in more than 5% of solutions.

2.4.5. Local-field gene regulatory network model for gene modules

In order to build a quantitative model relating the gene modules, we write a N-component gene regulatory network governed by a set of differential equations:

$$\dot{m}_i = -\frac{m_i}{\tau_i} + r_i^0 + r_i(\vec{m}), \qquad i = 1, \ldots, N,$$

where τ_i and r_i^0 are respectively the life-time and basal production rate of module i; we will rescale $\tau_i = 1$ and $r_i^0 = 0$ with loss of generality. We denote the level of module i as m_i. We assume here that modules interact only by modulating each-other's rate of production, described here by rate functions $r_i(\vec{m})$ which depend on the state $\vec{m} = [m_1, \ldots, m_N]$ of the gene regulatory network. We model inherent biological variability through a Gaussian noise term $\eta_i(t)$ with mean 0 and variance σ^2_{noise}.

As above, we consider that the production rate $r_i(\vec{m})$ is the result of only direct interactions, in which each gene *j* exerts a drive on gene *i* which is equal to an interaction strength J_{ij} (positive or negative) multiplied by the level of module *j*. The total drive ϕ_i on gene *i* is the sum of the drives from the different modules:

$$\phi_i(\vec{m}) = \sum_{j=1}^{N} J_{ij} m_j \, .$$

We now assume r_i has a universal scaling form that is the same for all factors,

$$r_i(\vec{m}) = r[\mu(\phi_i - \phi_0)],$$

where $r(\phi; \phi_0, \mu)$ is a monotonic sigmoidal function centered at ϕ_0 and bounded by the limits

$$r(\phi) = \begin{cases} 0, & \phi \ll \phi_0 \\ 1, & \phi \gg \phi_0 \end{cases};$$

the sharpness of crossover is determined by the nonlinearity parameter μ. The upper bound of $r_i = 1$ sets the maximum sustainable expression at $m_i = 1$. In the limit $\mu \to \infty$, $r(\phi)$ becomes the Heaviside step function, and $m_i \in \{0,1\}$ is binary.

Suppose state $\vec{m}^\alpha = \{m_1^\alpha, \dots, m_{25}^\alpha\}$ with expression level m_i^α in module i is a stable state of the network. In the limit $\mu \to \infty$, the condition for $\vec{m}^\alpha$ to be a fixed point is:

$$m_i^\alpha = H\left(\sum_j J_{ij} m_j^\alpha - \phi_0\right), \quad m_i^\alpha, m_j^\alpha \in \{0,1\},$$

where H is the Heaviside step function. (Note that if ϕ_0>0 then $\vec{m} = \vec{0}$ is always a stable fixed point of the network, *i.e.* the network will not spontaneously come back from the dead.)

In this limit, each state $\vec{m}^\alpha$ of the network is associated with N constraints given by inequalities of the form

$$m_i^\alpha = 0 \;\Rightarrow \sum_j J_{ij} m_j^\alpha < \phi_0,$$

$$m_i^\alpha = 1 \;\Rightarrow \sum_j J_{ij} m_j^\alpha > \phi_0.$$

If $\vec{m}^\alpha$ is a fixed point, all N of its constraints must hold. If we know the fixed points of the network, then we can write down a system of inequalities that constrain possible

values for J_{ij}. Since gene-gene interactions cannot be infinitely strong, J_{ij} must be bounded. We take $|J_{ij}| < 1$ and $\phi_0 = 0.1$.

2.4.6. Linear programming

The constraints placed on J_{ij} by the fixed point condition are linear in J_{ij}. We can take advantage of this fact and use linear programming methods to obtain solutions for J_{ij} by extremalizing a linear objective function of the form

$$U(J_{ij}) = \sum_{i,j} a_{ij} J_{ij} = \text{constant},$$

where a_{ij} are constant coefficients. The system of constraints defines a N^2-dimensional polytope in J-space that encloses all solutions of J_{ij} consistent with the fixed-point constraints, and U defines a $N^2 - 1$ dimensional hyperplane. Linear programming returns a solution for J_{ij} (a point in J-space) where the polytope contacts a U-plane of extremal value. The solution will lie on the boundary of the polytope and is in general non-unique. There is no general principle with which to select any specific U-plane as the "best" objective function. Furthermore, one would like to sample points in the interior of the polytope, and not just on its surface. Here, guided by the fact that we seek pertubative solutions for J_{ij} that ideally lie close to the origin, we impose a fictitious additional constraint on the polytope in the form of a hyperplane that contains the origin

$$\sum_{i,j} a_{ij} J_{ij} \leq 0, \quad a_{ij} \in \{0,1\},$$

where the coefficients a_{ij} are randomly chosen; this in effect slices the polytope in two and exposes an interior plane. Then, using the same choices of a_{ij} to define a U-plane, we seek a linear programming solution that maximizes U, i.e. a solution that lies on the now-exposed interior plane (if possible). Because these fictitious constraints radiate from the origin, points in the polytope that lie closest to the origin are sampled more densely.

2.4.7. Common features of the sampled networks

By using many different randomly generated fictitious constraints to sample the polytope, we can study the ensemble of model networks that all satisfy the fixed point constraints, and attempt to determine whether they share any common regulatory motifs. As discussed in the main text, we sampled 5,000 solutions J_{ij} that satisfied the fixed-point constraints defined by the binarized expression patterns of the known cell types. We then calculated the mean and coefficient of variation (c.v.) for each coupling. We were thus able to discover a core network between the different modules that is shared by the majority of solutions (Figure 2.5D).

2.4.8. Spurious fixed points

By construction, the 11 observed cell types (and the null state $\vec{m} = \vec{0}$) are fixed points for all 5,000 sampled solutions for J_{ij}. However, each solution J_{ij} may have additional spurious fixed points. The number of fixed points associated with each solution J_{ij} varied between 12 and 95, with an average of 33 (Figure 2.6C). However, none of the spurious fixed points occurs in more than 5% of sampled solutions (Figure 2.6D).

2.4.9. Reprogramming predictions

Given a particular solution J_{ij}, any arbitrary state of the network $\vec{m}$ (not necessarily a fixed point) will have dynamics obeying

$$m_i(t+1) = H\left(\sum_j J_{ij} m_j(t) - \phi_0\right), \quad (1)$$

where $m_i(t)$ and $m_i(t+1)$ are the level of module i at successive discretized time points.

For each particular solution J_{ij}, cells will get stuck in spurious fixed points; yet these spurious fixed points are highly unlikely to exist since they are stable in only a small number of the sampled J_{ij}. We can capture the average dynamics of different states of the network given the set of sampled solutions $\{J_{ij}\}$ by calculating the probability over all sampled solutions of moving from one arbitrary state $\vec{m}^a$ to another arbitrary state $\vec{m}^b$. This allows us defines a 2^{25} x 2^{25} state-to-state transition matrix $\mathcal{T}$:

$$\mathcal{T}_{b \leftarrow a} = p\left(\vec{m}^a \rightarrow \vec{m}^b \middle| \{J_{ij}\}\right). \quad (2)$$

If we denote as $\vec{p}(t)$ the vector of probabilities of being in the 2^{25} different states at time t, then

$$\vec{p}(t+1) = \mathcal{T}\,\vec{p}(t). \quad (3)$$

In order to test reprogramming hypotheses, we calculated the probability of moving between fixed points $\vec{m}^\alpha$ and $\vec{m}^\beta$ when overexpressing some set of modules $\{m_i\}$. We calculated the dynamics using the transition matrix $\mathcal{T}$ and enforced the overexpression of

the set of modules at each time point, updating the probabilities $\vec{p}(t)$ accordingly. The probabilities shown in Figure 2.5 are after 1,000 time steps.

Acknowledgements

We thank Sandeep Choubey, Vilas Menon and Toshihiko Oki for scientific discussions. We also thank Christof Koch, Sean Eddy, Nathan Kutz, Andrew Murray, KC Huang, Jim Valcourt, and Dann Huh for detailed comments and feedback on this work.

References

Abrikosov, A.A. (2004). Nobel Lecture: Type-II superconductors and the vortex lattice. Rev. Mod. Phys. *76*, 975–979.

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C.T., Bryder, D., Yang, L., Borge, O.-J., Thoren, L.A.M., et al. (2005). Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell *121*, 295–306.

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature *404*, 193–197.

Anderson, P.W. (1978). Local moments and localized states. Rev. Mod. Phys. *50*, 191–201.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. Mol. Syst. Biol. *3*, 78.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. Nat. Genet. *37*, 382–390.

Best, R.B., and Hummer, G. (2005). Reaction coordinates and rates from transition paths. Proc. Natl. Acad. Sci. U. S. A. *102*, 6732–6737.

Best, R.B., and Hummer, G. (2011). Diffusion models of protein folding. Phys. Chem. Chem. Phys. *13*, 16902–16911.

Buckingham, M.E., and Meilhac, S.M. (2011). Tracing cells for tracking cell lineage and clonal behavior. Dev. Cell *21*, 394–409.

Busslinger, M. (2004). Transcriptional control of early B cell development. Annu. Rev. Immunol. *22*, 55–79.

Chung, H.S., Piana-Agostinetti, S., Shaw, D.E., and Eaton, W.A. (2015). Structural origin of slow diffusion in protein folding. Science *349*, 1504–1510.

Corson, F., and Siggia, E.D. (2012). Geometry, epistasis, and developmental patterning. Proc. Natl. Acad. Sci. U. S. A. *109*, 5568–5575.

Crispino, J.D. (2005). GATA1 in normal and malignant hematopoiesis. Semin. Cell Dev. Biol. *16*, 137–147.

Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T., and Shakhnovich, E.S. (1998). On the transition coordinate for protein folding. J. Chem. Phys. *108*, 334.

Ficara, F., Crisafulli, L., Lin, C., Iwasaki, M., Smith, K.S., Zammataro, L., and Cleary, M.L. (2013). Pbx1 restrains myeloid maturation while preserving lymphoid potential in hematopoietic progenitors. J. Cell Sci. *126*, 3181–3191.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G., and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. Cancer Res. *68*, 5924–5931.

Gazit, R., Garrison, B.S., Rao, T.N., Shay, T., Costello, J., Ericson, J., Kim, F., Collins, J.J., Regev, A., Wagers, A.J., et al. (2013). Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. Stem Cell Reports *1*, 266–280.

Gilbert, S.F. (2014). Developmental Biology (Sinauer).

Goossens, S., Janzen, V., Bartunkova, S., Yokomizo, T., Drogat, B., Crisan, M., Haigh, K., Seuntjens, E., Umans, L., Riedt, T., et al. (2011). The EMT regulator Zeb2/Sip1 is essential for murine embryonic hematopoietic stem/progenitor cell differentiation and mobilization. Blood *117*, 5620–5630.

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

Guo, G., Luc, S., Marco, E., Lin, T.-W., Peng, C., Kerenyi, M.A., Beyaz, S., Kim, W., Xu, J., Das, P.P., et al. (2013). Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. Cell Stem Cell *13*, 492–505.

Heng, T.S.P., and Painter, M.W. (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. *9*, 1091–1094.

Iwasaki, H., and Akashi, K. (2007). Myeloid lineage commitment from the hematopoietic stem cell. Immunity *26*, 726–740.

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.

Jojic, V., Shay, T., Sylvia, K., Zuk, O., Sun, X., Kang, J., Regev, A., Koller, D., Best, A.J., Knell, J., et al. (2013). Identification of transcriptional regulators in the mouse immune system. Nat. Immunol. *14*, 633–643.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M. V, Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

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.

Kondo, M., Weissman, I.L., and Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell *91*, 661–672.

Koschmieder, S., Rosenbauer, F., Steidl, U., Owens, B.M., and Tenen, D.G. (2005). Role of transcription factors C/EBPalpha and PU.1 in normal hematopoiesis and leukemia. Int. J. Hematol. *81*, 368–377.

Krivov, S. V (2013). On Reaction Coordinate Optimality. J. Chem. Theory Comput. *9*, 135–146.

Kurotaki, D., Osato, N., Nishiyama, A., Yamamoto, M., Ban, T., Sato, H., Nakabayashi, J., Umehara, M., Miyake, N., Matsumoto, N., et al. (2013). Essential role of the IRF8-KLF4 transcription factor cascade in murine monocyte differentiation. Blood *121*, 1839–1849.

Landau, L.D., and Lifshitz, E.M. (1951). Statistical Physics, Volume 5 (Elsevier).

Laurenti, E., Varnum-Finney, B., Wilson, A., Ferrero, I., Blanco-Bose, W.E., Ehninger, A., Knoepfler, P.S., Cheng, P.-F., MacDonald, H.R., Eisenman, R.N., et al. (2008). Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. Cell Stem Cell *3*, 611–624.

Laurenti, E., Doulatov, S., Zandi, S., Plumb, I., Chen, J., April, C., Fan, J.-B., and Dick, J.E. (2013). The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. Nat. Immunol. *14*, 756–763.

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc. Natl. Acad. Sci. U. S. A. *102*, 4936–4942.

Machta, B.B., Chachra, R., Transtrum, M.K., and Sethna, J.P. (2013). Parameter space compression underlies emergent theories and predictive models. Science *342*, 604–607.

Macian, F. (2005). NFAT proteins: key regulators of T-cell development and function. Nat. Rev. Immunol. *5*, 472–484.

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. (2015a). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

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. (2015b). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Margolin, A.A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., and Califano, A. (2006). Reverse engineering cellular networks. Nat. Protoc. *1*, 662–671.

McGibbon, R.T., and Pande, V.S. (2016). Identification of simple reaction coordinates from complex dynamics. 16.

Van der Meer, L.T., Jansen, J.H., and van der Reijden, B.A. (2010). Gfi1 and Gfi1b: key regulators of hematopoiesis. Leukemia *24*, 1834–1843.

Min, I.M., Pietramaggiori, G., Kim, F.S., Passegué, E., Stevenson, K.E., and Wagers, A.J. (2008). The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. Cell Stem Cell *2*, 380–391.

Miyawaki, K., Arinobu, Y., Iwasaki, H., Kohno, K., Tsuzuki, H., Iino, T., Shima, T., Kikushige, Y., Takenaka, K., Miyamoto, T., et al. (2015). CD41 marks the initial myelo-erythroid lineage specification in adult mouse hematopoiesis: redefinition of murine common myeloid progenitor. Stem Cells *33*, 976–987.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

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.

Pereira, C., Clarke, E., and Damen, J. (2007). Hematopoietic colony-forming cell assays. Methods Mol. Biol. *407*, 177–208.

Perrimon, N., Pitsouli, C., and Shilo, B.-Z. (2012). Signaling mechanisms controlling cell fate and embryonic patterning. Cold Spring Harb. Perspect. Biol. *4*, a005975.

Pina, C., May, G., Soneji, S., Hong, D., and Enver, T. (2008). MLLT3 regulates early human erythroid and megakaryocytic cell fate. Cell Stem Cell *2*, 264–273.

Pongubala, J.M.R., Northrup, D.L., Lancki, D.W., Medina, K.L., Treiber, T., Bertolino, E., Thomas, M., Grosschedl, R., Allman, D., and Singh, H. (2008). Transcription factor EBF restricts alternative lineage options and promotes B cell fate commitment independently of Pax5. Nat. Immunol. *9*, 203–215.

Qi, X., Hong, J., Chaves, L., Zhuang, Y., Chen, Y., Wang, D., Chabon, J., Graham, B., Ohmori, K., Li, Y., et al. (2013). Antagonistic regulation by the transcription factors C/EBPα and MITF specifies basophil and mast cell fates. Immunity *39*, 97–110.

Radomska, H.S., Huettner, C.S., Zhang, P., Cheng, T., Scadden, D.T., and Tenen, D.G. (1998). CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. Mol. Cell. Biol. *18*, 4301–4314.

Ragu, C., Boukour, S., Elain, G., Wagner-Ballon, O., Raslova, H., Debili, N., Olson, E.N., Daegelen, D., Vainchenker, W., Bernard, O.A., et al. (2010). The serum response factor (SRF)/megakaryocytic acute leukemia (MAL) network participates in megakaryocyte development. Leukemia *24*, 1227–1230.

Rebollo, A., and Schmitt, C. (2003). Ikaros, Aiolos and Helios: transcription regulators and lymphoid malignancies. Immunol. Cell Biol. *81*, 171–175.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Riddell, J., Gazit, R., Garrison, B.S., Guo, G., Saadatpour, A., Mandal, P.K., Ebina, W., Volchkov, P., Yuan, G.-C., Orkin, S.H., et al. (2014). Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. Cell *157*, 549–564.

Robert-Moreno, A., Espinosa, L., de la Pompa, J.L., and Bigas, A. (2005). RBPjkappa-dependent Notch function regulates Gata2 and is essential for the formation of intra-embryonic hematopoietic cells. Development *132*, 1117–1126.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Satoh, Y., Yokota, T., Sudo, T., Kondo, M., Lai, A., Kincade, P.W., Kouro, T., Iida, R., Kokame, K., Miyata, T., et al. (2013). The Satb1 protein directs hematopoietic stem cell differentiation toward lymphoid lineages. Immunity *38*, 1105–1115.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Solar, G.P., Kerr, W.G., Zeigler, F.C., Hess, D., Donahue, C., de Sauvage, F.J., and Eaton, D.L. (1998). Role of c-mpl in early hematopoiesis. Blood *92*, 4–10.

Stier, S., Cheng, T., Dombkowski, D., Carlesso, N., and Scadden, D.T. (2002). Notch1 activation increases hematopoietic stem cell self-renewal in vivo and favors lymphoid over myeloid lineage outcome. Blood *99*, 2369–2378.

Sugawara, T., Oguro, H., Negishi, M., Morita, Y., Ichikawa, H., Iseki, T., Yokosuka, O., Nakauchi, H., and Iwama, A. (2010). FET family proto-oncogene Fus contributes to self-renewal of hematopoietic stem cells. Exp. Hematol. *38*, 696–706.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.

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

Tamura, T., Nagamura-Inoue, T., Shmeltzer, Z., Kuwata, T., and Ozato, K. (2000). ICSBP directs bipotential myeloid progenitor cells to differentiate into mature macrophages. Immunity *13*, 155–165.

Thomson, M., Liu, S.J., Zou, L.-N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875–889.

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

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

Vaillant, F., Blyth, K., Andrew, L., Neil, J.C., and Cameron, E.R. (2002). Enforced expression of Runx2 perturbs T cell development at a stage coincident with beta-selection. J. Immunol. *169*, 2866–2874.

Wang, H., Lee, C.H., Qi, C., Tailor, P., Feng, J., Abbasi, S., Atsumi, T., and Morse, H.C. (2008). IRF8 regulates B-cell lineage specification, commitment, and differentiation. Blood *112*, 4028–4038.

Zeisel, A., Manchado, A.B.M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-.). *347*, 1138–1142.

Zhang, P., Behre, G., Pan, J., Iwama, A., Wara-Aswapati, N., Radomska, H.S., Auron, P.E., Tenen, D.G., and Sun, Z. (1999). Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. Proc. Natl. Acad. Sci. U. S. A. *96*, 8705–8710.

Chapter 3. Probabilistic model of gene networks controlling embryonic stem cell differentiation inferred from single-cell transcriptomics

[A large part of this chapter is in review as Sumin Jang*, Leon Furchtgott*, Sandeep Choubey*, Ling-Nan Zou, Adele Doyle, Vilas Menon, Ethan Loew, Anne-Rachel Krostag, Refugio A. Martinez, Linda Madisen, Boaz P. Levi, Sharad Ramanathan, "Probabilistic model of gene networks controlling embryonic stem cell differentiation inferred from single-cell transcriptomics." SR designed the study. SJ performed experiments and data analysis. LF and SC performed single-cell transcriptomic analysis based on method developed by LF. SC and LF implemented the gene regulatory network model and linear programming method developed by LNZ and SR. LNZ, AD, VM, EL, ARK, RM, LM and BL performed experiments. SJ, LF, SC and SR wrote the manuscript with input from all of the authors.]

Abstract

A quantitative understanding of how gene regulatory networks lead multipotent cells to acquire different cell fates has been challenging. Using a novel Bayesian framework to analyze single-cell transcriptomics data, we infer the gene expression dynamics of early differentiation of mouse embryonic stem cells, revealing discrete state transitions across nine cell states. Using a probabilistic model of the gene regulatory networks, we predict that these states are further defined by distinct responses to perturbations. We experimentally verify three predictions of such state-dependent behavior: that whether (i) *Sox2* overexpression represses *Oct4*, (ii) *Snai1* overexpression represses *Oct4*, and (iii) LIF and BMP promote pluripotency or differentiation into neural crest, all depend on cell

state. This study provides a framework to infer predictive models of the gene regulatory networks that drive cell fate decisions.

3.1. Introduction

During differentiation, cells repeatedly choose between alternative fates in order to give rise to a multitude of distinct cell types. A major challenge in developmental biology is to uncover the dynamics of gene expression and the underlying gene regulatory networks that lead cells to their different fates. Given the complexity of gene regulatory networks, with their large number of components and even larger number of potential interactions between those components, building detailed predictive mathematical models is challenging. The lack of sufficient data requires a large number of assumptions to be made in order to constrain all the parameters in such models (Karr et al., 2012).

A potential alternative approach to modeling gene regulatory networks underlying the dynamics of differentiation is through the application of tools developed and used for understanding the statistical physics of state transitions in complex physical systems (Anderson, 1984; Landau and Lifshitz, 1980). The key to understanding these physical systems is based on the discovery that in every one of them, instead of measuring the large number of variables associated with the constituents of the system, the experimentally measured value of low-dimensional order parameters (Figure 3.1A) was sufficient to accurately describe both the state and the state transitions of the underlying dynamical system. Mathematical models based on order parameters have led to a fundamental understanding of state transitions in a variety of physical systems (Ekspong, 1993, 2008; Lundqvist, 1992). We asked whether we could similarly determine the suitable parameters to quantitatively describe and model cell state transitions during development.

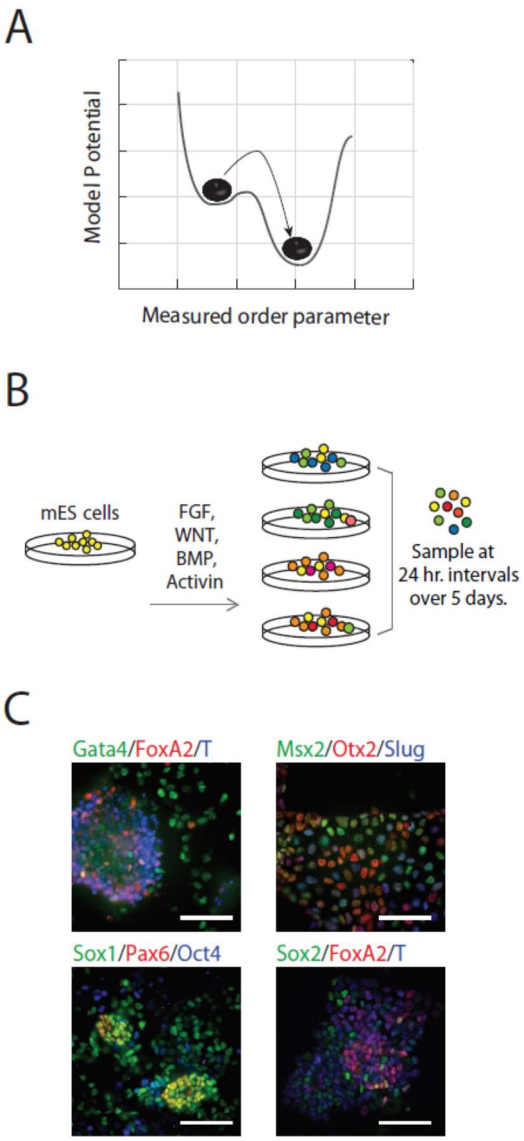


Figure 3.1: Single-Cell Gene Expression Profiling of mESCs during early germ layer differentiation. (A). States and state transitions in complex physical systems are characterized by order parameters whose value (x-axis) changes as the system changes state. Theoretical models based on effective potential functions (y-axis) of these order parameters are formulated to understand and make quantitative predictions about both the statistical mechanics and dynamics of the systems. (B). Mouse embryonic stem cells (mESCs) were exposed to various differentiation conditions to perturb FGF, WNT, and TGF-beta signaling for up to five days of differentiation. Single cells, collected every 24 hours during differentiation, were transcriptionally profiled using CEL-Seq. (C). Images of immunostained mESCs undergoing differentiation show cell-to-cell variability in their expression of known germ layer marker genes. (Scale bar = 100μm)

Here we study the specific example of early mammalian germ layer differentiation of pluripotent mouse embryonic stem (mES) cells, which are derived from the inner cell mass of the peri-implantation-stage embryo (see pictorial summary in Figure 3.2A). During this stage, both mES cells and cells *in vivo* express key pluripotency factors, such as *Nanog*, *Sox2*, *Oct4*, *Klf4*, *Jarid2*, and *Esrrb*, which mutually activate one another to form a pluripotency circuit (Kim et al., 2008; Young, 2011; Zhou et al., 2007). Following implantation, naïve pluripotent ES cells of the inner cell mass downregulate *Klf4* and upregulate *Otx2*, *Dnmt3a*, and *Dnmt3b*, as they transition into "primed" pluripotent cells found in the epiblast (Nichols and Smith, 2009). Over the next few days of differentiation, TGF-beta signaling factors, with the aid of WNT/beta-catenin signaling, promote and inhibit the differentiation of pluripotent cells into mesendodermal (characterized by genes such as *Brachyury* (*T*), *FoxA2*, *Mixl1* and *Gsc*) and ectodermal (characterized by *Eras*, *Sez6*, *Stmn3*, and *Stmn4*) cell fates, respectively (Gadue et al., 2006; Hart et al., 2002; Li et al., 2015; Lindsley et al., 2006; Tada et al., 2005; Watabe and Miyazono, 2009). Mesendodermal progenitors further differentiate into mesoderm and definitive endoderm progenitors. Mesoderm cells are usually distinguished by expression of *Gata4* and *Eomes*, and endoderm cells by *Sox17* and *FoxA2*, although in mouse these genes are shared between both lineages, with differences only in their timing and level of expression (Arnold and Robertson, 2009; Kanai-Azuma et al., 2002; Kim and Ong, 2012; Lumelsky et al., 2001; Rojas et al., 2005). Along the ectodermal lineage, BMP signaling pushes ectodermal cells toward epidermis, while in the absence of BMP signaling, ectodermal cells acquire a neural fate (Wilson and Hemmati-Brivanlou, 1995). Epidermal cells are characterized by Keratins, whereas neural cells express *Sox1* and *Pax6* (Koch and Roop, 2004; Pevny et al.,

1998; Sansom et al., 2009; Streit and Stern, 1999). The cells at the physical border between epidermal and neural cells give rise to neural crest cells (expressing *Sox10*, *Msx2*, *Snai1* and *Slug*) in response to WNT and BMP signaling, which are often described as a fourth germ layer because of the diverse range of tissues to which they give rise (Gans and Northcutt, 1983; Knecht and Bronner-Fraser, 2002; Le Douarin, 1991). Despite the detailed understanding of early embryonic development revealed by decades of work in genetics and developmental biology, a quantitative understanding of how the underlying gene regulatory network leads cells through a series of cell fate decisions has remained elusive.

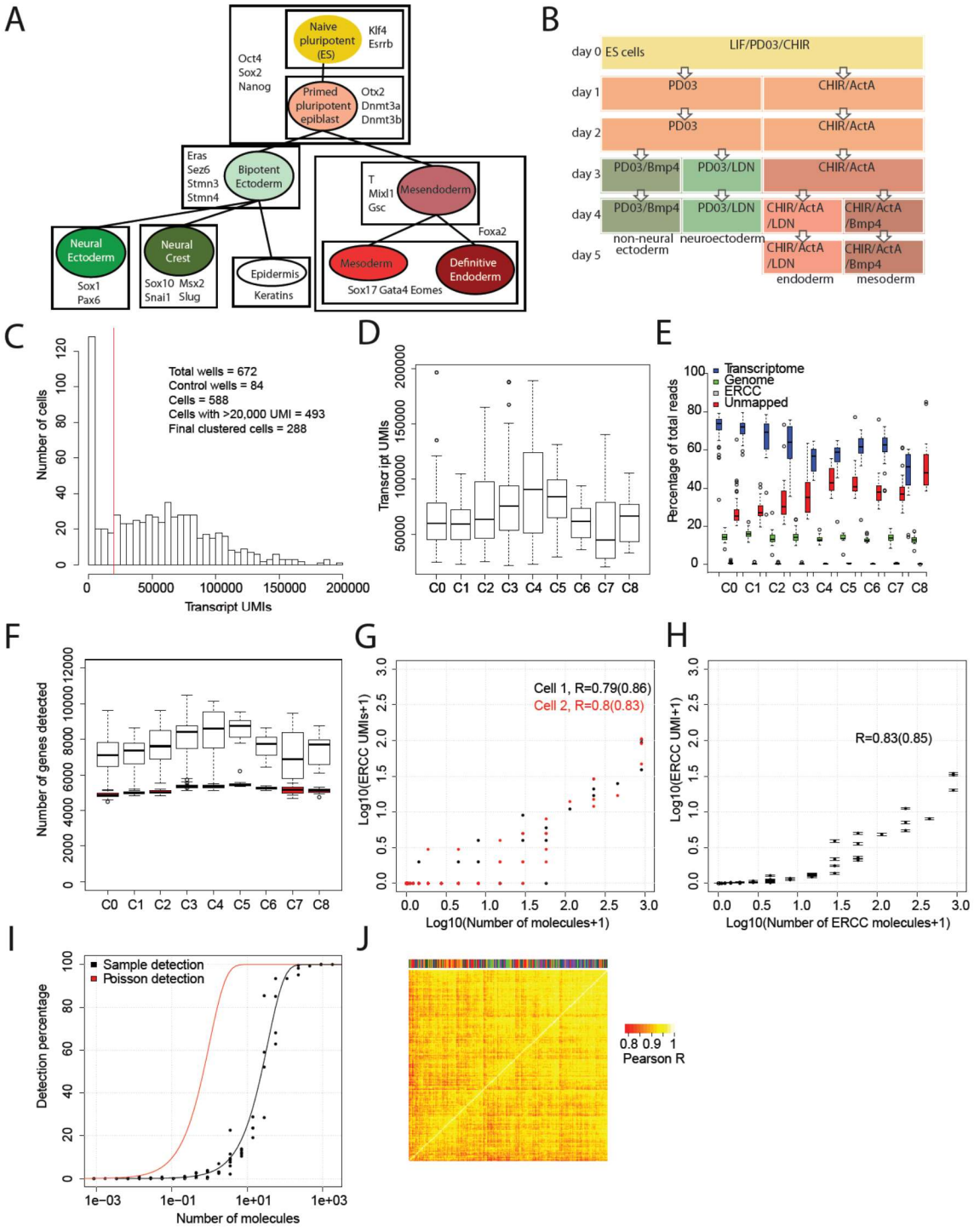


Figure 3.2: Literature summary and single-cell transcriptomic analysis. (A). Diagram summarizing the literature on cell types (each represented by a colored circle labeled by its name) that arise during early mouse germ layer differentiation, their lineage relationships (represented by lines connecting cell types), and genes that characterize each cell type (listed in boxes that surround the cell types in which they are expressed).

(Figure 3.2, continued) (B). Summary of cell culture conditions that were used to generate populations enriched with neural/non-neural ectoderm, definitive endoderm, or mesoderm-like cells over the course of up to five days. Undifferentiated ES cells were maintained in LIF/PD0/CHIR (i.e. Lif2i) conditions, and duration of differentiation was measured from the time at which these conditions were removed. (C). Histogram of the number of Unique Molecular Identifiers (UMIs) mapping to annotated genes per cell. Note that this histogram includes 84 control (empty or ERCC-only) wells. (D). Box and whisker plots of the number of UMIs mapping to annotated genes per cell, grouped by cell cluster (total cells = 288). (E). Percentage of reads mapping to the transcriptome, to the genome (i.e. regions outside the reference transcriptome annotation), and to ERCC spike-in control sequences, and percentage of reads unmapped per cell. Cells are grouped by cell cluster. (F). Box and whisker plots of the number of genes detected (UMI>0) per cell, grouped by cluster, using the full data set (clear boxes) and after subsampling cells to 20,000 transcriptome-mapping UMIs (red boxes). (G). Representative plots for two cells, showing the number of UMIs detected for each ERCC species versus the putative number of molecules spiked in. UMI counts are based on subsampling to 20,000 transcriptome-mapping UMIs. Pearson's R values in log space, using all ERCC species and using only ERCC species present at > 1 molecule (in parentheses) are shown for each cell. (H). Same as (E), but with mean and SEM values for all clustered cells, after subsampling to 20,000 transcriptome-mapping UMIs per cell. (I). Fraction of times a given ERCC species is detected (UMI>0) in all clustered cells, after subsampling to 20,000 transcriptome-mapping UMIs per cell, versus the putative number of ERCC molecules spiked in. The red line indicates expected detection fractions based on Poisson statistics of dilution, whereas the black line indicates the best fit through the experimental data. The fit suggests that the detection rate is 1 out of 35 molecules. (J). Clustered heatmap of Pearson's correlation coefficients among clustered cells based on ERCC UMI expression values. Clustering was done using average linkage with a distance metric of 1 – Pearson's R. The color bar at the top identifies the cluster membership of each cell; cells of the same type do not cluster together based on ERCC expression, suggesting a lack of process-related artifacts in the final clusters. All box and whisker plots use boxes to represent the 25th and 75th percentile, and whiskers represent 1.5 times the intraquartile range.

We use single-cell RNA-seq to determine how gene expression patterns change as mouse embryonic stem cells differentiate into different germ-layer progenitors. Motivated by approaches in statistical physics, we adopt a Bayesian framework to simultaneously infer cell states, the sequence of transitions between these states, and the key sets of genes whose expression patterns provide a parameter space in which the cell states and cell state transitions are inferred. Our computational analysis, together with flow cytometry and live-cell imaging of an *Otx2* reporter mES cell line, shows that cells reside in discrete states and

rapidly transition from one state to another. By requiring models of the underlying gene regulatory network to replicate the existence of the observed discrete cell states, we extract probability distributions of model parameters. Our probabilistic predictions using this model gene regulatory network show that the discrete cell states first inferred from their gene expression patterns are further defined by their unique responses to the same perturbations in signals and transcription factor levels. We experimentally validate these predictions using molecular tools, live-cell microscopy and flow cytometry. Finally, we discuss the biological implications of our results.

3.2. Results

3.2.1. Acquiring single-cell transcriptomics data during early differentiation

We differentiated populations of mES cells by exposing them to one of four combinations of signaling factors and small molecules to perturb FGF, WNT, and/or TGF-beta signaling for up to five days (Figure 3.1B; see also Figure 3.2B, Methods). Although cells in each population were differentiated in a monolayer culture and therefore exposed to nearly uniform conditions, we observed significant heterogeneity in the expression – as measured by immunofluorescence – of various known early germ layer marker genes (such as T, Pax6, Slug, FoxA2, and Gata4) in each population, suggesting a diversity of cell types under the same signaling conditions (Figure 3.1C). Further, undifferentiated pluripotent cells persisted in differentiating populations. Therefore, to capture the cell-to-cell variability within differentiating populations, we collected and transcriptionally profiled single cells every 24 hours over the course of five days of differentiation using a modified version of CEL-seq (Hashimshony et al., 2012). We obtained gene expression data from a

total of 288 cells (Figure 3.2C-J; Methods) with a median of 508,939 mapped reads, 48,475 transcripts and 7,032 genes detected per cell.

3.2.2. Bayesian statistical approach discovers appropriate coordinate systems to infer cell states and state transitions

We developed a Bayesian probabilistic framework (Chapter 2. ; Chapter 5.) that, given gene expression data from single cells$\{g_i\}$, simultaneously infers (i) cell cluster identities of the cells, $\{C\} \equiv \{c_1, c_2, \dots, c_N\}$, (ii) the sets of transitions $\{\mathrm{T}\}$ between these clusters, (iii) the key sets of marker genes $\{\alpha_i\}$ that define each cell cluster and (iv) the sets of genes $\{\beta_i\}$ that determine the transitions between clusters. We determined the maximum likelihood estimates of these variables using an iterative algorithm (Figure 3.3A; see also Figure 3.4A and Mathematical Appendix in Chapter 5).

We started by clustering the single-cell gene expression data for the 288 cells into 12 clusters $\{c_1^0, c_2^0, \dots, c_{12}^0\}$ using Seurat (Satija et al., 2015), restricting the analysis to transcription factors (2,672 total) because of their functional role in orchestrating global gene expression. Seurat identifies cell clusters by performing density-based clustering on a t-distributed Stochastic Neighboring (t-SNE) map of the gene expression data (Van der Maaten and Hinton, 2008). These clusters $\{C_0\} = \{c_1^0, c_2^0, \dots, c_{12}^0\}$, ranging in size from 14 to 47 single cells, served as a seed for our algorithm.

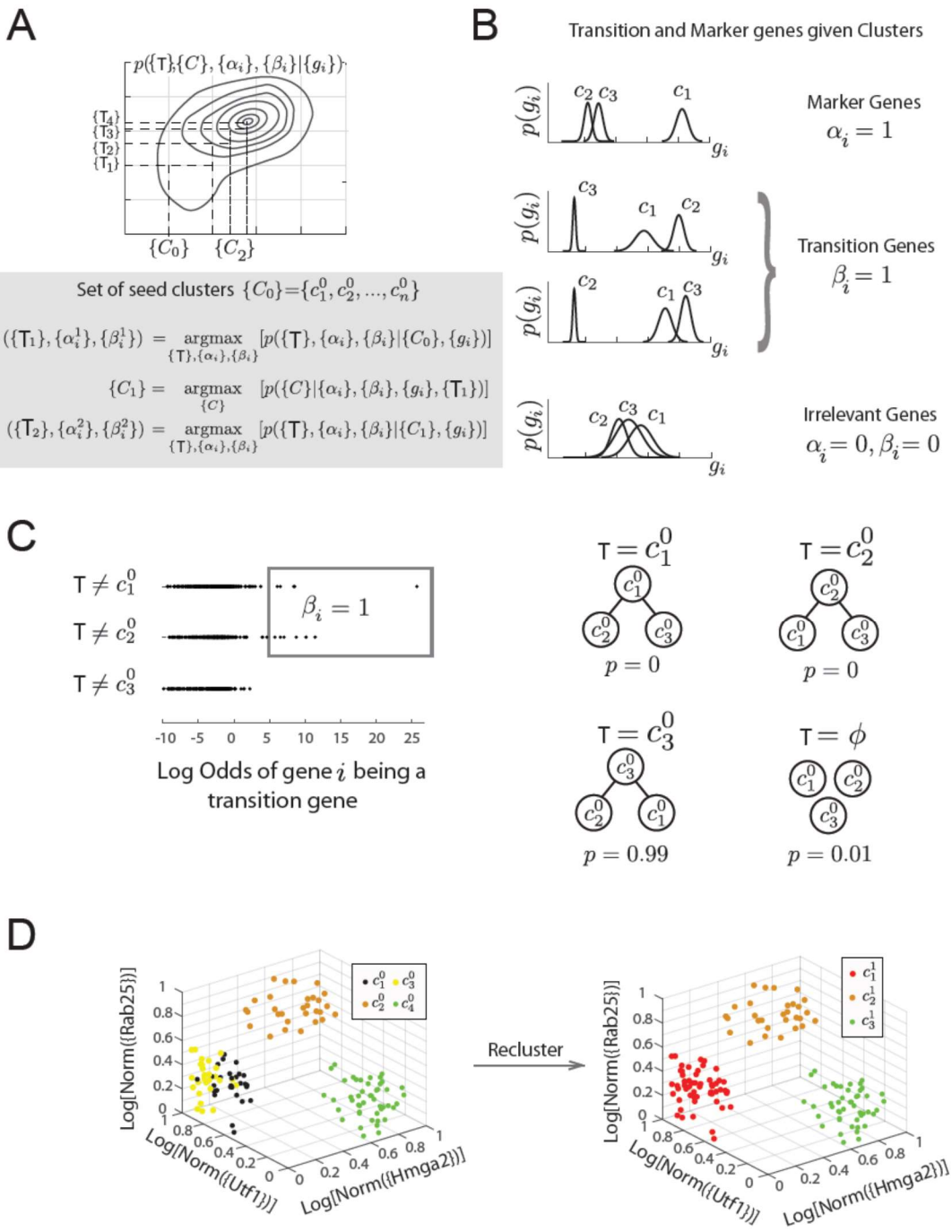


Figure 3.3: Bayesian framework to obtain cell cluster identities and transition relationships from single-cell transcriptomics data. (A). Maximization algorithm to determine most likely cluster identities $\{\boldsymbol{C}\} \equiv \{\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_n}\}$, sets of transitions $\{\mathbf{T}\}$, marker genes $(\boldsymbol{\alpha_i} = \mathbf{1})$ and transition genes $(\boldsymbol{\beta_i} = \mathbf{1})$, given single-cell gene expression data $\{\boldsymbol{g_i}\}$.

(Figure 3.3, continued) Starting from a seed clustering scheme $\{\boldsymbol{C_0}\}$, iterative maximization of the conditional probabilities $\boldsymbol{p}(\{\mathbf{T}\},\{\boldsymbol{\alpha_i}\},\{\boldsymbol{\beta_i}\}|\{\boldsymbol{g_i}\},\{\boldsymbol{C}\})$ and $\boldsymbol{p}(\{\boldsymbol{C}\}|\{\boldsymbol{g_i}\},\{\mathbf{T}\},\{\boldsymbol{\alpha_i}\},\{\boldsymbol{\beta_i}\})$ converges to most likely set $(\{\boldsymbol{C}\},\{\mathbf{T}\},\{\boldsymbol{\alpha_i}\},\{\boldsymbol{\beta_i}\})$. (See also Figure 3.3A and Chapter 5.) (B). Gene expression patterns of marker genes, transition genes, and irrelevant genes in cell clusters c_1, c_2, and c_3. Marker genes are highly expressed in only one cluster, whereas transition genes are highly expressed in two clusters and downregulated in the third. High probability transition genes alone are used for the determination of set of transitions; both high probability transition and marker genes are used for re-clustering. (C). For the three initial clusters $\boldsymbol{c_1^0}$, $\boldsymbol{c_2^0}$ and $\boldsymbol{c_3^0}$, plot of the odds of each gene (represented by a dot) being a transition gene (x-axis) and the cluster with the minimum expression of the gene (y-axis). In our framework, each gene's odds of being a transition gene is used to compute the probabilities of the sets of transitions $\mathbf{T}$ between the three clusters (Methods and Chapter 5). A gene whose expression is lowest in $\boldsymbol{c_k^0}$ casts a probabilistic vote against $\boldsymbol{c_k^0}$ being the intermediate state (i.e., against the relationships $\mathbf{T} = \boldsymbol{c_k^0}$), which is weighted by the odds that the gene is a transition gene, given the cluster identities. Two groups of genes (boxed) are the highest likelihood transition genes, casting a strong vote against $\boldsymbol{c_1^0}$ or $\boldsymbol{c_2^0}$ being the intermediate cell type. The computed probability of the topology given gene expression data indicates with .99 probability that $\boldsymbol{c_3^0}$ is the central node. (D). Left: single cells belonging to clusters $\boldsymbol{c_1^0} - \boldsymbol{c_4^0}$ (dots colored based on cluster identity) in the gene expression space defined by transition and marker genes (probability > 0.8) associated with triplet $\boldsymbol{c_2^0}$, $\boldsymbol{c_3^0}$, $\boldsymbol{c_4^0}$. Axes represent the normalized log expression values of, respectively, transition genes expressed in $\boldsymbol{c_2^0}$ and downregulated in $\boldsymbol{c_4^0}$ and $\boldsymbol{c_3^0}$, and marker genes for $\boldsymbol{c_2^0}$; the most likely gene of each class is represented in curly brackets. Right: after re-clustering cells in the subspace defined by high probability marker and transition genes, clusters $\boldsymbol{c_1^0}$ and $\boldsymbol{c_3^0}$ have merged.

We next considered every possible group of 3 clusters (e.g., c_1^0, c_2^0 and c_3^0) from a total of ${}^{12}C_3 = 220$ such combinations. For each triplet of clusters, we first determined the probability that each gene was a marker gene ($\alpha_i = 1$), a transition gene ($\beta_i = 1$) or neither ($\alpha_i, \beta_i = 0$) based on the distribution of their expression patterns in cells of each cluster. The marker and transition genes are defined as follows (Figure 3.3B):

(i) A marker gene g_i ($\alpha_i = 1$) has a distribution of expression levels that is highest in one cluster, and well separated from the distribution of its expression levels in the other two clusters (Figure 3.3B, the better this separation the higher the probability that the gene is a marker gene). Marker genes distinguish one of the clusters from the other two.

(ii) A transition gene g_j ($\beta_j = 1$) has a distribution of expression levels that is lowest in one cluster, and well separated from the distribution of its expression levels in the other two clusters (Figure 3.3B). Each such transition gene establishes relative relationships between the three clusters, implying a closer relationship between the two clusters in which it is expressed than with the cluster in which it is not.

(iii) Genes that are neither marker ($\alpha = 0$) nor transition genes ($\beta = 0$) do not follow constraints (i), (ii) on expression level distributions.

Computing the probability of each gene being a marker gene, a transition gene or neither allowed us to determine the most likely set of transitions T between each triplet of clusters. Each gene's contribution to the posterior probability of T is weighted by the odds ratio that the gene is a transition gene (Figure 3.3C; see also Supplemental Experimental Procedures). For example, for clusters c_1^0, c_2^0 and c_3^0, a gene whose expression is lowest in c_2^0 casts a vote against c_2^0 being the intermediate state (i.e., against the transition T = c_2^0, where c_2^0 is intermediate, Figure 3.3C, right) that is weighted by its odds of being a transition gene for those three clusters (Figure 3.3C, left). Our Bayesian framework led to a summation of these weighted votes to determine the most likely set of transitions between the three clusters and concomitantly the most likely marker and transition genes corresponding to these clusters and transitions (Figure 3.3C, right).

For the seed cluster set $\{C_0\}$, we determined 179 sets of transitions between clusters and identified 1,035 transcription factors that were high probability (probability > 0.5) marker or transition genes for at least one of the identified transitions. We next re-clustered the single cells in the gene expression space defined by these 1,035 marker or transition genes, using Seurat, to obtain a new cluster set $\{C_1\} = \{c_1^0, c_2^0, \ldots, c_{10}^0\}$, consisting of 10

clusters. In this process, cells changed cluster identities, and certain clusters merged (Figure 3.3D).

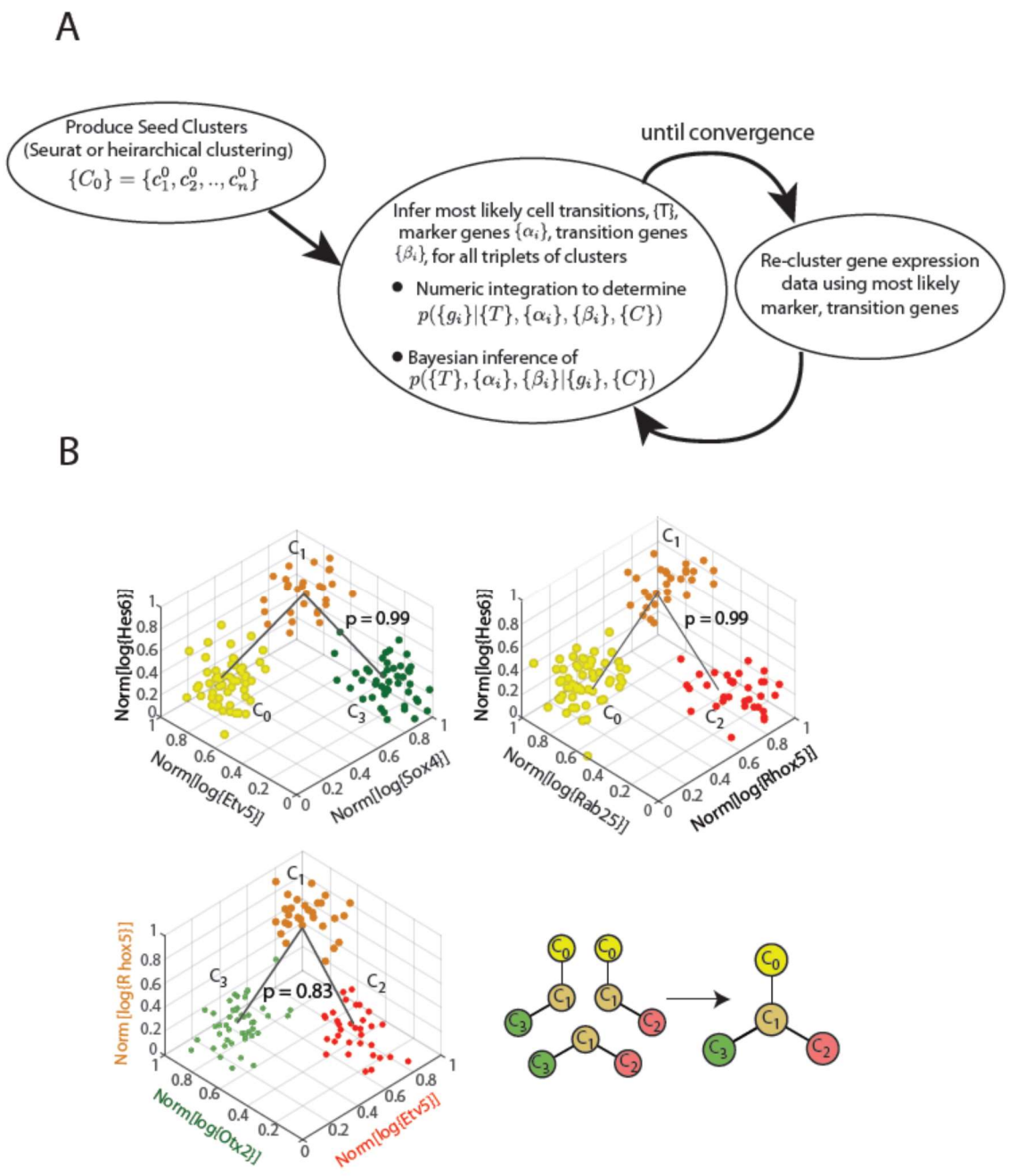


Figure 3.4: Details of clustering and lineage determination algorithm. (A). Workflow for clustering and lineage determination algorithm. (B). Three-dimensional plots for triplets (C_0, C_1, C_2), (C_0, C_1, C_3), and (C_1, C_2, C_3). Each dot in this space represents a single cell, and cells are colored based on their cluster identity.

(Figure 3.4, continued) The x- ,y- and z-axes are, respectively, the normalized log expression levels of the two classes of transition genes and the marker gene class for the intermediate. The probabilities of the inferred lineage topology are shown. In all three cases, cluster C_1 is inferred to be the intermediate cluster. Combining the inferred relationships between the four clusters, we obtain the relationships between the cell clusters C_0, C_1, C_2 and C_3.

By iteratively determining the most likely sets of transitions and the most likely marker and transition genes, and by re-clustering the cells within the subspace of these genes, our algorithm converged upon the most likely set of cell clusters, the sets of transitions between these cell clusters (Table 3.2), as well as the marker and transition genes for each set of three clusters after five iterations (Figure 3.5A). The final cluster set consists of 9 cell clusters ranging in size between 14 and 57 cells; every cell was mapped to a cluster. We combined the local sets of transitions between different triplets of clusters (Table 3.2) in order to infer the most parsimonious lineage tree between the clusters (Figure 3.5A; see also Figure 3.3B and Supplemental Experimental Procedures). Finally, we obtained identical final clusters starting with different seed cluster sets and using k-means clustering with the gap statistic, to show that our results were robust to the choice of seed clusters and clustering method (Figure 3.6; see also Supplemental Experimental Procedures).

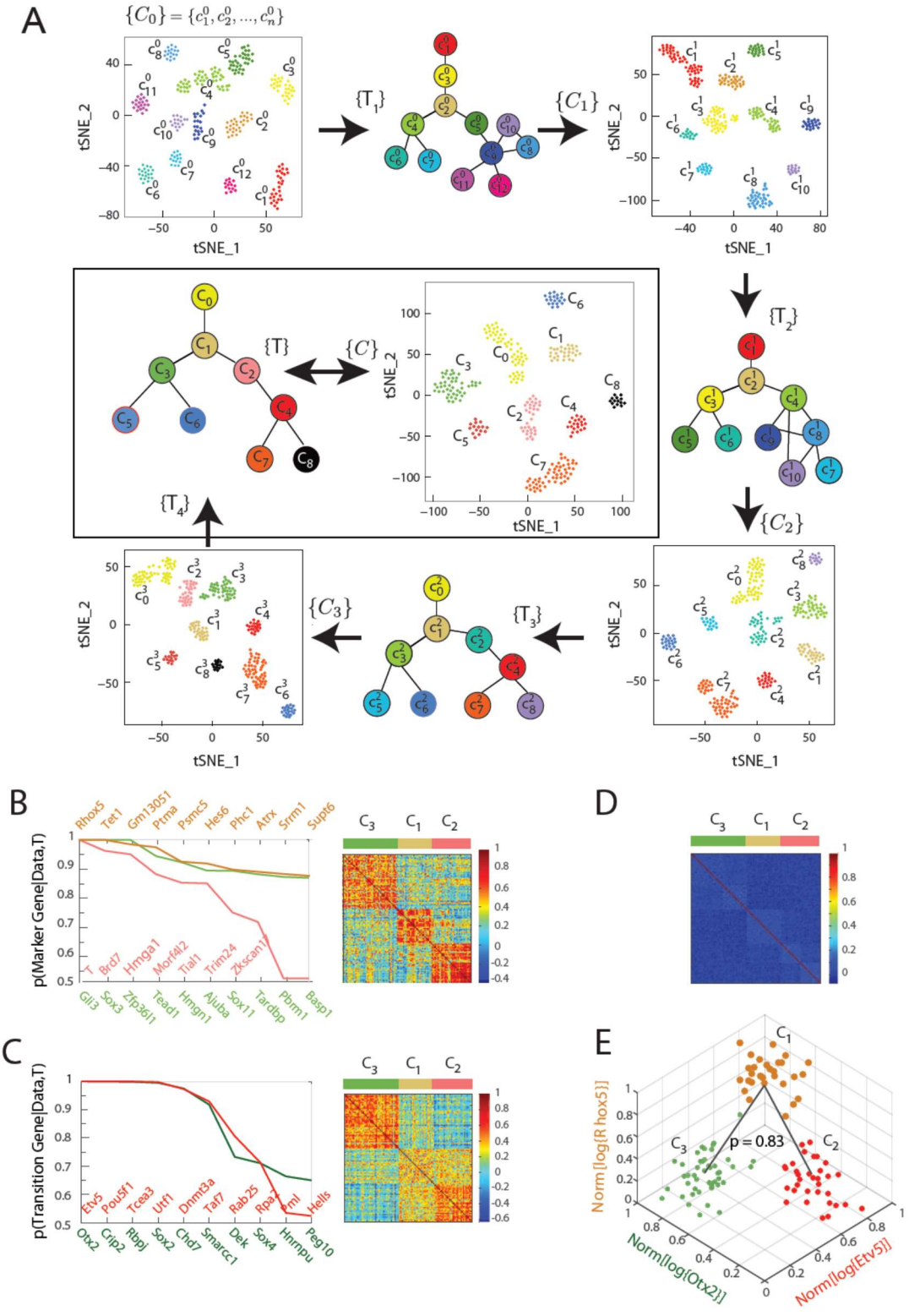


Figure 3.5: Iterative algorithm converges upon a set of cell clusters and local transitions that together define a multi-potent lineage tree. (A). Iterative determination

(Figure 3.5, continued) of the most likely sets of transitions {**T**} and re-clustering of cells in the resulting subspace of transition and marker genes, starting from a seed set of cluster identities $\{\boldsymbol{C_0}\}$. With each iteration, the cluster identities as well as the total number of clusters change, as shown by the Seurat t-SNE maps (each dot represents a cell, colored based on its cluster identity). The inferred sets of transitions between clusters at each iteration are represented as a lineage tree (each circle represents a cell cluster; see also Figure 3.4B). After five iterations, the algorithm converged upon a set of 9 clusters (shown in box). (See also Figure S3) (B). Left: Top ten genes (x-axis) with highest probability of being marker genes for clusters C_1 (yellow), C_2 (light red) and C_3 (light green) plotted against their probability of being marker genes. Right: The cell-cell correlation matrix computed using these 30 marker genes for the 108 cells belonging to clusters C_1, C_2 and C_3 shows three clear blocks of high correlation along the diagonal. (C). Left: Top ten genes (x-axis) with highest probability of being transition genes for clusters C_1, C_2 and C_3, plotted against their probability of being transition genes (y-axis). The transition genes belong to one of two classes, those that show high expression in cells belonging to C_1 and C_2 but low expression in C_3 (red), and those expressed at high levels in cells in clusters C_1 and C_3 but low levels in C_2 (green). The cell-cell correlation matrix (right) computed using these 20 transition genes shows that the 29 cells belonging to cluster C_1 have intermediate levels of correlation with cells in both C_2 and C_3, whereas the 46 cells in C_2 show low correlation levels with the 33 cells in C_3. (D). The cell-cell correlation matrix computed using all genes with coefficient of variation greater than 10, which includes transition and marker genes, shows a barely-detectable signature of the underlying cell clusters and the set of transitions between them. (E). The inferred clusters and their lineage relationships can be represented in a three-dimensional coordinate system where the x- and y- axes are the normalized log expression level of the two classes of transition genes (genes in (B), left) and the z-axis measures the normalized log expression level of the marker genes for cluster C_1 ((A) left in yellow). Each dot represents a single cell, and cells are colored based on their cluster identity.

The inferred lineage relationships between the final clusters could be visualized in the subspace of inferred marker and transition genes. We illustrate this first for the three clusters C_1, C_2, and C_3. We identified three classes of marker genes, each consisting of high-probability marker genes specific to one of the three clusters (Figure 3.5B). Each gene class is denoted by its highest probability member gene in curly brackets (e.g., {*Otx2*}). When the cell-cell Pearson correlation matrix between the cells in these three clusters was determined using all high-variance genes, the matrix showed a barely detectable structure (Figure 3.5D). In contrast, the same matrix computed using high-probability marker genes

for clusters C_1, C_2, and C_3 showed three distinct blocks of high correlation along the diagonal, each corresponding to a different cluster (Figure 3.5B). When the cell-cell correlations were measured using the two classes of inferred transition genes (Figure 3.5C, left), each consisting of high-probability transition genes present in C_1 and downregulated either in C_2 or in C_3, the correlation matrix showed intermediate correlation levels between C_1 and either C_2 or C_3, and low correlation levels between C_2 and C_3 (Figure 3.5C, right). The distribution functions of these transition genes in the different clusters led to the inference that clusters C_2 and C_3 are connected via cluster C_1 with a probability of 0.83.

We visualized the gene expression changes that characterize transitions from one cell cluster to another by plotting the cells in C_1, C_2 and C_3 in a three-dimensional gene expression subspace (Figure 3.5E), using as axes the mean normalized expression levels of the two transition gene classes down-regulated in C_2 or C_3 (in red and green in Figure 3.5B) and of the marker gene class specific to C_1 (Figure 3.5A in orange). Like the order parameters used to describe state transitions in physical systems, these axes constitute a low-dimensional coordinate system for the inferred set of transitions between C_1, C_2 and C_3.

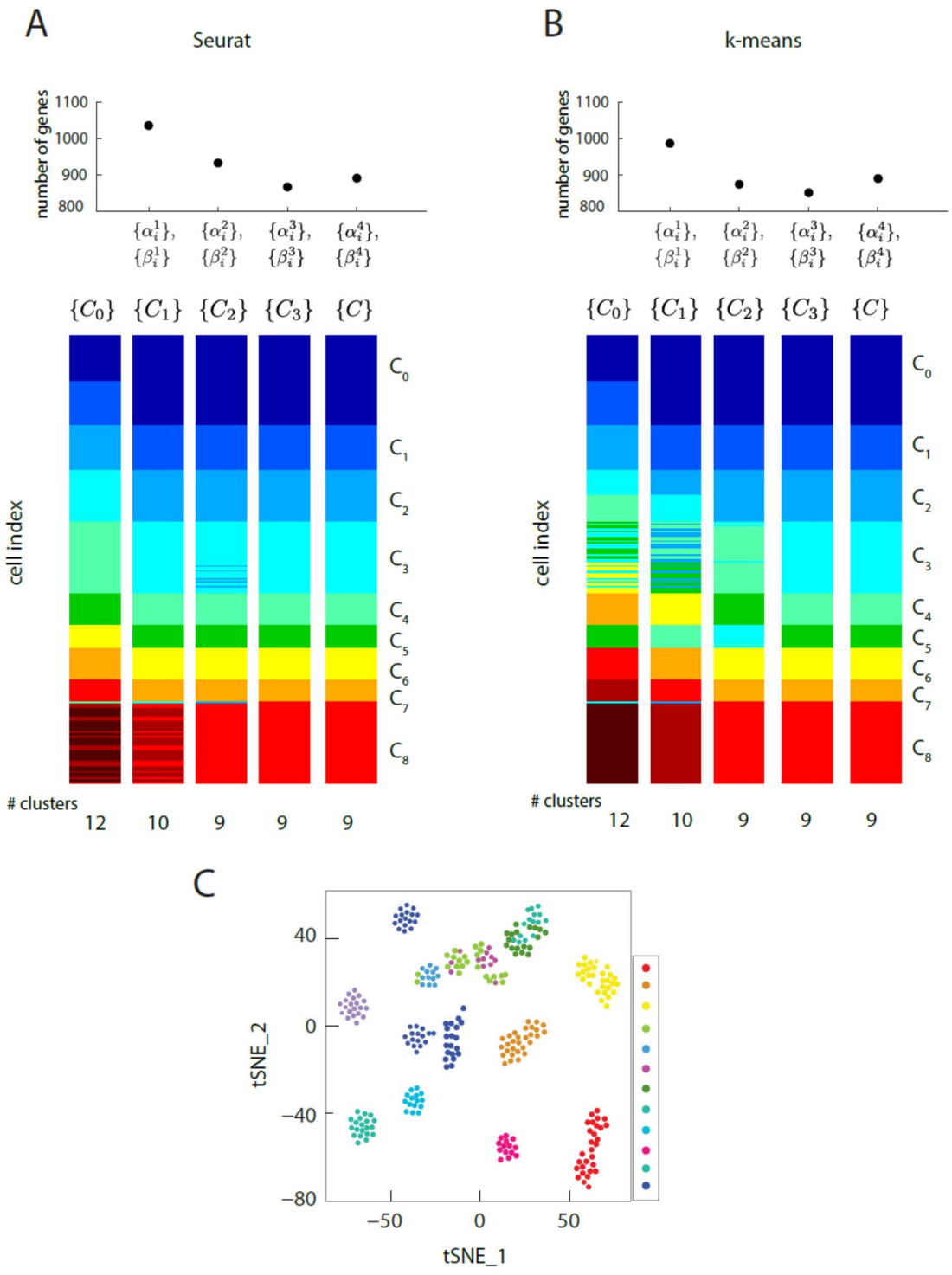


Figure 3.6: Iterative clustering and lineage determination algorithm using two different clustering methods for seed clusters and re-clustering : (A) Seurat, and (B) k-means clustering with the gap statistic. For each clustering method, the seed clusters were computed using the gene expression of all 2,762 transcription factors.

(Figure 3.6, continued) The top plot shows the number of marker and transition genes identified at each iteration and used for the subsequent reclustering. The cluster identities for the single cells in different iterations are represented in different colors below. Despite starting with different seeds, both clustering methods converge to the same cluster identities after 4 iterations. **(**C**)**. Two-dimensional projection of expression data from 288 single cells using Seurat. Each cell (single dot), colored by its cluster identity (there are 12 clusters in total), defined by k-means clustering, showing that t_SNE and k-means clustering lead to different seed cluster configurations.

Similarly, the inferred transitions across all sets of three clusters (Table 3.2) together form a lineage tree (Figure 3.7A) that spans all nine identified cell clusters, which can be visualized in gene expression space through a series of local transition and marker gene classes (Figure 3.7C). We next investigated the gene expression variability among cells within each cluster by performing principal component analysis (PCA) on the transcription factor gene expression for cells within each cluster. We found that for all clusters, no principal component is statistically significant (compared to randomizations of the data; Figure 3.7B), suggesting that within each inferred cluster, the cells have the same identity within the resolution of our data.

The inferred dynamics of differentiation can therefore be visualized in a low-dimensional subspace of gene expression, suggesting that differentiation occurs through a sequence of discrete cell state transitions.

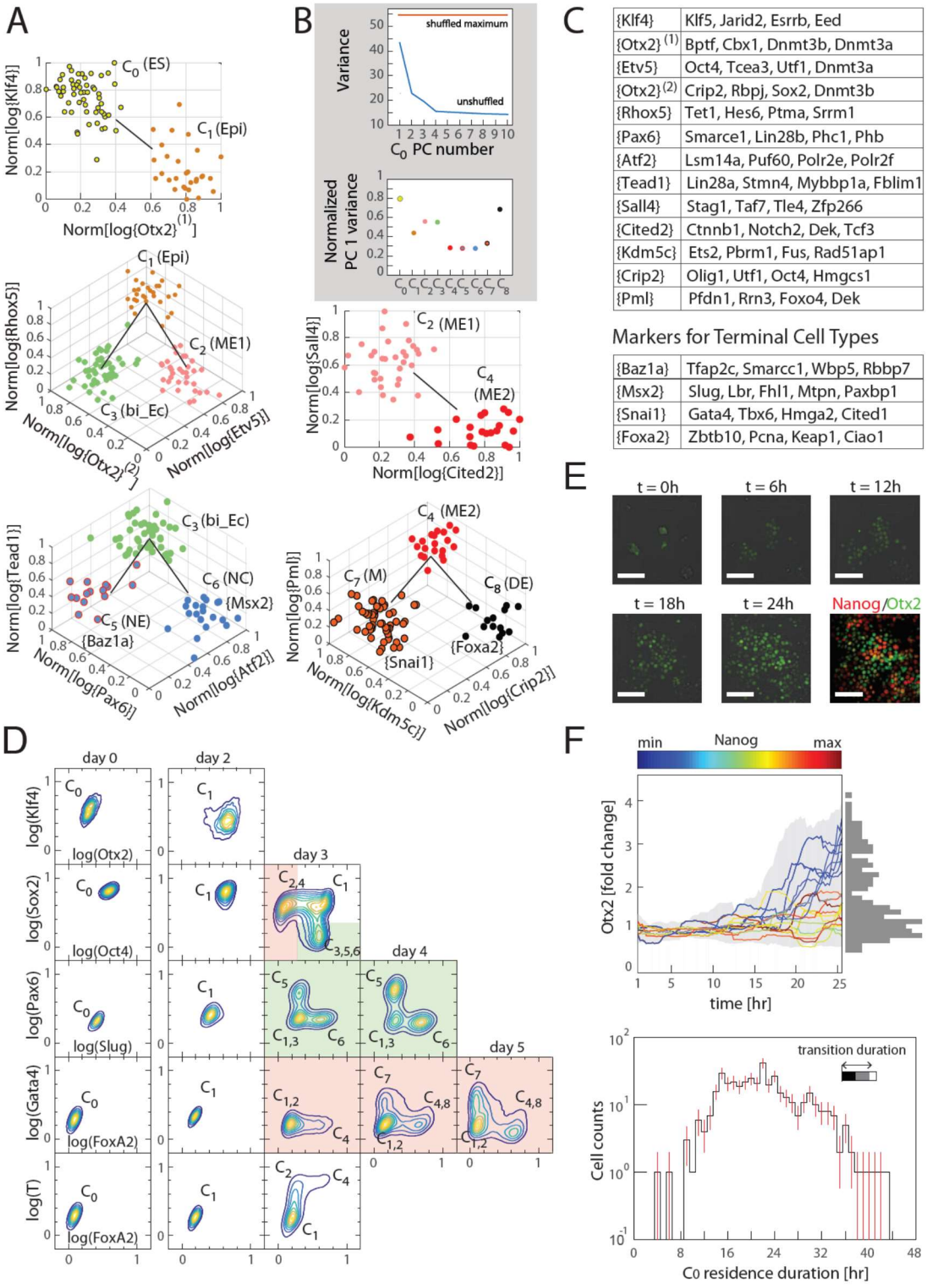


Figure 3.7: Cells transition from one discrete state to another during differentiation. (A). Computationally inferred cell clusters and sequence of transitions are shown in the appropriate subspace of gene expression.

(Figure 3.7, continued) Each dot represents a single cell, and cells are colored based on their cluster identity. For a linear transition sequence of cell states (such as from C_0 to C_1), the transitions are represented in a 2 dimensional plot with the axes defined by the normalized mean log of the unique reads of genes that are most differentially regulated in the two states, while for lineage bifurcations between alternative daughter cell states, the plots are shown in 3 dimensions, where the x and y axes are normalized mean log unique reads of the associated set of transition genes, and the z axes are the normalized mean log unique reads of the marker genes associated with the inferred progenitor state. Labeled in parenthesis next to each cluster are the abbreviated names of the putative corresponding cell types found *in vivo* (Epi: epiblast; bi_Ec: bi-potent ectoderm; ME: mesendoderm; NE: neural ectoderm; NC: neural crest; M: mesoderm; DE: definitive endoderm). (B). Top: Plot of the variances of the first ten principal components of the gene expression of cells in cluster C_0. The red line is the maximum principal component variance over 1000 randomizations of the data, showing that no principal component is statistically significant. Bottom: variances of the first principal component of each cluster, normalized by the maximum principal component variance of the randomized gene expression data for the corresponding cluster. (C). A list of high probability genes that belong to the various marker and transition gene classes that define the axes of the plots in Figure 3.7A, each represented by one gene in curly brackets. The curly brackets contain the gene name with the highest probability for that class, and other high probability genes (as in Figure 3.5A and Figure 3.5B) are listed in the table. While some of the genes are used only once, others such as *Otx2* and *Oct4* are repeatedly reused in different subspaces to describe the transition. (D). Flow cytometry analysis of cell populations sampled every 24 hours during differentiation and immunostained for nine genes (two shown at a time for each density contour plot): Klf4, Otx2, Oct4, Sox2, Slug, Pax6, FoxA2, Gata4 (each taken from a different gene class shown in Figure 4C), and T recapitulate the predicted structure and temporal ordering of transitions through discrete cell states. Axes represent the log of gene expression, normalized by the range between the minimum and maximum across each gene. Plots in pink and green represent C_2 and C_3 lineages following the split from C_1, respectively. (E). Live cell microscopy of *Otx2* reporter (mCitrine) cell line to infer the dynamics of cell state transition from C_0 to C_1. Sample images (shown) at t=0, 6, 12, 18, and 24 hours of differentiation. Cells were terminated at approximately 25 hours into differentiation and immunostained for Nanog (ES marker gene, Figure 3.8A), which shows an anti-correlation between Otx2 and Nanog expression levels. (Scale bar = 100μm) (F). Top: Time series (x-axis) traces of single-cell Otx2 (y-axis) expression dynamics taken every 15 minutes show that the duration of transition from Otx2-low (C_0) to Otx2-high (C_1) is approximately 4 hours, which is well within the time frame of one cell cycle (~10 hours). The end-point (t = 25h) Otx2 levels show a clear separation between high and low (histogram of ~200 cells shown to the right in gray), indicating that some cells have made the transition from C_0 to C_1 while others not. Each trace is colored by its relative end-point Nanog immunofluorescence intensity level. Otx2 levels are normalized by the mean level at t = 0. Bottom: Histogram (y-axis = log (cell count)) of residence durations of ~400 cells in the Otx2-low C_0 state, showing that transition times vary across multiple cell cycle lengths (time lapse length = 48 hours). Inset bar shows mean (gray) as well as upper (white) and lower (Rojas et al.) quartiles of the transition durations of cells.

3.2.3. Correspondence of cell states discovered ab initio from single-cell data to known in vivo cell types

Inspection of the genes that make up the local transition and marker gene classes (Figure 3.7C) allowed us to match clusters to embryonic cell types found *in vivo* that show similar gene expression.

Cluster C_0 is characterized by the high expression of pluripotency genes *Oct4*, *Sox2*, *Sall1*, *Etv5*, *Jarid2*, *Esrrb*, *Klf4* and *Klf5*, whereas cluster C_1 has lower *Jarid2*, *Esrrb*, *Klf4* and *Klf5*, and higher *Otx2*, *Bptf*, *Cbx1* and *Dnmt3a*/*b* expression compared to cluster C_0, suggesting that clusters C_0 and C_1 correspond to naïve ES and primed epiblast pluripotent cell types, respectively (Borgel et al., 2010; Goller et al., 2008; Kim et al., 2001; Nichols and Smith, 2009; Tesar et al., 2007; Zhou et al., 2007).

Clusters C_2 and C_3, which branch out from C_1, show differential expression of pluripotency genes relative to C_1; *Bptf* and *Cbx1* are downregulated in both C_2 and C_3, *Oct4*, *Etv5* and *Dnmt3a* are downregulated in cluster C_3 but maintained in C_2, and *Sox2*, *Otx2* and *Dnmt3b* are downregulated in cluster C_2 but maintained in cluster C_3. Cluster C_2 is further characterized by a high expression level of primitive streak markers *Mixl1* and *T* (Hart et al., 2002; Tada et al., 2005), whereas cluster C_3 is characterized by *Sez6*, *Stmn3* and *Stmn4*, which have recently been shown to characterize the previously elusive mammalian bi-potent ectoderm progenitor population (Li et al., 2015). Together, these patterns strongly suggest that clusters C_2 and C_3 represent mesendoderm and bi-potent ectoderm progenitor cell types, respectively.

The bi-potent ectoderm progenitor-like cluster C_3 is then followed by a lineage split into clusters C_5 and C_6. While *Stmn4* is downregulated in both C_5 and C_6 compared to C_3,

Sez6 is downregulated in only C_5, and *Stmn3* as well as neural progenitor marker Pax6 are downregulated in C_6 but maintained in C_5. Cluster C_5 is further characterized by *Smarce1* and *Zic2*, and cluster C_6 by *Slug* and *Msx2*, suggesting that C_5 and C_6 may be related to neural progenitor and neural crest cells, respectively (Brown and Brown, 2009; Le Douarin, 1991; Vogel-Ciernia and Wood, 2014).

Cluster C_4, although similar in its expression level of *Mixl1* and *T* to cluster C_2, shows higher expression of other primitive streak genes such as *FoxA2* and *Tcf3* (Merrill et al., 2004). Cluster C_4 is then followed by a bifurcation between clusters C_7 and C_8. Cluster C_8 shows high expression levels of *Gata4* and *Snai1*, indicative of its relation to mesoderm, and cluster C_7 is characterized by high *FoxA2* compared to clusters C_4 and C_8, suggestive of its relation to definitive endoderm (Kim and Ong, 2012; Rojas et al., 2005). We predict that cluster C_4 represents a primed bi-potent mesendoderm cell type relative to cluster C_2 (Nakanishi et al., 2009).

Together, these results suggest that the cell clusters and sets of transitions computationally inferred from single-cell transcriptomics data correspond to known *in vivo* cell types and their lineage relationships.

3.2.4. Differentiation occurs through a series of discrete cell state transitions

The fact that gene expression in each cell cluster does not vary significantly allows for genes to be sorted into a few gene classes that show highly correlated expression patterns across clusters (Figure 3.7C). This suggests that one can validate the inferred sequence of cell state transitions and its gene expression dynamics by measuring the expression of one gene from each class in differentiating cells over time.

In order to confirm the gene expression dynamics over the inferred sequence of cell state transitions, we assessed populations of cells for their expression levels of key transition and marker genes (each taken from a different gene class) via immunostaining and flow cytometry. We sampled mES cell populations every 24 hours during differentiation and immunostained each for Klf4, Otx2, Oct4, Sox2, Pax6, Slug, FoxA2, Gata4 and T. (Although *T* is not assigned to a specific gene class, it is highly expressed in the mesendoderm-like states C_2 and C_4, and it thus allows us to distinguish C_2 from the earlier epiblast-like state C_1.) The flow cytometry density contour plots shown (Figure 3.7D) are characterized by high-density peaks which are separated from one another by regions of low density, mirroring the discreteness of the cell states inferred from single-cell transcriptomics data. The relative locations of these high-density peaks and the time at which they appear and disappear recapitulate the inferred gene expression dynamics of the cell state transitions of the lineage tree.

During the first two days of differentiation, all cell populations downregulated Klf4 and upregulated Otx2, as shown in the first row of density contour plots in Figure 3.7D. This is consistent with the first observed state transition in our inferred lineage tree from the naïve ES C_0 state to the primed epiblast-like state C_1. On day three of differentiation (third column of plots in Figure 3.7D), Sox2 and Oct4 are asymmetrically downregulated relative to the preceding population, as is seen in mesendoderm-like state C_2 and bi-potent ectoderm-like state C_3 relative to the epiblast-like state C_1. Sox2-high, Oct4-low cells on day three are either high for Pax6 or for Slug, consistent with comparisons between the neural ectoderm-like state C_5 and neural crest-like C_6. On day 4, the Pax6-high and Slug-high populations become proportionally larger as the Pax6/Slug-low population shrinks,

supporting the inferred temporal ordering that C_5 and C_6 arise from the bi-potent ectoderm-like state C_3. Oct4-high, Sox2-low cells on day three of differentiation are high for T, but show two discrete levels of FoxA2, mirroring the difference between the two mesendoderm-like states C_2 (FoxA2-low) and C_4 (FoxA2-high). Finally, at days four and five, we observe FoxA2-high, Gata4-low and FoxA2-low, Gata4-high cell populations, which correspond to the primed mesendoderm and definitive endoderm-like states C_4 and C_8 and the mesoderm-like state C_7, respectively. We thus confirmed that differentiating cell populations recapitulate the gene expression dynamics of cell state transitions inferred from single-cell data (Figure 3.7A).

The observation that the majority of randomly sampled cells are found to belong to one of nine discrete cell states (both transcriptionally and at the protein level) suggests that cell state transitions occur within a relatively short timeframe compared to the amount of time cells spend within each state. We tested this hypothesis on the first cell state transition from the naïve ES C_0 state to the primed epiblast-like state C_1 (Figure 3.7A). To do so, we generated an *Otx2*-mCitrine fusion protein reporter mES cell line (Figure 3.7C, Methods) and observed the single-cell-resolution dynamics of *Otx2* expression for up to two days (Figure 3.7E and Figure 3.7F).

In agreement with our hypothesis, we observed that Otx2 levels, at the end of 24 hours of differentiation, show a bimodal distribution (Figure 3.7F), and cells tend to occupy either an Otx2-low state (corresponding to ES state C_0) or an Otx2-high state (corresponding to epiblast-like state C_1). We find that cells transition from an Otx2-low to an Otx2-high state well within the duration of a single cell cycle (mean transition duration of 4.52 hours compared to the cell-cycle length of approximately 10 hours). In contrast,

cells tend to stay in either Otx2-low or -high states for up to multiple cell cycles, with a large amount of cell-to-cell variability in the residence duration (Figure 3.7F). Together with our results from the analysis of single-cell transcriptomics data, these observations show that cells reside in discrete states in gene expression space and correspondingly undergo abrupt state transitions.

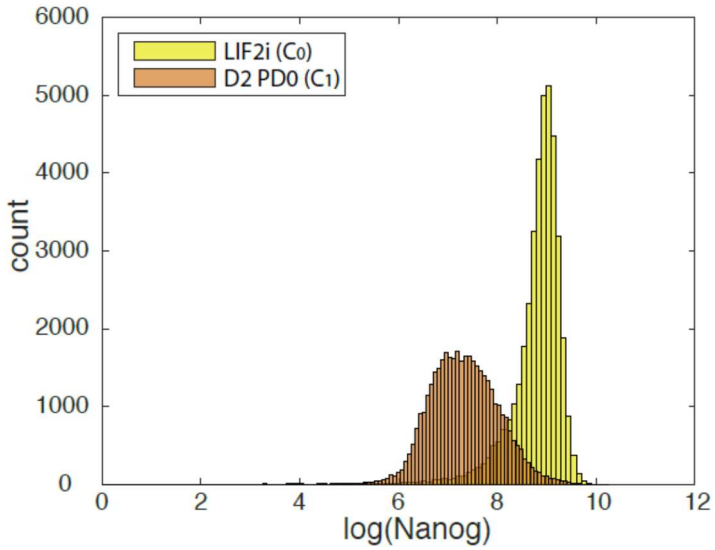


Figure 3.8: Nanog expression before and after differentiation. (A**)**. Histogram of Nanog expression (as measured by immunostaining and flow cytometry) before differentiation (yellow; Lif2i C_0 state) and after two days of differentiation (orange; D2 PD03 C_1 state) shows that Nanog expression is downregulated throughout most of the population during the first two days, similar to the observed changes of Klf4 during this time (Figure 3.7D). We thus use Nanog immunostaining, which produces better signal compared to Klf4 antibodies, to identify cells that are still remaining in the naïve pluripotent C_0 state.

3.2.5. A probabilistic model that replicates the observed discrete cell states predicts state-dependent interpretation of perturbations

Our analysis of single-cell gene expression data revealed a lineage tree composed of discrete cell states, and identified genes associated with individual cell states and transitions between cell states. Our analysis also inferred the coordinate system in which

to visualize these transitions. We next sought to build a predictive quantitative model of the underlying gene regulatory network based on the expression patterns of the marker and transition genes.

Since some transition genes inferred from our Bayesian analysis are re-used to infer multiple local state transitions (Figure 3.7C, e.g., *Oct4*, *Otx2*), we classified transcription factors and signaling genes based on their distinct binarized patterns of expression across all nine cell states, with genes showing the same patterns belonging to the same module (Supplemental Experimental Procedures, Figure 3.9A, Table 3.4). We categorized the 184 marker and transition genes and signaling gene groups into 23 gene modules, each of which showed distinct patterns of expression across the cell states. We denote each gene module by a representative gene in square brackets; for example, the gene module that uniquely characterizes the ES state C_0 is denoted as [*Klf4*] (

Table 3.3).

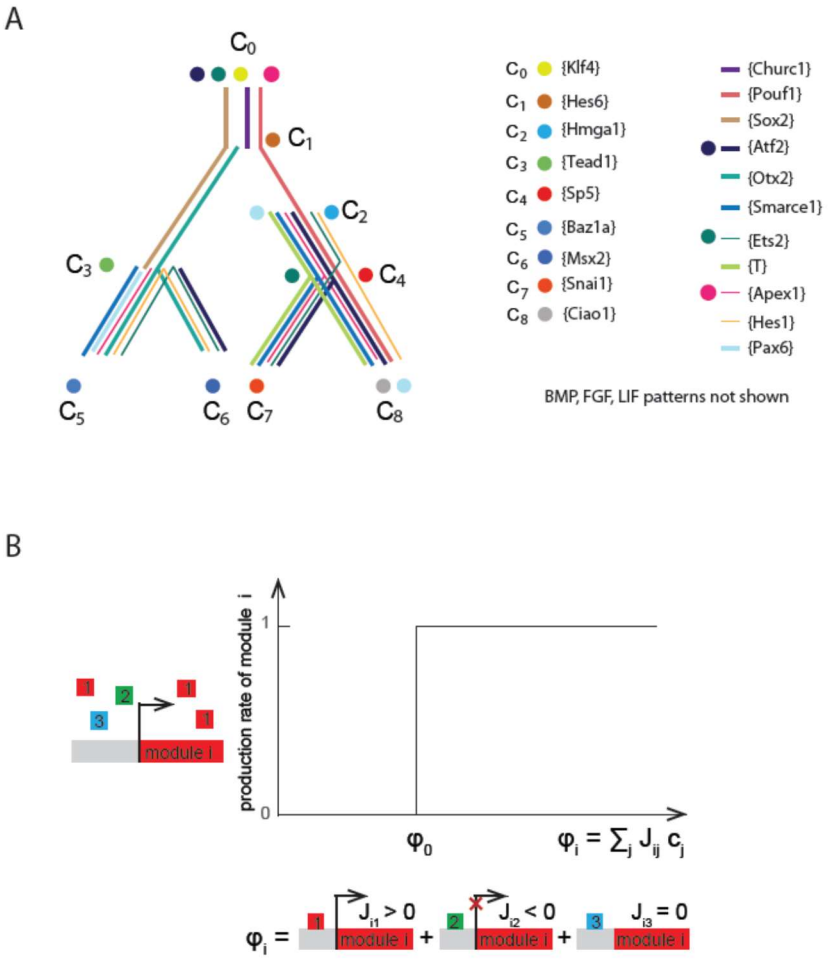


Figure 3.9: Construction of gene regulatory network. **(**A**)**. River diagram of the gene expression patterns of the 23 modules in the 9 cell clusters. Straight lines indicate asymmetric regulation favoring the colored branch; dots indicate symmetric downregulation in the subsequent two branches. **(**B**)**. Plot of the production rate of module $\boldsymbol{i}$, $\boldsymbol{r_i(\vec{m})}$, as a function of the drive from the other modules $\boldsymbol{\phi_i(\vec{m}) = \sum_{j=1}^{N} J_{ij} m_j}$. The production rate is equal to 1 if the drive is greater than a critical drive $\boldsymbol{\phi_0}$ and 0 otherwise.

By construction, any mathematical model of a network between these 23 modules must produce the nine cell states seen in Figure 3.7A. We considered a network that contains direct interactions, in which each module *j* exerts a drive on module *i*, which is equal to an interaction strength J_{ij} (positive or negative) multiplied by the concentration of module *j*. The total drive on module *i* is the sum of the drives from the different modules.

We further considered that the total drive on module i affects expression in a highly non-linear manner, with high gene expression for drives that exceed a critical drive ϕ_0, and low gene expression otherwise (Figure 3.9B). Thus the effective dynamics of expression levels m_i of each module i are given by the non-linear equation:

$$\frac{dm_i}{dt} = H\left(\sum_j J_{ij} m_j - \phi_0\right) - \frac{m_i}{\tau_i} \quad (1)$$

where H is the Heaviside step function and τ_i is the effective lifetime of module i (Supplemental Experimental Procedures).

We determined the set of interactions J_{ij} that are consistent with the observed cell states (C_0-C_8, Figure 3.7A) being stable fixed points of the network. If state $\vec{m}^\alpha = \{m_1^\alpha, \ldots, m_{23}^\alpha\}$ with expression level m_i^α in module i is a stable fixed point of the network, then the interactions J_{ij} must be such that the total drive on each module that is expressed in $\vec{m}^\alpha$ is greater than the critical drive, and the total drive on each module that is not expressed in $\vec{m}^\alpha$ is less than the critical drive:

$$m_i^\alpha = 1 \Rightarrow \sum_j J_{ij} m_j^\alpha \geq \phi_0,$$

$$m_i^\alpha = 0 \Rightarrow \sum_j J_{ij} m_j^\alpha < \phi_0. \quad (2)$$

Thus, for each stable state, we have 23 constraints on the possible values of J_{ij}, one for each module. Given that we have nine cell states, there are 23*9 = 207 inequalities that constrain the values of the 23^2 = 529 different parameters, J_{ij}. The problem is therefore underdetermined, and there are an infinite number of solutions that would allow for the observed cell states to be stable.

To overcome this problem, we exploited recent developments based on renormalization group approaches to determine which microscopic variables are relevant for the observed data (Machta et al., 2013). By using a linear programming method to obtain 10,000 sets of J_{ij} interactions (Supplemental Experimental Procedures), each satisfying the constraint that all nine cell states are stable fixed points, we estimated the probability distribution for the 529 parameters of the model, as shown in Figure 3.10A, giving us a probabilistic model of the underlying network.

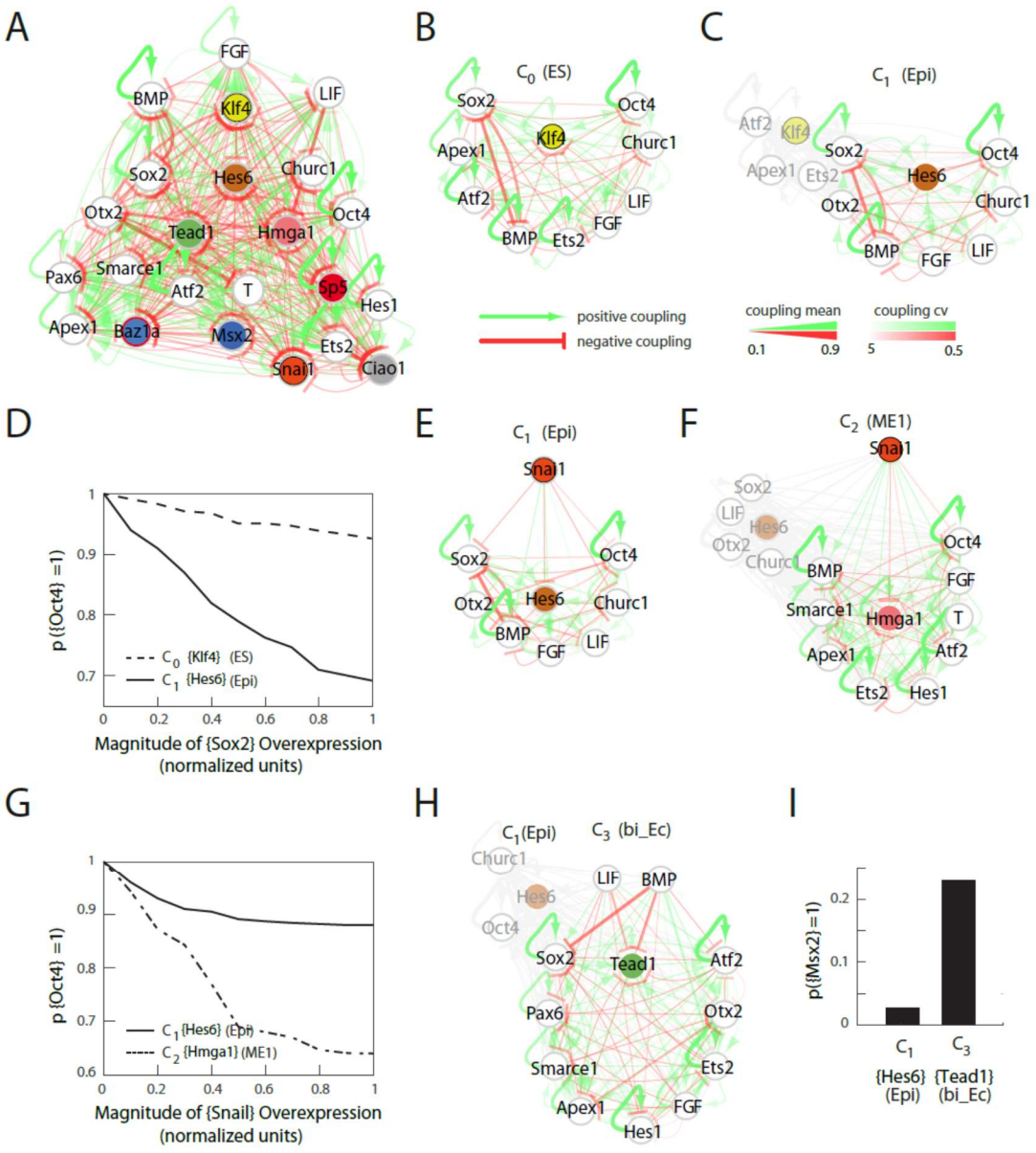


Figure 3.10: Quantitative modeling of the network underlying germ layer differentiation.

(Figure 3.10, continued) (A). The inferred gene regulatory network from 10,000 sampled solutions that stabilize each of the nine cell states. Each circle represents a gene module. Mean positive and negative interactions between the modules are shown in red and green, respectively, and their thickness and transparency are proportional to the absolute magnitude of the mean and the coefficient of variation (c.v.), respectively. The colored circles represent the gene modules expressed uniquely in only one of the cell states (color code matched with Figure 3.7A for each state). (B, C). Subsets of the network consisting of gene modules that are expressed in (and stabilize) the naïve ES C_0 state (B) and epiblast-like C_1 (C) state. As cells transition from C_0 to C_1, expression of [*Klf4*], [*Apex1*], [*Ets2*], [*Atf2*] modules is downregulated (shown in gray) while [*Hes6*] and [*Otx2*] modules are upregulated, leading to changes in the effective interaction strengths between gene modules that are common to both C_0 and C_1 states, such as [*Sox2*] and [*Oct4*]. (D). [*Sox2*] overexpression (x-axis) plotted against the probability of [*Oct4*] downregulation (y-axis) computed over 10,000 models (Supplemental Experimental Procedures). In the C_1 state (solid line), [*Oct4*] is downregulated in an increasing fraction of models following [*Sox2*] overexpression, while in C0, [*Oct4*] is stable in >95% of the models (dotted line). (E, F). Subsets of the model consisting of gene modules that are expressed in the epiblast-like C_1 (E) and mesendoderm-like C_2 (F) states, and their interactions with [*Snai1*], which is not normally expressed in C_1 or C_2. As cells transition from the C_1 to C_2 state, [*Hes6*], [*Sox2*], [*Otx2*], [*Churc*] are downregulated (shown in gray), while [*Hmga1*], [*T*], [*Atf2*], [*Hes1*], [*Ets2*], [*Apex1*], and [*Smarce1*] are upregulated, leading to changes in the effective interaction strengths between [*Snai1*] and modules that are common to both C_1 and C_2, such as [*Oct4*]. (G). The probability of [*Oct4*] being downregulated (y-axis) as a function of [*Snai1*] overexpression (x-axis). In the C_1 state (blue line), the over expression of [*Snai1*] has no effect on [*Oct4*] levels in ~90% of the 10,000 models whereas in the C_2 state (red line), the overexpression of [*Snai1*] leads to [*Oct4*] downregulation in up to 35% of the models. (H). The C_3 state shows a downregulation of [*Oct4*] and [BMP], and upregulation of [*Tead1*], [*Apex1*], [*Pax6*], [*Smarce1*], [*Ets2*], [*Atf2*], [*Hes1*] modules relative to C_1. (I). Cells in different states are predicted to respond differently to morphogens. Plot showing the percentage of models (y-axis) where states C_1 and C_3 (x-axis) transition to C_6 (characterized by unique marker gene module [*Msx2*]), in response to [LIF]+[BMP]. C_1 cells remain stable in response to [LIF]+[BMP] signaling in >97% of the models whereas C_3 cells are destabilized and move to the C_6 state in 23% of the models.

To functionally validate the cell states and the gene expression dynamics of cell state transitions inferred from our single-cell data, we used this probabilistic model to make testable predictions as to how different cell states respond to perturbations. We found that the individual cell states not only have distinct transcriptional profiles but are also predicted to have distinct phenotypic responses to the same perturbations.

We made three probabilistic predictions that we experimentally tested, each probing different aspects of the model gene regulatory network. We first considered changes in the effective interaction strengths between two gene modules as a function of cell state. To this end we looked at two kinds of gene module pairs: (i) gene modules that are co-expressed in two cell states and (ii) gene modules that are never co-expressed in any cell state.

Gene modules [*Sox2*] and [*Oct4*] are highly expressed in both the ES cluster C_0 and the epiblast-like C_1 cluster, after which they are asymmetrically downregulated in the mesendoderm-like C_2 and ectoderm-like C_3. We find that for 81% of the 10,000 sampled solutions, [*Sox2*] and [*Oct4*] have mutually inhibitory interactions (i.e., negative coupling constants). However, their effective interactions are altered in different ways in each cell state by the presence of other gene modules. As cells transition from state C_0 to C_1, they downregulate gene modules [*Klf4*], [*Atf2*], [*Apex1*] and [*Ets2*], and upregulate [*Hes6*] and [*Otx2*], among others (Figure 3.10B and Figure 3.10C), leading to changes in the effective interaction strength between [*Sox2*] and [*Oct4*]. By incrementally increasing [*Sox2*] levels and assessing the fraction of models that show [*Oct4*] downregulation, we found that [*Oct4*] levels are predicted to be more stable to [*Sox2*] overexpression in state C_0 than in C_1 (Figure 3.10D).

On the other hand, [*Snai1*] and [*Oct4*] are not expressed together in any of the nine cell states. We investigated the predicted effects of [*Snai1*] overexpression on [*Oct4*] in the epiblast-like state C_1 and mesendoderm-like state C_2, both of which normally express [*Oct4*] but not [*Snai1*]. Although [*Snai1*] has a negative interaction with [*Oct4*] in 87.5% of the models, the modules expressed in C_1 exert a greater positive drive on [*Oct4*] (Figure 3.10E

and Figure 3.10F) than those expressed in C_2. This leads to the prediction that [*Oct4*] is less sensitive to [*Snai1*] overexpression in state C_1 compared to C_2 (Figure 3.10G).

We next considered the effect of morphogen signals in different states. Specifically, we considered the LIF, BMP, WNT and FGF signaling pathways, which are known to play a significant role in patterning the early embryo. We grouped signaling genes by their respective pathways (defined by GO categories) and assigned each group to a module based on its average expression pattern across the nine cell states. Because WNT and FGF modules show no changes in expression across all cell states, we focused on investigating the effects of LIF and BMP signaling on cells in the epiblast-like C_1 and in the bi-potent ectoderm-like state C_3 (Figure 3.10H). Given an initial state C_1 or C_3, we calculated the probabilities that cells either remain in the same state or move to a different state in response to [LIF] and [BMP] (Methods). Our simulations found that cells that are initially in state C_1 either remain stabilized in C_1 or move to state C_0 in response to [LIF] and [BMP], with a probability of 82.4% and 15.8%, respectively. However, in response to the same perturbation, cells in the C_3 state transitioned to the neural crest-like state C_6 state with a probability of 19.6%, and remained in the C_3 state in 77.3% of the models (Figure 3.10I). To summarize, we predict that [*Oct4*] expression is less sensitive to [*Sox2*] overexpression in state C_0 than in C_1; [*Oct4*] expression is less sensitive to [*Snai1*] overexpression in state C_1 compared to C_2; and cells in state C_3, but not in C_1, can transition to state C_6 following [LIF]+[BMP] exposure.

Thus, by categorizing genes into different modules by their expression patterns across the observed cell states, these modules provide a starting point for modeling the gene regulatory network responsible for cell-fate decisions, allowing us to make and test

predictions for how the network gives rise to different phenotypic responses to perturbation across different cell states.

3.2.6. Interpretation of Sox2, Snai1, and LIF+BMP are cell state dependent

We next experimentally tested the model's predictions of state-dependence in cells' responses to perturbations. We first tested how cells' Oct4 levels respond to *Sox2* overexpression in the naïve ES and epiblast-like states C_0 and C_1. We transiently transfected cells with a plasmid containing a Tet-inducible bi-directional promoter, flanked by the open reading frames of *Sox2* and mCerulean, which we used as a fluorescent reporter of induction (Figure 3.11A). We induced overexpression in cells either in the undifferentiated C_0 state or the epiblast-like C_1 state (Figure 3.11B). As a control, we used identical populations that were transfected with a plasmid containing only mCerulean under the inducible promoter. In such experiments, we typically saw mCerulean fluorescence appear approximately three hours into induction and persist for about three to four days after transfection. We therefore induced overexpression for 24 hours to minimize the effect of plasmid loss but still allow for several cell cycles to occur during induction. Following induction, we fixed and immunostained the cells for Oct4, and analyzed the results via flow cytometry. In agreement with our predictions (Figure 3.10D), we found that *Sox2* overexpression correlates with downregulation of Oct4 in the epiblast-like state C_1 (significant relative to control, $p = 5.72 \times 10^{-31}$; see also Figure 3.11C), whereas this effect was not observed in undifferentiated cells (state C_0) (Figure 3.12A and Figure 3.12B).

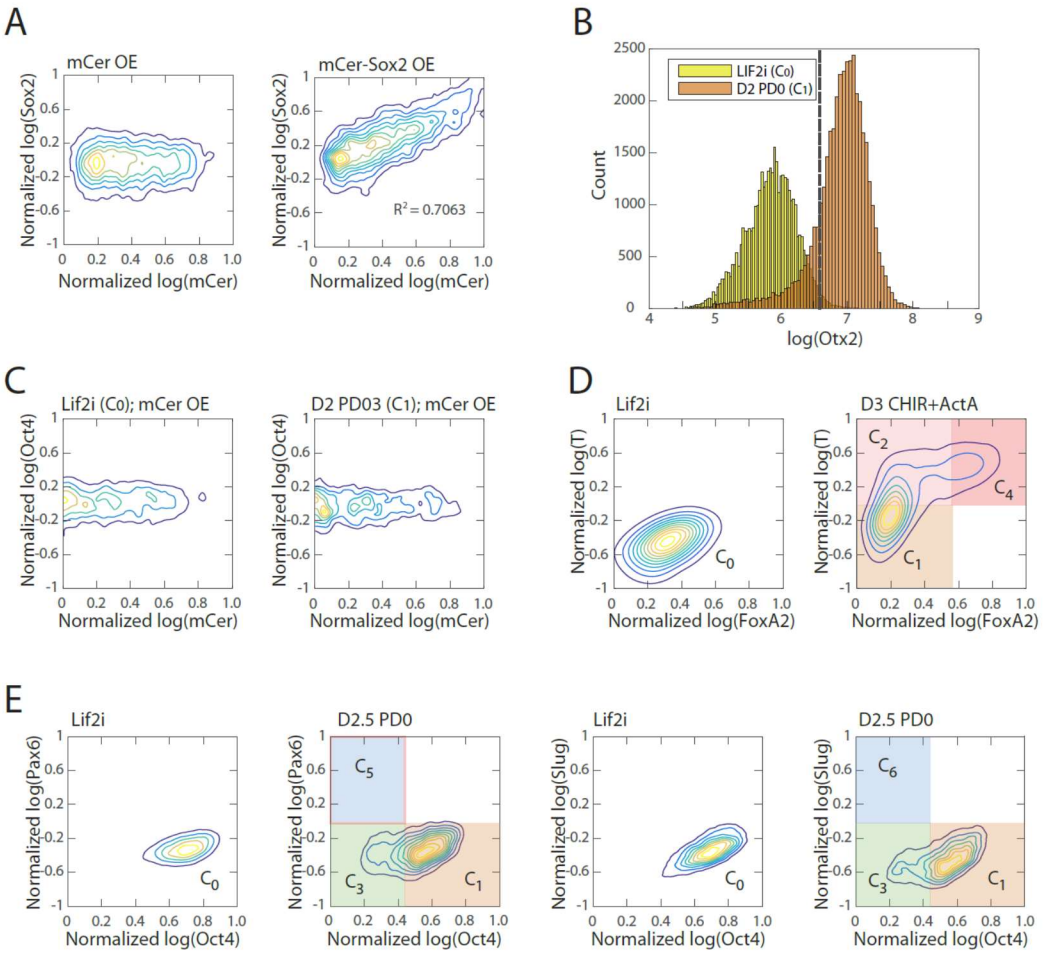


Figure 3.11: Overexpression experiments. **(**A**)**. (Right) mCerulean fluorescence level correlates with increasing total Sox2 levels, validating the use of mCerulean fluorescence as a measure for Sox2 overexpression. (Left) As a control, we show that Sox2 levels do not increase when only mCerulean is overexpressed. **(**B**)**. Histogram of Otx2 expression (as measured by immunostaining and flow cytometry) following 24 hours of Sox2 overexpression in either naïve ES C_0 cells (Lif2i; yellow) or epiblast-like C_1 cells (D2 PD0; orange). In determining the effects of Sox2 overexpression in the epiblast-like C_1 state (Figure 6A), we excluded cells that showed Otx2 expression less than two standard deviations above the mean of Otx2 levels in Lif2i (threshold shown in dotted line). **(**C**)**. Overexpression of only mCerulean does not show any effect on Oct4 levels in both C_0 (left; Lif2i) and C_1 (right; D2 PD0) cell states. **(**D**)**. At day 3 of differentiation using CHIR99021 and Activin A (Methods), populations consist of cells in C_1 (FoxA2-low, T-low) C_2 (FoxA2-low, T-high) and C_4 (FoxA2-high, T-high) states (right). The fraction of cells in C_4 is 17%. FoxA2 and T levels in undifferentiated C_0 state cells (Lif2i) are shown as a reference (left).

(Figure 3.11, continued) **(**E**)**. At day 2.5 of differentiation using PD0325901 (Methods), populations consist of cells in C_1 (Oct4-high) and C_3 (Oct4-low) states. At this point cells have not yet upregulated Pax6 (left two panels) or Slug (right two panels), showing that cells have not yet transitioned to either state C_5 or C_6.

We then tested the effects of *Snai1* overexpression on Oct4 in the epiblast-like state C_1 and mesendoderm-like state C_2, using the same experimental framework as described above. On day three of differentiation, cell populations either contain a mixture of C_1, C_2 and (minimally) C_4 cell states, or a combination of C_1, C_3 and C_5 (or C_6), depending on the signaling conditions (Figure 3.7D; see also Figure 3.11D). Using the signaling conditions that yield the former set of cell states, we transfected cells at 2.5 days into differentiation, and drove overexpression of *Snai1* 12 hours later in a population consisting primarily of cells in C_1 and C_2 states (Figure 3.12D). After 24 hours of *Snai1* overexpression and further differentiation, we fixed and immunostained the cells for T to distinguish cells in C_1 (T-low) and C_2 (T-high) states. We also immunostained the cells for Oct4 to distinguish the C_1 state from other T-low states that arise during the last 24 hours of differentiation following the initiation of induction. We found that the fraction of C_1 cells within the transfected population was significantly reduced relative to control ($p = 1.98 \times 10^{-13}$), suggesting that cells in this state had downregulated Oct4 levels in response to *Snai1* overexpression. On the other hand, the fraction of C_2 cells within the transfected population and their Oct4 levels were maintained relative to control, in agreement with our predictions (Figure 3.10G, Figure 3.12C, and Figure 3.12D).

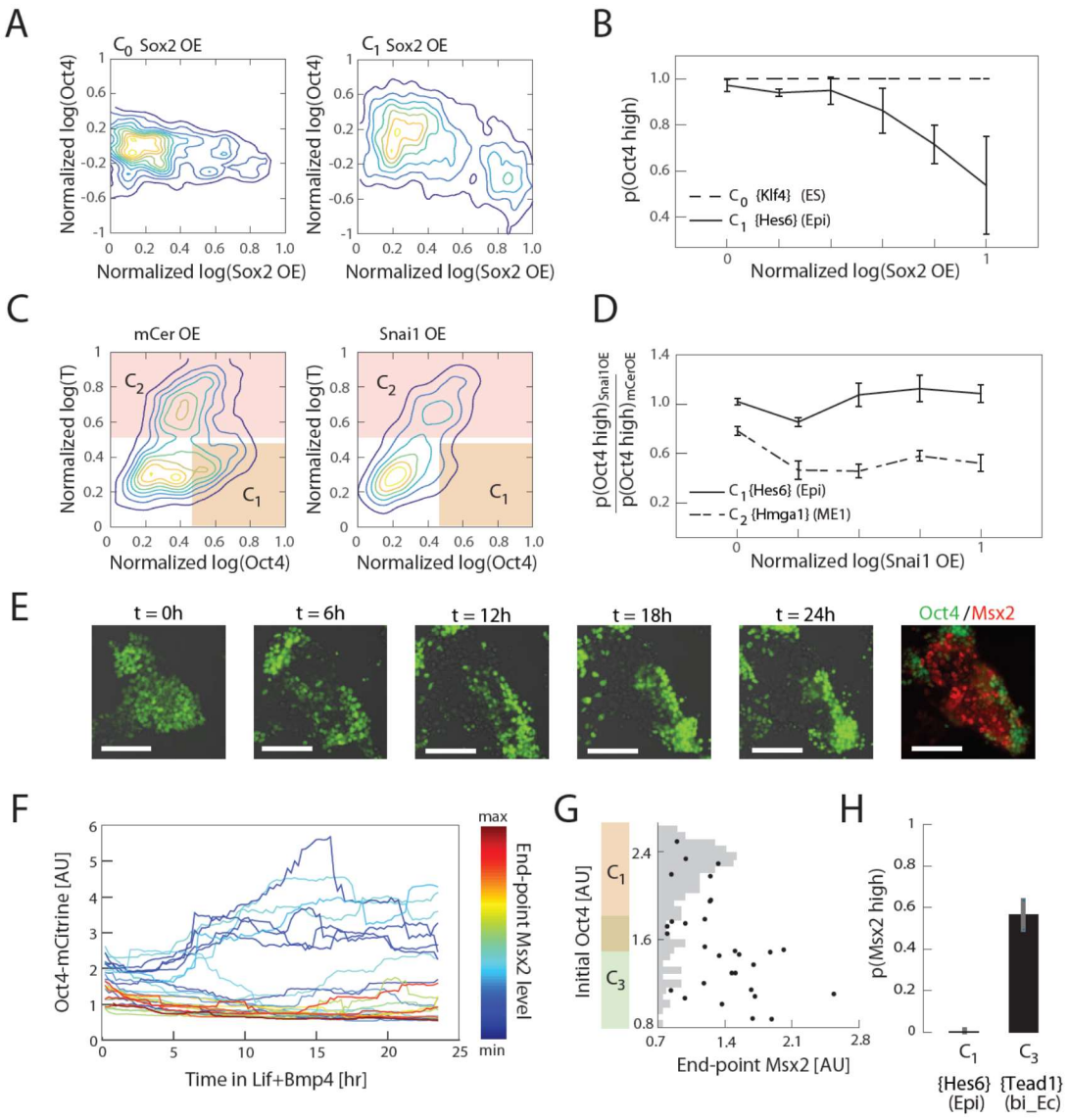


Figure 3.12: Experimental validation shows that interpretation of *Sox2*, *Snai1*, and LIF+BMP is cell state dependent. (A). Comparison of the effects of *Sox2* overexpression (x-axis) on Oct4 levels (y-axis) in the naïve ES state C_0 (left) and epiblast-like C_1 (right) state shows negative correlation between *Sox2* overexpression and Oct4 levels in the C_1 state, but not in C_0. Plots showing mCerulean (marker) -only overexpression in C_0 or C_1 are indistinguishable from *Sox2* overexpression in C_0 (Figure 3.11C) (B). Fraction of Oct4-high cells (y-axis; defined as greater than 2σ below the mean log of Oct4 of non-transfected control cells) plotted against binned *Sox2* overexpression level confirms model prediction (Figure 3.10D) that *Sox2* overexpression leads to downregulation of Oct4 in C_1 but not C_0. (C). Comparison of the effects of *Snai1* (right) and mCerulean-only (left) overexpression on Oct4 levels (x-axis) in the epiblast-like C_1 and mesendoderm-like C_2 states (y-axis; T-low and –high, respectively) shows downregulation of Oct4 in response to *Snai1* overexpression in the C_1 state but not in C_2.

(Figure 3.12, continued) (D). Fraction of Oct4-high cells in *Snai1* overexpressing cells, normalized by this fraction in mCerulean overexpressing control cells (y-axis), plotted against binned *Snai1* overexpression level (x-axis) confirms the prediction (Figure 3.10G) that *Snai1* overexpression leads to greater downregulation of Oct4 in C_1 compared to C_2. (E). Live cell images of Oct4-mCitrine cells at t= 0, 6, 12, 18, 24 hours of LIF+BMP exposure. At t= 0, cells are either in state C_1 (Oct4-high) or C_3 (Oct4-low) (Figure 3.11E). (Scale bar = 100μm) Cells were fixed at t=24 hours and immunostained for *Msx2*. (F). Time series (x-axis) traces of single-cell Oct4 expression (y-axis) taken every 15 minutes from live cells. Each trace is colored by its relative end-point *Msx2* immunofluorescence intensity level. (G). The initial Oct4 reporter (mCitrine) intensity (y-axis) and final Msx2 immunofluorescence (x-axis) are negatively correlated. Each dot represents a single cell. Histogram of Oct4 reporter intensity at t = 0 levels shown in gray. Based on this histogram, we defined a range of threshold values for determining Oct4-high and –low (shown in overlapping region of orange and green along y-axis). (H). Plot showing fraction of Msx2-high (y-axis; as defined by greater than 2σ above background) confirms prediction (Figure 3.10I) that Msx2 is upregulated with greater probability in the C_3 state compared to C_1 (x-axis) in response to LIF+BMP exposure.

Finally, we tested whether cells in epiblast-like C_1 and bi-potent ectoderm-like C_3 states respond differently to LIF+BMP signaling, as predicted by our model. In order to investigate the relationship between a cell's initial state and its final state in response to LIF+BMP exposure, we needed to assess cells' initial states non-invasively. We found that 2.5 days into differentiation, we could obtain populations that consist primarily of cells in epiblast-like state C_1 and bi-potent ectoderm-like state C_3 (Figure 3.11E), which have high and low expression of Oct4, respectively. We therefore utilized an Oct4-mCitrine mES cell line that we had previously engineered (Thomson et al., 2011) to distinguish cells in C_1 and C_3 states after 2.5 days of differentiation. At this point, 1200U/mL LIF and 25ng/mL BMP4 were added to the media, after which we followed individual cells' Oct4 expression dynamics for approximately 24 hours via live-cell microscopy, followed by fixing and immunostaining for Msx2, a unique marker gene for the neural crest-like cell state C_6 (Figure 3.12E and Figure 3.12F). As predicted by the model (Figure 3.10I), only cells that had low Oct4 levels (and were therefore in the bi-potent ectoderm-like state C_3) prior to

LIF+BMP exposure showed upregulation of Msx2 in response to LIF+BMP (Figure 3.12G and Figure 3.12H). Together, these results show that the inferred cell states reflect phenotypic discreteness in cells' responses to perturbations, and that the gene expression changes that define these responses mirror those predicted by our model gene regulatory network.

3.3. Discussion

In this study, we find that the challenges of clustering single-cell gene expression data to determine cell states, inferring the sequence of transitions between these states, and discovering the genes whose expression patterns best reflect these states and state transitions are intricately linked. However, by simultaneously inferring cell states, state transitions and the coordinate system that defines cell states and state transitions, we can analyze single-cell data to uncover the gene expression dynamics of differentiation. We also demonstrate that building a family of models to fit the data allows us to obtain the probability distributions of all the gene interaction parameters, which can be used to make experimentally testable predictions. We believe that both our Bayesian framework to infer the dynamics of cell state transitions as well as our modeling framework will be broadly useful to obtain a quantitative understanding of development from single-cell gene expression data.

Comprehensive interrogation of gene expression through RNA sequencing is impossible without the termination of cells, providing only static snapshots of gene expression during differentiation. Despite this and the complexity of the underlying network, we discover that both cell states and the sequence of cell state transitions can be

accurately determined by monitoring the levels of just a few transition or marker genes. Monitoring the expression dynamics of these key genes in live cells using microscopy will allow us in the future to continuously track the cell-fate decisions of individual cells. The inferred gene modules therefore represent the "order parameters" by which cell-state transition dynamics can be directly measured. Live cell microscopy experiments will also allow us to measure, in conjunction with cell state transition dynamics, changes in individual cells' spatial environment, movement, lineage history, and cell cycle dynamics in order to address fundamental biological questions as to how these factors affect cell fate decisions.

Requiring models to have discrete cell states leads to the prediction that each cell state has distinct responses to perturbations by signals and changing levels of gene expression. Our experimental tests show, as predicted by the model network, that *Oct4* is either downregulated or unaffected by overexpression of *Sox2* or *Snai1*, depending on the cell state. Previous studies have already shown that *Sox2* and *Oct4*, along with *Klf4*, constitute part of a positive feedback loop that stabilizes the pluripotent ground state (Kim et al., 2008; Young, 2011). It is also known that in undifferentiated cells, *Snai1* overexpression leads to downregulation of *Oct4* expression and, subsequently, to exit of pluripotency (Galvagni et al., 2015). However, our results demonstrate that these interactions are state-dependent by showing that the effective positive interactions between *Sox2* and *Oct4* become destabilized as *Klf4* levels drop and cells transition to a primed, epiblast-like pluripotent state. Similarly, the negative interaction exerted by *Snai1* on *Oct4* becomes attenuated in the presence of early primitive streak genes such as *T*. We also predict and show that LIF+BMP exposure pushes bi-potent ectoderm-like cells toward an

Msx2-positive neural crest-like state, but this effect is not seen in epiblast-like cells. These results are further supported by the fact that both LIF and BMP signaling pathways can be used to keep cells in the pluripotent cell state (Chambers, 2004; Tam et al., 2006; Ying and Smith, 2003), and that BMP signaling plays a significant role in the differentiation of neural crest cells (Knecht and Bronner-Fraser, 2002). Together, these findings signify that the inferred cell states directly reflect differences in cells' responses to perturbations and show that these cell states can also be defined by their unique responses to perturbations.

Finally, our results suggest that cell-to-cell heterogeneity within differentiating populations arises largely as a consequence of cells' variability in their timing of cell state transitions. Our inferred cell clusters show mixing of cells from different time points, suggesting that the observed states themselves do not change over time and that at the population level, differentiation occurs as a change in the proportions of cells in various cell states rather than through changes in the cell states themselves (Figure 3.7D). Since cells interpret perturbations differently even in consecutive states (Figure 3.12), this suggests that heterogeneity arising from timing variability is further amplified in response to signal addition or fluctuations in gene expression level. These findings emphasize the importance of understanding how the timing of cell state transitions is controlled during development.

3.4. Methods

3.4.1. ES-Cell Culture

v6.5 mouse embryonic cells were maintained and passaged in monolayer (non-embryoid body formation) in N2B27 basal media with signaling molecules and/or small molecules

added to the basal media. ES cells were maintained in a pluripotent cell state using 1200U/mL mLIF (murine leukemia inhibitory factor), 1μM PD0325901 (MEK inhibitor), and 3μM CHIR99021 (GSK inhibitor) conditions (a.k.a. "LIF+2i"; (Ying et al., 2008), and passaged every two days. To passage cells, we added 0.01% trypsin to cells after aspirating media and incubated the plate in 37'C for 1 ~ 2 minutes to detach cells. The trypsin was then quenched with 0.5mL of fetal bovine serum, and the resulting cell suspension was collected, counted, and pelleted at 200xg for 5 minutes at room temperature. The supernatant was aspirated and the cells were resuspended and re-seeded onto a gelatinized tissue culture dish at a density of 1e6 cells per 10cm diameter plate. All cell lines were depleted of feeders and transitioned to serum free medium over several passages prior to experiments (Ying and Smith, 2003). N2B27 is prepared as described in (Gaspard et al., 2008; Ying and Smith, 2003).

3.4.2. ES Cell differentiation

Cells were seeded at a density of 10^6 per 10cm diameter plate, and were not trypsinized again until they were harvested for analysis. We either exposed cells to 0.4μM PD0325901 or 3μM CHIR99021 and 10ng/mL Activin A (human, rat, mouse) for 2 days or 3 days, respectively, followed by either 25ng/mL *hBmp4* or 1μM LDN193189 (BMP antagonist) for up to two days. Media was replenished every 48 hours. Cells exposed to 0.4μM PD0325901 gave rise to ectodermal lineages, as characterized by expression of *Sox1*, *Pax6* (treated with LDN193189), *Slug*, and *Msx2* (treated with *hBmp4*) after three days of differentiation. Cells exposed to CHIR99021 and Activin A gave rise to mesendodermal lineages (Sumi et al., 2008), as characterized by expression of *T* after three days of

differentiation, and *FoxA2* (treated with LDN193189) and *Gata4* (treated with *hBmp4*) after four days of differentiation.

Table 3.1: Differentiation conditions and duration of single cells sorted into seven 96-well plates

	plate M1	plate M2	plate M3	plate M5	plate M6	plate M7	plate M8
row A	Lif2i	Day 1 ChAct	Day 2+2 PD0+LDN	Day 1 ChAct	Day 3+1 ChA+LDN	Day 3+2 ChA+LDN	Day 3+2 ChA+LDN
row B	Lif2i	Day 1 ChAct	Day 2+2 PD0+LDN	Day 1 ChAct	Day 3+1 ChA+LDN	Day 3+2 ChA+LDN	Day 3+2 ChA+LDN
row C	Day 1 PD0	Day 3 ChAct	Day 2+2 PD0+LDN	Das 3 ChAct	Day 3+1 ChA+LDN	Day 3+2 ChA+LDN	Day 3+2 ChA+LDN
row D	Day 1 PD0	Day 3 ChAct	Day 2+2 PD0+LDN	Day 3 ChAct	Day 3+1 ChA+LDN	Day 3+2 ChA+LDN	Day 3+2 ChA+LDN
row E	Day 2 PD0	Day 2+1 PD+LDN	Day 2+2 PD0+Bmp	Day 2+1 PD+LDN	Day 3+1 ChA+Bmp	Day 3+2 ChA+Bmp	Day 3+2 ChA+Bmp
row F	Day 2 PD0	Day 2+1 PD+LDN	Day 2+2 PD0+Bmp	Day 2+1 PD+LDN	Day 3+1 ChA+Bmp	Day 3+2 ChA+Bmp	Day 3+2 ChA+Bmp
row G	Day 2 ChAct	Day 2+1 PD+Bmp	Day 2+2 PD0+Bmp	Day 2+1 PD+Bmp	Day 3+1 ChA+Bmp	Day 3+2 ChA+Bmp	Day 3+2 ChA+Bmp
row H	Day 2 ChAct	Day 2+1 PD+Bmp	Day 2+2 PD0+Bmp	Day 2+1 PD+Bmp	Day 3+1 ChA+Bmp	Day 3+2 ChA+Bmp	Day 3+2 ChA+Bmp

3.4.3. Single-Cell RNA-Seq

CEL-seq libraries as previously reported (Hashimshony et al., 2012) with a few modifications. Single cells were sorted with a FACSAria (BD) into 96 well plates containing 1.2 µL 2 × CellsDirect Buffer (Life Technologies) with 0.1 µL of ERCCs diluted to 1×10^{-6} molecules (Life Technologies). Plates were frozen and stored at -80°C.

For library preparation, mRNA was reverse transcribed using 0.15625 pmol of oligoT primer carrying a cell-specific 8 NT barcode and a 5 NT unique molecular identifier (UMI) (Islam et al., 2014). Barcode design ensured at least two nucleotide differences from any other barcode. Samples were lysed at 70 °C for 5 minutes, then reverse transcribed using Superscript III for two hours at 50 °C, then primers digested with 1 µL of ExoSAP-IT (Affymetrix). Second strand synthesis was carried out with Second Strand Synthesis Buffer, dNTPs, DNA Polymerase, and RNAse H (NEB) at 16 °C for 2 hours. Single cell cDNAs were pooled by 24 wells per library, with each library containing a water-only well and one ERCC-only well. Pools were purified with an equal volume of RNA Clean Beads (Beckman Coulter) and amplified at 37°C for 15 h using the HiScribe T7 High Yield RNA Synthesis kit (NEB), and treated with DNAse I (Life Technologies). Amplified RNA was fragmented using the NEBNext RNA Fragmentation Module (NEB), purified with an equal volume of RNA Clean Beads, and visualized using the RNA Pico Kit on the Bioanalyzer 2100 (Agilent). The RNA fragments were repaired with Antarctic Phosphatase and Polynucleotide Kinase (NEB), and purified using an equal volume of RNA Clean Beads. cDNA libraries were made using the NEBNext Small Library Prep Kit according to the manufacturer's instructions, except Superscript III was used for the RT step. Index primers were used in PCR amplification. Approximately 160-200 nmol of a pool of libraries were size selected to exclude species smaller than 180 bp on a 2% Dye Free cassette on the Pippin Prep (Sage) and concentrated to approximately 14 µL. Pools were then quantified by qRT-PCR using p5 (5'-AATGATACGGCGACCACCGAGA-3') and p7 (5'-CAAGCAGAAGACGGCATACGAGAT-3') primers and by Bioanalyzer (DNA High Sensitivity Kit, Agilent), and sequenced on an Illumina HiSeq. The custom sequencing

primer: 5'-TCTACACGTTCAGAGTTCTACAGTCCGACGATC-3' was included with Illumina primer HP10 for sequencing. Standard Illumina primers HP12 and HP11 were used for the index read and the transcript read, respectively. PE50 kits (Illumina) were used for sequencing with read lengths of 25 nt, 6 nt, and 47 nt for read1 (cell barcode, UMI), index (library), and read2 (transcript), respectively. Following quantification, we discarded the data from wells that yielded below a total of 20,000 UMI (threshold based on empty well controls), which left us with 358 cells. Further, as others have recognized (Paul et al., 2015), we found that some well-to-well mixing was present with CEL-Seq multiplexed single-cell RNA-Seq. We used the data only from 288 cells because of this mixing artifact.

3.4.4. Immunofluorescence

Cells were grown on ibidi μ-bottom plates and fixed with 4% paraformaldehyde. Cells were permeabilized with ice-cold 100% methanol, blocked with 5% donkey serum, incubated with primary antibody, washed, and incubated with DAPI and secondary antibody coupled to Alexa488 Alexa568, or Alexa647. Images were acquired with a Zeiss 40× plan apo objective (NA 1.3) with the appropriate filter sets. Data was analyzed using custom written code in MATLAB. Antibodies and dilutions used in this study: *Klf4* (Abcam ab129473, 1:400); *Nanog* (eBiosciences 14-5761, 1:800); *Oct4* (Santa Cruz sc-8628, 1:800; Cell Signaling 2840, 1:400); *Sox2* (eBiosciences 14-9811, 1:800); *Otx2* (Neuromics GT15095, 1:400); *T* (*Brachyury*) (Santa Cruz sc-17745, 1:200); *FoxA2* (Cell Signaling 8186, 1:400); *Gata4* (eBiosciences 14-9980, 1:400); *Sox1* (Cell Signaling 4194, 1:200); *Pax6* (DSHB Pax6, 1:200); *Msx1+2* (DSHB 4G1, 1:200); *Slug* (Cell Signaling 9585, 1:200), *Snai1* (Cell Signaling 2879, 1:200).

3.4.5. Live-Cell Microscopy

For live-cell time-lapse microscopy, cells were plated into N2B27 without phenol-red (plus signaling molecules and small molecules) on ibidi µ-bottom plates. Cells were imaged on a Zeiss Axiovision inverted microscope with a Zeiss 40× plan apo objective (NA 1.3) with the appropriate filter sets with an Orca-Flash 4.0 camera (Hamamatsu). The microscope was enclosed with an environmental chamber in which CO_2 and temperature were regulated at 5% and 37°C, respectively. Images were acquired every 15 min for 12–48 hrs. Image acquisition was controlled by Zen (Zeiss); image analysis was done with ImageJ (NIH) and Matlab (MathWorks). 38 HE GFP/43 HE DsRed/46 HE YFP/47 HE CFP/49 DAPI/50 Cy5 filter sets from Zeiss. Transition duration of Otx2-mCitrine cells was defined as the time between the last image at which a cell's reporter intensity was equal to or below its intensity at t = 1 and the first image at which its intensity was equal to or above 2.2 (mean – σ of upper mode of Otx2 reporter intensity) on the normalized scale.

3.4.6. Plasmid Transfection

We cloned *Sox2* or *Snai1* cDNA to one side of a bi-directional Tet-on promoter (pTRE3G-BI; Clontech), to the other side of which we had cloned in mCerulean cDNA. Mini-prepped plasmid was ethanol-precipitated to further concentrate and remove any possible endotoxins. For *Sox2* overexpression, cells were seeded at 100,000 cells per 35mm diameter plate in 2mL of either LIF+2i conditions or differentiation media (0.4µM PD0325901 or 3µM CHIR99021) for 1 day. 200µL of FBS was then added to each plate and 1.8ug of plasmid was transfected using 5.4µL of JetPrime (Polyplus). Cells were incubated for 12 hours, then washed with PBS and replenished with fresh LIF+2i or differentiation media. We then added 3µL of Tet-Express mixed with 2.5µL of Intensifier

reagent (Clontech). Cells were incubated in induction media for 24 hours, after which they were harvested and fixed with 4% paraformaldehyde. Following fixation, they were permeabilized with ice-cold 100% methanol and rehydrated with 1% BSA. Cells were then stained for Oct4, Otx2 and Sox2 and analyzed using flow cytometry. For *Snai1* overexpression, cells were seeded at 100,000 cells per 35mm diameter plate in 2mL of 3μM CHIR99021 for 2.5 days. 200μL of FBS was then added to each plate and 1.8μg of plasmid was transfected using 5.4uL of JetPrime (Polyplus). Cells were incubated in transfection media for 12 hours, then washed with PBS and replenished with fresh N2B27 basal media. We then added 3μL of Tet-Express mixed with 2.5μL of Intensifier reagent (Clontech). Cells were incubated in induction media for 24 hours, after which they were harvested and fixed with 4% paraformaldehyde. Following fixation, they were permeabilized with ice-cold 100% methanol and rehydrated with 1% BSA. Cells were then stained for Oct4 and T and analyzed using flow cytometry.

3.4.7. Fluorescence-Activated Cell Sorting

Cells were trypsinized and fixed in suspension with formaldehyde (4% final concentration, diluted in PBS), permeabilized with ice cold 100% methanol and blocked with 5% donkey serum for 1 hour. Finally, cells are stained with primary antibodies diluted in PBS containing 1% BSA, and detected using fluorescent-tagged secondary antibodies. Flow cytometry was perfomed on a BD FACSAria flow cytometer equipped with 355nm, 405 nm, 488 nm, 561 nm, and 637 nm lasers. The data acquired were analyzed using custom programs written in MatLab.

3.4.8. Generation of mOTX2-Citrine reporter cell line

G4 mESCs, a 129S6 x B6 F1 hybrid line (Andras Nagy, University of Toronto) were maintained on DR4 mouse embryonic fibroblasts (MEFs). These cells (1×10^7) were electroporated (Transfection Buffer, Millipore; Bio-Rad set at 250V and 500 mF) with 5 µg each TALEN plasmid (AI-CN301 and AI-CN302 targeting TTCCAGGTTTTGTGAAGA and TTTAAAAATCACCCACAA, respectively) and 20 µg donor plasmid (AI-CN563). Following transfection, cells were placed on ice for 5 min, then plated onto 3 × 10 cm dishes with MEFs. Beginning 30 h after transfection, cells were selected with hygromycin at 150 µg/mL for 3 days, then 100 µg/mL for an additional 4 days. Approximately 48 hygromycin-resistant colonies were picked and expanded for freezing and DNA preparation and analysis. Five clones were identified with targeted integration by junction PCR (5' junction primers: AAGAGCTAAGTGCCGCCAACAGC, CATCAGCCCGTAGCCGAAGGTAG; 3' junction primers: CACGCTGAACTTGTGGCCGTTTA, CAGCTCACCTCCAGCCCAAGGTA). Following expansion and fluorescence-activated cell sorting (FACS), Cerulean$^+$ cells from two clones (2.1 and 2.4) were treated with Cre mRNA. After recovery and expansion, the Cerulean$^-$ cells were enriched by FACS and single-cell cloned. The resulting subclones were tested for removal of the selection cassette (primers: GGTGCCTATTCTGGTCGAACTGGATG, ATCACCTCTGCTTTGAAGGCCATGAC). The TALENs were kindly provided by the Joung lab synthesized using the FLASH method (Reyon et al., 2012).

3.4.9. Clustering and lineage inference algorithm

Our algorithm proceeds according to the following steps (Figure 3.4):

0. Find initial seed clustering configuration $\{C_0\}$ using Seurat (Satija et al., 2015)
1. For all triplets of clusters, find most likely T and $\{\alpha_i\}$ and $\{\beta_i\}$ given $\{C_0\}$:
    a. Compute $p\left(g_i^{A,B,C} \middle| T, \alpha_i, \beta_i, \{C_0\}\right)$ by integrating numerically over $p(\mu, \sigma)$. (Equations (8), (11) and (12) in Chapter 5. Appendix: Mathematical derivation of Bayesian Framework).
    b. Compute $p\left(T, \{\alpha_i\}, \{\beta_i\} \middle| \{g_i^{A,B,C}\}, \{C_0\}\right)$ using Equations (22) and (31) in Chapter 5. Appendix: Mathematical derivation of Bayesian Framework.
    c. Identify mostly likely topology T and set of $\{\alpha_i\}$ and $\{\beta_i\}$.
2. Recluster $\{g\}$ using Seurat in the space of all $\{\alpha_i\}$ and $\{\beta_i\}$ for the triplets with probability $p\left(T \middle| \{g_i^{A,B,C}\}, \{C_0\}\right)$>0.6 of being non-null. Determine new clustering configuration $\{C_1\}$.
3. Repeat steps 1 and 2 until convergence of $\{C\}$.
4. Determine most likely tree connecting cell clusters, recapitulating high-probability triplet topologies.

Steps 1 and 2 are described in the Mathematical Appendix in Chapter 5; the other steps are described below.

The probabilities of the triplets in the final clustering configuration are shown in the following table.

Table 3.2: Triplet probabilities of final tree.

triplet			probabilities for prior odds p(β_i=1)/p(β_i=0) = 1E-5				# non-null topologies with prob > 0.6	most likely topology	probability at max
A	B	C	p(**A**|{g},{C})	p(**B**|{g},{C})	p(**C**|{g},{C})	p(**0**|{g},{C})			
C0	C1	C3	3.05E-34	1.00E+00	1.01E-89	2.99E-04	1	C1	1.000
C0	C1	C8	1.44E-15	1.00E+00	6.00E-218	6.32E-07	1	C1	1.000
C0	C1	C5	5.33E-31	1.00E+00	5.37E-101	4.14E-05	1	C1	1.000
C0	C1	C6	9.52E-27	1.00E+00	1.89E-103	4.48E-04	1	C1	1.000
C0	C1	C7	4.14E-05	9.99E-01	1.48E-181	1.27E-03	1	C1	0.999
C0	C1	C4	3.90E-24	1.00E+00	6.95E-131	1.54E-04	1	C1	1.000
C0	C1	C2	9.20E-08	9.36E-01	5.13E-13	6.44E-02	1	C1	0.967
C0	C3	C8	1.16E-179	5.78E-10	1.03E-26	1.00E+00	0	null	1.000
C0	C3	C5	3.67E-253	5.41E-01	4.59E-01	6.99E-08	1	C3	0.613
C0	C3	C6	1.53E-181	8.72E-01	1.28E-01	2.99E-07	1	C3	0.976
C0	C3	C7	1.31E-83	1.40E-07	6.99E-47	1.00E+00	1	C3	0.798
C0	C3	C4	2.10E-120	1.97E-17	3.87E-47	1.00E+00	0	null	1.000
C0	C3	C2	2.01E-39	4.21E-57	9.94E-01	5.76E-03	1	C2	0.996
C0	C8	C5	4.77E-141	5.74E-01	5.99E-29	4.26E-01	1	C8	0.994
C0	C8	C6	2.98E-122	8.46E-11	5.60E-01	4.40E-01	2	C8	0.703
C0	C8	C7	2.00E-171	4.47E-01	5.53E-01	9.15E-05	1	C7	0.642
C0	C8	C4	1.28E-255	5.46E-07	1.00E+00	1.46E-06	1	C4	1.000
C0	C8	C2	1.02E-133	1.47E-65	1.00E+00	5.12E-05	1	C2	1.000
C0	C5	C6	1.59E-170	6.60E-01	3.40E-01	1.21E-08	1	C5	0.775
C0	C5	C7	2.55E-66	5.38E-31	2.84E-58	1.00E+00	0	null	1.000
C0	C5	C4	2.82E-93	3.36E-37	4.62E-39	1.00E+00	0	null	1.000
C0	C5	C2	4.30E-21	3.94E-67	9.89E-01	1.09E-02	1	C2	0.997

(Table 3.2, continued)

C0	C6	C7	1.31E-48	1.12E-01	1.35E-24	8.88E-01	1	C6	0.903
C0	C6	C4	1.64E-82	1.21E-08	3.26E-38	1.00E+00	1	C6	0.708
C0	C6	C2	8.53E-13	1.57E-41	9.89E-01	1.14E-02	1	C2	0.996
C0	C7	C4	1.01E-165	5.13E-01	4.87E-01	3.53E-05	0	C7	0.538
C0	C7	C2	1.03E-93	1.79E-39	9.92E-01	7.56E-03	1	C2	0.997
C0	C4	C2	6.05E-126	5.07E-03	9.92E-01	2.77E-03	1	C2	0.999
C1	C3	C8	8.60E-27	8.49E-07	4.52E-41	1.00E+00	0	null	1.000
C1	C3	C5	2.70E-52	5.52E-01	4.35E-01	1.38E-02	1	C3	0.660
C1	C3	C6	2.10E-14	9.28E-01	1.86E-02	5.29E-02	1	C3	0.980
C1	C3	C7	8.89E-04	1.94E-06	2.60E-45	9.99E-01	1	C1	0.800
C1	C3	C4	2.62E-09	4.21E-14	1.28E-48	1.00E+00	1	C1	0.705
C1	C3	C2	8.47E-01	4.01E-46	1.61E-04	1.53E-01	1	C1	0.879
C1	C8	C5	8.88E-29	1.54E-01	7.67E-32	8.46E-01	1	C8	0.820
C1	C8	C6	4.48E-26	1.51E-17	3.46E-03	9.97E-01	1	C6	0.816
C1	C8	C7	6.05E-177	6.89E-01	4.28E-16	3.11E-01	1	C8	0.949
C1	C8	C4	1.26E-215	1.74E-16	9.97E-01	3.17E-03	1	C4	0.998
C1	C8	C2	1.80E-173	8.61E-53	9.96E-01	3.84E-03	1	C2	0.998
C1	C5	C6	9.57E-24	7.14E-01	2.77E-01	8.76E-03	1	C5	0.882
C1	C5	C7	1.12E-05	7.42E-23	9.43E-43	1.00E+00	1	C1	0.874
C1	C5	C4	8.72E-23	2.47E-40	3.55E-18	1.00E+00	0	null	1.000
C1	C5	C2	1.52E-07	5.30E-62	9.22E-01	7.84E-02	1	C2	0.945
C1	C6	C7	1.13E-04	9.71E-04	7.74E-22	9.99E-01	0	C6	0.578
C1	C6	C4	1.71E-03	1.46E-13	1.58E-19	9.98E-01	1	C1	0.709
C1	C6	C2	7.99E-01	1.25E-36	1.58E-01	4.23E-02	1	C1	0.897
C1	C7	C4	1.04E-191	2.38E-09	9.91E-01	9.47E-03	1	C4	0.995
C1	C7	C2	2.67E-146	3.28E-16	9.50E-01	5.03E-02	1	C2	0.971

(Table 3.2, continued)

C1	C4	C2	3.29E-170	1.19E-01	5.87E-01	2.95E-01	1	C2	0.605
C3	C8	C5	1.02E-03	5.17E-86	9.89E-01	9.99E-03	1	C5	0.994
C3	C8	C6	9.91E-01	2.00E-55	5.49E-06	9.00E-03	1	C3	0.993
C3	C8	C7	6.83E-108	3.16E-02	6.22E-01	3.46E-01	1	C7	0.798
C3	C8	C4	3.67E-167	2.63E-20	9.88E-01	1.22E-02	1	C4	0.988
C3	C8	C2	7.72E-147	1.40E-12	5.53E-08	1.00E+00	1	C2	0.890
C3	C5	C6	3.30E-01	4.90E-01	1.72E-19	1.80E-01	0	C5	0.583
C3	C5	C7	1.45E-02	9.63E-01	8.46E-56	2.29E-02	1	C5	0.980
C3	C5	C4	1.92E-04	9.51E-01	2.21E-85	4.90E-02	1	C5	0.955
C3	C5	C2	2.15E-03	9.90E-01	1.64E-64	7.90E-03	1	C5	0.994
C3	C6	C7	9.19E-01	2.66E-04	5.57E-32	8.05E-02	1	C3	0.941
C3	C6	C4	6.98E-01	1.57E-01	6.29E-40	1.45E-01	1	C3	0.824
C3	C6	C2	7.89E-01	1.33E-01	1.54E-14	7.85E-02	1	C3	0.920
C3	C7	C4	1.04E-132	9.84E-01	3.90E-09	1.62E-02	1	C7	0.987
C3	C7	C2	1.86E-112	2.51E-07	9.97E-07	1.00E+00	1	C2	0.792
C3	C4	C2	3.36E-147	9.38E-01	3.31E-05	6.16E-02	1	C4	0.957
C8	C4	C2	4.43E-07	9.00E-01	7.45E-05	9.97E-02	1	C4	0.945
C8	C7	C2	7.78E-01	5.09E-02	3.73E-09	1.71E-01	1	C8	0.790
C8	C7	C4	1.10E-03	3.14E-12	7.79E-01	2.20E-01	1	C4	0.813
C8	C6	C2	1.36E-08	1.88E-143	2.08E-11	1.00E+00	1	C8	0.800
C8	C6	C4	3.15E-22	2.19E-149	9.65E-01	3.53E-02	1	C4	0.970
C8	C6	C7	2.51E-03	1.76E-123	9.32E-01	6.56E-02	1	C7	0.933
C8	C5	C2	4.75E-02	1.83E-130	1.17E-13	9.52E-01	1	C8	0.846
C8	C5	C4	1.99E-08	5.78E-123	8.73E-01	1.27E-01	1	C4	0.877
C8	C5	C7	5.81E-01	7.20E-113	1.28E-11	4.19E-01	1	C8	0.795
C8	C5	C6	1.54E-39	9.86E-01	2.38E-15	1.41E-02	1	C5	0.995

(Table 3.2, continued)

C5	C4	C2	5.18E-137	9.82E-01	6.09E-07	1.85E-02	1	C4	0.989
C5	C7	C2	5.72E-121	3.79E-09	4.53E-07	1.00E+00	1	C2	0.766
C5	C7	C4	1.82E-139	1.15E-01	1.19E-05	8.85E-01	1	C7	0.922
C5	C6	C2	9.46E-01	7.30E-08	3.64E-08	5.38E-02	1	C5	0.976
C5	C6	C4	8.57E-01	2.37E-19	7.55E-13	1.43E-01	1	C5	0.891
C5	C6	C7	9.28E-01	9.30E-05	4.45E-36	7.17E-02	1	C5	0.950
C6	C7	C4	1.40E-140	3.62E-01	6.10E-01	2.81E-02	1	C4	0.781
C6	C7	C2	6.57E-114	2.77E-01	6.61E-01	6.22E-02	1	C2	0.851
C6	C4	C2	5.76E-143	9.81E-01	1.84E-04	1.85E-02	1	C4	0.990
C7	C4	C2	4.02E-01	4.25E-01	3.90E-05	1.73E-01	0	C4	0.460

3.4.10. Clustering and re-clustering using Seurat

Clustering was performed using Seurat (Satija et al., 2015). For the initial seed clustering, we applied Seurat to the gene expression of all 2,672 transcription factors for the 288 single cells. For subsequent re-clustering steps, clustering was performed on a reduced set of genes for which $p\left(\alpha_i = 1 \text{ or } \beta_i = 1 \middle| \left\{g_i^{A,B,C}\right\}, T, \{C\}\right) > 0.5$ for at least one triplet at the previous iteration (assuming a prior odds of $\mathcal{O}_{\beta|T}(i) = 5 \times 10^{-2}$). This reduced set contained between 800 and 1050 genes at each of the reclustering steps (Figure 3.6A).

Seurat performs spectral t-SNE on the statistically significant principal components (PCs) of the gene expression dataset, and it determines the significance of each PC score using a randomization approach developed by Chung and Storey (Chung and Storey, 2015).

Our initial seed clustering was performed using the first 10 PCs; subsequent re-clusterings used the first 8 PCs.

Finally, Seurat performs density-based clustering on the t-SNE map; we used a density parameter of G=8.

3.4.11. Convergence of clustering configurations from different seed configurations

In order to test that our results were robust to the choice of seed clusters, we further used k-means clustering, a standard clustering method, which has previously been applied to identify different cell types using single cell transcriptomics data (Buettner et al., 2015).

We start with a seed clustering configuration of 12 clusters $\{c_1^0, c_2^0, \dots, c_{12}^0\}$ obtained using k-means clustering, which is distinct from the seed clustering configuration obtained via Seurat (Satija et al., 2015). The number of clusters was determined using the gap statistic (Tibshirani et al., 2001). We obtained 164 sets of transitions between clusters and identified 981 transcription factors that were high probability (probability > 0.5) marker or transition genes for at least one of the identified transitions. We next re-clustered the single cells in the gene expression space defined by these 981 marker or transition genes, using k-means clustering, to obtain a new cluster set $\{C_1\} = \{c_1^0, c_2^0, \dots, c_{10}^0\}$, consisting of 10 clusters. In the next iteration, the number of clusters went down to 9, and so on. By iteratively determining the most likely sets of transitions, the corresponding most likely marker and transition genes and re-clustering the cells within the subspace of these genes, our algorithm converged upon the most likely set of cell clusters (Figure 3.6B). We found that the eventual clustering configurations obtained using k-means clustering and Seurat are the same, confirming that the seed clusters do not affect the final outcome.

3.4.12. Determination of gene modules

Classifying genes based on their patterns of expression along the inferred lineage tree rather than by gene-gene correlations allowed us to identify gene modules (which included the transition and marker genes we inferred as well as signaling genes: BMP, WNT, LIF) with similar expression patterns in successive cell-fate decisions.

We partitioned the 184 transcription factors and signaling genes into a total of 36 subclasses, each of which is a unique intersection of the different transition and state genes. We determined binary gene expression profiles by calculating the mean log2 fold-change in expression level for each subclass.

We grouped subclasses with identical binary expression profiles together, leaving us with a total of 23 modules with unique binary gene expression profiles. We denote each module by a representative gene; the genes that belong to each module are shown in

Table **3.3** and their binary profiles in the 9 clusters in Table 3.4.

Table 3.3: Gene modules used for modeling the network.

Module names	Gene members									
[Klf4]	Ash2l	Esrrb	Eed	Kdm3a	Klf4	Klf5	Poldip2	Sin3b	Tfcp2l1	Zfp42
	Tsc22d1	Fblim1	Jarid2							
[Hes6]	Hes6	Hmgb2	Ncl	Zscan10	Pole	Cbx1	Prrc2c	Mycn	Rhox5	
[Churc1]	Churc1	Gm13051	Gm13154	Gm13157	Mtf2	Pa2g4	Psmc5	Ptma	Supt6	Suz12
	Tet1	Tomm6	Ttf1	Klf9						
[Apex1]	Apex1	Phc1	Drg1	Lin28b	Notch2	Phb	Plrg1	Zfp207		
[Pax6]	Pax6									
[Atf2]	Atf2	Lsm14a	Polr2e	Polr2f	Puf60					
[Sox2]	Chd4	Dnajc2	Rbpj	Set	Sox4	Upf1	Sox2			
[Baz1a]	Basp1	Baz1a	Exoc3	Fbxo18	Foxm1	Med14	Peg10	Zfp326	Tfap2c	Smarcc1
	Wbp5	Rbbp7								
[Msx2]	Lrrfip1	Naa15	Tal2	Zfp746	Msx2	Lbr	Mtpn	Paxbp1		
[Snai1]	Cand2	Cited1	Hmga2	Lef1	Pdlim4	Snai1	Tbx6			
[Ciao1]	Ciao1	Foxa2	Keap1	Msh3	Tdp2	Tsg101				
[Tead1]	Tead1									
[Hes1]	Ajuba	Basp1	Bbx	Cbx5	Fam58b	Hes1	Hmgcs1	Hmgn5	Kat7	Nlrp1a
	Olig1	Phc1	Rab15	Rest	Sap18	Sfmbt2	Sptbn1	Sra1	Taf10	Tsc22d4
	Tsn	Wiz	Zfp322a	Zfp593	Zmat3					
[Oct4]	Oct4	Utf1								
[Ets2]	Bclaf1	Cand1	Cenpa	Fus	Glyr1	Gtf2i	Kdm5c	Kif22	Pbrm1	Phf5a
	Polb	Preb	Rad51ap1	Rbms1	Smarca5	Top1	Utp6	Zmat2	Zmynd11	
[Hmga1]	Hmga1	Tial1	Sall4	Zfp266	Sod2	Son				
[Sp5]	Brd4	Chtf8	Dek	Etf1	Foxo4	Med10	Pfdn1	Pml	Ppp5c	Prkrir
	Pygo2	Qrich1	Zfp445	Rrn3	Sin3b	Ssbp3	Strn3	Ttf1		
[Otx2]	Otx2	Dnmt3b	Tceb2	Trim28	Crip2	Aplp2	Hnrnpu			
[T]	T	Hells								
[Etv5]*	Etv5	Ctbp2	Ddx3x	Aes	Foxo1	Rpa2	Pdlim7	Tcea3	Rab25	Rad23a
	Dnmt3a	Tomm6	Sp1	Top2a	Taf7					
[Smarce1]	Smarce1	Strbp								

(Table 3.3, continued)

[LIF]	Lif	Lifr	Il6st	Jak2	Jak3	Stat1	Stat3	Stat5a	Stat5b	
[FGF]**	Cep57	Ctgf	Ctnnb1	Dstyk	Dusp6	Fam20c	Fgf1	Fgf10	Fgf15	Fgf16
	Fgf17	Fgf18	Fgf2	Fgf20	Fgf21	Fgf22	Fgf23	Fgf3	Fgf4	Fgf5
	Fgf8	Fgf9	Fgfbp1	Fgfbp3	Fgfr1	Fgfr2	Fgfr3	Fgfr4	Flrt1	Flrt2
	Flrt3	Frs2	Frs3	Grb2	Hhip	Iqgap1	Kif16b	Kl	Klb	Lrit3
	Ndst1	Nog	Pdgfb	Prkd2	Rab14	Runx2	Setx	Shcbp1	Sos1	Trim71
	Fgf6	Fgf7								
[BMP]*	Bmp10	Bmp15	Bmp2	Bmp2k	Bmp3	Bmp4	Bmp5	Bmp6	Bmp7	Bmp8a
	Bmp8b	Bmpr1a	Bmpr1b	Bmpr2	Acvr1	Acvr1b	Acvr1c	Acvr2a	Acvr2b	Acvrl1
	Actr1a	Actr1b	Actr2	Actr3	Actr3b	Actr5	Actr6	Actr8	Actrt1	Actrt2
	Actrt3	Tgfbr1	Tgfbr2	Tgfbr3						
[WNT]**	Amer1	Ankrd10	Apc	Arntl	Aspm	Bambi	Bcl9	Bcl9l	Caprin2	Ccar2
	Cdh3	Cdk14	Cfc1	Col1a1	Csnk1d	Csnk1e	Ctdnep1	Ctnnd2	Dapk3	Disc1
	Eda	Egf	Emd	Folr1	Fzd10	Fzd2	Fzd3	Fzd4	Dvl2	Dvl3
	Fzd9	Gata3	Gprc5b	Gsk3b	Hoxb9	Ift20	Ilk	Ins2	Kdm6a	Lgr4
	Lrrk1	Lrrk2	Med12	Mesp1	Mgat3	Mitf	Mks1	Myc	Myh6	Ndp
	Otulin	Plpp3	Porcn	Prop1	Psen1	Pten	Ptk7	Ptpru	Rab5a	Rnf146
	Rspo3	Ryr2	Sdc1	Smad3	Sox7	Src	Stk11	Sulf2	Tbl1x	Tbl1xr1
	Tmem198	Tnks	Tnks2	Trpm4	Ube2b	Ubr5	Usp34	Uty	Vps35	Wls
	Wnt2b	Wnt3	Wnt3a	Wnt7a	Wnt7b	Wnt9a	Wnt9b	Xiap	Zbed3	Zfp703
	Ccnd1	Ccny	Cdc42	Fzd5	Fzd7	Fzd8	Nfkb1	Nle1	Nrarp	Tcf7
	Dixdc1	Dlx5	Dvl1	Lgr5	Lrp5	Lrp6	Rnf220	Rspo1	Rspo2	Wnt1
	Tcf7l1	Tdgf1	Wnt10b	Wnt2						

Table 3.4: Binary expression profiles of the gene modules used for modeling the network in the 9 cell clusters.

[Module name]	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
[Klf4]	1	0	0	0	0	0	0	0	0
[Hes6]	0	1	0	0	0	0	0	0	0
[Hmga1]	0	0	1	0	0	0	0	0	0
[Tead1]	0	0	0	1	0	0	0	0	0
[Sp5]	0	0	0	0	1	0	0	0	0
[Baz1a]	0	0	0	0	0	1	0	0	0
[Msx2]	0	0	0	0	0	0	1	0	0
[Snai1]	0	0	0	0	0	0	0	1	0
[Ciao1]	0	0	0	0	0	0	0	0	1
[Churc1]	1	1	0	0	0	0	0	0	0
[BMP]	1	1	1	0	1	1	1	1	1
[LIF]	1	1	0	0	1	1	1	1	1
[FGF]	1	1	1	1	1	1	1	1	1
[Pou5f1]	1	1	1	0	1	0	0	0	1
[Sox2]	1	1	0	1	0	0	0	0	0
[Atf2]	1	0	1	1	1	0	1	1	1
[Otx2]	0	1	0	1	0	1	1	0	0
[Smarce1]	0	0	1	1	1	1	0	1	1
[Ets2]	1	0	1	1	1	1	1	1	0
[T]	0	0	1	0	1	0	0	1	1
[Apex1]	1	0	1	1	1	1	0	1	1
[Hes1]	0	0	1	1	1	1	1	0	1
[Pax6]	0	0	0	1	0	1	0	0	0

3.4.13. Local-field gene regulatory network model for gene modules

We considered Hopfield-like gene regulatory networks and sampled parameters using a linear-programming approach, using the same framework as described in Section 2.2.5: Modeling the underlying network (*cf ad locum*).

3.4.14. Common features of the sampled networks

By using many different randomly generated fictitious constraints to sample the polytope, we can study the ensemble of model networks that all satisfy the fixed point constraints (Table 3.4), and attempt to determine whether they share any common regulatory motifs. As discussed in the main text, we sampled 10,000 solutions J_{ij} that satisfied the fixed-point constraints defined by the binarized expression patterns of the known cell states. We then calculated the mean and coefficient of variation (c.v.) for each coupling. We were thus able to discover a core network between the different modules that is shared by the majority of solutions (Figure 3.10A).

3.4.15. Predictions for Sox2 and Snai1 overexpression

Our model makes predictions for what happens to the level of Oct4 when Sox2 and Snai1 are overexpressed in different cell states. Sox2 and Oct4 are both present in the C_0 and C_1 clusters. On the other hand, Snai1 is not present in C_1 and C_2 but Oct4 is present in both clusters. We perturb the Sox2 and Snai1 levels by amounts Δs in the above mentioned states, which lead to a change in the field ϕ_i total drive on Oct4 level. Numerically we vary Δs in steps of 0.1 and for each step compute the number of models out of the 10000 total models, for which the Oct4 level decreases to zero. From this number we obtain the fraction of models for which the level of Oct4 goes down.

3.4.16. Predictions for BMP and LIF addition

In order to predict the effect of morphogen signals in different cell states, we considered the LIF, BMP, WNT, and FGF signaling pathways, which are known to play a significant role in patterning the early embryo. We assumed that no single gene in each given pathway is sufficient to evoke a signaling response, but a response rather requires the combined presence of the various constituent genes of the pathway. We therefore grouped genes by their respective signaling pathways and assigned each group to a module based on its average expression pattern across the nine cell states.

We next modeled the dynamics of BMP and LIF addition. By construction, the 9 observed cell states (and the null state $\vec{m} = \vec{0}$) are fixed points for all 10,000 sampled solutions for J_{ij}. However, each solution J_{ij} may have additional spurious fixed points. However, given that we only see 9 cell states, we would expect the spurious states to be unstable. In order to overcome this problem, we used the following method.

Given a particular solution J_{ij}, any arbitrary state of the network $\vec{m}$ (not necessarily a fixed point) will have dynamics obeying

$$m_i(t+1) = H\left(\sum_j J_{ij} m_j(t) - \phi_0\right),$$

where $m_i(t)$ and $m_i(t+1)$ are the levels of module i at successive discretized time points.

For each particular solution J_{ij}, cells will get stuck in spurious fixed points; yet these spurious fixed points are highly unlikely to exist since they are stable in only a small number of the sampled J_{ij}. We can capture the average dynamics of different states of the

network given the set of sampled solutions $\{J_{ij}\}$ by calculating the probability over all sampled solutions of moving from one arbitrary state $\vec{m}^a$ to another arbitrary state $\vec{m}^b$. This allows us defines a 2^{23} x 2^{23} state-to-state transition matrix $\mathcal{T}$:

$$\mathcal{T}_{b\leftarrow a} = p\left(\vec{m}^a \rightarrow \vec{m}^b \middle| \{J_{ij}\}\right).$$

If we denote as $\vec{p}(t)$ the vector of probabilities of being in the 2^{23} different states at time t, then

$$\vec{p}(t+1) = \mathcal{T}\,\vec{p}(t).$$

In order to figure out what happens to cells in different states to BMP and LIF addition, we calculated the probability of moving between fixed points $\vec{m}^\alpha$ and $\vec{m}^\beta$ when overexpressing some set of modules $\{m_i\}$. We calculated the dynamics using the transition matrix $\mathcal{T}$ and enforced the overexpression of the set of modules (BMP and LIF module respectively) at each time point, updating the probabilities $\vec{p}(t)$ accordingly. The probabilities shown in Figure 3.10 are after 1,000 time steps.

Acknowledgements

We thank Christof Koch, Ajamete Kayakas, Joshua Levi, Carol Thomson, John Phillips, Paola Arlotta, John Calarco, Leonid Mirny and Andrew Murray for their critical feedback. We thank the Allen Institute founders, P. G. Allen and J. Allen and the NIH Directors

Pioneer Award 5DP1MH099906-03 and National Science Foundation grant PHY-0952766

for support.

References

Anderson, P.W. (1984). Basic Notions of Condensed Matter Physics.

Arnold, S.J., and Robertson, E.J. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. Nature reviews Molecular cell biology *10*, 91-103.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schubeler, D., Sasaki, H., Forne, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nature genetics *42*, 1093-1100.

Brown, L., and Brown, S. (2009). Zic2 is expressed in pluripotent cells in the blastocyst and adult brain expression overlaps with makers of neurogenesis. Gene expression patterns : GEP *9*, 43-49.

Chambers, I. (2004). The molecular basis of pluripotency in mouse embryonic stem cells. Cloning and stem cells *6*, 386-391.

Ekspong, G. (1993). Nobel Lectures in Physics, 1981-1990.

Ekspong, G. (2008). Nobel Lectures in Physics, 2001-2005.

Furchtgott, L., Zou, L.-N., and Ramanathan, S. (2016). Simultaneous inference of lineage trees and gene regulatory networks from gene expression data. (In review).

Gadue, P., Huber, T.L., Paddison, P.J., and Keller, G.M. (2006). Wnt and TGF-beta signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *103*, 16806-16811.

Galvagni, F., Lentucci, C., Neri, F., Dettori, D., De Clemente, C., Orlandini, M., Anselmi, F., Rapelli, S., Grillo, M., Borghi, S.*, et al.* (2015). Snai1 promotes ESC exit from the pluripotency by direct repression of self-renewal genes. Stem cells *33*, 742-750.

Gans, C., and Northcutt, R.G. (1983). Neural crest and the origin of vertebrates: a new head. Science *220*, 268-273.

Gaspard, N., Bouschet, T., Hourez, R., Dimidschstein, J., Naeije, G., van den Ameele, J., Espuny-Camacho, I., Herpoel, A., Passante, L., Schiffmann, S.N., *et al.* (2008). An intrinsic mechanism of corticogenesis from embryonic stem cells. Nature *455*, 351-357.

Goller, T., Vauti, F., Ramasamy, S., and Arnold, H.H. (2008). Transcriptional regulator BPTF/FAC1 is essential for trophoblast differentiation during early mouse development. Molecular and cellular biology *28*, 6819-6827.

Hart, A.H., Hartley, L., Sourris, K., Stadler, E.S., Li, R., Stanley, E.G., Tam, P.P.L., Elefanty, A.G., and Robb, L. (2002). Mixl1 is required for axial mesendoderm morphogenesis and patterning in the murine embryo. Development *129*, 3597-3608.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports *2*, 666-673.

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

Kanai-Azuma, M., Kanai, Y., Gad, J.M., Tajima, Y., Taya, C., Kurohmaru, M., Sanai, Y., Yonekawa, H., Yazaki, K., Tam, P.P., *et al.* (2002). Depletion of definitive gut endoderm in Sox17-null mutant mice. Development *129*, 2367-2379.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389-401.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049-1061.

Kim, J.K., Huh, S.O., Choi, H., Lee, K.S., Shin, D., Lee, C., Nam, J.S., Kim, H., Chung, H., Lee, H.W., *et al.* (2001). Srg3, a mouse homolog of yeast SWI3, is essential for early embryogenesis and involved in brain development. Molecular and cellular biology *21*, 7787-7795.

Kim, P.T., and Ong, C.J. (2012). Differentiation of definitive endoderm from mouse embryonic stem cells. Results and problems in cell differentiation *55*, 303-319.

Knecht, A.K., and Bronner-Fraser, M. (2002). Induction of the neural crest: a multigene process. Nature reviews Genetics *3*, 453-461.

Koch, P.J., and Roop, D.R. (2004). The Role of Keratins in Epidermal Development and Homeostasis[mdash]Going Beyond the Obvious. J Investig Dermatol *123*, x-xi.

Landau, L.D., and Lifshitz, E.M. (1980). Course of Theoretical Physics: Vol. 5, Statistical Physics, Part 1.

Le Douarin, N.M.K., C. (1991). The Neural Crest.

Li, L., Song, L., Liu, C., Chen, J., Peng, G., Wang, R., Liu, P., Tang, K., Rossant, J., and Jing, N. (2015). Ectodermal progenitors derived from epiblast stem cells by inhibition of Nodal signaling. Journal of molecular cell biology *7*, 455-465.

Lindsley, R.C., Gill, J.G., Kyba, M., Murphy, T.L., and Murphy, K.M. (2006). Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. Development *133*, 3787-3796.

Lumelsky, N., Blondel, O., Laeng, P., Velasco, I., Ravin, R., and McKay, R. (2001). Differentiation of Embryonic Stem Cells to Insulin-Secreting Structures Similar to Pancreatic Islets. Science *292*, 1389-1394.

Lundqvist, S. (1992). Nobel Lectures in Physics 1971-1980.

Machta, B.B., Chachra, R., Transtrum, M.K., and Sethna, J.P. (2013). Parameter Space Compression Underlies Emergent Theories and Predictive Models. Science *342*, 604-607.

Merrill, B.J., Pasolli, H.A., Polak, L., Rendl, M., Garcia-Garcia, M.J., Anderson, K.V., and Fuchs, E. (2004). Tcf3: a transcriptional regulator of axis induction in the early embryo. Development *131*, 263-274.

Nakanishi, M., Kurisaki, A., Hayashi, Y., Warashina, M., Ishiura, S., Kusuda-Furue, M., and Asashima, M. (2009). Directed induction of anterior and posterior primitive streak by Wnt from embryonic stem cells cultured in a chemically defined serum-free medium. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *23*, 114-122.

Nichols, J., and Smith, A. (2009). Naive and Primed Pluripotent States. Cell Stem Cell *4*, 487-492.

Paul, F., Arkin, Y.a., Giladi, A., Jaitin, Diego 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.

Pevny, L.H., Sockanathan, S., Placzek, M., and Lovell-Badge, R. (1998). A role for SOX1 in neural determination. Development *125*, 1967-1978.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nature biotechnology *30*, 460-465.

Rojas, A., De Val, S., Heidt, A.B., Xu, S.M., Bristow, J., and Black, B.L. (2005). Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by

Forkhead and GATA transcription factors through a distal enhancer element. Development *132*, 3405-3417.

Sansom, S.N., Griffiths, D.S., Faedo, A., Kleinjan, D.J., Ruan, Y., Smith, J., van Heyningen, V., Rubenstein, J.L., and Livesey, F.J. (2009). The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis. PLoS genetics *5*, e1000511.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat Biotech *33*, 495-502.

Streit, A., and Stern, C.D. (1999). Neural induction: a bird's eye view. Trends in Genetics *15*, 20-24.

Sumi, T., Tsuneyoshi, N., Nakatsuji, N., and Suemori, H. (2008). Defining early lineage specification of human embryonic stem cells by the orchestrated balance of canonical Wnt/β-catenin, Activin/Nodal and BMP signaling. Development *135*, 2969-2979.

Tada, S., Era, T., Furusawa, C., Sakurai, H., Nishikawa, S., Kinoshita, M., Nakao, K., Chiba, T., and Nishikawa, S.-I. (2005). Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture. Development *132*, 4363-4374.

Tam, P.P., Loebel, D.A., and Tanaka, S.S. (2006). Building the mouse gastrula: signals, asymmetry and lineages. Current opinion in genetics & development *16*, 419-425.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196-199.

Thomson, M., Liu, S.J., Zou, L.N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875-889.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research *9*, 85.

Vogel-Ciernia, A., and Wood, M.A. (2014). Neuron-specific chromatin remodeling: A missing link in epigenetic mechanisms underlying synaptic plasticity, memory, and intellectual disability disorders. Neuropharmacology *80*, 18-27.

Watabe, T., and Miyazono, K. (2009). Roles of TGF-[beta] family signaling in stem cell renewal and differentiation. Cell Res *19*, 103-115.

Wilson, P.A., and Hemmati-Brivanlou, A. (1995). Induction of epidermis and inhibition of neural fate by Bmp-4. Nature *376*, 331-333.

Ying, Q.L., and Smith, A.G. (2003). Defined conditions for neural commitment and differentiation. Methods in enzymology *365*, 327-341.

Ying, Q.L., Wray J Fau - Nichols, J., Nichols J Fau - Batlle-Morera, L., Batlle-Morera L Fau - Doble, B., Doble B Fau - Woodgett, J., Woodgett J Fau - Cohen, P., Cohen P Fau - Smith, A., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal.

Young, Richard A. (2011). Control of the Embryonic Stem Cell State. Cell *144*, 940-954.

Zhou, Q., Chipperfield, H., Melton, D.A., and Wong, W.H. (2007). A gene regulatory network in mouse embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *104*, 16438-16443.

Chapter 4. Region-Specific Neural Stem Cell Lineages Revealed By Single-Cell RNA-Seq From Human Embryonic Stem Cells

[A large part of this chapter is in review at *Cell Stem Cell* as Sherman Ku[,], Vilas Menon[,], John K. Mich[,], Zizhen Yao[,], Anne-Rachel Krostag, Refugio A. Martinez, Leon Furchtgott, Heather Mulholland, Susan Bort, Magaret A. Fuqua, Ben W. Gregor, Rebecca D. Hodge, Anu Jayabalu, Ryan C. May, Samuel Melton, Angelique M. Nelson, N. Kiet Ngo, Nadiya V. Shapovalova, Soraya I. Shehata, Michael W. Smith, Leah J. Tait, Elliot R. Thomsen, Chaoyang Ye, Ian A. Glass, Ajamete Kaykas, Shuyuan Yao, Carol L. Thompson, John W. Phillips, Joshua S. Grimley[,], Boaz P. Levi[,,], Yanling Wang[,], Sharad Ramanathan, "Region-Specific Neural Stem Cell Lineages Revealed By Single-Cell RNA-Seq From Human Embryonic Stem Cells," Genome engineering was conducted by AJ, ARK, RAM, RCM, AMN, NKN, LJT, MS and planned by JSG and AK. Stem cell banking was conducted by AMN and NKN. YW and SY adapted the differentiation protocol. RAM, HM, MF, BWG and YW conducted all differentiations. ICC was conducted by YW, HM, MF and AMN. Calcium imaging was conducted by BWG and RAM, planned by YW. Live single-cell RNA-Seq was piloted by BPL and CY, and conducted by ARK with assistance from ERT. Bioinformatic analysis of RNA-Seq data was conducted by ZY and VM. The lineage algorithm was developed by LF and SR, and applied by VM with help from SM. Primary human tissue was obtained with help from IG, processed by JKM, RDH, and SIS. FRISCR analysis was conducted by ERT. Viral barcoding experiments were conducted by SK, and non-viral clonal analysis by JKM. All sorting was done by NVS and SB. Experiments were conceived by JSG, BPL, YW, SK, VM, JKM, ZY and SR. Paper was written by JKM, YW,

VM, SK, JSG, BPL and SR. CLT provided program management. SR and JWP provided program leadership.]

Abstract

The human brain is a complex organ composed of billions of neurons, representing a diverse array of interconnected cell types. The functions of the human brain are orchestrated through highly inter- or intra-connected regions that are developmentally defined as the forebrain, the midbrain and the hindbrain. Here, we report a human brain cell-type lineage tree through single-cell RNAseq analysis of differentiated human embryonic stem cells (hESCs). We isolated progenitors and neurons throughout a differentiation time-course using $DCX^{Cit/Y}$ and $SOX2^{Cit/+}$ engineered cell lines. Single-cell transcriptomic profiling was used to identify cell types, and their regional identities were determined by comparison to existing atlases and primary human fetal tissues. A Bayesian framework was used to infer a neural lineage tree and putative regulatory transcription factors from single-cell transcriptomic profiles. The lineage tree shows a prominent bifurcation between cortical and mid/hindbrain cell types, and the inferred lineage relationships were confirmed by clonal analysis experiments. In summary, we present an experimentally validated lineage tree that encompasses multiple brain regions, and our work sheds light on the molecular regulation of region-specific neural lineages during human brain development.

4.1. Introduction

The human brain is a complex and highly evolved structure. Mouse models do not fully recapitulate cell-type diversity or lineage trajectories of the human brain (Konopka et

al., 2012; Oberheim et al., 2009; Rakic, 2009; Reilly et al., 2015; Thomsen et al., 2016). Furthermore, human neurodevelopmental diseases such as autism spectrum disorders and schizophrenia are incompletely modeled in mouse. To better understand and combat these disorders, stem cell-based models of human brain development have been pursued (Habela et al., 2015; Hook et al., 2014; Ricciardi et al., 2012).

The molecular networks that drive the fate decisions and development of human neurons and glia are not fully understood, and some are likely to be evolutionarily unique (Lui et al., 2011; Lui et al., 2014). Until recently there had been technical obstacles to understanding the development of this complex tissue: developmental steps cells undergo to give rise to these neurons have been characterized using only a few molecular markers at a time. Recently, single-cell transcriptomics has been used to characterize cellular heterogeneity because it allows multidimensional molecular characterization at an increasing scale (Klein et al., 2015; Macosko et al., 2015); single-cell techniques have allowed the definition of new transcriptomic cell types from complex organs such as the gut (Grun et al., 2015), blood (Paul et al., 2015), and mouse brain (Tasic et al., 2016; Zeisel et al., 2015). In parallel, there has been recent progress in modeling early human brain development using human embryonic stem cells (hESCs) in neural differentiation protocols (Chambers et al., 2009; Espuny-Camacho et al., 2013; Lancaster et al., 2013; Shi et al., 2012); these systems promise to supply human neural tissue for analysis at developmental stages that are typically unavailable from donors. Although several studies have characterized differentiated cells by gene expression (Edri et al., 2015; van de Leemput et al., 2014), only one *in vitro* differentiation study has carried out single-cell transcriptomics (Camp et al., 2015). As these cultures presumably contain a mixture of cell

types at any given time point, single-cell resolution studies are essential to characterize the cell types produced in culture and to determine how well they compare to primary developing tissue.

Here, we study the early development of the human brain through single-cell transcriptomics. We computationally identified cell types and predicted the lineage relationships between them. To construct these lineage relationships of the cells from our *in vitro* neural differentiation system, we used a novel computational framework that we developed to infer cell states and the lineage relationships between them. We previously validated this framework on hematopoiesis and early germ layer specification (Jang et. al. in preparation, see supporting document). We validated the biological relevance of this lineage tree by direct comparison to cortical cells from mid-gestation human fetal embryos and experimentally tested the computational predictions using viral barcoding and clonal analysis. Our lineage tree captures some of the earliest regional patterning events of the brain, including the separation of cortical from posterior brain cell types and the appearance of excitatory and inhibitory forebrain neurons. Taken as a whole, these data constitute a deep and broad interrogation of hESC neural differentiation and highlight key steps in regional patterning, early brain development, and lineage specification.

4.2. Results

4.2.1. In vitro model of human brain excitatory cell development.

We developed and standardized an *in vitro* model of human brain development based on the neuralization of hESCs, adapted from previous protocols (Chambers et al., 2009; Espuny-Camacho et al., 2013; Shi et al., 2012). The cortical induction (CI) phase utilizes SMAD inhibition (Chambers 2009), the progenitor expansion (PE) phase includes

EGF and bFGF, and the neural differentiation (ND) phase includes neurogenic/neurotrophic factors BDNF, GDNF, NT3, and cAMP (Hu et al., 2010) (Figure 4.1A). At the end of CI (D12), most cells expressed both PAX6 and FOXG1, and 92 ± 3% co-express PAX6 and SOX2, suggesting efficient telencephalic induction (Figure 4.1B). By the end of PE (D26), 11 ± 2% of cells expressed the cortical intermediate progenitor marker EOMES (TBR2) (Figure 1B,S1D). Following ND (D54), many cells expressed the neuronal marker MAP2 and subtype-specific markers TBR1, BCL11B (CTIP2), POU3F2 (BRN2), and LHX2 (Figure 4.1B). In addition, we observed putative human-specific outer radial glial cells marked by HOPX (Figure 4.1B) (Pollen et al., 2015; Thomsen et al., 2016). Neuronal activity of D54 cells was confirmed using calcium imaging and pharmacological blocking experiments with tetrodotoxin (TTX), an action potential generation inhibitor. Out of 1148 recorded cells (3 biological replicates), 17 ± 6% demonstrated calcium activity with a frequency of 2.2 ± 0.5 events/min. Calcium activity was blocked in 42.3% of those cells by TTX (Figure 4.1C). These observations are comparable to data obtained by similar methods in recent reports (Edri et al., 2015; Espuny-Camacho et al., 2013; Gaspard et al., 2008; Lancaster et al., 2013; Mariani et al., 2012; Shi et al., 2012).

We profiled the different stages of our *in vitro* differentiation protocol and established its reproducibility (Figure 1D,SB-G) across replicate differentiations and cell lines (H1 and H9) by analyzing populations of cells from each time point (>1x10^6 cells/sample) using transcriptional analysis, flow cytometry, and immunocytochemistry (ICC). Pluripotency markers were rapidly down regulated, and developing cortex markers *EMX2, PAX6*, and *LHX2* appeared between days 6-12 (Figure 4.1E). Additionally, markers of ganglionic eminences (*DLX1*, *ASCL1,* and *GAD1*) as well as mid/hindbrain (*EN2*, *PAX7,*

and *TFAP2B*) were observed (Figure 4.1E). Analysis by flow cytometry showed 7 ± 3% of $SOX2^+$ cells lack PAX6 at D26. Since PAX6 is predominantly expressed in developing pallial progenitors whereas SOX2 is expressed in progenitors across many brain regions based on the Allen Developing Mouse Brain Atlas (Thompson et al., 2014), this suggests that these progenitors may have a non-cortical identity. Immunocytochemistry (ICC) corroborated the presence of $SOX2^+PAX6^-$ cells at D26 as well as the expression of TFAP2B, TH, GAD67 and PBX3 in cultured neurons at D54 (Figure 4.1B,S1H). The diversity of brain cell types present showed that full characterization would require single-cell resolution techniques.

Reporter lines were generated by TALEN-mediated genome engineering (Miller et al., 2011) to allow isolation of live progenitors and neurons, using citrine fluorescent protein gene fused to endogenous *SOX2* (marker of progenitors) or *DCX* (marker of immature neurons), respectively (Figure 4.1F). $SOX2^{Cit/+}$ cells exhibited near-uniform reporter expression in 93 ± 1% of hESCs (n = 3), which decreased to 51 ± 3% (n = 3) $citrine^+$ cells by D54 of differentiation (Figure 4.1G,S2C). In $DCX^{Cit/Y}$ cells, the citrine reporter was not detected in hESCs, was detectable in 3.3 ± 0.9% (n = 6) of differentiating cells at D12, and increased to 50 ± 8% (n = 6) of the cells at D54 (Figure 4.1G). Importantly, both reporters closely mimicked expression of the endogenous protein (Figure S2D), and both differentiated lines produced neurons with the same markers as the parental H1 line (Figure 4.1H). Additionally, reporter lines were generated for early neurogenesis markers *OTX2* and *PAX6* which closely reflected the temporal dynamics from the gene expression data (Figure 4.1E). These lines constitute an important set of tools for dissecting human neurogenesis.

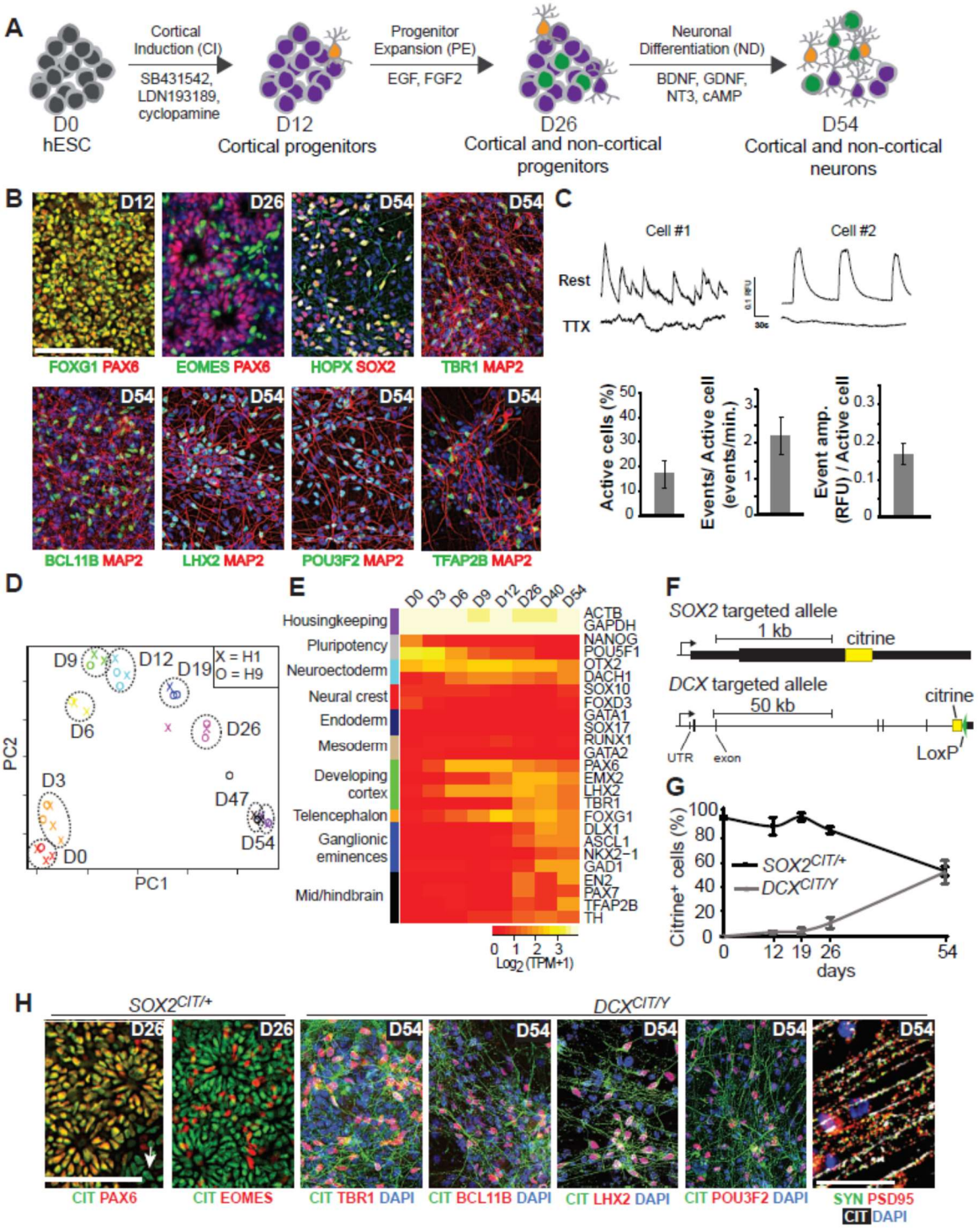


Figure 4.1: *In vitro* neural differentiation generates cortical and non-cortical cells. (A) Schematic representation of *in vitro* neural differentiation of hESCs and *in vivo* early brain patterning. (B) Representative images of ICC staining on D12, D26 and D54 of H1 differentiated cells. Scale bar: 100 μm.

(Figure 4.1, continued) (C) Representative traces of calcium activity as imaged with FURA2-AM (top traces) and after blockade by TTX (bottom traces). Data quantified from three representative experiments (n = 1148 cells at D54, from 3 biological replicates) (bottom); RFU = relative fluorescent units. (D) Principal component analysis of population RNAseq data demonstrates the reproducibility of differentiation methods across multiple experiments from both H1 and H9 stem cell lines. (E) Population RNAseq shows increased expression of genes that mark neuroectoderm, developing cortex, ganglionic eminences and mid/hindbrain. (F) Schematic of the relevant targeted loci of the $SOX2^{Cit/+}$ and $DCX^{Cit/Y}$ reporter cell lines. (G) Quantitation of flow cytometric analysis of percent citrine positive cells during differentiation. Mean ± SD is shown from 3 ($SOX2^{Cit/+}$) and 6 ($DCX^{Cit/Y}$) biological replicates. (H) Representative images of ICC staining of D26 $SOX2^{Cit/+}$ and D54 $DCX^{Cit/+}$ reporter lines. Arrow marks cells that express SOX2 but not PAX6. Scale bar: 100 **μ**m in B; 100 **μ**m in H except for SYN/PSD/CIT/DAPI micrograph where scale bar is 25 **μ**m.

4.2.2. Single-cell profiling and identification of cell types.

To characterize the heterogeneous cells generated in culture, a method based on multiplexed single-cell RNA-seq (Hashimshony et al., 2012) was used at multiple time points (D0, D12, D19, D26, D40, and D54) (Figure 4.2A). We isolated both progenitor-enriched (DCX^{Cit-} and $SOX2^{Cit+}$) and neuron-enriched cells (DCX^{Cit+} and $SOX2^{Cit-}$) resulting in 4368 cells harvested from at least six independent differentiations (Figure 4.2B). We analyzed only cells with >20,000 transcripts (n = 2,684) and subsampled all cells to 20,000 transcripts. PCA and hierarchical clustering separated progenitors and neurons by PC1 and further separated differentiation phase by PC2 (Figure 4.2C). Genes with variance greater than technical noise (estimated by ERCC spike-in controls) were used to drive iterative cell clustering using weighted gene coexpression network analysis (WGCNA (Zhang and Horvath, 2005)) (Figure 4.2B). The iterative WGCNA clustering resulted in 18 clusters of neuronal cell types and 18 clusters of progenitor cell types across all developmental time-points (Figure 4.2D-E).

We recognized, as other have (Paul et al., 2015), that some well-to-well mixing was present with CelSeq multiplexed single-cell RNA-Seq; thus we validated the presence of many transcriptomically identified cell types at D26 and D54 with antibody staining (Figure 4.2F) and an orthogonal method of single cell RNA-Seq (SmartSeq2) that has no multiplexing. Within progenitor and neuronal cell types found at and after D26, we observed frequent co-expression of markers suggestive of telencephalic identity (*LHX2*, *FOXG1,* and *FEZF2*, (Hanashima et al., 2004; Hirata et al., 2004; Porter et al., 1997)). At D26, a distinct progenitor cell type emerged that expressed markers suggestive of mid-hindbrain brain identity (*IRX2* and *POU3F2*, Figure 4.2D). At D40 and D54, we observed additional progenitor and neuronal cell types expressing mid/hindbrain markers such as *TFAP2B*, *PAX3*, *ROBO3*, and *PBX3*, but not *LHX2* (Figure 4.2D, Figure 4.2G). Antibody staining confirmed that LHX2 and POU3F2 have mutually exclusive expression patterns within neural progenitors at D26 and within neurons at D54 (Figure 4.2F). Surprisingly, although EOMES is often used as a marker for intermediate progenitor cells (IPCs), many $EOMES^{+}$ cells were also $TBR1^{+}$ (Figure 4.2F) and did not express cell cycle markers, suggesting these cells may be early-born preplate cells (Bulfone et al., 1999). Together, these data corroborate our RNA-seq data and suggest that we generated regionally diverse cell types.

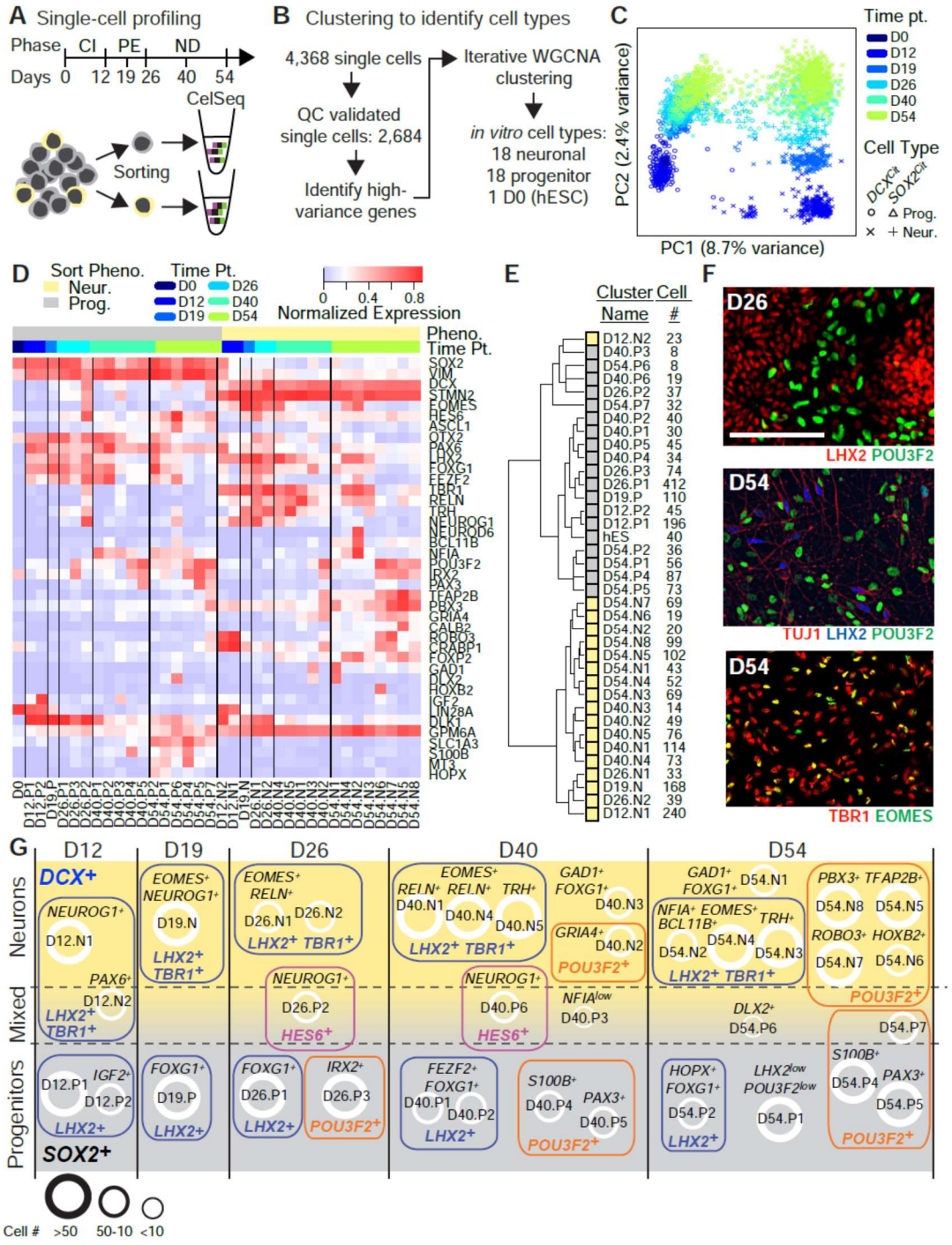


Figure 4.2: Identification of cell types through single-cell transcriptomics.(A) Single cell-profiling strategy from differentiation to single-cell library preparation. (B) Methodology of cell type identification from single cell RNA-Seq data. (C) Principal component analysis of all single cells used for analysis based on high variance genes (Table S1). (D) Normalized expression of marker genes for each cell type identified by single-cell transcriptomics.

(Figure 4.2, continued) (E) Dendrogram of hierarchical relationships of cell types and number of cells in each cell type. Hierarchical clustering by Ward's method using all genes differentially expressed between any two cell types (Table S1). (F) Fluorescence micrographs at D26 (*top*) and D54 (*middle*) show that POU3F2 and LHX2 mark primarily non-overlapping cell types. *Bottom*, EOMES and TBR1 expression in D54 cells. Scale in top panel is 100 μm and is the same for all panels. (G) Cell types identified at D12, D19, D26, D40, and D54. Cell type name, distinguishing molecular markers, and number of cells are indicated. Neuronal cell types (*yellow*) were defined as cell types with strong *DCX* expression, progenitors with strong *SOX2* expression (*gray*), and mixed cell types express both *SOX2* and *DCX* (*gray-yellow transition*).

4.2.3. Cell types show forebrain and mid/hindbrain regional identities

To characterize the cell types generated *in vitro*, we compared the single-cell transcriptomes to the BrainSpan Atlas of the Developing Human Brain (Miller et al., 2014) and the Developing Mouse Brain Atlas (Thompson et al., 2014). We focused on the conserved co-expressed gene modules identified by WGCNA (Langfelder and Horvath, 2007) that distinguished both the hESC-derived cell types and brain regions. We assessed the statistical significance of conserved gene modules based on permutation analysis (Langfelder et al., 2011). We estimated the similarity between differentiated D54 neurons and brain regions by comparing the Pearson correlation coefficient based on these conserved co-expressed gene modules. Neuron clusters D54.N2 and N4 (marked by *LHX2* and *TBR1*) showed strong correlation to cortex but weak correlation to mid/hindbrain regions in both human and mouse at E13.5 (Figure 4.3A-B). In contrast, *POU3F2*$^{+}$ neuron clusters D54.N5-N8 showed weak correlation to cortex but strong correlation to mid/hindbrain, while D54.N3 correlated best to diencephalon (forebrain). Finally, neuron cluster D54.N1 (*GAD1*$^{+}$) showed high correlation to the ganglionic eminences (Figure 4.3A-B). Based on this analysis, our data suggests that LHX2 is a good marker for human cortical progenitors. To test this hypothesis, we performed antibody staining on human

mid-gestational (122-132 days post conception (dpc)) brain tissue and confirmed that *LHX2* is a marker of human cortical progenitors and is not expressed in hindbrain progenitors (Figure 4.3C). Markers associated with the LHX2-expressing neuronal cell types, $BCL11B^{+}$ and $NFIA^{+}NeuN^{+}$, were also highly expressed in neurons in the cortical plate and intermediate zone of the cortex, respectively, but were rare and weakly expressed within hindbrain cells (Figure 4.3D). In contrast, markers of the *POU3F2*-expressing neuron clusters, TFAP2B and ROBO3, were present in hindbrain neurons but absent from cortex (Figure 4.3D). Thus, the *POU3F2*-expressing clusters of neurons likely correspond to posterior brain neuron types while *LHX2*-expressing clusters correspond best to forebrain cell types. As a whole, these data demonstrate progenitors and neurons belonging to multiple human brain regions are generated *in vitro*.

To assess the molecular similarity between the *in vitro* $LHX2^{+}$ cell types and primary cortical cell types at a single-cell level, we compared neurons and progenitors prospectively isolated from fixed primary human cortical samples using FRISCR (Thomsen et al., 2016). We directly compared 512 single cells from human cortical samples (two independent brains aged 108 days post conception (dpc) and 96 dpc) to the *in vitro* differentiated cells. FRISCR allowed targeted sampling of both progenitors ($SOX2^{+}PAX6^{+}TuJ1^{-}$) and neurons ($SOX2^{-}PAX6^{-}TuJ1^{+}$), which were then each further stratified by EOMES expression to reveal $EOMES^{+}$ intermediate progenitors and $EOMES^{+}$ neurons, respectively (Figure 4.4A). FRISCR data from fetal tissue was compared to SmartSeq2 single-cell RNAseq data from $LHX2^{+}$ cultured cells at D26 and D54, and good concordance was found between progenitors and neurons (Figure 4.4B). As cultured progenitors aged, they bore a stronger resemblance to primary progenitors. Interestingly,

primary IPCs appear distinct from neurons or progenitors, whereas cultured cells that express IPC genes (including *EOMES*) strongly resemble neurons (Figure 4.4B). Indeed, among *EOMES*$^{+}$ cells, fetal cells have higher expression of the progenitor eigengene (the average expression of the genes progenitor module) than hESC-derived cells ($P = 3 \times 10^{-7}$, t-test), and hESC-derived cells exhibit higher expression of the neuron eigengene than fetal cells ($P = 3 \times 10^{-20}$, t-test). Further, several collections of genes were specifically expressed in cultured cells or primary cells. Genes that mark oRG cells such as *HOPX* and *TNC* were largely absent from *in vitro* progenitors, indicating that although a few HOPX^{+} cells were observed by ICC and gene expression (Figure 4.1, Figure 4.2), the majority of hESC-derived progenitors lack oRG characteristics. Indeed, gene ontology analysis showed significant enrichment of gliogenesis and oligodendrocyte differentiation genes in primary progenitors, while embryonic development, extracellular matrix, and stress response gene classes were enriched in hESC-derived progenitors (Figure 4.4C). Lastly, the hESC-derived neurons lacked expression of *SATB2* (a marker of later-born callosal-projecting neurons) and express the Cajal-Retzius cell marker *RELN*. We were surprised by the lack of *SATB2* expression since it was detected with a commonly used anti-SATB2 antibody, suggesting the antibody may also detect other species, such as SATB1. These data show that at a single-cell level, hESC-derived progenitors and neurons are molecularly similar to primary cortical progenitors and are producing some of the earliest cortical neurons but are not yet producing more mature SATB2^{+} neurons.

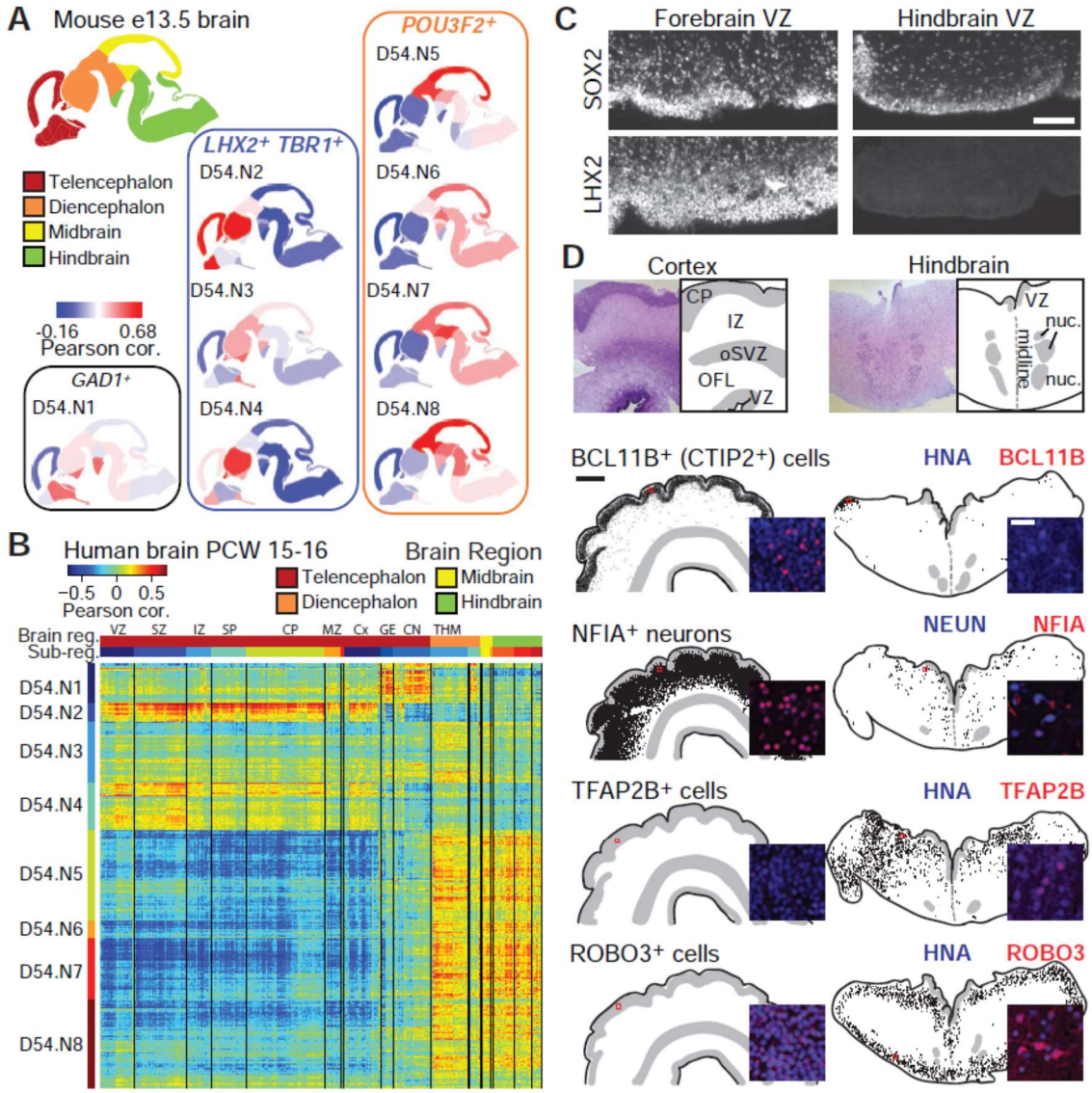


Figure 4.3: Stem cell-derived cell types resemble forebrain and mid/hindbrain cells types. Correlation of D54 neuronal cell types to e13.5 Allen Brain Atlas of the Developing Mouse Brain based on genes differentially expressed between cell types and tissue regions. Mouse regional gene expression levels derived from *in situ* hybridization staining intensity. (B) Correlation of single D54 neurons with regions of the human brain from the Brainspan Atlas of the Developing Human Brain. Correlation based on genes differentially expressed between cell types and tissue regions. (C-D) Fluorescence micrographs of 122 dpc cortex and 132 dpc hindbrain. (C) LHX2 marks human cortical but not hindbrain progenitors, while SOX2 marks progenitors in both regions. Scale is 100 µm.

(Figure 4.3, continued) (D) Immunohistochemistry of cortical and hindbrain cell type markers. *Top:* Nissl stain and representation of tissue architecture are shown; *below*: tissue representation based on DAPI staining. VZ ventricular zone, OFL outer fiber layer, oSVZ outer subventricular zone, IZ intermediate zone, CP cortical plate, nuc medullary nuclei. The entire tissue section was scored and each dot represents a positive cell. Scale is 1 mm. Inset: fluorescence micrograph showing a representative image (location indicated by red box); scale is 25 μm.

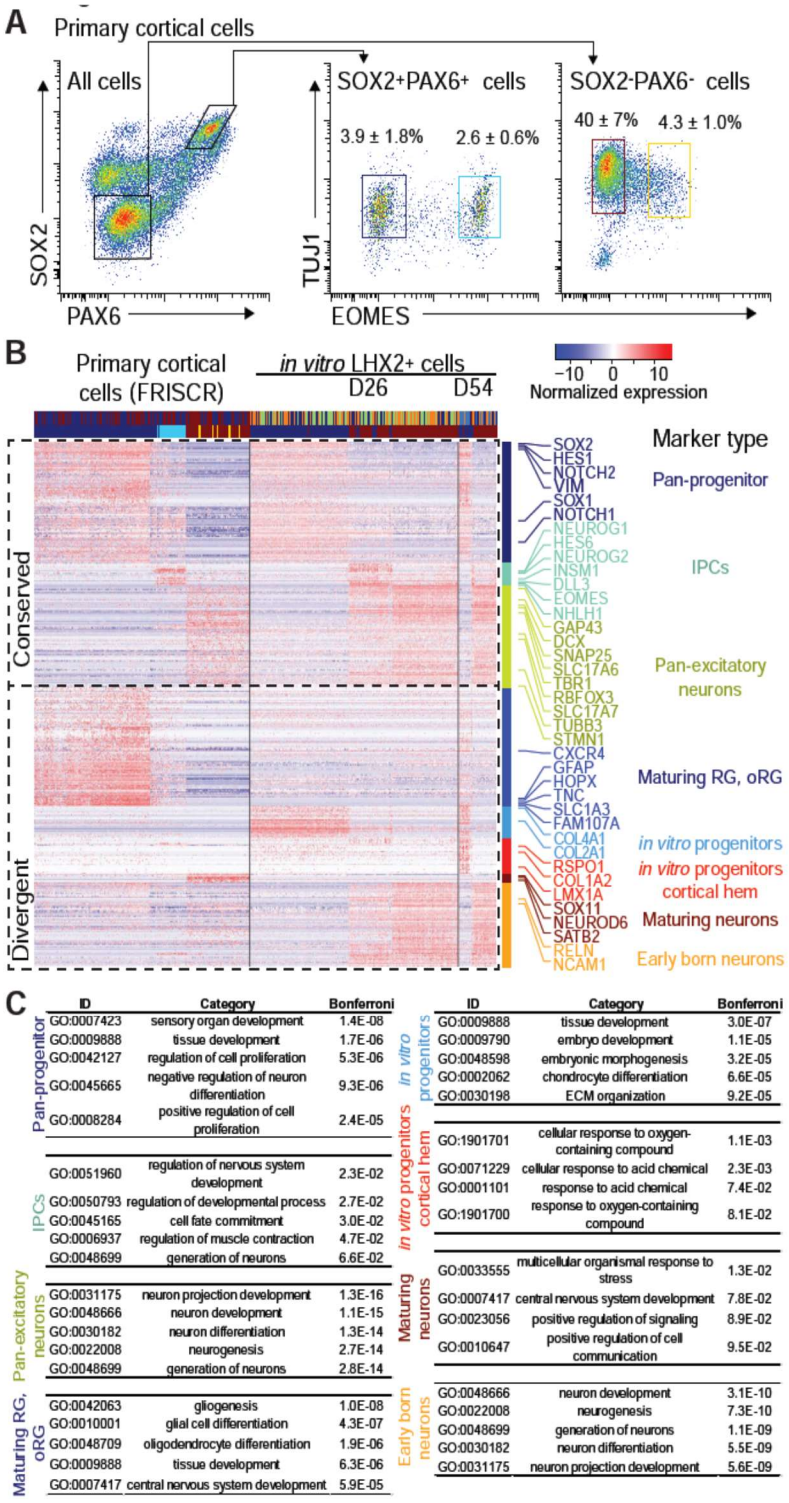


Figure 4.4: Comparison of single stem cell-derived forebrain cells to primary human single cells. (A) Flow cytometry plot showing primary cell populations that were sorted (n = 4) and profiled using FRISCR (n = 2). Mean ± SD of population percentage derived from four brains.

(Figure 4.4, continued) (B) Expression of conserved and divergent gene modules between primary human cortical single cells and *in vitro*-differentiated progenitors and neurons at D26 and D54. (C) Results of gene ontology analysis of conserved and divergent gene expression modules. The top five most significant biological processes with a Bonferroni correction value $<10^{-1}$ are shown.Correlation of D54 neuronal cell types to e13.5 Allen Brain Atlas of the Developing Mouse Brain based on genes differentially expressed between cell types and tissue regions.

4.2.4. An inferred lineage tree with forebrain and mid/hindbrain branches.

The clustering of single cell transcriptomes resulted in identification of different "cell types," which can be linked either as branches in a lineage tree or states along a differentiation trajectory. To resolve the hierarchy of these cell types with putative lineal relationships, typical techniques use distance metrics calculated from the high-dimensional molecular data (Shin et al., 2015; Trapnell et al., 2014); however, the number of master-like molecules instructing cell fate decisions may be few (Colasante et al., 2015; Takahashi and Yamanaka, 2006; Vierbuchen et al., 2010). Furthermore, lineage algorithms based on transcriptomic data usually model a progression and not bifurcations (Shin et al., 2015; Trapnell et al., 2014). Indeed, a low-dimensional projection of the single-cell data like PCA (Figure 4.2C) does not readily suggest a biologically meaningful linkage based on spanning-tree methods, which are common in modeling progressions as opposed to multiple bifurcations.

Here, we used a recently developed computational technique that uses a Bayesian approach to simultaneously infer cell clusters' lineage relationships between the clusters as well as the key set of markers and transition genes that define these relationships (Chapter 2, Chapter 3). This method analyzes the relationships between all the clusters, three at a time. To make these inferences, the forward model for the Bayes technique assumes that good marker genes are uniquely expressed in a cell cluster, while genes

establishing relationships shows shared expression between two of the three clusters. Using this forward model, this technique generates a limited number of high-confidence hypotheses about cellular relationships and the molecular drivers of patterning. Briefly, we determined relative relationships between all possible triplets of cell types at neighboring time points, assessed their putative lineage, and identified genes with expression patterns reflecting this relationship. In each case, the three types of cells were separated in a subset of transcription factor space, and one of the types is the intermediate. It is important to note that the existence of an intermediate state in a triplet represents two possibilities: the intermediate type is the parent, leading to two daughters, or alternatively, the intermediate type is a transition between the other two types. To demonstrate, we show the transcription factor expression for two triplets that show strong evidence of having an intermediate state (Figure 4.5A).

After identifying 118 high confidence triplets of cell types that showed evidence of a lineage relationship, we assembled these triplets into the most parsimonious putative *in silico* lineage tree (Figure 4.5B). The tree was rooted at D12 and assembled iteratively using information from triplets containing successively more mature time points. Cell types that could not be linked to other cell types with high confidence (such as D54.N1) we omitted from the lineage tree. Overall, our *in silico* tree suggests a major branch point separating the *POU3F2*-expressing and *LHX2*-expressing types and also identifies potential transcription factor candidates involved at specific branch points.

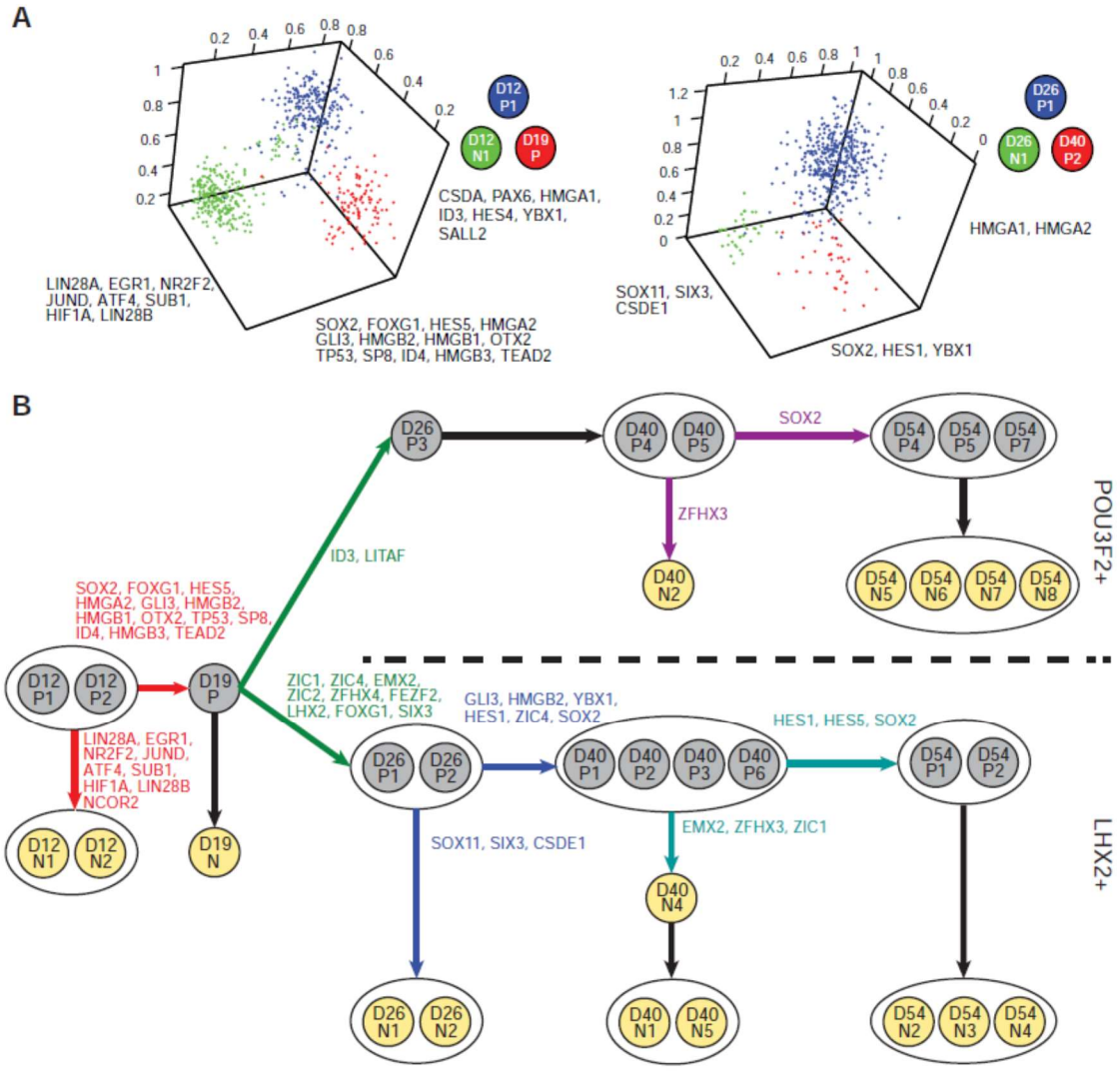


Figure 4.5: A lineage tree from single-cell transcriptomics (A) Examples of two triplets of transcriptomic cell types showing strong evidence for an intermediate state (in blue). In each triplet, the non-intermediate states (red and green) express genes only along one of the two horizontal axes, whereas the intermediate state expresses both sets of genes and also expresses a set of marker genes (vertical axis) that are not highly expressed in either of the other two states. Axis values represent means of normalized gene expression over all the genes on a given axis. Transcriptomic types are named as in Figure 2. (B) *In silico* lineage tree assembled from triplets showing strong evidence of an intermediate state. Arrows indicate proposed lineage/progression links, and key asymmetrically regulated genes are listed for each putative branch point. The tree segregates into two major branches, labeled as the POU3F2+ and LHX2+ branches. Circles around groups of types indicate that those types are not distinguishable in terms of lineage or progression using the tree building algorithm. Transcriptomic types are named and colored as in Figure 2 (progenitors in grey, post-mitotic neurons in yellow).

4.2.5. Predicted transcriptional regulators of the lineage tree.

The transcription factors identified by the Bayesian lineage framework tell a compelling narrative, with some inferences corroborated in the literature (Figure 4.5B). D12 progenitors give rise to either D12 neurons or D19 progenitors. The transcription factors predicted to control the progression to D19 progenitors include FOXG1, which suppresses early cortical neuron fates (Hanashima et al., 2004), and HMGA2 and HMGB3, known to regulate the balance of self-renewal and differentiation in multiple stem cell compartments (Nemeth et al., 2006; Nishino et al., 2008). From D19, the progenitors branch into *POU3F2*$^{+}$ and *LHX2*$^{+}$ progenitors. The transcription factors accompanying specification to the *LHX2*$^{+}$ branch are involved in cortical patterning: *ZIC1* and *ZIC2* are markers dorsal neural tube (Aruga et al., 1994; Brown et al., 1998), *EMX2* is expressed in an opposing gradient to *PAX6* to specify rostral cortical identity (Bishop et al., 2000; Hamasaki et al., 2004), and *SIX3* promotes anterior brain identity (Oliver et al., 1995; Wallis et al., 1999). At D26, the progenitors give rise to neurons that are likely to eventually mature into L5/6 subcortical projection neurons; SOX11 regulates expression of *FEZF2* (Shim et al., 2012), and SIX3 continues to maintain anterior brain identity. The D26 progenitors also give rise to D40 and D54 progenitors marked by HES1, which is known to oscillate to maintain progenitor state prior to selection of lineage (Imayoshi et al., 2013).

In the *POU3F2*$^{+}$ branch, the initial progression from D19 to D26 progenitors includes *ID3,* which presumably causes delayed differentiation. The D40 progenitors undergo a lineage branch that divides *ZFHX3* and *SOX2*; *ZFHX3* marks the mantle zone (where early neurons reside in e13.5 mouse) and is excluded from the SOX2^{+} VZ (Thompson et al., 2014). Of particular note, *ZFHX3* is expressed in immature neurons

throughout the brain but not the cortex. In summary, the Bayesian framework revealed regulators of brain development and thus suggests new testable hypotheses of human brain lineage and development.

4.2.6. Clonal analysis confirms forebrain cell types segregating from mid/hindbrain cell types

The *in silico* lineage tree predicts that *POU3F2*-expressing neuronal clusters (D54 clusters D54.N5-8) and *LHX2*-expressing neuronal clusters (D54.N2-4) arise from distinct progenitors by D26. To test this hypothesis, we undertook clonal analysis of cell fate from D26 progenitors using a viral barcoding strategy. We cloned a 10-bp degenerate barcode library in the 3' UTR of a *tdTomato* expression cassette and packaged this into VSV-G pseudotyped retrovirus, which showed no obvious tropism bias using classic markers of neurogenesis. Clones tracked by daily time-lapse microscopy display neuronal and non-neuronal morphologies, and clones analyzed by ICC demonstrate subtype diversity. In addition, strong transcriptomic correlation was observed between infected and uninfected cells.

To perform barcoded lineage tracing, cells infected at D27 (following re-plating at D26) were harvested at D54 (Figure 4.6A) and processed by SmartSeq2 to simultaneously recover the barcode (clone association) and the transcriptome (cell type). To match D54 cells profiled by SmartSeq2 to cell types identified by CelSeq, we developed a consensus gene set consisting of co-expression gene modules conserved between the two methods. From cultures containing ~50 total clones, 176 tdT^+ sorted cells were sequenced, with 111 cells yielding detectable barcodes. Of 29 unique clones (barcodes), 16 contained more than one cell, and 11 spanning more than one cell type. All five clones contain cell types from

the *POU3F2* lineage branch, showing that the indicated cell types are lineally related (Figure 4.6C). Further, we did not detect multicellular clones containing cells from both *POU3F2* and *LHX2* branches. Importantly, these data are consistent with the computational lineage tree, indicating that distinct progenitors at D27 have region-specific fates.

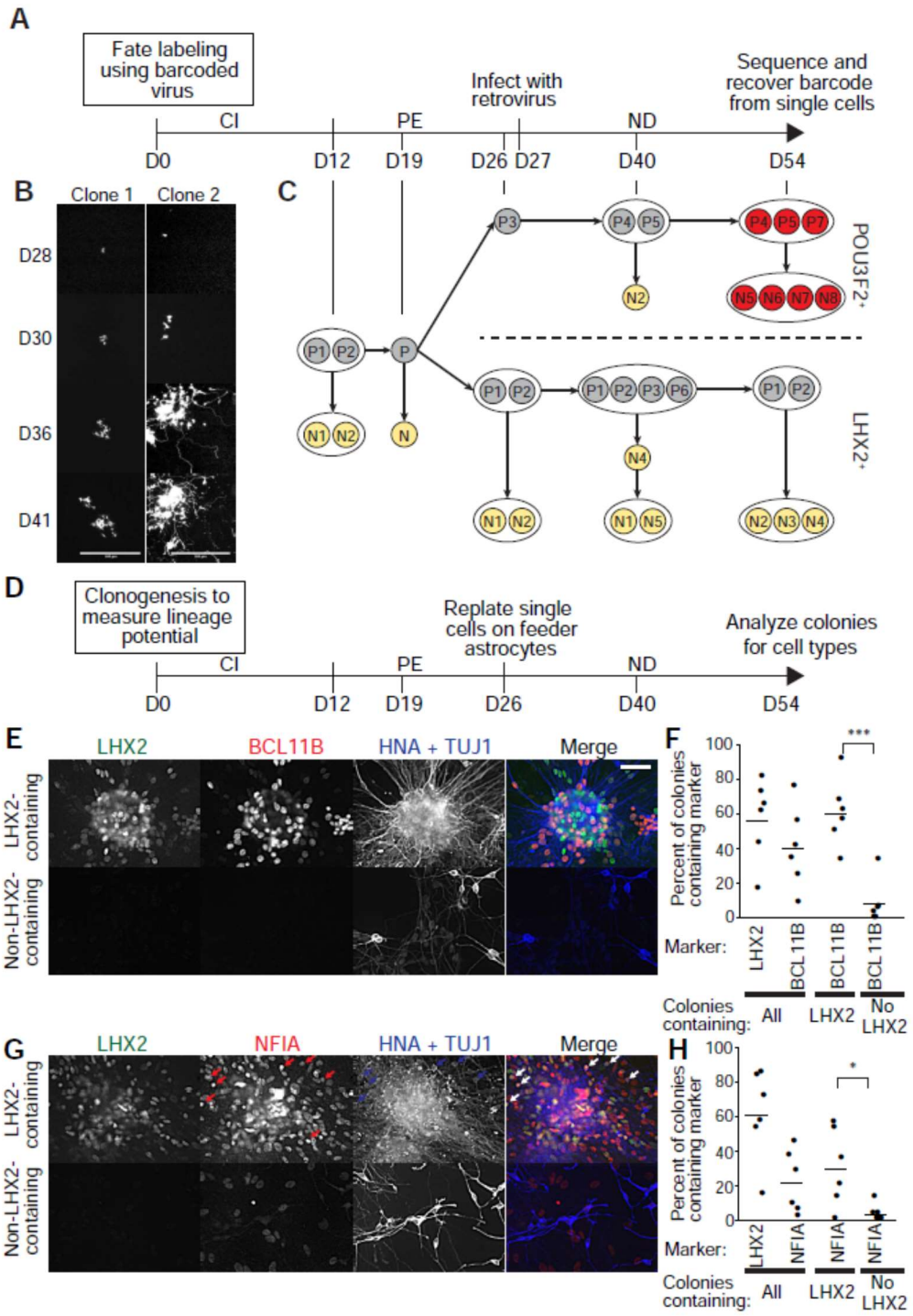


Figure 4.6: Clonal analysis confirms distinct POU3F2 and LHX2 branches of the human brain lineage tree. (A) Schematic of viral barcoding experiment, indicating time of infection and collection. (B) Representative fluorescence micrographs captured from live differentiating clones, exhibiting non-neuronal (Clone 1) and neuronal morphologies (Clone 2). Scale bar = 500 µm. (C) Schematic representation of aggregate barcoding analysis. Red indicates cell types with detected lineage relationships.

(Figure 4.6, continued) (D) Clonal analysis of cell-autonomous lineage potential was performed by re-plating D26 progenitors at clonal density on feeder mouse astrocytes, then analyzing outgrown colonies for cell composition at D54 by ICC. (E and G) Example colonies are stained with antibodies for HNA + TuJ1 (blue), LHX2 (green), and either CTIP2 (E, red) or NFIA (G, red). Colonies were grouped into categories of LHX2-containing and non-LHX2-containing. In G, arrows mark $NFIA^+TuJ1^+$ human neurons, but NFIA also marks $LHX2^+TuJ1^-$ human progenitors as well as HNA^- mouse astrocytes. Scale bar, 50 μm. (F) Colonies that contain $LHX2^+$ cells are more likely to contain $CTIP2^+$ cells as compared to colonies lacking $LHX2^+$ cells. (H) Colonies that contain $LHX2^+$ cells are more likely to contain $NFIA^+$ cells as compared to colonies lacking $LHX2^+$ cells. In F and H, six independent differentiations were analyzed, and 25-38 colonies per ICC staining cocktail per experiment were inspected for the presence of cell types. *** $P < .001$, * $P < .05$ by unpaired t-test.

To test whether the differences in cell fate of D26 progenitors were due to cell-autonomous differences in lineage potential and to bypass any possible viral tropism from the viral fate mapping analysis, we generated single-cell clones by seeding D26 progenitors at clonal density (10 cells/well in a 96-well plate) in differentiation medium for four weeks on mouse astrocytes (yielding 2.3 ± 1.4 spatially resolved colonies/ well, Figure 4.6D). Colonies varied in size (1-1000 cells), but those with cortical lineage potential (containing $LHX2^+$ cells) typically contained more cells than colonies lacking $LHX2^+$ cells. Cell-type-specific antibodies were used to assess cell-type composition within colonies (814 colonies analyzed over n = 6 independent experiments). We used human nuclear antigen (HNA) to distinguish human cells from mouse astrocytes, TUBB3 (TuJ1) to locate neurons, and LHX2 to identify clones with cortical lineage potential; these three markers were multiplexed with other cell-type markers identified from the sequencing dataset. One of these cell-type markers is BCL11B, which is a marker of cortical neurons that we observed in our D54 *LHX2*-expressing neuron cluster (Figure 4.2G,3D). Although the prevalence of LHX2-containing colonies and BCL11B-containing colonies varied greatly among independent experiments (ranging from <20% to >75% of clones with potential to generate

each cell type), colonies containing LHX2$^+$ cells much more likely included BCL11B$^+$ neurons than those lacking LHX2$^+$ cells (60 ± 20% versus 8 ± 13%, $P < .001$, unpaired t-test, Figure 6E-F). Antibody staining for NFIA similarly demonstrated that LHX2-containing colonies more frequently included NFIA$^+$TuJ1$^+$ cortical neurons (Figure 4.2G, Figure 4.3D) than non-LHX2-containing colonies (30 ± 20% versus 3 ± 5%, $P < .05$, unpaired t-test, Figure 4.6G-H). In contrast, when we stained colonies with markers of other (posterior brain) cell types that we observed in our sequencing dataset (POU3F2, FOXP2, CALB2 (Calretinin), and CRABP1), we detected no significant associations of these markers with LHX2 within colonies. Together, these clonal outgrowth results corroborate the presence of independent cortical and posterior regional branches of the human neural single-cell lineage tree. Moreover, they suggest that region-specific branches are caused by differences in cell-autonomous commitment among D26 progenitors rather than by stochastic and/or non-cell-autonomous phenomena that occur during neuronal differentiation.

4.3. Discussion

In this study, we present a comprehensive characterization of human brain cell types generated from hESCs using single-cell RNA-seq. We demonstrate the gradual production of multiple cortical and posterior brain neuronal and progenitor cell types with molecular similarity to primary progenitors and neurons. Furthermore, a unified lineage tree predicts that the cell types differentiate along divergent region-specific trajectories, and also highlights a series of known and unknown lineage-specific regulators. We confirmed portions of this lineage tree through clonal analyses of fate and lineage potential which had not been done previously in a hESC differentiation system. Our clonal analyses confirmed

a major predicted branch point during differentiation and, moreover, directly demonstrated that this branch point is a cell-autonomous property of neural progenitors. In total, this study charts human brain region-specific developmental pathways, which is essential to understanding the logic and uniqueness of human neurocircuitry.

Our lineage tree branches mimic established regional differences in neurogenesis. Transcription factors driving cortical patterning include FOXG1 (Hanashima et al., 2004), LHX2 (Porter et al., 1997), and FEZF2 (Hirata et al., 2004). Other transcription factors pattern the mid/hindbrain region such as PAX2/PAX5 (Schwarz et al., 1997) and EN1/EN2 (Liu and Joyner, 2001). We observed both of these transcription factor classes at D26 and beyond, but only cortical markers at D12 (Figure 4.2). Dual SMAD inhibition could cause cortical fate-specification but not fate-restriction, permitting regional plasticity during the progenitor expansion phase. Similarly, cortical identity was shown to be specified by e10.5 but not committed until after e13.5 in mice (Olsson et al., 1997). Alternatively, a small population of cells could escape dorsal telencephalic specification and disproportionally expand during the neural differentiation phase. Regardless the field will require optimized techniques to produce more uniform populations of targeted brain region cell types from hESCs.

Our lineage tree shows how early-born human neural progenitors give rise to early neurons with different regional identities in an *in vitro* system. Neural pan-progenitor genes like *VIM, NES,* and *SOX2* were expressed in progenitors from D12, but markers of maturing radial glia and outer radial glia (oRG) such as *SLC1A3* (*GLAST*), *HOPX*, and *TNC* (Pollen et al., 2015; Thomsen et al., 2016) only begin to be expressed at the latest time point. Since the oRG markers are first expressed throughout the germinal zone before

cells stratify into the outer-subventricular zone (oSZ), these cells likely reflect radial glia, prior to oSZ formation. This staging is consistent with the fact that cultured progenitors best-resemble the earliest progenitor regions of the human developmental atlas. We also detected evidence of the production of very early human neurons. The earliest (D12) DCX^+ neurons express *TBR1* and weak levels of *EOMES and RELN.* Perhaps these D12 DCX^+ neurons correspond to predecessor cells that express TBR1 and are formed during neural tube closing (Bystron et al., 2006). Cells with highest *RELN* levels (presumptive Cajal-Retzius cells (Hevner et al., 2003)) were detected at D26 and D40. Many of these cells also expressed *EOMES*. Although *EOMES* is a canonical marker of proliferative IPCs *in vivo*, most *in vitro* $EOMES^+$ cells were not proliferative and instead had a strong neuronal signature. We believe these are preplate cells where co-expression of EOMES and TBR1 has been reported (Bulfone et al., 1999). Comprehensive single-cell profiling of early human embryonic neural tissues would be needed to confirm this prediction.

One hallmark of the developing cortex is sequential creation of cellular layers that exhibit unique molecular and morphological properties. Although we detected multiple neuronal types, we did not detect many classical (especially upper) layer markers (e.g., SATB2, CUX2). There are several possible reasons for this. First, the culture duration may be limiting (54 days), and therefore progenitors may still be producing early born neurons. Second, classical layer markers may have poor predictive utility at mid-gestation stages: they are only weakly expressed in fetal macaque with dramatic regional and temporal variation (T. Bakken and E. Lein. in review, NIH Blueprint Non-Human Primate Atlas), and furthermore neurons from primary mid-gestational tissues appear relatively homogeneous (Figure 4.4B) as was previously reported (Camp et al., 2015; Pollen et al.,

2014). Together these considerations spotlight single cell RNA-Seq as an unbiased and high-value methodology for cultured cell characterization, rather than classical marker analysis.

In summary our observations suggest our culture system models the earliest steps of human brain development including regional patterning, which is vital because primary samples at these stages are exceedingly rare. This study represents an advance for the field in terms of breadth and depth of cell characterization and provides a vital benchmark dataset to understand the origins and diseases of the human brain. Future synthetic models of human brain will require a similarly atomistic view of brain structure and function to understand its emergent properties, and to uncover its fundamental molecular logic.

4.4. Methods

4.4.1. Genome engineering and hESC culture.

Human H1 or H9 ESCs (WiCell) were maintained with mTeSR1 media (Stem Cell Technologies) on Matrigel (BD) or hES media (DMEM/F12 with 20% KSR; Life Technologies) on CF-1 MEFs (GlobalStem). The TALEN genes targeting *SOX2*, *DCX*, and *OTX2* were made by the Joung lab using the FLASH method (Reyon et al., 2012), and those for *PAX6* were prepared using the REAL method (Sander et al., 2011). Mutations were introduced in HDR donor AI-CN409 to retain protein coding identity and disfavor repeated endonuclease activity following HDR. Engineered lines we generated as previously described (Martinez et al., 2015). See Supplemental Experimental Procedures for more details.

4.4.2. hESC neural differentiation.

hESCs were seeded for a 12-day cortical induction (CI) phase in NIM media with SMAD inhibitors (Chambers 2009) and cyclopamine (Stemgent); reseeded for the progenitor expansion (PE) phase at D12 and D19 in neural stem cell culture media (NSCM) with EGF (Thermo Fisher) and bFGF (Ciccolini and Svendsen, 1998; Tropepe et al., 1999); and finally at D26, re-seed for neural differentiation (ND) with neurogenic/neurotrophic factors BDNF (R&D Systems), GDNF (R&D Systems), NT3 (R&D Systems), and cAMP (Sigma) (Hu et al., 2010) (Figure 1A). For detail, see Supplemental Experimental Procedures. Differentiations were validated by ICC at D12, D26 and D54.

4.4.3. Antibody staining.

For immunocytochemistry (ICC), cells were fixed in 4% PFA for 15 min, blocked and permeabilized in 10% normal goat serum with 0.1% TritonX-100. Primary antibodies were incubated overnight at 4°C and secondary antibodies are listed (Table S3). Immunohistochemistry experiments were performed on 20 µm frozen sections as above. Brain pieces were fixed overnight in 4% PFA/PBS at 4°C, then cryoprotected in 30% sucrose in PBS for 24-48 hours, then embedded in OCT. Cryosections were cut on the coronal plane at 20 µm thickness, then stained as above.

4.4.4. Calcium imaging.

Cells were loaded with 4 µM FURA-2AM (Thermo Fisher) in ND at room temperature for 30 min. After rinsing cells, Ca^{2+} activity was recorded using a 40× objective for 5 min intervals. Images were captured with 300 ms exposures at both 340 nm and 380 nm. Nikon NIS-Elements software was used to analyze events with measurements greater than 0.006 RFU above baseline.

4.4.5. Fetal brain tissue processing.

Human fetal tissue was donated with written informed consent and requirements of the Uniform Anatomical Gift Act and National Organ Transplant Act were followed. Sample age in days post conception was estimated by foot length and dated menstrual cycles. Cortical tissue was identified by morphology, physically disrupted, and then enzymatically dissociated to single cells. Cells were then washed, filtered, counted, and fixed for storage at -80°C until processing by FRISCR.

4.4.6. Single cell transcriptomics.

Single cells were sorted on a FACS Aria (BD) into 96-well collection plates and stored at -80°C. We prepared libraries as previously reported (Hashimshony et al., 2012) with a few modifications (See Supplemental Experimental Procedures), FRISCR was carried out as previously described (Thomsen et al., 2016), and SmartSeq2 sequencing libraries were prepared as previously reported (Picelli et al., 2013). Libraries were the quantified and sequenced on a HiSeq (Illumina).

4.4.7. Lineage inference.

We used the method described in Chapter 2. and Chapter 3. . Briefly, we selected all possible triplets of cell types, then bases on only transcription factor expression data, used a Bayesian formulation to identify the highest-probability topology. Triplets that showed strong evidence of a hierarchical relationship, were assembled in an interative fashion into a tree rooted the tree at D12. For each successive time point, we selected only triplets containing any types from that time point, the previous time point, and the following time point, and linked these triplets to build the tree. Where conflicting triplets were obtained,

the topology with higher probability was selected. We then identified asymmetrically expressed transcription factors.

4.4.8. Viral clonal analysis.

A barcoded retroviral plasmid library was constructed by transferring the tdT expression construct from the Ai9 plasmid (Addgene plasmid 22799) to the pMXs backbone (Addgene plasmid 13367), followed by a 10 bp barcode library. The barcoded plasmid library pMXs-Ai9-BC was packaged into retrovirus particles (pseudotyped with the VSV-G) and concentrated and titered on 293T cells. At D27, 4 to 6 x 10^3 IFU were inoculated per well of a 24-well plate (1.5 x 10^6 cells) and differentiation as normal to D54, to yield 20-50 tdT^+ colonies. Daily fluorescent images were taking for some clones to monitor expansion. Single cells dissociated and processed into SmartSeq2 cDNA libraries and sequenced as described above (see Experimental Methods, Single cell transcriptomics). Beyond standard data alignment and processing, raw FASTQ data were aligned to a CAG-tdT-WPRE-polyA reference index and aligned with Bowtie2 and barcodes were identified. Random forest classification was used to identify the cell types that best corresponded to the barcoded cells. See Supplemental Experimental Procedures for more details.

4.4.9. Progenitor potential assay by clonal outgrowth.

Single cells from D26 were plated onto mouse astryocytes at clonal density (10 cells/well of a 96 well plate). Single cell colonies were grown in modified NSCM media for two weeks, followed by differentiation for 4 weeks in ND media. Colonies were fixed and analyzed by ICC using antibodies against HNA, TUBB3, POU3F2, LHX2, CRABP1, Calretinin, FOXP2, GAD67, CTIP2, and NFIA.

Acknowledgements

We wish to thank the Allen Institute founders, P. G. Allen and J. Allen, for their vision, encouragement and support. We wish to thank A. Bernard and E. Lein for assistance with human specimen procurement, J. Miller for assistance in analysis of BrainSpan Atlas of the Developing Human Brain. Human primary samples were received from the "Laboratory of Developmental Biology," supported by NIH Award Number 5R24HD000836 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. SR was supported in part by the NIH Directors Pioneer Award 5DP1MH099906-03 and National Science Foundation grant PHY-0952766.

References

Aruga, J., Yokota, N., Hashimoto, M., Furuichi, T., Fukuda, M., and Mikoshiba, K. (1994). A novel zinc finger protein, zic, is involved in neurogenesis, especially in the cell lineage of cerebellar granule cells. Journal of neurochemistry *63*, 1880-1890.

Bishop, K.M., Goudreau, G., and O'Leary, D.D. (2000). Regulation of area identity in the mammalian neocortex by Emx2 and Pax6. Science *288*, 344-349.

Brown, S.A., Warburton, D., Brown, L.Y., Yu, C.Y., Roeder, E.R., Stengel-Rutkowski, S., Hennekam, R.C., and Muenke, M. (1998). Holoprosencephaly due to mutations in ZIC2, a homologue of Drosophila odd-paired. Nature genetics *20*, 180-183.

Bulfone, A., Martinez, S., Marigo, V., Campanella, M., Basile, A., Quaderi, N., Gattuso, C., Rubenstein, J.L.R., and Ballabio, A. (1999). Expression pattern of the Tbr2 (Eomesodermin) gene during mouse and chick brain development. Mechanisms of Development *84*, 133-138.

Bystron, I., Rakic, P., Molnar, Z., and Blakemore, C. (2006). The first neurons of the human cerebral cortex. Nature neuroscience *9*, 880-886.

Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Brauninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M.*, et al.* (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proceedings of the National Academy of Sciences of the United States of America *112*, 15672-15677.

Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotech *27*, 275-280.

Ciccolini, F., and Svendsen, C.N. (1998). Fibroblast growth factor 2 (FGF-2) promotes acquisition of epidermal growth factor (EGF) responsiveness in mouse striatal precursor cells: identification of neural precursors responding to both EGF and FGF-2. The Journal of neuroscience : the official journal of the Society for Neuroscience *18*, 7869-7880.

Colasante, G., Lignani, G., Rubio, A., Medrihan, L., Yekhlef, L., Sessa, A., Massimino, L., Giannelli, S.G., Sacchetti, S., Caiazzo, M.*, et al.* (2015). Rapid Conversion of Fibroblasts into Functional Forebrain GABAergic Interneurons by Direct Genetic Reprogramming. Cell stem cell *17*, 719-734.

Edri, R., Yaffe, Y., Ziller, M.J., Mutukula, N., Volkman, R., David, E., Jacob-Hirsch, J., Malcov, H., Levy, C., Rechavi, G.*, et al.* (2015). Analysing human neural stem cell ontogeny by consecutive isolation of Notch active neural progenitors. Nature communications *6*, 6500.

Espuny-Camacho, I., Michelsen, K.A., Gall, D., Linaro, D., Hasche, A., Bonnefont, J., Bali, C., Orduz, D., Bilheu, A., Herpoel, A.*, et al.* (2013). Pyramidal neurons derived from

human pluripotent stem cells integrate efficiently into mouse brain circuits in vivo. Neuron *77*, 440-456.

Gaspard, N., Bouschet, T., Hourez, R., Dimidschstein, J., Naeije, G., van den Ameele, J., Espuny-Camacho, I., Herpoel, A., Passante, L., Schiffmann, S.N., *et al.* (2008). An intrinsic mechanism of corticogenesis from embryonic stem cells. Nature *455*, 351-357.

Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251-255.

Habela, C.W., Song, H., and Ming, G.L. (2015). Modeling synaptogenesis in Schizophrenia and Autism using human iPSC derived neurons. Molecular and cellular neurosciences.

Hamasaki, T., Leingartner, A., Ringstedt, T., and O'Leary, D.D. (2004). EMX2 regulates sizes and positioning of the primary sensory and motor areas in neocortex by direct specification of cortical progenitors. Neuron *43*, 359-372.

Hanashima, C., Li, S.C., Shen, L., Lai, E., and Fishell, G. (2004). Foxg1 suppresses early cortical cell fate. Science *303*, 56-59.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell reports *2*, 666-673.

Hevner, R.F., Neogi, T., Englund, C., Daza, R.A., and Fink, A. (2003). Cajal-Retzius cells in the mouse: transcription factors, neurotransmitters, and birthdays suggest a pallial origin. Brain research Developmental brain research *141*, 39-53.

Hirata, T., Suda, Y., Nakao, K., Narimatsu, M., Hirano, T., and Hibi, M. (2004). Zinc finger gene fez-like functions in the formation of subplate neurons and thalamocortical axons. Developmental dynamics : an official publication of the American Association of Anatomists *230*, 546-556.

Hook, V., Brennand, Kristen J., Kim, Y., Toneff, T., Funkelstein, L., Lee, Kelly C., Ziegler, M., and Gage, Fred H. (2014). Human iPSC Neurons Display Activity-Dependent Neurotransmitter Secretion: Aberrant Catecholamine Levels in Schizophrenia Neurons. Stem Cell Reports *3*, 531-538.

Hu, B.Y., Weick, J.P., Yu, J., Ma, L.X., Zhang, X.Q., Thomson, J.A., and Zhang, S.C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. Proceedings of the National Academy of Sciences of the United States of America *107*, 4335-4340.

Imayoshi, I., Isomura, A., Harima, Y., Kawaguchi, K., Kori, H., Miyachi, H., Fujiwara, T., Ishidate, F., and Kageyama, R. (2013). Oscillatory control of factors determining multipotency and fate in mouse neural progenitors. Science *342*, 1203-1208.

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.

Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, Michael C., Gao, F., Chen, L., Wang, G.-Z., Luo, R., Preuss, Todd M.*, et al.* (2012). Human-Specific Transcriptional Networks in the Brain. Neuron *75*, 601-617.

Lancaster, M.A., Renner, M., Martin, C.A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373-379.

Langfelder, P., and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. BMC systems biology *1*, 54.

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is my network module preserved and reproducible? PLoS computational biology *7*, e1001057.

Liu, A., and Joyner, A.L. (2001). Early anterior/posterior patterning of the midbrain and cerebellum. Annual review of neuroscience *24*, 869-896.

Lui, J.H., Hansen, D.V., and Kriegstein, A.R. (2011). Development and evolution of the human neocortex. Cell *146*, 18-36.

Lui, J.H., Nowakowski, T.J., Pollen, A.A., Javaherian, A., Kriegstein, A.R., and Oldham, M.C. (2014). Radial glia require PDGFD-PDGFRbeta signalling in human but not mouse neocortex. Nature *515*, 264-268.

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.

Mariani, J., Simonini, M.V., Palejev, D., Tomasini, L., Coppola, G., Szekely, A.M., Horvath, T.L., and Vaccarino, F.M. (2012). Modeling human cortical development in vitro using induced pluripotent stem cells. Proceedings of the National Academy of Sciences of the United States of America *109*, 12770-12775.

Martinez, R.A., Stein, J.L., Krostag, A.R., Nelson, A.M., Marken, J.S., Menon, V., May, R.C., Yao, Z., Kaykas, A., Geschwind, D.H.*, et al.* (2015). Genome engineering of isogenic human ES cells to model autism disorders. Nucleic Acids Res *43*, e65.

Miller, J.A., Ding, S.L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K.*, et al.* (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199-206.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J.*, et al.* (2011). A TALE nuclease architecture for efficient genome editing. Nature biotechnology *29*, 143-148.

Nemeth, M.J., Kirby, M.R., and Bodine, D.M. (2006). Hmgb3 regulates the balance between hematopoietic stem cell self-renewal and differentiation. Proceedings of the National Academy of Sciences of the United States of America *103*, 13783-13788.

Nishino, J., Kim, I., Chada, K., and Morrison, S.J. (2008). Hmga2 promotes neural stem cell self-renewal in young but not old mice by reducing p16Ink4a and p19Arf Expression. Cell *135*, 227-239.

Oberheim, N.A., Takano, T., Han, X., He, W., Lin, J.H., Wang, F., Xu, Q., Wyatt, J.D., Pilcher, W., Ojemann, J.G.*, et al.* (2009). Uniquely hominid features of adult human astrocytes. The Journal of neuroscience : the official journal of the Society for Neuroscience *29*, 3276-3287.

Oliver, G., Mailhos, A., Wehr, R., Copeland, N.G., Jenkins, N.A., and Gruss, P. (1995). Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. Development *121*, 4045-4055.

Olsson, M., Campbell, K., and Turnbull, D.H. (1997). Specification of Mouse Telencephalic and Mid-Hindbrain Progenitors Following Heterotopic Ultrasound-Guided Embryonic Transplantation. Neuron *19*, 761-772.

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.

Picelli, S., Bjorklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature methods *10*, 1096-1098.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A.*, et al.* (2015). Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell *163*, 55-67.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P.*, et al.* (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature biotechnology *32*, 1053-1058.

Porter, F.D., Drago, J., Xu, Y., Cheema, S.S., Wassif, C., Huang, S.P., Lee, E., Grinberg, A., Massalas, J.S., Bodine, D.*, et al.* (1997). Lhx2, a LIM homeobox gene, is required for eye, forebrain, and definitive erythrocyte development. Development *124*, 2935-2944.

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. Nature reviews Neuroscience *10*, 724-735.

Reilly, S.K., Yin, J., Ayoub, A.E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., and Noonan, J.P. (2015). Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science *347*, 1155-1159.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nature biotechnology *30*, 460-465.

Ricciardi, S., Ungaro, F., Hambrock, M., Rademacher, N., Stefanelli, G., Brambilla, D., Sessa, A., Magagnotti, C., Bachi, A., Giarda, E.*, et al.* (2012). CDKL5 ensures excitatory synapse stability by reinforcing NGL-1–PSD95 interaction in the postsynaptic compartment and is impaired in patient iPSC-derived neurons. Nature cell biology *14*, 911-923.

Sander, J.D., Cade, L., Khayter, C., Reyon, D., Peterson, R.T., Joung, J.K., and Yeh, J.R. (2011). Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nature biotechnology *29*, 697-698.

Schwarz, M., Alvarez-Bolado, G., Urbanek, P., Busslinger, M., and Gruss, P. (1997). Conserved biological function between Pax-2 and Pax-5 in midbrain and cerebellum development: evidence from targeted mutations. Proceedings of the National Academy of Sciences of the United States of America *94*, 14518-14523.

Shi, Y., Kirwan, P., Smith, J., Robinson, H.P., and Livesey, F.J. (2012). Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. Nature neuroscience *15*, 477-486, S471.

Shim, S., Kwan, K.Y., Li, M., Lefebvre, V., and Sestan, N. (2012). Cis-regulatory control of corticospinal system development and evolution. Nature *486*, 74-79.

Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G.L.*, et al.* (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell stem cell *17*, 360-372.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663-676.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T.*, et al.* (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature neuroscience.

Thompson, C.L., Ng, L., Menon, V., Martinez, S., Lee, C.K., Glattfelder, K., Sunkin, S.M., Henry, A., Lau, C., Dang, C.*, et al.* (2014). A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. Neuron *83*, 309-323.

Thomsen, E.R., Mich, J.K., Yao, Z., Hodge, R.D., Doyle, A.M., Jang, S., Shehata, S.I., Nelson, A.M., Shapovalova, N.V., Levi, B.P.*, et al.* (2016). Fixed single-cell transcriptomic characterization of human radial glial diversity. Nature methods *13*, 87-93.

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. Nature biotechnology *32*, 381-386.

Tropepe, V., Sibilia, M., Ciruna, B.G., Rossant, J., Wagner, E.F., and van der Kooy, D. (1999). Distinct neural stem cells proliferate in response to EGF and FGF in the developing mouse telencephalon. Developmental biology *208*, 166-188.

van de Leemput, J., Boles, N.C., Kiehl, T.R., Corneo, B., Lederman, P., Menon, V., Lee, C., Martinez, R.A., Levi, B.P., Thompson, C.L.*, et al.* (2014). CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. Neuron *83*, 51-68.

Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Sudhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. Nature *463*, 1035-1041.

Wallis, D.E., Roessler, E., Hehr, U., Nanni, L., Wiltshire, T., Richieri-Costa, A., Gillessen-Kaesbach, G., Zackai, E.H., Rommens, J., and Muenke, M. (1999). Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. Nature genetics *22*, 196-198.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C.*, et al.* (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138-1142.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol *4*, Article17.

Chapter 5. Appendix: Mathematical derivation of Bayesian Framework

Bayesian Framework for inferring cluster identities, state transitions, and marker and transition genes simultaneously

5.1. Notation; Bayes' Rule

Given gene expression data from single cells $\{g_i\}$, we built a probabilistic framework to simultaneously infer cell cluster identities, $\{C\} \equiv \{c_A, c_B, \dots\}$, the sequence of transitions T between these clusters, the key sets of marker genes $\{\alpha_i\}$ that define each cell cluster, and genes $\{\beta_i\}$ that determine the sequence of transitions between clusters. We maximized the joint probability distribution of these variables given the gene expression data (Figure 2A), $p(T, \{c_A, c_B, \dots\}, \{\alpha_i\}, \{\beta_i\} | \{g_i\})$ to determine the maximum likelihood estimates of these parameters.

We first consider how to solve this problem in the case in which there are three cell clusters, and we will later build a tree using all possible combinations of three cell clusters. Let the set of three cell clusters be c_A, c_B and c_C with gene expression data $\{g_i^{A,B,C}\}$ for all genes ($i = 1\ to\ N$) and all cells. The term $g_i^{A,B,C}$ denotes the expression data for just gene i in cells in clusters c_A, c_B, and c_C. The topology T of the relationships between cell clusters c_A, c_B and c_C can take on four possible values: $T = \mathcal{A}$: cell cluster c_A is in the middle (either c_A is the progenitor of c_B and c_C, or c_A is an intermediate cell type between c_B and c_C); $T = \mathcal{B}$: cell cluster c_B is in the middle; $T = \mathcal{C}$: cell cluster c_C is in the middle; or $T = \emptyset$: we cannot determine the topology. Complementarily, for each gene i we define variables α_i

and β_i, where $\alpha_i = 1$ and $\beta_i = 0$ if the gene is a marker gene, $\alpha_i = 0$ and $\beta_i = 1$ if the gene is a transition gene, and $\alpha_i = \beta_i = 0$ otherwise. Our task is to determine the probability $p(T, \{c_A, c_B, \dots\}, \{\alpha_i\}, \{\beta_i\} | \{g_i^{A,B,C}\})$ given gene expression data for all genes $\{g_i^{A,B,C}\}$.

According to Bayes' rule, $p(T, \{c_A, c_B, \dots\}, \{\alpha_i\}, \{\beta_i\} | \{g_i^{A,B,C}\})$ is proportional to the probability of the gene expression data given T, $\{C\}, \{\alpha_i\}$ and $\{\beta_i\}$:

$$p(T, \{c_A, c_B, \dots\}, \{\alpha_i\}, \{\beta_i\} | \{g_i^{A,B,C}\})$$
$$= \frac{p(\{g_i^{A,B,C}\} | T, \{C\}, \{\alpha_i\}, \{\beta_i\})\, p(\{\alpha_i\}, \{\beta_i\} | T, \{C\})\, p(T | \{C\})\, p(\{C\})}{p(\{g_i^{A,B,C}\})} \quad (4)$$

The denominator of the right hand side of Equation (4) is a normalization constant. Expressions $p(\{\alpha_i\}, \{\beta_i\} | T, \{C\})$, $p(T | \{C\})$, and $p(\{C\})$ are respectively the prior probabilities of $\{\alpha_i\}$ and $\{\beta_i\}$ given T and $\{C\}$, the prior probability of T given $\{C\}$, and the prior probability of $\{C\}$. We assume that in the absence of any expression data, the probability that a gene is a transition or marker gene is independent of that for any other gene and of the topology and clustering configuration: $p(\{\alpha_i\}, \{\beta_i\} | T, \{C\})\, p(T | \{C\})\, p(\{C\}) = p(\{C\}) p(T) \prod_i p(\alpha_i, \beta_i)$, and $p(T) = 1/4$.

5.2. Conditional independence

In our model, we assume that knowing the clustering configuration $\{C\}$, the topology T and whether or not a gene is a marker or transition gene is sufficient to determine the probability distribution for its expression levels in each of the cell clusters. Therefore, the gene expression patterns of different genes are conditionally independent given the topology, clustering and gene type:

$$p\left(\{g_i^{A,B,C}\}\middle|T,\{C\},\{\alpha_i\},\{\beta_i\}\right)=\prod_i p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i,\beta_i\right) \qquad (5)$$

Thus, Equation (4) becomes

$$p(T,\{c_A,c_B,\dots\},\{\alpha_i\},\{\beta_i\}|\{g_i^{A,B,C}\}) = \frac{p(\{C\})p(T)\prod_i p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i,\beta_i\right)p(\alpha_i,\beta_i)}{p\left(\{g_i^{A,B,C}\}\right)} \qquad (6)$$

We maximize the evaluated $p(T,\{c_A,c_B,\dots\},\{\alpha_i\},\{\beta_i\}|\{g_i^{A,B,C}\})$ with respect to T, $\{C\}$ and each of the α_i and β_i to obtain the most likely relationships between cell types c_A, c_B and c_C, as well as the genes most likely to be marker and transition genes.

5.3. Expression for $p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i=0,\beta_i=1\right)$ (transition genes)

To infer $(T, \{c_A, c_B, \ldots\}, \{\alpha_i\}, \{\beta_i\} | \{g_i^{A,B,C}\})$, we need a model to compute the probability of the gene expression data for each gene, $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i, \beta_i\right)$, given T, $\{C\}$, α_i and β_i, following Equation (6). Our model for the probability distribution of the expression of a single asymmetric gene i in the three cell types $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 0, \beta_i = 1\right)$ is defined solely by the geometry of the arrangement of the cell types in gene expression space, as described in the main text. For example, for $T = \mathcal{A}$ and $\beta_i = 1$, our model is that the distribution of the expression levels of $g_i^{A,B,C}$ in the three cell types A, B and C has the smallest mean value in either B or C but not in A (Figure 2B). If the distribution of the expression of gene i in cell type A is $D_A\left(g_i^A \middle| \mu_A^i, \sigma_A^i, \{C\}\right)$ (we assume a log-normal distribution) with a mean μ_A^i and standard deviation σ_A^i, with analogous expressions for cell types B and C, then our model defining $p\left(g_i^{A,B,C} \middle| T = \mathcal{A}, \beta_i = 1, \{C\}\right)$ is that either $\mu_B^i < \mu_C^i$ and $\mu_B^i < \mu_A^i$ or $\mu_C^i < \mu_B^i$ and $\mu_C^i < \mu_A^i$, where μ_A^i, μ_B^i and μ_C^i are the mean values of the expression levels of g_i in cell types A, B and C. Thus,

$$p\left(g_i^{A,B,C} \middle| T = \mathcal{A}, \beta_i = 1, \{C\}\right) = \frac{1}{2}\begin{Bmatrix} p\left(g_i^{A,B,C} \middle| \mu_B^i < \mu_A^i, \mu_B^i < \mu_C^i, \{C\}\right) \\ +p\left(g_i^{A,B,C} \middle| \mu_C^i < \mu_A^i, \mu_C^i < \mu_B^i, \{C\}\right) \end{Bmatrix} \qquad (7)$$

The terms in Equation (7) can be calculated by integrating over the prior probability distribution of the means μ_A^i, μ_B^i and μ_C^i and standard deviations σ_A^i, σ_B^i and σ_C^i, with the conditions on the means constraining the domains of integration:

$$p\left(g_i^{A,B,C}\middle|T=\mathcal{A},\beta_i=1,\{C\}\right)$$
$$=\frac{1}{2}\iiint_{\mu_B^i<\mu_A^i,\mu_B^i<\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} D_A\left(g_i^A\middle|\mu_A^i,\sigma_A^i\right)D_B\left(g_i^B\middle|\mu_B^i,\sigma_B^i\right)D_C\left(g_i^C\middle|\mu_C^i,\sigma_C^i\right)p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma\right.$$
$$+\frac{1}{2}\iiint_{\mu_C^i<\mu_A^i,\mu_C^i<\mu_B^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} D_A\left(g_i^A\middle|\mu_A^i,\sigma_A^i\right)D_B\left(g_i^B\middle|\mu_B^i,\sigma_B^i\right)D_C\left(g_i^C\middle|\mu_C^i,\sigma_C^i\right)p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma\right. \quad (8)$$

Probabilities $p\left(g_i^{A,B,C}\middle|T=\mathcal{B},\beta_i=1,\{C\}\right)$ and $p\left(g_i^{A,B,C}\middle|T=\mathcal{C},\beta_i=1,\{C\}\right)$ are defined similarly.

In addition to topologies $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, we consider a null hypothesis $\emptyset$ in which asymmetric genes have differential expression levels between states, but these levels are not correlated with any particular topology of states. This corresponds to having gene expression levels from cell-types A, B and C coming from three distributions with no restrictions on the relative order of the three means:

$$p\left(g_i^{A,B,C}\middle|T=\emptyset,\beta_i=1,\{C\}\right)$$
$$=\iiint_{\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} D_A\left(g_i^A\middle|\mu_A^i,\sigma_A^i\right)D_B\left(g_i^B\middle|\mu_B^i,\sigma_B^i\right)D_C\left(g_i^C\middle|\mu_C^i,\sigma_C^i\right)p\left(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\right. \quad (9)$$

Note that the probability of the data given the null hypothesis is the average of the probabilities of the data given the non-null hypotheses:

$$p\left(g_i^{A,B,C}\middle|T=\emptyset,\beta_i=1,\{C\}\right)=\frac{1}{3}\sum_{T=A,B,C}p\left(g_i^{A,B,C}\middle|T,\beta_i=1,\{C\}\right) \qquad (10)$$

Note that $p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i=0,\beta_i=1\right)$ depends on both T and $\{C\}$.

5.4. Expression for $p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i=1,\beta_i=0\right)$ (marker genes)

Our model for marker genes assumes that the probability distribution for the expression level of such genes, $p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i=1,\beta_i=0\right)$ to be independent of T and to be generated from distributions with two cell-types having a low value and the third a high value (for example, $D_{AB}\left(g_i^{AB}\middle|\mu_{AB}^i,\sigma_{AB}^i\right)$ for cell-types A and B and $D_C\left(g_i^C\middle|\mu_C^i,\sigma_C^i\right)$ for cell-type C, with the constraint $\mu_{AB}^i<\mu_C^i$):

$$
\begin{aligned}
& p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 1, \beta_i = 0, \right) \\
& = \frac{1}{3} \iiint\limits_{\mu_{AB}^i < \mu_C^i, \sigma_{AB}^i, \sigma_C^i} D_{AB}\left(g_i^{AB} \middle| \mu_{AB}^i, \sigma_{AB}^i\right) D_C\left(g_i^C \middle| \mu_C^i, \sigma_C^i\right) p\left(\mu_{AB}^i, \mu_C^i, \sigma_{AB}^i, \sigma_C^i\right) \\
& + \frac{1}{3} \iiint\limits_{\mu_{AC}^i < \mu_B^i, \sigma_{AC}^i, \sigma_B^i} D_{AC}\left(g_i^{AC} \middle| \mu_{AC}^i, \sigma_{AC}^i\right) D_B\left(g_i^B \middle| \mu_B^i, \sigma_B^i\right) p\left(\mu_{AC}^i, \mu_B^i, \sigma_{AC}^i, \sigma_B^i\right) \qquad (11) \\
& + \frac{1}{3} \iiint\limits_{\mu_{BC}^i < \mu_A^i, \sigma_{BC}^i, \sigma_A^i} D_{BC}\left(g_i^{BC} \middle| \mu_{BC}^i, \sigma_{BC}^i\right) D_A\left(g_i^A \middle| \mu_A^i, \sigma_A^i\right) p\left(\mu_{BC}^i, \mu_A^i, \sigma_{BC}^i, \sigma_A^i\right)
\end{aligned}
$$

Note that $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 1, \beta_i = 0\right) = p\left(g_i^{A,B,C} \middle| \{C\}, \alpha_i = 1, \beta_i = 0\right)$ does not depend on T but does depend on $\{C\}$.

5.5. Expression for $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 0, \beta_i = 0\right)$ (irrelevant genes)

Our model for genes that are neither marker nor transition genes is that the expression levels of such genes, $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 0, \beta_i = 0\right)$, is generated from one single distribution $D_{ABC}\left(g_i^{A,B,C} \middle| \mu^i, \sigma^i\right)$:

$$
p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 0, \beta_i = 0\right) = \iint\limits_{\mu^i, \sigma^i} D_{ABC}\left(g_i^{A,B,C} \middle| \mu^i, \sigma^i\right) p\left(\mu^i, \sigma^i\right) \qquad (12)
$$

Note that $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 0, \beta_i = 0\right) = p\left(g_i^{A,B,C} \middle| \alpha_i = 0, \beta_i = 0\right)$ does not depend on T or $\{C\}$.

5.6. Numerical Integration

Each of the probabilities on the right hand side of Equation (6) is evaluated numerically as above. We assume the distribution of the expression of gene i in cluster c_A $D_A(g_i^A|\mu_A^i, \sigma_A^i)$ to be log-normal. Given m log2-transformed replicate measurements g_i^A of gene expression of gene i in cells belonging to cluster c_A, the probability of the data assuming mean μ_A^i and standard deviation σ_A^i is:

$$D_A(g_i^A|\mu_A^i, \sigma_A^i) = \left(\frac{1}{\sqrt{2\pi\sigma_A^{i\,2}}}\right)^m \prod_{g_i^A}^{m} e^{-\frac{\left(g_i^A-\mu_A^i\right)^2}{2\sigma_A^{i\,2}}}. \quad (13)$$

Distributions $D_B, D_C, D_{AB}, D_{AC}, D_{BC}$ and D_{ABC} are defined analogously.

We take the *a priori* probability distribution of μ^i and σ^i, $p(\mu^i, \sigma^i)$ as uniform over a certain range of means and standard deviations. For the log2-transformed gene expression data, we take $0 < \mu^i < 6$ and $0 < \sigma^i < 1$.

We take the prior probabilities for the distributions in different cell types to be independent: $p(\mu_A^i, \mu_B^i, \mu_C^i, \sigma_A^i, \sigma_B^i, \sigma_C^i) = p(\mu_A^i, \sigma_A^i)\, p(\mu_B^i, \sigma_B^i)\, p(\mu_C^i, \sigma_C^i)$. The constraints on the order of the means are enforced by the domain of integration, and the prior must be properly normalized over this domain. For example, in Equation (8),

$$\frac{1}{2}\iiint_{\mu_B^i<\mu_A^i,\mu_B^i<\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} p(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i)$$

$$+\frac{1}{2}\iiint_{\mu_C^i<\mu_A^i,\mu_C^i<\mu_B^i,\sigma_A^i,\sigma_B^i,\sigma_C^i} p(\mu_A^i,\mu_B^i,\mu_C^i,\sigma_A^i,\sigma_B^i,\sigma_C^i)=1. \quad (14)$$

Integrals are evaluated numerically in MATLAB using trapezoidal integration with step-sizes $\delta\mu = 0.05$ and $\delta\sigma = 0.01$.

5.7. Probability of topology given gene expression and cluster identities $p(T|\{g_i^{A,B,C}\},\{C\})$

We can derive the probability of the topology given the gene expression data and cluster identities $p(T|\{g_i^{A,B,C}\},\{C\})$ by summing over all the $\{\alpha_i\}$ and $\{\beta_i\}$ to find the probability of the data given topology $p(\{g_i^{A,B,C}\}\,|\,T,\{C\})$:

$$p\left(\left\{g_i^{A,B,C}\right\} \middle| T, \{C\}\right)$$

$$= \sum_{\{\alpha_i,\beta_i\}} p\left(\left\{g_i^{A,B,C}\right\} \middle| T, \{\alpha_i\}, \{\beta_i\}, \{C\}\right) p(\{\alpha_i\}, \{\beta_i\} \mid T, \{C\})$$

$$= \sum_{\beta_1} \sum_{\beta_2} \dots \sum_{\beta_N} \prod_i p\left(g_i^{A,B,C} \middle| T, \alpha_i, \beta_i, \{C\}\right) p(\alpha_i, \beta_i) \quad (15)$$

$$= \prod_i \left(\sum_{\beta_i} p\left(g_i^{A,B,C} \middle| T, \alpha_i, \beta_i, \{C\}\right) p(\alpha_i, \beta_i) \right)$$

$$p\left(\left\{g_i^{A,B,C}\right\} \middle| T, \{C\}\right) = \prod_i p\left(g_i^{A,B,C} \middle| T, \{C\}\right), \quad (16)$$

where the probability of gene expression data for gene i given topology $p\left(g_i^{A,B,C} \middle| T, \{C\}\right)$ is obtained by summing $p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i, \beta_i\right)$ over α_i and β_i:

$$\begin{aligned} & p\left(g_i^{A,B,C} \middle| T, \{C\}\right) \\ & \quad = p\left(g_i^{A,B,C} \middle| \{C\}, \alpha_i = 1, \beta_i = 0\right) p(\alpha_i = 1, \beta_i = 0) \\ & \quad + \; p\left(g_i^{A,B,C} \middle| T, \{C\}, \alpha_i = 0, \beta_i = 1\right) p(\alpha_i = 0, \beta_i = 1) \\ & \quad + \; p\left(g_i^{A,B,C} \middle| \alpha_i = 0, \beta_i = 0\right) p(\alpha_i = 0, \beta_i = 0). \end{aligned} \quad (17)$$

The probability of topology T given data is proportional to the probability of the data given topology T (using Bayes' rule):

$$p(T \mid \{g_i^{A,B,C}\}, \{C\}) = \frac{p\left(\left\{g_i^{A,B,C}\right\} \middle| T, \{C\}\right) p(T)}{p\left(\left\{g_i^{A,B,C}\right\} \middle| \{C\}\right)}, \quad (18)$$

where

$$p\left(\left\{g_i^{A,B,C}\right\}|\,\{C\}\right) = \sum_T p(T)\, p\left(\left\{g_i^{A,B,C}\right\}\middle|\, T,\{C\}\right). \qquad (19)$$

Therefore, using equation (16), we obtain the following expression for $p(T \mid \{g_i^{A,B,C}\},\{C\})$:

$$p\left(T\middle|\left\{g_i^{A,B,C}\right\},\{C\}\right) = \frac{p(T)\prod_i p\left(g_i^{A,B,C}\middle|T,\{C\}\right)}{\sum_T p(T)\prod_i p\left(g_i^{A,B,C}\middle|T,\{C\}\right)}. \qquad (20)$$

Equation (20) can be written more explicitly by rewriting equation (17) as follows:

$$\begin{aligned} p\left(g_i^{A,B,C}\middle|T,\{C\}\right) &= p\left(g_i^{A,B,C}\middle|T,\beta_i=0,\{C\}\right)p(\beta_i \\ &= 0) + \; p\left(g_i^{A,B,C}\middle|T,\beta_i=1,\{C\}\right)p(\beta_i=1) \\ &= p\left(g_i^{A,B,C}\middle|\beta_i=0,\{C\}\right)p(\beta_i \\ &= 0)\left(1+\frac{p\left(g_i^{A,B,C}\middle|T,\beta_i=1,\{C\}\right)p(\beta_i=1)}{p\left(g_i^{A,B,C}\middle|\beta_i=0,\{C\}\right)p(\beta_i=0)}\right). \end{aligned} \qquad (21)$$

Here we have used the fact noted earlier that in our generating model, $p\left(g_i^{A,B,C}\middle|T,\alpha_i=1,\,\beta_i=0,\{C\}\right) = p\left(g_i^{A,B,C}\middle|\,\alpha_i=1,\,\beta_i=0,\{C\}\right)$ and $p\left(g_i^{A,B,C}\middle|T,\alpha_i=0,\,\beta_i=0,\{C\}\right) = p\left(g_i^{A,B,C}\middle|\alpha_i=0,\,\beta_i=0\right)$ do not depend on T (Equation (12)). The terms $\prod_i p\left(g_i^{A,B,C}\middle|\beta_i=0,\{C\}\right)p(\beta_i=0)$ cancel out in the numerator and denominator of equation (20), and we can write equation (20) in terms of ratios of the probabilities of the data given transition-gene and non-transition-gene status:

$$p\left(T\middle|\left\{g_i^{A,B,C}\right\},\{C\}\right)$$

$$= \frac{p(T)\prod_i\left(1+\frac{p\left(g_i^{A,B,C}\middle|T,\beta_i=1,\{C\}\right)p(\beta_i=1)}{p\left(g_i^{A,B,C}\middle|\beta_i=0,\{C\}\right)p(\beta_i=0)}\right)}{\sum_T p(T)\prod_i\left(1+\frac{p\left(g_i^{A,B,C}\middle|T,\beta_i=1,\{C\}\right)p(\beta_i=1)}{p\left(g_i^{A,B,C}\middle|\beta_i=0,\{C\}\right)p(\beta_i=0)}\right)}. \quad (22)$$

We can rewrite equation (22) as:

$$p\left(T\middle|\left\{g_i^{A,B,C}\right\},\{C\}\right)$$

$$= \frac{p(T)\prod_i\left(1+\frac{1}{p(T)}\,\mathcal{O}_{\beta|\{C\}}(i)\,p\left(T\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)\right)}{\sum_T p(T)\prod_i\left(1+\frac{1}{p(T)}\,\mathcal{O}_{\beta|\{C\}}(i)\,p\left(T\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)\right)}, \quad (23)$$

where $\mathcal{O}_{\beta|\{C\}}(i)$ is the odds that gene i is a transition gene, given clustering:

$$\mathcal{O}_{\beta|\{C\}}(i) = \frac{p\left(\beta_i=1\middle|g_i^{A,B,C},\{C\}\right)}{p\left(\beta_i=0\middle|g_i^{A,B,C},\{C\}\right)} = \frac{p\left(g_i^{A,B,C}\middle|\beta_i=1,\{C\}\right)}{p\left(g_i^{A,B,C}\middle|\beta_i=0,\{C\}\right)}\frac{p(\beta_i=1)}{p(\beta_i=0)} \quad (24)$$

and $p\left(T\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)$ is the probability of T given only gene expression data for gene i, clustering and that gene i is a transition gene:

$$p\left(T\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)=\frac{p\left(g_i^{A,B,C}\middle|\beta_i=1,\ T,\{C\}\right)p(T)}{p\left(g_i^{A,B,C}\middle|\beta_i=1,\{C\}\right)}. \qquad (25)$$

Thus each gene's contribution $p\left(T\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)$ to the probability of the topology given total gene expression $p\left(T\middle|\{g_i^{A,B,C}\},\{C\}\right)$ is weighted by the odds $\mathcal{O}_{\beta|\{C\}}$ that it is transition gene.

5.8. Rewriting Equation (23) in terms of negative votes

Let us denote the probability of gene expression data for gene i given that cell cluster ξ has the distribution with minimum mean expression as $p\left(g_i^{A,B,C}\middle|\mu_\xi^i \text{ is min, } \{C\}\right)$. For example, $p\left(g_i^{A,B,C}\middle|\mu_B^i \text{ is min , } \{C\}\right)=\ p\left(g_i^{A,B,C}\ \middle|\ \mu_B^i<\ \mu_A^i,\mu_B^i<\ \mu_C^i,\{C\}\right)$. Then, using $p(T)=1/4$ and Equations (7) and (10), we can write:

$$p\left(g_i^{A,B,C}\middle|\beta_i = 1,\{C\}\right) = \frac{1}{4}\left[\sum_{T=\mathcal{A},\mathcal{B},\mathcal{C},\emptyset} p\left(g_i^{A,B,C}\middle|T,\beta_i = 1,\{C\}\right)\right]$$

$$= \frac{1}{4}\left[\underbrace{\sum_{\xi=A,B,C} p\left(g_i^{A,B,C}\middle|\mu_\xi^i \text{ is min, } \{C\}\right)}_{T=\mathcal{A},\mathcal{B},\mathcal{C}}\right.$$

$$\left.+\underbrace{\frac{1}{3}\sum_{\xi=A,B,C} p\left(g_i^{A,B,C}\middle|\mu_\xi^i \text{ is min, } \{C\}\right)}_{T=\emptyset}\right] \quad (26)$$

$$= \frac{1}{3}\left[\sum_{\xi=A,B,C} p\left(g_i^{A,B,C}\middle|\mu_\xi^i \text{ is min, } \{C\}\right)\right]$$

Therefore, for $T = \mathcal{A},\mathcal{B},\mathcal{C}$, we can rewrite Equation (7) as:

$$p\left(g_i^{A,B,C}\middle|T,\beta_i = 1,\{C\}\right)$$

$$= \frac{1}{2}\left[3\,p\left(g_i^{A,B,C}\middle|\beta_i = 1,\{C\}\right) - \left(g_i^{A,B,C}\middle|\mu_T^i \text{ is min, } \{C\}\right)\right]. \quad (27)$$

Combining Equations (23) and (27), we derive, for $T \neq \emptyset$:

$$p\left(T\middle|\left\{g_i^{A,B,C}\right\},\{C\}\right) \propto p(T)\prod_i\left(1+\frac{p\left(g_i^{A,B,C}\,\middle|\,T,\beta_i=1,\{C\}\right)}{p\left(g_i^{A,B,C}\,\middle|\,\beta_i=1,\{C\}\right)}\,\mathcal{O}_{\beta|\{C\}}(i)\right)$$

$$\propto p(T)\prod_i\left(1+\frac{\frac{1}{2}\left[3\,p\left(g_i^{A,B,C}\middle|\beta_i=1,\{C\}\right)-p\left(g_i^{A,B,C}\middle|\mu_T^i\text{ is min},\{C\}\right)\right]}{p\left(g_i^{A,B,C}\,|\,\beta_i=1,\{C\}\right)}\,\mathcal{O}_{\beta|\{C\}}(i)\right) \qquad (28)$$

$$\propto\ p(T)\prod_i\left(1+\frac{3}{2}\mathcal{O}_{\beta|\{C\}}(i)\left[1-p\left(\mu_T^i\text{is min}\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)\right]\right),$$

where $p\left(\mu_T^i\text{is min}\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)$ is the probability that cell cluster T (the intermediate cluster in topology T) has the distribution with the minimum mean for gene i:

$$\begin{aligned} &p\left(\mu_T^i\text{is min}\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)\\ &\qquad=\frac{p\left(g_i^{A,B,C}\,\middle|\,\mu_T^i\text{is min},\{C\}\right)p\left(\mu_T^i\text{is min}\,\middle|\,\beta_i=1,\{C\}\right)}{p\left(g_i^{A,B,C}\,\middle|\,\beta_i=1\right)}\\ &\qquad=\frac{1}{3}\frac{p\left(g_i^{A,B,C}\,\middle|\,\mu_T^i\text{is min},\{C\}\right)}{p\left(g_i^{A,B,C}\,\middle|\,\beta_i=1,\{C\}\right)}. \end{aligned} \qquad (29)$$

Every gene can be thought of as casting a vote $-p\left(\mu_T^i\text{is min}\middle|g_i^{A,B,C},\beta_i=1,\{C\}\right)$ against cell type T being the intermediate, and this vote is weighted by the odds $\mathcal{O}_{\beta|\{C\}}(i)$ of the gene i being a transition gene and having a unique minimum, given the clustering.

5.9. Expression for $p\left(T,\{\alpha_i\},\{\beta_i\}\middle|\left\{g_i^{A,B,C}\right\},\{C\}\right)$

Once $p\left(T\middle|\{g_i^{A,B,C}\},\{C\}\right)$ is calculated, it is straightforward to find $p\left(T,\{\alpha_i\},\{\beta_i\}\middle|\{g_i^{A,B,C}\},\{C\}\right)$:

$$\begin{aligned} & p\left(T,\{\alpha_i\},\{\beta_i\}\middle|\{g_i^{A,B,C}\},\{C\}\right) \\ = \; & p\left(\{\alpha_i\},\{\beta_i\}\middle|\{g_i^{A,B,C}\},T,\{C\}\right) p\left(T\middle|\{g_i^{A,B,C}\},\{C\}\right), \end{aligned} \quad (30)$$

where $p\left(\{\alpha_i\},\{\beta_i\}\middle|\{g_i^{A,B,C}\},T,\{C\}\right)$ is the probability of $\{\alpha_i\}$ and $\{\beta_i\}$ given the particular topology T, clustering $\{C\}$ and gene expression. Because we have assumed that gene expression patterns $p\left(g_i^{A,B,C}\middle|T,\{C\},\alpha_i,\beta_i\right)$ are conditionally independent given T, $\{C\}$, α_i and β_i (Equation (5)), the probabilities of being marker or transition genes α_i or β_i are also conditionally independent given gene expression, clustering and the topology:

$$\begin{aligned} & p\left(\{\alpha_i\},\{\beta_i\}\middle|\{g_i^{A,B,C}\},T,\{C\}\right) \\ & = \frac{p\left(\{g_i^{A,B,C}\}\middle|T,\{\alpha_i\},\{\beta_i\},\{C\}\right) p(\{\alpha_i\},\{\beta_i\}|T,\{C\})}{p\left(\{g_i^{A,B,C}\}\middle|T,\{C\}\right)} \\ & = \prod_i \frac{p\left(g_i^{A,B,C}\middle|T,\alpha_i,\beta_i,\{C\}\right) p(\alpha_i,\beta_i\,|T,\{C\})}{p\left(g_i^{A,B,C}\middle|T,\{C\}\right)} \\ & = \prod_i p\left(\alpha_i,\beta_i\middle|T,\{C\},g_i^{A,B,C}\right), \end{aligned} \quad (31)$$

where $p\left(\alpha_i,\beta_i\middle|T,\{C\},g_i^{A,B,C}\right)$ is the probability that gene i is a marker or transition gene given its gene expression, the clustering, and that the topology is T.

5.10. Probability of clustering given gene expression and topology $p(\{C\}|\{g_i^{A,B,C}\},T)$

We can write an analogous expression to equation (20) for $p(\{C\}|\{g_i^{A,B,C}\},T)$, using Bayes' rule and equation (16):

$$p(\{C\}|\{g_i^{A,B,C}\},T) = \frac{p(\{C\})\prod_i p(g_i^{A,B,C}|T,\{C\})}{\sum_{\{C\}} p(\{C\})\prod_i p(g_i^{A,B,C}|T,\{C\})}. \quad (32)$$

Equation (32) can be written more explicitly by rewriting equation (17) as follows:

$$\begin{aligned} p(g_i^{A,B,C}|T,\{C\}) &= p(g_i^{A,B,C}|\alpha_i,\beta_i=0)\,p(\alpha_i,\beta_i \\ &\quad =0) + p(g_i^{A,B,C}|T,\alpha_i \text{ or } \beta_i=1,\{C\})\,p(\alpha_i \text{ or } \beta_i=1) \\ &= p(g_i^{A,B,C}|\alpha_i,\beta_i=0)\,p(\alpha_i,\beta_i \\ &\quad =0)\left(1+\frac{p(g_i^{A,B,C}|T,\alpha_i \text{ or } \beta_i=1,\{C\})\,p(\alpha_i \text{ or } \beta_i=1)}{p(g_i^{A,B,C}|\alpha_i,\beta_i=0)\,p(\alpha_i,\beta_i=0)}\right). \end{aligned} \quad (33)$$

We can rewrite equation (32)(22) as:

$$p\left(\{C\}\middle|\{g_i^{A,B,C}\},T\right)$$

$$\propto p(\{C\})\prod_i\left(1 + \frac{1}{p(\{C\})}\,\mathcal{O}_{\alpha\beta|T}(i)\,p\left(\{C\}\middle|g_i^{A,B,C},\alpha_i \text{ or } \beta_i = 1,T\right)\right), \quad (34)$$

where $\mathcal{O}_{\alpha\beta|T}(i)$ is the odds that gene i is a transition gene, given clustering:

$$\mathcal{O}_{\alpha\beta|T}(i) = \frac{p\left(\alpha_i \text{ or } \beta_i = 1\middle|g_i^{A,B,C},T\right)}{p\left(\alpha_i,\beta_i = 0\middle|g_i^{A,B,C},T\right)} = \frac{p\left(g_i^{A,B,C}\middle|\alpha_i \text{ or } \beta_i = 1,T\right)}{p\left(g_i^{A,B,C}\middle|\alpha_i,\beta_i = 0,T\right)}\frac{p(\alpha_i \text{ or } \beta_i = 1)}{p(\alpha_i,\beta_i = 0)} \quad (35)$$

and $p\left(\{C\}\middle|g_i^{A,B,C},\alpha_i \text{ or } \beta_i = 1,T\right)$ is the probability of $\{C\}$ given only gene expression data for gene i, topology T and that gene i is a marker or transition gene:

$$p\left(\{C\}\middle|g_i^{A,B,C},\alpha_i \text{ or } \beta_i = 1,T\right) = \frac{p\left(g_i^{A,B,C}\middle|T,\alpha_i \text{ or } \beta_i = 1,\ \{C\}\right)p(\{C\})}{p\left(g_i^{A,B,C}\middle|\alpha_i \text{ or } \beta_i = 1,T\right)}. \quad (36)$$

Thus each gene's contribution $p\left(\{C\}\middle|g_i^{A,B,C},\alpha_i \text{ or } \beta_i = 1,T\right)$ to the probability of the clustering given total gene expression and topology $p\left(\{C\}\middle|\{g_i^{A,B,C}\},T\right)$ is weighted by the odds $\mathcal{O}_{\alpha\beta|T}(i)$ that it is a marker or transition gene given topology.

In practice we do not explicitly calculate $p\left(\{C\}\middle|g_i^{A,B,C},\alpha_i \text{ or } \beta_i = 1,T\right)$ or $\mathcal{O}_{\alpha\beta|T}(i)$, but we recluster using high-probability marker and transition genes.

5.11. Determination of lineage tree from triplet topologies

5.11.1. Selection of triplets

In order to build lineage trees from the topologies we determine for each cell type, we select the triplets for which our determination of the topology is most robust. There is one free parameter in our determination of topology: the prior odds for a gene to be a transition gene in the absence of gene expression data, $p(\beta_i = 1)/p(\beta_i = 0)$. For each triplet, we vary this parameter between 10^{-10} and 10^{0} and calculate the probability of the topology given gene expression data $p\left(T\middle|\left\{g_i^{A,B,C}\right\},\{C\}\right)$ as a function of the prior odds.

We want to consider only triplets which showed a single dominant topology. We exclude triplets which show a weak probability for a particular topology or ones which depend on a particular choice of prior odds. We also do not consider triplets which show a strong probability for two different topologies, depending on the choice of prior odds.

We consider triplets for which only one non-null topology has probability $p\left(T\middle|\left\{g_i^{A,B,C}\right\}\right)$ greater than 0.6. Probabilities of the different topologies for each triplet are shown in Table S3.

5.11.2. Pruning rule

We assemble the triplets with known topology into a final undirected graph. Since we determined topologies by considering cell types three at a time, we obtain topological relationships involving both cell types that are nearest neighbors and cell types that are

more distantly related. In order to reconstruct the tree, we must determine which cell types are nearest neighbors and which ones are separated by one or more intermediate cell types.

The set of inferred topologies allows us to determine which cell types are separated by intermediates. For every pair of cell types, we ask whether any of the inferred topologies features an intermediate between the two cell types (See Figure 3.4B). If such a topology has been inferred, we consider that the two cell types are not nearest neighbors, and that at least one other cell type is an intermediate. For example, we can ignore triplet C_3 – **C_1** – C_4 because triplet C_1 – **C_2** – C_4 testifies that there exists an intermediate between C_1 and C_4. Similarly we can ignore triplet C_3 – **C_2** – C_4 because triplet C_2 – **C_1** – C_3 testifies that there exists an intermediate between C_2 and C_3.