Metastability and dynamics of stem cells: from direct observations to inference at the single cell level

Thesis by

Zakary S. Singer

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2015

(Defended May 13, 2015)

Acknowledgements

First, I would like to thank my advisor, Michael Elowitz, for his exceptional guidance, support, patience, and instruction over the years. His passion and zeal for science, unique perspectives, critical analysis, deep intuition, and high standards have been central to the development of all his students, and I am grateful to have had the opportunity to learn from his many talents.

My committee, Long Cai, Kathrin Plath, Bruce Hay, and Mitch Guttman, have consistently aided me with their expertise and advice, and I am deeply thankful to have benefited from their guidance, both personal and professional, over the years.

While heterogeneity is a central pursuit of the lab, one thing that is not variable within the group is the strong curiosity, generosity, and insight of its members. It has been an honor to work with such a deeply talented set of individuals. In particular, I would like to thank John Yong, Sahand Hormoz, and James Linton whose tireless efforts helped make the following chapters a reality. Additionally, it has been a pleasure to work with my fellow labmates and friends Yaron Antebi, Lacra Bintu, Mark Budde, Emily Capra, Grace Chow, Chiraj Dalal, Fangyuan Ding, Kirsten Frieda, Amit Lakhanpal, Lauren LeBon, Joseph Levine, Pulin Li, Yihan Lin, James Locke, Joe Markson, Sandy Nandagopal, Jin Park, Adam Rosenthal, Shaunak Sen, Fred Tan, and Jonathan Young. Keeping the lab running has also been no small feat, and a special thank you is due to the hard work of Leah Santat, Michelle Fontes-Shah, Tara Orr, and Jo Leonardo. I've also been extremely fortunate to have had remarkable collaborators outside our lab, including Jordi Garcia-Ojalvo, Daniel Kim, Hao-Yuan Kueh, Sheel Shah, and Eric Lubeck. All of these individuals have made doing science not only possible, but a truly enjoyable pursuit as well.

I'd like to thank my parents for their support and encouragement to follow my dreams, even when I couldn't explain what they were. And finally, a big thanks to all of my friends in and out of Caltech who helped me through the troughs, and celebrated with me at the peaks.

Abstract

Organismal development, homeostasis, and pathology are rooted in inherently probabilistic events. From gene expression to cellular differentiation, rates and likelihoods shape the form and function of biology. Processes ranging from growth to cancer homeostasis to reprogramming of stem cells all require transitions between distinct phenotypic states, and these occur at defined rates. Therefore, measuring the fidelity and dynamics with which such transitions occur is central to understanding natural biological phenomena and is critical for therapeutic interventions.

While these processes may produce robust population-level behaviors, decisions are made by individual cells. In certain circumstances, these minuscule computing units effectively roll dice to determine their fate. And while the 'omics' era has provided vast amounts of data on what these populations are doing en masse, the behaviors of the underlying units of these processes get washed out in averages.

Therefore, in order to understand the behavior of a sample of cells, it is critical to reveal how its underlying components, or mixture of cells in distinct states, each contribute to the overall phenotype. As such, we must first define what states exist in the population, determine what controls the stability of these states, and measure in high dimensionality the dynamics with which these cells transition between states.

To address a specific example of this general problem, we investigate the heterogeneity and dynamics of mouse embryonic stem cells (mESCs). While a number of reports have identified particular genes in ES cells that switch between 'high' and 'low' metastable expression states in culture, it remains unclear how levels of many of these regulators combine to form states in transcriptional space. Using a method called single molecule mRNA fluorescent *in situ* hybridization (smFISH), we quanti-

tatively measure and fit distributions of core pluripotency regulators in single cells, identifying a wide range of variabilities between genes, but each explained by a simple model of bursty transcription. From this data, we also observed that strongly bimodal genes appear to be co-expressed, effectively limiting the occupancy of transcriptional space to two primary states across genes studied here. However, these states also appear punctuated by the conditional expression of the most highly variable genes, potentially defining smaller substates of pluripotency.

Having defined the transcriptional states, we next asked what might control their stability or persistence. Surprisingly, we found that DNA methylation, a mark normally associated with irreversible developmental progression, was itself differentially regulated between these two primary states. Furthermore, both acute or chronic inhibition of DNA methyltransferase activity led to reduced heterogeneity among the population, suggesting that metastability can be modulated by this strong epigenetic mark.

Finally, because understanding the dynamics of state transitions is fundamental to a variety of biological problems, we sought to develop a high-throughput method for the identification of cellular trajectories without the need for cell-line engineering. We achieved this by combining cell-lineage information gathered from time-lapse microscopy with endpoint smFISH for measurements of final expression states. Applying a simple mathematical framework to these lineage-tree associated expression states enables the inference of dynamic transitions. We apply our novel approach in order to infer temporal sequences of events, quantitative switching rates, and network topology among a set of ESC states.

Taken together, we identify distinct expression states in ES cells, gain fundamental insight into how a strong epigenetic modifier enforces the stability of these states, and develop and apply a new method for the identification of cellular trajectories using scalable *in situ* readouts of cellular state.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cellular Specialization

From single-celled organisms operating in communities, to complex multi-cellular creatures, a ubiquitous process is cellular specialization. Bacterial populations develop specialized subsets that improve population fitness by aiding immune evasion, environmental robustness, motility, and metabolic labor division all through non-genetic variation [1–5]. In more complex organisms, stem cells in adult tissues are tasked with producing progenies that must decide between distinct terminally differentiated fates [6–11]. Defects in the fundamental regulatory processes underlying these decision-making processes can lead to cancer [12–19]. As a result, understanding how cells choose specific cellular states, the kinetics of these decision making processes, and how these populations interact to ultimately control basic processes is a central problem in modern biology and medicine.

In all of these contexts, cells move between phenotypic states that express unique sets of proteins. In one of the simplest known examples, following infection of *E. coli*, the bacteriophage lambda can decide to either enter a stable dormant state (lysogenic) or begin to produce and release more phage following lysis of the cell (lytic). Lyosgenic to lytic transitions are achieved through the downregulation of cI, a repressor of lytic activity [20, 21]. Transitions between these states can be triggered through environmental cues and cellular damage, but are also inherently stochastic.

In more complex systems, transitions between states is achieved by traveling along

defined transcriptional trajectories [22, 23]. For example, during developmental trajectories of higher organisms, precursor cells can act as common progenitors for multiple terminally differentiated fates by engaging distinct transcriptional programs [24, 25] (Figure 1A). Interconversion between cellular states also occur in pathology, normal tissue homeostasis, and state selection during adult stem cell differentiation. Examples include hematopoiesis [26, 27]; cancer [12–19]; tissue maintenance in the central nervous system [6, 11, 24, 28, 29], epithelial lining [8, 9], adipocytes [30], and epidermis[10, 31]; wound healing [32, 33]; and diabetes [34], among others. These programs are controlled by transcription factors in gene regulatory networks that activate or repress target genes necessary to enter into and maintain these cellular states [7, 24, 28]. Interactions within these networks determine the ultimate dynamics and robustness of such state switching behaviors. In particular, the rates with which such networks produce alternative states can directly impact tumor resistance to therapy [13, 14], biofilm resistance to antibiotics [3], bacterial population fitness under nutrient limitation [4], and tissue homeostasis [6, 9, 35]. Thus, it is a central problem to gain a better understanding of how cells adopt unique phenotypic states, identify genes or transcriptional programs that distinguish such phenotypic states, and the rates with which these transitions occur.

In one specific example of cell-state transitions, Gupta et al. [13] began with a cell line derived from primary breast cancer tissue in which the existence of three states had been identified on the basis of distinct gene expression profiles: basal, luminal, and stem-like. Furthermore, live cells in each of these states could be separated by FACS using a unique set of cell-surface proteins. By isolating and re-plating each of these subpopulations, they were able to demonstrate that each individual type was capable of returning to an equilibrium comprised of all three cell states. Critically, these experiments demonstrated that differentiated cells were capable of giving rise to stem-like cells. This data was consistent with a simple Markovian process wherein a cell's next transition rate depended only upon its current state. Significantly, these experiments show that therapeutics targeting any single cell state would not be able to eradicate the tumor.

Further demonstrating the significance of cell states in cancer, Leder et al. [14] showed how knowledge of cell states in a glioblastoma (GBM) could be leveraged to improve cancer therapy. They began by modeling metastability between two phenotypes characterized by either a stem-like quiescent radiation-resistant state, as well as a differentiated radiation-sensitive state. These mathematical models were combined with radiation treatment of an in vivo mouse GBM used to tune the model, as well as to demonstrate efficacy. Through combined theoretical and experimental work, they identified a dosing schedule that effectively limited the growth rate of the overall tumor. This was achieved by matching the treatment schedule to the dynamics of cellular interconversion, in order to maximize the number of quiescent cells. Thus, defining states using gene expression and phenotypic behavior is necessary in order to construct models of dynamic state transitions that can ultimately produce effective therapies.

1.2 Transcriptional noise and dynamics of activation complicate the identification of states

To begin, we must know what states exist within a heterogeneous population of cells. Averaging over a mixed population of cells would blur the differences between states, precluding the ability to readily identify a transcriptional program. The readout of 'states' must therefore be performed at the level of the individual cell (Figure 1B, top) [36, 37]. Additionally, even though single-cell readouts of static measurements can enable the identification of what states exist in a population, distinct regimes of dynamics could each give rise to the same equilibrium distribution, and this is a problem we will also seek to address.

The readout of cellular state is complicated by the fundamentally noisy process of transcription. mRNA is produced sporadically in short bursts whose initiation is intrinsically random. Such stochasticity can lead to large cell-to-cell variation, resulting in a distribution of transcript counts among otherwise identical cells. Under

certain assumptions, these distributions can often be modeled using a gamma or negative binomial distribution [38–47]. The effects of this bursty expression can in some cases be buffered by slower degradation rates of its protein. In other instances however, noisiness in expression can also trigger transitions between cellular states [4, 48–52]. Thus, transcriptional noise can make the identification of cellular states a challenging experimental problem.

In addition to the stochastic nature of transcription, cells can also behave probabilistically as a function of environmental cues. For example, when cells respond to TNFα through the activation of NF-κB, individual cells do not respond more strongly to increasing ligand concentrations; instead, the fraction of activated cells increases [53]. This again would confound results gathered from population level statistics such as bulk time-point assays.

1.3 Current methods for measuring dynamics

Once transcriptional states are identified within a population, there are several common methods to measure dynamics. There are a class of techniques that can obtain phenotypic state information at the single-cell level (through, for example, single-cell RNA-seq, qPCR, and flow cytometry), but are unable to follow the same cell over time. The lack of single cell histories prevent incorporation of lineage or spatial relatedness, limiting the application of these techniques to cases where a perturbation must be triggered at the beginning of the experiment. Therefore, these cannot address transition events that occur during equilibrium or homeostasis.

Another way to measure gene expression dynamics across multiple divisions in single cells is through the use of quantitative time-lapse microscopy with knock-in reporters. Time-lapse microscopy has provided a powerful means of exploring the temporal dynamics of gene expression and cellular decision-making among individual cells, and through *in vivo* lineage tracking [10, 31, 54]. In these examples, however, the target genes were identified *a priori* and required cell- or mouse-line construction. Thus, they are limited by availability of spectrally distinct fluorescent proteins, time-

consuming and potentially perturbing reporter engineering, invasive imaging, and pre-existing knowledge of genes of interest.

Alternatively, transition rates can be measured through the isolation of sub-populations via fluorescence-activated cell sorting (FACS), followed by growth where the cells are allowed to re-equilibrate [13, 52, 55]. However, cell transition events may not be purely stochastic, and may additionally rely on signaling interactions among cellular sub-populations. As a result, destroying the native context of the sample may inherently lead to transition rates that are biased. In the context of cancer therapeutics, this may be a significant limitation as the *in vivo* dynamics underlying tumor repopulation of heterogeneity following treatment may vary from what is observed from purified populations.

A recent single-cell sequencing approach takes advantage of somatic mutations to identify related cells, and maps their topology throughout the brain [56]. If this technique can be expanded to confer higher resolution of cellular lineages, and be coupled to high-throughput mRNA quantification, our new method as presented in Chapter 3 could also be performed on *in vivo* samples directly, and yield deep insight into dynamics within whole tissues.

1.4 Embryonic stem cells provide an ideal model system for studying metastability

In order to develop and demonstrate tools to better quantitatively analyze heterogeneous populations, we require a system where functional heterogeneity of single components have been demonstrated to occur naturally *in vitro* without external perturbations (Figure 1C). As such, we take advantage of mouse embryonic stem cells (mESCs, or ESCs) wherein a number genes studied individually have appeared heterogeneous under constant, standard growth conditions, and demonstrated interconversion between purified subpopulations. Furthermore, this variability has been shown to be functional, with effects on self-renewal and propensities to differenti-

ate into distinct lineages [55, 57–62]. Despite this deep literature, however, clear definitions of the underlying transcriptional states in multidimensional space, what contributes to state stability, or the trajectories along which cells move during transitions remain unresolved.

ESC are defined as the outgrowth of the inner cell mass (ICM) in culture from E3.5 embyros. When grown in serum and LIF, a number of genes including Tbx3, Rex1, Dppa3, Nanog, Prdm14, Pecam1, SSEA1, and Hex, appear highly variable, and are associated with specific, although sometimes redundant, differentiation propensities [55, 58–65]. For example, starting with a frequently used marker of pluripotency, Rex1, Toyooka et al [60] sorted high- and low-Rex1 cells using a knockin reporter and performed subpopulation qPCR. These experiments revealed that the 'low' cells expressed increased levels of Fgf5 and Sox17, markers of epiblast-like cells, corresponding to a slightly later stage of development than the ICM. Conversely, the Rex1-high cells appeared to express comparatively higher levels of other genes associated with pluripotency such as Nanog, Tcl1, Tbx3, Klf4, and Dppa3. Functionally, using a non-FACS-based system to select highly purified positive and negative states of Rex1, they demonstrated that, under differentiation conditions, Rex1-lows were much more efficient at differentiation into both mesendoderm and neuroectoderm than Rex1-highs. Significantly, both of these subpopulations showed the ability to interconvert between one another, re-equilibrating over the timescale of days.

At the same time as this study, Hayashi et al [58] also reported similar findings for a different gene called Dppa3 (or Stella). Intriguingly, they also concluded that Dppa3-low cells appeared closer to epiblast than ICM, were more likely to express Fgf5 and were lower in Rex1, Nanog, and other pluripotency associated genes than Dppa3-high sorted cells. They also demonstrated that Dppa3-low cells were more readily differentiated into neural cells than Dppa3-high cells. To explore the dynamics of these states, they also purified subpopulations and observed re-equilibration over the timescale of days, but also noticed a difference in plating efficiency between subpopulations after sorting; this is significant, as it may suggest that rates of interconversion and viability are dependent on context, and that studying dynamics

under perturbed conditions may be different than those under equilibrium, especially if subpopulations communicate with one another [66].

Another example of phenotypic stem cell heterogeneity is the surprising spontaneous entrance of ESCs into the so-called 2-cell (2C)-embryonic state. Developmentally, the 2C state is characterized by a major shift towards Zygotic Genome Activation (ZGA), where the embryo begins to produce transcripts from its own genome, as opposed to using those inherited maternally. At this stage of development, cells are totipotent, capable of giving rise to extraembyronic lineages in addition to all three germ layers. One of the most specific transcriptional markers of this state is known as Zscan4, which is only expressed in 2C cells [67] and sporadically in 5% of ESCs at any given time [57, 68].

Zscan4 is a set of nine paralogous genes in a subtelomeric cluster on Chromosome 7. It contains a SCAN domain that mediates self-association and heterodimerization with other SCAN-domain-containing proteins [69]. Zscan4 also has a series of four zinc-fingers, likely enabling specific DNA binding [67]. Perturbations to Zscan4 at this stage of development lead to delayed division and failure to implant [67]. Mechanistic studies of Zscan4 have demonstrated a connection to telomere extension by rapid induction of telomeric sister chromatid exchange; maintenance of genome integrity and euploidy; and prevention of culture crisis [68]. This state, as marked by Zscan4 expression, is also associated with the strong upregulation of endogenous retrovirus activity, specifically murine endogenous retrovirus with leucine tRNA primer (MuERV-L) [57].

As these MuERV-L elements were strongly correlated with Zscan4, Macfarlan et al [57] developed a live-cell reporter using MuERV-L regulatory elements in order to separate populations demonstrating this retrotransposon activity. Doing so, they found that 50% of cells displaying high levels of activity quickly shut down MuERV-L activity with 24 hours, indicating a relatively short state half-life [57]. Additionally, Zalzman et al [68] learned using a Zscan4 reporter that almost every cell in culture can pass through this state at least once in nine days. Furthermore, they demonstrated the unique ability for these 2C-like cells isolated in culture to contribute to extraembryonic lineages [57], whereas negative cells contributed only to ICM. Thus, much like was

demonstrated for Rex1 and Dppa3, Zscan4 displays phenotypic differences between metastable states.

Intriguingly, because Zscan4 is located in a subtelomeric region compacted by high levels of DNA and histone methylation, it may exhibit 'telomere position effect'. It was postulated that as a result, these loci may be tightly controlled epigenetically, and be susceptible to demethylating perturbations [70]. Indeed, treatment with histone or DNA methylation inhibitors increased the fraction of cells found to be expressing Zscan4 [70]. Furthermore, evidence of a positive feedback loop for Zscan4 has also been identified [71]. Therefore, regulators of chromatin modifications may influence the entry into, or stability of, these metastable states.

This was made all the more plausible by a recent connection to another heterogeneously expressed gene, Tbx3, which acts as an activator of Zscan4 [71]. Tbx3 is a transcription factor shown to be heterogeneously expressed [63, 72], associated with maintenance of pluripotency [63, 73] and represses ectoderm differentiation [66, 74]. One way in which Tbx3 is thought to assist in maintenance of pluripotency is through the upregulation of Nanog via cross-activation of PI(3)K from LIF signaling [63]. Further evidence of the strong interconnectivity of the network underlying pluripotency is Tbx3's direct repression of *de novo* DNA methyltransferase Dnmt3b, which can lead to a slight increase in the fraction of cells expressing 2C-like genes such as Zscan4, as well as elongated telomeres [71]. Thus, as the known network of heterogeneous regulators grows increasingly complex, so do the apparent interactions.

There are paradoxical reports that overexpression of pluripotency regulators can also lead to differentiation. For example, it was reported that while loss of Tbx3 can lead to differentiation towards ectoderm and trophectoderm differentiation [74], overexpression of Tbx3 can actually lead to activation of Gatat6 and differentiation into primitive endoderm [74]. As a result, it appears that Tbx3 dosage must be tightly controlled in order to maintain pluripotency; this unusual 'dual role' behavior has also been observed for several pluripotency genes in mouse ESCs including Oct4 and Sox2 [75–79].

As a result, it becomes even more clear that knowing the relative expression levels

of genes within individual cells is crucial to understanding decision making in ESCs. The literature also provides evidence that multiple genes, each studied individually, can represent highly similar phenotypes and dynamics, suggesting that a higher-dimensional analysis is first necessary to address specifically what states exist among population, whether these genes might represent overlapping substates, and how to resolve differences between these states and other known states, such as EpiSCs [80, 81].

Efforts to gain a systems-level understanding of ES cells have used a variety of measurement techniques such as chromatin immunoprecipitation sequencing (ChIP-Seq), bulk RNA-seq, gain- and loss-of function perturbation, mathematical modeling, and small-sample sized analyses of single cells by RNA-seq and qPCR [65, 73, 82–85]. These analyses have primarily led to the identification of novel stem cell regulators, highlighted the complexity and interconnectivity of the underlying machinery, and have postulated how dynamics might arise from known physical regulatory interactions.

However, several questions remain unanswered. First, how can we distinguish states of gene expression from noise in transcription? Because production of mRNA is inherently a stochastic processes, it is best characterized by its distribution [86–88]. Fitting quantitative models of transcription to distributions requires a large number of cells, but can enable the disambiguation of states from noise (Chapter 2). Second, in high-dimensional space, it is unclear what trajectories are taken through state space under natural, unperturbed conditions at equilibrium. This is critical, as the study of cell transitions frequently occur at equilibrium in order to maintain homeostasis, and as such, applying a forced external perturbation may yield systematically biased rates. Here, building on a rich literature of previously identified functional heterogeneity, interactions between pluripotency regulators, and the importance of several signaling pathways in ESCs, we aim to construct states of pluripotency that appear in culture using a unified experimental and quantitative context, as well as to reveal the dynamics of transitions between these states.

1.5 How can we improve the identification of transcriptional states and their transition rates?

While individual genes in ES cells have been show to be heterogeneous (Figure 1C), it is not known how these combine in a high-dimensional transcriptional space to define cellular states. This is critical as individual genes do not act alone, but operate within a complex regulatory network.

A truly direct, quantitative, and high-dimensional investigation of stem cell heterogeneity across many hundreds to thousands of cells could yield mRNA distributions that provide insight into transcriptional kinetics, as well as the relationships between genes. Additionally, a method that can be performed directly on unperturbed samples could provide information about spatial context, local signaling, clonal expansion, and more. Furthermore, the use of popular sequencing approaches can introduce strong technical noise such as amplification bias [89]. Thus, we propose using a technique known as 'single molecule mRNA fluorescence *in situ* hybridization', abbreviated here as smFISH [38, 87, 90], to address these difficulties. Not only does this approach enable the collection of large sample sizes directly without nucleic acid amplification (Figure 2A), it can also be performed in an intact specimen, where the spatial information of the sample can provide further insight into the tissues' behavior.

While static distributions begin to provide insight into the occupancy of transcriptional state space, it cannot reveal the dynamics of gene expression or cellular state transitions. Excitingly, we demonstrate that simply adding on top of these measurements the relatedness of cells can reveal insight into the dynamics underlying the distribution. To take advantage of this, we use a combination of time-lapse movies tracking cells across multiple generations (Figure 2B), followed by end-point single-cell measurements [91–93] in the spatially unperturbed sample (Figure 2C). Generally, these can encompass smFISH, immunostaining, morphological parameters, or any imaging-based *in situ* single cell readout. Using this method, we demonstrate the ability to infer transition rates and network topology connecting expression states across many generations [94]. Overlaying static gene-expression or phenotype data

on lineage trees *in situ* provides a powerful, scalable platform to identify novel, high-dimensional dynamics in systems ranging from cell culture to model organisms.

1.6 How do cells implement state stability?

In addition to estimating the rates of transition between states, it is important to understand the sources that drive and resist transitions. Thus, the mechanisms that control the stability of states and generate transitions among them are areas of interest for both therapeutics and basic biology. In stem cells, signaling pathways [95, 96] and epigenetic regulatory systems [97] have been shown to play key roles in these processes. A particular epigenetic modification, DNA methylation (which is found in most plant, animal and fungal systems [98]) is particularly stable and heritable over long time periods [99]. As a result, in the context of cell state transitions, this modification is frequently studied in major unidirectional developmental changes [100–103]. The recent discovery of Tet hydroxymethylases [104] opens up the possibility that methylation could be regulated in a dynamic fashion on faster timescales than are usually considered. However, it has remained unclear what role methylation plays in reversible state-switching events.

mESCs here also provide an ideal model system to address this question. They switch stochastically between metastable states in standard growth conditions. Moreover, expression of enzymes that regulate DNA methylation, as well as overall methylation levels themselves, is strongly affected by media that remove this heterogeneity, such as culture under the signaling-perturbing '2i' media [105, 106]. Conversely, reducing the expression of Tet hydroxymethylases leads to decreased expression of the pluripotency regulator Nanog [107, 108], but this point has also been debated [109]. Furthermore, while the role of methylation in ES cells has previously been explored [102, 110], the ES cell population has generally been treated as a homogenous population, potentially obscuring how methylation might be regulated or necessary for the reported heterogeneity [58, 60, 62]. Together, there exists the possibility that dynamic changes in DNA methylation might be involved with state switching in metastable

mouse ES cells. To test this possibility, it is necessary to determine the extent to which DNA methylation varies between states, the timescale over which methylation levels respond to state-switching events, and the causal role of DNA methyltransferases and Tet hydroxymethylases in controlling transitions between states. These questions are addressed in Chapter 2.

Together, using mouse ES cells, we attempt to ask how noise and distinct gene expression states together give rise to heterogeneity observed in ES cells, how cells stabilize the resulting expression states, and how we can identify the rates of transition and trajectories between these characterized states.

1.7 Figures

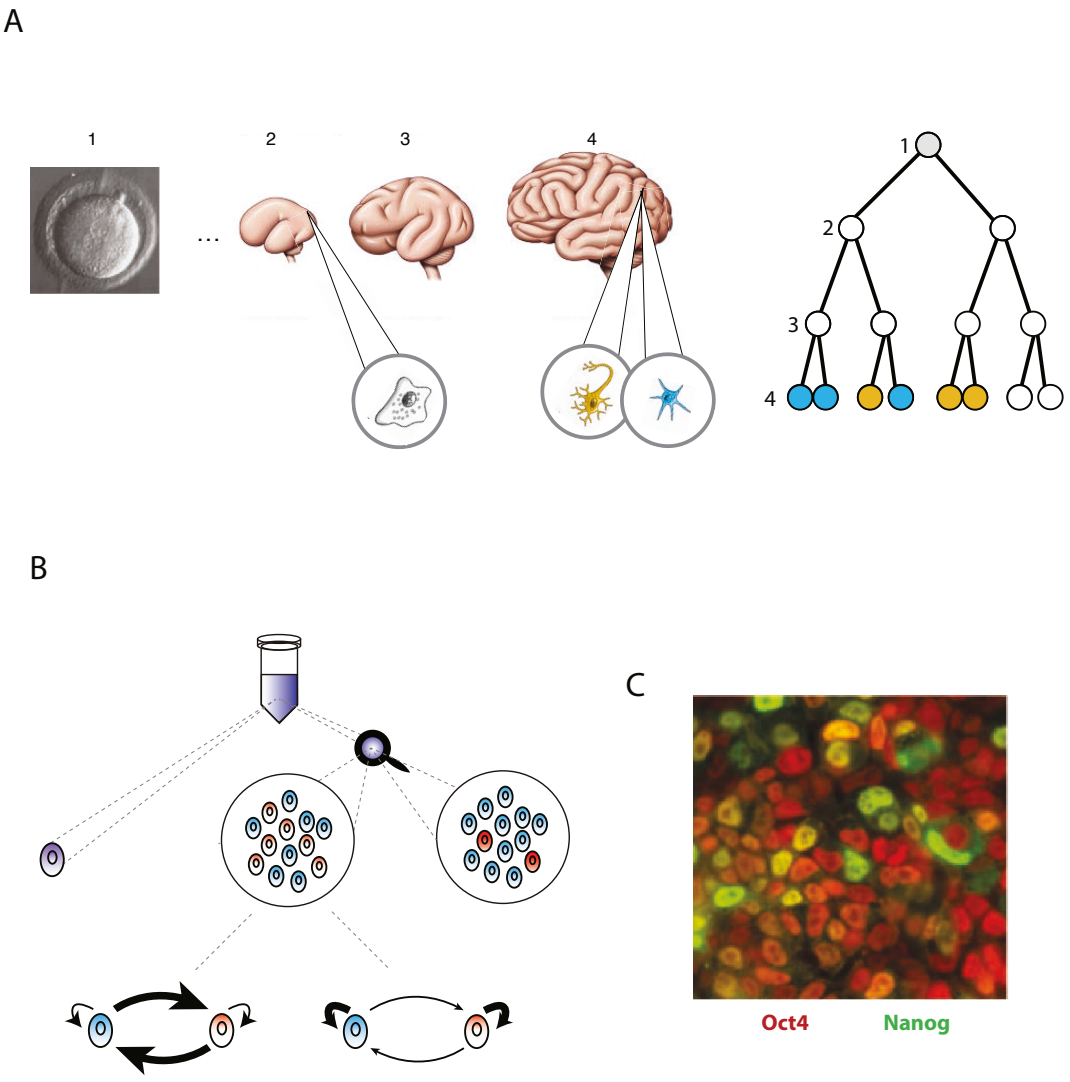

Figure 1.1: Examples of phenotypic state switching

(A) A cartoon schematic of a developing embryo from zygote to fully developed brain. The lineage tree (right) shows how an early progenitor (depicted in white) may later choose between two alternate fates (neuron or glia, in orange and blue). (B) A 'tube' of cells may, as a bulk average, appear purple leading one to assume that all cells in the tube are purple. Closer examination would reveal that the purple-appearing tube is made up of a combination of blue and red cells. The single-cell distribution of half red and half blue cells can, in turn, be generated from two different underlying dynamical models: one with fast switching between blue and red, and one with slow switching between blue and red. (C) ESCs show homogeneous antibody stains for Oct4 expression (red), but heterogeneous Nanog (green). Image by Fred Tan.

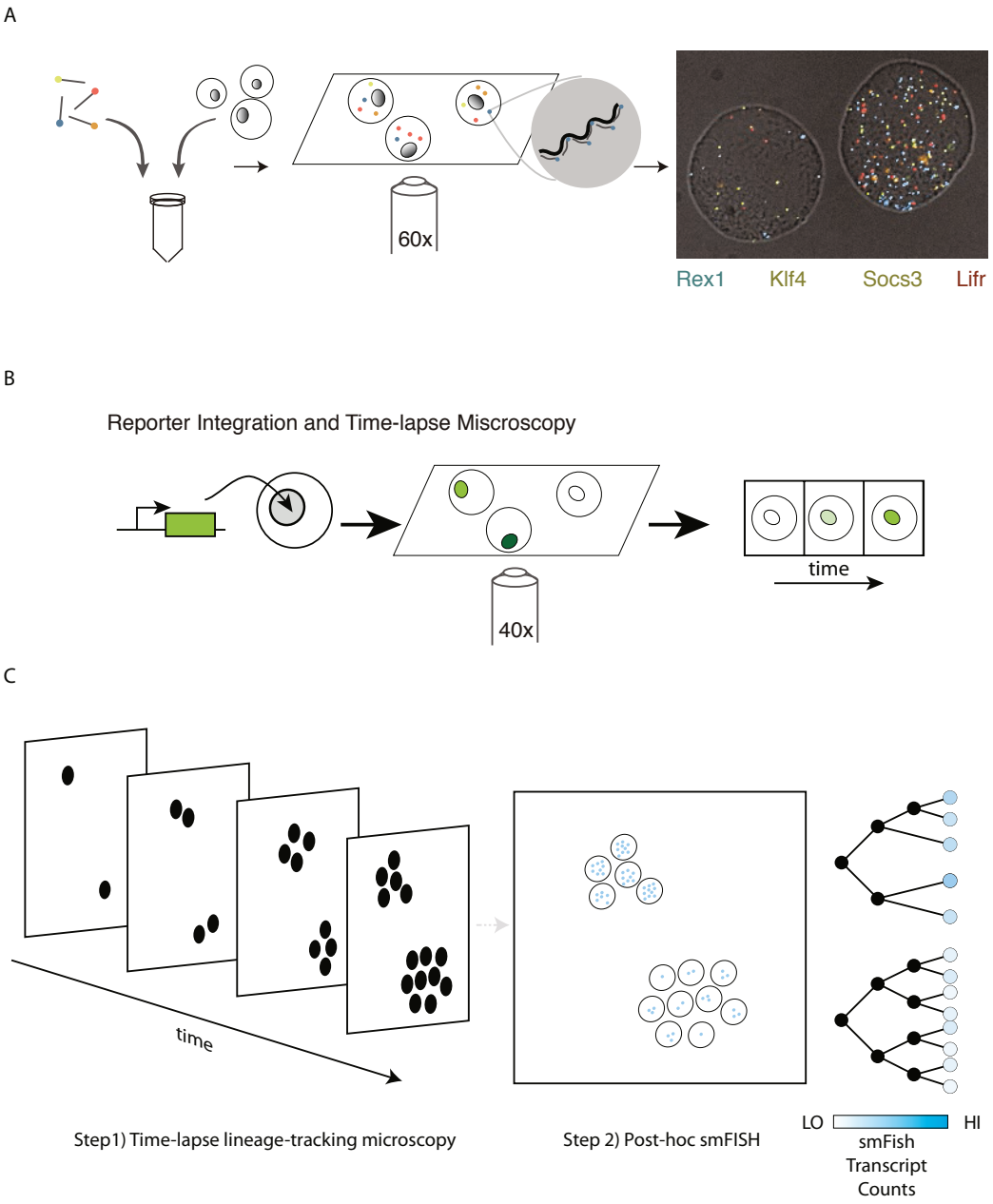

Figure 1.2: Schematic of methods to be applied

(A) Protocol for application of smFISH; probes are mixed with cells in suspension and spotted onto a slide at low density for imaging. Resulting transcripts will appear as single, diffraction limited dots. (B) Schematic for knock-in reporter creation for use in time-lapse microscopy. (C) Proposed fusion of time-lapse with endpoint smFISH to connect terminal transcriptional state with lineage information.

Primary References

1. Van Gestel, J., Vlamakis, H. & Kolter, R. From Cell Differentiation to Cell Collectives: Bacillus subtilis Uses Division of Labor to Migrate. *PLoS Biology* **13,** e1002141 (Apr. 2015) (cit. on p. 1).
2. Casadesus, J. & Low, D. A. Programmed Heterogeneity: Epigenetic Mechanisms in Bacteria. *Journal of Biological Chemistry* (2013) (cit. on p. 1).
3. Claessen, D., Rozen, D. E., Kuipers, O. P., Søgaard-Andersen, L. & van Wezel, G. P. Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies. *Nature Reviews Microbiology* (2014) (cit. on pp. 1, 2).
4. Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M. & Elowitz, M. B. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* (2006) (cit. on pp. 1, 2, 4).
5. Balázsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* (2011) (cit. on p. 1).
6. Sakamoto, M. *et al.* Continuous neurogenesis in the adult forebrain is required for innate olfactory responses. *Proc Natl Acad Sci USA* (2011) (cit. on pp. 1, 2).
7. Imayoshi, I. *et al.* Roles of continuous neurogenesis in the structural and functional integrity of the adult forebrain. *Nat Neurosci* (2008) (cit. on pp. 1, 2).
8. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nat Rev Mol Cell Biol* (2014) (cit. on pp. 1, 2).
9. Leushacke, M. & Barker, N. Lgr5 and Lgr6 as markers to study adult stem cell roles in self-renewal and cancer. *Oncogene* (2012) (cit. on pp. 1, 2).
10. Rompolas, P., Mesa, K. R. & Greco, V. Spatial organization within a niche as a determinant of stem-cell fate. *Nature* (2013) (cit. on pp. 1, 2, 4).
11. Doetsch, F., Caillé, I., Lim, D. A., García-Verdugo, J. M. & Alvarez-Buylla, A. Subventricular Zone Astrocytes Are Neural Stem Cells in the Adult Mammalian Brain. *Cell* (1999) (cit. on pp. 1, 2).
12. Wagenblast, E. *et al.* A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis. *Nature* (2015) (cit. on pp. 1, 2).
13. Gupta, P. B. *et al.* Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* (2011) (cit. on pp. 1, 2, 5).

14. Leder, K. *et al.* Mathematical modeling of PDGF-driven glioblastoma reveals optimized radiation dosing schedules. *Cell* (2014) (cit. on pp. 1–3).

15. Navin, N. E. Tumor evolution in response to chemotherapy: phenotype versus genotype. *CellReports* (2014) (cit. on pp. 1, 2).

16. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *CellReports* (2014) (cit. on pp. 1, 2).

17. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* (2013) (cit. on pp. 1, 2).

18. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* (2012) (cit. on pp. 1, 2).

19. Zuber, J. *et al.* RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* (2011) (cit. on pp. 1, 2).

20. Bednarz, M., Halliday, J. A., Herman, C. & Golding, I. Revisiting bistability in the lysis/lysogeny circuit of bacteriophage lambda. *PLoS ONE* (2014) (cit. on p. 1).

21. Golding, I. Decision Making in Living Cells: Lessons from a Simple System. *Annu. Rev. Biophys.* (2011) (cit. on p. 1).

22. Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys Rev Lett* (2005) (cit. on p. 2).

23. Chang, H. H., Oh, P. Y., Ingber, D. E. & Huang, S. Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biol* (2006) (cit. on p. 2).

24. Gokoffski, K. K. *et al.* Activin and GDF11 collaborate in feedback control of neuroepithelial stem cell proliferation and fate. *Development (Cambridge, England)* **138,** 4131–4142 (Oct. 2011) (cit. on p. 2).

25. Ihrie, R. A. & Alvarez-Buylla, A. Cells in the astroglial lineage are neural stem cells. *Cell Tissue Res* (2008) (cit. on p. 2).

26. Kueh, H. Y. *et al.* Positive feedback between PU.1 and the cell cycle controls myeloid differentiation. *Science* (2013) (cit. on p. 2).

27. Weston, W., Zayas, J., Perez, R., George, J. & Jurecic, R. Dynamic equilibrium of heterogeneous and interconvertible multipotent hematopoietic cell subsets. *Scientific reports* **4,** 5199 (2014) (cit. on p. 2).

28. Imayoshi, I. & Kageyama, R. bHLH Factors in Self-Renewal, Multipotency, and Fate Choice of Neural Progenitor Cells. *Neuron* **82,** 9–23 (Apr. 2014) (cit. on p. 2).

29. Bonaguidi, M. A. *et al.* In vivo clonal analysis reveals self-renewing and multipotent adult neural stem cell characteristics. *Cell* **145,** 1142–1155 (June 2011) (cit. on p. 2).

30. Ahrends, R. *et al.* Controlling low rates of cell differentiation through noise and ultrahigh feedback. *Science* (2014) (cit. on p. 2).

31. Clayton, E. *et al.* A single type of progenitor cell maintains normal epidermis. *Nature* (2007) (cit. on pp. 2, 4).

32. Tata, P. R. *et al.* Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature* **503,** 218–223 (Nov. 2013) (cit. on p. 2).

33. Senyo, S. E. *et al.* Mammalian heart renewal by pre-existing cardiomyocytes. *Nature* (2013) (cit. on p. 2).

34. Talchai, C., Xuan, S., Lin, H. V., Sussel, L. & Accili, D. Pancreatic β cell dedifferentiation as a mechanism of diabetic β cell failure. *Cell* **150,** 1223–1234 (Sept. 2012) (cit. on p. 2).

35. Lander, A. D., Gokoffski, K. K., Wan, F. Y. M., Nie, Q. & Calof, A. L. Cell lineages and the logic of proliferative control. *PLoS Biol* (2009) (cit. on p. 2).

36. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* (2014) (cit. on p. 3).

37. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* (2013) (cit. on p. 3).

38. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biol* (2006) (cit. on pp. 4, 10).

39. Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* (2011) (cit. on p. 4).

40. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* (2003) (cit. on p. 4).

41. Elowitz, M., Levine, A., Siggia, E. & Swain, P. Stochastic Gene Expression in a Single Cell. *Science* (2002) (cit. on p. 4).

42. Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* (2006) (cit. on p. 4).

43. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nat Genet* (2002) (cit. on p. 4).

44. Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* (2000) (cit. on p. 4).

45. Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theoretical Population Biology* (1995) (cit. on p. 4).

46. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* (2008) (cit. on p. 4).

47. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* (2011) (cit. on p. 4).

48. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* (2006) (cit. on p. 4).

49. Choi, P. J., Cai, L., Frieda, K. & Xie, X. S. A Stochastic Single-Molecule Event Triggers Phenotype Switching of a Bacterial Cell. *Science* (2008) (cit. on p. 4).

50. Maamar, H., Raj, A. & Dubnau, D. Noise in gene expression determines cell fate in Bacillus subtilis. *Science* (2007) (cit. on p. 4).

51. Zong, C., So, L.-h., Sepúlveda, L. A., Skinner, S. O. & Golding, I. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Mol Syst Biol* (2010) (cit. on p. 4).

52. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* (2008) (cit. on pp. 4, 5).

53. Tay, S. *et al.* Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* (2010) (cit. on p. 4).

54. Bothma, J. P. *et al.* Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 10598–10603 (July 2014) (cit. on p. 4).

55. Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol* (2009) (cit. on pp. 5, 6).

56. Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85,** 49–59 (Jan. 2015) (cit. on p. 5).

57. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* (2012) (cit. on pp. 6, 7).

58. Hayashi, K., Lopes, S. M. C. d. S., Tang, F. & Surani, M. A. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* (2008) (cit. on pp. 6, 11).

59. Singh, A. M., Hamazaki, T., Hankowski, K. E. & Terada, N. A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells* (2007) (cit. on p. 6).

60. Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K. & Niwa, H. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development* (2008) (cit. on pp. 6, 11).

61. Yamaji, M. *et al.* PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell Stem Cell* (2013) (cit. on p. 6).

62. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* (2007) (cit. on pp. 6, 11).

63. Niwa, H., Shimosato, D. & Adachi, K. A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature* (2009) (cit. on pp. 6, 8).

64. Canham, M. A., Sharov, A. A., Ko, M. S. H. & Brickman, J. M. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS Biol* (2010) (cit. on p. 6).

65. MacArthur, B. D. *et al.* Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. *Nat Cell Biol* (2012) (cit. on pp. 6, 9).

66. Weidgang, C. E. *et al.* TBX3 Directs Cell-Fate Decision toward Mesendoderm. *Stem Cell Reports* (2013) (cit. on pp. 7, 8).

67. Falco, G. *et al.* Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Developmental Biology* (2007) (cit. on p. 7).

68. Zalzman, M. *et al.* Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature* (2010) (cit. on p. 7).

69. Edelstein, L. C. & Collins, T. The SCAN domain family of zinc finger transcription factors. *Gene* **359,** 1–17 (Oct. 2005) (cit. on p. 7).

70. Dan, J. *et al.* Rif1 maintains telomere length homeostasis of ESCs by mediating heterochromatin silencing. *Developmental Cell* **29,** 7–19 (Apr. 2014) (cit. on p. 8).

71. Dan, J. *et al.* Roles for Tbx3 in regulation of two-cell state and telomere elongation in mouse ES cells. *Sci Rep* (2013) (cit. on p. 8).

72. Faddah, D. A. *et al.* Single-Cell Analysis Reveals that Expression of Nanog Is Biallelic and Equally Variable as that of Other Pluripotency Factors in Mouse ESCs. *Cell Stem Cell* (2013) (cit. on p. 8).

73. Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nat Cell Biol* (2006) (cit. on pp. 8, 9).

74. Lu, R., Yang, A. & Jin, Y. Dual functions of T-box 3 (Tbx3) in the control of self-renewal and extraembryonic endoderm differentiation in mouse embryonic stem cells. *Journal of Biological Chemistry* (2011) (cit. on p. 8).

75. Shimozaki, K., Nakashima, K., Niwa, H. & Taga, T. Involvement of Oct3/4 in the enhancement of neuronal differentiation of ES cells in neurogenesis-inducing cultures. *Development (Cambridge, England)* **130,** 2505–2512 (June 2003) (cit. on p. 8).

76. Zeineddine, D. *et al.* Oct-3/4 dose dependently regulates specification of embryonic stem cells toward a cardiac lineage and early heart development. *Developmental Cell* **11,** 535–546 (Oct. 2006) (cit. on p. 8).

77. Thomson, M. *et al.* Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* (2011) (cit. on p. 8).

78. Niwa, H., Miyazaki, J. & Smith, A. G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genetics* (2000) (cit. on p. 8).

79. Kopp, J. L., Ormsbee, B. D., Desler, M. & Rizzino, A. Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells. *Stem cells (Dayton, Ohio)* **26,** 903–911 (Apr. 2008) (cit. on p. 8).

80. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* (2007) (cit. on p. 9).

81. Brons, I. G. M. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* (2007) (cit. on p. 9).

82. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* (2010) (cit. on p. 9).

83. Kumar, R. M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* (2014) (cit. on p. 9).

84. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* (2006) (cit. on p. 9).

85. Chen, X. *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* (2008) (cit. on p. 9).

86. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* (2010) (cit. on p. 9).

87. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Meth* (2008) (cit. on pp. 9, 10).

88. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* (2008) (cit. on p. 9).

89. Van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research* **322,** 12–20 (Mar. 2014) (cit. on p. 10).

90. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* (1998) (cit. on p. 10).

91. Lee, R. E. C., Walker, S. R., Savery, K., Frank, D. A. & Gaudet, S. Fold Change of Nuclear NF-κB Determines TNF-Induced Transcription in Single Cells. *Mol Cell* (2014) (cit. on p. 10).

92. Kellogg, R. A. & Tay, S. Noise Facilitates Transcriptional Control under Dynamic Inputs. *Cell* (2015) (cit. on p. 10).

93. Purvis, J. E. *et al.* p53 dynamics control cell fate. *Science* (2012) (cit. on p. 10).

94. Hormoz, S., Desprat, N. & Shraiman, B. I. Inferring epigenetic dynamics from kin correlations. *Proceedings of the National Academy of Sciences of the United States of America* (Apr. 2015) (cit. on p. 10).

95. Wray, J., Kalkan, T. & Smith, A. G. The ground state of pluripotency. *Biochem Soc Trans* (2010) (cit. on p. 11).

96. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* (2012) (cit. on p. 11).

97. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* (2007) (cit. on p. 11).

98. Feng, S., Jacobsen, S. E. & Reik, W. Epigenetic reprogramming in plant and animal development. *Science* (2010) (cit. on p. 11).

99. Schübeler, D. *et al.* Genomic targeting of methylated DNA: influence of methylation on transcription, replication, chromatin structure, and histone acetylation. *Mol Cell Biol* (2000) (cit. on p. 11).

100. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* (2012) (cit. on p. 11).

101. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat Biotechnol* (2007) (cit. on p. 11).

102. Mohn, F. *et al.* Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* (2008) (cit. on p. 11).

103. Hackett, J. A. *et al.* Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* (2013) (cit. on p. 11).

104. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* (2009) (cit. on p. 11).

105. Ficz, G. *et al.* FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell Stem Cell* (2013) (cit. on p. 11).

106. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* (2013) (cit. on p. 11).

107. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* (2010) (cit. on p. 11).

108. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* (2011) (cit. on p. 11).

109. Koh, K. P. *et al.* Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* (2011) (cit. on p. 11).

110. Fouse, S. D. *et al.* Promoter CpG Methylation Contributes to ES Cell Gene Regulation in Parallel with Oct4/Nanog, PcG Complex, and Histone H3 K4/K27 Trimethylation. *Cell Stem Cell* (2008) (cit. on p. 11).

Chapter 2

Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells

2.1 Summary

Cell populations can be strikingly heterogeneous, composed of multiple cellular states, each exhibiting stochastic noise in its gene expression. A major challenge is to disentangle these two types of variability, and to understand the dynamic processes and mechanisms that control them. Embryonic stem cells (ESCs) provide an ideal model system to address this issue because they exhibit heterogeneous and dynamic expression of functionally important regulatory factors. We analyzed gene expression in individual ESCs using single-molecule RNA-FISH and quantitative time-lapse movies. These data discriminated stochastic switching between two coherent (correlated) gene expression states and burst-like transcriptional noise. We further showed that the '2i' signaling pathway inhibitors modulate both types of variation. Finally, we found that DNA methylation plays a key role in maintaining these metastable states. Together, these results show how ESC gene expression states and dynamics arise from a combination of intrinsic noise, coherent cellular states, and epigenetic regulation.

2.2 Introduction

Many cell populations appear to consist of mixtures of cells in distinct cellular states. In fact, interconversion between states has been shown to underlie processes ranging from adult stem cell niche control [1, 2] to bacterial fitness [3] and cancer development [4]. A central challenge is to identify transcriptional states, along with the mechanisms that control their stability and generate transitions among them.

Single-cell transcriptional studies have revealed substantial gene expression heterogeneity in stem cells [5–9]. Moreover, subpopulations expressing different levels of Nanog, Rex1, Dppa3, or Prdm14 show functional biases in their differentiation propensity [10–13]. This heterogeneity could in principle arise from stochastic fluctuations, or 'noise', in gene expression [14–16]. Alternatively, it could reflect the coexistence of multiple cellular states, each with a distinct gene expression pattern showing correlation between a set of genes [4, 8, 17, 18]. Disentangling these two sources of variation is important for interpreting the transcriptional states of individual cells and understanding stem cell dynamics.

A related challenge is to understand the mechanisms that stabilize cellular states despite noise. DNA methylation has been shown to be heritable over many generations, is critical for normal development [19], and may help stabilize irreversible cell fate transitions [20–23]. However, the role of DNA methylation in the reversible cell state transitions that underlie equilibrium population heterogeneity has been much less studied [24, 25]. Recently, it was reported that exposing ES cells to inhibitors of MEK and GSK3$beta (called 2i) abolishes heterogeneity and induces a 'naive' pluripotent state [26, 27] with reduced methylation [28–30]. However, a causal role linking methylation, heterogeneity, and 2i remains to be elucidated.

Together, these observations provoke several fundamental questions: First, how do noise and states together determine the distribution of expression levels of individual regulatory genes (Fig. 1A)? Second, how do gene expression levels vary dynamically in individual cells, both within a state and during transitions between states (Fig. 1B)? Finally, how do cells stabilize metastable gene expression states, and what role

does DNA methylation play in this process?

Using single-molecule RNA-FISH (smFISH), we analyzed the structure of heterogeneity in the expression of key cell fate regulators, finding that distinct cell states account for most variation in some genes, while others are dominated by stochastic bursts. Using time-lapse movies of individual cells, we observed abrupt, step-like dynamics due to cell state transitions and transcriptional bursts. Finally, using perturbations, we observed that DNA methylation modulates the population fraction of cells in the two states, consistent with reciprocal expression of the methyltransferase Dnmt3b and the hydroxymethylase Tet1. Together, these results suggest how noise, dynamics, and epigenetic regulatory mechanisms contribute together to the overall distribution of gene expression states in stem cell populations.

2.3 Results

2.3.1 Mouse ESCs show three distinct types of gene expression distributions

The process of mRNA transcription is inherently stochastic. As a result, even a clonal cell population in a single state is expected to display variability in the copy number of each mRNA [31–39], potentially leading to phenotypic differences between otherwise identical cells [3, 40–43]. In order to accurately measure mRNA copy numbers in large numbers of individual ESCs, we developed an automated platform for smFISH (Supp. Info.). This system enables rapid analysis of four genes per cell across 400 cells per sample (Figs. 6A-D). We validated the system by comparing three measures of expression of the same gene in the same cells using a Rex1-dGFP reporter line [44] (Fig. 6E).

Using this platform, we analyzed 36 pluripotency associated regulators that play critical roles in ESCs or are heterogeneously expressed, as well as several markers of early cell fates and housekeeping genes. The resulting mRNA distributions exhibited a range of distribution shapes and degrees of heterogeneity (Fig. 2A). We analyzed

these distributions within the framework of bursty transcriptional dynamics. In this model, mRNA production occurs in stochastic bursts that are brief compared to the mean inter-burst interval, and are exponentially distributed in size. Bursty dynamics produce negative binomial (NB) mRNA distributions [37, 45], whose shape is determined by the frequency and mean size of bursts.

Genes exhibited three qualitatively distinct types of mRNA distributions. First, most genes were unimodal and well-fit by a single NB distribution (Figs. 2B & 7A, maximum-likelihood estimation (MLE), χ^2 Goodness of Fit (GOF) test p¿0.05). This class included Oct4, Rest, Tcf3, Smarcc1, Sall4, and Zfp281. Coefficients of variation (CV) were typically 0.5 for the most homogeneous genes (Fig. 2A).

Second, a subset of unimodal genes exhibited "long-tailed" distributions, in which most cells had few, if any, transcripts, while a small number of cells displayed many transcripts. These distributions were also well fit by a single NB distribution, but with resulting distributions that generally decreased monotonically with increasing mRNA concentration (Figs. 2B & 7A, χ^2 G.O.F. p¿0.05). The most heterogeneous long-tailed genes had burst frequencies of less than one burst per mRNA half-life. These included Tbx3 (CV=2.13±0.23, mean±s.e.m.), Dppa3 (CV=1.76±0.31), and Prdm14 (CV=1.599±0.20). Other long-tailed genes such as Pecam1, Klf4, Blimp1, Socs3, Nr0b1, and Fgfr2 had higher burst frequencies and less skew. Long-tailed genes arising from rare bursts could provide a source of stochastic variation that could propagate to downstream genes.

Third, there were some genes whose mRNA distributions were significantly better fit by a linear combination of two NB distributions than by one (Supp. Info., Akaike's Information Criteria (AIC) and log-likelihood ratio test, p¡0.05). These genes included Rex1, Nanog, Esrrb, Tet1, Fgf4, Sox2, Tcl1, and Lifr (Figs. 2B & 7A). In some cases, the two components of these distributions were well separated from one another (e.g., Rex1 and Esrrb), while in other cases they overlapped strongly (e.g., Nanog and Lifr), such that the absolute number of transcripts for a single gene did not accurately indicate to which state the cell belonged. These bimodal distributions suggested the existence of multiple cell states (see below).

Markers of most differentiated fates including Pax6 (neuroectoderm), Fgf5 (epiblast), Sox17 and FoxA2 (definitive endoderm), and Gata6 (primitive endoderm) showed no detectable expression (data not shown). However, the mesendodermal regulator Brachyury (T) was expressed at a level of 5-20 transcripts in 6% of Rex1-low cells. Similarly, the two-cell-like state marker Zscan4c [46] showed 3-60 transcripts in 3% of cells (Fig. 7A). These genes did not fit well to NB distributions, suggesting that processes other than transcriptional bursting impact their expression in this small fraction of cells.

2.3.2 Bimodal genes vary coherently

We next used the smFISH data to determine whether the bimodal genes were correlated, which would suggest their control by a single pair of distinct cell states, or varied independently, which would suggest a multiplicity of states. The data revealed a cluster of bimodal genes that correlated with one another. Rex1, Nanog, and Esrrb displayed the strongest correlations (r≈0.7, Figs. 2C, 7B), while genes with strong overlap between modes, such as Tcl1, Lifr, Sox2, and Tet1, displayed somewhat weaker, but still significant, correlations (r≈0.5, Fig. 2C, 7B), beyond those observed between bimodal and non-bimodal genes (e.g., r=0.2 for Rex1 and Oct4). Thus, a cell in the high or low expression state of one bimodal gene is likely to be in the corresponding expression state of others. Some correlations were negative: expression of the de novo methyltransferase Dnmt3b was reduced in the Rex1-high state (r=-0.46, Fig. 2C). Note that cell cycle effects did not explain these correlations (Fig. 7C). Together, these data suggest that bimodal genes appear to be broadly co-regulated in two distinct states.

Long-tailed genes exhibited more complex relationships. Those with very large variation (CV¿1.5) were correlated with the expression state of the bimodal gene cluster, but not with one another (Fig. 2C, 2D, 7B). For example, genes like Dppa3, Tbx3, and Prdm14 burst predominantly in the Rex1-high state, but even in this state, most cells showed no transcripts of these genes (p¡0.001, see Supp. Info. for statistical

analysis). Thus, it appears that these genes are expressed in infrequent, stochastic bursts that occur mainly in one of the two cellular states.

Interestingly, expression of Socs3, a negative regulator and direct target of Stat signaling [47], appeared conditional on expression of its bimodally expressed receptor Lifr (note absence of Socs3 expression in low Lifr cells in Fig. 7B). Analysis of additional regulators not measured here could in principle reveal additional states or more complex distributions. Overall, however, the multi-dimensional mRNA distributions measured here are consistent with a simple picture based on two primary states and stochastic bursting.

2.3.3 The two primary states exhibit distinct DNA methylation profiles

These data contained an intriguing relationship between three factors involved in DNA methylation: the de novo methyltransferase Dnmt3b, the hydroxylase Tet1, which has been implicated in removing methylation [48–51], and Prdm14, which represses expression of Dnmt3b [13, 30, 52, 53]. While Rex1 was anticorrelated with Dnmt3b expression, and positively correlated with Tet1 (Fig. 3A), Prdm14 showed a long-tailed distribution conditioned on the Rex1-high state (Fig. 2D). Based on these relationships and the observation that methylation increases during early development [54], we hypothesized that the Rex1-low state might exhibit increased methylation compared to the Rex1 -high state.

To test this hypothesis, we sorted Rex1-high and -low cells using the Rex1-dGFP reporter line, and performed locus-specific bisulfite sequencing at known targets of methylation Dazl, Mael, and Sycp3 [55] (Figs. 3B, 8A,B). These promoters exhibited 2-3 times greater methylation in Rex1-low cells compared to Rex1-high cells, indicating that the two states are differentially methylated in at least some genes. In contrast, Rex1-low cells that subsequently reverted to the Rex1-high state recovered the methylation levels of Rex1-high cells, indicating that methylation is reversible. Similarly, quantitative ELISA analysis demonstrated both differential methylation

and reversibility in global methylation levels (Fig. 3C).

We next asked more generally which genes exhibited differential promoter methylation. We again sorted Rex1-high and -low cells and assayed DNA methylation by Reduced Representation Bisulfite Sequencing analysis (RRBS), analysing regions 2kb upstream to 500bp downstream of each ESC-expressed mRNA transcriptional start site [26, 54]. The distributions of methylation levels across genes were bimodal in both Rex1 states, with the more highly methylated peak shifted to even higher methylation levels in Rex1-low cells (Fig. 3D). By analyzing the shift in methylation on a gene by gene basis, we found that the increase in methylation in Rex1-low cells occurred predominantly through increased methylation of the promoters that were more highly methylated in Rex1-high cells (Figs. 3E, 8C). Thus, the change in promoter methylation occurs in a specific subset of promoters. Furthermore, the overall methylation level of a gene was related to the number of CpGs in its promoter, such that the larger the CpG content of a promoter, the lower its methylation in both states. Not all gene promoters were well covered by RRBS. However, among those that were, several key ES regulators including Esrrb, Tet1, and Tcl1 all showed increased levels of methylation in the Rex1 -low state. Figs. 3E (inset) and 8C show methylation levels of individual CpGs for 17 gene promoters. These results provide a view of the change in promoter methylation that occurs during transitions between the Rex1-high and -low states.

2.3.4 Bursty transcription generates dynamic fluctuations in individual cells

Evidently, cells populate two transcriptional states, each characterized by distinct methylation profiles. To understand the dynamic changes in gene expression that occur as individual cells switch between these states, we turned to time-lapse microscopy. We analyzed transcriptional reporter cell lines for Nanog and Oct4, each containing a histone 2B (H2B)-tagged fluorophore expressed under the control of the corresponding promoter (Figs. 9A & B; see also Supp. Info.). We imaged the reporter

cell lines for 50-hour periods with 15-minute intervals between frames, and segmented and tracked individual cells over time in the resulting image sequences. For each cell lineage, we quantified the instantaneous reporter production rate, defined as the rate of increase of total fluorescent protein in the cell, and corrected for the partitioning of fluorescent protein into daughter cells during cell division (Supp. Info.). The H2B-fluorescent protein degradation rate is negligible under these conditions (Fig. 9C), enabling us to use the reporter production rate as a measure of instantaneous mRNA level. An advantage of this approach is that it provides relatively strong fluorescence signals per cell, but still enables high time resolution analysis (Locke et al., 2009). Consistent with static smFISH distributions, the production rate distributions of the Nanog and Oct4 fluorescent reporters were bimodal and unimodal, respectively (Fig. 4A).

Dynamically, cells remained in either one of two distinct Nanog expression states for multiple cell cycles (Fig. 4B). During these periods, expression levels varied over the full range of Nanog expression levels within each state, with no evidence for persistent sub-states. However, closer examination revealed fluctuations within a single state, which typically occurred in discrete steps separated by periods of steady expression (Fig. 4C). Using a computational step detection algorithm (Fig. 9D, Supp. Info.), we found that Nanog and Oct4 reporters exhibited 0.38 and 0.29 steps per cell cycle, respectively. These steps occurred in a cell cycle phase-dependent manner (Fig. 4D), with down-steps clustered around cell division events and up-steps more broadly distributed across cell cycle phases.

Could these step-like dynamics arise simply from transcriptional bursting? To address this question, we simulated single cell mRNA and protein traces using the bursty transcription model, with parameters determined from the NB fits of the static mRNA distributions (Supp. Info.). These simulations generated dynamic traces resembling those observed experimentally (Figs. 4E, 9E). In the simulations, mRNA half-life and burst frequency determine the characteristics of detectable steps (Fig. 9F); in general, steps were most prominent at low burst frequencies and short mRNA half-lives, and became difficult to discriminate at high burst frequencies and long

mRNA half-lives.

Step-like dynamics appear to be a natural consequence of stochastic expression, with up-steps reflecting burst-like production of mRNA, and down-steps resulting from 2-fold reduction in mRNA copy number at cell division (Fig. 9G). This interpretation is consistent with the observed clustering of down-steps around cell division events, and a more uniform cell cycle distribution of up-steps (Fig. 4D). Because large bursts can effectively cancel out mRNA dilution at cell division events, they may appear under-represented near cell division events. Note that most cell cycles showed no up-steps, suggesting that they are not due to increased gene dosage after chromosome replication [**Brewster:2014ez**, 56].

2.3.5 Dynamic transitions between cellular states

We next asked how cells transition dynamically between states. Previous work has relied on cell sorting, which can distort the signaling environments. By contrast, movies enable direct observation of switching events within a mixed cell population. Since the Nanog reporter production rate fluctuates even within a single state, we used a Hidden Markov Model (HMM) to classify each cell into either Nanog-high or Nanog-low at every point in time (Supp. Info.). We trained the HMM using time-series data of Nanog reporter production rates, sampled at fixed intervals across all tracked cell cycles, and used it to identify switching events and estimate switching frequencies.

Transitions from the Nanog-low to the Nanog-high state, or vice versa, occurred at a rate of 2.3 $\pm$ 0.25, or 7.9 $\pm$ 1.2, transitions per 100 cell cycles (mean $\pm$ SD), respectively (Figs. 4F & G). These events did not correlate between sister cells (Table 1), consistent with independent, stochastic events. Inter-state switching on average showed bigger and longer-lasting fold-changes than intra-state steps in production rates (Fig. 9H). Together, these results suggest that gene expression dynamics are dominated by a combination of step-like changes due to bursty transcription on shorter timescales, and abrupt, apparently stochastic inter-state switching events on

longer timescales.

2.3.6 '2i' inhibitors modulate bursty transcription and state-switching dynamics

We next asked how gene expression heterogeneity and dynamics change in response to key perturbations. Dual inhibition of MEK and GSK3$beta, known as '2i', were shown to enhance pluripotency and reduce Rex1 and Nanog heterogeneity [26, 44]. However, it has remained unclear how 2i affects the distribution of other heterogeneously expressed regulatory genes, and what impact it has on dynamic fluctuations in gene expression.

We found that addition of 2i to serum+LIF media reduced variability in the mRNA levels of most genes (Fig. 5A). In principle, this could reflect the elimination of one cellular state and/or changes in burst parameters. In 2i, the bimodal genes from Fig. 2A became unimodal, suggesting that 2i suppresses one of the two cellular states (Figs. 5A, 10A). In the case of Nanog, the remaining state increased its mean transcript level by 1.5-fold, to what we term Nanog-SH (Super High). Tet1, Sox2, and Tcl1 also became unimodal, but displayed an overall decrease in absolute expression. With long-tailed genes, we found that mean Dppa3 expression decreased slightly, while Prdm14 and Tbx3 became more homogenously expressed, exhibiting an increase in mean expression by 300% and 1000%, respectively. These changes reflect the fact that nearly all cells were now observed to express Prdm14 and Tbx3. Thus, 2i appeared to reduce variability in most genes, either by eliminating bimodality or by increasing their burst frequency.

Recently, it was shown that 2i-treated cells exhibit differentiation propensity similar to sorted Rex1-high subpopulations in embryoid body formation, suggesting they may represent similar functional states (Marks et al., 2012). We used the time-lapse movie system to compare the dynamic behavior of 2i-treated cells to that of cells in the Rex1-high subpopulation. Consistent with mRNA measurements, 2i shifted most cells into a Nanog-SH state (96% of total), characterized by 3-fold higher median

production rate compared to the Nanog-high state in serum+LIF (Fig. 5B). Only a small fraction of cells showed expression overlapping with the Nanog-low state in serum+LIF at the beginning of the movies (after 6 days in treatment). Moreover, these cells switched to the Nanog-SH state at a ¿40-fold higher rate than the Nanog-low to Nanog-high switching rate measured in serum+LIF, with no reverse transitions observed (Figs. 5C & 10B). These observations suggest that 2i increases the Nanog-low to Nanog-high switching rate and reduces or eliminates Nanog-high to Nanog-low transitions (Fig. 5C).

What effect, if any, does 2i have on the dynamics of gene expression noise? Static distributions suggested that 2i increased Nanog burst frequency 45% from 0.39 to 0.55 burst/hour, using Nanog mRNA half-life previously estimated (Table S1 in [57]) and assuming no change between conditions. To analyze the effects on dynamics, we computed the "mixing time", previously introduced to quantify the mean timescale over which a cell maintains a given expression level relative to the rest of the cell population (Sigal et al., 2006). Simulations of the bursty gene expression model showed that higher burst frequencies lead to faster mixing times, while burst size has little effect (Fig. 10C). Together with smFISH measurements, this model predicted that Nanog mixing times should be faster in 2i. Qualitatively consistent with this prediction, the mixing time of Nanog production rate was reduced from 8.5 hours in Nanog-high in serum+LIF media to 1.7 in Nanog-SH cells in 2i-containing media (Figs. 5D & 10D).

Together, these results indicate that 2i impacts ESC heterogeneity at three levels: first, it reduces gene expression variation in many, but not all, genes. Second, it eliminates one cell state by increasing the rate of transitions out of the Nanog-low state and inhibiting the reverse transition. Third, as shown for Nanog, 2i increases burst frequency and reduces mixing times, effectively speeding up the intra-state equilibration rate within the cell population.

2.3.7 DNA methylation modulates metastability

Previous work has shown that in addition to reducing heterogeneity, 2i also diminishes global levels of DNA methylation [29, 30, 48]. While the Rex1-high and -low states appear differentially methylated (Figs. 3B-E), it remains unclear whether methylation plays a functional role in stabilizing these states. To address this issue, we used a triple-knockout (TKO) cell line lacking the active DNA methyltransferases Dnmt1, Dnmt3A, and Dnmt3B [58]. We compared the expression distribution of Rex1, Nanog, and Esrrb in TKO cells to its parental line using smFISH. The TKO cell lines had 35±2% fewer cells in the Rex1-low state (Fig. 5E), with similar results observed for Nanog and Esrrb. This change did not reflect global up-regulation of all genes, as expression of the houskeeping gene SDHA was indistinguishable between the two cell lines. These results suggest that DNA methyltransferases increase the population of the Rex1-low state.

To see if these results could be recapitulated with acute rather than chronic perturbations to methylation, we assayed changes in heterogeneity in Rex1-dGFP reporter cells exposed to 70nM 5-azacytidine (5-aza), an inhibitor of DNA methylation. Within six days, the number of cells in the Rex1-low state diminished by more than half from 29% to 13% of all cells (Fig. 5F). Thus, acute as well as chronic methylation inhibition reduced the occupancy of the low state.

Finally, we asked whether methylation was similarly required for cells to return to the low state after removal of 2i from 2i+serum+LIF conditions. The Rex1-low population began to emerge within 48h of 2i removal (Fig. 5G). However, when 2i was removed and replaced with 5-aza, the emergence of Rex1-low cells was severely delayed and diminished. After 6 days, 5-aza treated cells only showed 6% Rex1-low cells, compared to 25% in DMSO-treated cells. Together, these results suggest that methylation is required for normal exit from the 2i state. Reduced methylation in 2i thus contributes to the stability of the 2i 'ground state' [26, 29, 30, 48].

2.4 Discussion

Recent work on ESC biology from a systems perspective has highlighted the apparent complexity and strong interconnectivity of the circuit governing pluripotency [59, 60]. But it has been unclear how variably this circuit behaves in different cells, and to what extent population average measurement techniques may obscure its single-cell dynamics. Because gene expression is a stochastic process, levels of both mRNA and protein in each cell are effectively random variables, best characterized by their distributions. The framework of stochastic gene expression provides a tool to more rigorously and quantitatively separate stochastic fluctuations inherent to gene expression from variation due to multiple cell states specified by the underlying transcriptional and signaling circuit. While the simplified model of bursty transcription used here can explain the data, other models, including the "telegraph" model of transcription, may provide other insights [36, 61].

Our data suggest that heterogeneity emerges in three distinct ways: first, gene expression is inherently noisy, occurring in stochastic bursts, even in genes such as Oct4 that are distributed in a relatively uniform fashion. Second, cells switch stochastically between distinct states that impact the expression of many genes in a coordinated manner. Third, 'long-tailed' regulators such as Prdm14, Tbx3, and Dppa3 are uncorrelated with one another and show low burst frequencies and large burst sizes, leading to very high variability. Live cell imaging will be required to determine the absolute burst kinetics for these genes. However, an mRNA distribution in which only a small subpopulation of cells exhibit a large number of mRNA molecules for a particular gene need not, and most likely does not, indicate a distinct cellular state. Moreover, infrequent bursting may provide a potential mechanism for stochastic priming of cell fate decision-making [3, 42]. Further investigation of this possibility will require determining whether these bursts propagate to influence subsequent cell fate decision-making events [62, 63].

The data above implicate methylation as a key regulatory mechanism affecting stochastic switching between states. Methylation was previously explored in ES cells

at the population level [24, 25, 29, 30, 48, 49], but it remained unclear whether methylation itself contributes to heterogeneity [6, 10, 12, 13]. smFISH data revealed a strong reciprocal relationship between the hydroxymethylase Tet1 and the DNA methyltransferase Dnmt3b, with Tet1 expressed more highly in the Rex1-high state, and Dnmt3b expressed more highly in the Rex1-low state. This difference in expression correlates with a differential global DNA methylation and in the methylation of the promoters of key pluripotency regulators. Finally, methylation appears to be functionally required for transitions, since either genetic deletion of DNA methyltransferases or pharmacological inhibition both impact the populations of the two cell states and the underlying dynamics of state-switching (Figs. 5E-G). It will be interesting to see whether methylation plays similar functional roles in other stochastic state-switching systems.

These data provoke further questions about the molecular mechanisms through which methylation is regulated and through which it modulates metastability. For example, while known methyl binding proteins that aid in methylation-dependent chromatin compaction and silencing are expressed in ES cells [26], DNA methylation may also inhibit binding of transcription factors [64–66], and can control mRNA isoform selection via alternative splicing [64]. The Esrrb gene, whose activity is central to maintenance of pluripotency [67, 68], may provide a good model system to investigate the effects of methylation, since its methylation levels and expression levels are both strongly state-dependent. Regulation of this methylation likely involves Prdm14, which is known to inhibit Dnmt3b expression [13, 28–30, 52, 53]. Given the long-tailed expression pattern of Prdm14 observed here in serum+LIF and its strong up-regulation in 2i, it will be interesting to see how much of the variation in Dnmt3b/Tet1, and methylation more generally, can be attributed to bursty expression of Prdm14.

Previous studies of ESC gene expression dynamics have focused on the equilibration of FACS-sorted subpopulations of high and low Nanog and Rex1 expression [6, 12]. Two groups explored transcriptional circuit models to explain the long timescales of state-switching dynamics [69, 70]. These included noise-induced bistable switches,

oscillators, and noise-excitable circuits [71]. Our dynamic data demonstrate that both Nanog-high and Nanog-low states in serum+LIF conditions typically persist for ≥4 cell cycles, and that state-switching events are abrupt at the level of promoter activity. Depending on protein and mRNA half-lives, the timescale of resulting protein level changes may follow somewhat more slowly. State-switching events are also infrequent (¡10% per cell cycle), and uncorrelated between sister cells. Together, these findings appear incompatible with oscillatory or excitable models, which predict deterministic state-switching or an unstable excited state, respectively, but are consistent with the stochastic bistable switch model previously proposed [70]. These properties could make this state-switching system a useful model for understanding the circuit level dynamics of spontaneous cell state transitions in single cells.

Several competing explanations were proposed for the apparent heterogeneity in Nanog expression in serum+LIF conditions. These models suggest that heterogeneity is an artifact of knock-in reporters [72], or that it arises at least in part from monoallelic regulation [73] or is manifested biallelically [74, 75]. Our smFISH data support the existence of Nanog expression heterogeneity in wild-type cells in a standard feeder-free culture condition. Further, both static and dynamic measurements indicate that intra-state heterogeneity in Nanog is consistent with bursty transcription, with a relatively low burst frequency of 0.39 burst/hour. Thus, active transcriptional loci analyzed by smFISH against nascent transcripts [73] would be expected to 'flicker' on and off stochastically due to bursting. Such bursting could also lead to the misleading appearance of weak correlations between alleles in static snapshots, and in measurements based on destabilized fluorescent reporters. On the other hand, the Nanog protein fusion reporters analyzed by Filipczyk et al showed correlated static levels between alleles, likely because the longer lifetime of their reporters allowed integration of signal over many transcriptional bursts, and because transitions between cellular states are rare and affect both alleles in a correlated fashion. The results of Faddah et al with dual transcriptional reporters similarly showed general correlations between the two alleles, consistent with the smFISH correlations shown here (Figs. 6E & 9B). Taken together, these previous studies and data presented here appear

to converge on a relatively simple picture of heterogeneity based on two states and stochastic bursting.

The quantitative measurement and analysis platform described above should enable further investigation of the structure of static and dynamic heterogeneity in single ESCs. With the advent of higher dimensionality smFISH [76, 77], single-cell RNA-Seq, and microfluidic high-throughput qPCR approaches, as well as improved methods for rapidly and accurately constructing knock-in reporters [78], it will soon be possible to explore the dynamics of ESC components in higher dimensions in individual cells, both within metastable states and during cell state transitions [79]. Ultimately, this should provide the capability of better understanding the dynamic architecture of cell fate transition circuits.

2.5 Experimental Procedures

2.5.1 Culture Conditions and Cell Lines

E14 cells (E14Tg2a.4) obtained from Mutant Mouse Regional Resource Centers were used for smFISH studies. NKICit cells, created by Kathrin Plath, were generated by targeting the endogenous Nanog locus in V6.5 cells with H2B-Citrine-IRES-Neo-SV40pA (Fig. 9A). NKICit+Cer cells were generated by randomly integrating into NKICit cells a linearized PGK-H2B-Cerulean-BGHpA-SV40-Hygro-SV40pA vector. OBACCer cells were generated by randomly integrating into E14 cells (a kind gift from Bill Skarnes and Peri Tate) a linearized bacterial artificial chromosome (BAC) containing the Oct4 locus (BACPAC (CHORI)), in which H2B-Cerulean-SV40pA-PGK-Neo-BGHpA was inserted before the coding sequence (Fig. 9A). Rex1-dGFP cells were kindly provided by the Austin Smith lab (Wray et al., 2011). All cells were maintained on gelatin-coated dishes without feeders.

2.5.2 smFISH Hybridization, Imaging, and Analysis

The RNA FISH protocol from Raj et al 2008 was adapted for fixed cells in suspension. See supplemental experimental procedures for details. Semi-automated dot detection and segmentation were performed using custom Matlab software. A Laplacian-of-Gaussian (LoG) Kernel was used to score potential dots across all cells. The distribution of these scores across all potential dots was thresholded by Otsu's method to discriminate between true dots and background dots (see Figs. 6A-D).

2.5.3 mRNA Distribution Fitting

The Negative Binomial Distribution is defined as

$$P(n, r, p_o) = \binom{n + r - 1}{n} p_o^r (1 - p_o)^n$$

where n=number of transcripts per cell, p_o=probability of transition from the on promoter state to the off promoter state, and r=number of bursting events per mRNA half-life. The average burst size is computed as b=(1-p_o)/p_o. Using this model, individual mRNA distributions were fit using maximum likelihood estimation. To discriminate between unimodal and bimodal fits, two tests were used to ensure that the improvement of the fit is counterbalanced by the additional degrees of freedom from the added parameters. To be considered bimodal, a distribution was required to pass both Akaike Information Criteria (AIC) and the log-likelihood ratio test ($p<0.05$).

2.5.4 Fluorescence time-lapse microscopy and data analysis

Reporter cells were mixed with unlabeled parental cells at 1:9 ratio and plated at a total density of 20,000 cells/cm2 on glass-bottom plates (MatTek) coated with human laminin-511 (BioLamina) ¿12 hours before imaging. Images were acquired every 15 minutes for 50 hours with daily medium change. Cells were segmented and tracked from the acquired images using our own Matlab code (see supplementary for image

analysis methods).

2.5.5 2i Perturbation and Analysis

For 2i treatment we supplemented serum+LIF media with MEK inhibitor PD0325901 at 1uM and GSK3 inhibitor CH99021 at 3uM. Cells grown in serum+LIF media were treated with 2i for 6 days before harvesting for smFISH assay and imaging for movies.

2.5.6 Methylation Analysis and Perturbation

RRBS preparation and high-throughput sequencing were performed by Zymo Research. Analysis was performed using Bismark and Galaxy, with a single CpG coverage threshold $\geq$5. 5-azacytidine (Sigma) was used at a final concentration of 70nM. 5mC ELISA was performed with ELISA 5mC kit (Zymo).

2.6 Accession Information

Sequencing data has been deposited in NCBI's GEO under accession number GSE58396.

2.7 Acknowledgements

Z.S.S., J.Y., J.T., L.C., M.A.S., and M.B.E. conceived experiments. Z.S.S., J.Y., J.T., and J.A.H. performed experiments and analyzed data, with Z.S.S. leading smFISH and methylation experiments, and J.Y. leading the movie experiments and modeling. A.A. contributed computational algorithms. M.A.S. and M.B.E. supervised research. Z.S.S., J.T., J.Y., and M.B.E. wrote the manuscript with substantial input from all authors. We thank Jordi Garcia-Ojalvo, Xiling Shen, Georg Seelig, and David Sprinzak for helpful comments on the manuscript; the Kathrin Plath Lab, the Austin Smith Lab, and Riken for kindly providing reporter and knockout cell lines; and the Caltech FACS Facility for assistance with cell sorting. This work was supported by the National Institutes of Health grants R01HD075605A, R01GM086793A,

and P50GM068763; the Weston Havens Foundation; Human Frontiers Science Program; the Packard Foundation; a Wellcome Trust Investigators Grant to MAS; and a KAUST, APART, and CIRM Fellowship to JT. This work is funded by the Gordon and Betty Moore Foundation through Grant GBMF2809 to the Caltech Programmable Molecular Technology Initiative.

2.8 Figures

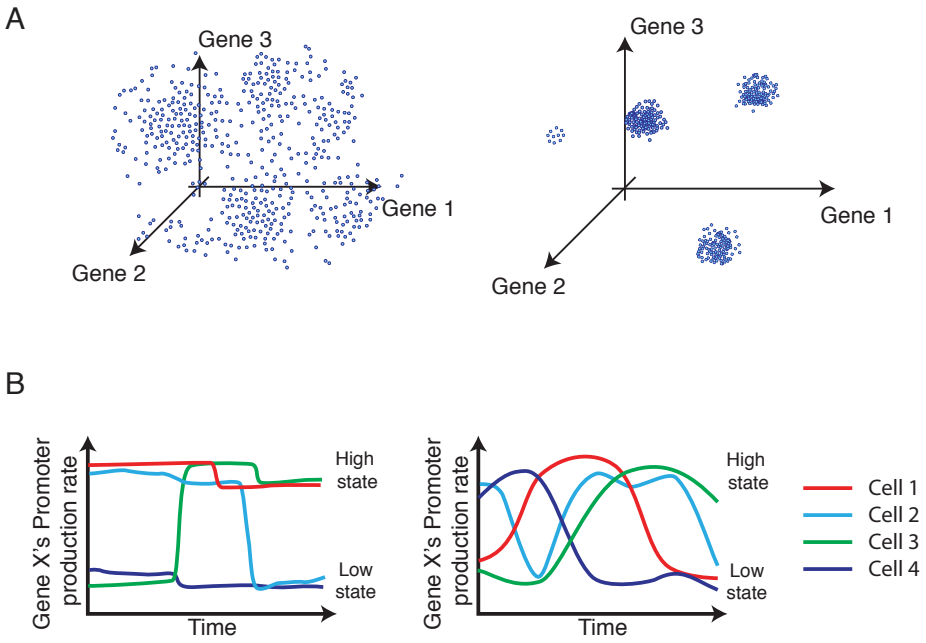

Figure 2.1: Different types of gene expression heterogeneity

(A) Intrinsic noise in gene expression can lead to uncorrelated variation (left), while the coexistence of distinct cellular states can produce correlated variability in gene expression (right). Both panels depict schematic static 'snapshots' of gene expression. (B) Dynamically, gene expression levels could vary infrequently and abruptly (left) or more frequently and gradually (right) both within and between cellular states (schematic).

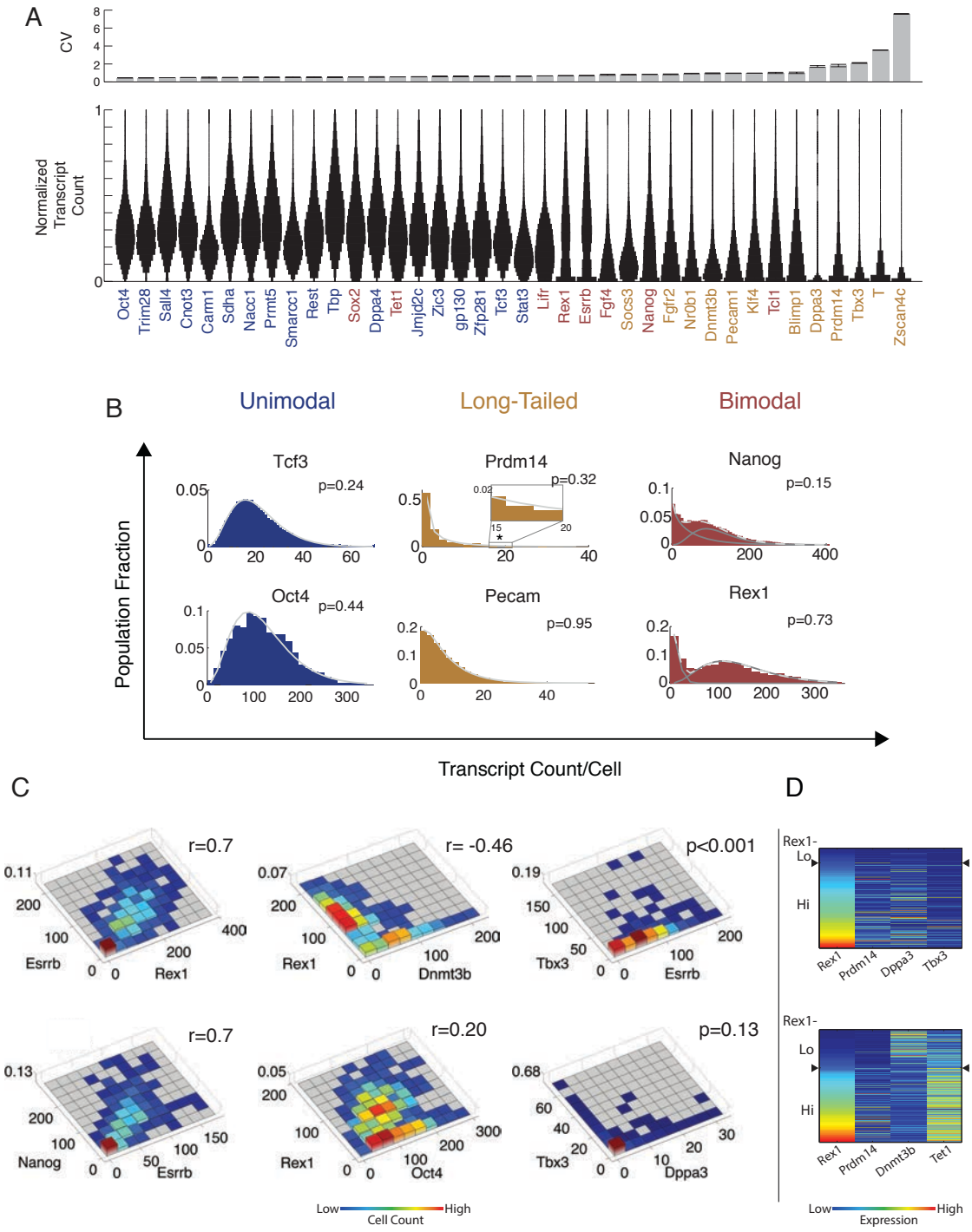

Figure 2.2: smFISH reveals gene expression heterogeneity and correlation

(A) Top: coefficients of variation (CV, mean±SEM) for ESC-associated regulators and housekeeping genes. Bottom: Distributions (violin plots) normalized by maximum expression level reveal qualitatively distinct gene expression distributions. Genes are sorted by increasing CV. (B) Smoothed histograms for mRNA distributions overlaid with NB fits. Solid lines show individual NB distributions. Dashed gray lines show their sum (for bimodal genes). * denotes 95th percentile for Prdm14. p-value: χ^2 goodness of fit test. (C) Pairwise relationships between genes, analyzed by smFISH (r, Pearson correlation coefficient; p-value by 2D K-S test (methods, figs. 7A,B). (D) Heat maps show examples of 4-dimensional data sets.

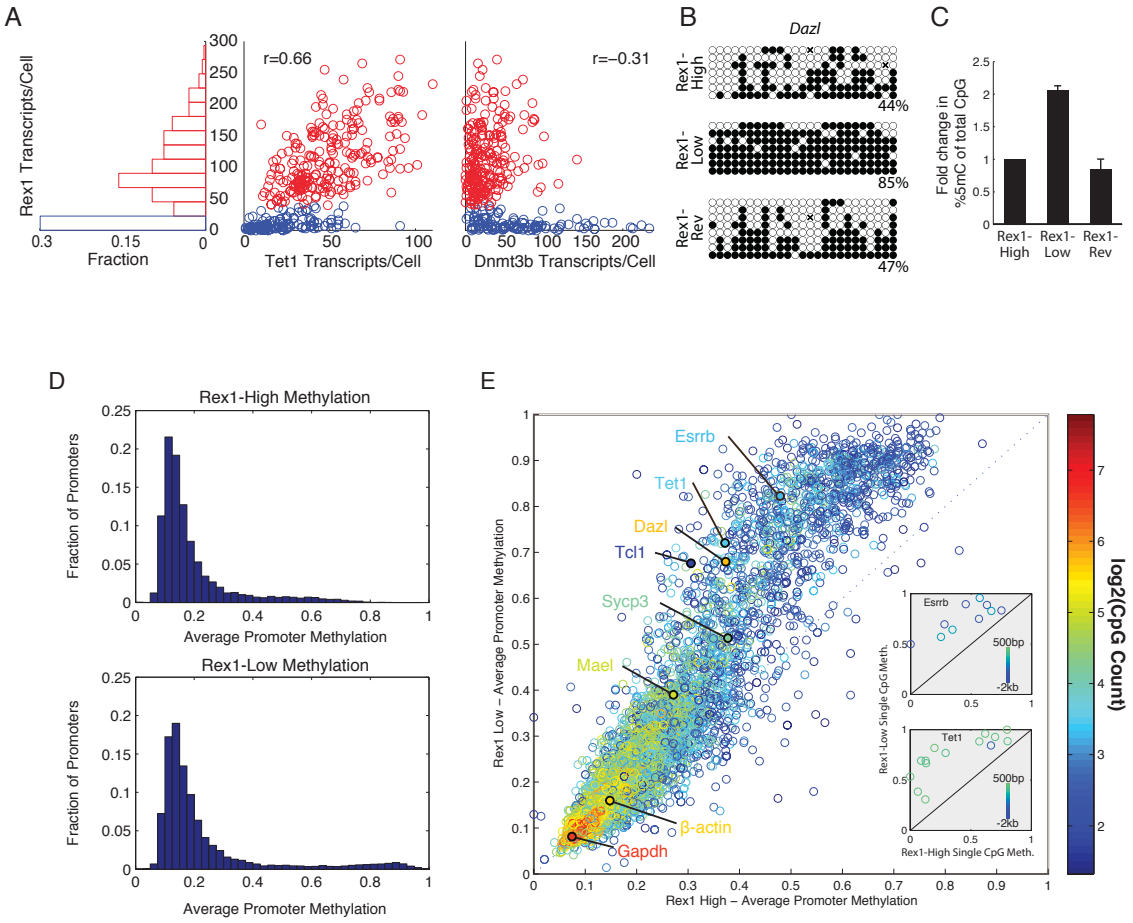

Figure 2.3: The two Rex1 states are differentially methylated

(A) smFISH measurements show Rex1 bimodality is correlated with Tet1, and anticorrelated with Dnmt3b expression. (B) Locus-specific bisulfite sequencing of the Dazl promoter. Methylation levels shown are in the Rex1-high (top), Rex1-low (middle), and Rex1-low-to-high reverting (bottom) populations. (C) Global levels of 5mC measured by quantitative ELISA in the Rex1-high, -low, and -low-to-high reverting cells. (D) Histogram of promoter methylation shows bimodality in the Rex1-high (top) and -low (bottom) states, as quantified by RRBS. (E) Scatter plot of promoter methylation between Rex1-high and -low states. Each point is the methylation fraction of a single gene promoter, color-coded by the number of CpGs in that promoter. Divergence from the diagonal implies differential methylation between states. Inset) Single CpGs in the promoter of the specific gene labeled, color coded by distance from TSS; see Fig. 8C for additional genes.

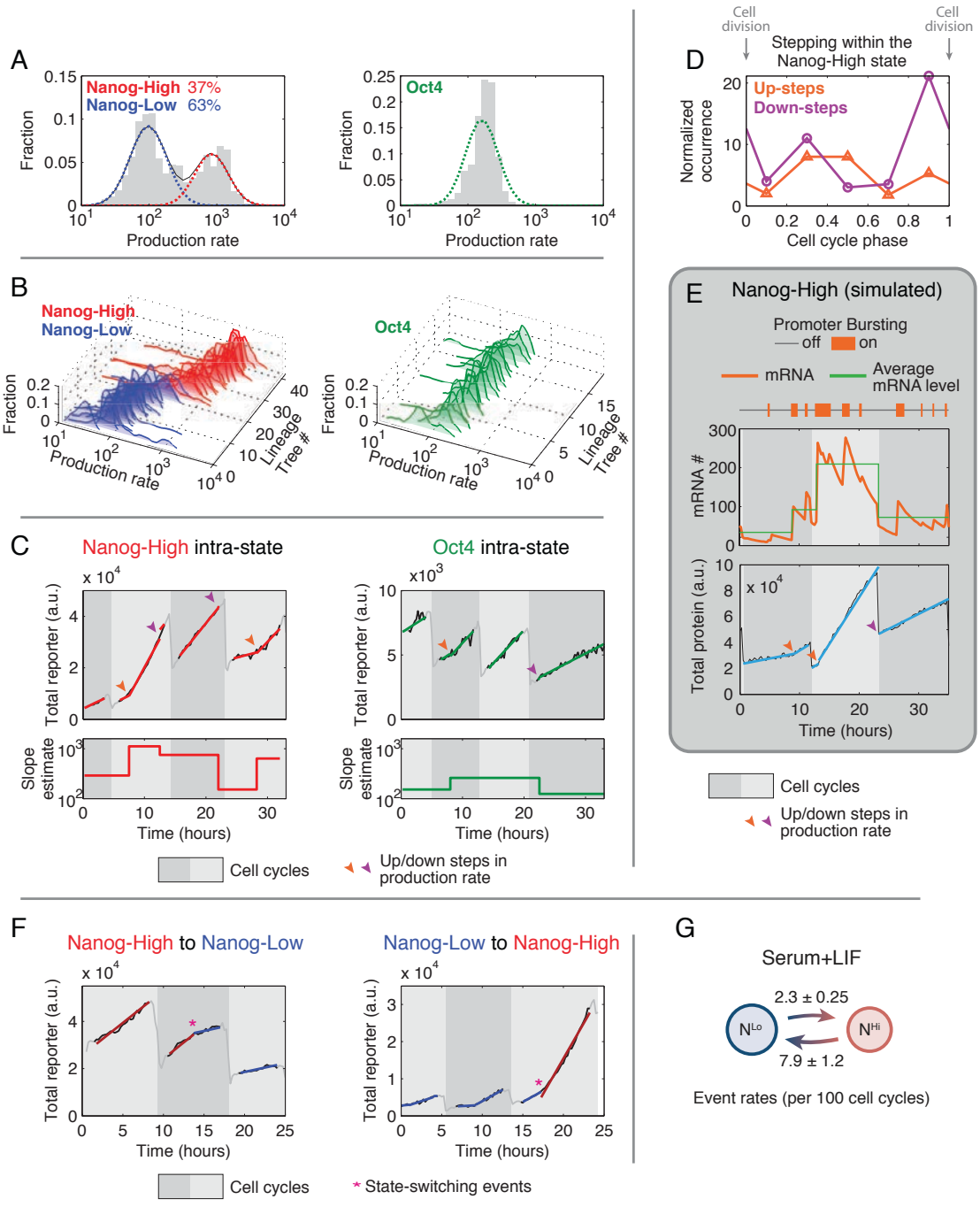

Figure 2.4: Movies reveal transcriptional bursting and state-switching dynamics in individual cells

(A) Distribution of Nanog and Oct4 production rates from representative movies in serum+LIF, and Gaussian fits to the components. Production rates were extracted from a total of 376 and 103 tracked cell cycles for Nanog and Oct4, respectively. (B) Production rate distributions of individual cell lineage trees, each consists of closely related cells descending from a single cell. Lineage trees are color-coded by the state they spend the majority of time in. (C) Example single lineage traces exhibiting step-like changes in production rates within a state. (D) Cell cycle phase distribution of steps within the Nanog-high state. Step occurrences are normalized by the frequencies of each cell cycle phase observed in the tracked data. (E) Representative trace showing apparent steps from simulations under the bursty transcription model, using parameters estimated from mRNA distribution for the Nanog-high state (see Supp. Info.; see Fig. 9E for simulation of Oct4 dynamics). (F) Example traces of individual cells switching between Nanog-low and Nanog-high states. (G) Empirical transition rates (mean $\pm$ SD) between the two Nanog states (NHi, Nanog-high; NLo, Nanog-low).

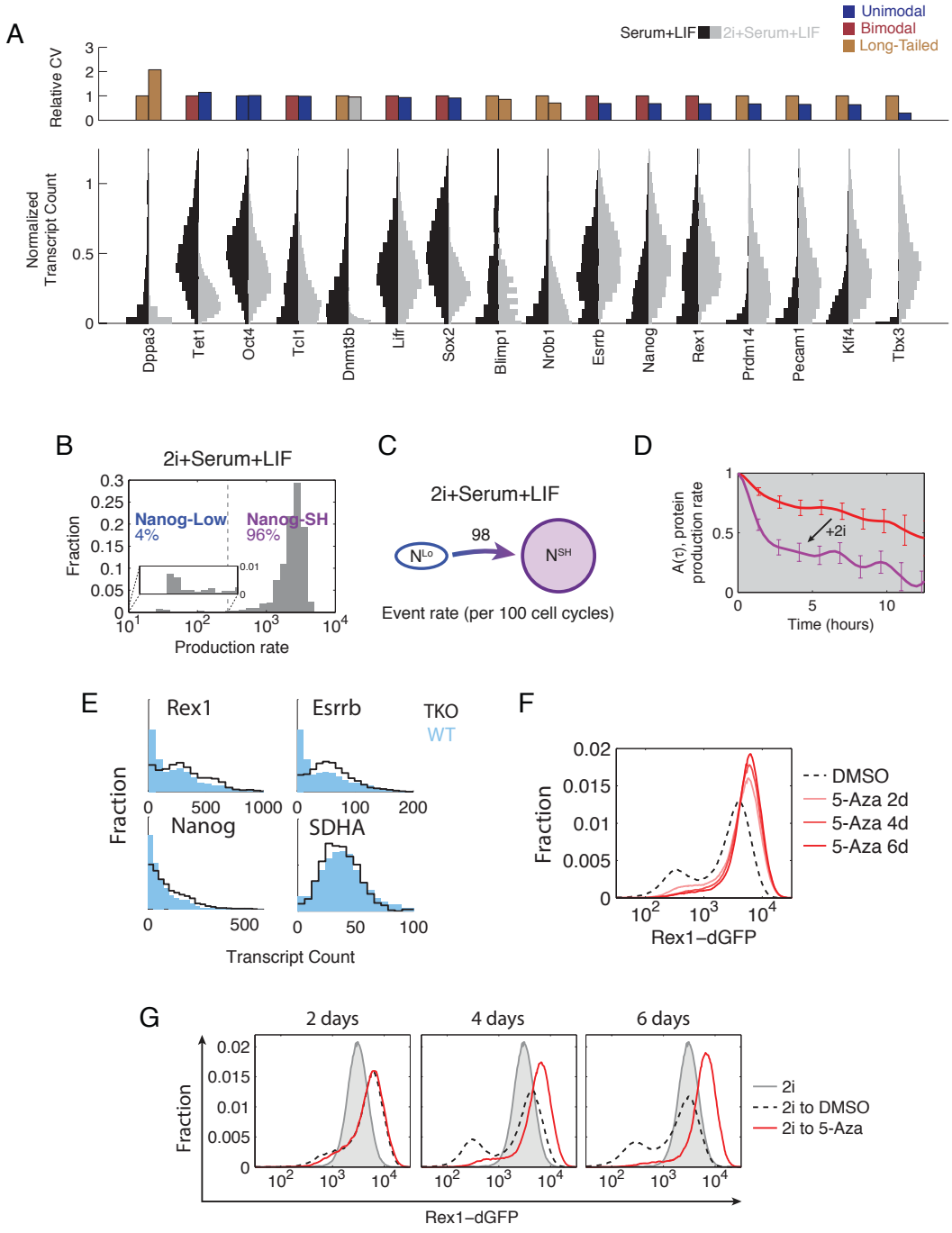

Figure 2.5: 2i and DNA methylation modulate bursty transcription and state-switching dynamics

(A) Comparison of mRNA distributions and CV between cells grown in serum+LIF and 2i+serum+LIF. Top: For each gene, the CV in serum+LIF is plotted on the left, and the CV for 2i+serum+LIF is plotted on the right. Dnmt3b in 2i+serum+LIF is represented in gray to reflect its marginal case of poor quality of fit in both bimodal and long-tailed models. Bottom: The left half of each violin represents the mRNA distribution in serum+LIF, while the right represents 2i+serum+LIF. For each gene, both conditions are normalized by the same value that is the larger of the pairs 95th percentile expression level. (B) Distribution of Nanog production rates from movies in 2i+serum+LIF. (C) Empirical transition rates between the two Nanog states in the presence of 2i (NLo, Nanog-low; NSH, Nanog-SH). (D) Mixing time in each condition is estimated from auto-correlation, $A(\tau)$, of production rate ranks shown in Fig. 10D, right panels. Red: serum+LIF; Purple: 2i+serum+LIF; Error bars: standard deviation, bootstrap method. (E) Comparison of transcriptional heterogeneity between Dnmt TKO (black line) and the parental line (blue bars) as measured by smFISH for Rex1, Nanog, Esrrb, and SDHA. Note that for Rex1/Nanog/Essrb, there are fewer off cells in the leftmost bins for the TKO than WT. (F) Rex1-dGFP distribution as measured by flow cytometry grown in serum+LIF with 5-aza or DMSO (carrier control). Time-points were taken after 2, 4, and 6 days. (G) Cells were grown in 2i+serum+LIF, and subsequently re-plated into serum+LIF with 5-aza or DMSO (carrier control). Time-points were taken after 2, 4, and 6 days. GFP levels were measured by flow-cytometry.

Primary References

1. Lander, A. D., Gokoffski, K. K., Wan, F. Y. M., Nie, Q. & Calof, A. L. Cell lineages and the logic of proliferative control. *PLoS Biol* (2009) (cit. on p. 26).
2. Rompolas, P., Mesa, K. R. & Greco, V. Spatial organization within a niche as a determinant of stem-cell fate. *Nature* (2013) (cit. on p. 26).
3. Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M. & Elowitz, M. B. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* (2006) (cit. on pp. 26, 27, 37).
4. Gupta, P. B. *et al.* Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* (2011) (cit. on p. 26).
5. Canham, M. A., Sharov, A. A., Ko, M. S. H. & Brickman, J. M. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS Biol* (2010) (cit. on p. 26).
6. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* (2007) (cit. on pp. 26, 38).
7. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* (2008) (cit. on p. 26).
8. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* (2010) (cit. on p. 26).
9. Yamanaka, Y., Lanner, F. & Rossant, J. FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. *Development* (2010) (cit. on p. 26).
10. Hayashi, K., Lopes, S. M. C. d. S., Tang, F. & Surani, M. A. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* (2008) (cit. on pp. 26, 38).
11. Singh, A. M., Hamazaki, T., Hankowski, K. E. & Terada, N. A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells* (2007) (cit. on p. 26).
12. Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K. & Niwa, H. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development* (2008) (cit. on pp. 26, 38).

13. Yamaji, M. *et al.* PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell Stem Cell* (2013) (cit. on pp. 26, 30, 38).

14. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* (2010) (cit. on p. 26).

15. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* (2008) (cit. on p. 26).

16. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* (2008) (cit. on p. 26).

17. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* (2014) (cit. on p. 26).

18. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* (2013) (cit. on p. 26).

19. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* (1999) (cit. on p. 26).

20. Hackett, J. A. *et al.* Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* (2013) (cit. on p. 26).

21. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* (2007) (cit. on p. 26).

22. Schübeler, D. *et al.* Genomic targeting of methylated DNA: influence of methylation on transcription, replication, chromatin structure, and histone acetylation. *Mol Cell Biol* (2000) (cit. on p. 26).

23. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* (2012) (cit. on p. 26).

24. Fouse, S. D. *et al.* Promoter CpG Methylation Contributes to ES Cell Gene Regulation in Parallel with Oct4/Nanog, PcG Complex, and Histone H3 K4/K27 Trimethylation. *Cell Stem Cell* (2008) (cit. on pp. 26, 38).

25. Mohn, F. *et al.* Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* (2008) (cit. on pp. 26, 38).

26. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* (2012) (cit. on pp. 26, 31, 34, 36, 38).

27. Wray, J. *et al.* Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. *Nat Cell Biol* (2011) (cit. on p. 26).

28. Ficz, G. *et al.* FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell Stem Cell* (2013) (cit. on pp. 26, 38).

29. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* (2013) (cit. on pp. 26, 36, 38).

30. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* (2013) (cit. on pp. 26, 30, 36, 38).

31. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* (2003) (cit. on p. 27).

32. Elowitz, M., Levine, A., Siggia, E. & Swain, P. Stochastic Gene Expression in a Single Cell. *Science* (2002) (cit. on p. 27).

33. Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* (2006) (cit. on p. 27).

34. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nat Genet* (2002) (cit. on p. 27).

35. Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* (2000) (cit. on p. 27).

36. Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theoretical Population Biology* (1995) (cit. on pp. 27, 37).

37. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biol* (2006) (cit. on pp. 27, 28).

38. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* (2008) (cit. on p. 27).

39. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* (2011) (cit. on p. 27).

40. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* (2006) (cit. on p. 27).

41. Choi, P. J., Cai, L., Frieda, K. & Xie, X. S. A Stochastic Single-Molecule Event Triggers Phenotype Switching of a Bacterial Cell. *Science* (2008) (cit. on p. 27).

42. Maamar, H., Raj, A. & Dubnau, D. Noise in gene expression determines cell fate in Bacillus subtilis. *Science* (2007) (cit. on pp. 27, 37).

43. Zong, C., So, L.-h., Sepúlveda, L. A., Skinner, S. O. & Golding, I. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Mol Syst Biol* (2010) (cit. on p. 27).

44. Wray, J., Kalkan, T. & Smith, A. G. The ground state of pluripotency. *Biochem Soc Trans* (2010) (cit. on pp. 27, 34).

45. Paulsson, J., Berg, O. G. & Ehrenberg, M. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc Natl Acad Sci USA* (2000) (cit. on p. 28).

46. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* (2012) (cit. on p. 29).

47. Auernhammer, C. J., Bousquet, C. & Melmed, S. Autoregulation of pituitary corticotroph SOCS-3 expression: characterization of the murine SOCS-3 promoter. *Proc Natl Acad Sci USA* (1999) (cit. on p. 30).

48. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* (2011) (cit. on pp. 30, 36, 38).

49. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* (2010) (cit. on pp. 30, 38).

50. Koh, K. P. *et al.* Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* (2011) (cit. on p. 30).

51. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* (2011) (cit. on p. 30).

52. Grabole, N. *et al.* Prdm14 promotes germline fate and naive pluripotency by repressing FGF signalling and DNA methylation. *EMBO Rep* (2013) (cit. on pp. 30, 38).

53. Ma, Z., Swigut, T., Valouev, A., Rada-Iglesias, A. & Wysocka, J. Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat Struct Mol Biol* (2011) (cit. on pp. 30, 38).

54. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* (2008) (cit. on pp. 30, 31).

55. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet* (2010) (cit. on p. 30).

56. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* (2005) (cit. on p. 33).

57. Sharova, L. *et al.* Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA research* (2009) (cit. on p. 35).

58. Tsumura, A. *et al.* Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* (2006) (cit. on p. 36).

59. Chen, X. *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* (2008) (cit. on p. 37).

60. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* (2006) (cit. on p. 37).

61. Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* (2009) (cit. on p. 37).

62. Dunlop, M. J., Cox, R. S., Levine, J. H., Murray, R. M. & Elowitz, M. B. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet* (2008) (cit. on p. 37).

63. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* (2005) (cit. on p. 37).

64. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* (2011) (cit. on p. 38).

65. Takizawa, T. *et al.* DNA methylation is a critical cell-intrinsic determinant of astrocyte differentiation in the fetal brain. *Dev Cell* (2001) (cit. on p. 38).

66. You, J. S. *et al.* OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proc Natl Acad Sci USA* (2011) (cit. on p. 38).

67. Festuccia, N. *et al.* Esrrb Is a Direct Nanog Target Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell Stem Cell* (2012) (cit. on p. 38).

68. Martello, G. *et al.* Esrrb Is a Pivotal Target of the Gsk3/Tcf3 Axis Regulating Embryonic Stem Cell Self-Renewal. *Cell Stem Cell* (2012) (cit. on p. 38).

69. Glauche, I., Herberg, M. & Roeder, I. Nanog variability and pluripotency regulation of embryonic stem cells–insights from a mathematical model analysis. *PLoS ONE* (2010) (cit. on p. 38).

70. Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol* (2009) (cit. on pp. 38, 39).

71. Furusawa, C. & Kaneko, K. A dynamical-systems view of stem cell biology. *Science* (2012) (cit. on p. 39).

72. Faddah, D. A. *et al.* Single-Cell Analysis Reveals that Expression of Nanog Is Biallelic and Equally Variable as that of Other Pluripotency Factors in Mouse ESCs. *Cell Stem Cell* (2013) (cit. on p. 39).

73. Miyanari, Y. & Torres-Padilla, M.-E. Control of ground-state pluripotency by allelic regulation of Nanog. *Nature* (2012) (cit. on p. 39).

74. Filipczyk, A. *et al.* Biallelic expression of nanog protein in mouse embryonic stem cells. *Cell Stem Cell* (2013) (cit. on p. 39).

75. Hansen, C. H. & van Oudenaarden, A. Allele-specific detection of single mRNA molecules in situ. *Nat Meth* (2013) (cit. on p. 39).

76. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat Meth* (2014) (cit. on p. 40).

77. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Meth* (2012) (cit. on p. 40).

78. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* (2013) (cit. on p. 40).

79. Kueh, H. Y. *et al.* Positive feedback between PU.1 and the cell cycle controls myeloid differentiation. *Science* (2013) (cit. on p. 40).

2.9 Supplemental Data and Figures

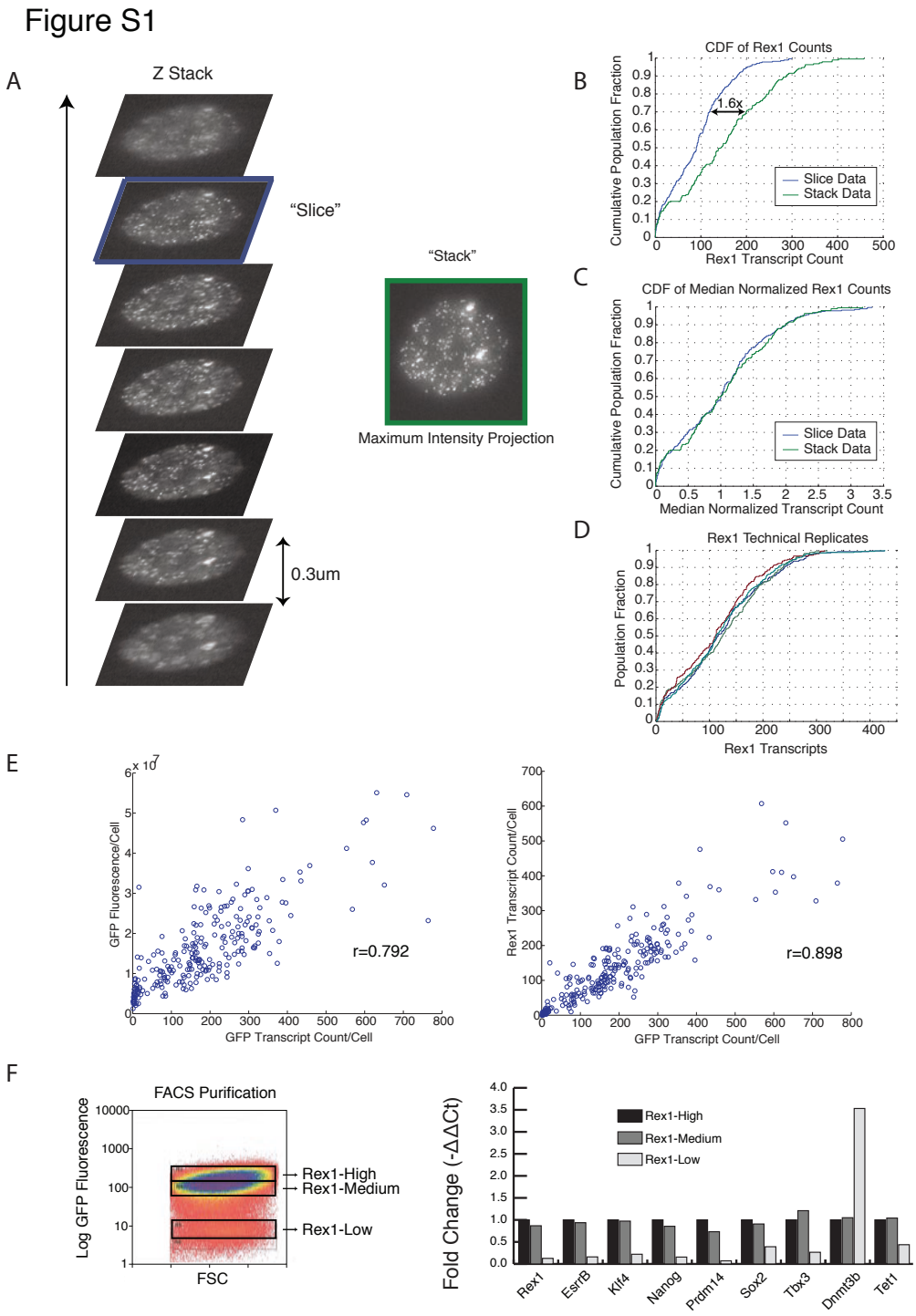

Figure 2.6: Validation of smFISH

(A) A stack of snapshots taken through the whole volume of a single cell; the resulting maximum-intensity projection (green box), and a single slice (blue box) are fed into the image-processing algorithm for dot-detection. (B) Cumulative Distributions of dot counts for each of the two imaging approaches is shown across a population of cells. (C) Same distributions as in B, but normalized by the sample median. (D) Technical replicates for the single-slice approach. (E) Correlation between dGFP protein fluorescence as measured simultaneously with dGFP transcripts. (left), and correlation between Rex1 (unmodified allele) and dGFP (knock-in reporter on second allele) transcripts (right). r is the Pearson correlation coefficient. (F) (Left) Sorted subpopulations of the bimodal Rex1-dGFP knock-in reporter. (Right) qPCR results on these subpopulations for a subset of target genes also examined by smFISH. Values were normalized to expression levels of the housekeeping gene Gapdh, and are represented as $2^{(-\Delta\Delta C_t)}$ with respect to the 'Rex1-high' subpopulation

Figure S2.

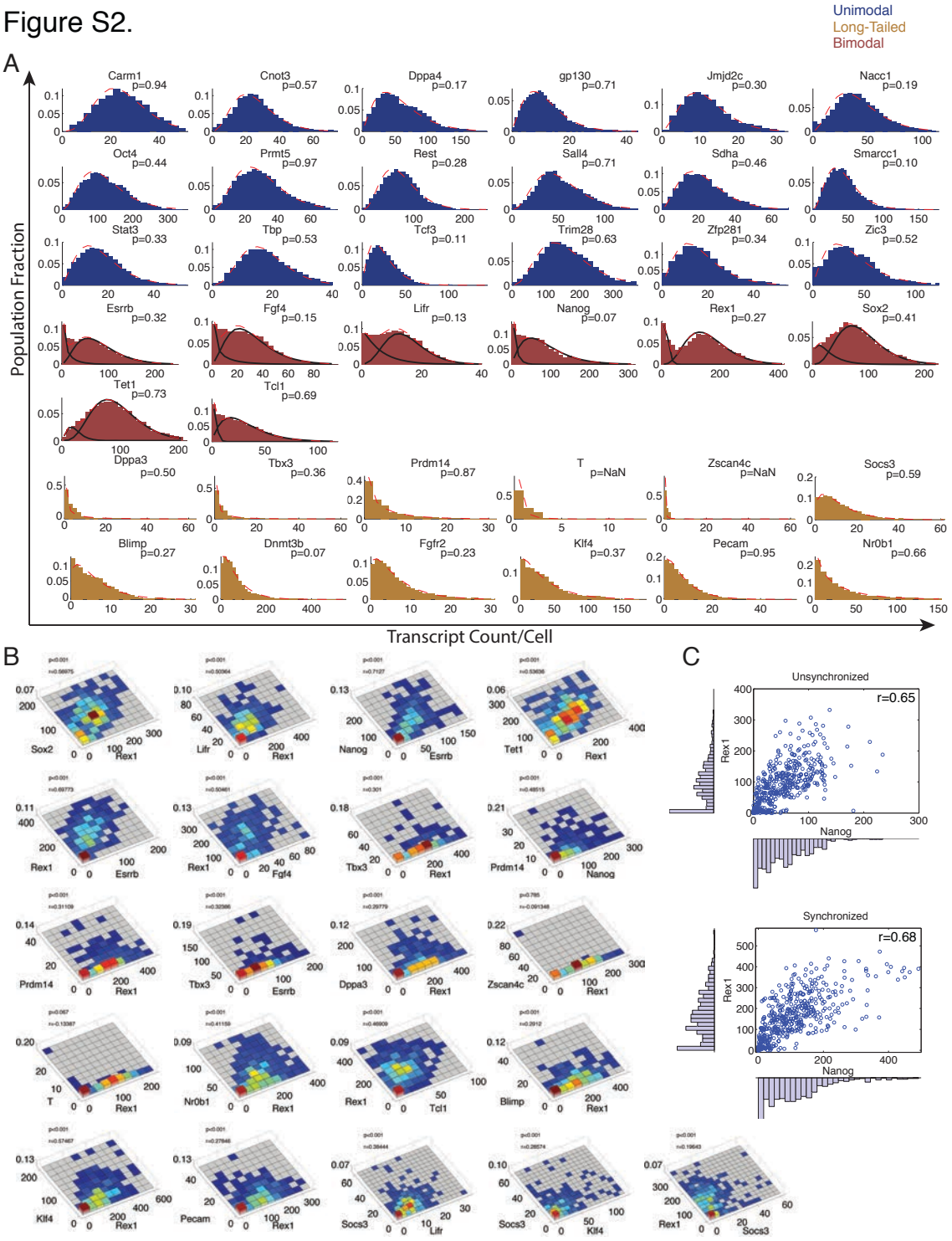

Figure 2.7: mRNA distributions and correlations by smFISH

(A) Empirical distributions and MLE fits for unimodal, bimodal, and long-tailed genes. p-values are for χ^2 GOF tests. p>0.05 indicates that the fit to the distribution is indistinguishable from the empirically measured distribution. Where present, solid lines represent components of the fit. Dashed line represents the overall fit to the distribution. (B) Pairwise relationships between heterogeneously expressed genes. p-values are from the 2D KS-test. r is the Pearson correlation coefficient. (C) Correlation and marginal distributions of Rex1 and Nanog in a control population (top) and population synchronized by a double thymidine block fixed immediately following the block (bottom). r is the Pearson correlation coefficient.

Figure S3

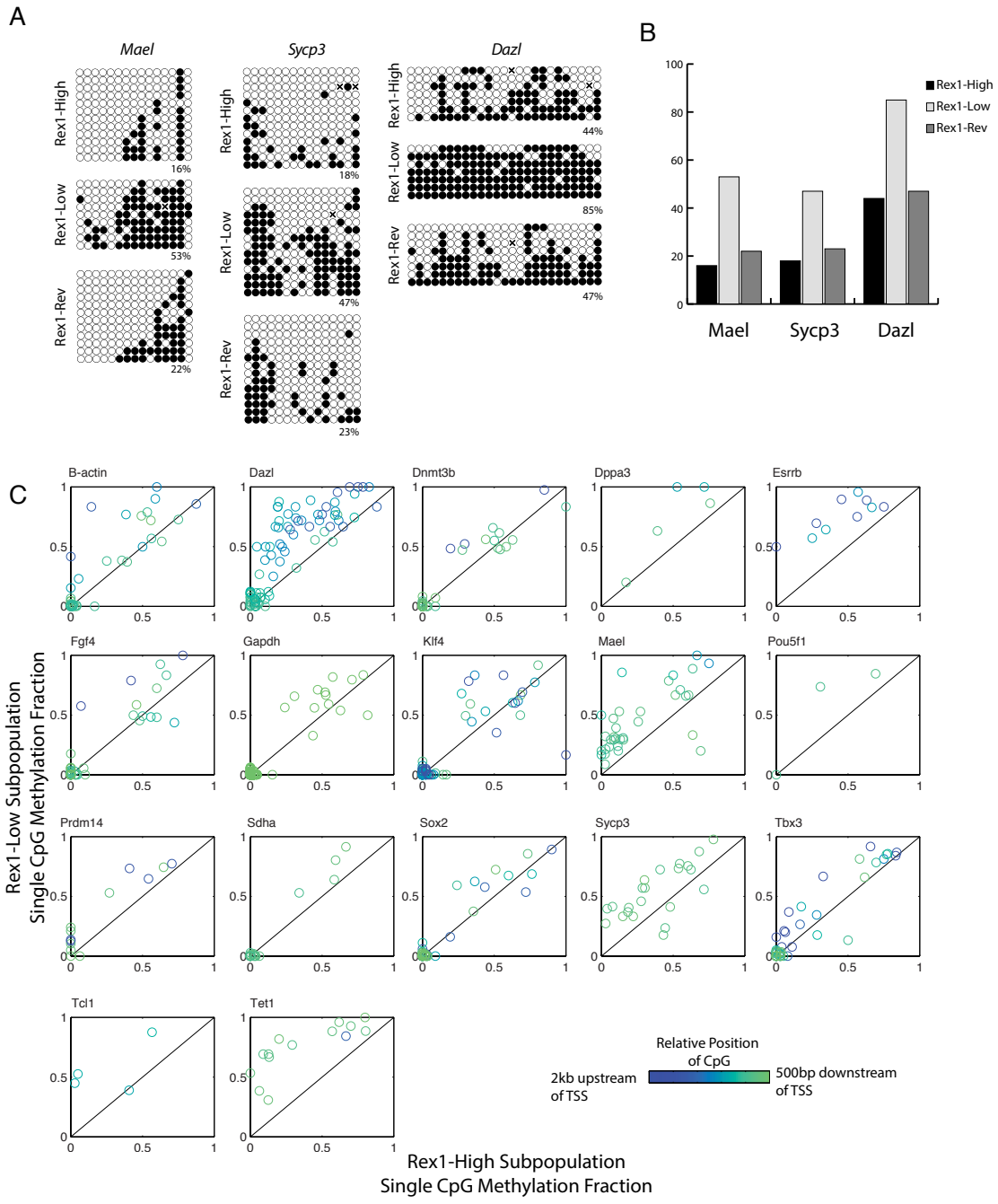

Figure 2.8: Differential methylation between Rex1 states

(A) Locus specific bisulfite sequencing plots between Rex1-high, -low, and -low-to-high-reverting cells at three targets of methylation. Open circles are unmethylated, filled circles are methylated, and x's are unknown. (B) Measurements from A are plotted as bar graphs for comparison. (C) Scatter plots showing how single CpGs in the promoters of a given gene change between Rex1-high and -low states. Color coding represents the position of a base relative the transcriptional state site.

Figure S4.

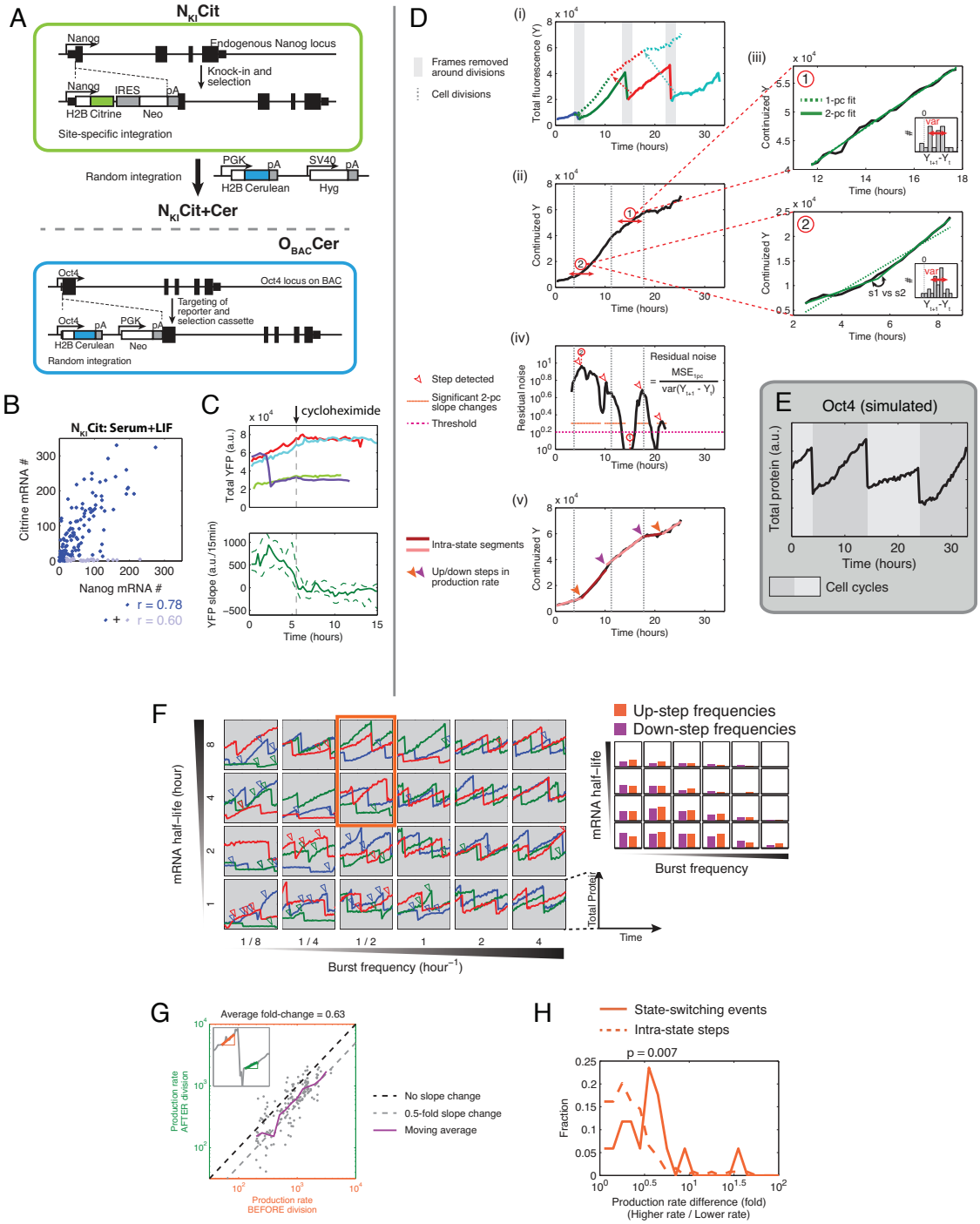

Figure 2.9: Construction and analysis of live cell reporters, and simulations based on observed kinetics

(A) Schematic of Nanog reporter (top) and Oct4 reporter (bottom) construction. (B) Correlation between Nanog (unmodified allele) and Citrine (knock-in reporter on second allele) transcripts in NKICit cell line. r, Pearson correlation coefficient. Light blue, presumed fraction of cells with silenced reporter cassettes (∼10% of all cells; see Supp. Info. for discussion); dark blue, remaining cell population. (C) H2B-Citrine protein degradation rate assayed by blocking translation during movie at time indicated. Total YFP became flat (top) with negligible slope (bottom) shortly after cycloheximide treatment. (D) Identification of sharp inflections in total fluorescence traces. (i) First, frames around cell divisions are removed and fluorescence lost during divisions is added back to the daughter trace to create a continuous trace for each lineage (ii), where a step detector spanning a 6-hour window is applied across consecutive frames. (iii) For each window, a one-piece linear fit is compared with a two-piece fit that is flexible at the midpoint. A two-piece fit is considered better than a one-piece fit when two criteria are met. 1) Residual noise of the one-piece fit is higher than a threshold (see Supp. Info.), and 2) the slopes of the two-piece fit are significantly different between the two pieces. iv) For each stretch of frames meeting both criteria 1 (magenta line indicates threshold) and 2 (orange line indicates where two-piece fit yields significantly different slopes), the window with the highest residual noise is assigned to be the inflection. v) Continuized trace approximated into linear segments between identified points of inflection. (E) Apparent steps from simulated Oct4 expression under the bursty transcription model using parameters estimated from smFISH. (F) Protein traces were simulated under the bursty transcription model over various mRNA half-life and burst frequency combinations; mean burst size was kept constant at 35 mRNA/burst. Gaussian noise proportional to the total protein level and equivalent to the magnitude of frame-to-frame variation empirically observed was added to the simulated traces for comparability. Arrowheads indicate detected steps on simulated trace of the corresponding color. Note that changes in production rate around cell division events can be identified as steps either before or after the division. Red box: Estimated regime for Nanog-Hi in serum+LIF. Right: Variation in the frequency of detected steps over the same parameter space. (G) Production

rates decrease by an average of 0.63-fold across cell divisions. Each point represents a division event. Average production rates of the 4-hour windows before and after each cell division are compared. Black dotted line: zero change; grey dotted line: 0.5-fold change; purple line: average trend; Inset) example trace indicating slope before and after division. (H) Changes in production rate over state-switching events or intra-state steps. 'Higher rate'-to-'lower rate' ratios are plotted for all steps and events, i.e., down-steps and Nanog-high-to-Nanog-low switching events are represented by the reciprocals of rate change. (p-value, KS test)

Figure S5.

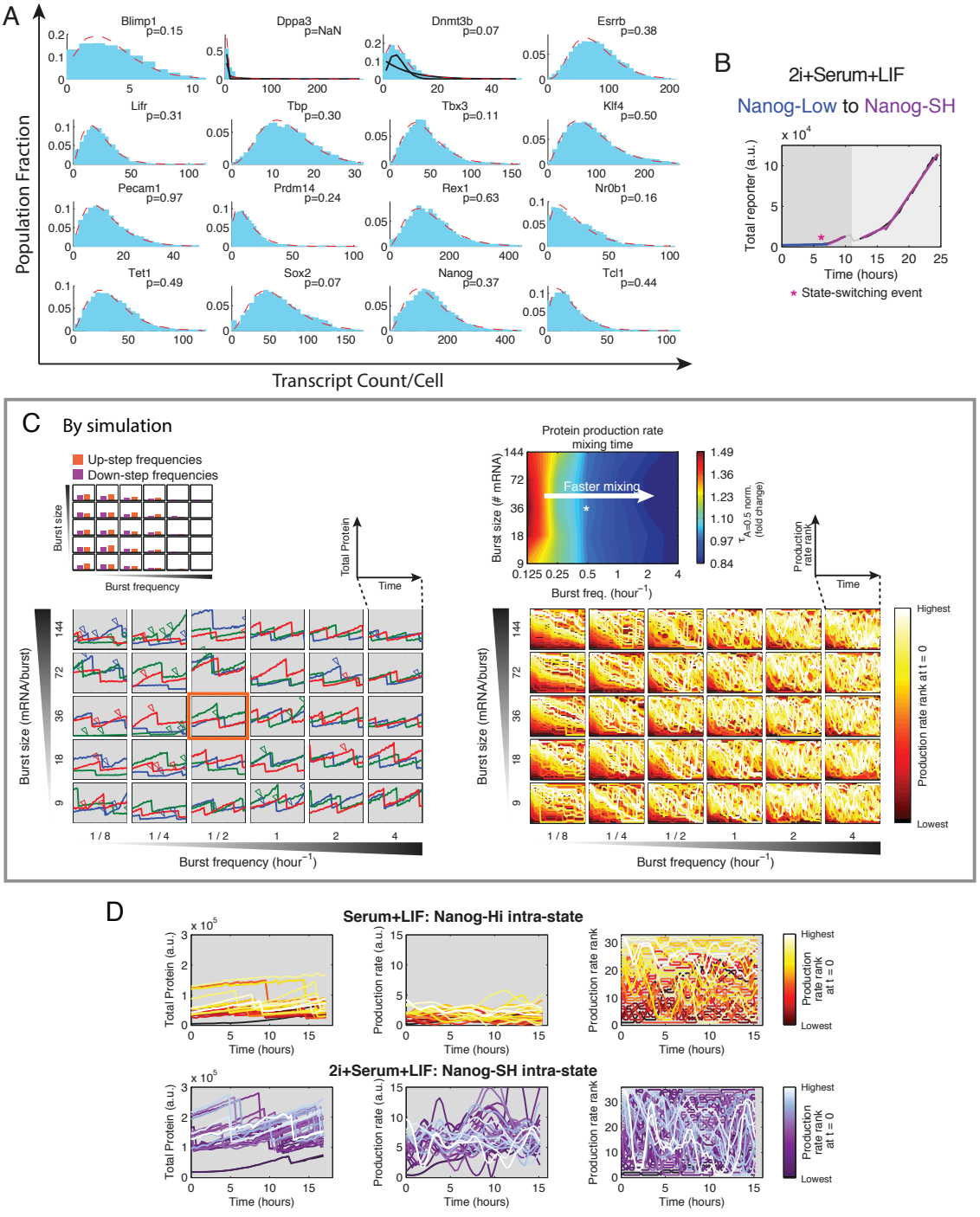

Figure 2.10: Quantitative analysis of how 2i+serum+LIF affect static distributions and dynamics of gene expression for pluripotency regulators

(A) smFISH transcript count distribution of factors in 2i+serum+LIF with MLE fits overlaid. p-values are for χ^2 GOF tests. p>0.05 indicates that the fit to the distribution is indistinguishable from the empirically measured distribution. Where present, solid lines represent components of the fit. Dashed line represents the overall fit to the distribution. (B) Example trace of cells switching from Nanog-low to Nanog-SH in 2i+serum+LIF. (C) Left: simulated traces similar to Fig. 9F, except over various combinations of burst size and burst frequency; mRNA half-life was kept constant at 4 hours. Bottom right: rank of production rate of 30 randomly selected traces (out of a total of 200) in each simulation under the corresponding parameter combination. Traces are color-coded by the initial rank at t = 0 as in D. Top right: mixing time of protein production rate, defined as the time where auto-correlation of rank drops below 0.5. (D) Nanog expression dynamics of cells in serum/LIF with or without 2i. Each trace represents one cell randomly picked from a tracked lineage tree. Production rates are normalized by cell size and ranked within the group for each time point. Traces are color-coded by the initial rank at t = 0.

Direction of switch	Neither sister switched	Only one sister switched	Both sisters switched	Expected number of sister pairs that both switched**
NLo-to-NHi	169	7	7	[0 - 1]
NHi-to-NLo	139	15	2	[0 - 2]

Table 2.1: State-switching events show no correlation between sister cells

Data shown in Table 2.1 are combined results from two independent experiments. Analysis of individual data sets yields the same conclusion. * Data points are discarded if one of the cells in a sister pair was lost or not traceable in the movie ** Confidence interval obtained by random permutation test with 100,000 trials. Green indicates observed frequency of sister pairs in which both cells switched falls within the 95% C.I.

2.10 Supplemental Experimental Procedures

2.10.1 Detailed Culture Conditions

All cells were maintained in humidity-controlled chamber at 37^{o}C, with 5% CO2 in serum+LIF media [Glasgow Minimum Essential Medium (GMEM) supplemented with 10% FBS (HyClone, Thermo Scientific), 2 mM glutamine, 100 units/ml penicillin, 100 μg/ml streptomycin, 1 mM sodium pyruvate, 1000 units/ml Leukemia Inhibitory Factor (LIF, Millipore), 1X Minimum Essential Medium Non-Essential Amino Acids (MEM NEAA, Invitrogen), and 50 uM β-Mercaptoethanol.

2.10.2 Correlation between Citrine and Nanog transcripts in Nanog knock-in reporter cells (NKICit)

We validated the Nanog knock-in reporter by performing smFISH for correlation between Nanog (unmodified allele) and Citrine (knock-in reporter on second allele)

(Fig. 9B). We observed that when grown in serum+LIF conditions, ∼10% of cells contained Nanog but no Citrine transcripts, likely due to silenced expression of their reporter cassettes during prolonged propagation without antibiotics. The remaining cell population showed even stronger correlation between Nanog and Citrine transcripts (r=0.78). We corrected for the potential systematic error that may result in the calculation of low-to-high switching rate such that an observed rate of 1.9±0.29 transitions per 100 cell cycles was adjusted to the reported 2.3±0.25 (mean±SD). We note that the magnitude of this error does not alter key conclusions, including those about the relative stabilities of the two states. Furthermore, the asymmetry of this silencing behavior (we did not find a corresponding fraction of cells expressing Citrine but no Nanog transcripts) suggests that this is unlikely a result of mono-allelic regulation.

2.10.3 smFISH Procedure and imaging system

Up to 48 20mer DNA probes per target mRNA were synthesized and conjugated to Alexa fluorophore 488, 555, 594, or 647 (Life Technologies) and then purified by HPLC. Cells for smFISH experiments were plated at 40,000/cm^2 and harvested after 48 hours. Trypsinized cells were washed in PBS and fixed in 4% formaldehyde at room temperature for 5 mins. Fixed cells were resuspended in 70% ethanol and stored at -20^{o}C overnight. The next day, cells were hybridized with 4nM probe per target species at 30^{o}C, in 20% Formamide, 2X SSC, 0.1g/ml Dextran Sulfate, 1mg/ml E.coli tRNA, 2mM Vanadyl ribonucleoside complex, and 0.1% Tween 20 in nuclease free water. The following morning, cells were washed in 20% Formamide, 2x SSC, and 0.1% Tween 20 at 30^{o}C, followed by two washes in 2x SSC + 0.1% Tween 20 at room temperature. Hybridized cells were placed between #1 coverslips and flattened by applying pressure evenly across the glass.

After flattening cells between coverslips, dots typically span two distinct focal planes. However, to maximize the number of cells imaged in a given acquisition time, only one of these focal planes was captured. This results in approximately 60% of

each cell's transcripts being captured in a single slice, as compared to taking a stack of images across the entire volume of each cell (Fig. 6A-D).

Imaging was performed on a Nikon Ti-E with Perfect Focus, Semrock FISH filtersets, Lumencor Sola illumination, 60x 1.4NA oil objective, and a Coolsnap HQ2 camera. Snapshots were taken using an automated grid-based acquisition system on a motorized ASI MS-2000 stage.

2.10.4 Monte-Carlo Bivariate Kolmogorov-Smirnov Test

The 1D Kolmogorov-Smirnov test was extended to 2D dimensions [1] to determine whether an empirical bivariate distribution showed any dependence between variables; the 2D Cumulative Distribution Function (CDF) is computed in each possible quadrant of the 2D plane $P(x<x_0)$, $P(y<y_0)$; $P(x>x_0)$, $P(y<y_0)$; $P(x<x_0)$, $P(y>y_0)$; and $P(x>x_0)$, $P(y>y_0)$. The 2D KS test statistic is thus defined as the largest difference between empirical and theoretical distributions across each of these possible regions. In order to generate a test-statistic distribution under the null hypothesis, we performed a Monte-Carlo simulation where sets of random pairs of data points are sampled from the PDF formed by the product of the marginal distributions. The resulting bivariate CDF is compared to the theoretical CDF and the maximal difference is taken. This is performed repeatedly in order to generate a distribution of maximal differences that would occur by chance. Finally, the test statistic is computed from the empirical distribution, and compared to this distribution at a 95% confidence level.

2.10.5 Movie acquisition system

Images were acquired on the IX81 inverted microscope system (Olympus) using the Metamorph acquisition software (Molecular Devices) with the iKon Charge Coupled Device (CCD) camera (Andor). Fluorophores were excited using X-Cite XLED1 light source (Lumen Dynamics) equipped with the BLX, BGX, and GYX modules.

2.10.6 Movie data analysis: Segmentation and tracking

The Schnitzcells script package [2] was used to segment and track cells from the acquired images. This package performs a number of procedures as described below. Briefly, cells were segmented with Matlab built-in edge detection script, using Laplacian of Gaussian method. Segmented cells in individual frames were then tracked across all time points by performing a point-matching algorithm on successive pairs of frames to generate a cell lineage data structure. To obtain the total fluorescence level of each cell, the images were "flattened" by correcting for the nonuniformity of illumination, followed by local background correction that takes into account the camera acquisition background, autofluorescence from the medium and fluorescence contribution from neighboring cells.

2.10.7 Movie data analysis: Production rate estimation and step detection

To enable the continuous estimation of production rates (slopes), frames around cell divisions are removed and fluorescence lost during divisions (to sister cells) is added back to the trace of interest to create a continuous total fluorescence trace for each lineage. Instantaneous fluorescence production rates were estimated by fitting the continuous total fluorescence of a 6-hour window to a linear section using the linear least squares method. Distributions of reporter production rates (Figs. 4A,B) were obtained by sampling the instantaneous fluorescence production rates of all cell lineages at 1-hour intervals. To characterize abrupt changes in production rates, we identified sharp inflections of the continuous total fluorescence traces by applying a custom-built step detector on overlapping and consecutive 6-hour windows 15 minutes apart (Fig. 9D). For each window, we obtained fits to a linear polynomial model and a continuous two-piece linear polynomial model with a joint at midpoint using the linear least squares and non-linear least squares methods, respectively. The continuous two-piece linear model can be represented as follows:

$$y = \begin{cases} m_a x + c & , \quad x < x_{mid} \\ m_a x_{mid} + c + m_b(x - x_{mid}) & , \quad x \geq x_{mid} \end{cases}$$

where x_{mid} is the midpoint of the window.

We used two criteria to determine whether a given window fits better to the one-piece or two-piece linear fits: (1) whether the noisiness of the trace can explain the deviation of the data from the one-piece fit (mean sum of squared errors, M.S.E.), and (2) whether the two slopes obtained from the two-piece fit are significantly different from each other. For (1), we define the noisiness of the trace as the variance of the distribution of frame-to-frame fluctuations in total fluorescence, i.e., var(Y_t+1 - Y_t). For a perfectly linear trace without noise, the mean of Y_t+1 - Y_t equals the slope of the trace. As the observation noise increase, the SSE of one-piece fit increases even if the underlying trace has a constant slope. We therefore estimated the portion of SSE of one-piece fit unexplained by the noisiness of the trace as the residual noise, defined as MSE1pc / var(Y_t+1 - Y_t), where n is the number of frames within a window. For (2) we obtained the 95% confidence bounds of the two slopes in the two-piece fit and determined if they overlap. Using (1) and (2), we identified stretches of frames where two-piece fit is significantly better than one-piece. The frame with the highest residual noise among each of these stretches was designated as the point of inflection and the rest of the trace was approximated by linear segments between these points.

2.10.8 Movie data analysis: Hidden Markov Model and Viterbi Algorithm

We set up a two-state HMM to estimate the frequency of state-switching events between the higher and lower Nanog states. We assume each of the two states can produce an independent Gaussian distribution of production rates, with specified mean and variance, including potential overlap between the two states. Over each unit time, a cell can either stay at its current state or switch to the other state with specified probabilities. Thus, given a specific parameter set, there exists for

the production rate time-series of each cell a corresponding series of underlying states that has the maximum likelihood. This likelihood is a balance between the probability of observing a production rate at the corresponding state and that of switching to another state, such that a cell that transiently exhibits a production rate far from the mean of its current state is more likely to be fluctuating rapidly within a state than switching away and back. The Baum-Welch algorithm [3] maximizes the sum of this likelihood over all cells by iteratively changing the parameters in small increments, improving the total likelihood each time.

Prior to training the model with data, initial transition rates between the states in both directions were set at 0.0001/hour. Initial parameters for each state were set with the mean value drawn from the range of observed production rates and variance. Re-initializing the random parameters in the model yielded similar results. We employed the HMM toolbox for Matlab (Murphy, 1998), which generated maximum likelihood estimate of the model parameters using the Baum-Welch algorithm. Since the production rate sequences used to train the HMM contained repeated time-series when multiple lineages shared the same ancestor, the state-transition rates generated directly from HMM could be an overestimation. We applied the Viterbi algorithm [4] to combine the model parameter estimates and the observed data to infer the most likely state sequence for each cell lineage. From this we reported the empirical state-transition rates, normalized to the average length of a cell cycle.

2.10.9 Bursty transcription simulation and mixing time analysis

Bursty transcription was simulated using the model previously described [5]. In this model, a promoter can transit stochastically between an active and an inactive form. This is not to be confused with a cellular state, which is usually maintained over a longer timescale and within which a gene bursts in a characteristic burst size and frequency. Transcription occurs only when the promoter is in its active form, producing a burst of mRNA molecules, which decay exponentially. To aid comparison

between the simulated transcription dynamics and our experimental observations, we added protein production to the simulation. Further, since our fluorescence protein is stable, and to restrict the source of heterogeneity in our simulation to stochastic transcription, we assumed zero protein decay rate and deterministic protein production at a constant rate. Lastly, both mRNA and protein are partitioned when cells divide, which were set to have division rates similar to experimental data. This model can be described by the following reactions:

$$\begin{aligned} \mathrm{I} &\xrightarrow{K_{on}} \mathrm{A} \\ \mathrm{A} &\xrightarrow{K_{off}} \mathrm{I} \\ \mathrm{A} &\xrightarrow{\beta_m} \mathrm{A+m} \\ \mathrm{m} &\xrightarrow{\alpha_m} \phi \\ \mathrm{m} &\xrightarrow{\beta_p} \mathrm{m+p} \end{aligned}$$

Here, A and I denote the promoter in its active and inactive forms, respectively; m - mRNA level; p - protein level; k_{on} and k_{off} - activating and inactivating rates of the promoter, respectively; α_m - mRNA degradation rate; β_m - mRNA production rate; ϕ - mRNA degradation; β_p - protein production rate.

A cellular state is thus characterized by the frequency of mRNA bursts and the mean number of mRNA molecules produced per bursts. Here, we considered one limiting case of this model, where k_{off} is significantly larger than k_{on} and somewhat larger than α_m. This assumption can be related physically to a scenario where bursts are relatively infrequent and have short durations, and the distribution of mRNA levels produced under these assumptions can be described with a single gamma (Raj et al., 2006) or NB function [6]. A cell changes state in a gene when one or more of the parameters k_{on}, k_{off}, or β_m for that gene is changed, thus resulting in different burst frequencies and sizes.

To simulate mRNA and protein dynamics for the Nanog-high state in serum+LIF condition (shown in Figs. 4E, 9E), we used the following parameters estimated from mRNA distributions in smFISH: for Nanog – burst size = 33 mRNA/hour, burst frequency = 0.39 bursts/hour; for Oct4 – burst size = 87 mRNA/hour, burst frequency

= 0.52 bursts/hour. These assume that mRNA half-lives of Nanog and Oct4 are 5.85 and 7.4 hours, respectively [7].

We utilized computer simulations of this model to explore whether changes in state affect the intra-state dynamics of heterogeneity. Varying burst frequency and burst size results in traces with various frequencies of apparent steps when analyzed using the same step detector, which identified regimes in the bursty transcription parameter space where steps of similar quality to the ones observed can be generated (Figs. 9F, 10C). Furthermore, the resulting dynamics also display a wide range of shapes of fluctuation and levels of expression. We quantified these variations with an objective measurement, the "mixing time", a population metric adapted from Sigal et al. [8]. For each simulated population (n = 200 traces) using a single parameter set, we ranked all traces by their production rate at each time point. Thus, a cell starting with the lowest production rate among the population may change in this rank when its production rate changes over time. We computed the autocorrelation function $A(\tau)$ of this rank for each population and the mixing time is defined as the time lag τ at which $A(\tau)$ decayed to 0.5. We opted to calculate the mixing time using production rate but not total fluorescence level because the stable fluorescent reporter facilitates accurate production rate estimate but may not reflect the physiological level of endogenous proteins. Additionally, for more direct comparison between the mixing times calculated from simulated and observed data, the production rates in simulation were computed using the simulated protein traces after Gaussian noise similar to the level observed was added.

Supplemental References

1. Peacock, J. A. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society* (1983) (cit. on p. 75).
2. Young, J. W. *et al.* Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols* (2011) (cit. on p. 76).
3. Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nat Biotechnol* (2008) (cit. on p. 78).
4. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* (1989) (cit. on p. 78).
5. Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theoretical Population Biology* (1995) (cit. on p. 78).
6. Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* (2000) (cit. on p. 79).
7. Sharova, L. *et al.* Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA research* (2009) (cit. on p. 80).
8. Sigal, A. *et al.* Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nat Meth* (2006) (cit. on p. 80).

Murphy, K.P. (1998). Hidden Markov Model (HMM) Toolbox for Matlab. Retrieved Dec 6, 2012, from http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html.

Chapter 3

Lineage-based inference of dynamics and network architecture from static single cell measurements

3.1 Abstract

Elucidating the dynamics of cellular decision-making is critical to understanding processes ranging from organismal development to immunology to cancer proliferation. As a result, knowing the frequency, nature, and extent of transitions among cellular states can inform both clinical and fundamental biology. Here, we demonstrate the ability to infer transition rates and network topology connecting expression states across many generations, by combining cell-phylogeny with multi-dimensional endpoint smFISH measurements. Overlaying static gene-expression data on lineage trees *in situ* provides a powerful, scalable platform to identify novel, high-dimensional dynamics in systems ranging from cell culture to model organisms.

3.2 Introduction

Identifying the dynamics of gene regulatory networks underlying single cell decision-making is a central problem in systems biology ranging from development to adult tissue maintenance and to cancer homeostasis. Thus, understanding the states that

cells exist in and the transition rates between them in natural undisturbed tissue would lead to significant insights in a range of contexts.

Cells exhibit dynamic transitions among distinct states, characterized in part by expression levels of one or more genes. In some systems, cells can remain in a state for multiple cell cycles before transitioning to another state. For these systems, one would like to know what transitions are permitted, how frequently they occur, and whether there are specific temporal sequences. Knowing this information would provide fundamental insights into important processes such as phenotypic switching in cancer cells and cell type decision-making processes in stem cells [1–3].

Despite the many systems of interest that exhibit these behaviors, there is no general method to measure these dynamics. While existing techniques have provided critical insights, each have inherent limitations. These include bulk and single-cell qPCR and RNA-seq using sophisticated computational algorithms [4], FACSed subpopulation re-equilibration, *in vivo* and *in vitro* time-lapse microscopy for lineage tracking or direct gene expression measurements [5–11], and spontaneous-mutation-based lineage reconstruction [12]. However, while each has been valuable, they all present their own individual challenges, such as limited temporal, spatial, or breadth of target resolution; *a priori* knowledge; perturbations from natural contexts or expression level; or averaging over populations.

Here we describe a new approach that circumvents aforementioned difficulties and provides the capability of inferring temporal event sequences, quantitative switching rates, and network topology. We use a combination of time-lapse movies followed by multiplexed end-point single-cell measurements [13–15] in the spatially unperturbed sample. Generally, these can encompass smRNA-FISH, immunostaining, morphological parameters, or any imaging-based *in situ* single cell readout. We demonstrate the ability to infer transition rates and network topology connecting expression states across many generations [16], by combining cell-phylogeny with multi-dimensional endpoint smFISH measurements. Overlaying static gene-expression data on lineage trees in situ provides a powerful, scalable platform to identify novel, high-dimensional dynamics in systems ranging from cell culture to model organisms.

Quantitatively measuring distributions of gene expression can yield insight into the nature of the underlying heterogeneity. Simply adding on top of these measurements the relatedness of those cells can reveal the dynamics of the underlying distribution. The dynamics of state transitions captured during growth and division would thus confer a signature as clusters among the leaves of the trees (Fig 1A). Consider a case where cells switch between two states. In the first case, imagine the situation where transition events occur frequently; as the lineage grows, closely related cells may quickly adopt independent states, leading to little or no clustering (Fig 1A, left). In an opposing scenario, imagine that the probability of staying in a specific state is much greater than transitioning out of it. This would produce a more persistent phenotype, and as such, would lead to stronger clustering of states among the offspring (Fig 1A, right). Notably, the final distribution among the leaves of both trees are identical (1/2 low, 1/2 high), but the differences in the underlying transition probabilities yield markedly distinct dynamics that are measurable once the relatedness of each cell is incorporated into the measurement.

In order to test the system, we applied the technique to mouse ES cells that display metastability and heritability over multiple cell cycles under standard, unperturbed growth conditions. While many transcription factors have been identified as key regulators of pluripotency or fate-choice, it remains unclear how the underlying gene regulatory networks gives rise to the observed dynamics, and what trajectories cells take. Thus, even without requiring external perturbations, a rich set of dynamical transition is observed between these states. Specifically, we explore the dynamical relationships between three ES genes previously identified as heterogeneous [8, 17, 18]. We demonstrate the utility of such a system by coupling movies with multiplexed endpoint single-cell RNA-FISH in order to (1) infer and additionally validate through a knock-in reporter directly the stochastic transitions of pluripotency marker Esrrb; and (2) extend the analysis to infer the transcriptional trajectory across a set of transcriptional states defined by observed occupancy of Esrrb, Tbx3, and Zscan4c.

3.3 Results

3.3.1 The Framework

First, we explain how to infer transition rates between distinct cell states from measured lineage trees and end state gene expression data. To begin, we define the lineage distance u between two cells at the final time point as the number of generations to their common ancestor, as shown in Figure 2A. Thus, $u = 1$ for sisters, $u = 2$ for first cousins, etc. We assume that cell state transition dynamics can be represented as a Markovian memory-less process; that is, the rate of their next stochastic transition is controlled only by their current state, and are constant through time. They can be represented as a transition rate matrix T, whose elements $T(I|M)$ represent the probability of observing a daughter cell in state I given that its parent cell was in state M (Figure 2B, top left). Diagonal elements then represent the probability of a daughter remaining in the same state as her parent. With these definitions, the columns of T sum to one. For simplicity in this figure, we assume that all states are equally likely. We also focus on the symmetric case ('detailed balance') in which $T(I|M) = T(M|I)$. However, this assumption is not strictly necessary, and the framework can be adapted for situations that do not behave like this.

Our first step is to compute a multi-generational transition matrix. The probability of observing a cell in state I conditional on state M of its ancestor u generations back can be obtained by successively applying matrix T u times: $T_u(I|M)$, where u denotes matrix T taken to the power u.

In order to connect the dynamic transitions to the observed frequencies of each state, we compute the correlation matrix, $C_{IJ}(u)$, defined as the probability of finding a pair of cells separated by lineage distance u in states I and J (Figure 2B, top right). If T is known, then $C_{IJ}(u)$ can be computed directly:

$$C_{ij}(u) = \frac{1}{N}\sum_M T_u(I|M)T_u(J|M) = \frac{1}{N}\sum_M T_u(I|M)T_u(M|J) = \frac{1}{N}T^{2u}(I|J)$$

where the summation is over all possible states M of the ancestor, which, by as-

sumption, occur with probability $\frac{1}{N}$. The correlation function $C(u)$ is a set of NxN matrices, one for each value of u. Because any pair must be in one of theN^2 possible pairs of states, the elements of C must sum to one.

Experimentally, $C(u)$ can be read off from lineage-associated end state measurements by counting occurrence of all pairs of states on the tree at lineage distance u. The rate of decay of these correlation functions should be undone by compensating for successive applications of a fixed transition rate, enabling an estimate of the transition rates between states from the correlation functions alone. To infer the dynamics from measured correlation functions, we work backwards to recover T by computing $T_{inferred}(u) := C(u)^{1/(2u)}$ (Figure 2B). If the assumptions are correct, $T_{inferred}(u)$ should match the actual T, and be independent of u. Conversely, dependence of $T_{inferred}$ on u indicates that some underlying assumptions are incorrect.

3.3.2 Transition rates between states of Esrrb can be inferred using this method

To test this framework, we first turned to the dynamics of a mouse stem cell pluripotency regulator Esrrb, which has been shown to play a critical role in the core pluripotency network [19], facilitate fibroblast reprogramming into iPS [20], and be central to maintaining the naive pluripotent state [21], in addition to displaying a heterogeneous expression pattern in stem cell culture [8]. To validate our inference method, we set out to measure both directly and through inference the dynamics of Esrrb as it moves between its two expression states. A non-perturbing knock-in reporter for Esrrb was constructed by incorporating a self-cleaving peptide followed by a Histone2B-mCitrine after the last exon of Esrrb, without affecting its 3' UTR.

We followed single cells using quantitative fluorescence timelapse microscopy (Fig 3A) to directly measure the frequency at which cells switch states, as well as to construct lineage trees. When the movie ended, the cells were fixed and hybridized *in situ* with smFISH probes for the Esrrb transcript as well as the housekeeping control gene β-actin. Cells and transcripts were segmented in 3D with the aid of a membrane-

targeted fluorescent protein, enabling accurate and direct measurement of cytoplasmic transcripts without disturbing the sample (Fig 3A). An example lineage tree overlaid with extracted movie-fluorescence and state-assignment (hi/lo) confidence derived from smFISH counts are shown in Fig 3B.

As a first validation, we checked to see if the rate of change in total fluorescence of Esrrb in movies (hereafter referred to as promoter activity) across the final cell cycle correlated with the transcript counts measured at the movie's endpoint. Indeed, a strong correlation was observed between Esrrb promoter activity with Esrrb transcript counts, but not with β-actin transcript counts (Figure 5), suggesting internal consistency between these measurement modalities. Furthermore, the endpoint Esrrb smFISH marginal distribution across all positions appeared bimodal (Figure 3A) as previously described [8].

Using the correlation functions compiled from 14 trees (299 cells), we estimated transition rates that were constant over time, demonstrating self-consistency and Markovian transitions. Specifically, we measured the stability of the low sate to be 89±6% and the hi state as 93±3% (Figure 3B). Similarly, in the movies, we observed cells 90±1% to stay in the low state, (Fig 3C) and 92±1% staying in the high state. Together, these data show that the inference method yielded the same Markovian dynamics as those measured directly using the more laborious and invasive knock-in reporter.

3.3.3 Inferring trajectories among higher dimensional transcriptional states

Having validated the inference approach, we next turned to a set of genes that have previously been characterized as heterogeneous [8, 17, 22], but whose state transitions and temporal persistence remain unknown. In addition to Esrrb, Tbx3 is biologically important for the maintenance of pluripotency in ES cells as overexpression or underexpression both destabilize the pluripotent state [23–27], and also shows a long-tail distributions in ES cells yielding the highest CV among ES regulators [8]. Second,

Zscan4c displays long-tailed heterogeneity and is associated with totipotency, or the ability to generate all three germ layers, as well as extra-embryonic lineages [8, 17, 28]. These tailed distributions are consistent either with burst-like mRNA production [29] or with Markovian two-state switching, among other models. Moreover, regardless of the type of dynamics, the timescales for activating, de-activating, and remembering a state are not known.

Tracking single cells growing into colonies using fluorescent protein expression like that shown in Figure 4Ai, we collected data from 32 trees (446 cells) followed by quantitative endpoint smFISH for Tbx3, Esrrb, and Zscan4c (Figure 4Aii). We then combine these trees with single cell transcript counts as measured at the leaves (Figure 4C). Immediately, we observe both entire and partial colonies of cells that appeared to express Tbx3 or Zscan4c, as opposed to a salt-and-pepper pattern in every colony. This result largely rules out the stochastic model, where all cells have some low probability of expression, thus showing the importance of performing experiments *in situ*.

Using the population scatter data (Fig 4D), we observe that in the Esrrb+ state, a subset of cells express Tbx3, while in Esrrb- state, a subset express Zscan4c, thus yielding four main states. Additionally, a minority of cells appear to be in a state between these others. Thus, after classifying each cell into one of the resulting five states, we can proceed to construct correlation functions. We assume a five-state Markovian process (Figure 4E) where all transitions are allowed, and count the frequency with which we observe pairs of sisters ($u = 1$), cousins ($u = 2$) etc., that occupy all possible combinations of states (Figure 4F), yielding our correlation functions. By compensating for the dilutive affect over time (See Figure 2), we use this data to compute the transition rates among these classes. Indeed, Figure 4G shows flat lines representing the probability that staying within each of these states over time is constant, validating the models assumptions. Additionally, some of the transition rates appear to drop out of the system due to the lack of observed pairs in those states. This suggests that cells move largely along a linear chain of states (Figure 4H), where cells tend to generally make a subset of the possible transitions, as opposed to the starting model

which allowed any-to-any connectivity (Figure 4E). This does not strictly rule out these other transition events, but suggests that in general, cells are much more likely to follow the linear chain trajectory between opposing ends of the network. Notably, these data suggest that in order to access the Esrrb+/Tbx3+/Zscan4c+ state, a cell must first pass through the Esrrb+/Tbx3+ state. An inspection of the 3D scatter data alone would incorrectly lead one to assume that the Zscan4+ state emerges from the persistent Esrrb- state, showing the utility of applying lineage information.

3.4 Discussion

Here we have demonstrated and validated directly the ability to extract the rates of a two-state Markovian process, as well as the likely trajectory of cells switching between a set of ESC states. As an inference, this data makes a set of predictions that can and should be tested to both validate the model, and extend the utility of the approach; specifically, the qualitative, highly non-trivial prediction would be that Zscan4c+ is most readily accessed from the Esrrb+/Tbx3+ state. Furthermore, these two states may also be more epigenetically similar, as both of these are likely to have reduced levels of DNA methylation [8, 30]. This would also be biologically plausible given evidence from the literature connecting Tbx3 to Zscan4c [26]. Intuitively, the linear path of transitions as inferred allow a cell to move from least pluripotent to totipotent.

Using only cell lineage information and a single snapshot of gene expression data, we can determine rates of transitions and cellular trajectories. By applying novel methods to expand the number of mRNAs assayed [31, 32], this platform presents a powerful way to identify the dynamics among gene expression states.

3.5 Figures

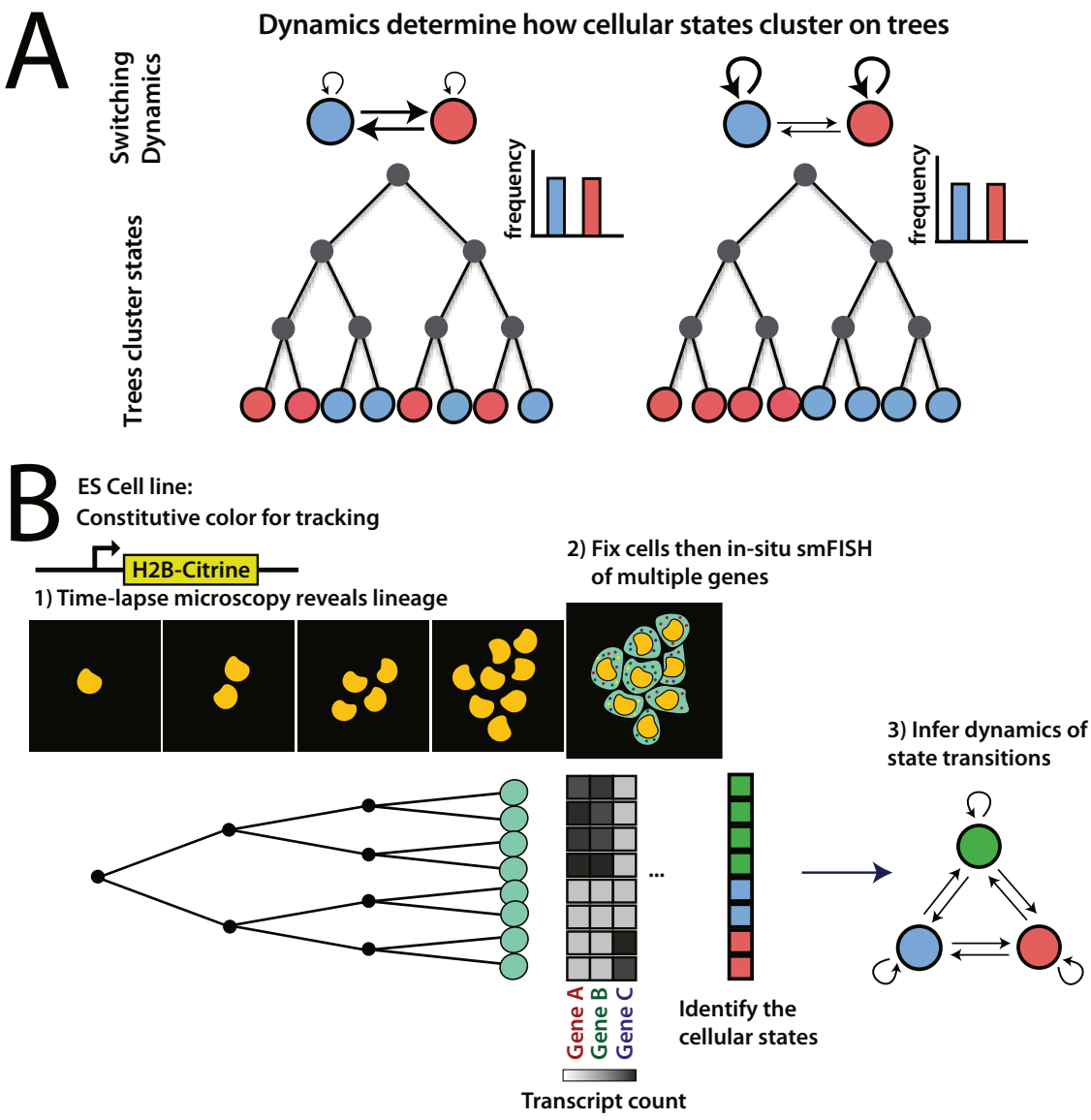

Figure 3.1: Cartoon example of dynamics on trees, and how the method is applied in order to measure the rates

(A) Examples of switching between two states differing only by the rate of switching. Notice the strong clustering of red and blue leaves on the right, while this is effect is washed out on the left due to the faster mixing. (B) A schematic of the experimental technique. 1) Live cells are first tracked using a fluorescent protein reporter, and

are then fixed, stained, and imaged by smFISH. The lineage tree is reconstructed directly from observing cell division events depicted underneath each movie frame. Transcript counts are then used to classify cellular states. These statistics are used in correlation functions to identify transition rates using an assumed Markovian model of the network.

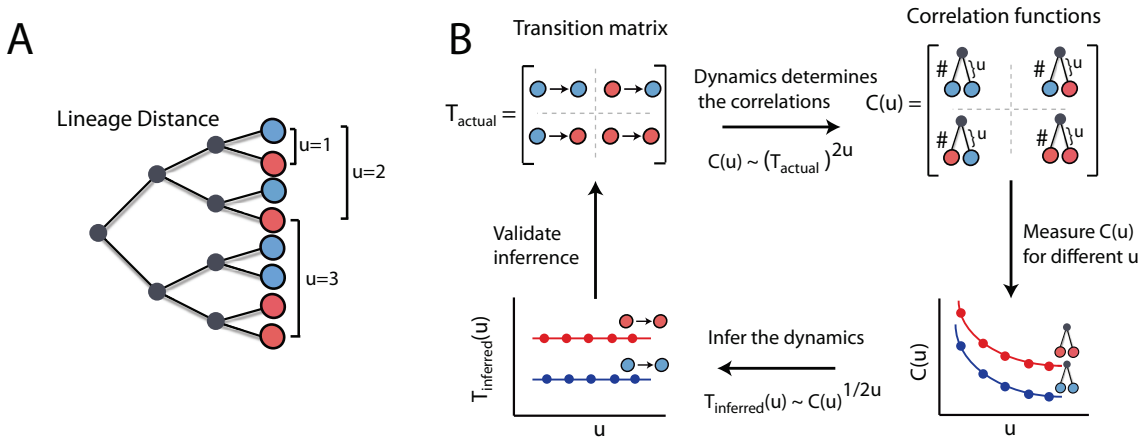

Figure 3.2: Visual depiction of mathematical workflow for rate inference

(A) Schematic showing an example tree following three divisions. Leaves are colored red or blue to indicate state assignment. Brackets betweens two cells demonstrate the definition of pairwise relatedness u, where sisters are $u = 1$, etc. (B) The transition matrix T is broken into four entries, each defining a particular transition; for example, the top left is the probability of staying in the blue state. Columns must sum to one. Below the matrix is an example plotting the blue-to-blue and red-to-red rate as a function of relatedness. If the Markovian model explains the data, these should appear flat over time. Experimentally, we measure the right half of panel B; $C(u)$ is the number of pairs of cells at a fixed u in that set of observed states. This is expected to decay as cells become less related and have had more time to transition out of their starting state. $C(u)$ and $T(u)$ are related through taking T to the power of 2u, where intuitively you are applying some fixed rate of transition u times, one for each cellular division.

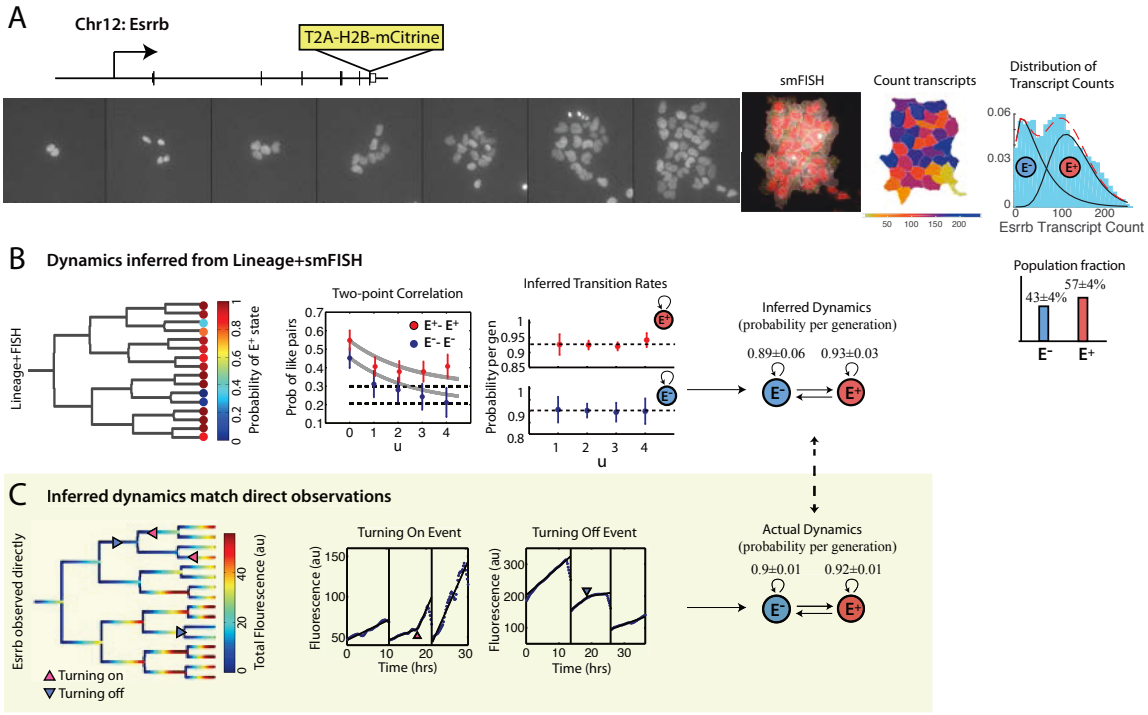

Figure 3.3: Inferred transition rates between Esrrb states match those measured from a direct reporter, and are well modeled as a two-state Markov process

(A) Left: a schematic of the Esrrb-H2B-mCitrine knockin reporter above an example filmstrip of mCitrine levels. The filmstrip is followed by a composite image of the membrane-mTurquoise (white), DAPI (red), and Esrrb transcripts by smFISH (cyan dots). Note the off cells in the lower right corner. Dots are detected and attributed to cells in 3D, with the resulting Esrrb transcript count mask displayed. Combining cells from all colonies analyzed yields a bimodal distribution of Esrrb with the indicated population fractions. (B) The lineage tree resulting from one of the two starting cells in the example movie shown in A is shown in blue; leaves of the tree are colored by confidence of cells in the Esrrb Hi state due to overlap of modes. Correlation functions are plotted for pairs of cells in the on/on state and off/off state. These correlation functions are transformed into transition rates by taking the appropriate roots of the correlation function. Note that the rates appear constant (within error bars) over time The two state model with resulting rates as inferred using correlation functions is shown at the right. Error bars are standard deviation from bootstrapping over individual colonies. (C) Analogous measurements to B, but as performed using direct readouts of Esrrb promoter activity. The tree itself is colored to show total fluorescence over time. Arrows indicate where on-to-off or off-to-on transition events are detected on the tree. Middle) Example traces of an on- and off- transition event. Right) The two state model with rates measured directly from the knock-in reporter. Note that rates estimated from both methods fall within each others error, showing excellent corroboration.

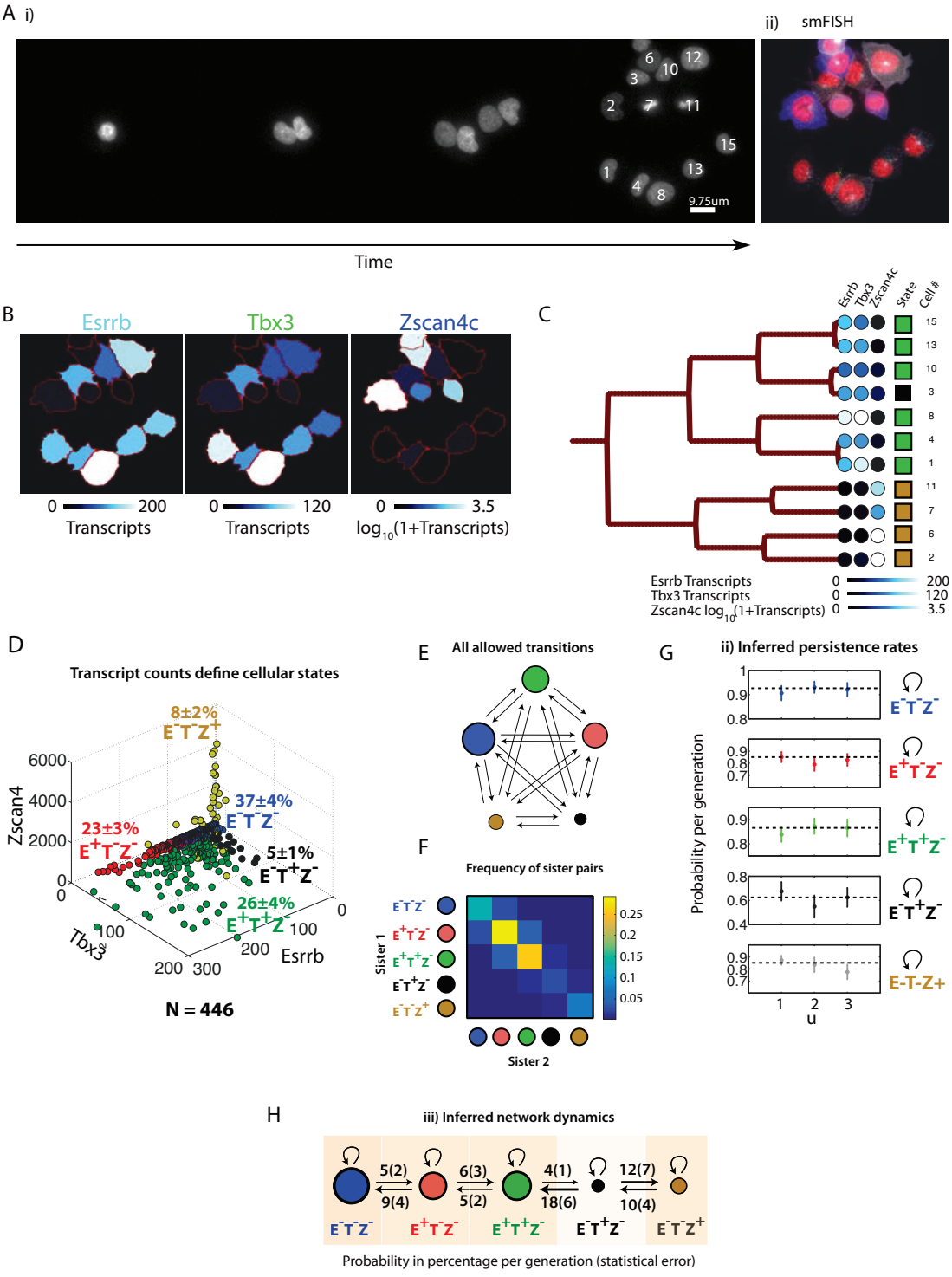

Figure 3.4: Inferring the topology of a complex ES network

(A) i) An example movie with fluorescence used only for tracking cells as they divide. The last frame is overlaid with cell number as used in C. ii) Multiplexed smFISH following the movie. White is membrane-mTurquoise2, red is DAPI, blue dots are Esrrb-smFISH, green dots are Tbx3-smFISH, and cyan dots are Zscan4c-smFISH. B) Masks of transcript counts per cell for each gene individually. Shading represents transcript abundance where the dynamic range is indicated under each frame (C) Lineage tree constructed from cells tracked in A.i, with transcript counts for each cell indicated on the leaves; the colormap for each gene is the same as B. (D) A 3D scatter plot of transcript counts, where each dot is a cell. Color coding indicates state (E) The full matrix of possible transitions between the observed states; the size of each node is proportional to the fraction of cells in that state (F) To compute correlation functions, we first measure the frequency with which we observe pairs of cells in a given combination of states. For this example, we show pairs of sisters, or $u = 1$. Each element is the fraction of all pairs caught in the state defined by the row and column. (G) Stability of each state as a function of lineage distance u. (H) Five-state Markov model with rates for transitions between each state

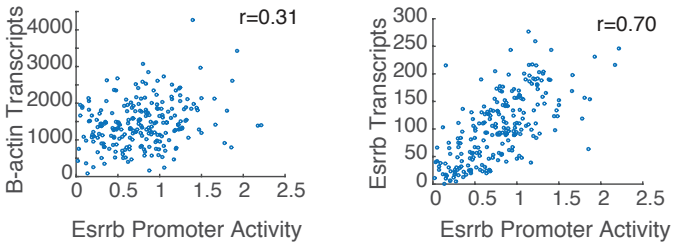

Figure 3.5: Esrrb Knock-In Reporter Validation

Primary References

1. Gupta, P. B. *et al.* Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. *Cell* (2011) (cit. on p. 83).
2. Leder, K. *et al.* Mathematical modeling of PDGF-driven glioblastoma reveals optimized radiation dosing schedules. *Cell* (2014) (cit. on p. 83).
3. Lander, A. D., Gokoffski, K. K., Wan, F. Y. M., Nie, Q. & Calof, A. L. Cell lineages and the logic of proliferative control. *PLoS Biol* (2009) (cit. on p. 83).
4. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* (2014) (cit. on p. 83).
5. Rompolas, P., Mesa, K. R. & Greco, V. Spatial organization within a niche as a determinant of stem-cell fate. *Nature* (2013) (cit. on p. 83).
6. Clayton, E. *et al.* A single type of progenitor cell maintains normal epidermis. *Nature* (2007) (cit. on p. 83).
7. Bothma, J. P. *et al.* Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 10598–10603 (July 2014) (cit. on p. 83).
8. Singer, Z. S. *et al.* Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol Cell* (2014) (cit. on pp. 83, 84, 86–89).
9. Locke, J. C. W., Locke, J. C. W., Elowitz, M. B. & Elowitz, M. B. Using movies to analyse gene circuit dynamics in single cells. *Nature Reviews Microbiology* (2009) (cit. on p. 83).
10. Locke, J. & Elowitz, M. B. Using movies to analyse gene circuit dynamics in single cells. *Nature Reviews Microbiology* (2009) (cit. on p. 83).
11. Levine, J. H., Lin, Y. & Elowitz, M. B. Functional roles of pulsing in genetic circuits. *Science (New York, NY)* **342,** 1193–1200 (Dec. 2013) (cit. on p. 83).
12. Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85,** 49–59 (Jan. 2015) (cit. on p. 83).
13. Albeck, J. G., Mills, G. B. & Brugge, J. S. Frequency-Modulated Pulses of ERK Activity Transmit Quantitative Proliferation Signals. *Mol Cell* (2012) (cit. on p. 83).

14. Lee, R. E. C., Walker, S. R., Savery, K., Frank, D. A. & Gaudet, S. Fold Change of Nuclear NF-κB Determines TNF-Induced Transcription in Single Cells. *Mol Cell* (2014) (cit. on p. 83).

15. Kellogg, R. A. & Tay, S. Noise Facilitates Transcriptional Control under Dynamic Inputs. *Cell* (2015) (cit. on p. 83).

16. Hormoz, S., Desprat, N. & Shraiman, B. I. Inferring epigenetic dynamics from kin correlations. *Proceedings of the National Academy of Sciences of the United States of America* (Apr. 2015) (cit. on p. 83).

17. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* (2012) (cit. on pp. 84, 87, 88).

18. Faddah, D. A. *et al.* Single-Cell Analysis Reveals that Expression of Nanog Is Biallelic and Equally Variable as that of Other Pluripotency Factors in Mouse ESCs. *Cell Stem Cell* (2013) (cit. on p. 84).

19. Festuccia, N. *et al.* Esrrb Is a Direct Nanog Target Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell Stem Cell* (2012) (cit. on p. 86).

20. Jiang, J. *et al.* Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* (2009) (cit. on p. 86).

21. Martello, G. *et al.* Esrrb Is a Pivotal Target of the Gsk3/Tcf3 Axis Regulating Embryonic Stem Cell Self-Renewal. *Cell Stem Cell* (2012) (cit. on p. 86).

22. Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* (2008) (cit. on p. 87).

23. Han, J. *et al.* Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature* (2010) (cit. on p. 87).

24. Weidgang, C. E. *et al.* TBX3 Directs Cell-Fate Decision toward Mesendoderm. *Stem Cell Reports* (2013) (cit. on p. 87).

25. Lu, R., Yang, A. & Jin, Y. Dual functions of T-box 3 (Tbx3) in the control of self-renewal and extraembryonic endoderm differentiation in mouse embryonic stem cells. *Journal of Biological Chemistry* (2011) (cit. on p. 87).

26. Dan, J. *et al.* Roles for Tbx3 in regulation of two-cell state and telomere elongation in mouse ES cells. *Sci Rep* (2013) (cit. on pp. 87, 89).

27. Niwa, H., Shimosato, D. & Adachi, K. A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature* (2009) (cit. on p. 87).

28. Carter, M. G. *et al.* An in situ hybridization-based screen for heterogeneously expressed genes in mouse ES cells. *Gene Expr. Patterns* (2008) (cit. on p. 88).

29. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biol* (2006) (cit. on p. 88).

30. Dan, J. *et al.* Rif1 maintains telomere length homeostasis of ESCs by mediating heterochromatin silencing. *Developmental Cell* **29,** 7–19 (Apr. 2014) (cit. on p. 89).

31. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat Meth* (2014) (cit. on p. 89).

32. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Meth* (2012) (cit. on p. 89).