12-14-2015

Single Cell Resolution Mapping of mRNA and Protein Expression Dynamics During Human Somatic Cell Reprogramming to Pluripotency

Frederick W. Kolling IV
University of Connecticut, fkollingiv@gmail.com

Single Cell Resolution Mapping of mRNA and Protein Expression Dynamics in Human Somatic Cell Reprogramming to Pluripotency

Frederick W. Kolling IV, PhD

University of Connecticut, 2015

In recent years, the fields of regenerative medicine and developmental biology have been revolutionized by the ability to reprogram adult somatic cells back to a pluripotent state, producing induced pluripotent stem cells (iPSCs) through the induction of the reprogramming factors OCT4, SOX2, KLF4 and c-MYC (OSKM). The promise of this technique has been somewhat limited by the low efficiency and long duration of the process which produces heterogeneous mixtures of cells in culture, many of which fail to fully reprogram over a 3-4 week period. Research in recent years has focused on the underlying mechanisms of reprogramming and identifying rate-limiting steps in the process. Most studies to date utilize bulk measurements of cells undergoing reprogramming however, these techniques cannot measure the changes occurring in rare cells that will become iPSCs and are inherently biased towards unsuccessful reprogramming events. Here we apply single cell technologies to measure the dynamics of mRNA and protein expression and develop mathematical models to precisely describe these behaviors. We find that productively reprogramming cells activate genes in an ordered, probabilistic fashion but do so independently of one another, lacking hallmarks of gene regulatory network activity. Some genes, despite their expression as mRNAs, are not immediately translated into protein, identifying post-transcriptional mechanisms as a potential rate limiting step. In contrast, cells moving along an alternate trajectory away from fibroblast but not towards iPSC, fail to activate pluripotency genes and do not express the full complement of OSKM. This is due to premature inactivation of the individual factors, causing cells to drop off the productive trajectory and fail to reprogram. Performing these analyses under two different delivery methods of OSKM and in two cell types reveals that while the timing of gene activation varies between conditions, the probabilistic order of gene activation is conserved, suggesting a common reprogramming trajectory. Taken together these findings represent the first descriptions of

mRNA and protein expression dynamics in single, human cells undergoing reprogramming. We also provide a robust mathematical framework for identifying rate-limiting steps in the process and dissecting the mechanism of action of treatments known to enhance the efficiency of reprogramming.

Single Cell Resolution Mapping of mRNA and Protein Expression Dynamics in Human Somatic Cell Reprogramming to Pluripotency

Frederick W. Kolling IV

B.S., University of Rhode Island, 2009

A Dissertation

Submitted in Partial Fulfilment of the

Requirement for the Degree of Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Single Cell Resolution Mapping of mRNA and Protein Expression Dynamics in Human Somatic Cell Reprogramming to Pluripotency

Presented by

Fred Kolling IV, B.S

Major Advisor _____

Craig Nelson

Associate Advisor _____

Barbara Mellone

Associate Advisor _____

David Goldhamer

Associate Advisor _____

Charles Giardina

Associate Advisor _____

Michael O'Neill

University of Connecticut

2015

Acknowledgements

I would first like to thank Craig for affording me the opportunity to pursue my thesis work in his lab. He has always given me the freedom to make my own decisions (and mistakes) and has forced me to think critically about every experiment I undertake. Craig has also entrusted me with great responsibility to manage the many academic and industrial collaborations flowing through our lab and this has been one of the most rewarding parts of my graduate career.

To Barbara, thank you for the countless discussions about science, graduate school and life in general. You have provided me with clarity many times when I wasn't sure where I was headed or how I would get there. Thanks for helping me keep my head up.

To David, you have always set the bar high for me since our first meeting before I was even accepted to UCONN. We have had many fruitful discussions, both scientific and otherwise and it's been a pleasure to have you as a mentor the past six years.

To Mike and Charlie, thank you both for taking the time to hear me ramble on about my experiments and future goals / escape plans during committee meetings. You have asked great questions and helped me to identify weak points in my analyses and forced me to view my data in a new light and I'm a better scientist for it today.

To Albert, we have had some serious ups and downs with our various projects but somehow we kept each other afloat. Thanks for all your hard work and for being a great teammate; I can't believe I'll be working somewhere else and not have your energy to keep things exciting. I wish you all the best wherever you end up, I know you'll do great.

To Asav, my partner in crime and science. We've worked hard to help each other design and execute experiments but most importantly you've been one of the best friends I've had in grad school. You have always been there through the rough patches and would stop everything at a moment's notice to grab a celebratory beer or to commiserate over the pains of science or life outside the lab. Grad school would not

have been the same without you around and I thank you for making this experience a million times better. I know we'll make it a point to stay in touch, but I'll miss not having you around buddy.

To the rest of the Nelson lab: Jay, Caroline, Ajay, Steve, Ed and others. It has been fantastic to have you all as lab mates, from great scientific discussion to casual conversation, you have helped make the lab an easy and fun place to come to, day in a day out. Having such an awesome environment to work in has made the grad school experience that much better, I can't imagine what it would have been like without all of you!

To my parents, for believing in me since day one and providing me with a nurturing environment to grow into the person I am today. I can confidently say I wouldn't be here without your constant support and I hope I have done everything I can to make you proud.

To my partner Shannon, thanks for being my teammate for the last three years and for giving me the strength to solve my own problems. You have taught me so much about how to face, rather than run from difficult challenges. This has been a big hurdle to overcome and I know it's not the last. I look forward to facing many more challenges with you as we build our life together.

Table of Contents

Table of Figures

Chapter 1 Introduction

1.1 Project Context and Significance

A pluripotent, embryonic stem cell (ESC) cell is defined by its ability to differentiate into all three embryonic germ layers (endoderm, mesoderm and ectoderm) and can thus contribute to all cell types of an adult organism[1]. In 2005 Takahashi et al demonstrated that overexpression of the transcription factors OCT4, SOX2, KLF4 and c-MYC (OSKM) could revert differentiated somatic cells to a pluripotent state [2]. This process of somatic cell reprogramming generates induced Pluripotent Stem Cells (iPSCs), which promise to revolutionize the field of regenerative medicine. While this technique was developed nearly a decade ago, many barriers still exist to the efficient and reproducible derivation of iPSCs by this method [3–5]. At the time this project was initiated much of the focus in the reprogramming field was on developing different methods to improve the efficiency of the process by manipulating the delivery method of OSKM and introducing new factors or small molecules to the cocktail [6,7]. These efforts yielded small increases in efficiency and provided some insight into the barriers to successful reprogramming, however little was known about the molecular changes occurring during the transition from the somatic to pluripotent state. Early efforts aimed at studying the process itself relied on bulk measurements of cells undergoing reprogramming and were performed primarily in mouse [8–11]. Because reprogramming is temporally asynchronous and produces considerable cellular heterogeneity [11–14], these approaches are not optimal for understanding this process in detail. In addition, many variables contribute to the overall rate and efficiency of reprogramming including the species and cell type of the starting population, as well as the method used to initiate the process.

For these reasons, the goal of this project is to use single cell genomic and proteomic technologies to profile individual cells at the transcript and protein level as they are driven from a somatic to a pluripotent state in human cells. Since the initiation of the project six years ago, the field has advanced at a dizzying pace and other single-cell resolution studies of the process have been reported.

However, this work represents the first single cell analysis of mRNA and protein expression in human reprogramming and the first to compare the dynamics of the process under different reprogramming conditions. This methodology provides novel insights into how human cells acquire pluripotent characteristics, a necessary step for the therapeutic implementation of the technique. In addition, the statistical methods utilized herein contribute to the nascent field of single-cell data analysis and provide a framework for deconstructing complex biological populations that can be extended to a variety of systems.

1.2 The Epigenetic Landscape

During normal development, pluripotent cells with the ability to generate all tissues of the organism progressively lose their developmental potential as they differentiate into increasingly specialized cell types [10,15–18]. Known as canalization, this process was first articulated in 1948 by Conrad Waddington rapidly becoming a dogma of developmental biology. He likened the developmental potential of a cell to a ball rolling down a hill. Just as a ball loses potential energy as it reaches the base of the hill, so too does a cell lose developmental potential as it differentiates into increasingly specialized cell types. This epigenetic landscape as he called it, becomes more restrictive as development proceeds, inactivating regions of the genome and thus canalizing the developmental potential of the cell. It is this epigenetic restriction that produces stable cell types and prevents their spontaneous reversion to a more primitive developmental state. Conversely, reprograming to pluripotency represents a reversal of developmental potential, akin to rolling the ball back up the hill. This process requires a strong external force to perturb a cell from its stable somatic state followed by a suite of molecular changes including resetting of the epigenetic landscape that must occur en route to pluripotency [4,8,19,20].

1.3 Methods of Reprogramming

Reprogramming was first demonstrated using Somatic Cell Nuclear Transfer (SCNT) to generate pluripotent cells by transferring the nucleus of a differentiated cell type into an enucleated oocyte [21]. SCNT remains the most rapid and efficient method of generating pluripotent cells *in vitro* and has been

widely used to clone animals from a variety of species [22–26]. This method was recently demonstrated to work in human cells [27] however, it remains highly controversial and requires an abundance of oocyte donors in order to be feasible for therapeutic purposes.

In 2005, Takahashi et al used retroviruses to overexpress the transcription factors OCT4, SOX2, KLF4 and c-MYC (OSKM) to reprogram skin fibroblasts to pluripotency at an efficiency of 0.01% over a period of 30 days [2,28]. This finding was heralded as a boon for regenerative medicine, promising to provide patient-derived stem cells for therapeutic purposes while avoiding the ethical concerns of SCNT. Because retroviruses containing the 4 factors integrate randomly and may induce unintended alterations to the host genome, they are not ideal for use in therapeutic applications [29,30]. To address this issue, methods have been developed to reprogram cells using floxed OSKM cassettes or piggy-bac vectors to generate "footprint-free" iPSCs [31–34], as well as non-integrating methods such as episomal vectors or mini-circle constructs [31,35–37]. Cells have also been reprogrammed by DNA-free approaches including mRNA/miRNA transfection, protein transduction and recently, small molecule cocktails [38–42]. In addition to altering the delivery method of OSKM, other factor combinations have also been used to successfully reprogram cells. In particular, the Thompson cocktail of OCT4, SOX2, NANOG and LIN28 has been particularly effective and has been shown to act synergistically with OSKM, suggesting these factors promote reprogramming by distinct pathways [6].

1.4 Interrogating the Reprogramming Process

Great strides have been made in recent years to uncover the molecular changes occurring during reprogramming however, our ability to interpret and synthesize these results into meaningful models is complicated by three key features of the process. First, the low efficiency of reprogramming results in 1:1,000 – 1:10,000 cells being successfully reprogrammed. Second, infection with OSKM produces considerable cellular heterogeneity as evidenced by the variety of stable, partially reprogrammed intermediates that can be isolated at various stages of the process and the transcriptional and epigenetic

variation observed in fully reprogrammed iPSCs [2,9,43–48]. Lastly, reprogramming is temporally asynchronous with cells responding to OSKM and progressing towards pluripotency at different rates [11].

The majority of the studies to date rely on the analysis of bulk populations of reprogramming cells and thus are inherently biased towards measuring events that do not lead to pluripotency. In addition the asynchrony and heterogeneity of the process results in bulk expression profiles that reflect an averaging of a continuum of cell states, rather than a discrete population. Despite these complications, population-level measurements have shed light on the ensemble protein and mRNA expression changes during reprogramming and provide insight into the major molecular changes required to achieve pluripotency. Understanding the events occurring in the minority of cells that become iPSCs however, requires techniques with resolution at the single-cell level.

The first single-cell technology used to interrogate the reprogramming process was time-lapse microscopy, whereby cells infected with OSKM were recorded over time and monitored for changes in cell morphology, cell cycle rate and the expression of reporter genes linked to successful reprogramming [49]. In particular, Nanog-GFP and Oct4-GFP reporters have been commonly used to monitor the endogenous activation of these key pluripotency loci, events that have been shown to occur late in the process [11,12,50,51]. Retroactive tracing of these reporter-positive populations back to their initial infection with OSKM has revealed a stepwise progression of cells through the reprogramming process, characterized by an early increase in cell-cycle rate and decrease in cell size, followed by a mesenchymal to epithelial transition and lastly, the activation of the GFP reporter construct [51–55]. The fact that these events occur in a reproducible order suggests that reprogramming is an ordered process, however the molecular resolution of this technique is limited and cannot interrogate mRNA and protein-level changes with appreciable throughput.

The advent of single-cell genomics and proteomics solutions has allowed genome/proteome-wide profiling of individual cells undergoing reprogramming. While this thesis project was being completed, single-cell mRNA profiling was performed in both mouse and human cells utilizing FACS to enrich for

individual productive reprogramming events followed by qPCR or mRNA-seq [56,57]. Recently, single cell proteomics methodologies such as CyTOF, which allows the simultaneous analysis of up to 40 proteins in 1 X 10^5 – 1 X 10^6 individual cells has been applied to mouse reprogramming, adding an additional level of resolution to our understanding of the process [58,59]. The specific findings and models derived from these studies are elaborated below.

1.5 Molecular Dynamics during Reprogramming to Pluripotency

Reprogramming to pluripotency requires an exquisite coordination of molecular/cellular events to revert a "terminally" differentiated somatic cell back to an embryonic pluripotent state, a feat equivalent to developmental time-travel. These alterations in cell state encompass every biological process imaginable, from mRNA, protein, miRNA and non-coding RNA expression, to epigenetics, metabolism and morphology. What follows is an attempt to summarize the vast array of knowledge accumulated to date about these processes but which is likely to evolve rapidly as the field continues to progress at an astonishing rate.

1.5.1 Epigenetics

Perhaps the greatest rate-limiting step to successful reprogramming is the resetting of the epigenetic landscape from a somatic to a pluripotent state. This includes DNA methylation and histone modifications which can control the expression of individual loci and thus, most if not all molecular changes occurring during reprogramming (mRNA, miRNA, lncRNA, protein etc.) first require a change in epigenetic state. These changes are not merely a consequence of reprogramming, but can directly influence the course of reprogramming as well. This is evidenced by the effects of small molecule chromatin modifiers and altered expression of chromatin-modifying enzymes on the efficiency of the process. In general, molecules/factors that promote open chromatin structure such as the HDAC inhibitors valproic acid (VPA) and trichostatin A (TSA) or over-expression of the trithorax subunit WDR5 (involved in H3K4me3) improve efficiency [60–62], while interventions that reinforce repressive chromatin like knockdown of the H3K9 demethylase LSD1, reduce efficiency [63,64]. In addition to these generalized

observations, much is known about the specific chromatin state changes taking place during reprogramming as well.

Almost as soon as reprogramming factor expression is initiated, a plethora of epigenetic modifying enzymes are upregulated, including histone methyltransferases, histone deacetylases/acetyltransferases and nucleosome remodeling complexes[65]. Early in reprogramming, somatic genes that will eventually become silenced lose the activating H3K4me3 mark and gradually acquire repressive H3K27me3 [8,66]. The converse happens at loci activated during the process, particularly genes that are peripheral to the "core" pluripotency network, where H3K27me3 is lost in favor of H3K4me3. Often, these changes occur first in the enhancer regions, followed by changes at the promotor of the corresponding loci [67]. Later in reprogramming (around day 10 in mouse), H3K9me3 and DNA methylation changes occur. These modifications are typically associated with long-term silencing of genes and are commonly found among the core pluripotency factors including OCT4, SOX2, NANOG and others [68,69]. While it is unclear how DNA methylation is removed during reprogramming, it is thought to occur through both active and passive means; passively through the dilution of the modification by DNA replication without maintenance of the methylation mark and actively through the expression of the Tet and Aid demethylation enzymes [70,71]. Loss of DNA methylation at core pluripotency loci by either means is a key late-step in reprogramming, as evidenced by the conversion of stable, partially reprogrammed cells to pluripotency after treatment with the DNA methylation inhibitor 5-aza-cytidine [9,61]. In conjunction with the aforementioned epigenetic changes, a subset of loci, mostly transcription factors involved in development, acquire a bivalent chromatin state characterized by co-occupancy of active H3K4me3 and repressive H3K27me3. This occurs gradually over the course of reprogramming and the dynamics of this change is poorly understood [72,73].

1.5.2 Transcription

Transcriptional changes occurring early in reprogramming are governed by the binding of the OSKM transcription factors to DNA as dictated by the epigenetic landscape. An early wave of gene

induction is associated with promoter regions decorated with the activating histone mark H3K4me3 and the binding of c-MYC and KLF4 [67,74]. This includes a set of "early pluripotency" genes as well as loci associated with apoptosis. Indeed, apoptosis is a common end-point for many reprogramming cells and knockdown of p53 by shRNA has been used to increase reprogramming efficiency [31,75,76]. Simultaneously, late pluripotency loci that possess repressive epigenetic marks such as H3K27me3 and DNA methylation in somatic cells, are bound by a combination of O, S, K and/or M at their distal enhancer regions but remain transcriptionally inactive at this stage. During the second wave of transcriptional activity a concomitant loss of H3K27me3 and DNA methylation and gain of H3K4me3 occurs at genes that will become active in the pluripotent state (i.e. late pluripotency genes). At this stage, the expression of pluripotency loci is governed by combinations of transcription factors similar but not identical to those observed in true hESCs/iPSCs.

While the above primarily refers to changes in mRNA expression, miRNA and lncRNA expression have also been measured during the course of reprogramming [40,41,77–80]. The expression of miRNAs exhibits a biphasic pattern similar to that of mRNA expression, with miRNAs targeting fibroblast genes induced early and those targeting pluripotency genes down-regulated late in the process. This results in the expected inverse correlation between the expression of miRNAs and their corresponding mRNA targets. For example, miR-294 is upregulated early in reprogramming and its target TGFBR2 is subsequently downregulated as cells reprogram[39,78,81]. The opposite is true for the let-7 miRNA which is turned off late in reprogramming, allowing the pluripotency loci LIN28 and SALL4 to be activated in the iPSC state[82]. Similarly, the expression of lncRNAs have recently been shown to shift from a fibroblast to pluripotent expression profile [83]. lncRNAs recruit epigenetic modifying complexes to target loci and thus play an important role in establishing the epigenetic landscape and stabilizing cellular identity. Taken together, the dynamics of mRNA, miRNA and lncRNA expression demonstrate the concerted transcriptional changes that must occur as cells transition to a pluripotent state, each of which represents a potential barrier to successful reprogramming.

1.5.3 Protein Expression

In bulk samples subjected to LC/MS-MS, protein expression dynamics reflect what is seen at the mRNA level, with an early wave of activation associated with MYC target loci and a late wave of pluripotency gene activation [84,85]. Interestingly, by comparing these data with bulk mRNA-seq measurements collected in parallel, these authors also find that genes involved in cell adhesion, androgen/oestrogen signaling and mitochondrial function have a lower correlation at the protein compared with the mRNA level. These findings were extended by Hansson et. al. who demonstrated mRNA and protein expression exhibit greater correlation early in reprogramming and that this relationship deteriorates later in the process. Together, this suggests that post-transcriptional regulation of protein expression may be a rate limiting step in the reprogramming process, however it is unclear if this phenomenon occurs in the few cells undergoing productive reprogramming which would require single-cell resolution to establish definitively.

Recently, mass cytometry (CyTOF) has emerged as a technology to profile protein expression in individual cells using a combination of flow cytometry and mass spectrometry which can measure ~40 proteins in hundreds of thousands of individual cells[86]. This technique was recently used to profile mouse fibroblasts undergoing reprogramming for a panel of 34 proteins[58,59]. While these studies did not report waves of expression as indicated by bulk analysis, they did identify apparent intermediate states on the route to pluripotency. This includes an early $OCT4^{high}KLF4^{high}$ population that later activates CD73 and CD104 and finally EpCAM before acquiring an iPSC-like profile. This late activation of EpCAM, a marker of MET is in contrast to previous reports at the mRNA level that MET is an early event in the process. Again, the discordance between mRNA and protein expression suggest post-transcriptional regulation may play an important role in the reprogramming process. In addition to the above-mentioned trajectory, these authors also identify a subset of reprogramming cells that exit the cell cycle (as evidenced by low Ki67 expression) and altogether fail to productively reprogram. Most importantly, these studies have identified a unique intermediate $CD73^{+}$, $CD49d^{+}$, $CD200^{+}$ population of cells, of which 12%

become iPSCs. This represents the greatest enrichment for productive reprogramming intermediates to date and demonstrates the power of CyTOF to identify useful biomarkers in highly heterogeneous populations of cells. Furthermore, CyTOF can interrogate post transcriptional modifications such as phosphorylation and methylation of histones and other proteins, permitting single cell analysis of entirely new levels of cellular activity such as epigenetics and the kineome.

1.5.4 Cell Cycle

Human fibroblasts (and many other somatic cell types) exhibit doubling times on the order of 24-36hrs whereas ESCs double once every 16-18hrs[53,87–89]. This represents a challenge for reprogramming cells to increase their progression through the cell cycle and proliferate more rapidly. It is thought that a major role of c-MYC in the OSKM cocktail is to increase proliferation of cells and while reprogramming can succeed in its absence, the efficiency is ~100-fold lower [32,74,90]. Not only is an increase in cell cycle progression critical for the maintenance of pluripotency, it is also thought to facilitate changes in the epigenetic landscape [11,91]; the replication of the genome represents an opportunity to dilute somatic epigenetic modifications while marks associated with pluripotency are established on the newly synthesized DNA. This is corroborated by a report from Hanna et al., whose modeling of the reprogramming process showed that the number of cell divisions is a better predictor of successful reprogramming than days post OSKM induction[11].

1.5.5 Mesenchymal to Epithelial Transition

Many cell types used for reprogramming including fibroblasts are mesenchymal in nature, making limited cell-to-cell contacts and producing extra cellular matrix components to facilitate adhesion to and mobility across surfaces. These cells also express characteristic mesenchymal genes including N-Cadherin, Vimentin, Fibronectin, Snai2 and Twist, the maintenance of which is dependent on TGFB signaling[54,92,93]. In contrast, embryonic stem cells form an epithelium with tight cell-to-cell junctions and form colonies in culture with basal-apical polarity, an activity primarily coordinated by the expression of E-Cadherin and EpCAM[94,95]. Reprogramming cells must undergo a mesenchymal to epithelial transition

(EMT), a critical step in the conversion to pluripotency, but there are conflicting reports as to whether this is an early or late event in the process[55,58,96].

1.5.6 Inactivation of OSKM Viruses

For reprogramming methods that rely on viral delivery of OSKM, inactivation of the four factors is a critical late step in forming iPSCs[4,97]. The persistence of OSKM expression is associated with the generation of non-iPSC, self-renewing cell types in reprogramming cultures as well as aberrant differentiation of iPSCs into target cell types[33,98–100]. This is due to the sensitivity of ESCs and iPSCs to transcription factor dosage which control the balance between pluripotency and differentiation[101–103]. For example, persistent OCT4 expression prevents the differentiation of iPSCs, while as little as 2.5-fold overexpression of SOX2 downregulates pluripotency targets and promotes differentiation to neuronal and mesodermal lineages[104–106]. While the importance of inactivating OSKM in iPSCs is well understood, the mechanisms by which this occurs are not. It has been found that silenced retroviral elements in ESCs are associated with H3K9me3 and DNA methylation and binding of the MeCP2 and HP1 proteins in the viral LTRs. However, none of the known mediators of these modifications are specifically expressed in cell types where silencing occurs and thus the factors involved in this process have yet to be identified[97,107–109].

1.6 Models of the Reprogramming Process

Through the extensive molecular data generated in recent years, a model of the reprogramming process has emerged. The variety of datasets (mRNA vs protein, single-cell vs bulk, etc.) somewhat complicates a unified model, however several common themes have been observed. In 2013, Buganim et al proposed a two-phased model of reprogramming following single cell mRNA analysis in the mouse system. They observed that cells responding to OSKM infection initially transition through a “stochastic phase,” whereby gene expression patterns appear random and uncoordinated. This is followed by a “deterministic phase” with cells activating key pluripotency loci in a stepwise, hierarchical fashion[56,110]. While the deterministic phase is thought to occur rapidly as cells become locked into a pluripotent fate, the stochastic phase is protracted and probabilistic, resulting in the asynchrony or “variable latency” that

is characteristic of the process. This is concordant with the observations that the binding of the OSKM factors early in reprogramming (48hrs post-infection) is dictated by the epigenetic state of the target loci and that changes in the epigenetic state is required for factor binding to some loci, in particular pluripotency genes with repressive chromatin states[67]. Remodeling of the epigenetic landscape is facilitated by DNA replication and cell division and thus increased in cell cycle have been shown to improve reprogramming efficiency. Consistent with this notion, factors that increase cell proliferation and promote permissive chromatin states (ie histone acetylation) also improve the efficiency of reprogramming[111].

The deterministic phase was identified by comparing the expression profiles of single-cells late in the reprogramming process that had not activated an Oct4-GFP reporter, with Oct4-GFP+ reprogramming cells. Bayesian analysis of these populations revealed a hierarchy of pluripotency gene activation initiated by the expression of Sox2. Successful reprogramming, albeit at a lower efficiency, was observed by overexpression of downstream members in this cascade in the absence of Sox2, further corroborating these findings.

Taken together, this model implicates the stochastic phase as the major rate limiting step in reprogramming and provides a target for improving the rate and efficiency of the process. In contrast, dissection of the hierarchical phase is likely to provide informative biomarkers that can be used to identify successful reprogramming events. This model, while supported by data from human reprogramming experiments is largely comprised of studies performed in mouse and it is unclear how these findings translate between the two species.

1.7 Mouse vs Human

The therapeutic potential of reprogramming is dependent on our ability to understand and optimize the technique in human cells. Despite this fact, the overwhelming majority of research into the mechanisms and dynamics of the process comes from studies in mouse due to the abundance of genetic

tools available to manipulate these cells and the increased speed of reprogramming in this system[112–114]. Reprogramming in mouse occurs in as little as 10 days, compared with 30 days in human cells and the efficiency can be up to 20-fold higher in mouse[91]. In addition, many studies of reprogramming in mouse take advantage of secondary reprogramming systems whereby mouse fibroblasts are infected with a doxycycline-inducible OSKM construct, reprogrammed to iPSCs and subsequently re-differentiated into fibroblasts. Addition of doxycycline to these cultures allows for a more homogeneous reprogramming process and facilitates molecular dissection of the events leading to pluripotency. It is unclear however, whether the prior reprogramming of these cells faithfully represents the process as it occurs in primary fibroblasts. Thus, much of what we know about the molecular underpinnings of reprogramming come from artificial mouse systems and it is unclear how these finding translate to human cells. While it is expected that the cellular processes involved will be consistent between species, the particular genes implicated in these processes may be different. Indeed, the wiring of the gene regulatory networks governing early development are quite varied between the two species[115]. Nonetheless, much of the information to date comes from mouse reprogramming, illustrating the need for similarly detailed studies to be performed in human cells.

1.8 Scope of the Thesis

The primary goal of this thesis is to profile the transcript and protein dynamics in human cells undergoing reprogramming by OSKM at single-cell resolution and to model progression of cells through the stochastic phase. After establishing a baseline model, this analysis is performed under two delivery methods of the OSKM virus (monocistronic and polycistronic) and in two different cell types (BJ and MRC-5 fibroblasts) to better understand how differences in reprogramming conditions impact the acquisition of pluripotency.

Because much of our current knowledge of reprogramming comes from transgenic mouse systems and a variety of reprogramming methods, it remains to be seen how these findings translate to the human reprogramming system and whether the method of reprogramming has an impact on the process.

By examining both mRNA expression dynamics under different reprogramming conditions as well as protein expression, this work provides a unified model of the reprogramming process that is directly relevant to the therapeutic application of the technique. In addition, the statistical models developed herein provide powerful tools for the field of single-cell analysis to dissect and interrogate heterogeneous mixtures of cells.

Chapter 2 Single Cell Analysis and Modeling of Monocistronic Reprogramming by OSKM

This work was published in PLOS ONE in 2014

2.1 Introduction

Methods of reprograming somatic cells to a pluripotent state (iPSC) have enabled the direct modeling of human disease and ultimately promise to revolutionize regenerative medicine [5,116]. While iPSCs can be consistently generated through viral infection with the Yamanaka Factors OCT4, SOX2, KLF4, and c-MYC (OSKM) [2], infected cells rapidly become heterogeneous with significant differences in transcriptional and epigenetic profiles, as well as developmental potential [14,46,117–119]. This heterogeneity, the low efficiency of iPSC generation (0.1-0.01%) and the fact that many iPSC lines display karyotypic and phenotypic abnormalities [44,120,121] has hindered the production of iPSCs that can be used safely and reliably in a clinical setting. A thorough mechanistic understanding of the reprogramming process is critical to overcoming these barriers to the clinical use of iPSC.

In the past several years, ChIP-seq and RNA-Seq experiments have revealed ensemble gene expression and epigenetic changes that occur during reprogramming by OSKM, and have greatly enhanced our understanding of the process [5,10,50,66,122]. These studies require the use of populations of cells comprised of heterogeneous mixtures undergoing reprogramming (0.01-

0.1% of which will become iPSC) or stable, partially reprogrammed self-renewing lines arrested in a partially reprogrammed state, unlikely to ever become iPSCs without additional manipulation [14,117–119]. Because these techniques rely on either the ensemble properties of mixed populations, or upon the analysis of cell lines arrested at partially reprogrammed states that may not be representative of normal intermediate steps in a functional reprogramming process, they have limited ability to reveal the changes that appear to be essential to successful reprogramming.

Longitudinal single-cell imaging studies provide a powerful complement to ensemble, population level analyses. Live imaging studies have identified a number of key morphological and cell cycle related changes that occur during reprogramming to iPSC [55,123]. These observations suggest that an ordered set of phenotypic changes precede acquisition of the fully pluripotent state [10]. However, these studies are necessarily limited in their molecular-genetic resolution, and they provide little insight to the transcriptional changes accompanying key morphological and developmental transitions in the reprogramming process.

Recently, a single-cell transcriptional analysis of reprogramming of mouse fibroblasts by OSKM revealed that reprogramming proceeds in two major phases: an early stochastic phase followed by a rapid "hierarchical" phase [110]. While the latter phase appears deterministic and is characterized by the coordinated expression of pluripotency genes in an ordered fashion, the early phase exhibits apparently random gene expression patterns that persist through the majority of the process [110,124]. This conclusion is further supported by two key pieces of evidence from other studies: 1) transgenic OSKM activity is required for the majority of the reprogramming process, indicating that most of this process is not governed by the concerted action of the endogenous pluripotency gene regulatory network (GRN) [51,52,123]; and 2) a mechanistically

undescribed period of variable ‘latency’ of cells in the stochastic phase results in significant temporal variability in the appearance of fully reprogrammed iPSC colonies [11]. Some insight to pluripotency gene activation during the stochastic phase was provided by a recent study in mouse fibroblasts that describes the ‘gradual activation of pluripotency genes’ between the initial response to OSKM induction and the activation and stabilization of the pluripotency GRN [8]. Together, these findings suggest that the stochastic phase is a major rate-limiting step in the reprogramming process, but provide little mechanistic insight into the molecular underpinnings of these events. In addition, it has not yet been determined how these findings translate to the reprogramming of human cells, which will be required prior to clinical application of iPSCs.

Several studies have attributed the protracted stochastic phase to the requirement for extensive chromatin remodeling during reprogramming [21,125]. These changes involve the complex coordination of factors to deposit and remove histone modifications and DNA methylation at specific loci to achieve a pluripotent epigenetic state. The need to reset the epigenetic landscape appears to delay the coordinated activation of the pluripotency GRN and is likely to be a major barrier to rapid and efficient reprogramming. Indeed, it has been shown that OSKM binding in the early stages of reprogramming is greatly impeded by the presence of repressive chromatin, and initial binding is largely restricted to existing open chromatin domains [5,50,66,67,69]. Subsequent remodeling of somatic cell chromatin clearly occurs, but the order and mechanism of remodeling events during the stochastic phase is not fully understood. Accurate mapping of gene expression dynamics during the stochastic phase can provide a framework for the molecular dissection of these rate-limiting events in reprogramming.

In this study we perform single-cell transcript analysis of MRC-5 human lung fibroblasts undergoing reprogramming by OSKM and find that cells appear to follow two trajectories: one

toward an ESC-like state (the “productive” trajectory) and the other away from both ESC and fibroblasts (the “alternative” trajectory). These trajectories can be differentiated by the concerted consolidation of expression of a suite of chromatin modifiers in cells entering the productive trajectory and the down-regulation of these same genes in cells entering the alternative trajectory. By analyzing the dynamics of gene expression changes along the productive trajectory (toward pluripotency) we demonstrate that changes in gene expression in the stochastic phase of reprogramming are not simply gradual and random; rather, genes are activated and inactivated at specific points during the progression from fibroblast to iPSC. Coupling single-cell transcript profiling with mathematical modeling we show that the gradual acquisition of pluripotency gene expression during reprogramming occurs as an ordered, probabilistic, gene-specific process that shows no signatures of interdependence between genes. This finding is consistent with the hypothesis that gene-specific chromatin states in the starting cells control gene activation dynamics during the reprogramming process. Our map of reprogramming also provides a robust model that can be used to dissect the precise mechanisms and chromatin modifications that limit the rate and efficiency of conversion of somatic cells to iPSC. This work represents a rigorous single cell transcript analysis of the reprogramming process in human cells and lays the foundation for the precise measurement and mechanistic dissection of this critical rate-limiting step in reprogramming.

2.2 Results

2.2.1 Experimental Design

In this report we combine qualitatively and quantitatively robust single-cell transcript profiling [126] with FACS to measure the progression of individual MRC-5 human fetal lung fibroblasts through the reprogramming process. To make our results as broadly relevant as

possible we used viral delivery of the OSKM transgene cocktail, the most widespread method applied to human cell reprogramming [28,127]. At select time points after transduction, cells were dissociated, stained, analyzed and collected by FACS. FACS markers used in this study include GFP (virus derived), αSSEA4, αTRA-1-60, and αCDH1 (see Materials and Methods). These markers were essential and allowed for enrichment of the rare cells exhibiting hallmarks of productive reprogramming. For example, SSEA4 and TRA-1-60 routinely provide ~30 and 3,000 fold enrichment, respectively (data not shown). While very few SSEA4+ cells are likely to become true iPSCs, they provide a measurement of cells that have begun to exit the fibroblast state in response to OSKM transduction. In contrast, isolation of TRA-1-60+ cells later in reprogramming (Day 14) is likely to yield a large number of cells destined to become iPSC. In fact, >90% of these cells remain TRA-1-60+ after sorting and subsequent culture and this stability of the TRA-1-60+ phenotype has been shown to be a major determinant for the potential of cells to become iPSC [57]. Single cells with defined FACS phenotypes were collected into cell lysis buffer and subject to single-cell RT-qPCR as previously described [126] (Figure 1A and Supplemental Figure 1). Throughout the course of this study we isolated and pre-screened 576 cells in total, using 172 cells that passed quality control for our final analysis (see Materials and Methods and Supplemental Table 1). This includes many partially reprogrammed cells, as well as an un-transduced set of MRC-5 fibroblasts and H9 human embryonic stem cells (H9-hESC), which represent the beginning and end states of the process, respectively.

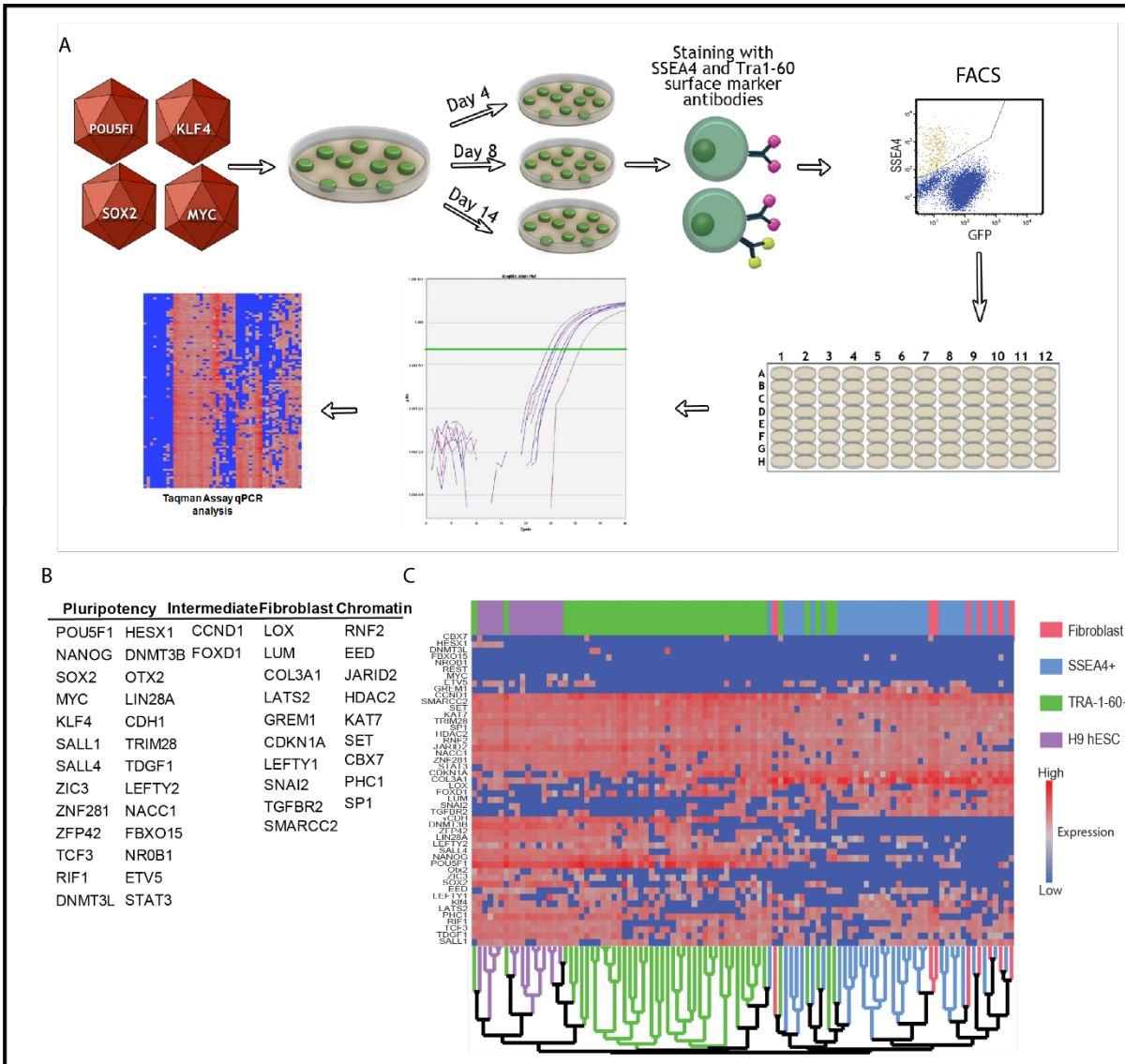


Figure 1: Schematic representation of the pipeline used to isolate and analyze single cells undergoing OSKM-mediated reprogramming. A) Cells were infected with OSKM (MOI = 5) and cultured for 4, 8 or 14 days prior to harvest. Cells were then singularized and stained with SSEA4 and TRA-1-60 antibodies and subjected to FACS. SSEA4+/TRA-1-60- (SSEA) and SSEA4+/TRA-1-60+ (TRA-1-60) single cells were sorted directly into lysis buffer in 96-well plates followed by RT and linear pre-amplification. Amplified cDNA samples were used for Taqman qPCR analysis of 48 genes on an Applied Biosystems 7900HT real time machine and data analysis was performed in JMP. B) Table of the 48 gene panel used for qPCR analysis, categorized as fibroblast-associated, pluripotency-associated, intermediate marker or chromatin modifier gene. C) Unsupervised hierarchical clustering analysis illustrating the effective isolation of single cells by FACS for SSEA4 and TRA-1-60 surface markers. While some overlap is observed between the two populations, they are largely transcriptionally separable. GFP+-only and CDH1+ populations have been excluded for illustrative purposes.

In order to monitor progress toward pluripotency, and away from the fibroblast state, we assembled a 48-gene qPCR panel including genes expressed in fibroblasts [55,128,129], a large number of genes involved in the maintenance of pluripotency (including various chromatin modifiers) [122,130–132] and genes previously suggested to be intermediate markers of the reprogramming process [9,49]. For a complete list of qPCR markers see (Figure 1B and Supplemental Table 2). Initial visualization of the full dataset by unsupervised hierarchical clustering reveals that our FACS sorting strategy, and qPCR marker panel, isolates statistically separable populations that capture a range of transcriptional phenotypes between the fibroblast and pluripotent states (Figure 1C). We then performed a series of statistical analyses to: 1) describe probable trajectories followed by OSKM-infected cells; 2) measure the progress of cellular transcriptional profiles toward a pluripotent transcriptional phenotype; and 3) determine the order of gene activation during the reprogramming process.

2.2.2 Mapping the Trajectory of OSKM-Infected Cells Throughout Reprogramming

As a first step in visualizing our single cell transcription dataset, we used principal components analysis (PCA) to assess the complexity and major sources of variation in gene expression between all cells collected in our study. This analysis reveals that the first two PCA dimensions account for 33.1% of the observed variation, where PC1 primarily represents a cell's distance from hESC, and PC2 primarily captures distance from fibroblasts (Figure 2A). In addition, these two axes appear to represent distinct trajectories followed by cells transduced with OSKM The first is a roughly linear productive trajectory between the fibroblast and hESC groups (R^2=0.60, Figure 2B) and the second is an orthogonal trajectory leading away from fibroblast but not towards a pluripotent phenotype (herein referred to as the alternate trajectory, or ALT). Because the productive and alternate trajectory are well correlated with the PC1 and PC2 dimensions respectively (Figure 2C) and capture much of the variation in our dataset, we developed

a metric to analyze our data in a 2-dimensional Euclidean space that maps each cell's distance (relative similarity) to the centroids of both the Fibroblast and hESC groups. In addition, we construct a Euclidean diagonal between Fibroblast and hESC which we term the "reprogramming progression axis". This axis serves as a useful measurement of a given cell's progression towards pluripotency and is a metric used in all subsequent analysis presented here.

It is important to note that our analysis constructs likely reprogramming trajectories by sampling partially reprogrammed cells. This approach is common among many efforts to sample dynamic processes and is particularly ubiquitous in attempts to dissect the reprogramming process[9,49,74]. We apply the standard parsimonious assumption that the shortest path defined by these samples represents the most likely trajectory of the process. One caveat of this approach is that we cannot exclude the possibility that progression within the observed state-space is non-linear, and may be complex and/or cyclical. These possibilities will need to be ruled out with longitudinal live cell studies beyond the scope of this work. Another important consequence is that while cells clearly take time to traverse the trajectory, we do not expect progress along a trajectory to have a linear relationship with time. However, progress may be loosely thought of as a surrogate for time but should not be strictly interpreted as such.

Interestingly, when mapping the FACS-sorted phenotypes onto our Euclidean similarity graph we noticed that, while SSEA4 and TRA-1-60 appear in the expected order ($SSEA4^+$ before $TRA\text{-}1\text{-}60^+$), the $SSEA4^+$ and $SSEA4^+/TRA\text{-}1\text{-}60^+$ populations exhibit considerable transcriptional heterogeneity (Figure 2D). SSEA4 positive cells are found in both the productive and alternative trajectories suggesting that, while SSEA4 may be a reliable marker of exit from the fibroblast state, it does not necessarily indicate that cells have moved toward a pluripotent transcriptional phenotype. Even more pronounced is the diversity of TRA-1-60 positive cells. The transcriptional phenotype of these cells extends from a nearly fibroblast-like profile, to a nearly ESC-like profile. The extremely high degree of transcriptional heterogeneity we observe, even within well-defined and widely utilized FACS profiles, underscores the

utility of single cell analysis to dissect fine differences in gene expression between partially reprogrammed cells.

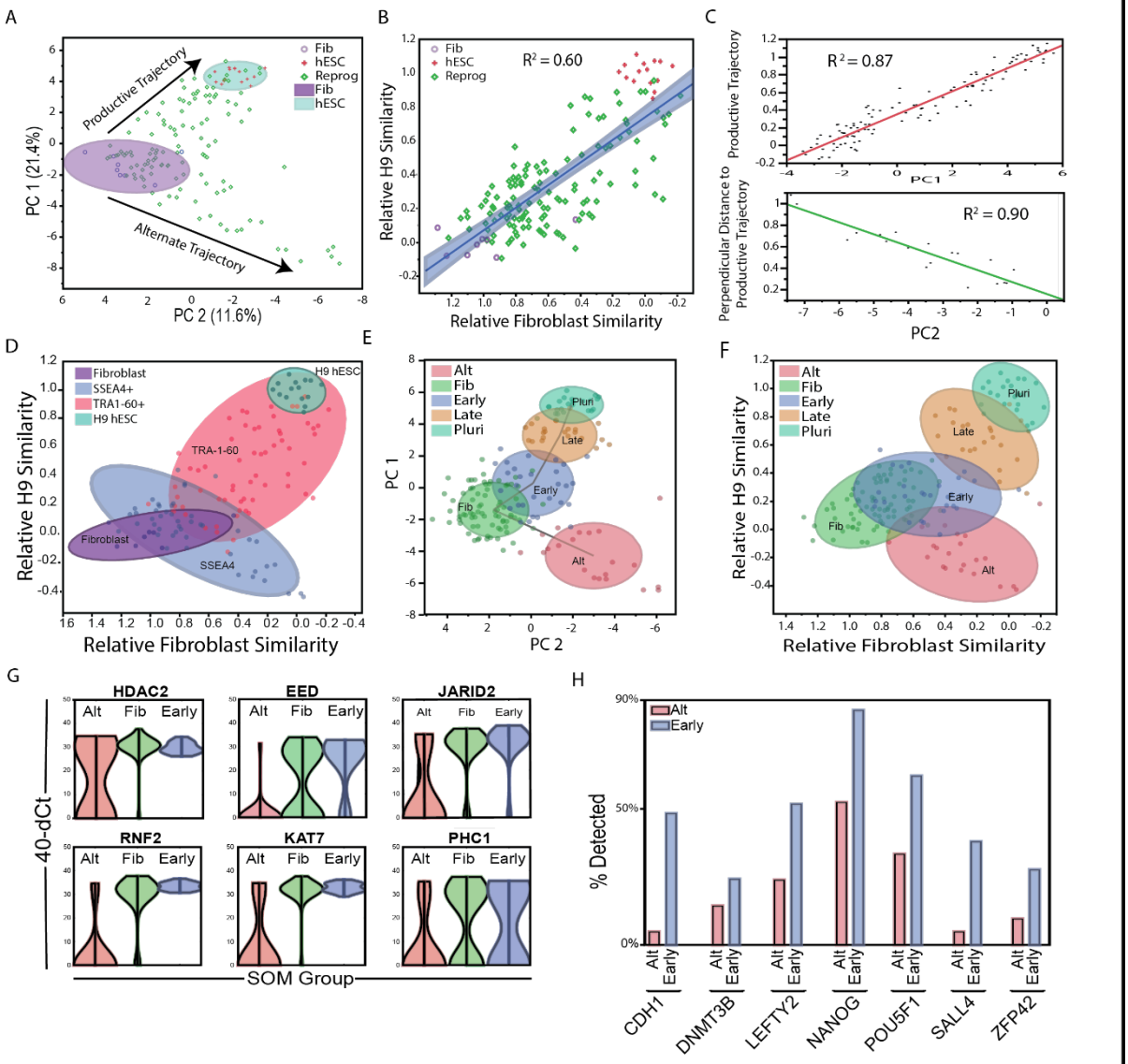


Figure 2: Mapping the trajectories of OSKM- infected cells. A) Principle Components Analysis (PCA) shows the two trajectories followed by OSKM-infected cells. One productive trajectory leading away from the starting fibroblast population (purple oval) and towards the hESC group (teal oval) and a second, orthogonal trajectory leading away from both fibroblast and hESC, denoted as the "alternate trajectory"." B) Regression analysis showing the linear nature of the productive trajectory. C) Correlation analysis between PC1 and the productive trajectory (C, top panel) and PC2 and the perpendicular distance to the productive trajectory. D) Mapping of cell types onto a Euclidean distance graph shows the broad range of transcriptional phenotypes observed for SSEA4+ (blue oval) and TRA-1-60+ (pink oval) FACS-sorted cells. Also included are untransfected MRC-5 fibroblasts (purple oval) and pluripotent H9 hESC cells (teal oval). Self-Organizing Map (SOM) analysis identifies transcriptionally separable groups within our dataset in PCA (E) and Euclidean (F) space. This includes 4 groups along the productive trajectory (Fib, Early, Late and Pluri) as well as one group comprised of cells in the alternate trajectory (Alt). G) Violin plots comparing expression of chromatin modifier genes between the Alt (red), Fib (green) and Early (blue) groups. Gene expression levels are plotted on the y axis, with the width of the graph representing the prevalence of cells at a given expression level. H) Bar graph illustration differences in pluripotency gene expression between the Alt and Early groups.

With the phenotypic diversity of commonly utilized cell surface markers in mind, we utilized a Self-Organizing Map (SOM) to identify separable groups along the two previously described reprogramming trajectories in both PCA and Euclidean space (Figure 2E and F, respectively). Four of these groups (Fib, Early, Late and Pluri) lie along the productive trajectory from Fibroblast to ESC and the fifth encompasses cells in the alternate trajectory. It is important to note that while these groups can be statistically distinguished from one another, we do not believe these represent discrete stages in the reprogramming process. Further inspection reveals that progression along the productive trajectory is characterized by the consolidation of chromatin modifier expression, an increased probability of pluripotency gene expression, a progressive decrease in the expression of fibroblast markers and transient expression or repression of predicted intermediate markers [9,11]. Among the earliest distinctions between the productive and alternate trajectories (Early vs Alt) is the induction of chromatin-modifying enzyme expression. While many of these genes are expressed at low levels in fibroblasts, they are coordinately up-regulated in the "Early" group, and become expressed at uniformly high levels in all cells progressing towards pluripotency. In contrast, cells in the alternate trajectory down-regulate or eliminate expression of these genes (Fig. 2G). In addition, "Alt" cells fail to upregulate the expression of early pluripotency genes (Figure 2H) and are found at all of the time points examined, suggesting that these cells are unlikely to be on a trajectory that ultimately leads to pluripotency. Because "Alt" cells appear to be following an orthogonal trajectory that may lead to fates unrelated to ESC (such as transformation or apoptosis [4,133]) they were excluded from further analysis of the productive reprogramming trajectory.

Taken together these data indicate that OSKM infected cells exit the fibroblast state along two distinct trajectories, and that the upregulation of chromatin modifiers marks a key early step towards successful reprogramming. The rapid upregulation of chromatin modification genes is consistent with the need for extensive chromatin remodeling prior to establishment of the endogenous pluripotent GRN [5,18,72].

2.2.3 Mapping Coarse Changes in Gene Expression along the Productive Trajectory

In order to provide a rough benchmark for other literature examining transcriptional changes in ensemble samples of partially reprogramed cells, we identified quantitative expression differences between SOM groups along the productive trajectory (Figure 3). It is clear from this data that specific changes in gene expression occur along different portions of the trajectory, which suggests an underlying order to the gradual acquisition of pluripotency gene expression during the reprogramming process. However, closer analysis reveals that there does not appear to be tight covariance between genes activated along the progression toward pluripotency. Representative bubble plots illustrating transcript presence and absence (Figure 3 and Supplemental Figure 2) show that genes being activated during reprogramming exhibit a period of heterogeneity in transcript detection prior to being detected in all cells approaching pluripotency. Quantitative analysis of gene expression levels also supports this finding (Figure 3, Supplemental Figure 3). These plots depict gene expression levels on the y-axis, overlain with a distribution graph showing the range of expression values within the population. A unimodal distribution indicates uniform expression around a mean within the population, whereas a bimodal distribution demonstrates a transcriptionally heterogeneous population (e.g. high/low) for the gene in question. Nearly all the genes in our study exhibit this bimodal behavior at some point along the reprogramming trajectory, before achieving a unimodal distribution as they approach the fully reprogrammed state, however the point of bimodality varies in a gene-specific manner. These findings demonstrate that the activation or inactivation of gene expression during reprogramming proceeds through a probabilistic intermediate step, resulting in transcriptionally heterogeneous cell populations, and that the timing of this transition occurs with gene specific dynamics.

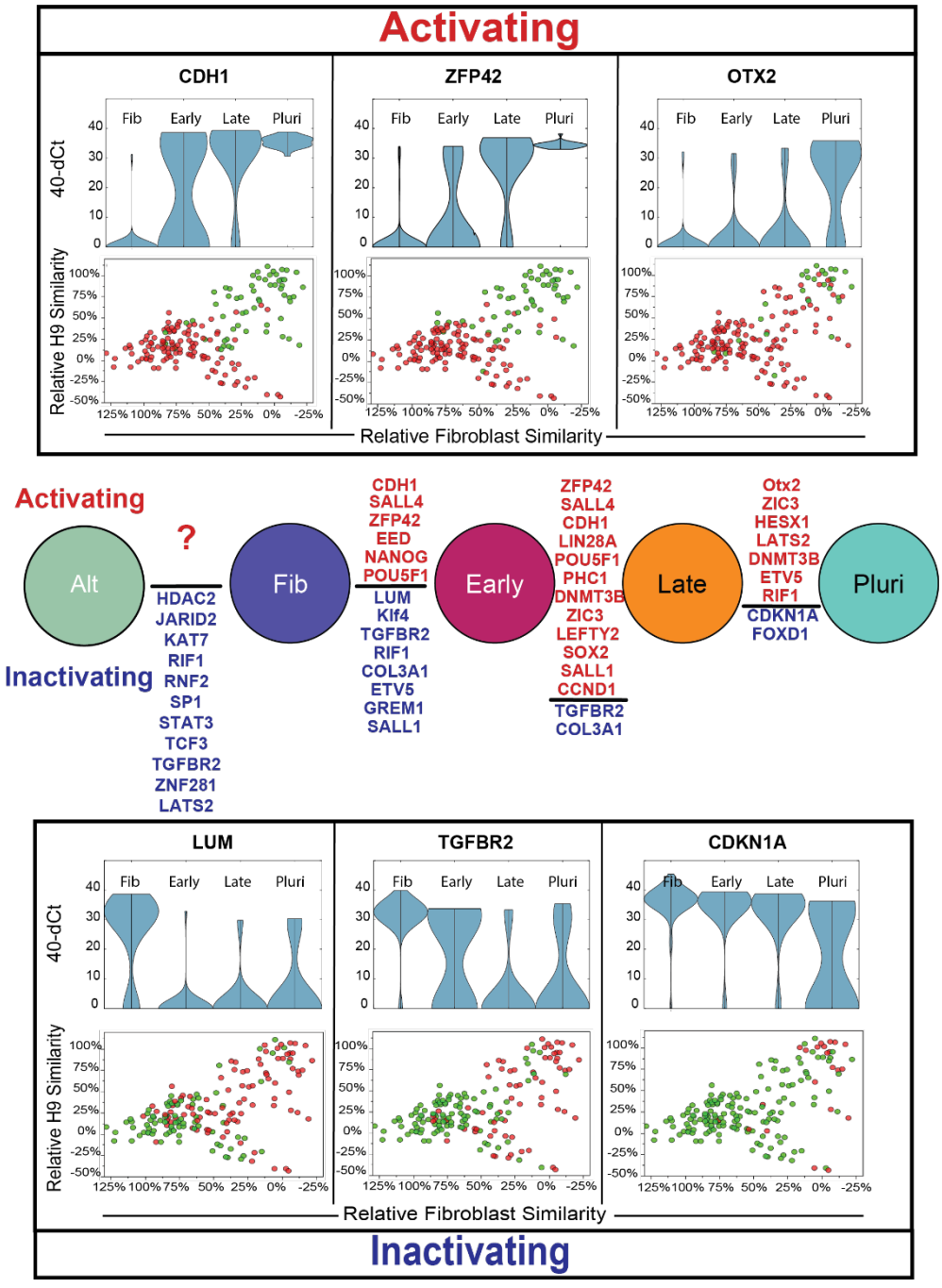


Figure 3: Gene Expression Changes between SOM Groups. (Middle panel) Tukey-Kramer test results showing significant increases or decreases in gene expression between the groups identified in the PC-SOM analysis ($p>0.05$). Genes are ranked in order of significance from highest to lowest. Violin and bubble plots (above and below) show qualitative and quantitative changes (respectively) in per-cell gene expression for the genes with the greatest change between groups. Top panel shows genes whose level and probability of expression undergo an "activating" effect during reprogramming, while genes with decreased probability of expression during reprogramming are labeled "inactivating" and shown in the bottom panel.

In order to scan for potential differences in reprogramming gene expression dynamics between species (mouse and human) we processed our data so that it would be roughly comparable to that generated by Polo et al [8]. As in the present study, Polo and coworkers used FACS to isolate and measure the transcriptional profiles of a large number of partially reprogrammed mouse fibroblasts and clustered genes based on their expression dynamics. We compared these clusters to the dynamics of the human orthologs [122,130] represented in our dataset (Supplemental Figure 4). While high-resolution comparison was not possible with the publically available mouse data, most genes shared between datasets appear to exhibit similar dynamics in the stochastic phase. That is, early mouse genes change expression early in the human trajectory, while late genes change later in the trajectory. However, despite the coarse limits of resolution in this comparison, several genes, including NANOG, LIN28A, POU5F1 and STAT3, appear to change at different stages of the reprogramming process in these two species. These disparities, while requiring more direct comparison and detailed confirmation, are consistent with distinct differences between regulation of the pluripotent state in mouse and human cells as well as probable differences in the starting chromatin state of loci in mouse and human fibroblasts.

2.2.4 Reprogramming is a Loosely Ordered Probabilistic Process Effectively Modeled by Gaussian Distributions

Our observation that distinct transcriptional differences exist between PC-SOM clusters indicates that gene expression changes during the stochastic phase of reprogramming appears to occur in an ordered fashion. However, the coarse grained nature of this differential analysis between statistically identifiable, but not necessarily biologically relevant groups, provides little insight to the exact nature of the order of gene expression dynamics during the stochastic phase. In particular, we wanted to address two specific questions: 1) Is the acquisition of pluripotency gene expression random and gradual, with all genes approaching a pluripotent profile at a

uniform rate over the course of the process?; and 2) Is there sub-structure within the patterns of gene activation that would suggest the activation of modules within the pluripotency GRN? We addressed these questions by differentiating between null and alternative hypotheses (in the form of distribution models) predicting gene expression frequencies along the reprogramming trajectory from MRC-5 to H9-ESC and comparing these to what we observe in our experiments.

In order to formally address the first question we modeled random gradual change in gene expression by assigning each fibroblast and pluripotency marker a uniform rate (probability) of change along the trajectory from MRC-5 to H9-ESC that would result in predicted gene expression frequencies that match the observed frequencies at the start (MRC-5) and end (H9-ESC) of the process [8]. In contrast, our alternative hypothesis was that genes change expression at specific stages of the process; in other words, gene expression during the stochastic phase is *ordered*. This alternative scenario was modeled by fitting Gaussian probability distributions to each gene such that the probability distribution was centered at the point of greatest change in gene expression frequency along the reprogramming trajectory. In order to model the behavior of transient genes, and to help calibrate differences between goodness of fit between models, we also built more complex models with two probability distributions, which allowed us to model genes that change expression at two points in the process. Changes in gene expression frequency predicted by our null model are linear, while the alternative model with one probability distribution predicts sigmoidal changes and the two distribution model allows for more complex dynamics of change in gene expression frequency, such as transient activation or inactivation. The goodness of fit of each model to our observed data was then measured for each gene in both PCA and Euclidean space using an F-test statistic. Because goodness of fit typically scales with the number of parameters in a model, the Gaussian models were penalized for added

parameters using a corrected Akaike Information Criterion (AIC, see Materials and Methods). The results of these tests can be found in (Figure 4A-D and Supplemental Table 3).

As demonstrated in Figure 4B, the vast majority of genes reject the null hypothesis (F-statistic > F-Critical) in favor of a Gaussian model. Note that many genes that reject the null hypothesis do so very strongly, while the few genes that better fit linear dynamics do so only marginally (Figure 4C). In addition, most genes that do not reject the uniform model exhibit little or no change over the course of reprogramming or have noisy expression profiles. Both of these observations suggest that most gene expression changes occurring during the stochastic phase are not simply gradual acquisition of an ESC-like expression frequency, rather they turn on and off at specific points in the process.

To further assess the confidence with which random change (uniform probability distribution) in gene expression during the stochastic phase can be rejected by our models (Gaussian probability distribution) is to compare the explanatory power of each model, as adjusted for the additional parameters required in each more progressively complex scenario. Figure 4D shows that while one normal distribution significantly improves AIC (lower is better), two normal (or even three normal - data not shown) do not add much explanatory power. One exception is for genes that exhibit transient expression changes, the fits for which are shown in Supplemental Figure 5. For this reason, we suggest that gene expression dynamics during the stochastic phase are best described as events occurring at specific points in the process, where most gene's expression dynamics are well described by a single normal probability distribution centered at the point of maximal rate of change. Genes that change at very specific points in the process have very tight probability distributions, while genes with less precise dynamics display broader probability distributions (approaching the uniform distribution of our null model).

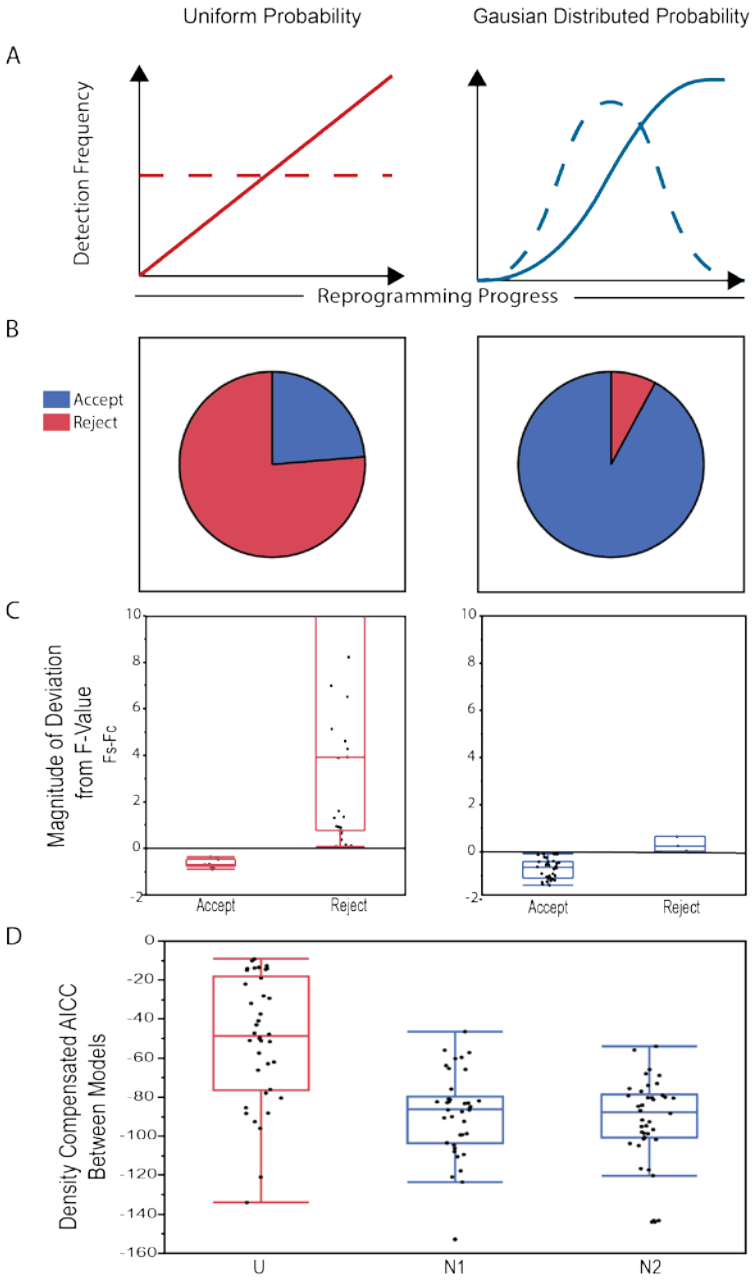


Figure 4: Rejection of a uniform model and justification of modeling using Gaussian distributions. (A) Predicted outcomes of gene expression probabilities associated with uniform (left panel) or Gaussian (right panel). Uniform and Gaussian probability distributions (dashed line) give rise to cumulative probabilities (solid line) that describe the population of cells at a given point in time. A Uniform probability results in the gradual activation / inactivation of a gene throughout the process, while Gaussian distributions suggest a bias in expression change towards a particular point in the process. (B) Pie charts showing the relative number of genes that accept or reject the Uniform (left panel) or Gaussian model (right panel) as determined using an F-statistic test. The strength with which these genes accept or reject each model is shown in (C). (D) Comparison of AICC value for all genes between the Uniform model and a Gaussian model using one or two normal distributions. While considerable improvement is observed for the Gaussian vs Uniform model, the addition of a second normal distribution does not dramatically improve model fit.

In order to compare dynamics between genes, we modeled each gene in our study using single Gaussian probability distributions as described above. All model fits are illustrated in the Supplemental Figure 6. One example fit is illustrated for CDH1 in Figure 5A. In this figure the black dots represent measured expression frequencies of CDH1 in sliding windows along the inferred reprograming trajectory. The red curve shows gene expression dynamics modeled as a Gaussian probability distribution fit to the experimental data and the blue line illustrates expression frequencies predicted by that probability curve.

When the dynamics of several genes are compared in one graph (Figure 5B-E) it is readily apparent that: 1) genes are activated or inactivated at different points during the reprogramming process; 2) genes have specific stringencies in their activation dynamics (some genes change at fairly specific stages, while others change over almost the entire course of the process); and 3) there is considerable overlap in the expression probabilities of individual genes. Most genes are activated or repressed with diffuse dynamics, while several (NANOG, CDH1, ZFP42, ZIC3 and OTX2) change at more specific stages of the reprogramming process. The diffuse dynamics and broad windows of activation observed for most pluripotency markers is consistent with the longitudinal observation that the expression of the surface antigens SSEA4 and TRA-1-60 in iPSC colonies are not strongly predictive of successful reprogramming events [49,51]. Taken together, this data strongly supports the hypothesis that rather than being a strictly ordered or strictly random process, the stochastic phase of reprogramming is an ordered probabilistic process. Seen in this light, prior ordered and random models can be coherently united [56,124,134].

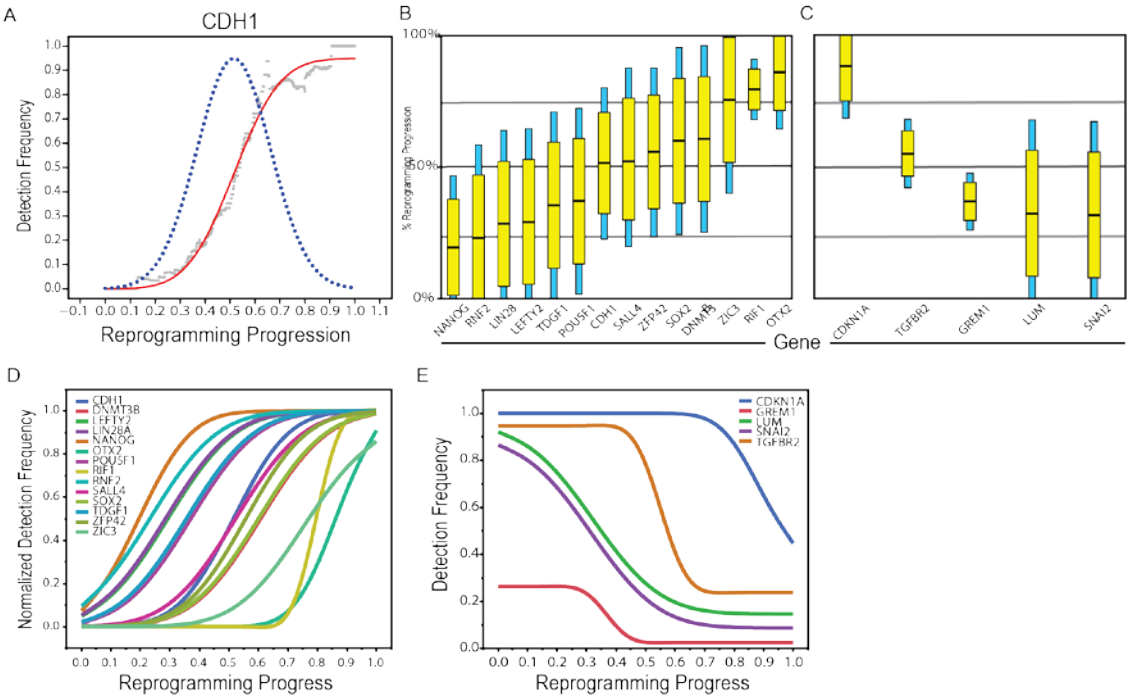


Figure 5: Modeling Gene Expression Dynamics with Gaussian Distributions. (A) Goodness of fit of a Gaussian model using activation of the CDH1 gene as an example. Gaussian distributions are represented as box and whisker plots for activating (B) and inactivating (C) genes. Yellow boxes and blue whiskers represent the 50% and 95% confidence intervals of the normal curve respectively, with the means shown as black lines. Cumulative distributions derived from the Gaussian model are overlaid for genes that are activated (D) or inactivated (E) during the course of reprogramming.

2.2.5 Changes in Pluripotency Gene Expression During the Stochastic Phase Do Not Show Hallmarks of Activation of the Pluripotency Gene Regulatory Network

Having observed ordered dynamics in the stochastic phase, we sought to determine if there was any indication that this order might arise from the partial activation of the endogenous pluripotency GRN. Current models suggest that partially reprogrammed cells enter a late, rapid deterministic phase that is controlled by activation of the endogenous pluripotency GRN and may be marked (in mouse cells) by the activation of the endogenous Sox2 locus[56,135]. Alternatively, order could emerge gradually or piecemeal during the stochastic phase. A hallmark of concerted gene regulation as exerted by a GRN, is strong correlation (or anti-correlation) between gene expression patterns [8,110,124]. Our model provides a powerful way to detect correlated gene expression that lies above the background correlations inherent during reprogramming (i.e. pluripotency markers all become expressed in fully reprogrammed cells). In this case, our null hypothesis is that during the stochastic phase there is no dependency between genes and that all correlation between gene expression in individual cells results simply from the increase in frequency of pluripotency markers as cells approach an ESC-like transcriptional profile. Our alternative hypothesis is that some pluripotency genes may be co-regulated (or cross-regulate) during the stochastic phase and would thus display higher than background levels of co-expression (as measured by correlation). To test these hypotheses we used the probability profiles of each gene to generate a simulated data set in which gene expression is determined only by the probability profile of each gene, with no dependencies between genes. The resulting dataset accurately recapitulates the individual dynamics of each gene in our dataset, and provides pairwise correlation values that are solely dependent upon the convergence of all pluripotency markers on uniform expression in ESC. We then compared pairwise correlations between genes

in this background data set with the real correlations observed in our single-cell transcript data (Figure 6).

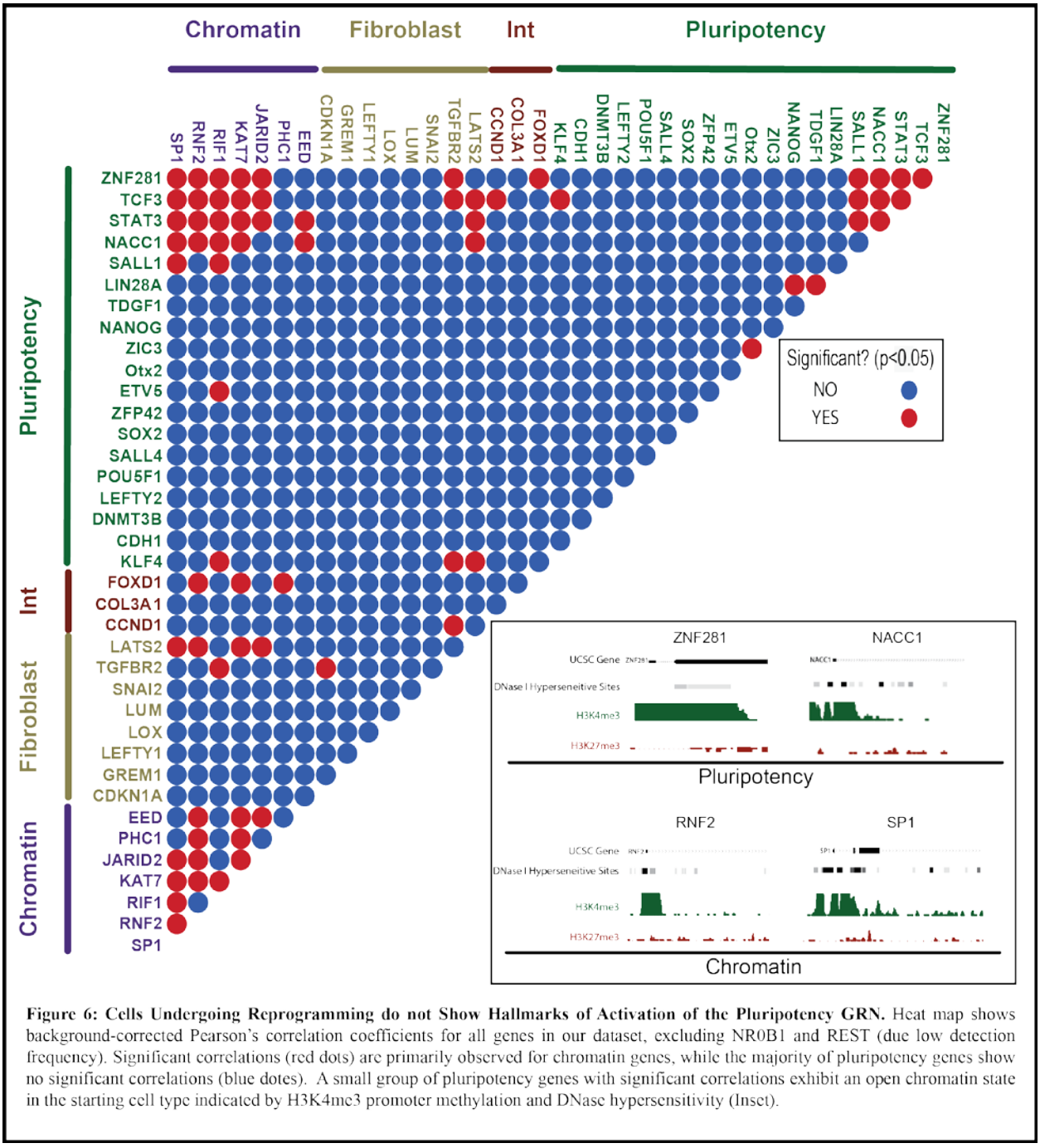


Figure 6: Cells Undergoing Reprogramming do not Show Hallmarks of Activation of the Pluripotency GRN. Heat map shows background-corrected Pearson's correlation coefficients for all genes in our dataset, excluding NR0B1 and REST (due low detection frequency). Significant correlations (red dots) are primarily observed for chromatin genes, while the majority of pluripotency genes show no significant correlations (blue dotes). A small group of pluripotency genes with significant correlations exhibit an open chromatin state in the starting cell type indicated by H3K4me3 promoter methylation and DNase hypersensitivity (Inset).

Interestingly, the only correlations we find rise above background expectations occur between a set of chromatin regulators that distinguish between entry into the productive trajectory and entry into the alternative trajectory (Figure 6). This coordinated activity is likely the result of activation of the c-MYC GRN, which is known to be activated upon OSKM induction, and is largely limited to genes with a permissive chromatin state in fibroblasts as is the case for many chromatin modifier genes [136,135] (Figure 6, inset). In contrast, none of the correlations between members of the pluripotency GRN rise above background expectations, despite their overall increase in expression frequency as cells approach an ESC-like expression profile. We therefore accept the null hypothesis: that despite the ordered activation of genes in the pluripotency GRN during the reprogramming process, there is no evidence for gradual or modular activation of the pluripotency GRN during the stochastic phase of reprogramming. An important corollary that follows from this result is that the dynamics of gene activation during the stochastic phase appear to depend only upon the local properties of each gene, rather than the sequential activation of precursors in the GRN. Of course, the numbers of genes we analyze in our study somewhat limits the power of this analysis, and a more comprehensive single-cell study measuring many more genes might uncover obligate relationships between genes that are not apparent in our core pluripotency GRN gene set.

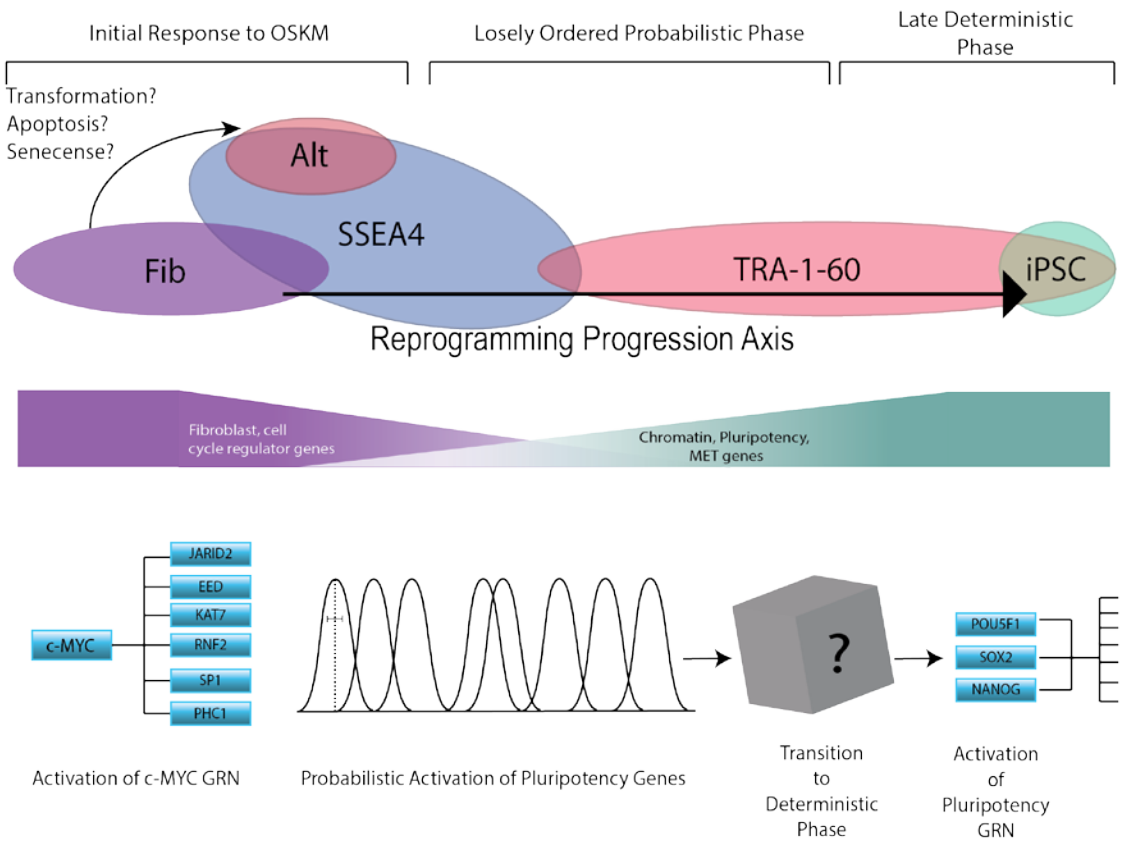


Figure 7: Combined Models Describing the Trajectories and Transcriptional Phenotypes Observed during Somatic Cell Reprogramming. (Top panel) Two trajectories are observed for cells undergoing reprogramming by OSKM, a productive trajectory leading towards pluripotency and an alternate trajectory away from fibroblast but not towards a hESC phenotype. The productive trajectory is characterized by the expression of the surface markers SSEA4 early and TRA-1-60 late in the process, and in general, involves the down-regulation of fibroblast and cell cycle-associated genes and simultaneous up-regulation of chromatin modifier and pluripotency genes. Putting our results in the context of the current literature, we observe an early wave of gene induction involving chromatin modifying enzymes and other loci with an open chromatin state that is likely the result of cMYC and KLF4 activity at these promoters. This initial wave is followed by a period of independent probabilistic gene expression, which we have model using a series independent Gaussian distributions. This probabilistic phase of pluripotency gene activation will eventually lead to an as yet unknown event that allows transition into the deterministic phase and the subsequent acquisition of pluripotency.

2.3 Discussion

In this study we present a rigorous single cell analysis of reprogramming in human cells and show that the stochastic phase of reprogramming of human fibroblasts by OSKM is an ordered probabilistic process which can be simply modeled using independent Gaussian distributions. An advantage of our approach lies in the fact that it makes no *a priori* assumptions about the progression of cells toward pluripotency, based on time or surface marker expression, both of which are poor indicators of reprogramming progress. In addition, the simplicity of our model and its exceptional fit to our observed expression dynamics provide a tractable framework for further dissecting the rate-limiting aspects of reprogramming. The results of this work also unify existing ordered and random models of the stochastic phase of reprogramming [11,49,52,55,110,123] and are consistent with observations from both population level and single cell studies of gene expression changes during reprogramming [9,49,56]. The ordered nature of the stochastic phase is readily apparent in the distinct, gene-specific expression dynamics we observe during reprogramming, while the probabilistic nature of the process is evident in broad gene-specific expression dynamics over large portions of the reprogramming trajectory (Figure 5 and Figure 7), and the apparently independent control of gene expression dynamics during the stochastic phase (Figure 6). These findings are consistent with a recent study by Tanabe et al. [57] that suggests the TRA-1-60+ phenotype is unstable and transcriptionally heterogeneous and that stabilization of the TRA-1-60+ population is a critical rate limiting step in reprogramming. Note we suggest retaining the term “stochastic” for this phase of the reprogramming process, in that stochastic can be used to describe ordered probabilistic events, and does not necessarily imply

complete randomness. The use of the term stochastic is especially appropriate given the independence of activation dynamics of key genes in the core pluripotency GRN.

One consequence of the independent activation of genes during reprogramming is that an extremely wide variety of cell states are present during the reprogramming process, which gives the overt appearance of disorder. Thus, while any given partially reprogrammed cell's gene expression pattern may appear to be random, the probabilities of expression of individual genes are clearly biased towards specific points along the reprogramming trajectory. One implication of these findings is that any single marker is unlikely to be effective at determining the extent to which a given cell has been reprogrammed [49,137].

We note that variations in the cell cycle could contribute to the transcriptional heterogeneity of a subset of genes in our dataset. However recent studies in hESC have shown that the transcription of genes associated with pluripotency does not fluctuate during the cell cycle [138], suggesting that cell cycle status is unlikely to have a major impact on our analysis of the activation of the pluripotency GRN. In addition, the persistence of cyclin transcripts throughout the cell cycle and their considerable post-transcriptional regulation in ESC's [139], precludes strong inference of cell cycle status from transcriptional measurement of a single cell-cycle regulator.

Another possible source of transcriptional heterogeneity between partially reprogrammed cells in our cultures could be the delivery of O, S, K, and M on individual vectors (as is standard in widely utilized human reprogramming protocols). However the broad agreement of expression dynamics over the course of reprogramming between our results using individual viral delivery, and those reported by Polo et al using an inducible, polycistronic construct in a clonal cell line,

suggests that viral heterogeneity does not fundamentally affect the order of gene expression dynamics, or the shape of the trajectory of cells undergoing the reprogramming process. Furthermore, the initial description of the highly heterogeneous nature of the stochastic phase by Buganim et al was also derived from data using clonal cells expressing OSKM from an inducible polycistronic OSKM construct. Thus, the stochastic nature of this phase does not appear to be a direct consequence of OSKM heterogeneity. However, these results do not rule out the possibility that each of the OSKM factors have distinct roles in various stages of the reprogramming process, nor that heterogeneity in OSKM content will be observed across the partially reprogrammed population of cells. Indeed, understanding the role of each factor in the reprogramming process and the critical window for the action of each represents an important goal of future work.

A likely explanation for the apparent lack of deterministic behavior during the stochastic phase may be the existence of as yet unidentified, gene-specific factors that restrict the rate of transcription activation by OSKM. One compelling candidate for these factors is the local chromatin architecture of the pluripotency genes in the starting somatic cell type. Indeed, epigenetic remodeling was implicated as a major rate limiting step in even the earliest days of somatic cell reprogramming using nuclear transfer [21,125] and is almost certainly one of the most important probabilistic events limiting the rate and efficiency of reprogramming. Many reports have experimentally validated this hypothesis by demonstrating that global chromatin reorganization is critical for successful reprogramming [5,50,66,69]. Because many of the required changes in chromatin state appear to occur in a slow and probabilistic fashion [43,140,141] it is likely that these changes limit the rate at which exogenous OSKM can activate the endogenous

pluripotency GRN thus limiting the efficiency and speed of reprogramming and endowing the majority of the process with stochastic dynamics.

Our finding, that enhanced expression of chromatin modifiers is a hallmark of entry into productive reprogramming complements several studies demonstrating that successful reprogramming requires the gradual erosion of epigenetic barriers to activation of the pluripotency GRN by OSKM [4,5,65,67,69]. This event is likely governed by the activity of c-MYC, which together with KLF4, acts early in reprogramming to activate loci with permissive chromatin states, including many chromatin modifier loci in fibroblasts [66,67]. In addition, many treatments known to enable chromatin remodeling have been shown to enhance the rate and/or efficiency of the reprogramming process [61,65,142,143], while, conversely, knocking down factors required for such epigenetic changes can inhibit or prevent successful reprogramming [61,65,111,142,144,145]. However, with the exception of some very early events [66,67] the order and precise identity of chromatin modifications required for successful reprogramming is not yet well known. By precisely describing and modeling gene expression dynamics during the stochastic phase the present study provides a quantitative framework for dissecting these key rate limiting steps and will enable the mechanistic dissection of interventions known to accelerate or enhance the efficiency of the reprogramming process.

Chapter 3 Comparison of Monocistronic and Polycistronic Reprogramming Methods in Two Cell Types

3.1 Introduction

Reprogramming terminally differentiated cells to a pluripotent state by exogenous expression of the Yamanaka factors OCT4, SOX2, KLF4 and c-MYC (OSKM) has the potential to revolutionize many aspects of modern medicine. However, despite years of research this process remains highly inefficient

and produces considerable cellular heterogeneity, problems that must be overcome before this technique can be used clinically. In the years following the first reports of OSKM-mediated reprogramming, several methods have been developed to reprogram cells to pluripotency in an effort to increase the efficiency and quality of iPSC generation. This includes using different methods of delivering the OSKM factors, including retroviral, lentiviral and episomal vectors, mRNA/miRNA transfection, as well as the use of additional factors such as NANOG, LIN28, SALL4 and others [6,31,38,77,146]. While many of these approaches have increased the efficiency and/or rate of reprogramming, at present it is unclear how these results manifest at the molecular level and such knowledge could provide insight into common mechanisms necessary to acquire pluripotency.

Many efforts to illuminate the molecular underpinnings of reprogramming have been complicated by the inefficiency and temporal asynchrony of the process. Only 0.01-1% of cells reaches the pluripotent state and do so at different rates over the course of a 3-4 week period. As a result, the majority of studies conducted to date that rely on bulk measurement of heterogeneous populations of cells are inherently biased towards analyzing unsuccessful reprogramming events. Thus, measurement of transcriptional or other events leading to pluripotency may be obscured. To overcome this limitation of bulk analysis our group and others have used single cell analysis and mathematical modeling to deconstruct the transcriptional and protein-level changes occurring in cells undergoing reprogramming [10,56,58,147]. By profiling individual cells en route to pluripotency we are better able to assess how the pluripotency gene regulatory network (GRN) becomes activated in response to the OSKM factors. Specifically, whether this activation happens as a series of concerted deterministic events, or occurs gradually over the length of the process. Equally important is the ability to measure what appear to be unsuccessful reprogramming events leading to trajectories other than pluripotency. Identification of common features in divergent cells can reveal events preventing cells from becoming iPSCs.

This previous work proposed a model where acquisition of pluripotency is primarily limited by an early probabilistic or stochastic phase. During this phase, genes associated with pluripotency activate

independently , lacking coordinated expression characteristic of a stable, pluripotent state [11,55,56,134]. This period can persist for a variable length of time, after which cells that have made the requisite epigenetic and transcriptional changes, activate the pluripotency gene regulatory network (GRN) and are stabilized in the iPSC state [56,91,148]. The stabilization of this network requires precisely controlled levels of OSKM expression[67,101]. Premature inactivation of exogenous OSKM fails to generate iPSCs[52,104] and conversely, failure to inactivate the OSKM cassette forces cells to an alternate ESC-like state, distinct from iPSC[149]. Given the relationship between factor stoichiometry and efficiency, it is important to assess how variations in reprogramming method impact the acquisition of pluripotency.

Comparison of monocistronic and polycistronic viral delivery of the 4 factors is of particular interest, as this remains the most widely utilized reprogramming strategy in the human system[2,150]. Monocistronic delivery allows flexibility in the stoichiometry of factor delivery due to random integration of the individual constructs, however many cells receive combinations of factors that are suboptimal for reprogramming, or may cause cells to take a different trajectory to the pluripotent state[101,151]. In contrast, polycistronic delivery fixes the ratio of factor delivery at 1:1:1:1, a ratio that may not be optimal for successful reprogramming, but guarantees that all transfected cells will carry a full complement of the reprogramming factors. In separate studies it has been demonstrated that mono and polycistronic systems reprogram cells at different efficiencies in mouse, 0.01% and 0.5%, respectively, and human 0.2% and 1.5%,respectively [2,28,32,33,145], however no direct comparison of these methods exists currently. Furthermore, species-specific differences in the molecular events leading to pluripotency exist between mouse and human[152], further complicating the comparison of these two techniques and underscoring the importance of studying reprogramming in human cells for clinical purposes.

The majority of studies to date have focused on reprogramming fibroblasts due to their simplicity of isolation, however dozens of other cell types have been successfully reprogrammed. The starting cell type has been demonstrated to have a significant effect on both the efficiency of the process, as well as the differentiation capacity of the resulting iPSCs. There is evidence to suggest that these effects are due

to unique epigenetic landscapes in different cell types which can affect the accessibility of pluripotency loci and consequently their ability to be activated by reprogramming factors[5,50,69,153]. This same epigenetic landscape also results in a 'memory' of the cell's starting identity, making differentiation back to the cell type of origin more efficient than generation of more therapeutically relevant alternatives[118,154,155]. Thus, the starting cell type can have a dramatic influence on the outcome of the reprogramming process but again, no analysis of whether this affects the acquisition of pluripotency has been performed.

In this study, we apply single cell transcript analysis to compare the transcriptional dynamics underlying the acquisition of pluripotency in monocistronic and polycistronic OSKM systems. These two delivery methods were tested in both MRC-5 and BJ fibroblasts. We demonstrate that polycistronic viral delivery produces significantly higher reprogramming efficiencies compared with monocistronic delivery and that this effect is due in part to premature inactivation of the individual O, S, K or M vectors in the monocistronic method. In addition, we show that the activation of key pluripotency loci such as NANOG, OCT4, LIN28 and DNMT3B occurs earlier in the polycistronic condition and that these cells progress more uniformly towards pluripotency. Finally, we compare polycistronic reprogramming between MRC-5 and BJ fibroblast cells and reveal that while the order of gene activation is similar between cell types, MRC-5 and BJ cells take divergent paths upon factor induction, followed by convergence later in the reprogramming process.

3.2 Results

3.2.1 Monocistronic and Polycistronic Reprogramming Efficiency

As an initial assessment of reprogramming efficiency between the monocistronic (Mono) and polycistronic (Poly) reprogramming methods, we analyzed the percent of SSEA4 single-positive (S+T-) and SSEA4/TRA-1-60 double-positive (S+T+) cells by FACS, markers associated with early and late reprogramming respectively[49,51,137]. We observe a significant enrichment of S+T- cells in the Poly, compared with the Mono condition which increased from a 2-fold difference at D4, to >8-fold at D14.

This trend is apparent for TRA-1-60+ cells as well, where Poly exhibits ~15-fold increase at both time points analyzed (Figure 8C). To determine whether the difference in SSEA4 and TRA-1-60 expression between conditions correlated with reprogramming efficiency, we stained and counted AP+ colonies at D21 and D28. Poly cells have 10-fold more AP+ colonies than Mono cells at D21, and this increase is even more pronounced at D28, the point at which colonies are typically picked to establish iPSCs (Figure 8D). This corresponds to an efficiency of ~5% and 0.5% for Poly and Mono, respectively. This is consistent with previous reports showing a 10-fold increase in reprogramming efficiency between the two conditions, albeit in separate studies[2,146,150]. In our experience, Mono colonies tend to be broad and cover more area than Poly colonies which are small and punctate. Example colonies are shown in Figure 1B. To ensure this difference in morphology did not skew our colony counting results we also measured the total area of the plate covered by AP+ cells. We still observe significantly higher AP-positivity in Poly compared with Mono (Figure 8E). These findings also hold true in BJ fibroblasts reprogrammed with Mono and Poly as shown in Supplemental Figure 7.

3.2.2 Experimental Design

In order to measure transcripts in individual cells at various points in the reprogramming process, we infected MRC-5 fibroblasts with a polycistronic construct containing all four Yamanaka factors (Poly). We then isolated cells by FACS at D4, D7, D11, D14 and D21 using the surface markers SSEA4 and TRA-1-60 to enrich for early (SSEA4+/TRA-1-60-) and late (SSEA4+/TRA-1-60+) reprogramming events, respectively (Figure 8A). These cells were sorted into 96-well PCR plates and processed through our single cell pipeline and qPCR was performed on the Fluidigm Biomark against a panel of 96 markers (Supplemental Table 4). In addition to profiling 80 reprogramming cells, we also profiled 16 MRC-5 fibroblasts and 32 H9 and H1 hESCs to represent the beginning and end points of the process, respectively (Supplemental Table 5). The Poly dataset was trimmed for comparison with our previously published MRC-5 Mono data which contains cells sampled at days 4, 7 and 14, measured for the

expression of 48 genes, all of which are present in the larger 96 gene panel analyzed in the Poly experiment[149].

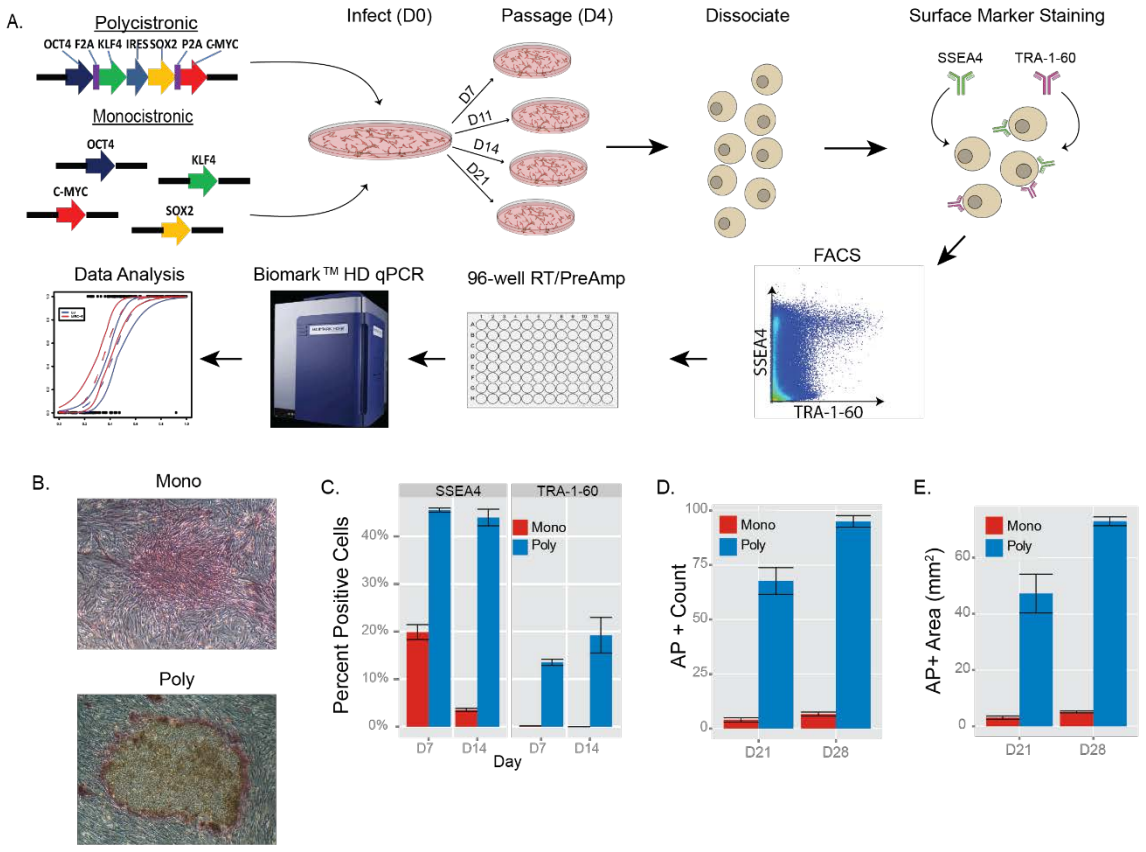


Figure 8: Polycistronic Delivery of OSKM Increases Efficiency Compared with Monocistronic Vectors. (A) Schematic summarizing somatic reprogramming experimental approach. Cells are infected with either Mono or Poly viruses and passaged on day 4. Cells are cultured in hESC media until the date of harvest where they are dissociated, stained with SSEA4 and TRA-1-60 antibodies and enriched single cells are sorted by FACS into 96-well plates. Following RT and PreAmp, mRNA expression is measured on the Fluidigm Biomark and analyzed in R. (B) Representative images of AP+ MRC-5 colonies reprogrammed with either monocistronic (top) or polycistronic (bottom) viruses. (C) Quantification of SSEA4+ / Tra1-60- and SSEA4+ / Tra1-60+ cells at Day 7 and Day 14 in both monocistronic and polycistronic conditions. Polycistronic reprogramming produces 10-fold more AP+ colonies in terms of both number (D) and area covered (E).

3.2.3 Progression of Individual Cells in Mono and Polycistronic Reprogramming

To visualize the progression of cells from the fibroblast to the pluripotent state, we used our previously described method of plotting cells based on their relative distance from both the fibroblast and hESC populations[149] (Figure 9A), overlaid with the surface markers used to isolate the cells. This method is agnostic to the time point of collection since progression through the reprogramming process is asynchronous and poorly correlated with time[58]. Using this approach, we observe a striking increase in progress of S+T- cells in the Poly condition, with some cells overlapping the hESC population, whereas S+T- Mono cells are only present in the first half of the reprogramming trajectory. We also notice that S+T+ Poly cells are very tightly clustered around the hESC population while S+T+ Mono cells span a large portion of the reprogramming trajectory. The increased progression in the Poly condition is accompanied by greater reprogramming synchrony compared with Mono, as revealed by the tighter distribution of cells along the reprogramming trajectory, maintained over time (Figure 9B). The distribution of S+T- Mono cells across the reprogramming trajectory broadens between D4 and D14, suggesting that either some cells are initiating reprogramming at the later time point, or that not all cells expressing SSEA4 are progressing through the process at the same rate, commonly referred to as variable latency. This is in contrast to Poly cells, which all progress towards an ESC-like transcriptional profile by D14.

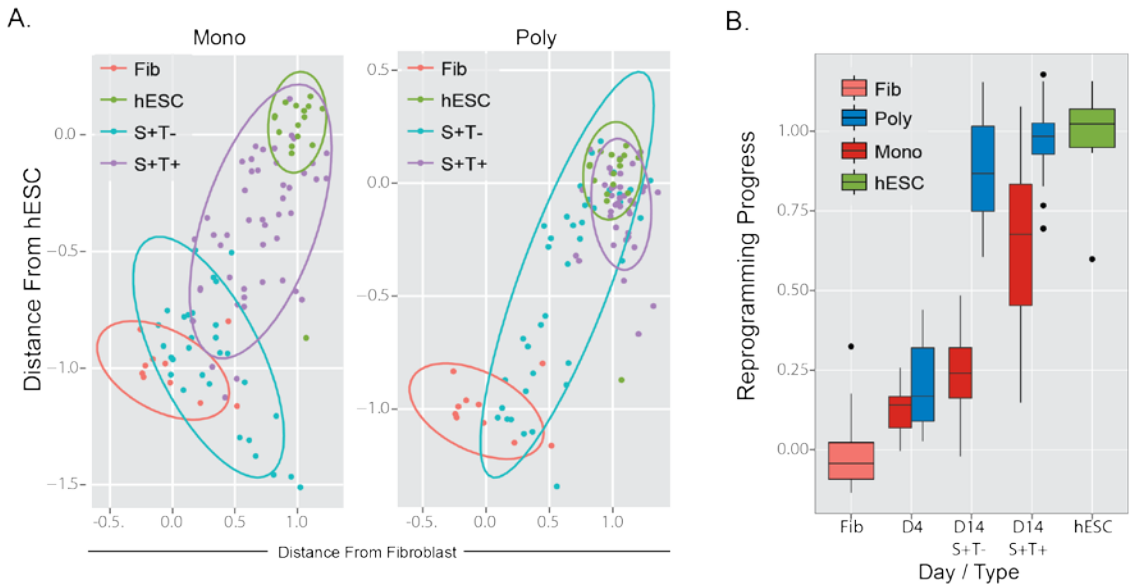


Figure 9: Polycistronic Reprogramming Exhibits Uniform Progression towards Pluripotency. (A) Reprogramming trajectory of mono (left) and poly (right) cells plotted by euclidean distance from fibroblast (x-axis) and hESC (y-axis). Fibroblasts and hESC are marked by pink and green ovals respectively, while SSEA4+ and TRA-1-60+ cells are shown in teal and purple, respectively. (B) Boxplot shows the progression of cells from each condition as a function of time. Both SSEA4+ and TRA-1-60+ Poly cells are more progressed and tightly distributed at D14 than comparable cells from Mono.

3.2.4 Generating a Logistic Regression Model

It has been proposed that the period of variable latency results from the stochastic and uncoordinated activation of pluripotency loci required to drive cells towards the pluripotent state. Because cells reprogrammed by the polycistronic method progress more uniformly towards the ESC state than monocistronic reprogramming cells, we asked whether the activation of pluripotency loci, or inactivation of fibroblast-associated loci, was more tightly coordinated in Poly cells. To this end, we improved upon our published method [149](Methods) to model the expression changes of genes along the reprogramming trajectory from fibroblast to hESC. Our new method gains higher accuracy while reducing the number of parameters to minimize bias compared with our previous model. We define the reprogramming trajectory by projecting cells into a 2-dimensional PCA space and fitting a polynomial curve through the dataset. We then find the shortest distance from each point to the curve and assign a value for that cell along the trajectory. These values are scaled between 0 and 1, representing the beginning and end of the process, respectively. For each gene in our dataset, we reduce the data to presence/absence calls and fit a logistic regression to the data, representing a continuous measure of the probability of detecting a given gene over the course of reprogramming. In addition, we gain information about when a gene is activated in the majority of samples and how rapidly that change occurs based on the point of greatest change in probability and the steepness of the curve. An example fit curve is shown in Figure 10A with dashed lines representing bootstrapped confidence intervals around the fit curve. Model fits for all genes can be found in Supplemental Figure 8. The expectation of this model is that conditions where gene expression changes rapidly correspond to a reprogramming process with fewer barriers to the transcriptional activation/inactivation events necessary to reach pluripotency and more closely resembles a deterministic rather than probabilistic process.

3.2.5 Assessment of Two Reprogramming Methods Using a Logistic Regression Model

To compare the model fits between conditions we separated activating and inactivating genes and plotted the point of greatest slope and the bootstrapped confidence intervals in Figure 10B. We notice

significantly earlier points of activation in Poly compared with Mono reprogramming (Figure 10C and E) for a subset of genes in our panel. This includes several key pluripotency loci such as POU5F1, NANOG and DNMT3B and may in part explain the improvement in reprogramming efficiency. Interestingly, despite earlier changes in gene expression in Poly cells, the order in which these genes are activated/inactivated is strongly correlated between the two conditions (Spearman's $r = 0.75$). This is further supported by the high correlation of gene loadings from independent PC analysis of Mono and Poly cells in the PC1 dimension (Figure 10D). The loadings provide a measure of when and how strongly each gene contributes to progression through the process and therefore, strong correlation in the loading scores indicates a common path to pluripotency for Mono and Poly reprogramming. It is important to note that while the two methods as a whole follow a similar path to the pluripotent state, the activation of a given gene remains a probabilistic event under our model. Thus, the order in which an individual cell activates/inactivates these loci is not fixed (ie is not deterministic). Consistent with this notion, we do not see a narrowing of the activation window as there is no significant difference in the slope of the activation curves between conditions (Figure 10F), nor do we observe increased correlation between genes in the Poly condition (Figure 11A-C). Taken together, these results suggest that while some pluripotency genes are activated earlier in the process in Poly reprogramming, coordinated GRN activity or deterministic behavior is not observed.

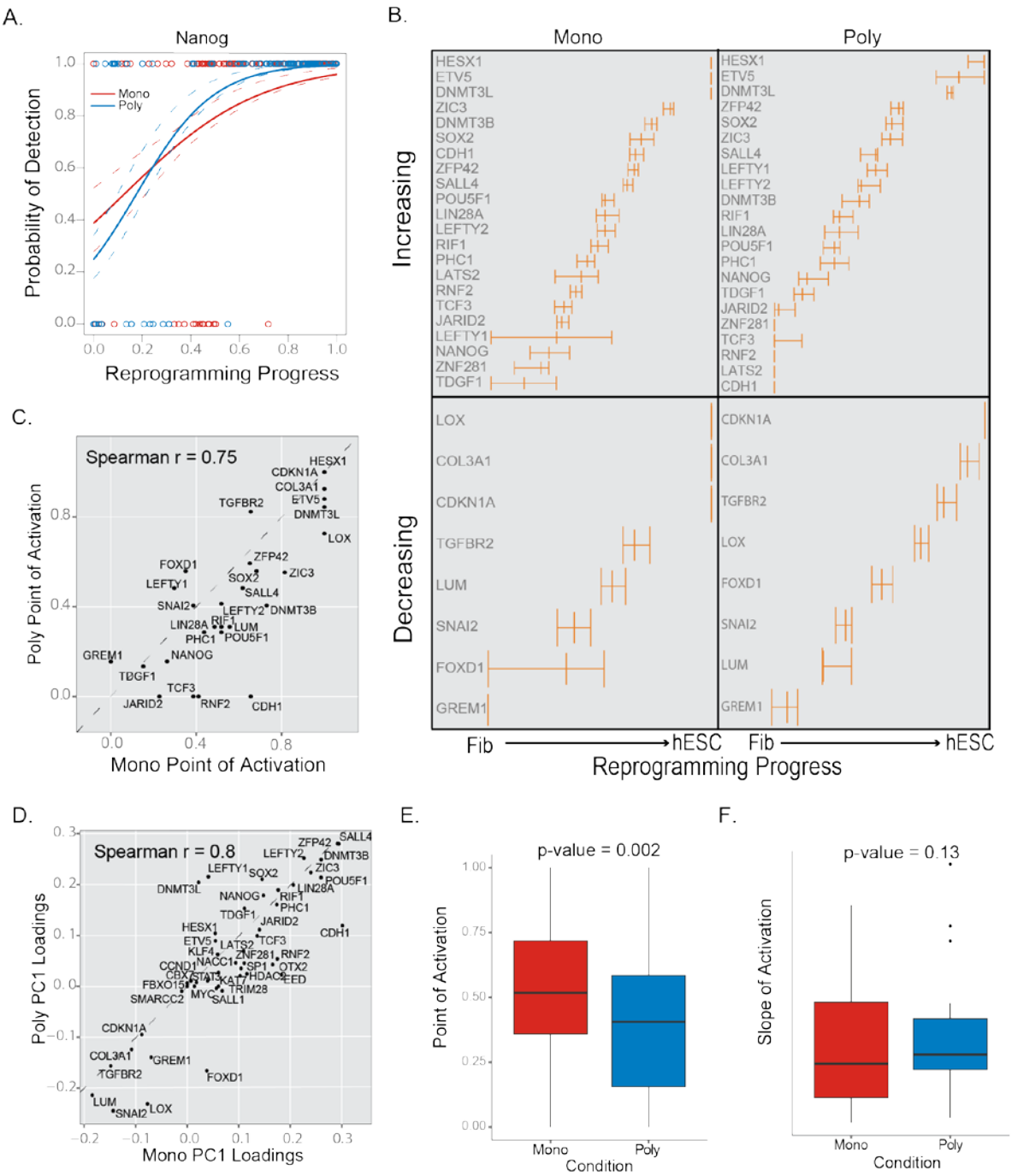


Figure 10: Delayed but Conserved Dynamics of Pluripotency Loci Activation in Monocistronic Reprogramming.(A) Example logistic regression fit of NANOG expression in Mono (red) and Poly (blue) reprogramming with bootstrapped confidence intervals (dashed lines). Points represent the binary expression data for each cell used in the model fitting. (B) Box and whisker plots depict point of greatest change with confidence intervals for activating and inactivating genes. The order of activation is highly correlated between the two conditions (C, Spearman = 0.75) as are the gene loadings from independent PCA analyses (D, Spearman = 0.8). The point of activation is shifted significantly earlier in Poly (E, paired t-test p = 0.002), however the rate of activation as given by the slope of the logistic curve is the same (F, paired t-test p = 0.13).

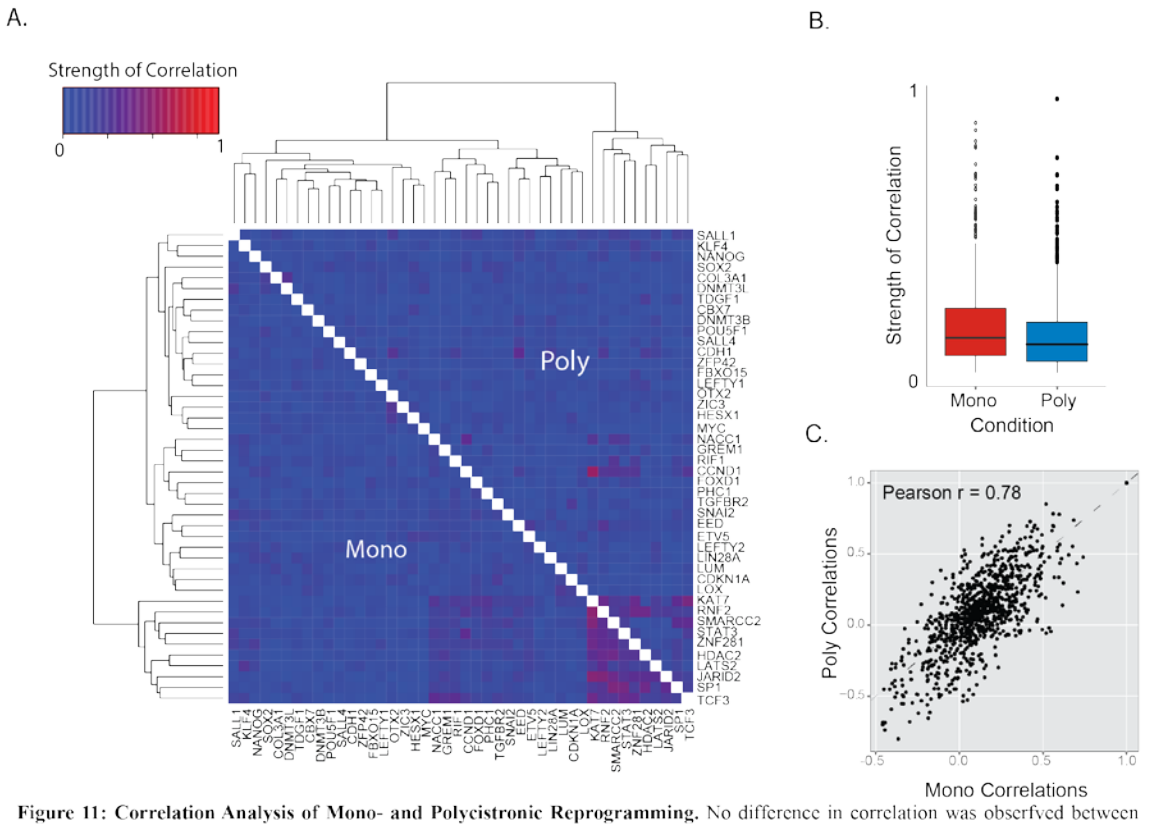


Figure 11: Correlation Analysis of Mono- and Polycistronic Reprogramming. No difference in correlation was observed between genes in the two conditions when correcting for background expectations (A,B) or using the raw correlation values (C, Pearson = 0.78).

3.2.6 Heterogeneous Expression of Exogenous OSKM Factors

Given that gene activation/inactivation is only slightly enhanced in polycistronic reprogramming and that the overall dynamics of the process appear similar between conditions, we looked for other factors contributing the poor efficiency of Mono reprogramming in both MRC-5 (Figure 12) and BJ (Supplemental Figure 9) fibroblasts. We hypothesized that OSKM heterogeneity could contribute to the low efficiency of Mono reprogramming since the factors are delivered on separate viral particles. To this end, we included SYBR primers targeting synthetic 3'-UTR regions present in the individual OSKM constructs (Supplemental Table 6) allowing us to measure the expression of the transgenes in all single cells collected for this experiment in addition to the 48-gene panel analyzed above. Looking at all four factors collectively it is apparent that a vast minority of cells express all four exogenous factors, with most cells expressing only one or two of the transgenes, including cells close to the hESC state (Fig 12A). Interestingly, cells that express the full complement of reprogramming factors are clustered early in the reprogramming trajectory, with no 4-factor containing cells progressed beyond the 50% mark. In contrast, most cells late in the trajectory express only one or two factors, typically either OCT4, MYC or both (Fig 12B). As expected, nearly all cells progressing along a previously described alternate trajectory away for both fibroblast and hESC lack expression of all reprogramming factors except MYC, illustrating the requirement of OSK expression for productive reprogramming.

3.2.7 Expression Patterns of Endogenous and Exogenous OSKM in Monocistronic Reprogramming

The considerable heterogeneity of transgene expression in Mono reprogramming cells led us to compare the expression of the endogenous (ENDO) and exogenous (EXO) copies of the OSKM factors to see if cells lacking transgene expression exhibited activation of the endogenous copy (Fig 12B). Nearly all Mono cells express exogenous MYC, while only three cells express the endogenous form. This can likely be attributed to the profound proliferative effects of high levels of MYC expression[4,66,136,135], which results in expansion of this population and increases the likelihood they are sampled in our experiment. In contrast, EXO-KLF4 is detected in very few cells, however the endogenous form is present in the

majority of samples. This is consistent with the role of KLF4 in promoting MET, an essential step in reprogramming that occurs late in the process[74,156,157]. Also, in agreement with previous reports that OCT4-high SOX2-low is an optimal factor stoichiometry for reprogramming[58,101], we notice the expression of exogenous OCT4 and SOX2 exhibit opposite patterns, with EXO-SOX2 expressing cells confined to the first half of the reprogramming trajectory while EXO-OCT4 cells persist until the later stages of reprogramming (Fig 12B). In addition, many late-reprogramming cells expressed both the ENDO and EXO forms of OCT4 further supporting this notion. Surprisingly, ~50% of late reprogramming cells fail to express either ENDO or EXO-SOX2. In mouse, SOX2 is required for entry into the deterministic phase and stabilization of the pluripotent state. The absence of SOX2 in some of our late reprogramming cells begs the question of whether or not these cells will successfully reprogram. In addition, it is unclear whether these cells were capable of progressing to the late stages of reprogramming in the absence of SOX2 expression, or if the SOX2 virus was prematurely inactivated prior to completing the process.

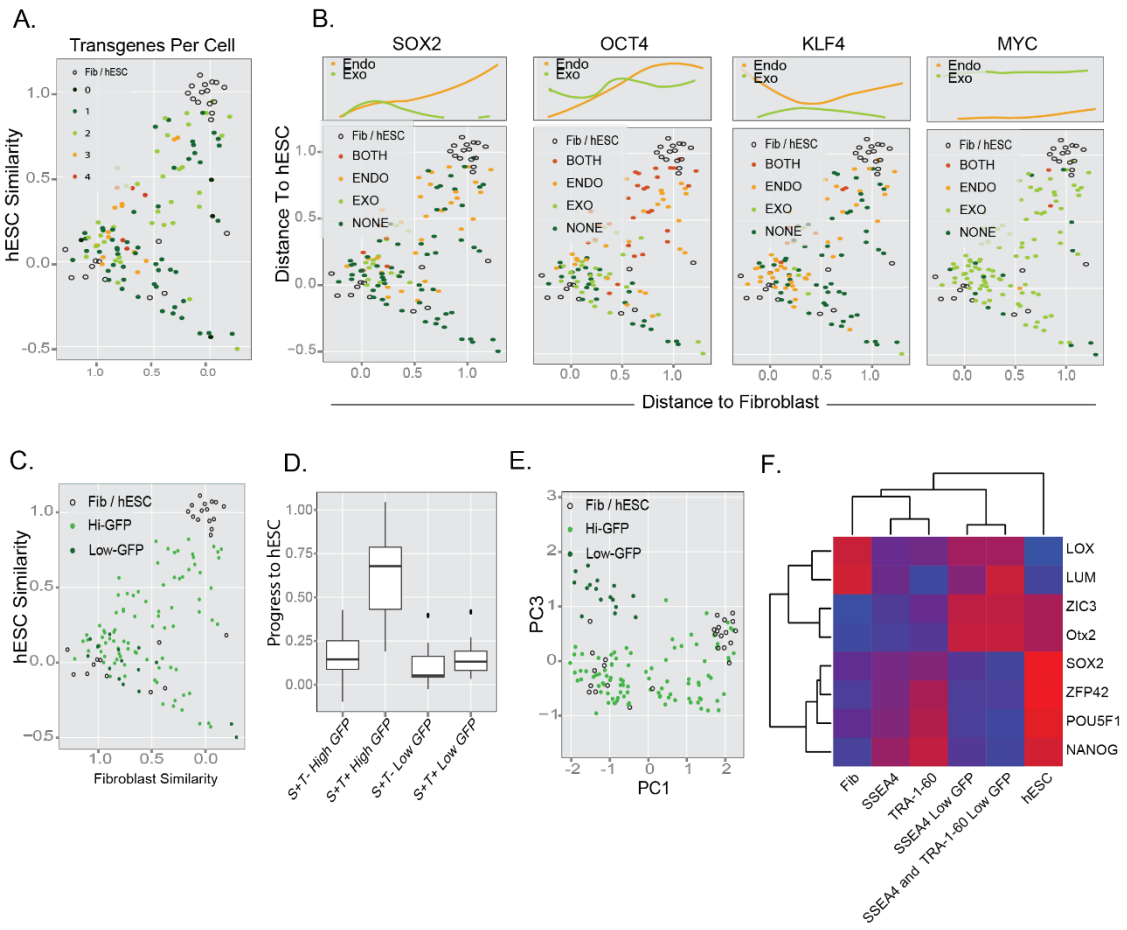


Figure 12: Expression of OSKM transgenes is Heterogeneous in Monocistronic Reprogramming. (A) Reprogramming trajectory overlaid with number of transgenes expressed within each cell as determined by SYBR green qPCR. Few cells express all four factors while most contain only one or two. (B) Trajectory plots with total exogenous and/or endogenous OSKM content displayed. Splines of endogenous and exogenous factor content along the trajectory are shown above. Mapping of High- and Low-GFP expressing cells on reprogramming trajectory demonstrates Low-GFP cells exhibit a fibroblast-like expression pattern (C) and fail to progress towards pluripotency when compared with High-GFP cells (D). Principal Component Analysis reveals Low-GFP cells are distinct from all other cells in our experiment (E). This is due to the expression of late reprogramming genes ZIC3 and OTX2 despite failure to activate core pluripotency genes and the persistence of fibroblast gene expression as shown in (F).

3.2.8 Transcriptional Analysis of Low-GFP Reprogramming Cells

To address this question we took advantage of the fact that our monocistronic OSKM vectors co-express GFP along with each reprogramming factor, allowing selection of cells with low viral content as indicated by low-GFP expression. These cells were sorted by FACS and we assessed their transcriptional profile using our 48-gene panel. When added to our reprogramming trajectory (Fig 12C), these cells look nearly identical to fibroblasts and both S+T- and S+T+ Low-GFP cells exhibit impaired progression compared with their High-GFP counterparts (Fig 12D). This would suggest these cells failed to respond to the OSKM cocktail however, Principal Component Analysis reveals the Low-GFP cells are a distinct population and separate from non-reprogrammed fibroblasts along the PC3 axis (Fig 12E). This separation is due to the expression of the late reprogramming genes ZIC3 and OTX2 in these cells, despite the failure to activate core pluripotency loci including OCT4, NANOG, SOX2 and ZFP42 (REX1) and the persistence of fibroblast gene expression (LOX and LUM) (Fig 12F). The expression of ZIC3 and OTX2 in the Low-GFP population indicates that these cells have reached the late stages of reprogramming, but collapsed back to a Fibroblast-like state due to premature loss of transgene expression. Alternatively, cells may have only been infected with a subset of the reprogramming factors, following a reprogramming trajectory not typical of cells receiving the full complement of OSKM.

3.2.9 Reprogramming Using Three-Factor Combinations OSKM

We tested this hypothesis explicitly by generating all possible 3-factor combinations (SKM, OKM, OSM and OSK), removing each factor from the STEMCCA polycistronic vector and measuring the reprogramming trajectory of infected cells. Attempts to reprogram cells with any of these 3-factor combinations failed to produce any AP+ colonies and resulted in a significant reduction in SSEA4+ cells (Supplemental Figure 10A). These cells also fail to productively reprogram, moving away from fibroblast but not towards hESC (Fig 13A) as evidenced by minimal expression of both fibroblast and pluripotency genes in our panel. This reduced expression is not due to lack of cell viability or the induction of apoptosis however, indicating these cells follow a trajectory that cannot be measured using our existing

marker set (Supplemental Figure 10B). K-means clustering of the 3-factor reprogramming conditions along with the MRC-5 Mono dataset reveals that this trajectory is equivalent to the Alternate trajectory we identified previously[149] and suggests that Mono cells in this Alternate group are also cells that failed to receive the full complement of OSKM (Fig 13C).

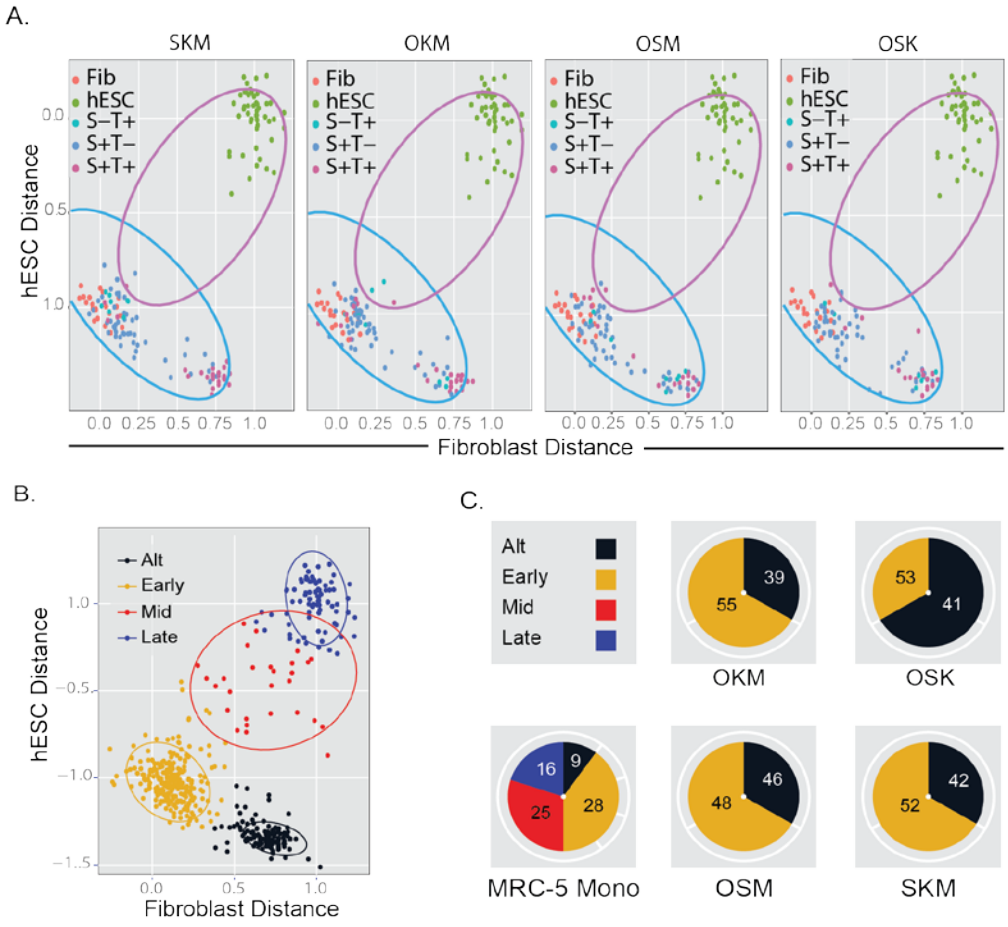


Figure 13: Premature Inactivation of the Individual OSKM Factors Prevents Productive Reprogramming. (A) Reprogramming trajectory of cells reprogrammed with polycistronic virus carrying all possible 3-factor combinations (SKM, OKM, OSM and OSK), overlaid with the S+T- and S+T+ coverage ellipses from MRC-5 Mono reprogramming. All cells move away from fibroblast but not towards hESC indicating a failure to successfully reprogram. (B) Overlaying all 3 factor reprogramming conditions with MRC-5 monocistronic reprogramming shows that 3-factor reprogramming sends cells to the 'Alternate' group, defined in our previous publication. Groups were determined using kmeans clustering with k=4.Other groups include Early, Mid and Late groups along the path to pluripotency. (C) Quantification of the number of cells found in each group described in (B).

3.2.10 Trajectory of BJ and MRC-5 Fibroblast Lines in Polycistronic Reprogramming

Having determined that premature inactivation of the individual OSKM factors is a major weakness of Mono reprogramming, we turned our attention to comparing the dynamics of pluripotency gene expression between two fibroblast cell lines, BJ and MRC-5, using the polycistronic method. To this end, we compared MRC-5 and BJ cells reprogrammed with polycistronic OSKM and analyzed the expression of 96 genes (as described above). In our system, BJ fibroblasts exhibit ~3x greater efficiency than MRC-5, as determined by the number of AP+ colonies at days 7, 14, and 21 (Supplemental Figure 7). Thus, we next sought to determine whether this difference in efficiency was evident in the trajectories of each reprogramming cell type, or in the expression of individual genes. Comparing the progression of cells over the time course of reprogramming shows little difference in the S+T- and S+T+ cells from both cell types and, as expected, S+T+ cells progress uniformly towards pluripotency, while S+T- exhibit a larger distribution due to variable latency(Fig 14 A-B). In addition, visualizing the trajectories in PCA space shows that the majority of the process looks identical between cell types in the first 2 PC dimensions, which cumulatively capture ~35% of the variance(Fig 14C). Including the PC3 dimension (5% variance) however, reveals a slight divergence of the two trajectories early in the process, followed by convergence near the hESC state (Fig 14D). This observation is reiterated by plotting each cell type side-by-side in its own PC space. An initial comparison of the PCA shows both cell types exhibit a similar distribution of reprogramming intermediates as determined by the amount of variation captured by each PC dimension (Fig 14E). Comparison of the gene loadings between the MRC-5 and BJ PC analyses reveals strong correlation in PC1 (Spearman r = 0.95) and PC2 (Spearman r = 0.72) dimensions, while correlation in PC3 is weak (Spearman r = 0.59) (Fig 14F). This again suggests nearly identical gene expression dynamics between the two cell types with subtle differences.

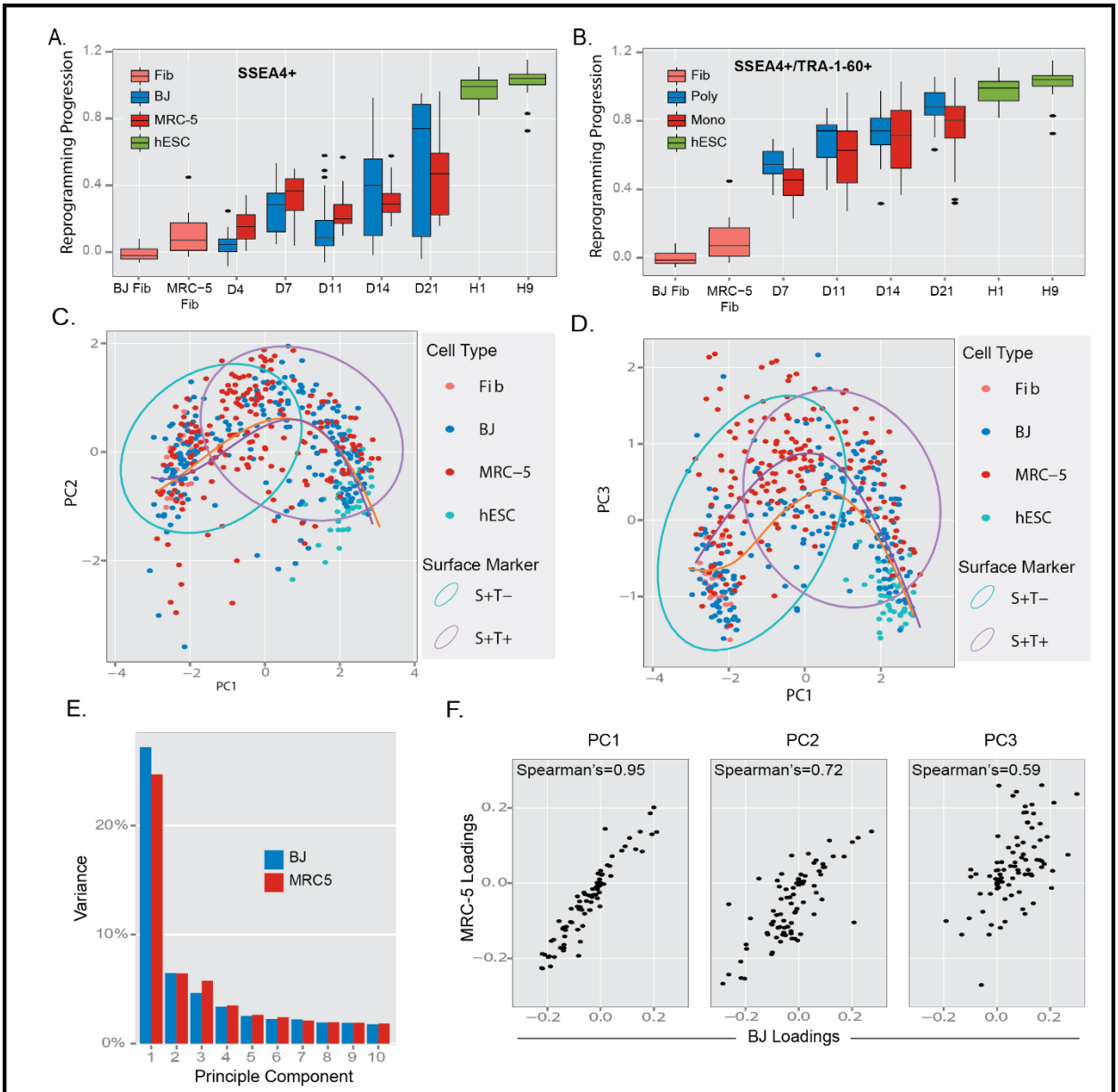


Figure 14: MRC-5 and BJ Fibroblast Trajectories Diverge Early and Converge Late in Reprogramming. Plotting the reprogramming progression of each cell type as a function of time shows both MRC-5 and BJ S+T+ cells are broadly distributed along the trajectory but progress towards hESC over time (A). The same analysis of S+T+ cells shows a tighter distribution of cells at all time points regardless of cell type (B). PCA shows MRC-5 and BJ fibroblast follow nearly identical reprogramming trajectories in the first two components (C), however PC3 reveals a divergence of the trajectories early in reprogramming followed by convergence later in the process (D). This separation is minimal as PC3 only captures ~5% of the variability (E). (F) Comparison of the gene loadings between the MRC-5 and BJ PC analyses in PC1(Spearman r = 0.95), PC2(Spearman r = 0.72) and PC3 (Spearman r = 0.59) demonstrates that the same genes define the trajectories in both cell types, suggesting a common route to pluripotency

3.2.11 Gene Expression Dynamics in BJ and MRC-5 during Polycistronic Reprogramming

To determine what genes specifically contribute to the slight differences in the reprogramming trajectories of BJ and MRC-5 fibroblasts, we utilized our model to compare the point of activation of genes between conditions. Example fit curves for increasing and decreasing genes are shown in Fig 15A and E, respectively. The complete set of fit curves is displayed in Supplemental Figure 11. We again use box and whisker plots to represent the mean and bootstrapped confidence intervals of the fit curves for both activating and inactivating genes (Fig 15B and F). A delay in the activation of several genes (Fig 15B and D, red highlight) is immediately apparent in MRC-5 cells early in the trajectory. These include key pluripotency genes such as NANOG, POU5F1, DNMT3B and LIN28. As expected, genes late in the trajectory exhibit nearly identical activation patterns, consistent with the observation that the trajectories converge near the ESC state. We also observe delayed inactivation of the fibroblast marker LOX and an inhibitor of MET, SNAI2, in MRC-5 cells (Fig 15F and H, red highlight). For both activating and inactivating genes, we see the same degree of correlation between genes in both conditions (Fig 15C and G), indicating the interactions between genes are consistent in BJ and MRC-5 reprogramming.

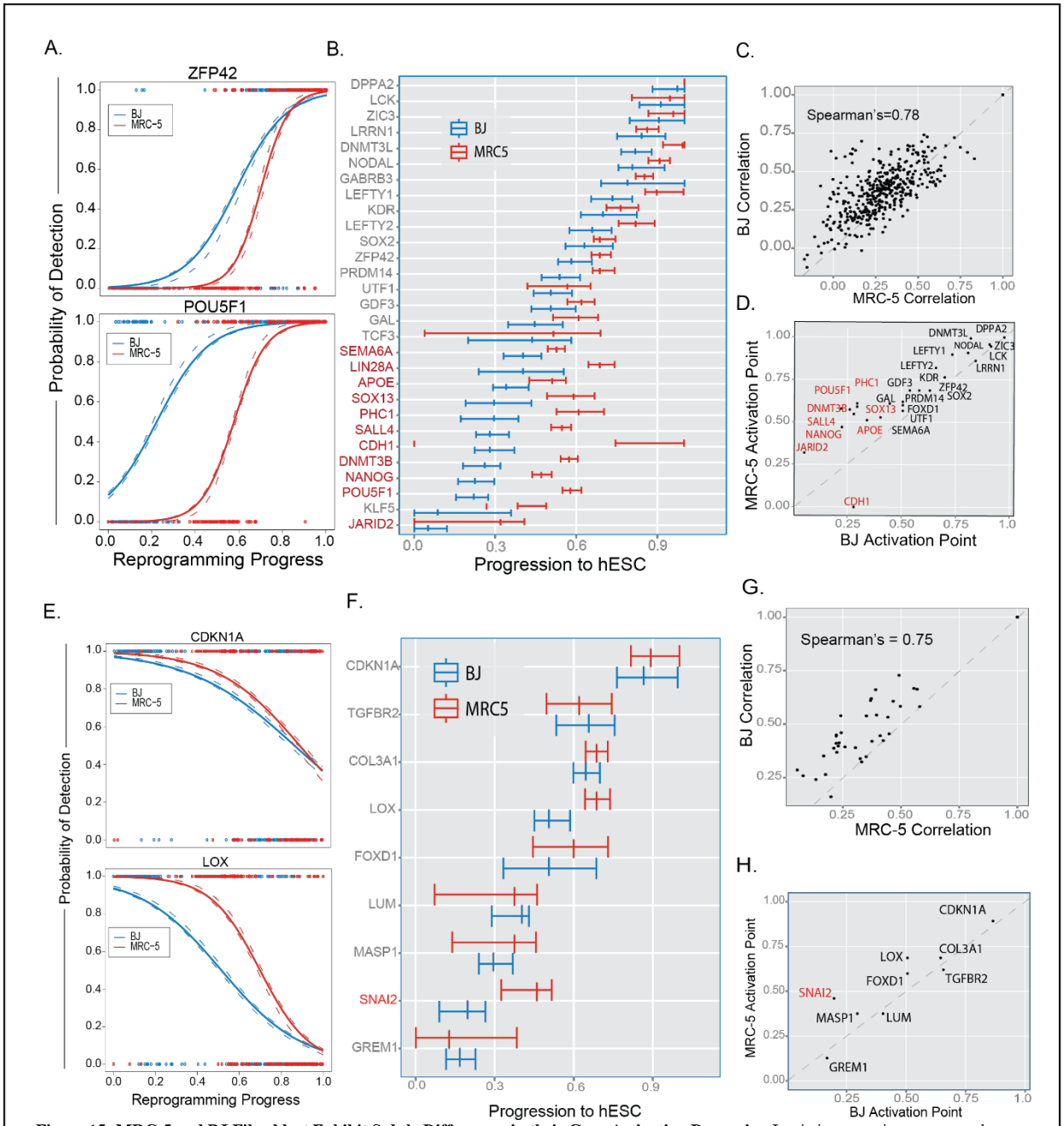


Figure 15: MRC-5 and BJ Fibroblast Exhibit Subtle Differences in their Gene Activation Dynamics. Logistic regressions were used to model the probability of detecting a given gene along the reprogramming trajectory. Example model fits for activating and inactivating genes are shown in (A) and (E), respectively. Box and whiskers plot are used to represent the mean and bootstrapped confidence intervals of the fit curves for both activating (B) and inactivating genes (F) and indicate small group of genes (red highlights) exhibit a delay in expression in MRC-5 fibroblasts. However, the overall order of activation/inactivation is highly correlated between cell types (D and H). We also observe a similar correlation between genes in both conditions (C and G), although decreasing genes appear more tightly regulated in BJ than MRC-5 (F).

3.3 Discussion

In this study we performed a side-by-side comparison of Polycistronic and Monocistronic reprogramming in human fibroblast cells. Our primary finding is that reprogramming by the Polycistronic method results in a 10-fold increase in efficiency over Monocistronic reprogramming and that this difference is due in part to the premature inactivation of the individual OSKM factors in the Monocistronic condition. While it has been previously documented that factor expression decreases over the course of reprogramming[52,158], this was thought to represent cells entering the pluripotent state. Inactivation of the reprogramming factors is generally considered to be a late event in the reprogramming process and is associated with the generation of stable iPSC lines[49,97]. Our study represents the first report that premature inactivation of OSKM can occur amidst productive reprogramming and results in a collapse of cells back to a fibroblast-like state. These cells exhibit signatures of productive reprogramming, in particular, the expression of the late reprogramming genes ZIC3 and OTX2, however they fail to activate the core pluripotency circuitry and continue to express markers of the fibroblast state. We exclude the possibility that these failed reprogramming events arise from cells receiving an incomplete complement of OSKM by demonstrating that cells lacking any one of the 4 factors fail to reprogram. Our analysis of transgene content in productively reprogramming cells demonstrates that a particular stoichiometry is optimal for pushing cells towards the pluripotent state. Specifically, high levels of OCT4 and low levels of SOX2 are favored in cells that have reached an ESC-like transcriptional profile, while KLF4 expression is consistent throughout the process, in line with previous reports[58,151]. The robust detection of MYC throughout the process is expected due to the rapid expansion of MYC expressing cells and this increase in cell cycling has been shown to greatly enhance the efficiency of reprogramming[135]. Importantly, it has been previously shown that there is no selective inactivation of any of the 4 reprogramming factors in iPSCs[151] and thus, the differences we see in OSKM content reflect a

bias for particular combinations selected for in late reprogramming events (ie. TRA-1-60+ cells) and do not result preferential inactivation of any one factor.

By examining the trajectories followed by cells reprogrammed with either Mono or Poly viruses, we notice that cells from the Mono condition exhibit a delay in the activation of several pluripotency loci including POU5F1, NANOG, DNMT3B and LIN28 compared with Poly reprogramming. We also observe a similar delay in the activation of these same loci in MRC-5 versus BJ fibroblasts. The period preceding the activation of the core pluripotency circuitry is referred to as latency, which is thought to be a major rate-limiting step in generating iPSCs. Our observation that latency is increased in the two conditions with relatively lower efficiencies (Mono and MRC-5) lends support to this notion and to our knowledge, is the first time this phenomenon has been measured between distinct conditions. It is commonly believed that latency results from the remodeling of the epigenetic landscape to allow the activation of pluripotency loci and it is expected that different factor stoichiometries or starting cell types would affect the rate at which this occurs[5,50,66,101,125,158]. If true, this would imply BJ cells have a more permissive chromatin state at some loci than MRC-5 fibroblasts, facilitating their activation. A rigorous comparison of reprogramming in cell types with divergent chromatin states would directly address this hypothesis and represents an important future direction of this work.

Despite the differences observed early in the process, cells from all conditions activate pluripotency genes in a similar probabilistic order following the period of latency, suggesting a common mechanism in establishing pluripotency. The convergence on a common trajectory late in the process resembles the deterministic phase described by Buganim et al, however in our system these gene activation events are independent and probabilistic and thus we don't believe that our observation represents a strictly deterministic process. It is also somewhat surprising that different reprogramming conditions result in similar reprogramming trajectories given the variation in quality and differentiation potential of iPSCs derived from these different methods[31,159]. This implies that variation in iPSC phenotype results not from differences in how the pluripotency network is established but is likely due to

differences not analyzed in our study. This could result from differential expression genes whose expression alter iPSC phenotype, or could occur at the epigenetic level, as has been shown for the Dio-Dlk3 locus in mouse[101].

In contrast to the similarities between BJ and MRC-5 at the transcript level, the surface markers SSEA4 and TRA-1-60 label slightly different populations between cell types. While the reason for this is unclear, it illustrates the impact of cell type on selecting informative biomarkers to isolate cells from different parts of the reprogramming process. SSEA4 and TRA-1-60 are also unique in that they are the only markers examined in our study that exhibit a strict order of activation; SSEA4 turns on before TRA-1-60 and all TRA-1-60+ cells are also SSEA4+. This is very different from the probabilistic order observed at the transcript level for the majority of genes in our panel and it begs the question as to whether the process is more highly ordered or deterministic at the level of protein expression. Recently several groups have begun to explore the dynamics of the proteome during reprogramming, however this has not yet been coupled with transcriptional analysis and all studies to date have been performed in mouse. Indeed, this remains an important area of study in the field of reprogramming.

Taken together, our study demonstrates that different reprogramming paradigms have the greatest effect early in the process during the period of variable latency. Once pluripotency gene expression is initiated, cells from all conditions follow a similar path to the pluripotent state as long as OSKM expression is maintained. The establishment and maintenance of factor expression is a critical challenge in monocistronic reprogramming as not all cells receive the full OSKM cocktail, nor do they maintain their expression throughout the process in all cells. This is a key advantage of the polycistronic method which ensures delivery of all four factors on a single construct. Our ability to make these conclusions relies on the single-cell resolution of our analysis and the comparison between multiple reprogramming conditions and demonstrates the need for rigorous comparison between protocols in order to determine the effect of procedural variables on the reprogramming process.

Chapter 4 Protein-Level Analysis of Human Reprogramming

4.1 Introduction

Reprogramming of adult somatic cells to a pluripotent embryonic-like state can be achieved through overexpression of the four Yamanaka factors OCT4, SOX2, KLF4 and c-MYC and has been successfully performed in both human and mouse [2,160]. The development of human reprogramming in particular represents a promising means to obtain patient-matched stem cells for therapeutic purposes and for the modeling of disease states where the cells of interest are not readily accessible such as Alzheimer's and Parkinson's disease. The great potential of this technique is limited however, by the low efficiency of the process as well as the variability in the quality and differentiation potential of the resulting iPSCs[46,118,119]. This has spurred a flurry of research in recent years to define the molecular changes occurring as cells acquire a pluripotent phenotype with the goal of increasing the efficiency and reproducibility of the process.

Many efforts to understand the reprogramming process have focused on transcriptional dynamics both in bulk samples as well as in single cells [9,38,56,149]. This work from our group and others has resulted in a two phase model of reprogramming whereby the expression of transcripts early in the process is loosely ordered and probabilistic, with genes behaving independently of one another. Late in the process mRNA expression shows increased dependency and coordination, characteristic of a gene regulatory network (GRN) that stabilizes cells in the pluripotent state. It is thought that the early probabilistic expression of mRNAs is a rate-limiting step in reprogramming as this phase is protracted and only a minority of cells progress past this point in the process. In addition, compounds that increase the rate and efficiency of reprogramming result in greater coordination between transcripts, suggesting a more rapid and robust activation of the pluripotency GRN facilitates reprogramming[161].

The independent behavior of transcripts during the probabilistic phase of reprogramming is often attributed to locus-specific epigenetic barriers that prevent the activation of individual loci despite the expression of other members of the GRN and indeed, the rate limiting effects of the epigenetic landscape

on reprogramming are well established[7,9,63,65,68]. However, the expression of members of the pluripotent GRN is largely based on mRNA expression analysis and it is unclear whether these transcripts are translated into protein, which is requisite to exert their function. At the time this project was initiated, very little was known about protein expression during the course of reprogramming. Reprogramming systems in mouse utilizing GFP reporters for Nanog and Oct4 demonstrated that expression of these proteins is a late event in the process, however detection of these mRNAs occurs as early as Day 4 of reprogramming[50,51,113,149,162]. In addition, we and others had observed that two surface markers commonly used to enrich for reprogramming intermediates, SSEA-4 and TRA-1-60 are expressed in a dependent manner (i.e. SSEA-4 is always activated before TRA-1-60). This raised the question of whether protein expression is delayed compared with mRNA expression and if once expressed, proteins exhibit more coordinated and dependent behavior than their cognate transcripts.

In the year since this project began, three groups have performed protein expression analysis on mouse cells during reprogramming[58,59]. Zunder et. al. used mass cytometry to examine the expression of 37 proteins in secondary reprogramming mouse fibroblasts and were able to define a number of intermediate states of reprogramming that they present as evidence for reprogramming being more ordered at the protein level[58,59]. However, it is clear from their data that cells exist in a continuum between these states and no rigorous analysis of correlation or dependency between the proteins was performed. In addition, the mass cytometry data stands on its own and is not directly compared to the expression of the corresponding mRNAs. In two separate studies, Benevento et. al. and Hansson et al. performed a more direct comparison of mRNA and protein using LC/MS-MS to profile the entire proteome in bulk reprogramming samples at different time points while collecting mRNA-seq data from the same experiment in parallel[84,85]. These data demonstrate that proteins involved in the same process or that exist is multi-subunit complexes are highly co-regulated while those involved in opposing processes exhibit an inverse expression pattern. Interestingly however, when these proteins are compared with their corresponding mRNA transcripts, they only exhibit strong correlations early in the process, with a greater

discordance between the two late in reprogramming. Taken together, these studies provide strong evidence that post-transcriptional mechanisms exist as cells approach pluripotency and may indeed be a key step for generating iPSCs.

Because these data were generated in the mouse system and in bulk samples, it remains unclear how these data translate to human reprogramming and whether these observations apply to the minority of cells undergoing productive reprogramming. To this end, we utilized CyTOF technology to profile the expression of 37 proteins at time points throughout the reprogramming process and compared these results with our previously generate single-cell mRNA expression data. Here we directly compare the reprogramming trajectories of cells at the mRNA and protein level and identify genes putatively subjected to post-transcriptional regulation. Additionally, we construct mathematical models to assess the degree of dependency between mRNAs and proteins and provide evidence for tighter regulation at the protein level. This dataset, while limited in scope, suggests translation of stochastic transcripts into protein may also be rate limiting in human reprogramming and that in general, the reprogramming process is more ordered at the protein level. This work lays the foundation for a more extensive study of mRNA and protein expression dynamics in single human cells during reprogramming in future work.

4.2 Results

4.2.1 Experimental Design

To profile the dynamics of protein expression during reprogramming, BJ fibroblasts were infected with polycistronic lentivirus containing the OSKM reprogramming factors. Cells were harvested at days 4, 7, 11, 14 and 21, the same time points used in our previous mRNA analysis. We also collected BJ fibroblasts and H9 hESCs to represent the beginning and end points of the process, respectively. Following harvest, cells were stained with metal-labeled antibodies, barcoded and run in parallel on the Helios™ CyTOF instrument to measure the expression of 27 proteins in ~1 million individual cells. Barcoding allows different samples to be combined into a single staining reaction and run simultaneously

on the CyTOF to minimize technical variation (Figure 16). These samples are then computationally de-multiplexed after acquisition using the manufacturer's software. The proteins analyzed were chosen from our 96-gene mRNA expression panel where antibodies were available (Supplemental Table 7). Preference was given to antibodies previously validated for use in flow cytometry to simplify their optimization for CyTOF. Antibodies were further validated and titrated on BJ fibroblasts and H9 hESCs to confirm their specificity and expression in the appropriate cell type. This resulted in 13 targets that could be directly compared to the mRNA dataset. Additional targets were chosen based on the work of Zunder et al.[58,59] and include the intermediate markers CD73, CD49d, CD200 and PDGFRa, as well as EpCAM, an epithelial marker expressed in hESCs.

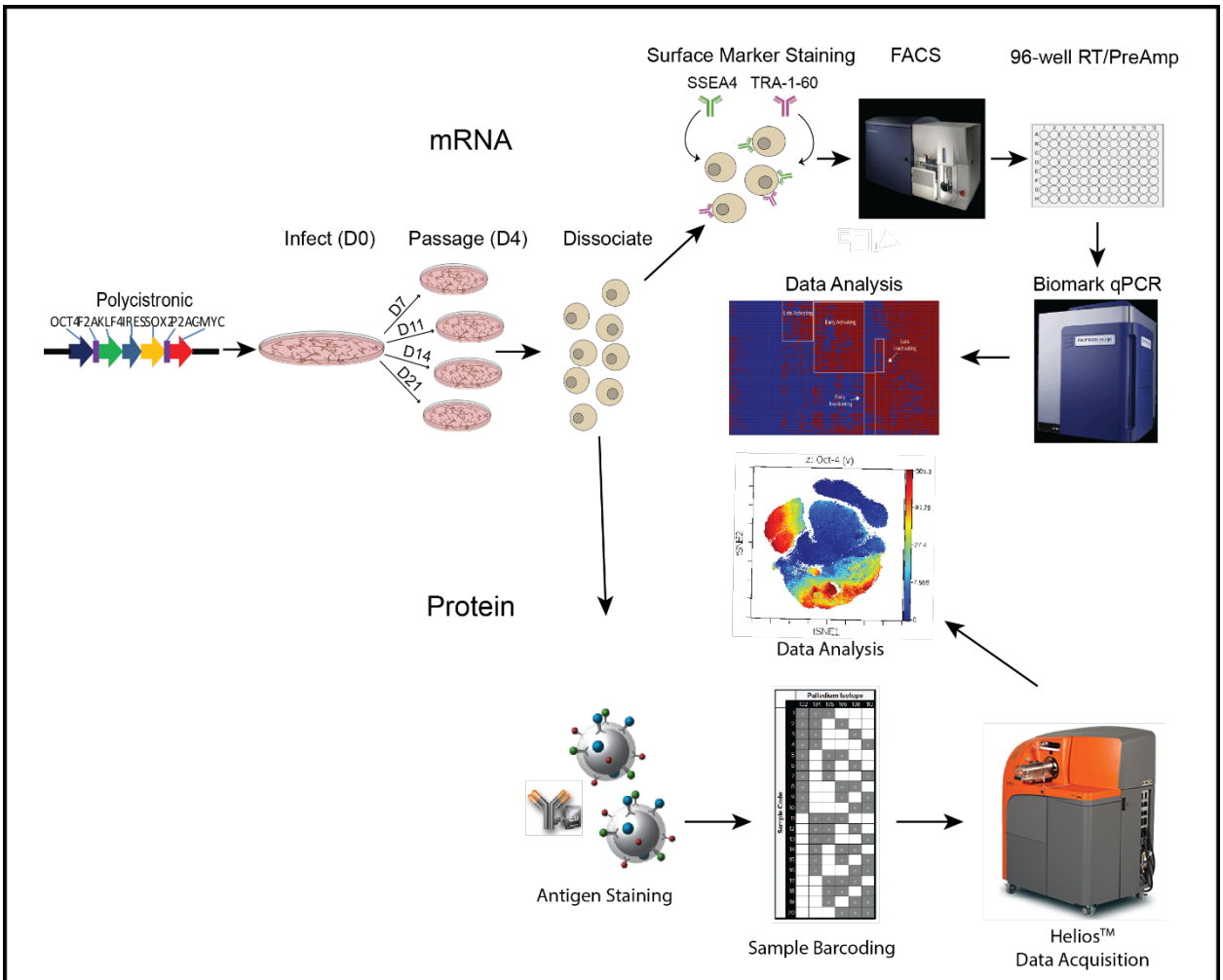


Figure 16: Schematic of CyTOF Experimental Design. Cells BJ Fibroblasts were reprogrammed with Polycistronic virus at different days such that harvest of cells at D4, D7, D11, D14 and D21 could be performed simultaneously. Cells were harvested and stained with surface antigens and fixed for shipment to Fluidigm Sciences. Fluidigm performed intracellular staining and barcoding of the samples prior to acquisition on the Helios instrument. Data analysis was performed at the University of Connecticut. Single Cell mRNA analysis was performed as described in Chapter 3.

4.2.2 Defining the Productive vs Alternate Reprogramming Trajectories

In order to model the expression of proteins during reprogramming it is essential to profile only the cells undergoing productive reprogramming, as unproductive events will likely exhibit different behavior and will confound the analysis. To determine which cells were productively reprogramming we used the viSNE algorithm to project our data into 2 dimensions, while conserving the relationships between cells in the original high-dimensional dataset[163] (Figure 17A). Importantly, viSNE is intended for use with non-linear data, particularly flow cytometry data, making it ideal for this application. Examination of the viSNE projection shows a clear separation of the fibroblast and hESC groups, with a wide range of reprogramming cells spanning these two populations. This reprogramming population is rapidly distinguished from the fibroblast group as early as day 4, and progresses toward a hESC-like profile at subsequent days.

The expression profiles of the reprogramming factors themselves distinguish cells that have received the virus from those that did not, since the antibodies recognize both the endogenous and exogenous proteins. Based on these results, there is a clear population of cells at the bottom of the graph that robustly express all four reprogramming factors (Figure 17B, red boxes). This is expected because the reprogramming factors are translated from a polycistronic mRNA and thus all cells that receive the virus should express the full complement of OSKM. In contrast, a population at the top of the map exhibits low or no expression of OSKM and is distinct from the fibroblast group, suggesting these are cells that received but pre-maturely inactivated the virus during the reprogramming process (Figure 17B, blue boxes). The low expression we detect in this group could reasonably be attributed to residual OSKM proteins that have yet to be turned over in the cell.

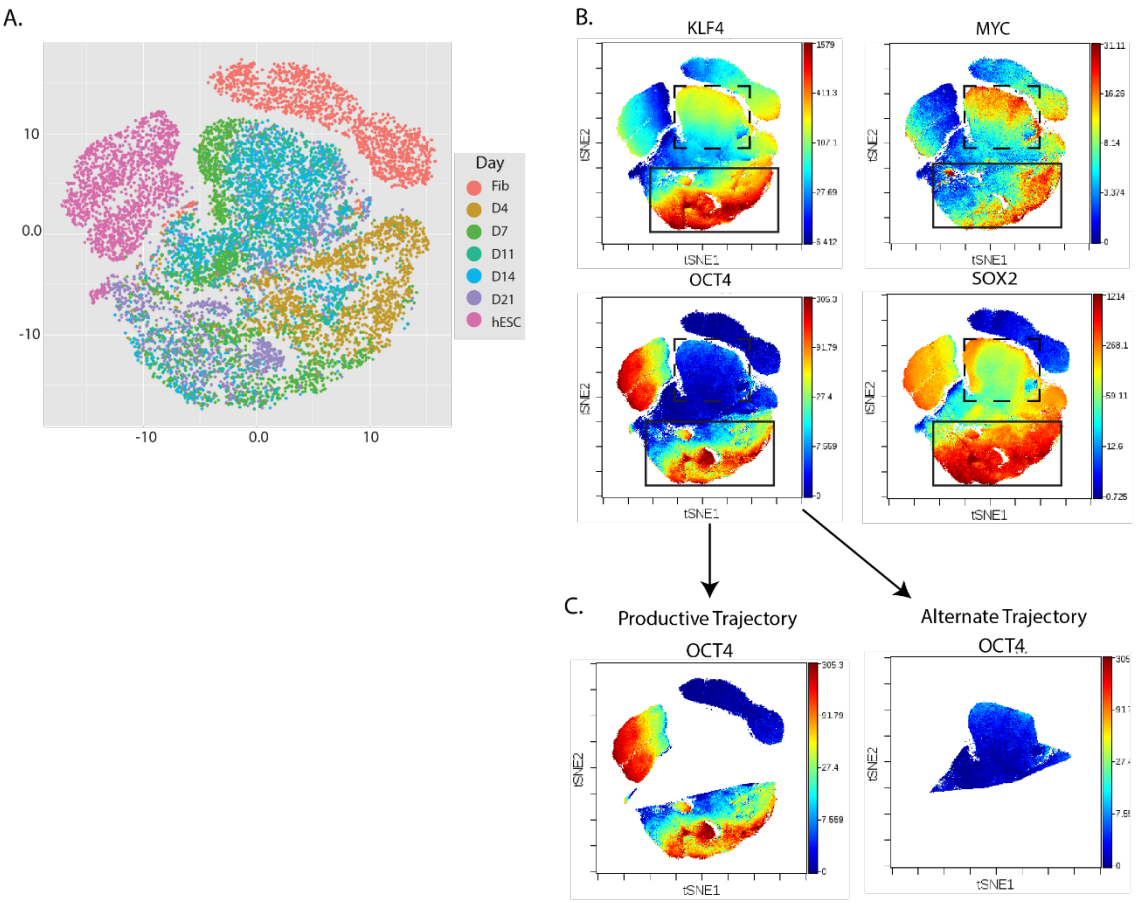


Figure 17: viSNE Projection and Transgene Analysis Reveals Productive and Non-Productive Reprogramming Trajectories. (A) viSNE Projection of CyTOF data colored by the day of collection shows a roughly clockwise progression from fibroblast (top right) to hESC (top left) at successive reprogramming time points. (B) Overlay of OSKM transgene expression on the viSNE map shows high expression of the factors at the bottom of the graph (solid boxes) and low expression at the top of the graph that lies at an intermediate state between fibroblast and hESC (dashed boxes). We define the productive and non-producctive reprogramming populations based on these transgene expression profiles (C).

Having identified the productive population, we sought to measure the dynamics of gene expression as cells progress through the reprogramming process. Using the progression of cells over time as an approximation for reprogramming progress, we gated populations (Figure 18A) and measured the median expression of each protein in the panel. Unsupervised hierarchical clustering separates these gates into two major groups; an early group expressing fibroblast genes and lacking pluripotency gene expression and a late group with the inverse expression profile (Figure 18B). Within these broad classifications there is a progressive inactivation of the fibroblast genes LUM and LCK, and progressive activation of pluripotency genes such as EpCAM, DNMT3B and NANOG. Strikingly, several surface markers expressed in fibroblast cells are abruptly down-regulated between the early and late groups including CD13, CD44, CD49d and CD73.

Because the above analysis relies on temporal changes in the data, which are known to be asynchronous between cells, we turned to a recently developed, unsupervised algorithm wanderlust, which was developed to identify cell-state transitions from high dimensional CyTOF datasets[164]. Briefly, wanderlust uses a graph-based approach to construct a *k* nearest neighbor graph of all the cells in the dataset, based on their similarity (in this case, protein expression profiles). The algorithm then computes the shortest distance from each cell to a pre-defined starting cell population (fibroblast) and a randomly chosen waypoint cell, using nearest neighbors as steps along the path. The cells are then ordered based on these distance measurements. This process is repeated iteratively using a randomized subset of k-nearest neighbors until convergence, providing a robust trajectory measurement. The utility of this approach is clear when the wanderlust trajectory is mapped onto our viSNE graph (Figure 18A). This trajectory meanders through the viSNE space, ordering cells based on protein expression similarity in a way that could not be achieved through manual gating and provides a continuous measure of reprogramming progress.

Measuring protein expression dynamics along the wanderlust trajectory reveals two key differences between the productive and non-productive trajectories. First, the non-productive population exhibits low expression of the proliferation marker Ki67, a marker that is progressively upregulated in the productive reprogramming populations as cells approach the hESC state (Figure 18D, right panel). This finding is in line with data from Zunder et. al. in mouse that identified a Ki67-null population that had exited the cell cycle and when enriched by FACS and cultured for several days, failed to produce iPSC colonies. Second, the non-productive population robustly expresses PDGFRa, a marker that is absent in fibroblasts cells but activated early in the productive trajectory (Figure 18D, left panel). PDGFRa is eventually downregulated in the productive cells as they approach the hESC state, indicating that the alternate group initiated, but will fail to complete reprogramming. Because PDGFRa is downregulated in the productive group late in the trajectory but remains expressed in the alternate group at all time points, it may be useful for negative selection against cells that are refractory to reprogramming at later stages of the process.

Fibroblast and pluripotency genes exhibit the expected patterns of up- or down-regulation along the wanderlust trajectory but groups of genes appear to change in concert at particular points in the process (Figure 18 E and F). This is very different from the highly variable points of activation/inactivation observed from genes in our previous mRNA expression analysis and suggests potentially different expression dynamics at the mRNA and protein levels.

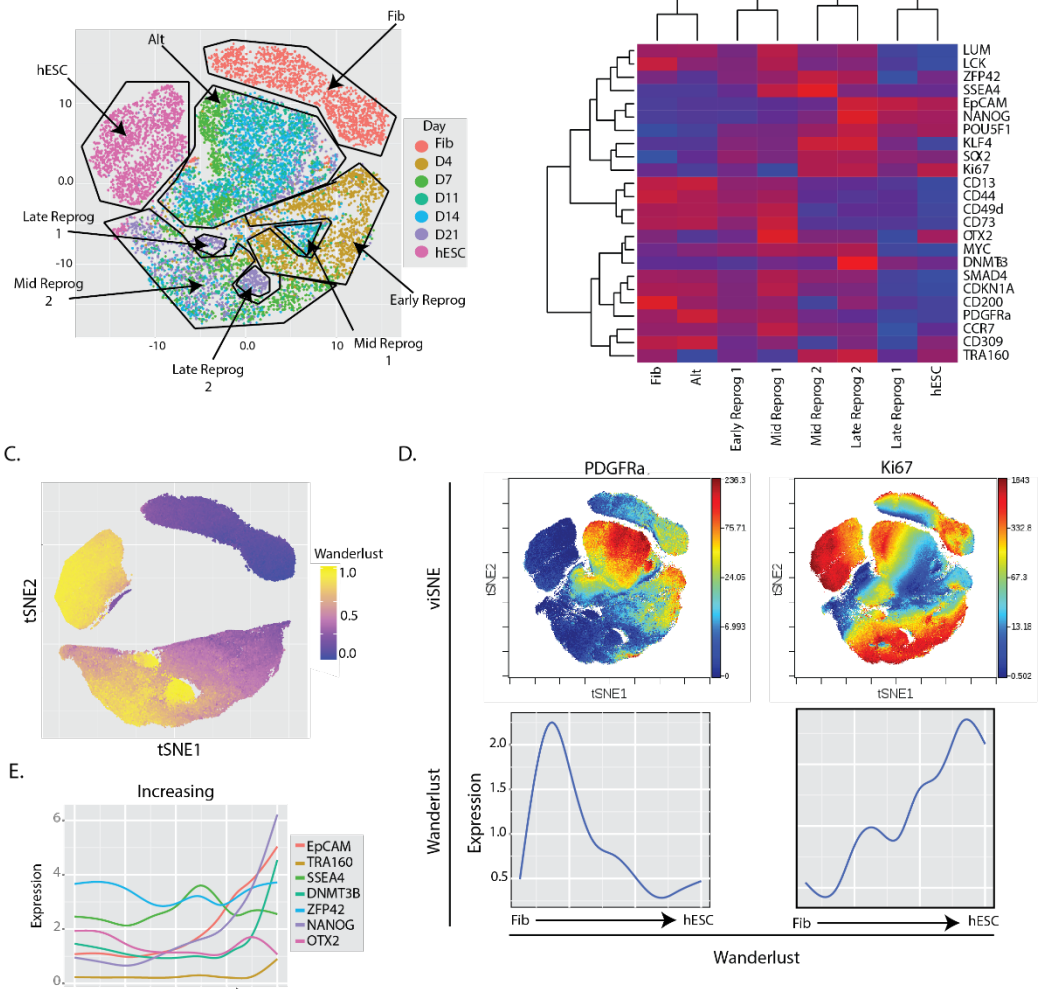


Figure 18: Measuring Protein Expression Along the Productive Trajectory. The progression of cells from Fib to hESC over the timecourse of reprogramming was manually gated in viSNE space (A) and the mean expression of each protein was measured and visualized by hierarchical clustering (B). Unsupervised trajectory detection was perfomed using the wanderlust algorithm which orders cells between 0 (fibroblast-like) and 1 (hESC-like) based on their expression profiles. These values are overlaid on the viSNE trajectory in (C). (D) Robust expression of PDGFRa is obseved in the alternate trajectory (top left) but is early and transient in the productive trajectory (bottom left). In contrast, Ki67 expression is low in alternate cells (top right) but increases steadily along the productive trajectory. Splines depict expression profiles for increasing (E) and decreasing (F) genes along the produictive, wanderlust trajectory.

4.2.3 Delayed Protein Expression of Key Pluripotency Genes

In order to directly compare mRNA and protein expression dynamics we first defined a common reprogramming trajectory between the two datasets, with the alternate trajectories removed. For the protein dataset, we examined only $SSEA4^+$ events, since all cells collected for mRNA analysis were enriched for SSEA4 expression. We then used Euclidean distance to measure each cell's similarity to both the fibroblast and hESC groups within a given dataset such that the mRNA reprogramming trajectory is relative to fibroblast and hESC mRNA profiles, while the protein trajectory is relative to the fibroblast and hESC protein expression profiles (Figure 19A and B). It is clear that in both datasets there is a continuum of cell states between the fibroblast and hESC groups, although the protein dataset is slightly more diffuse in an orthogonal direction to the productive reprogramming axis.

We next applied our Gaussian modeling pipeline to the two datasets to assess differences in the rate or timing of protein/mRNA expression. Because the protein expression dataset has nearly $1X10^5$-fold more data points than the mRNA dataset, we down-sampled and bootstrapped the CyTOF data to reduce sampling bias. Comparing the model fits between the two datasets as a whole, we observe a slight tightening of the distributions in the protein dataset suggesting more rapid up- or down-regulation at the protein level, however this difference is not significant ($\alpha = 0.05$) for all markers across the dataset (Figure 19C). A similar trend exists when comparing the mean point of activation/inactivation along the trajectory, where globally, there exists a slight delay in protein expression but again this difference is not significant (Figure 19D). When examining individual targets, it is clear that loci such as DNMT3B, NANOG and LUM have a significant delay in protein activation while other targets do not show an effect (Figure 19E and F).

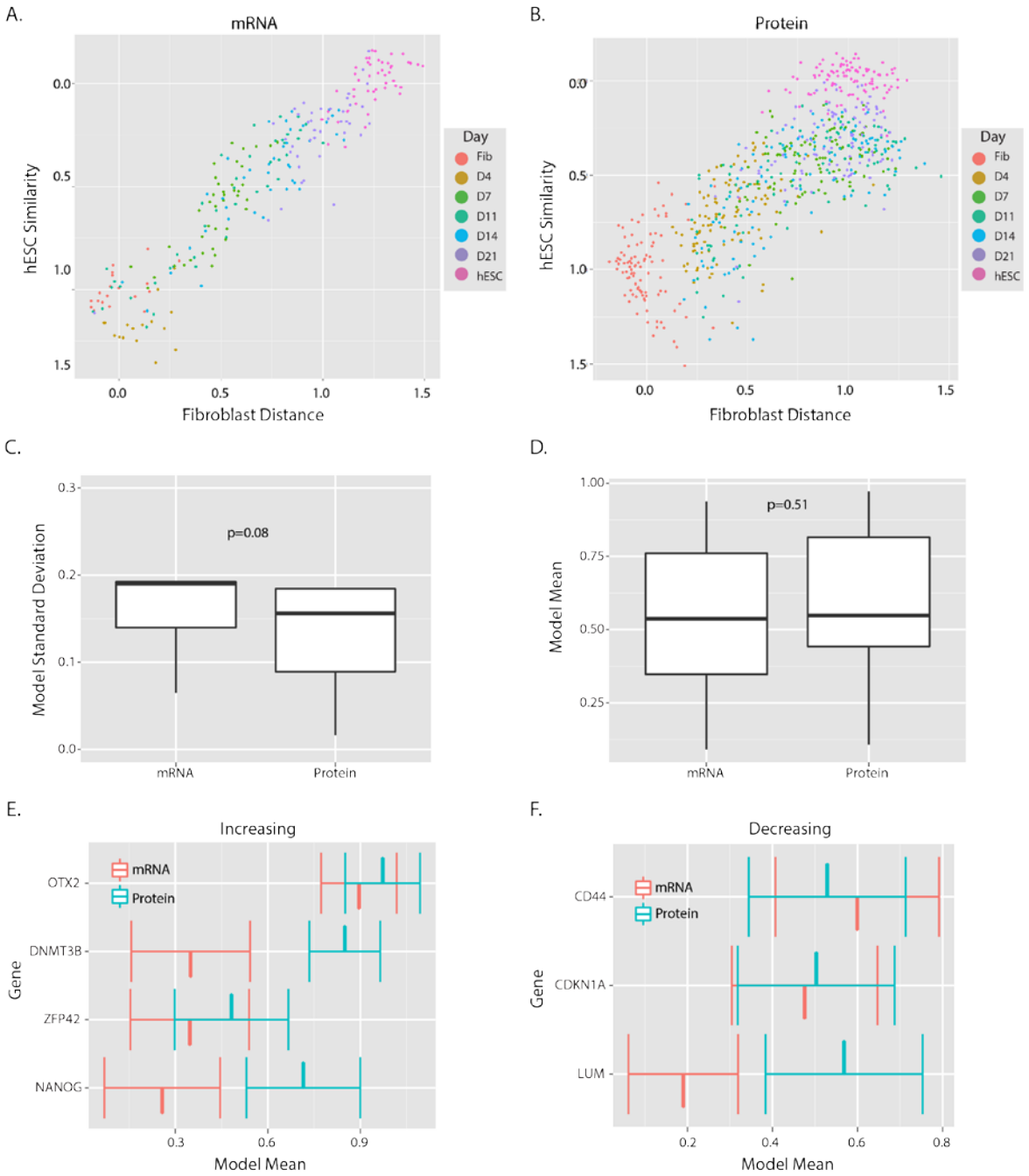


Figure 19: Comparisson of mRNA and Protein Expression Dynamics. Euclidean distance projection of mRNA (A) and Protein (B) expression datasets colored by day of collection. Analysis of all targets shows a more narrow window for protein compared with mRNA expression changes determined by the standard deviation of the Gaussian model fit individually to each target's expression data (C). A slight but insignificant delay in protein expression is observed based on the mean of the Gaussian model fitting (D). Examination of individual genes shows a significant delay in protein expression for a subset of activating (E) and inactivating (F) targets including NANOG, DNMT3B and LUM.

4.2.4 Greater Co-Regulation of Protein than mRNA Expression

To address whether protein expression exhibits greater co-regulation than mRNA expression, we measured the degree of correlation between markers within each of these datasets. First, we generated a background correlation that assumes independent behavior of all the genes. We then measured the Pearson correlations for all of the genes and subtracted the background to obtain our correlation-above-background measurement. Figure 20 A and B demonstrates that when looking across all targets for each dataset, there is considerably greater correlation between proteins, compared with genes at the mRNA level. As expected, we observe strong correlation between the reprogramming factors, as these are co-transcribed and co-translated from the polycistronic virus and the CyTOF antibodies recognize both the endogenous and exogenous species (Figure 20C). We also observe many strong correlations for increasing and decreasing loci in the protein dataset (Figure 20 D and E), much greater than what is observed when compared directly with correlations from mRNA expression data (Figure 20F).

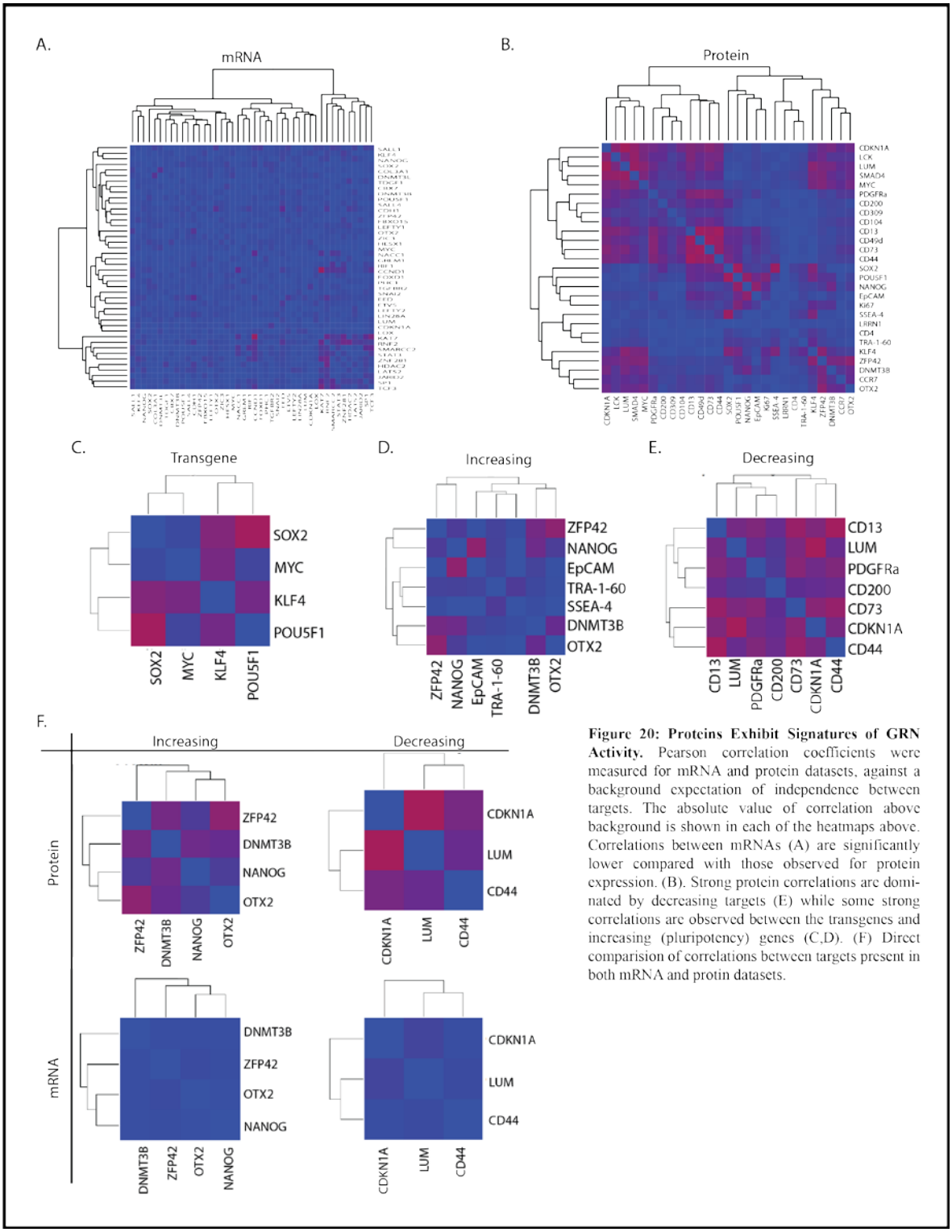


Figure 20: Proteins Exhibit Signatures of GRN Activity. Pearson correlation coefficients were measured for mRNA and protein datasets, against a background expectation of independence between targets. The absolute value of correlation above background is shown in each of the heatmaps above. Correlations between mRNAs (A) are significantly lower compared with those observed for protein expression. (B). Strong protein correlations are dominated by decreasing targets (E) while some strong correlations are observed between the transgenes and increasing (pluripotency) genes (C,D). (F) Direct comparision of correlations between targets present in both mRNA and protin datasets.

4.3 Discussion

Years of research on the molecular mechanisms of reprogramming have yielded valuable insights into key rate limiting steps but have been primarily focused on genomic analyses such as mRNA expression an epigenetic modifications. Until recently, the only studies of reprogramming at the protein level were limited to techniques such as immunohistochemistry, FACS or reporter constructs, which can only analyze a few markers at a time. The low throughput nature of these technologies limits their ability to identify novel protein biomarkers and to compare directly to the wealth of genomics data available. As a result, little is known about how proteins behave during reprogramming and whether post-transcriptional mechanisms may be an additional rate-limiting step in the process.

Here we present the first single cell protein expression analysis of reprogramming in human cells using CyTOF technology and directly compare these data with single-cell mRNA expression data. The single cell resolution of these datasets is essential for accurately comparing their dynamics during a highly heterogeneous and temporally asynchronous process and provides greater power to assess the co-regulation of these molecules. The use of human cells is more directly relevant for therapeutic approaches and provides a necessary point of comparison for similar studies performed in mouse to determine the commonalities and differences between these two systems.

A major finding of this work is that of the proteins examined, the pluripotency markers NANOG and DNMT3B are expressed significantly later in reprogramming than their mRNA transcripts. From our previous work we know that mRNA transcripts associated with pluripotency are expressed in a probabilistic fashion prior to being upregulated in the majority of cells by the end of the process, however it was unclear whether these transcripts were effectively translated into protein. The delayed expression of these markers at the protein level indicates their translation is a late event in reprogramming.

There are several reasons why this delay in translation might occur. One possibility is that the probabilistic expression of transcripts early in the process is ephemeral and does accumulate sufficient mRNA to be effectively translated, or that the amount translated is below the threshold of detection. This is unlikely to be the case however, as the expression levels of these transcripts is comparable to what is seen in hESCs, where expression of these genes as proteins is abundant. A more plausible explanation is the presence of miRNAs or other negative regulators of mRNA translation acting within the cell. Indeed many miRNAs are expressed in fibroblasts and the repression of aberrant transcripts is a common means of stabilizing cellular identity[40,165–167], however it is unclear whether any of these target pluripotency genes. Analysis of miRNA expression during reprogramming has been reported previously[42,152,168] and mining these datasets for miRNAs present in fibroblasts but downregulated in iPSCs coupled with a target prediction algorithm such as mirPath could identify putative regulators of this phenomenon. An alternative mechanism preventing the translation of probabilistically expressed mRNAs into protein is the absence of appropriate RNA binding proteins to stabilize the transcript or facilitate splicing into isoforms appropriate to the cell type. Indeed many stem-cell specific RNA binding proteins have been identified which presumably must be reactivated during reprogramming to achieve pluripotency[169]. The field of RNA biology is in its infancy however, and the role of cell-type specific RNA binding proteins is largely unknown.

Regardless of the mechanism, delayed protein expression maybe be a rate limiting step in the process and manipulating regulators of this phenomenon could improve reprogramming efficiency. It is important to note that not all genes in our dataset follow this pattern and it is possible that the delay in protein expression may be the exception, rather than the rule in reprogramming. This has been demonstrated in proteomics analysis of mouse reprogramming where only a subset of genes, mainly those involved in cell adhesion and cell-to-cell signaling, exhibited poor correlation between mRNA and protein expression[85]. To determine how many targets exhibit this effect, a larger-scale analysis would need to be conducted. Indeed, a major limitation of the present study is the number of targets in the panel

that can be compared directly between the mRNA and protein datasets. When developing our panel for CyTOF, we started with a list of 96 genes for which we had already generated single-cell mRNA expression data. After selecting targets with reliable antibodies and validating on our own cell populations, we were left with 13 genes from the original panel. Additional markers of interest were chosen from the literature and brought the total number of proteins analyzed to 27, however these additional markers could not be compared with mRNA expression data. An important future direction of this work will be to perform single cell mRNA-seq analysis of the reprogramming process and to repeat the CyTOF experiment using a larger panel which is currently limited by the availability of rare metal conjugates to ~40 targets. This will provide a much larger dataset and allow for more confidence and robust statistical power to complement the data presented here.

This work also demonstrates that greater coordination exists at the protein level when compared with the corresponding mRNA transcripts. This finding is somewhat expected given that not all transcripts are effectively translated into protein since the presence of the protein is typically required to exert function, for example regulating target genes. The increased correlation between proteins however does not directly imply that the process is more deterministic at the protein level. Simply visualizing the reprogramming trajectory derived from the protein expression data suggests a continuum of cell states rather than a series of discrete intermediates, however assessing the degree of deterministic behavior between proteins/mRNAs is a hypothesis that must be explicitly tested. A simple method of measuring determinism between genes is to compare the frequency with which one genes is activated before the other. A completely random system with no determinism between the genes would result in a frequency near 50%; an equal chance of either gene being activated, independently of the other. In a strictly deterministic situation, the probability of one gene activating before the other would be much greater and would approach 100%.

Performing this test in a pairwise manner is straight forward, however looking at a complex system such as our reprogramming datasets, requires more sophisticated analyses. Bayesian analysis is a

powerful tool to estimate the probability of an event (expression of gene A) given other information about the system (expression of genes B, C, D etc.). This analysis is particularly powerful when combined with single cell analysis because each single cell collected represents a measurement of the system, and the confidence of the predictions increases as more information is collected. Once all of the conditional probabilities of the system have been computed, a network diagram can be assembled to understand which genes are deeply integrated in the system and which are peripheral. It will be important to apply this method to both our mRNA and protein datasets to see if these networks are the same, and whether there is a difference in the dependence or connectivity of these systems. A caveat to this analysis is the need to reduce the continuous measurements obtained from the instrument, to a binary measure of presence or absence of expression. Single cell gene expression data lends itself well to this technique since cells with little/no expression lack detectable amounts of transcript and are thus have inherent zero values. In contrast, protein expression data from the CyTOF has no absolute zero values as even cells with no expression of a marker show some background staining. Several algorithms exist to binarize continuous datasets such as kmeans analysis or Euclidean distance stepping, both of which have been applied successfully to similar datasets such as microarray experiments. These analysis coupled with a larger mRNA and protein dataset for comparison will provide useful information about how effectively mRNAs are translated into protein during reprogramming and whether chemical or other interventions can overcome this hurdle. In addition, the analysis of a large protein dataset holds promise to identify new biomarkers of productive reprogramming that can be used to enrich or isolate cells with increased potential to become iPSCs.

Chapter 5 Discussion and Future Directions

5.1 Discussion

The ability to reprogram somatic cells to pluripotent iPSCs was first demonstrated in 2006 and shows incredible promise as an autologous source of cells for transplantation and to model human diseases *in vitro*. The low efficiency of the process and heterogeneity of the resulting iPSCs however, has

limited the translation of this technique to a clinical setting. This led our group and many others to study the mechanisms governing the reprogramming process to identify rate-limiting factors and biomarkers of productive reprogramming.

During the time this work was conducted the field of iPSC research progressed at a dizzying rate and many reports were published defining the molecular events preceding the acquisition of pluripotency. Many of these early studies profiled cell populations in bulk, but in recent years single cell analysis has been used to effectively describe the dynamics of reprogramming. Our group was the first to report single cell transcript analysis in human reprogramming and to our knowledge, are also the first to measure protein expression in human reprogramming as well. Through the measurement of 96 mRNAs and 27 proteins (in separate studies) we mapped the progression of cells through the reprogramming process. Using mathematical models to describe the behavior of these molecules, we demonstrate that gene expression dynamics are ordered yet probabilistic, reconciling previous ordered and stochastic models of the reprogramming process. We also identify an alternate reprogramming trajectory of cells moving away from a fibroblast state but not towards pluripotency. Because cells on this trajectory fail to express many of the markers in our panel their exact molecular signature is unclear, however we confirm these cells are viable and lack detection of pro-apoptotic markers.

We have also used our experimental pipeline to profile mRNA expression using different methods of reprogramming. At a time when many methods have been used to reprogram a variety of cell types, it is important to understand how and if variations in the technique impact the way in which cells achieve pluripotency. By comparing two delivery methods of the OSKM factors, we show that polycistronic reprogramming is more efficient than monocistronic reprogramming and that this is due at least in part to premature inactivation of the reprogramming factors in the monocistronic condition. This phenomenon also results in the alternate trajectory mentioned previously, which is present in the monocistronic but not polycistronic dataset. In a separate experiment reprogramming BJ and MRC-5 fibroblasts with polycistronic virus, we find that the more efficient BJ cell type activates pluripotency

genes earlier in the trajectory however, the order of gene activation remains the same. This suggest a common reprogramming trajectory between these two cell types despite differences in efficiency.

Outside the context of reprogramming, this body of work demonstrates the power of single cell analysis and mathematical modeling to dissect complex biological processes and gain insights not possible at the bulk-population level. Processes such as embryonic development and cell signaling exhibit temporal asynchrony and differential response between individual members of the population and are amenable to the methods presented here. By taking many molecular measurements (ie gene or protein expression) of each cell, we determine the relationship between cells, create a map of the process independent of time and determine a parsimonious trajectory along which to model the behavior of individual genes or proteins. This approach has since been applied by other groups[164,170] and is quickly becoming a common tool for analyzing complex systems.

5.2 Future Directions

The findings presented here represent a first step in the analysis of human reprogramming at the single cell level and provide a framework to dissect the dynamics of the reprogramming process under different conditions. In this work, we employed the most common reprogramming methods to establish a "baseline" reprogramming trajectory in a relevant and widely applicable system. This same analysis however, can and should be expanded to a variety of reprogramming systems. One approach would be to determine the targets of small molecule compounds known to enhance the efficiency of the process. Many small molecules such as chromatin modifying compounds (TSA, VPA,5-Aza etc.), have known targets, while others such as OAC1, a compound identified in a screen for OCT4 activating agents, have no known mechanism of action[71]. A direct comparison of reprogramming with and without these molecules at the single cell level combined with our modeling technique is likely to reveal how these molecules enhance reprogramming efficiency and could help identify novel barriers to reprogramming. Importantly, single-cell mRNA-seq is quickly becoming a trivial technique and applying this technology to the

reprogramming paradigm avoids the barriers associated with selecting a panel with a limited number of targets.

One important finding of this work is the independent behavior of individual loci during the stochastic phase of reprogramming, which suggests a local property of the gene governs its activity. In line with this conclusion, the epigenetic state of pluripotency loci is known to influence the order and timing of their activation. This implies that cell types with known differences in their epigenetic profiles should exhibit different gene expression dynamics during reprogramming, however this hypothesis has not been explicitly tested. We find a common reprogramming trajectory between BJ foreskin and MRC-5 lung fibroblasts, however the similarity of their epigenetic profiles is unclear and little epigenetic profiling has been performed on MRC-5 cells. A more rigorous test of the hypothesis would be performed in many divergent cell types by single-cell mRNA-seq.

The ENCODE project has performed ChIP-seq on dozens of commonly reprogrammed cell types, five of which were performed by the same lab and have perfectly overlapping datasets[171]. These cell types which include Human Skeletal Muscle Myoblasts (HSMM), Human Umbilical Vein Endothelial Cells (HUVEC), Human Mammary Epithelial Cells (HMEC), Normmal Human Epidermal Karatinocytes (NHEK) and Normal Human Lung Fibroblasts (NHLF) have been profiled for 12 different epigenetic modifications, many of which have established roles in reprogramming. These cell types and their accompanying epigenetic data represent a perfect system to assess whether variations in chromatin state impact reprogramming gene expression dynamics and whether this contributes to differences in efficiency between cell types. For example, cell types such as HSMM that have a largely repressive epigenetic environment may take longer to activate certain pluripotency loci and thus may reprogram at lower efficiency. In contrast, we would expect the opposite effect in HUVEC cells which have greater developmental plasticity and thus a more permissive epigenetic environment. These cell-type level effects would likely result from the facilitated or impaired activation of individual pluripotency loci in different

cell types. Identification of novel loci whose activation/repression impacts reprogramming efficiency would also be an important goal of future work.

The protein expression analysis presented in Chapter 4 provides exciting preliminary findings that translation of stochastic transcripts into protein may be a rate limiting step in reprogramming and that the process as a whole may be better coordinated at the protein level. This analysis however relies on too few markers to definitively make this conclusion and several follow-up studies are necessary. First, a more comprehensive CyTOF experiment must be performed where there is greater overlap between the mRNA and protein expression datasets. This could be facilitated by mRNA-seq analysis of the reprogramming process to obtain data for the whole transcriptome. From this data a large panel of targets could be selected that are expressed/repressed at different points along the reprogramming trajectory. Starting with a large set of markers will ensure that a sufficient number of protein targets will remain after quality control and validation of the antibodies. With both the mRNA and protein data in hand, the degree of determinism in the system can be inferred by Bayesian methods and our modeling approach can be used to compare expression dynamics.

Genes or proteins that appear dependent can also be validated by orthogonal experiments, such as those used by Buganim et. al. to validate their hierarchical model of the reprogramming process. This method uses single molecule Fluorescent *In Situ* Hybridization (smFISH) to label target mRNAs coupled with flow cytometry to measure the frequency with which one gene's expression precedes another in thousands of individual cells. When coupled with antibody staining, this technique could also be used to visualize instances where the translation of an mRNA appears delayed, as observed for NANOG and DNMT3B in our existing dataset. Furthermore, smFISH can be combined with fluorescence microscopy to provide an absolute quantification of mRNA transcripts to determine whether translation of mRNAs into protein is subject to a threshold effect, or whether other post-transcriptional mechanisms are involved.

At a time when most reprogramming studies are performed in bulk populations of mouse cells, our profiling individual human cells undergoing reprogramming at the mRNA and protein levels represents an important contribution to the field. In addition, the mathematical modeling approach presented here can be used for future studies to interrogate the reprogramming process or other complex biological processes at the single cell level. Here we provide a powerful framework for similar analyses that will be required as the feasibility of, and necessity for, these experiments increase in the years to come.

Chapter 6 Materials and Methods

6.1 Production of Monocistronic OSKM Retrovirus

Retroviral vectors (pMIG) containing OCT4, SOX2, KLF4, c-MYC (OSKM) along with helper plasmids (VSV-G and Gag-pol) were obtained from I.H.Park (Yale University, New Haven, CT). To generate viral particles, individual retroviral vectors were co-transfected with VSV-G and Gag-pol into 293T cells seeded at 2 x 10^6 cells per 10-cm^2 using FuGENE 6 transfection reagent (Roche Applied Science). After 72-hour induction, supernatants were collected, filtered through 0.45µm filter and concentrated using Vivaspin 300,000 MWCO PES filter columns (Sartorius). Viral titer was determined using FACS analysis for GFP expression (encoded in the pMIG vector). An MOI of 5 was used for all experiments.

6.2 Production of Polycistronic OSKM Lentivirus

Simultaneous delivery of the four reprogramming factors OCT4, SOX2,KLF4 and c-MYC was achieved using the STEMCCA-LoxP polycistronic lentiviral vector, a generous gift from Dr. Gustavo Mustolovsky (U. Mass Boston Medical School). Virus was produced by cotransfection of STEMCCA-LoxP along with helper plasmids VSV-G, Gag-pol, Rev and TAT into 293T cells in 5 x 10cm^2 dishes using Xtreme-GENE9 transfection reagent. 72hrs post-transfection, supernatant was harvested and

concentrated to 1mL using Vivaspin 300,000 MWCO PES filter columns (Sartorius), and 100ul of the concentrated virions were used for each reprogramming experiment.

6.3 Construction of 3-Factor Reprogramming Lentiviruses

The 3-factor polycistronic vectors were constructed by modifying STEMCCA-LoxP as follows. First, we removed either the OCT4-F2A-KLF4 or SOX2-E2A-cMYC cassette from the STEMCAA-LoxP vector using the NotI/BamHI and NdeI/BsaBI sites, respectively. The deleted cassettes were then replaced with one of the two original cDNAs to generate a 3-factor-containing polycistronic vector. The individual human cDNAs encoding the four reprogramming factors were amplified from the STEMCCA-LoxP vector using following primers: OCT4 NotI Poly F (5'-GCGGCCGCATGGCGGGACACCTGGCTTC-3'); OCT4 BamHI Poly R (5'-GGATCCTCAGTTTGAATGCATGGGAGAG-3'); KLF4 NotI Poly F (5'-GCGGCCGCATGGCTGTCAGCGACGCGCTG-3'); KLF4 BamHI Poly R (5'-GGATCCTTAAAAATGCCTCTTCATGTG-3'); Sox2 NdeI Poly F (5'-CATATGATGTACAACATGATGGAGACGG-3'); Sox2 BsaB1 Poly R (5'-GATCCTAATCCTATGTGTGAGAGGGGCAGTGTG-3'); c-Myc NdeI Poly F (5'-CATATGATGCCCCTCAACGTTAGCTTCACC-3'); c-Myc BsaB1 Poly R (5'-GATCCTAATCTTACGCACAAGAGTTCCGTAGCTG-3').

6.4 Cell culture and Fibroblast Reprogramming

MRC-5 human fetal lung fibroblasts were obtained from I.H. Park (Yale University, New haven, CT) at passage 8 and BJ human foreskin fibroblasts were purchased from Global Stem (GSC-3002) at passage 6. No cells beyond passage 10 were used for reprogramming. Briefly, fibroblast cells were expanded in human fibroblast (hFib) media (DMEM (Gibco), 10% FBS (Milipore), 1% L-glutamine (Gibco) and 1X Penn-Strep (Gibco)). One day prior to infection, 1 x 10^5 MRC-5 fibroblasts were seeded into one well of a 6-well dish containing hFib media. The next day, cells were incubated in RI media

(MEM alpha (Mediatech) and 10% FBS (Millipore)) containing 5ug/mL protamine sulfate (Sigma) and OSKM virions for 24hrs followed by replacement with fresh RI media. Cells were cultured for 72hrs post-infection and passaged to $10cm^2$ dishes pre-seeded with 7.5×10^5 inactivated feeders in hESC media supplemented with 10uM ROCK inhibitor (Y-27632, Calbiochem). Cells were split 1:2 for all monocistronic experiments and 1:3, 1:10, 1:20 and 1:30 for day 7, 11, 14 and 21 polycistronic reprogramming, respectively. After passaging, fresh hESC media was added daily until the end of the experiment. H9 human embryonic stem cells (WiCell) were maintained either on MEFs in hESC media (DMEM F-12 (Gibco), 20% Knockout-Serum Replacement (Gibco), 1% L-Glutamine (Gibco), 1% Non-Essential Amino Acids (Gibco), 5μM β-mercaptoethanol (Gibco), and 2ng/mL b-FGF), or feeder-free on matrigel coated plates with mTesr-1 media following the manufacturer's instructions. hESCs were passaged as single cells using hES Cloning Recovery Reagent (Stemgent) to enhance clonal survival.

6.5 Antibody Staining and FACS Sorting of Reprogramming Cells

Reprogramming fibroblast cells were harvested with 1mL Accumax (Millipore) per well (6-well dish) for 10 minutes at 37°C. Cells were pelleted, washed with PBS (Gibco) and wash buffer (2% FBS in HBSS (Invitrogen)), and resuspended in wash buffer. Cells were then stained for 30min using αSSEA-4 (Biolegend, Cat# 330405) αTRA-1-60 (Biolegend, Cat# 330605) and αMEF (Miltenyi, 130-102-900) antibodies, washed 3 times and resuspended in FACS buffer (1% FBS in PBS). For FACS, cells were live/dead stained and gated on GFP (for monocistronic reprogramming) and appropriate surface markers as indicated and single cells sorted into 96 well PCR plates. All FACS was performed using a BD Bioscience FACS Aria II.

6.6 AP Staining and Surface Marker Quantification

Alkaline phosphatase staining was performed in 6-well plates using the alkaline phosphatase detection kit (Millipore) per the manufacturer's instructions. Plates were images in bright field on an Olympus SZ61 dissecting microscope and colony number and total area were counted using ImageJ. To accurately quantify the percentage of SSEA4+ and TRA-1-60+ cells from each condition, we stained

with biotinylated α-SSEA4 (Biolegend, 330404) or α-TRA-1-60 (Biolegend,330604) primary, followed by Brilliant Violet-421 secondary (Biolegend 405226) and APC α-MEF (Miltenyi) at reprogramming day 14 and 21. All experiments above were performed in biological triplicate.

6.7 Quality Control and Single Cell qRT-PCR

Single cell qRT-PCR was performed as previously described [126]. Single cells were sorted by FACS into lysis solution (5% NP-40, 1000U RNasin plus (Promega, PRN2615)) and denatured by incubating at 70°C for 10 minutes and then cooled to 4°C. Cells were then reverse transcribed with MMLV (Promega, PR-M1705) and pre-amplified with TAKARA PCR Master Mix (TAKR004A) using gene specific primers (0.25X pooled TaqMan assays). qPCR was performed either using TaqMan chemistry in 384 well plates on an ABI 7900HT Fast Real-Time system, or in 96.96 Dynamic Arrays on the BioMark-HD® system (Fluidigm). Average cycle threshold (Ct) values obtained from qPCR reactions were normalized to GAPDH (ΔCt), and inverted by taking the (40 – ΔCt) value. To reduce technical error and ensure robust sample quality, all cells with a GAPDH Ct value of 25 or greater were excluded from further analysis. TaqMan assays for endogenous OCT4, SOX2, KLF4 and c-MYC were directed against the 3'-UTR region of the transcript, which is distinct from the synthetic UTRs incorporated in the monocistronic OSKM transgenes, conferring their specificity to the endogenous transcripts. To test for the presence of the viral transgenes in monocistronic reprogramming, primers targeting the synthetic O,S,K and M UTRs were used for RT, Pre-Amp and analysis by SYBR green qPCR on an ABI 7900HT. These primers are listed in (Supp)

6.8 Marker Panel Selection

Genes selected for inclusion in our 96 marker panel were chosen based on several criteria. For pluripotency and chromatin modifier genes we selected those whose role in the establishment or maintenance of the pluripotent state was well documented and experimentally validated. This decision was further informed using the dataset of Dowell et. al. [172] which assigns a self-renewal score to genes based on their integration in the pluripotency gene regulatory network (as determined by direct binding of

O, S, K and/or M) as well as their degree of co-expression with well-established pluripotency genes. Fibroblast genes were selected based on their expression in fibroblasts and absence from hESCs as determined in [128,173].

6.9 Data Analysis from Chapter 2

6.9.1 Mapping the Reprogramming Trajectory

The Euclidean reprogramming trajectory was determined by reducing gene expression to 0(undetected) and 1(detected, Ct < 40) and calculating the average Euclidean distance for each cell to the FIB and PLURI groups, ignoring self-comparisons. Similarity was computed for each group distance by taking the ratio of the distance between FIB and PLURI minus each cell's distance to the group in question, over the distance between FIB and PLURI minus the average distance of that group to itself. The average of the similarity to PLURI and the complement of the similarity to FIB was taken as an estimate of the progression of each cell along the reprogramming trajectory. Distance off of the trajectory was taken as the Euclidean distance from the FIB and PLURI similarities to the trajectory value.

6.9.2 Self-Organizing Map Analysis

PCA-based SOM analysis was performed in JMP, Version 10 (a SAS product)[174] using a 5-by-1 matrix and visualizing on a biplot (PC1 vs PC2). Cells within the "Alt" group were considered to be outliers (as described above) and were excluded from subsequent analysis, unless otherwise indicated.

6.9.3 Hierarchical Clustering

Hierarchical clustering was also performed in JMP, using Ward's method with no standardization, on (40-ΔCT) values. Coverage ellipses on the Euclidian distance graphs represent 90% coverage of the data points from the group indicated. For correlation analysis Pearson's correlation coefficients within a defined SOM grouping were taken for the entire 48x48 matrix of genes analyzed in this study. Network graphs were constructed in Cytoscape using a force-directed layout derived from the top 100 Pearson correlations between all of the cells, excluding outliers, in our analysis (n = 117).

6.9.4 Model Generation

To generate accurate models, the data was first interpolated to generate a high resolution training set. The entire sample population was included, except for outliers considered as the cells with the highest distance off of the trajectory (10%, N=17). The training data represented the percentage of cells expressing a gene at any point along the PLURI trajectory, and was measured by uniformly placing overlapping bins of fixed width across the range and directly counting the number of cells expressing each gene. Models were generated to then predict the percentage expressing at any trajectory location. 'Uniform' models were generated by assigning a 'Baseline' value at the start of the trajectory (=0), and fixing a slope such that a straight line passed from the 'Baseline' to the value at the end of the trajectory (=1). 'Normal' models were then fit to this data using the 'optim' function in R, attempting to minimize the mean squared error, using the constraint, $StdDev \leq 3/16$ and the following form:

$$dNormal(t) = Baseline + \sum_{i=1}^{d} Scale_i * NormalCDF(t, Mean_i, StdDev_i)$$

In order to verify model quality and compare fitting between different models, AICc was calculated and a bootstrapping test was performed. AICc was calculated by:

$$AICc = n \ln MSE + 2k + \frac{2k(k+1)}{n-k-1'}$$

where *n* is the effective number of sample points present in the original data, *k* is the number of free model parameters, and *MSE* is the mean squared error from the model prediction to the training data. Bootstrapping was performed by repeatedly simulating the training data but using only *n* bins and randomly resampling a fixed number of cells from each bin's range. The error between the model prediction and the resampled data was compared to the expected error using an F-test to predict if the error induced by lack-of-fit exceeded the pure error of the data by a significant level, and this was tracked as a percentage of all tests done against the model.

6.9.5 Correlation Analysis

First, simulated populations of an equal size were generated by sampling a set of points along the reprogramming progression axis such that they matched the distribution of values in the original dataset. For each sampling point, representative of a single simulated cell, each gene was set to detected or undetected independently, using the frequency curves generated from our Gaussian model. Pearson correlation coefficients were then computed for this reference population, and averaged over repeated runs (n=1000000). Differences in correlation between this background dataset and those calculated for our observed data were then tested for significance using the 'r.test' function of the R package 'psych'.

6.10 Data Analysis and Modeling from Chapter 3

qPCR data from the Biomark was binarized such that detected genes (Ct<35) and undetected genes were converted to 0 and 1 values, respectively. This dataset was then used for all subsequent analysis in R v3.0.1. We then developed a modeling pipeline to describe the expression changes occurring during reprogramming. First, we use PCA to reduce the data to two dimensions (PC1 and PC2) and we fit a polynomial regression curve to the data and define the reprogramming trajectory. We then project each cell to a point on the curve based on the shortest distance, providing a value for each call along the trajectory. The trajectory values are then scaled between 0 (fibroblast) and 1 (hESC) for easier interpretation. To model expression of each gene, we use the binary data and reprogramming trajectory to fit a logistic regression that describes the probability of detection as reprogramming progresses. Confidence intervals were created by bootstrapping the logistic fitting procedure 100 times and sampling with replacement to ensure robustness of the method.

6.11 Cell Death Analysis

To evaluate the degree of cell death due to inactivation of reprogramming factors, BJ fibroblast cells were reprogrammed using both polycistronic and 3-factor reprogramming vectors and cells were analyzed at day 14 and day 21 in triplicate. Cells were stained either with propidium iodide or α-cleaved

Caspase-3 (Cell Signaling, 9664P) to measure live/dead and apoptosis, respectively. All staining was performed at the manufacturer's recommended dilution and measured on a BD FACSCalibur instrument.

6.12 Collection of Cells for CyTOF

For CyTOF, we collected reprogramming cells at days 4, 7, 11, 14 and 21 as well as BJ fibroblasts and H9 hESCs. In order to minimize technical variation, reprogramming was initiated on different days and all time points were collected and processed in parallel. On the day of collection cells were harvested with Accumax reagent and stained with IdU (Fluidigm, 201127) and Cisplatin (Fluidigm, 201064) following the manufacturer's protocol. A surface antibody cocktail was prepared by combining all surface markers in our panel in 100ul MaxPar Cell Staining Buffer (Fluidigm, 201068). Concentrations used for each antibody can be found in Supplemental Table 7. $3X10^6$ cells from each time point were incubated in the antibody cocktail for 30min at 4C, then fixed and permeabilized using the MaxPar Nuclear Antigen Staining Buffer (Fluidigm, 201063) according to the manufacturer's instructions. Cells were then barcoded using the Cell-ID 20-plex Pd Barcoding Kit (Fluidigm, 201060), combined into a single tube and stained with a cytoplasmic/nuclear antibody cocktail (100ul volume) following the Fluidigm Nuclear Antigen Staining Protocol. Nuclear staining and CyTOF data acquisition were performed by Fluidigm Sciences Inc (San Francisco, CA).

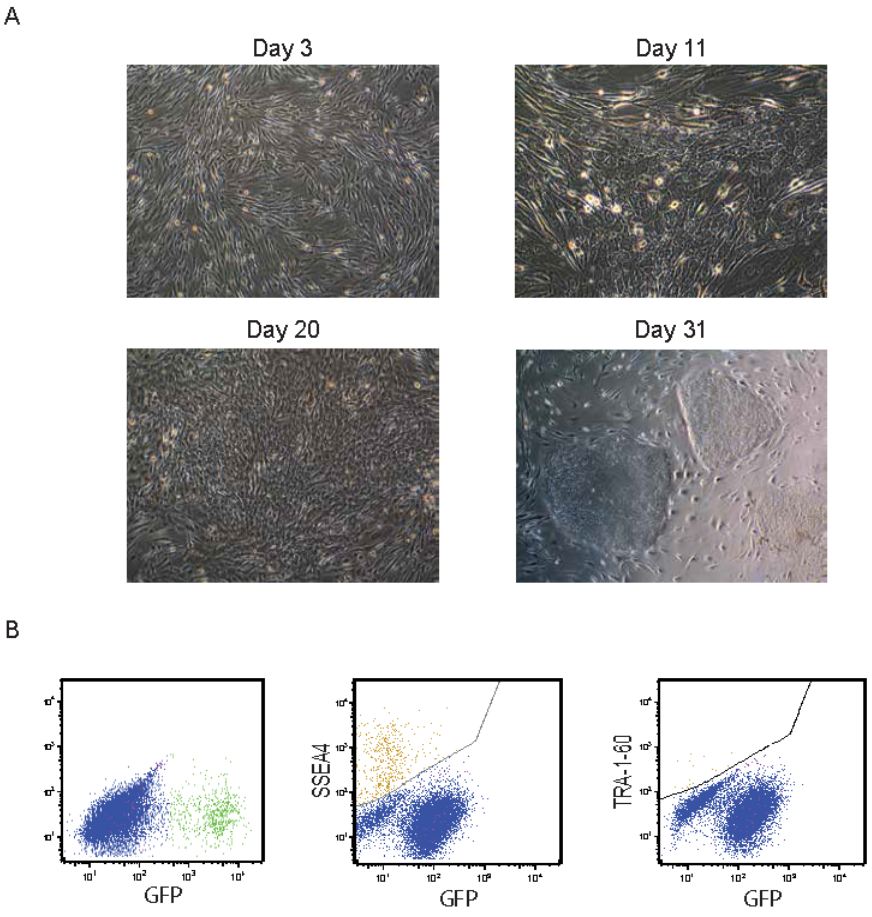
6.13 Data Analysis for Chapter 4

Visualization of the CyTOF dataset by viSNE and gating of cell populations was performed in Cytobank (www.cytobank.org). Hierarchical clustering, correlation analysis and modeling were performed as descried above with the exception that CyTOF dataset was downsampled to the same number and per-sample representation of cells and bootstrapped 100 times to avoid biases when comparing the two datasets. The results of the correlation analysis and model fits plot the average of these bootstrapped runs.

Chapter 7 Appendices

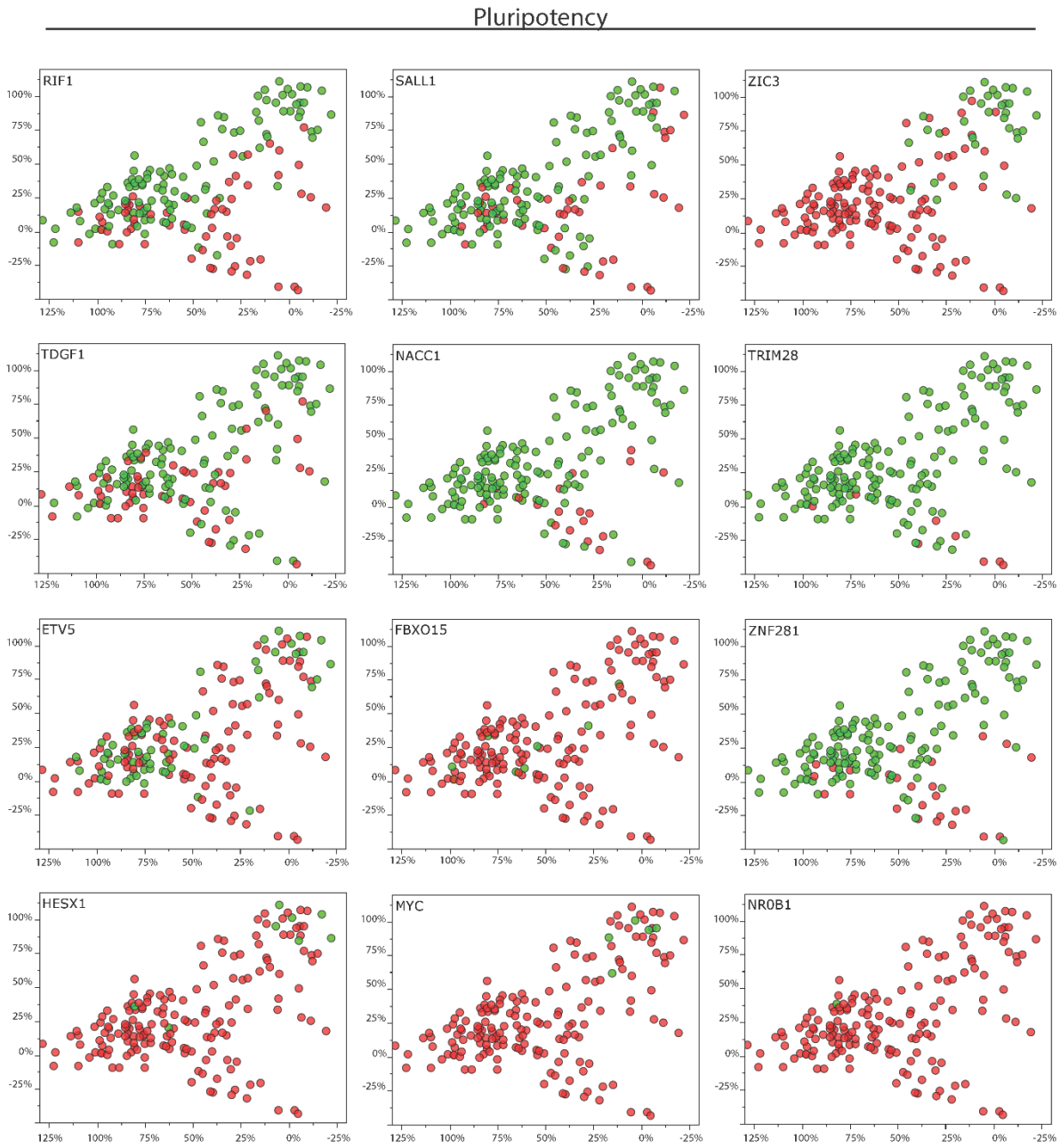
6.14 Supplemental Figures

6.14.1 Supplemental Figure 1

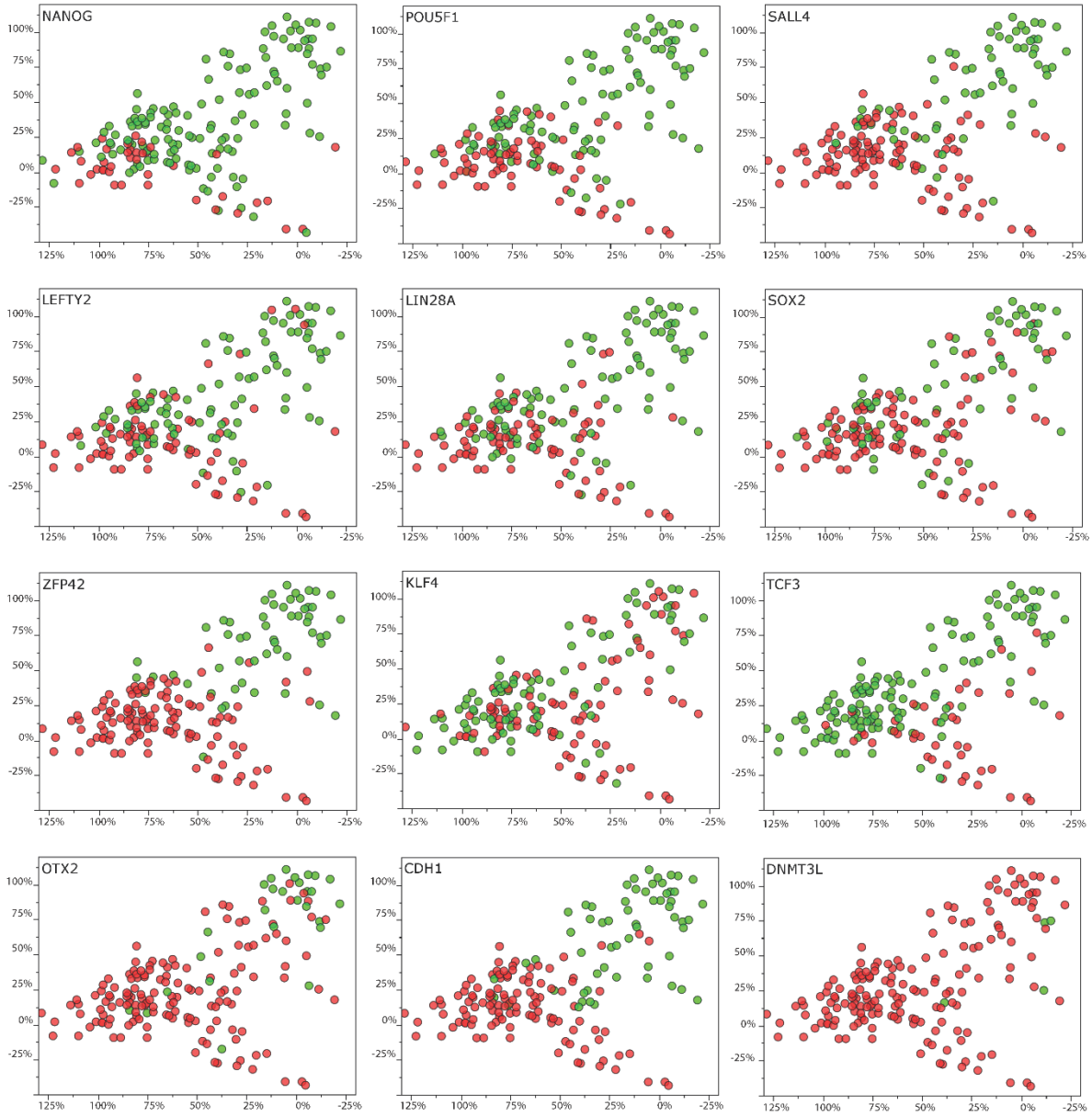


Supplemental Figure 1: **Overview of Experimental Design**
(A) Representative images of Human MRC-5 cells undergoing reprogramming at indicated time points post-infection with OSKM virus. (B) Representative FACS plots showing the gating scheme used for the isolation of GFP$^+$, SSEA4$^+$ and TRA-1-60$^+$ cells.

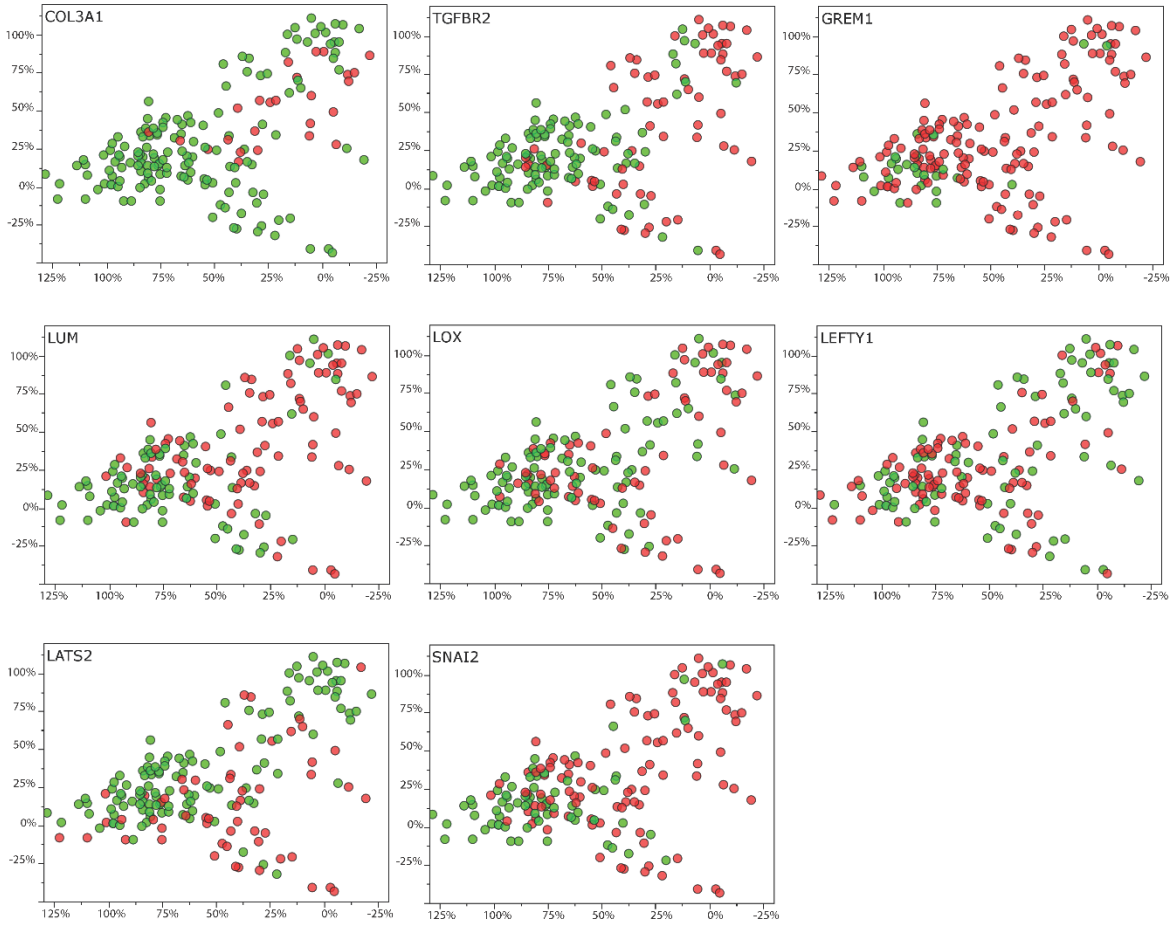
6.14.2 Supplemental Figure 2



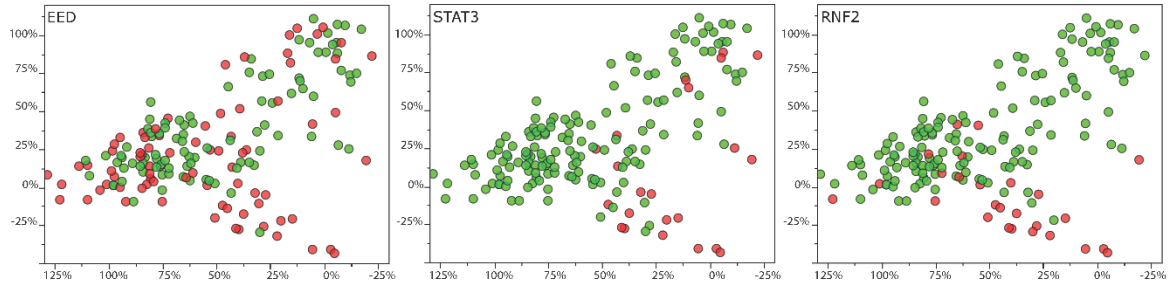
Pluripotency



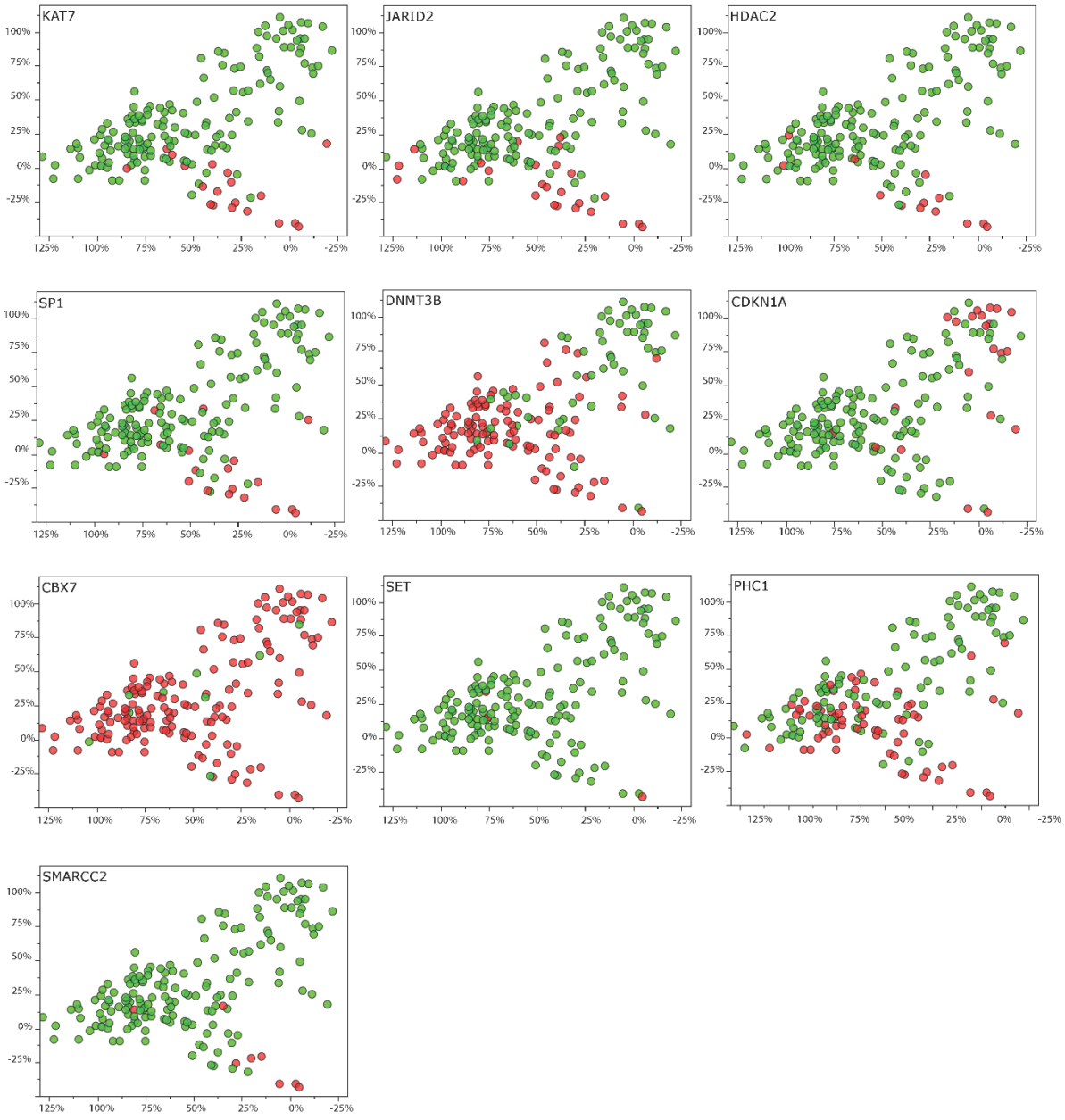
Fibroblast

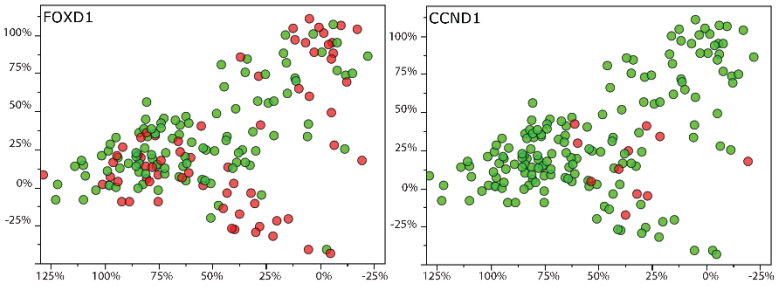


Chromatin Modifiers



Chromatin Modifiers

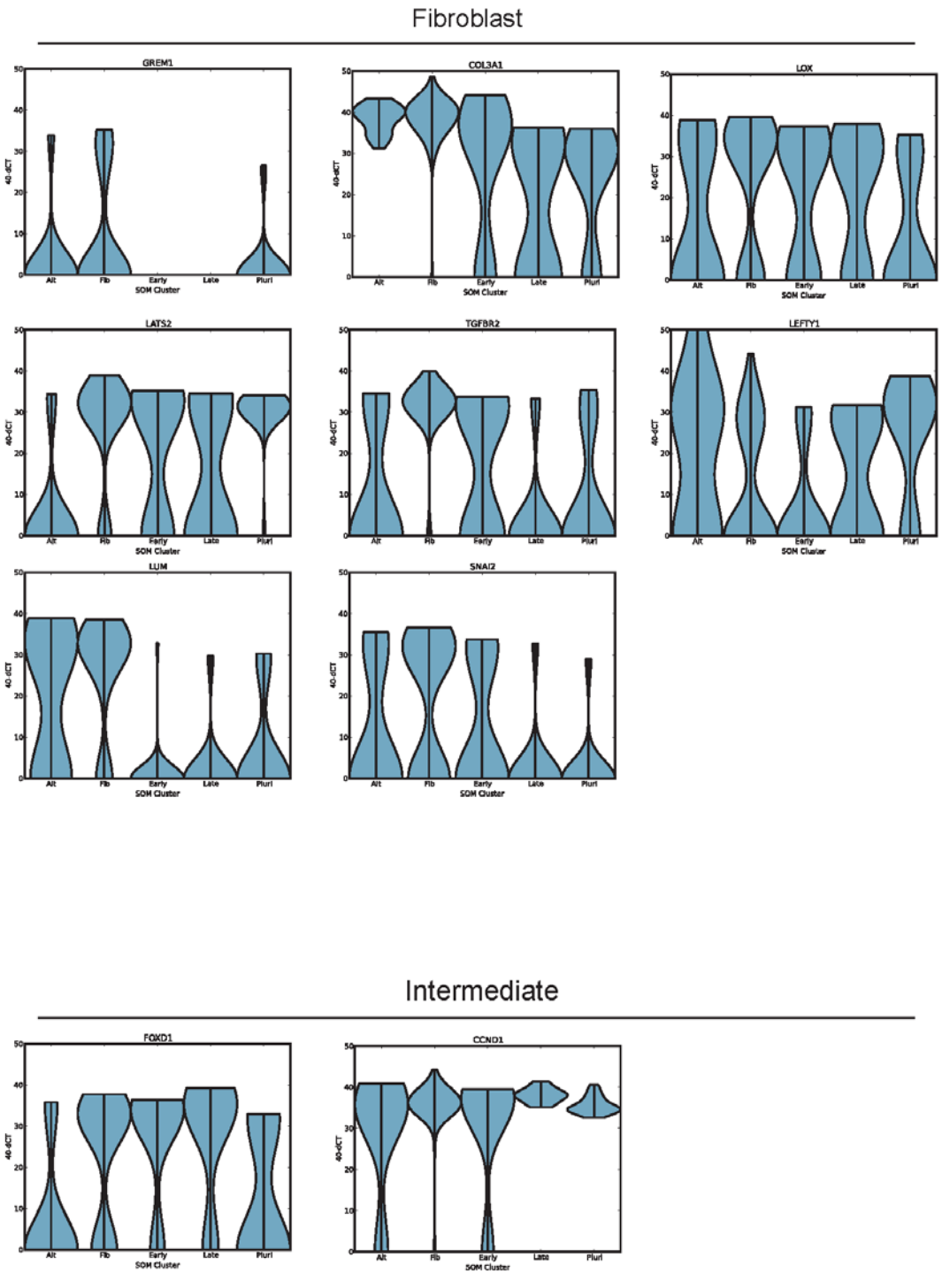


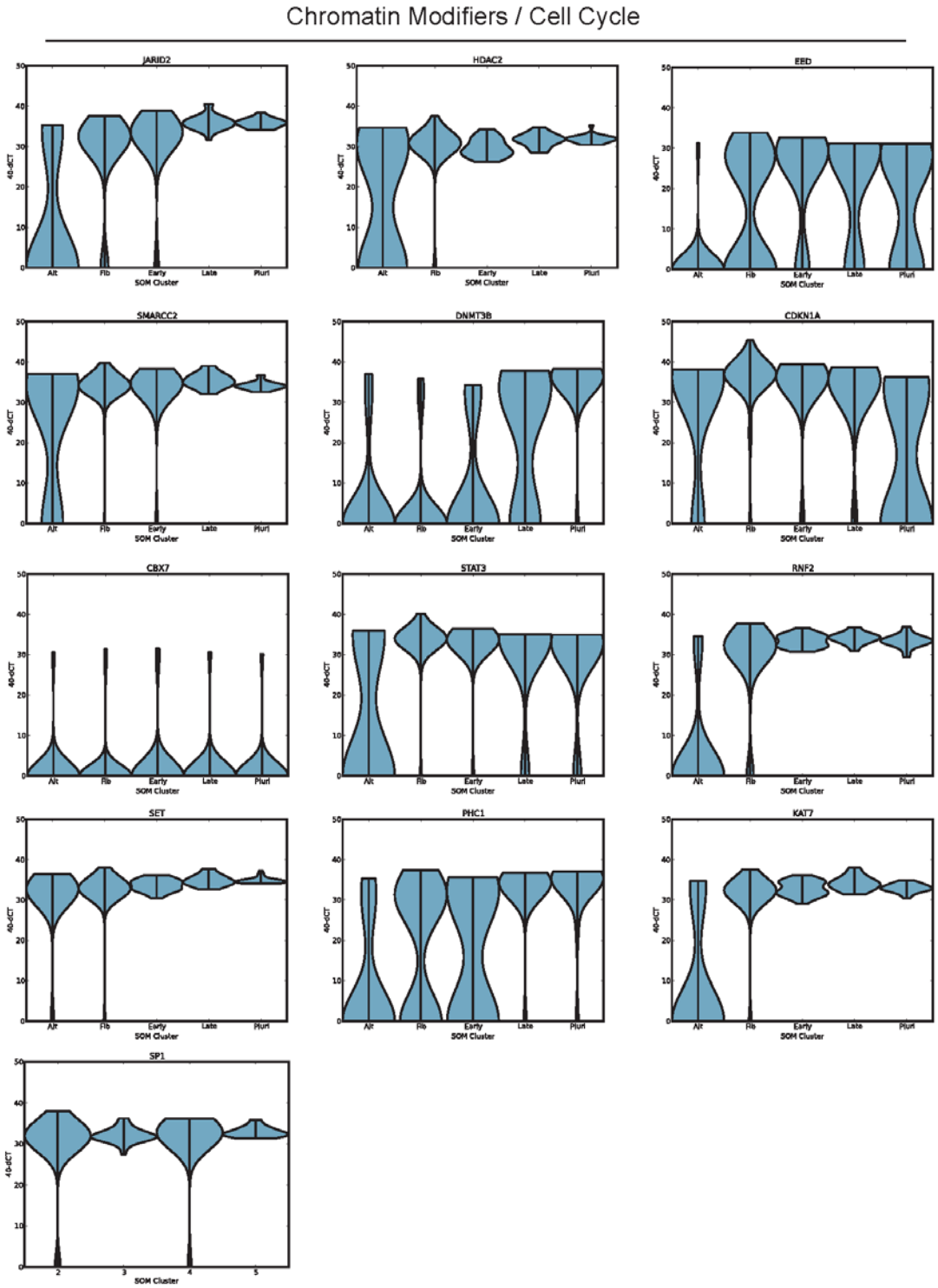


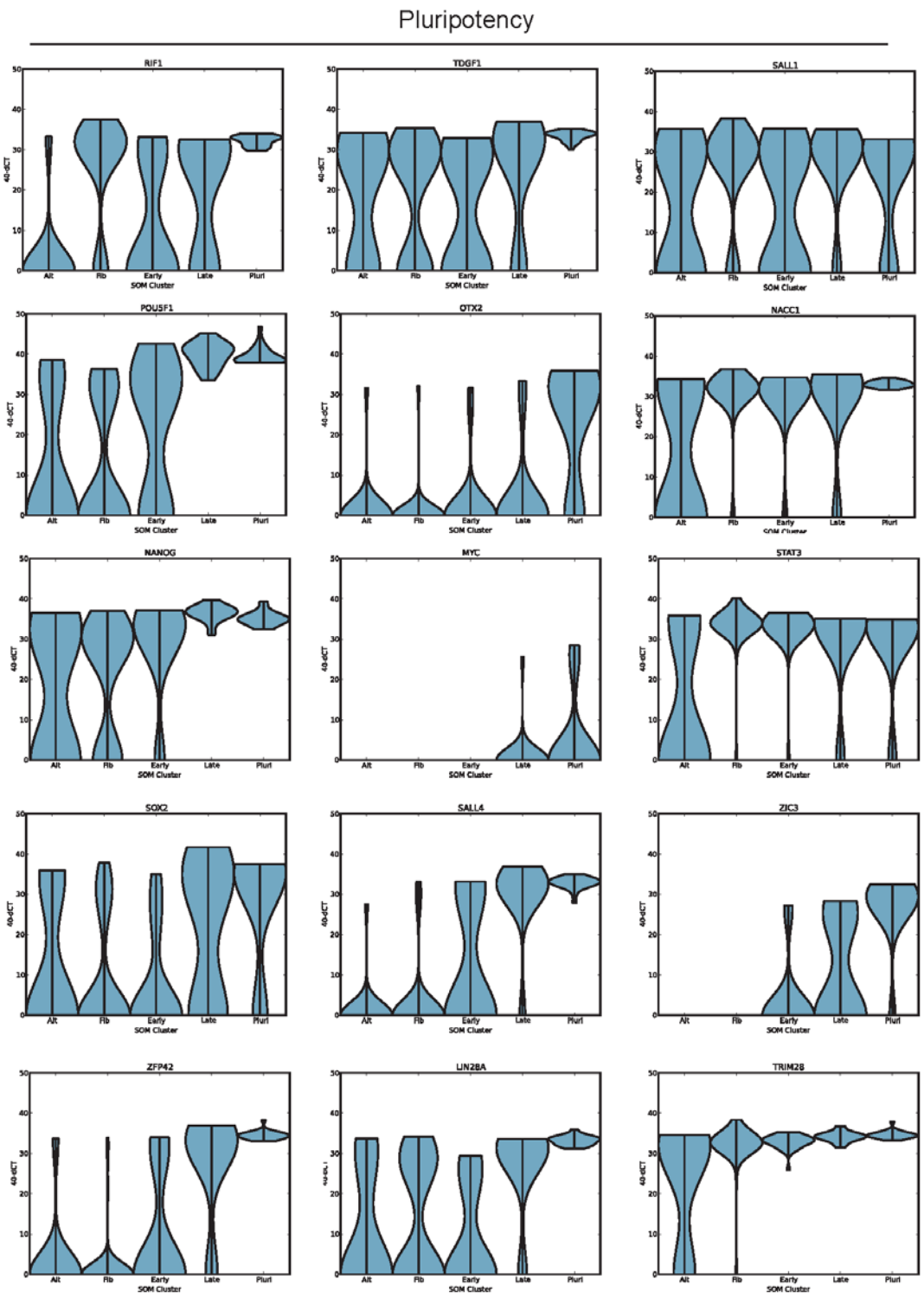
Supplemental Figure 2: Bubble Plots Demonstrating Qualitative Changes in Gene Expression During Reprogramming

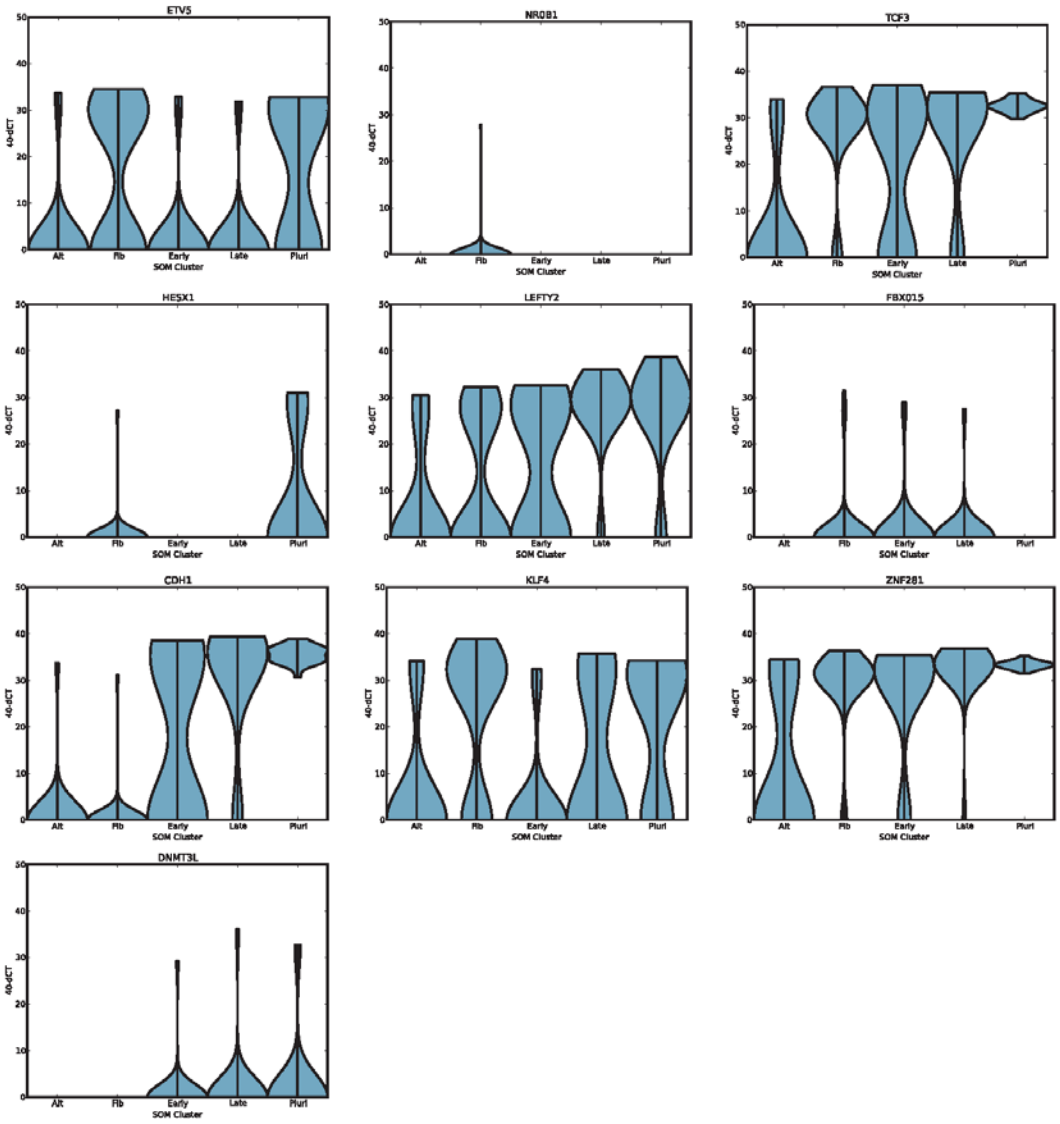
Bubble plots were generated using the relative fibroblast and relative H9 similarity metrics to plot the presence (green) or absence (red) of genes expression in a given cell. This view of the transcriptional dynamics during reprogramming reveals genes activated early, intermediately, or late in the process.Genes with no qualitative changes in expression are also observed.

6.14.3 Supplemental Figure 3




Chromatin Modifiers / Cell Cycle
JARID2
HDAC2
EED
SMARCC2
DNMT3B
CDKN1A
CBX7
STAT3
RNF2
SET
PHC1
KAT7
SP1
40-dCT
SOM Cluster
Alt
Fib
Early
Late
Pluri

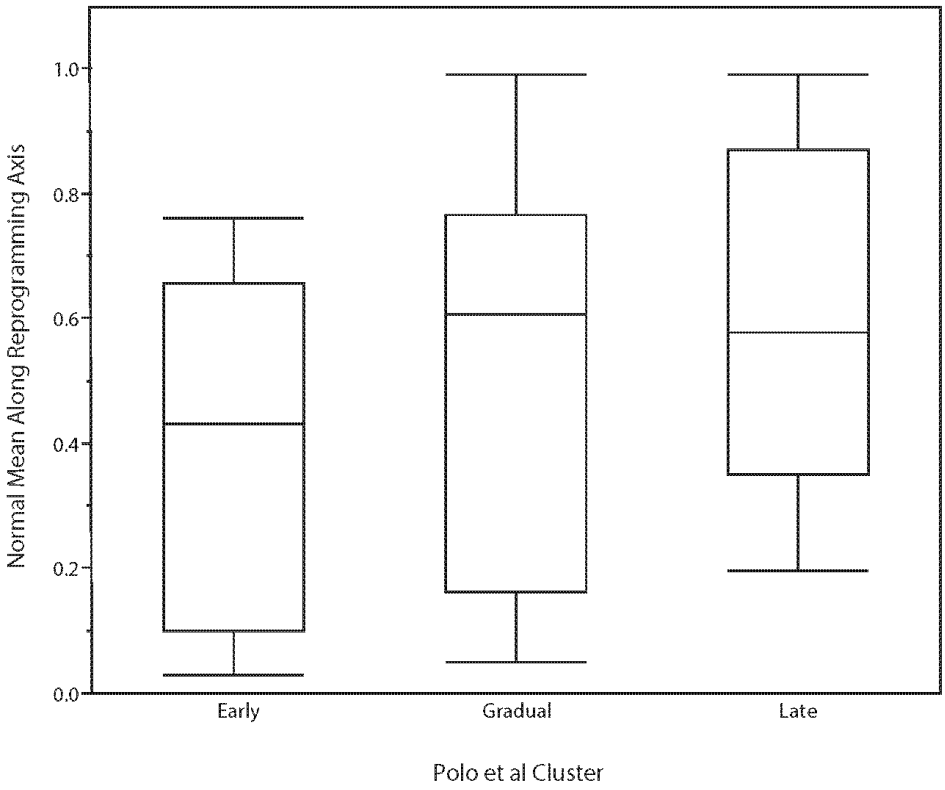

Pluripotency
RIF1
TDGF1
SALL1
POU5F1
OTX2
NACC1
NANOG
MYC
STAT3
SOX2
SALL4
ZIC3
ZFP42
LIN28A
TRIM28
40-dCT
SOM Cluster
Alt
Fib
Early
Late
Pluri



Supplemental Figure 3: Violin Plots Depicting Quantitative Changes in Gene Expression During Reprogramming

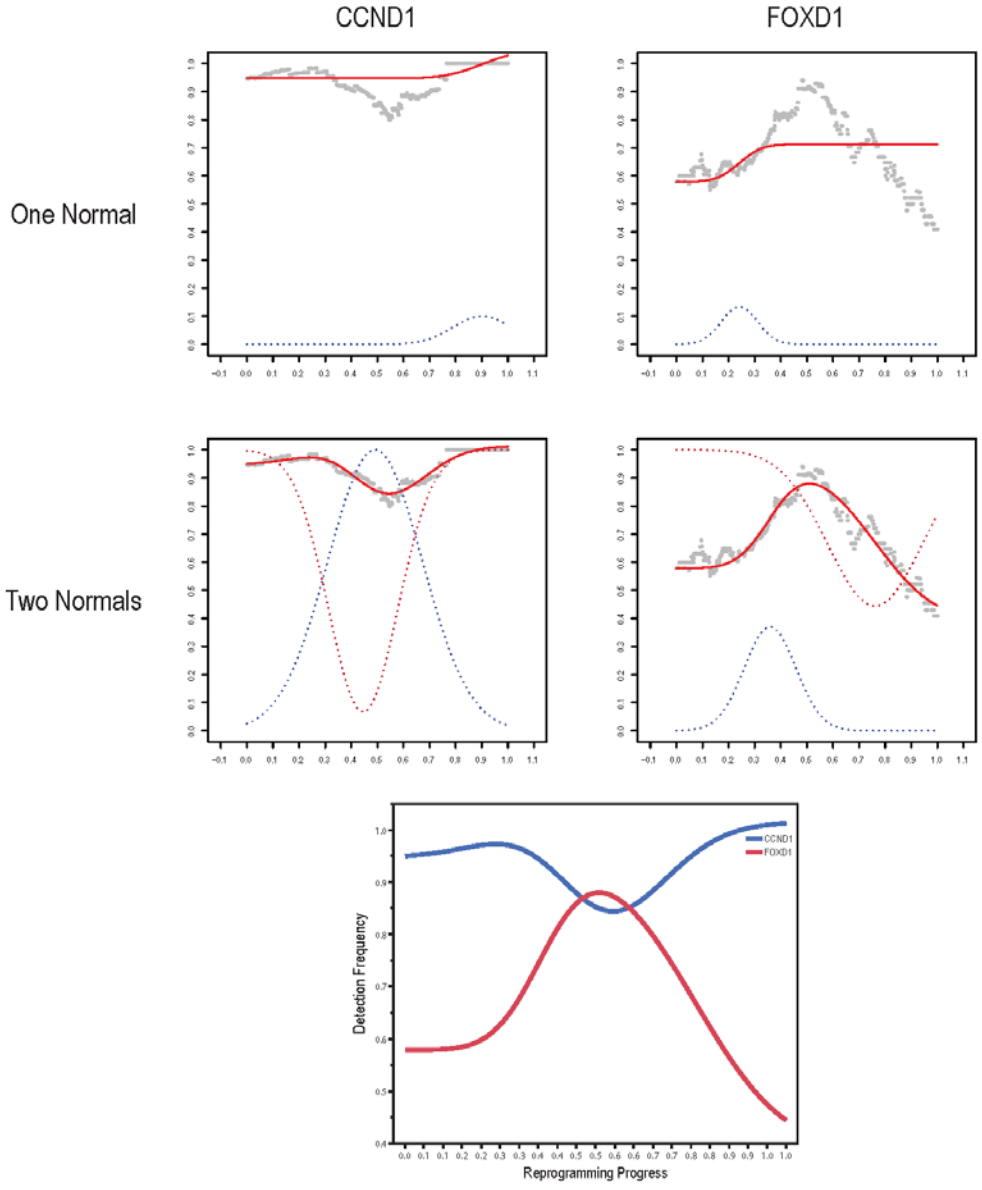
Cells were separated by SOM grouping (outliers excluded) and gene expression levels were plotted for each cell within the group. The width of the violin represents the distribution of cells across expression levels. For the majority of genes, there is a clear inflection point in these graphs where a large number of cellsup or down regulate a given transcript when transitioning from one SOM group to another, eventually coalescing around high or low expression levels as cells approach pluripotency and cellular phenotypes stabilize.

6.14.4 Supplemental Figure 4



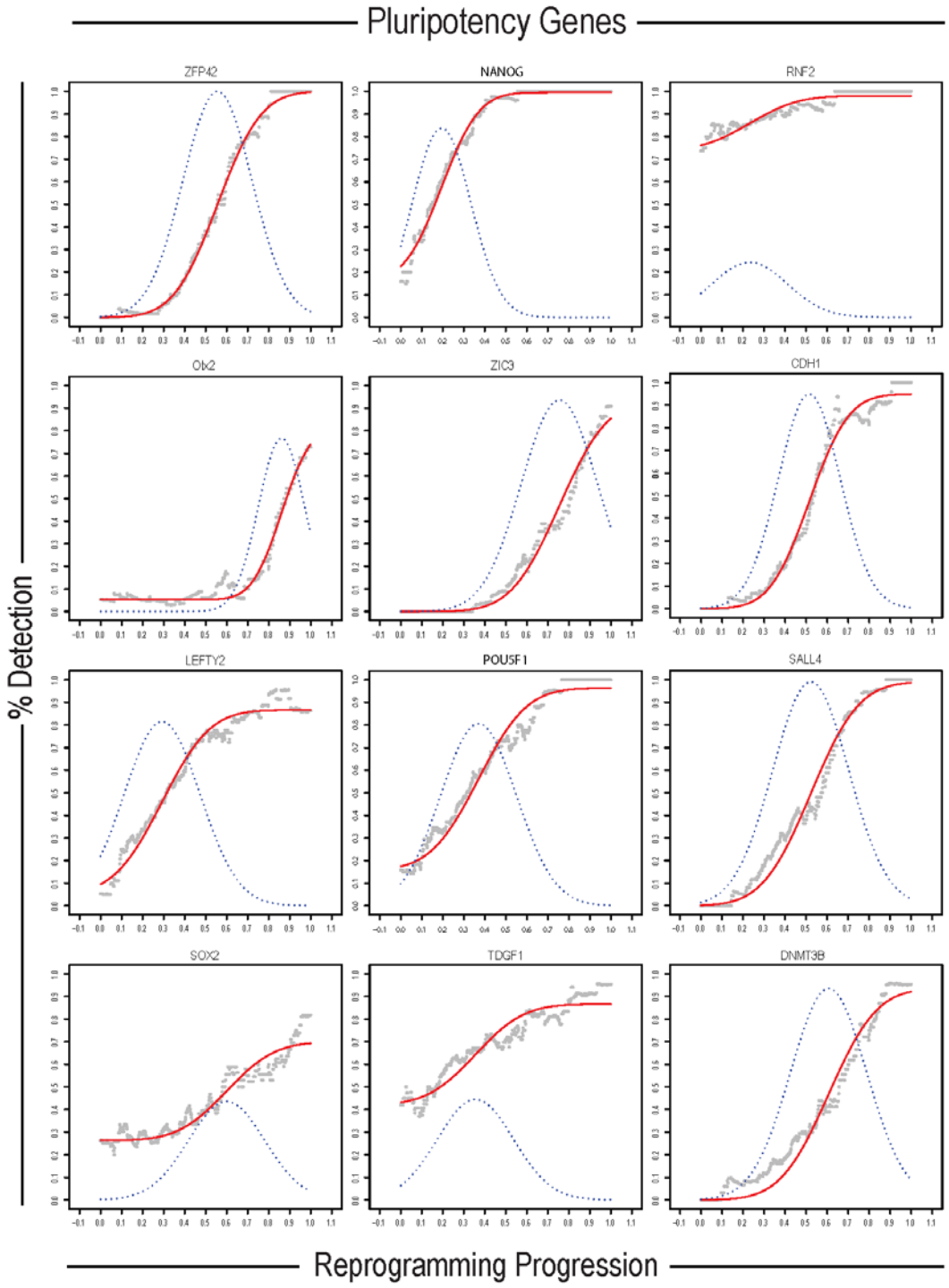
Supplemental Figure 4: Comparison of timing of gene activation / inactivation with Polo et all 2012. Using the cluster definitions provided in Polo et al genes whose expression increased or decreased Early, Gradually or Late were collapsed into a single Early, Gradual or Late cluster. We then compared these clusters to the mean of the normal distribution for each gene as defined in our model as shown in the box and whisker plots above.

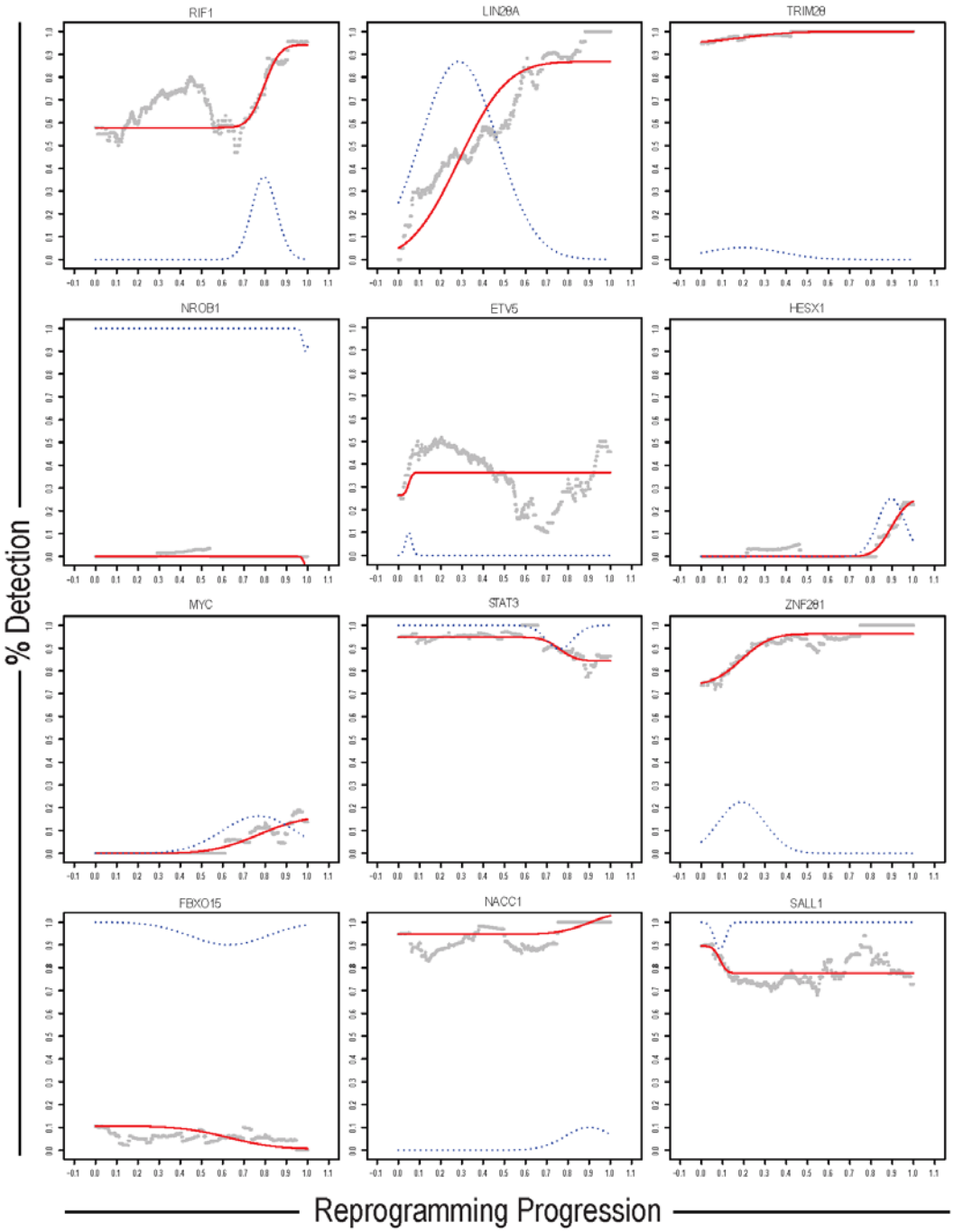
6.14.5 Supplemental Figure 5

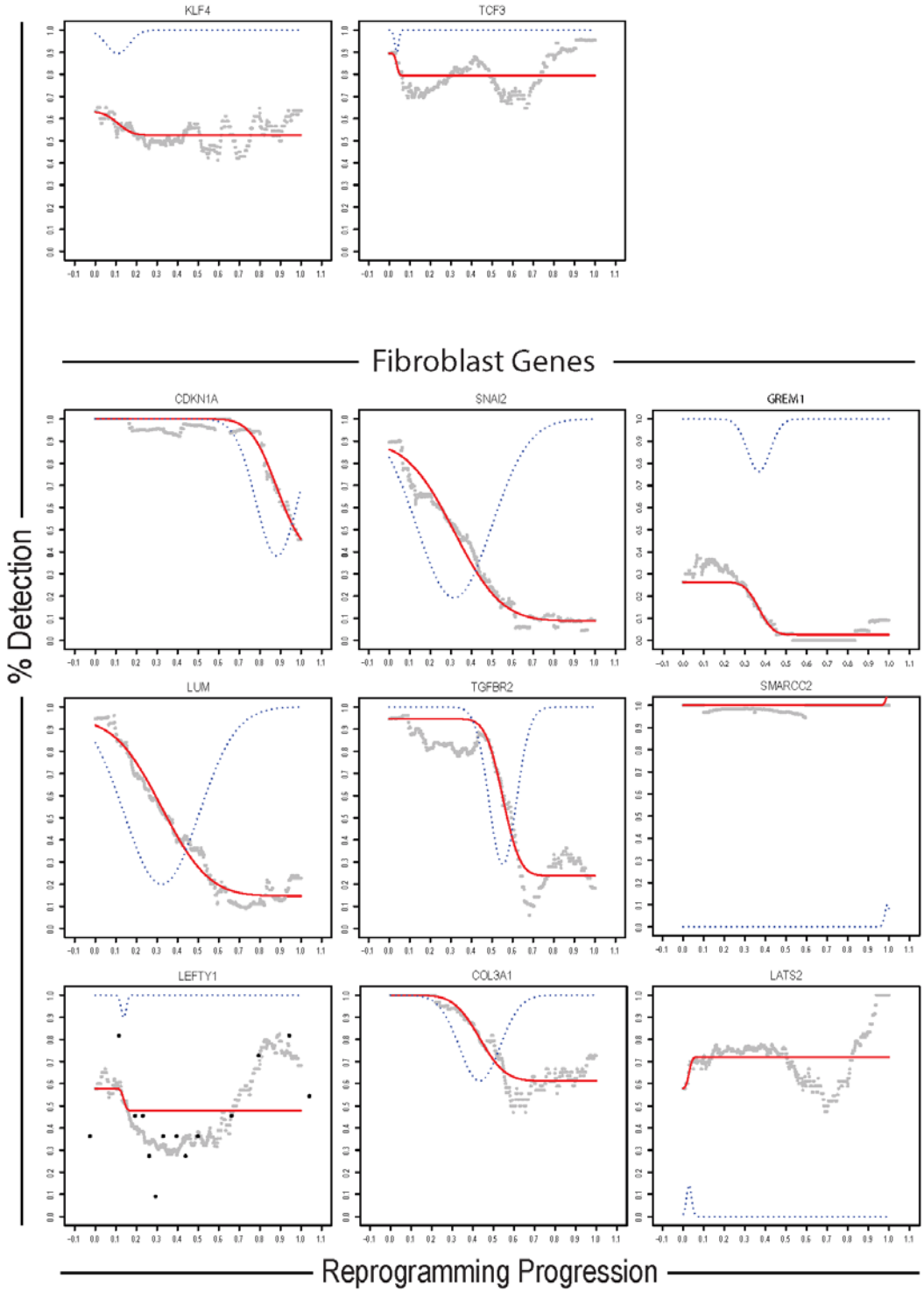


Supplemental Figure 5: Modeling Complex Gene Behavior with Two Gaussian Distributions. Genes with complex behavior (transiently activated or inactivated) were modeled using one (top panel) or two (bottom panel) Gaussian distributions. Poor fit is observed using a single Gaussian curve, however accounting for both activating (blue dotted line) and repressive (red dotted line) events using two curves results in excellent fit to the data. A combined view of these expression probabilities is shown in the bottom panel.

6.14.6 Supplemental Figure 6




RIF1
LIN28A
TRIM28
NR0B1
ETV5
HESX1
MYC
STAT3
ZNF281
FBXO15
NACC1
SALL1
% Detection
Reprogramming Progression


KLF4
TCF3
Fibroblast Genes
CDKN1A
SNAI2
GREM1
LUM
TGFBR2
SMARCC2
LEFTY1
COL3A1
LATS2
% Detection
Reprogramming Progression

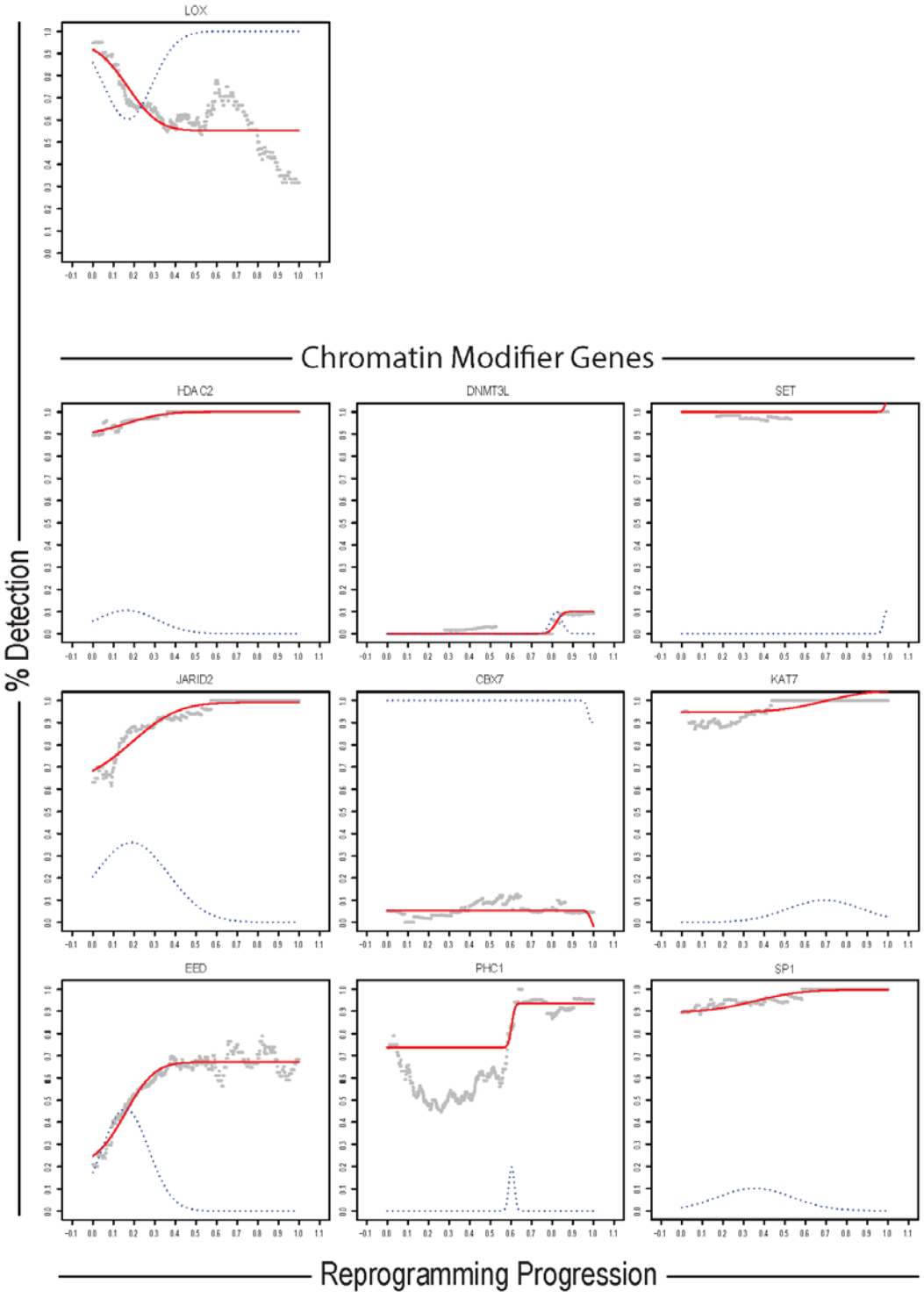
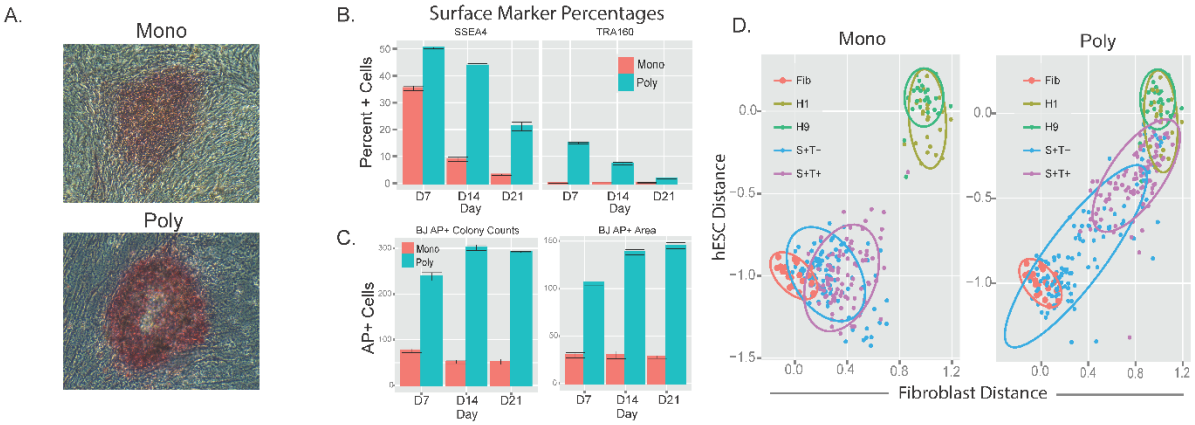

LOX
Chromatin Modifier Genes
HDAC2
DNMT3L
SET
JARID2
CBX7
KAT7
EED
PHC1
SP1
% Detection
Reprogramming Progression

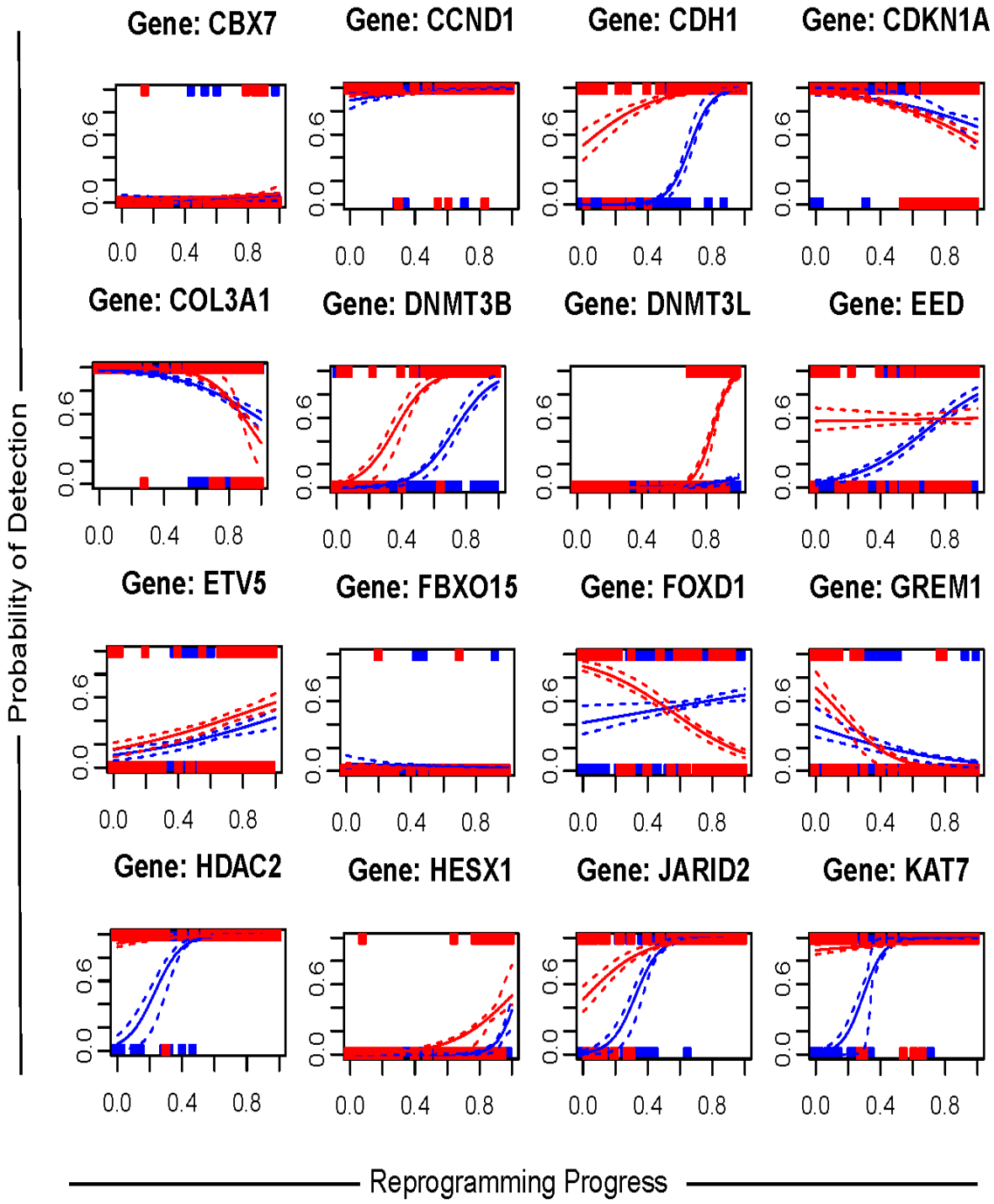
Figure S5: **Modeling Gene Expression Dynamics Using Gaussian Distributions.** Models depict the observed detection frequency (grey dots) along the Reprogramming Progression Axis using a sliding window analysis as described in Methods. Red lines depict the model fit resulting from the underlying normal distribution (blue dotted line).

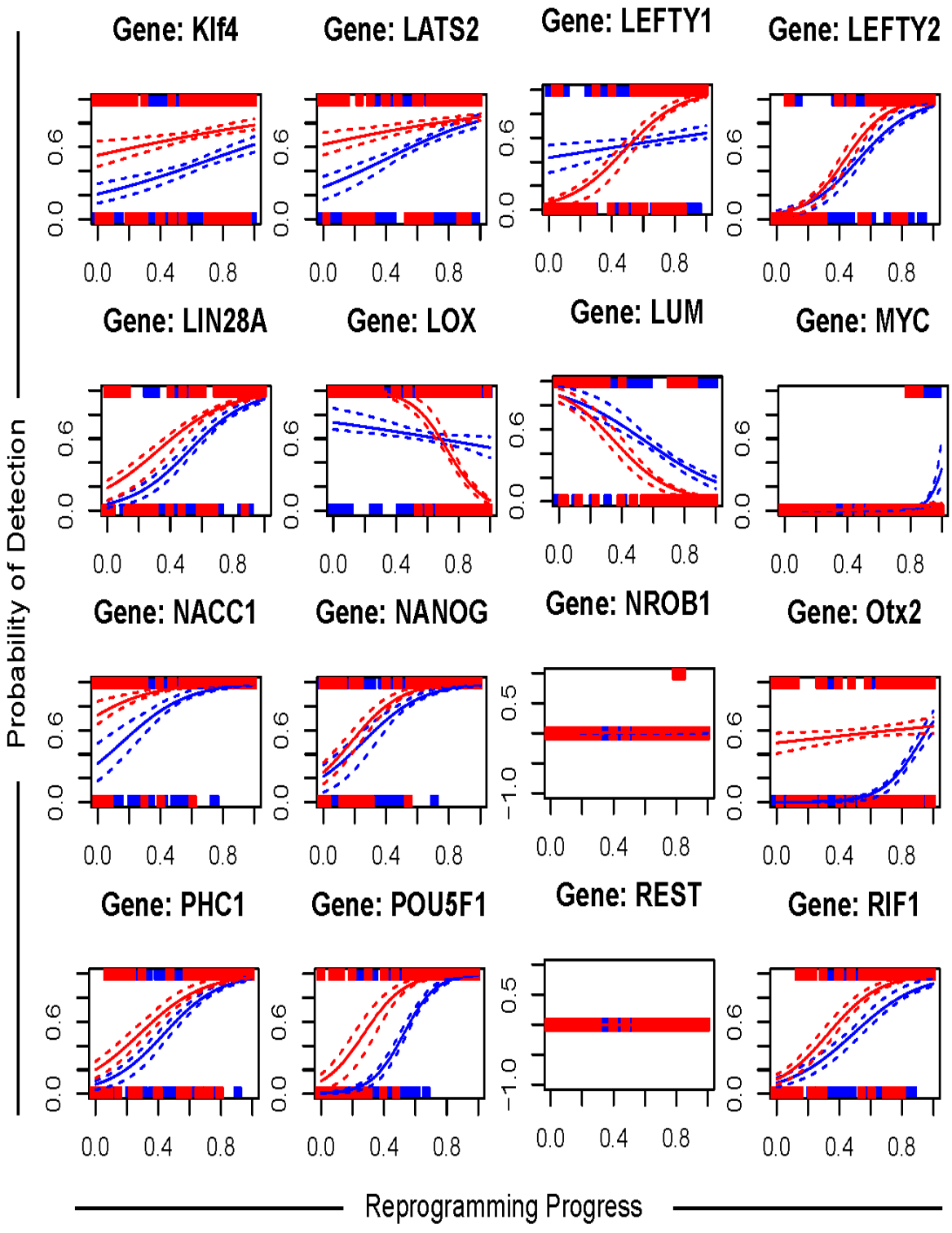
6.14.7 Supplemental Figure 7

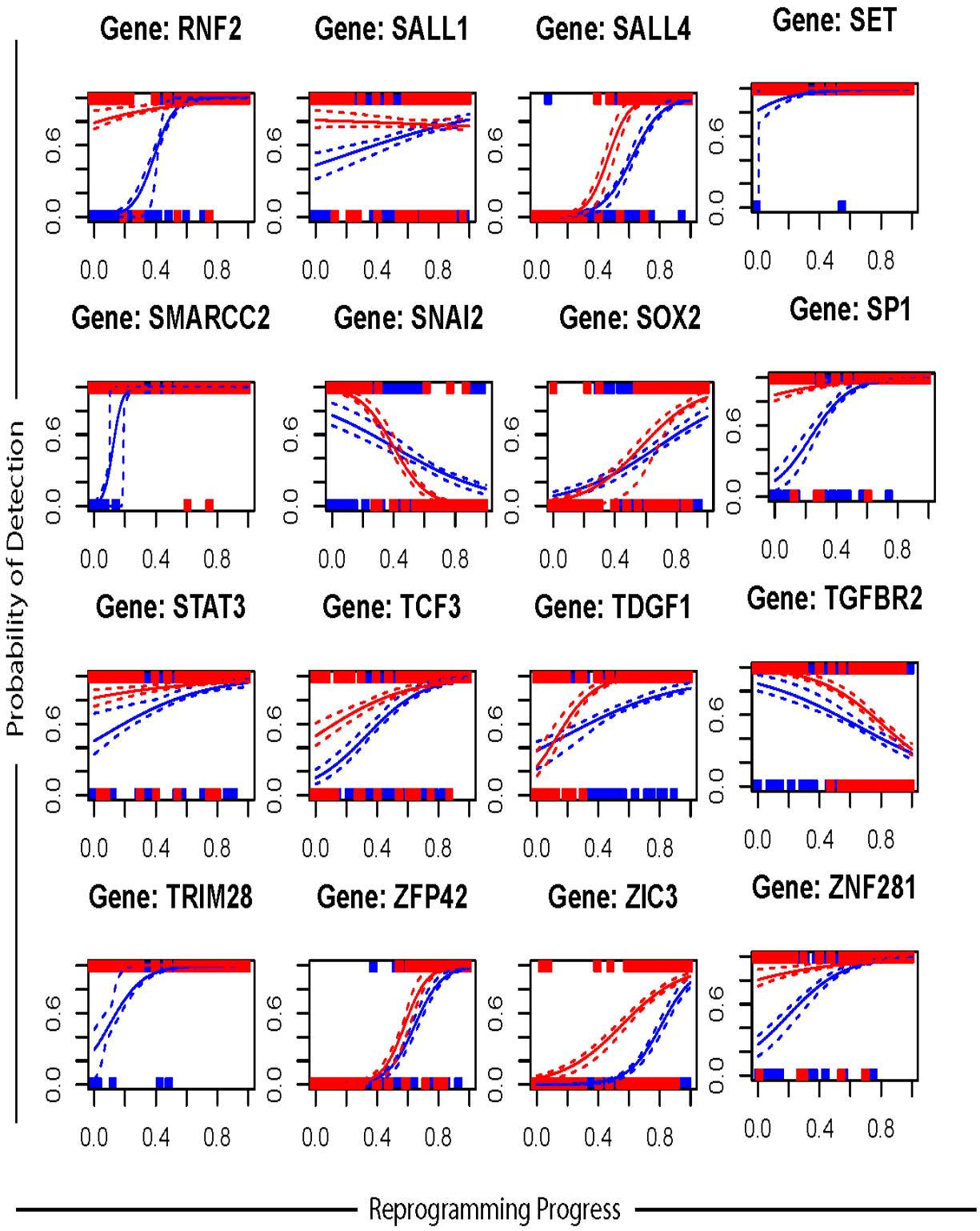


Supplemental Figure 7: Efficiency and trajectory of mono- and polycistronic reprogramming in BJ fibroblasts. (A) Representa-tive images of AP+ colonies resulting from Mono- (Top) and Polycistronic (Bottom) reprogramming. (B) Quantification of SSEA4 (left) and TRA-1-60+ (right) cells in Mono and Poly conditions by antibody staining and FACS. (C) Total number (left) and area (right) of AP+ colonies measured at days 7, 14 and 21 of reprogramming. (D) Euclidean trajectory of BJ fibroblasts reprogrammed with Mono (left) and Polycistronic (right) viruses. Ovals represent 90% coverage of the indicated population.

6.14.8 Supplemental Figure 8

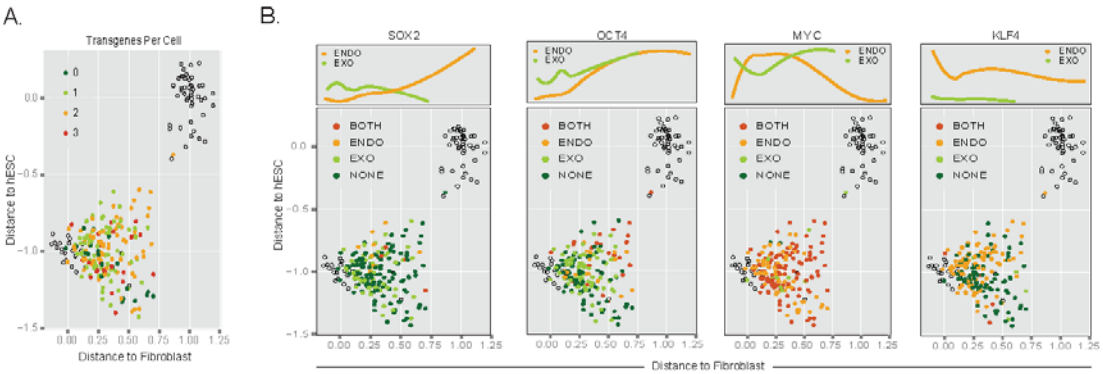



Gene: Klf4
Gene: LATS2
Gene: LEFTY1
Gene: LEFTY2
Gene: LIN28A
Gene: LOX
Gene: LUM
Gene: MYC
Gene: NACC1
Gene: NANOG
Gene: NROB1
Gene: Otx2
Gene: PHC1
Gene: POU5F1
Gene: REST
Gene: RIF1
Probability of Detection
Reprogramming Progress


Gene: RNF2
Gene: SALL1
Gene: SALL4
Gene: SET
Gene: SMARCC2
Gene: SNAI2
Gene: SOX2
Gene: SP1
Gene: STAT3
Gene: TCF3
Gene: TDGF1
Gene: TGFBR2
Gene: TRIM28
Gene: ZFP42
Gene: ZIC3
Gene: ZNF281
Probability of Detection
Reprogramming Progress

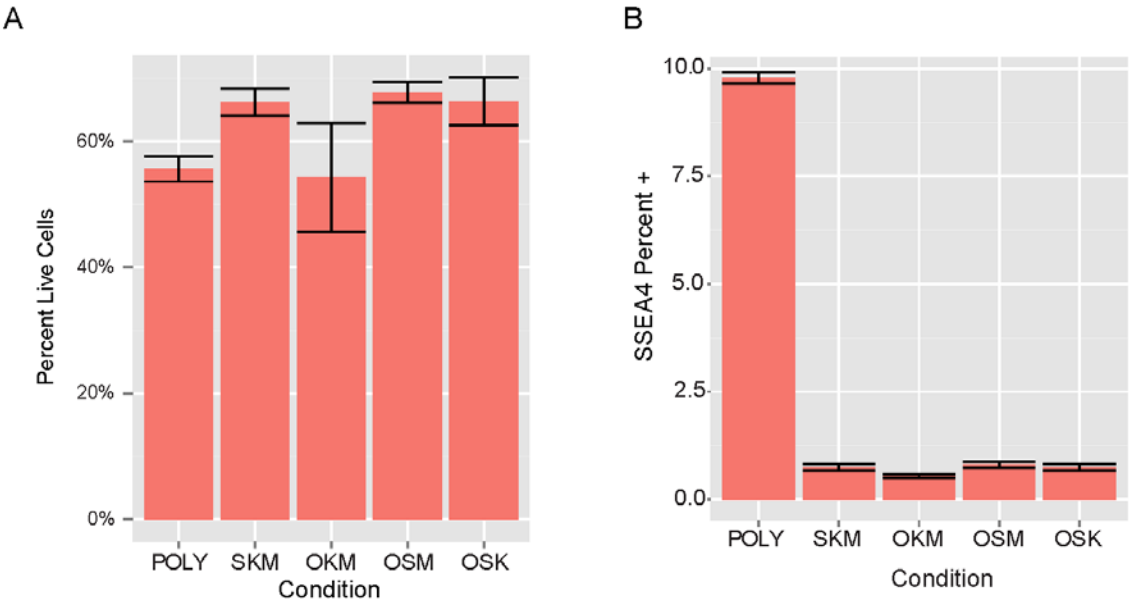
Supplemental Figure 8: Gene expression models of mono- and polycistronic reprogramming. Gene expression data for cells along the reprogramming traecjtory was binarized and fit with logistic regression curves to model their probability of expression as reprogramming progresses. Fit curves are shown for both monocistronic (blue) and polycistronic (red) reprogramming conditions.

6.14.9 Supplemental Figure 9



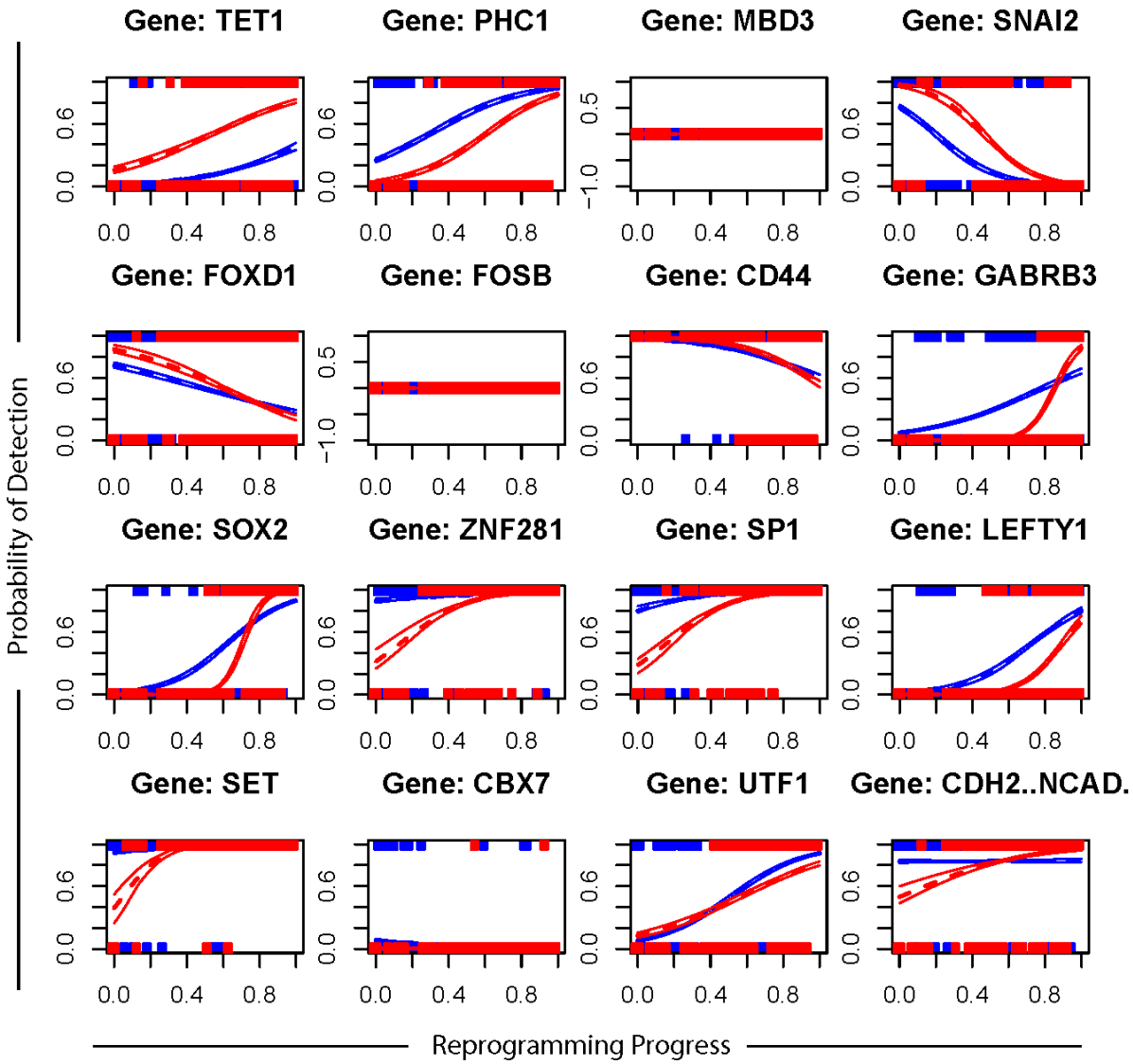
Supplemental Figure 9: Heteroeneous expression of OSKM factors in BJ fibroblasts. (A) OSKM trans-gene content was measured in each cell by qPCR throughout the reprogramming trajectory, using SYBR green primers specific to the exogenous copy. (B,bottom panels) Total (endoge-nous and exogenous) transgene content measured in each cell for each factor individually. Splines depict a smoothed distribution of endogenous (orange) and exogenous (green) transgene content along the reprogramming progress axis (B, bottom panel).

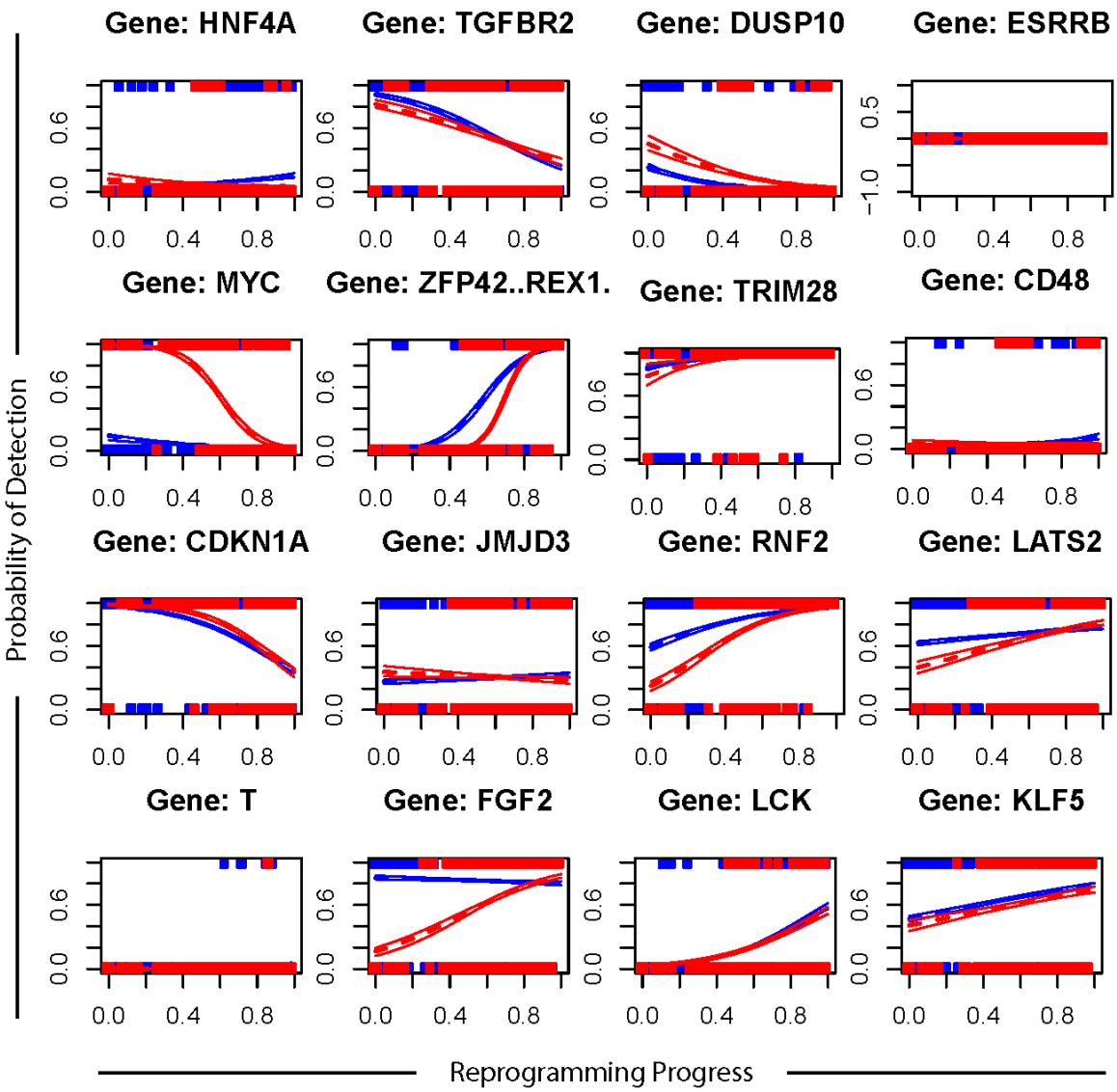
6.14.10Supplemental Figure 10

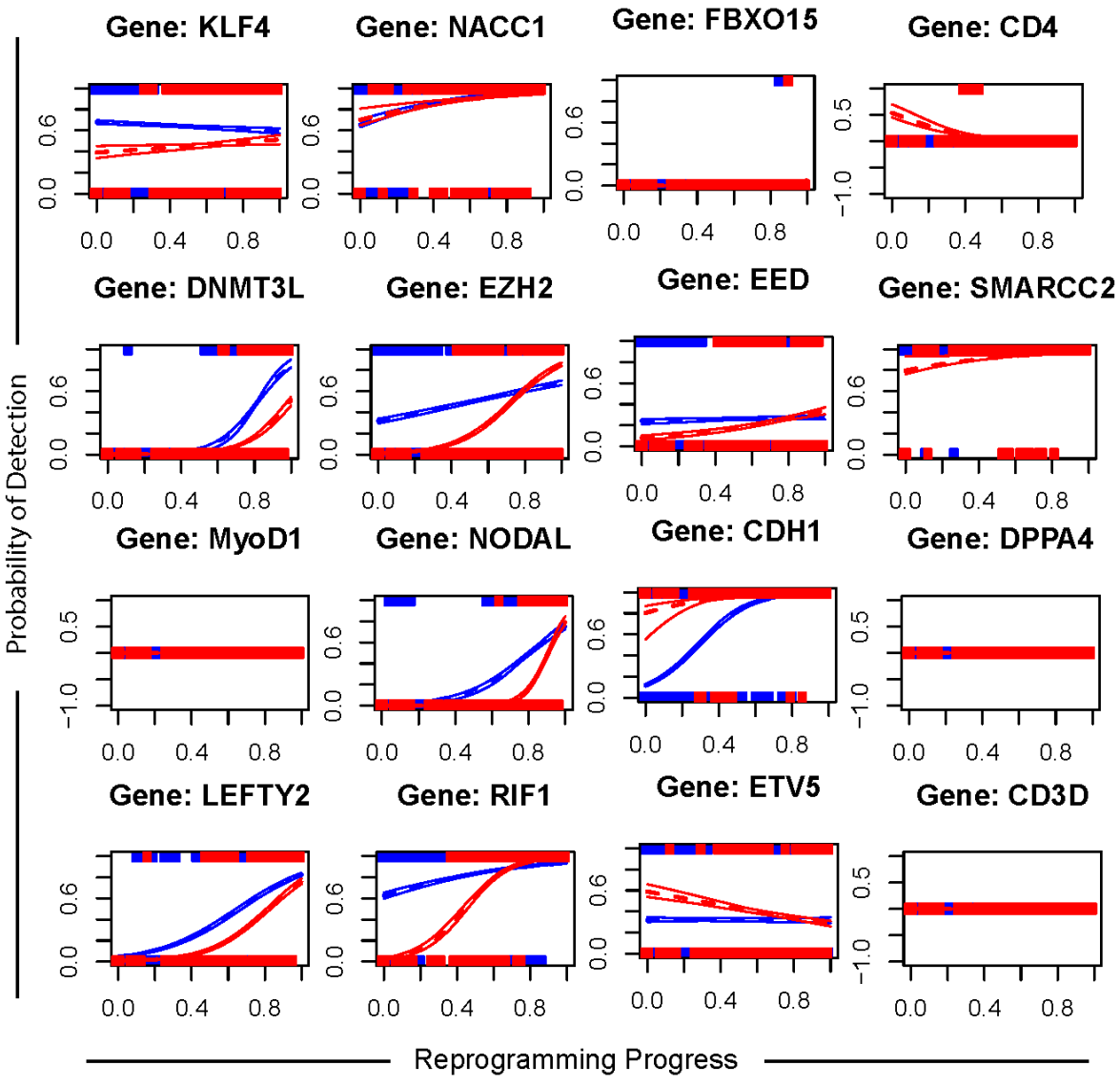


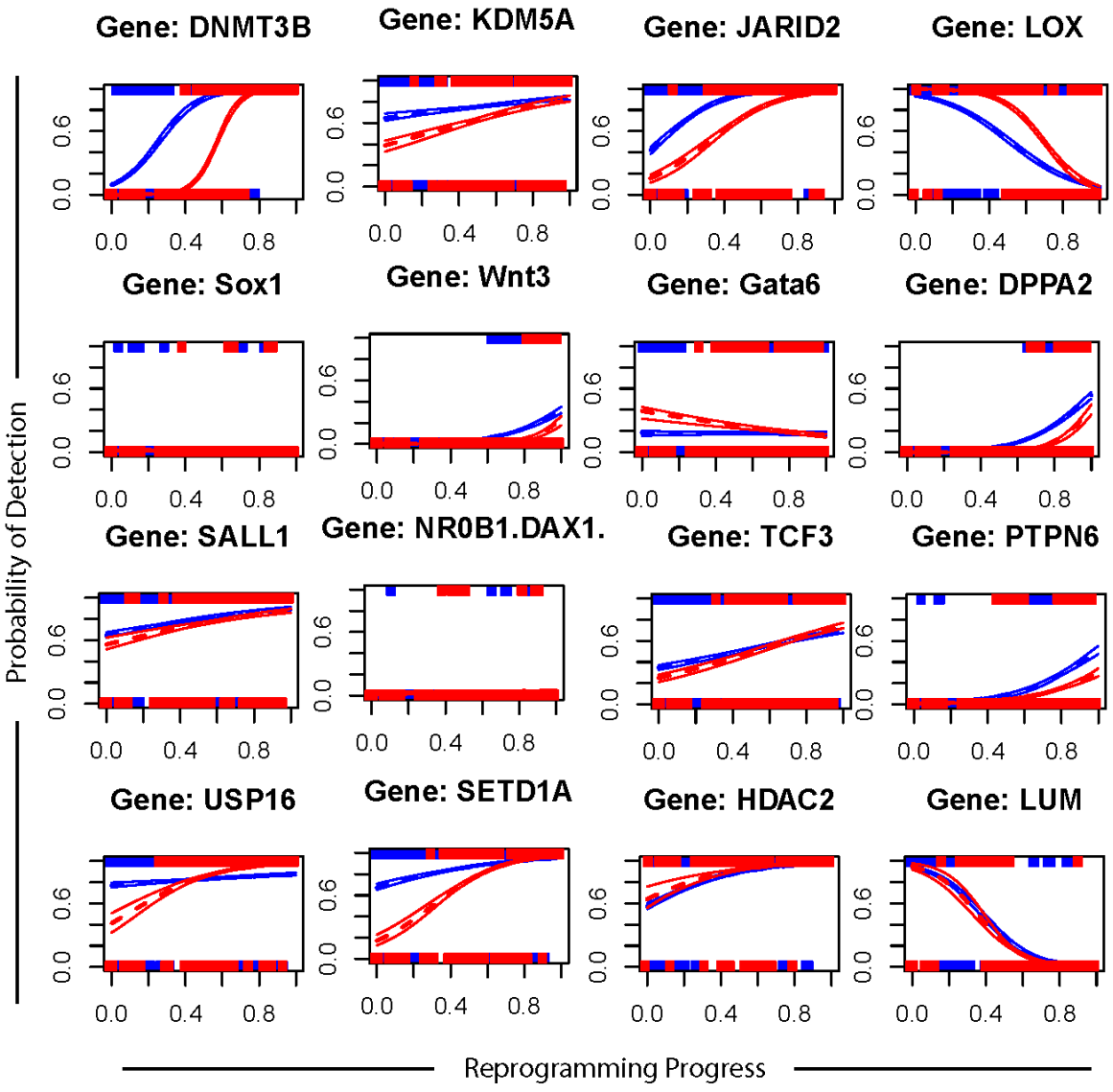
Supplemental Figure 10: FACS data from three-factor reprogramming experiments. The percent live cells from each condition was determined by FACS analysis of PI negative events (A). Cells were stained with anti-SSEA4 antibody and the percent positive cells were quantified by FACS (B).

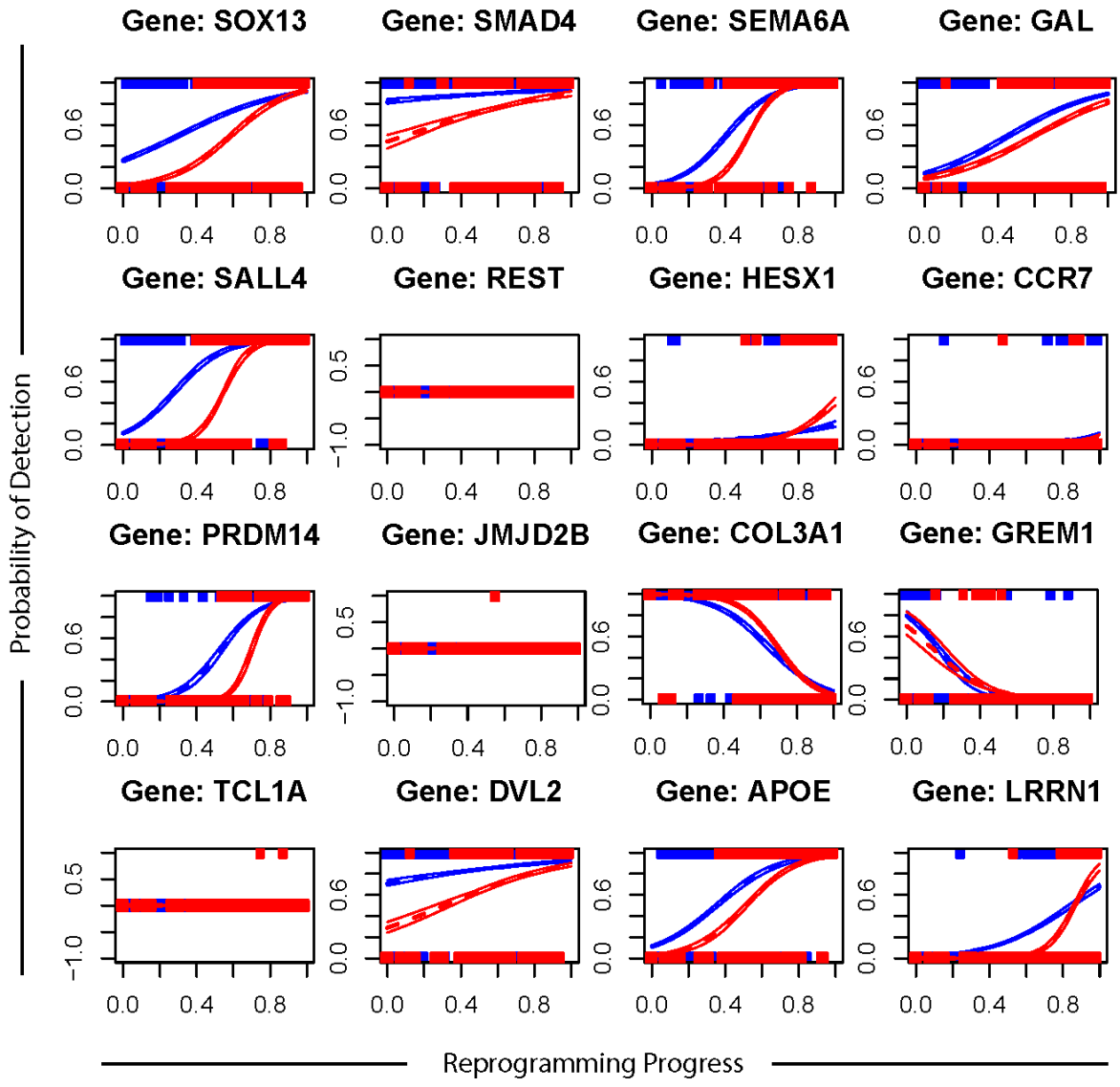
6.14.11Supplemental Figure 11

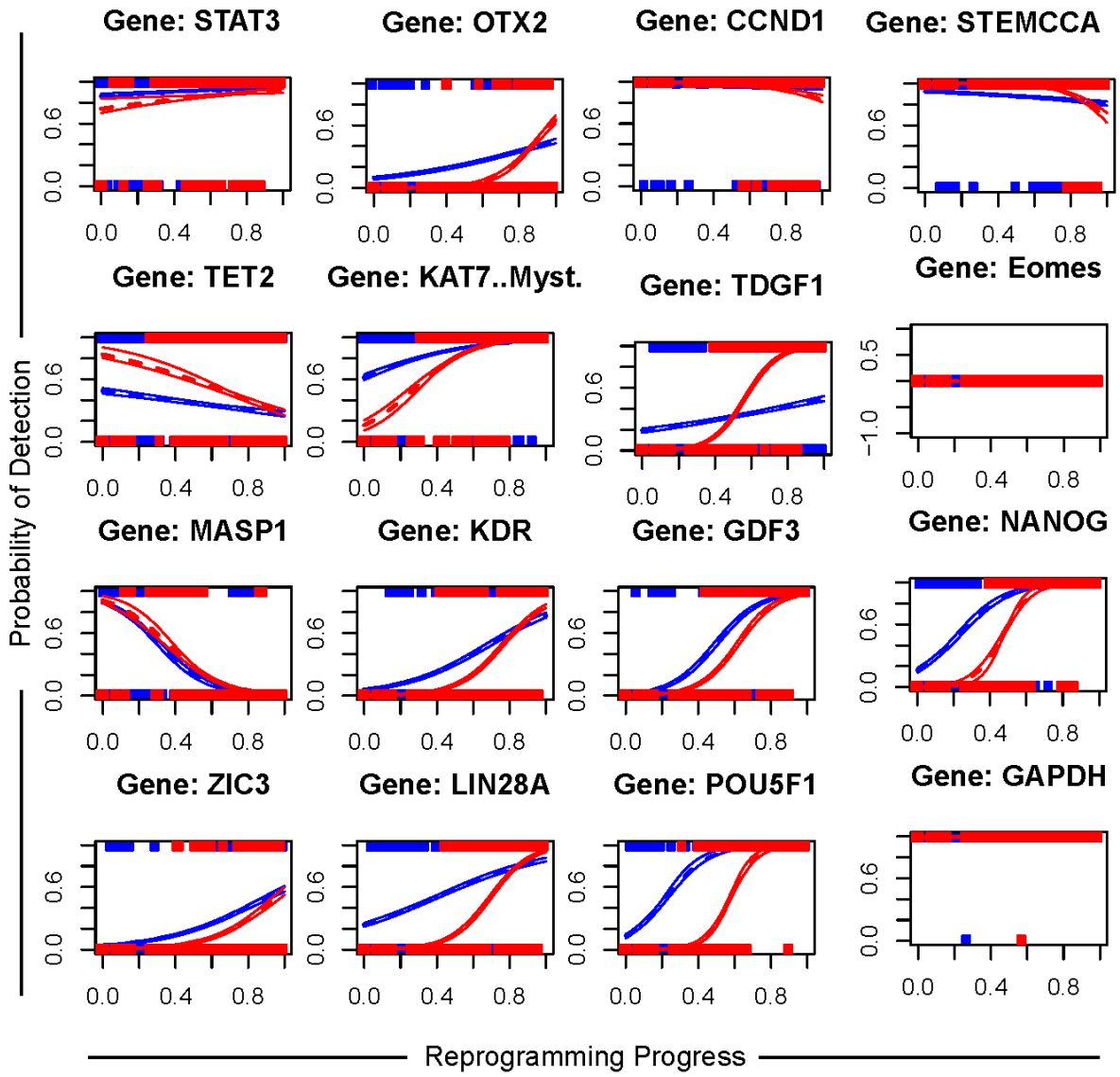



Gene: HNF4A
Gene: TGFBR2
Gene: DUSP10
Gene: ESRRB
Gene: MYC
Gene: ZFP42..REX1.
Gene: TRIM28
Gene: CD48
Gene: CDKN1A
Gene: JMJD3
Gene: RNF2
Gene: LATS2
Gene: T
Gene: FGF2
Gene: LCK
Gene: KLF5
Probability of Detection
Reprogramming Progress


Gene: KLF4
Gene: NACC1
Gene: FBXO15
Gene: CD4
Gene: DNMT3L
Gene: EZH2
Gene: EED
Gene: SMARCC2
Gene: MyoD1
Gene: NODAL
Gene: CDH1
Gene: DPPA4
Gene: LEFTY2
Gene: RIF1
Gene: ETV5
Gene: CD3D
Probability of Detection
Reprogramming Progress


Gene: DNMT3B
Gene: KDM5A
Gene: JARID2
Gene: LOX
Gene: Sox1
Gene: Wnt3
Gene: Gata6
Gene: DPPA2
Gene: SALL1
Gene: NR0B1.DAX1.
Gene: TCF3
Gene: PTPN6
Gene: USP16
Gene: SETD1A
Gene: HDAC2
Gene: LUM
Probability of Detection
Reprogramming Progress
0.0
0.6
0.4
0.8


Gene: SOX13
Gene: SMAD4
Gene: SEMA6A
Gene: GAL
Gene: SALL4
Gene: REST
Gene: HESX1
Gene: CCR7
Gene: PRDM14
Gene: JMJD2B
Gene: COL3A1
Gene: GREM1
Gene: TCL1A
Gene: DVL2
Gene: APOE
Gene: LRRN1
Probability of Detection
Reprogramming Progress



Supplemental Figure 11: Model fits of BJ and MRC5 Polycistronic Reprogramming. Logistic regression curves were fit to binarized expression data for each gene along the reprogramming axis from fibroblast to hESC for both MRC-5 (red) and BJ(blue) fibroblast reprogrammed by the polycistronic method.

6.15 Supplemental Tables

6.15.1 Supplemental Table 1

	FIBROBLAST	GFP+ DAY 4	GFP+ DAY 8	GFP+ DAY 14	SSEA4+ DAY 4	SSEA4+ DAY 8	SSEA4+ DAY 14	TRA-1-60+	CDH1+	HESC
# USED IN ANALYSIS	15	14	15	15	16	16	13	48	16	15

Supplemental Table 1: Phenotype and number of cells collected by FACS and used for analysis

6.15.2 Supplemental Table 2

Gene Name	Taqman Assay ID
CBX7	Hs00545603_m1
CCND1	Hs00765553_m1
CDH1	Hs01023894_m1
CDKN1A	Hs00355782_m1
COL3A1	Hs00943809_m1
DNMT3B	Hs00171876_m1
DNMT3L	Hs01081364_m1
EED	Hs00537777_m1
ETV5	Hs00231790_m1
FBXO15	Hs00380856_m1
FOXD1	Hs00270117_s1
GAPDH	Hs99999905_m1
GREM1	Hs01879841_s1
HDAC2	Hs00231032_m1
HESX1	Hs00172696_m1
JARID2	Hs01004460_m1
KAT7 (MYST2)	Hs01561260_m1
KLF4	Custom
LATS2	Hs00324396_m1
LEFTY1	Hs00764128_s1
LEFTY2	Hs00745761_s1
LIN28A	Hs00702808_s1
LOX	Hs00942480_m1
LUM	Hs00158940_m1
MYC	Hs01570247_m1
NACC1	Hs00369413_m1
NANOG	Hs02387400_g1
NR0B1(DAX1)	Hs03043658_m1
OTX2	Hs00222238_m1
PHC1	Hs01051497_m1
POU5F1	Custom
REST	Hs00958503_m1
RIF1	Hs00871714_m1
RNF2	Hs00200541_m1
SALL1	Hs00231307_m1
SALL4	Hs00360675_m1
SET	Hs00853870_g1
SMARCC2	Hs00161961_m1
SNAI2	Hs00950344_m1
SOX2	Hs01053049_s1
SP1	Hs00916521_m1
STAT3	Hs01047580_m1
TCF3	Hs01012685_m1
TDGF1	Hs02339499_g1
TGFBR2	Hs00234253_m1
TRIM28	Hs00232212_m1
ZFP42 (REX1)	Hs00399279_m1
ZIC3	Hs00185665_m1
ZNF281	Hs00273550_s1

Supplemental Table 2 List of 48 Taqman Assays Used for Single-Cell qRT-PCR

6.15.3 Supplemental Table 3

Gene	Baseline	Scale	Mean	Stdev	AICC
CBX7	0.052632	-0.10051	0.991556	0.016493	-7011.49
CCND1	0.947368	0.100026	0.90085	0.112272	-6049.31
CDH1	0	0.948792	0.515163	0.147273	-6367.52
CDKN1A	1	-0.61862	0.882695	0.101018	-6707.1
COL3A1	1	-0.38748	0.43193	0.098134	-6001.72
DNMT3B	0	0.935446	0.607084	0.180596	-5600.2
DNMT3L	0	0.100011	0.818973	0.026747	-8780.74
EED	0.210526	0.460534	0.15776	0.112932	-6484.99
ETV5	0.263158	0.100248	0.030285	0.009632	-4278.97
FBXO15	0.105263	-0.1	0.622454	0.180596	-6545.29
FOXD1	0.578947	0.133341	0.241371	0.069019	-4202.41
GREM1	0.263158	-0.23747	0.368077	0.054695	-6266.1
HDAC2	0.894737	0.105578	0.163729	0.144948	-9691.05
HESX1	0	0.254509	0.895169	0.063659	-7838.06
JARID2	0.631579	0.359664	0.188672	0.178532	-7099.61
KAT7	0.947368	0.100034	0.690427	0.180596	-6692.84
KIf4	0.631579	-0.10577	0.109557	0.055471	-5956.13
LATS2	0.578947	0.141871	0.028376	0.012383	-4351.91
LEFTY1	0.578947	-0.10029	0.138609	0.012123	-3633.38
LEFTY2	0.052632	0.814088	0.292862	0.180596	-6309.38
LIN28A	0	0.8684	0.284766	0.180596	-4518.9
LOX	0.947368	-0.39521	0.169431	0.11828	-4560.11
LUM	0.947368	-0.80084	0.32286	0.180596	-6203.03
MYC	0	0.163619	0.769866	0.171468	-7734.03
NACC1	0.947368	0.100054	0.900837	0.113807	-6061.31
NANOG	0.157895	0.837598	0.195052	0.139553	-7441.85
NROB1	0	-0.10058	0.991556	0.012586	-8602.77
Otx2	0.052632	0.766356	0.861305	0.110117	-7161.05
PHC1	0.736842	0.198745	0.602832	0.013183	-3898.86
POU5F1	0.157895	0.805615	0.370688	0.180596	-6094.48
REST	0	3.65E-21	0.991556	0.009632	-99460.3
RIF1	0.578947	0.362344	0.795243	0.059163	-4795.98
RNF2	0.736842	0.243195	0.231981	0.180596	-7166.14
SALL1	0.894737	-0.11851	0.087099	0.025809	-6033.18
SALL4	0	0.990657	0.522721	0.180596	-6078.84
SET	1	0.100226	0.991556	0.012596	-8068.11
SMARCC2	1	-0.10006	0.991556	0.012787	-7911.87
SNAI2	0.894737	-0.80742	0.315916	0.180596	-6311.56
SOX2	0.263158	0.436022	0.599727	0.180596	-6027.71
SP1	0.894737	0.101535	0.354385	0.180596	-8395.92
STAT3	0.947368	-0.10426	0.760637	0.06302	-7621.89
TCF3	0.894737	-0.10068	0.028747	0.009632	-4920.81
TDGF1	0.421053	0.445005	0.355582	0.180596	-5938.2
TGFBR2	0.947368	-0.70933	0.55147	0.065972	-4894.66

TRIM28	0.947368	0.052569	0.194608	0.180596	-10737.5
ZFP42	0	1	0.556834	0.164067	-7508.61
ZIC3	0	0.936096	0.754576	0.180596	-6585.11
ZNF281	0.736842	0.225216	0.18891	0.108623	-7123.8

Supplemental Table 3: Parameters for Single Gaussian Distribution Model.

6.15.4 Supplemental Table 4

Gene	Assay ID	Classification
TET1	Hs00286756_m1	Chromatin
SET	Hs00853870_g1	Chromatin
CDKN1A	Hs00355782_m1	Chromatin
DNMT3L	Hs01081364_m1	Chromatin
DNMT3B	Hs00171876_m1	Chromatin
USP16	Hs0017079_m1	Chromatin
PRDM14	Hs01119056_m1	Chromatin
TET2	Hs00325999_m1	Chromatin
PHC1	Hs01051497_m1	Chromatin
CBX7	Hs00545603_m1	Chromatin
JMJD3	Hs00389738_m1	Chromatin
EZH2	Hs01016789_m1	Chromatin
JARID1A	Hs00231908_m1	Chromatin
SETD1A	Hs00322315_m1	Chromatin
JMJD2B	Hs00943636_m1	Chromatin
MYST2	Hs01561260_m1	Chromatin
MBD3	Hs00922219_m1	Chromatin
UTF1	Hs00864535_s1	Chromatin
RNF2	Hs00200541_m1	Chromatin
EED	Hs00537777_m1	Chromatin
JARID2	Hs01004460_m1	Chromatin
HDAC2	Hs00231032_m1	Chromatin
COL3A1	Hs00943809_m1	Fibroblast
TDGF1	Hs02339499_g1	Fibroblast
SNAI2	Hs00950344_m1	Fibroblast
CDH2	Hs00983062_m1	Fibroblast
LATS2	Hs00324396_m1	Fibroblast
SMARCC2	Hs00161961_m1	Fibroblast
LOX	Hs00942480_m1	Fibroblast
LUM	Hs00158940_m1	Fibroblast
GREM1	Hs00171951_m1	Fibroblast
Eomes	Hs01015629_m1	Lineage Marker
FOXD1	Hs00270117_s1	Lineage Marker
HNF4A	Hs00604435_m1	Lineage Marker
T	Hs00610080_m1	Lineage Marker
MyoD1	Hs02330075_g1	Lineage Marker
Sox1	Hs01057642_s1	Lineage Marker
SOX13	Hs00232193_m1	Lineage Marker
TCL1A	Hs00951350_m1	Lineage Marker
MASP1	Hs00373559_m1	MSC
FOSB	Hs00171851_m1	MSC
TGFBR2	Hs00234253_m1	Pluripotency
FGF2	Hs00266645_m1	Pluripotency
NODAL	Hs00415443_m1	Pluripotency
Wnt3	Hs00229135_m1	Pluripotency
SMAD4	Hs00929647_m1	Pluripotency
DVL2	Hs00182901_m1	Pluripotency
KDR	Hs00911700_m1	Pluripotency
CD44	Hs01075861_m1	Pluripotency
DUSP10	Hs00200527_m1	Pluripotency
LCK	Hs00178427_m1	Pluripotency
CDH1	Hs01023894_m1	Pluripotency
Gata6	Hs00232018_m1	Pluripotency
SEMA6A	Hs00221174_m1	Pluripotency
APOE	Hs00171168_m1	Pluripotency
GDF3	Hs00220998_m1	Pluripotency
GABRB3	Hs00241459_m1	Pluripotency
ESRRB	Hs01584024_m1	Pluripotency
KLF5	Hs00156145_m1	Pluripotency
DPPA4	Hs00216968_m1	Pluripotency
DPPA2	Hs00414515_m1	Pluripotency
GAL	Hs00544355_m1	Pluripotency
LRRN1	Hs00979743_m1	Pluripotency
NANOG	Hs02387400_g1	Pluripotency
SOX2	Hs01053049_s1	Pluripotency
MYC	Hs01570247_m1	Pluripotency
KLF4	CUSTOM	Pluripotency
LEFTY2	Hs00745761_s1	Pluripotency
SALL1	Hs00231307_m1	Pluripotency
SALL4	Hs00360675_m1	Pluripotency
STAT3	Hs01047580_m1	Pluripotency
ZIC3	Hs00185665_m1	Pluripotency
ZNF281	Hs00273550_s1	Pluripotency
ZFP42	Hs00399279_m1	Pluripotency
NACC1	Hs00369413_m1	Pluripotency
RIF1	Hs00871714_m1	Pluripotency
NR0B1	Hs03043658_m1	Pluripotency
REST	Hs00958503_m1	Pluripotency
OTX2	Hs00222238_m1	Pluripotency
LIN28	Hs00702808_s1	Pluripotency
SP1	Hs00916521_m1	Pluripotency
TRIM28	Hs00232212_m1	Pluripotency
FBXO15	Hs00380856_m1	Pluripotency
ETV5	Hs00231790_m1	Pluripotency
TCF3	Hs01012685_m1	Pluripotency
HESX1	Hs00172696_m1	Pluripotency
CCND1	Hs00765553_m1	Pluripotency
POU5F1	Custom	Pluripotency
LEFTY1	Hs00764128_s1	Pluripotency
CD48	Hs00914738_m1	T-cell
CD4	Hs01058407_m1	T-cell
CD3D	Hs00174158_m1	T-cell
PTPN6	Hs00169359_m1	T-cell
CCR7	Hs01013469_m1	T-cell
STEMCCA-P2A	Custom	Transgene
GAPDH	Control	

Supplemental Table 4: List of 96 Taqman Assays used for single cell qRT-PCR

6.15.5 Supplemental Table 5

		SSEA4+						Tra1-60+		
		Day 4	Day 7	Day 11	Day 14	Day 21	Day 7	Day 11	Day 14	Day 21
BJ Fibroblast	Monocistronic	24	20	16	20	16	24	24	24	24
	Polycistronic	24	24	24	24	24	24	24	24	24
MRC-5	Monocistronic	16	16	0	13	0	0	0	48	0
	Polycistronic	16	16	16	16	16	24	24	24	24

BJ	MRC-5	H1	H9
20	16	24	24

Supplemental Table 5: Number of single cells analyzed for each cell type, surface marker and day of collection.

6.15.6 Supplemental Table 6

Gene	Forward primer	Reverse Primers
POU5F1 (OCT4)	5'-CCTCACTTCACTGCACTGTA-3'	5'-CCTTGAGGTACCAGAGATCT-3'
SOX2	5'-CCCAGCAGACTTCACATGT-3'	5'-CCTTGAGGTACCAGAGATCT-3'
KLF4	5'-GATGAACTGACCAGGCACTA-3'	5'-CCTTGAGGTACCAGAGATCT-3'
C-MYC	5'-TGCCTCAAATTGGACTTTGG-3'	5'-CGCTCGAGGTTAACGAATT-3'

Supplemental Table 6: Primers used for SYBR Green detection of exogenous OSKM factors.

6.15.7 Supplemental Table 7

Mass	Metal	Target	Clone	Cell Location	Concentration Per 3M Cells
127	IdU	S-phase/Proliferation	N/A	Nucleus	50uM
141	Pr	EpCAM	9C4	Surface	1ul/ test
145	Nd	CD4	RPA-T4	Surface	1ul / test
148	Nd	TRA-1-60	TRA-1-60	Surface	0.25ug
149	Sm	SMAD4	253343	Nucleus	0.125ug
150	Nd	SOX2	O3O-678	Nucleus	1ul / test
151	Eu	SSEA-4	MC813	Surface	0.125ug
152	Sm	CD13	WM15	Surface	1ul / test
153	Eu	LCK	Lck-01	Cytoplasm	0.125ug
154	Sm	CD309	7D4-6	Surface	0.25ug
159	Tb	p21/CIP1/WAF1	12D1	Nucleus	0.25ug
160	Gd	PDGFRa	D13C6	Surface	1ul / test
161	Dy	Ki67	B56	Nucleus	1ul / test
162	Dy	KLF4	D1F2	Nucleus	1ul / test
163	Dy	LUM	EPR8898	Cytoplasm	0.125ug
164	Dy	DNMT3b	832121	Nucleus	0.25ug
165	Ho	POU5F1	40/Oct-3	Nucleus	1ul / test
166	Er	ZFP42	polyclonal	Nucleus	0.125ug
167	Er	CCR7	G043H7	Surface	1ul / test
168	Er	CD73	AD2	Surface	0.25ug
169	Tm	NANOG	N31-355	Nucleus	1ul / test
171	Yb	CD44	IM7	Surface	1 ul / test
172	Yb	CD200	OX-104	Surface	0.25ug
173	Yb	CD104	58XB4	Surface	0.25ug
174	Yb	CD49d	9F10	Surface	0.125ug
175	Lu	OTX2	246826	Nucleus	0.125
176	Yb	c-MYC	9E10	Nucleus	1ul / test

Supplemental Table 7: Mass chanel assignmeent, antibody clone and concentrations used for CyTOF analysis

6.16 Authored Publications

6.16.1 Single Cell Analysis Reveals the Stochastic Phase of Reprogramming to Pluripotency is an Ordered Probabilistic Process

Chung, K.-M., **Kolling, F. W.**, Gajdosik, M. D., Burger, S., Russell, A. C., Nelson, C. E. Single cell analysis reveals the stochastic phase of reprogramming to pluripotency is an ordered probabilistic process. *PloS One*, *9*(4), e95304. (2014)

Abstract: Despite years of research, the reprogramming of human somatic cells to pluripotency remains a slow, inefficient process, and a detailed mechanistic understanding of reprogramming remains elusive. Current models suggest reprogramming to pluripotency occurs in two-phases: a prolonged stochastic phase followed by a rapid deterministic phase. In this paradigm, the early stochastic phase is marked by the random and gradual expression of pluripotency genes and is thought to be a major rate-limiting step in the successful generation of induced Pluripotent Stem Cells (iPSCs). Recent evidence suggests that the epigenetic landscape of the somatic cell is gradually reset during a period known as the stochastic phase, but it is known neither how this occurs nor what rate-limiting steps control progress through the stochastic phase. A precise understanding of gene expression dynamics in the stochastic phase is required in order to answer these questions. Moreover, a precise model of this complex process will enable the measurement and mechanistic dissection of treatments that enhance the rate or efficiency of reprogramming to pluripotency. Here we use single-cell transcript profiling, FACS and mathematical modeling to show that the stochastic phase is an ordered probabilistic process with independent gene-specific dynamics. We also show that partially reprogrammed cells infected with OSKM follow two trajectories: a productive trajectory toward increasingly ESC-like expression profiles or an alternative trajectory leading away from both the fibroblast and ESC state. These two pathways are distinguished by the coordinated expression of a small group of chromatin modifiers in the productive trajectory, supporting the notion that chromatin remodeling is essential for successful reprogramming. These are the first results to show that the stochastic phase of reprogramming in human fibroblasts is an ordered,

probabilistic process with gene-specific dynamics and to provide a precise mathematical framework describing the dynamics of pluripotency gene expression during reprogramming by OSKM.

6.16.2 Development of intestinal organoids as tissue surrogates: Cell composition and the Epigenetic control of differentiation

Cao, L., Kuratnik, A., Xu, W., Gibson, J. D., **Kolling, F.**, Falcone, E. R., Ammar, M., Van Heyst, M.D., Wright, D.L., Nelson, C.E., Giardina, C. Development of intestinal organoids as tissue surrogates: Cell composition and the Epigenetic control of differentiation. *Molecular Carcinogenesis*, *202*(April 2013), 189–202. http://doi.org/10.1002/mc.22089 (2013)

Abstract: Intestinal organoids are multicellular crypt-like structures that can be derived from adult intestinal stem cells (ISCs), embryonic stem cells (ESCs) or induced pluripotent stem cells (IPSCs). Here we show that intestinal organoids generated from mouse ESCs were enriched in ISCs and early progenitors. Treatment of these organoids with a γ-secretase inhibitor increased Math1 and decreased Hes1 expression, indicating Notch signaling regulates ISC differentiation in these organoids. Lgr5 and Tert positive ISCs constituted approximately 10% and 20% of the organoids. As found in native tissue, Lgr5 and Tert expressing cells resolved into two discreet populations, which were stable over time. Intestinal organoids derived from cancer-prone Apc(Min/+) mice showed similar numbers of ISCs, but had reduced Math1 expression, indicating a suppressed secretory cell differentiation potential (as found in intestinal tissue). Apc(Min/+) organoids were used to screen epigenetically active compounds for those that increased Math1 expression and organoid differentiation (including HDAC inhibitors, Sirtuin (SIRT) modulators and methyltransferase inhibitors). Broad-spectrum HDAC inhibitors increased both Math1 and Muc2 expression, indicating an ability to promote the suppressed secretory cell differentiation pathway. Other epigenetic compounds had a diverse impact on cell differentiation, with a strong negative correlation between those that activated the secretory marker Muc2 and those that activated the absorptive cell marker Fabp2. These data show that ESC-derived intestinal organoids can be derived in large numbers, contain distinct ISC types and can be used to screen for agents that promote cell differentiation through different lineage pathways.

6.16.3 pH dependence of amylin fibrillization

Jha, S., Snell, J. M., Sheftic, S. R., Patil, S. M., Daniels, S. B., **Kolling, F. W.,** & Alexandrescu, A. T. pH dependence of amylin fibrillization. *Biochemistry*, *53*(2), 300–10. http://doi.org/10.1021/bi401164k (2014)

Abstract: In type 2 diabetics, the hormone amylin misfolds into amyloid plaques implicated in the destruction of the pancreatic β-cells that make insulin and amylin. The aggregative misfolding of amylin is pH-dependent, and exposure of the hormone to acidic and basic environments could be physiologically important. Amylin has two ionizable residues between pH 3 and 9: the α-amino group and His18. Our approach to measuring the pKa values for these sites has been to look at the pH dependence of fibrillization in amylin variants that have only one of the two groups. The α-amino group at the unstructured N-terminus of amylin has a pKa near 8.0, similar to the value in random coil models. By contrast, His18, which is involved in the intermolecular β-sheet structure of the fibrils, has a pKa that is lowered to 5.0 in the fibrils compared to the random coil value of 6.5. The lowered pKa of His18 is due to the hydrophobic environment of the residue, and electrostatic repulsion between positively charged His18 residues on neighboring amylin molecules in the fibril. His18 acts as an electrostatic switch inhibiting fibrillization in its charged state. The presence of a charged side chain at position 18 also affects fibril morphology and lowers amylin cytotoxicity toward a MIN6 mouse model of pancreatic β-cells. In addition to the two expected pKa values, we detected an apparent pKa of ~4.0 for the amylin-derived peptide NAc-SNNFGAILSS-NH2, which has no titratable groups. This pKa is due to the pH-induced ionization of the dye thioflavin T. By using alternative methods to follow fibrillization such as the dye Nile Red or turbidimetry, we were able to distinguish between the titration of the dye and groups on the peptide. Large differences in reaction kinetics were observed between the different methods at acidic pH, because of charges on the ThT dye, which hinder fibril formation much like the charges on the protein.

6.16.4 SCLD: a Stem Cell Lineage Database for the annotation of cell types and developmental lineages

Hemphill, E. E., Dharia, A. P., Lee, C., Jakuba, C. M., Gibson, J. D., **Kolling, F. W.**, & Nelson, C. E. SCLD: a stem cell lineage database for the annotation of cell types and developmental lineages. *Nucleic Acids Research*, gkq941. (2010)

Abstract: Stem cell biology has experienced explosive growth over the past decade as researchers attempt to generate therapeutically relevant cell types in the laboratory. Recapitulation of endogenous developmental trajectories is a dominant paradigm in the design of directed differentiation protocols, and attempts to guide stem cell differentiation are often based explicitly on knowledge of in vivo development. Therefore, when designing protocols, stem cell biologists rely heavily upon information including (i) cell type-specific gene expression profiles, (ii) anatomical and developmental relationships between cells and tissues and (iii) signals important for progression from progenitors to target cell types. Here, we present the Stem Cell Lineage Database (SCLD) (http://scld.mcb.uconn.edu) that aims to unify this information into a single resource where users can easily store and access information about cell type gene expression, cell lineage maps and stem cell differentiation protocols for both human and mouse stem cells and endogenous developmental lineages. By establishing the SCLD, we provide scientists with a centralized location to organize access and share data, dispute and resolve contentious relationships between cell types and within lineages, uncover discriminating cell type marker panels and design directed differentiation protocols.

6.16.5 Regulation of the Fanconi anemia pathway by a CUE ubiquitin-binding domain in the FANCD2 protein

Rego, M. A., **Kolling, F. W.**, Vuono, E. A., Mauro, M., & Howlett, N. G. Regulation of the Fanconi anemia pathway by a CUE ubiquitin-binding domain in the FANCD2 protein. *Blood*, *120*(10), 2109–2117. (2012)

Abstract: The Fanconi anemia (FA)-BRCA pathway is critical for the repair of DNA interstrand crosslinks (ICLs) and the maintenance of chromosome stability. A key step in FA-BRCA pathway activation is the covalent attachment of monoubiquitin to FANCD2 and FANCI. Monoubiquitinated FANCD2 and FANCI localize in chromatin-associated nuclear foci where they interact with several well-characterized DNA repair proteins. Importantly, very little is known about the structure, function, and regulation of FANCD2. Herein, we describe the identification and characterization of a CUE (coupling of ubiquitin conjugation to endoplasmic reticulum degradation) ubiquitin-binding domain (UBD) in FANCD2, and demonstrate that the CUE domain mediates noncovalent binding to ubiquitin in vitro. We show that although mutation of the CUE domain destabilizes FANCD2, the protein remains competent for DNA damage-inducible monoubiquitination and phosphorylation. Importantly, we demonstrate that the CUE domain is required for interaction with FANCI, retention of monoubiquitinated FANCD2, and FANCI in chromatin, and for efficient ICL repair. Our results suggest a model by which heterodimerization of monoubiquitinated FANCD2 and FANCI in chromatin is mediated in part through a noncovalent interaction between the FANCD2 CUE domain and monoubiquitin covalently attached to FANCI, and that this interaction shields monoubiquitinated FANCD2 from polyubiquitination and proteasomal degradation.

6.16.6 The p21Cip1/Waf1 cyclin-dependent kinase inhibitor is required for the activation of the FA-BRCA pathway

Rego, M. A., Mauro, M., Harney, J. A., Shen, M., **Kolling, F. W.**, & Howlett, N. G. Abstract LB-102: The p21Cip1/Waf1 cyclin-dependent kinase inhibitor is required for the activation of the FA-BRCA pathway. *Cancer Research*, *70*(8 Supplement), LB–102. (2010)

Abstract: Fanconi anemia (FA) is a rare disease characterized by congenital defects, progressive bone marrow failure and heightened cancer susceptibility. The FA proteins, BRCA1 and FANCD1/BRCA2 function cooperatively in the FA-BRCA pathway to repair damaged DNA. Activation of the FA-BRCA pathway occurs via the monoubiquitination of the FANCD2 and FANCI proteins, targeting these proteins to discrete nuclear foci where they function in DNA repair. The cellular regulation of FANCD2/I monoubiquitination, however, remains poorly understood. In this study, we have examined the roles of the p53 tumor suppressor protein, as well as its downstream target, the p21(Cip1/Waf1) cyclin-dependent kinase inhibitor, in the regulation of the activation of the FA-BRCA pathway. We demonstrate that, in contrast to p53, p21 has a major role in the regulation of the activation of the FA-BRCA pathway: p21 promotes S-phase and DNA damage-inducible FANCD2/I monoubiquitination and nuclear foci formation. Several lines of evidence establish that this effect is not a consequence of a defective G1-S checkpoint or altered cell-cycle progression in the absence of p21. Instead, we demonstrate that p21 is required for the transcriptional repression of the USP1 deubiquitinating enzyme upon exposure to DNA-damaging agents. In the absence of p21, persistent USP1 expression precludes the DNA damage-inducible accumulation of monoubiquitinated FANCD2 and FANCI. Consequently, p21(-/-) cells exhibit increased levels of mitomycin C-inducible complex chromosomal aberrations and elevated γH2AX nuclear foci formation. Our results demonstrate that p21 has a critical role in the regulation of the activation of the FA-BRCA pathway and suggest a broader role for p21 in the orchestration of DNA repair processes following exposure to DNA crosslinking agents.

6.16.7 Functional interaction between the Fanconi Anemia D2 protein and proliferating cell nuclear antigen (PCNA) via a conserved putative PCNA interaction motif

Howlett, N. G., Harney, J. A., Rego, M. A., **Kolling, F. W.**, & Glover, T. W. (2009). Functional interaction between the Fanconi Anemia D2 protein and proliferating cell nuclear antigen (PCNA) via a conserved putative PCNA interaction motif. *Journal of Biological Chemistry*, *284*(42), 28935–28942.

Abstract: Fanconi Anemia (FA) is a rare recessive disease characterized by congenital abnormalities, bone marrow failure, and cancer susceptibility. The FA proteins and the familial breast cancer susceptibility gene products, BRCA1 and FANCD1/BRCA2, function cooperatively in the FA-BRCA pathway to repair damaged DNA and to prevent cellular transformation. Activation of this pathway occurs via the mono-ubiquitination of the FANCD2 protein, targeting it to nuclear foci where it co-localizes with FANCD1/BRCA2, RAD51, and PCNA. The regulation of the mono-ubiquitination of FANCD2, as well as its function in DNA repair remain poorly understood. In this study, we have further characterized the interaction between the FANCD2 and PCNA proteins. We have identified a highly conserved, putative FANCD2 PCNA interaction motif (PIP-box), and demonstrate that mutation of this motif disrupts FANCD2-PCNA binding and precludes the mono-ubiquitination of FANCD2. Consequently, the FANCD2 PIP-box mutant protein fails to correct the mitomycin C hypersensitivity of FA-D2 patient cells. Our results suggest that PCNA may function as a molecular platform to facilitate the mono-ubiquitination of FANCD2 and activation of the FA-BRCA pathway.

6.16.8 The Fanconi anemia protein interaction network: Casting a wide net

Rego, M. A., **Kolling, F. W.**, & Howlett, N. G. The Fanconi anemia protein interaction network: Casting a wide net. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *668*(1), 27–41. (2009)

Abstract: It has long been hypothesized that a defect in the repair of damaged DNA is central to the etiology of Fanconi anemia (FA). Indeed, an increased sensitivity of FA patient-derived cells to the lethal effects of various forms of DNA damaging agents was described over three decades ago [A.J. Fornace, Jr., J.B. Little, R.R. Weichselbaum, DNA repair in a Fanconi's anemia fibroblast cell strain, Biochim. Biophys. Acta 561 (1979) 99-109; Y. Fujiwara, M. Tatsumi, Repair of mitomycin C damage to DNA in mammalian cells and its impairment in Fanconi's anemia cells, Biochem. Biophys. Res. Commun. 66 (1975) 592-598; A.J. Rainbow, M. Howes, Defective repair of ultraviolet- and gamma-ray-damaged DNA in Fanconi's anaemia, Int. J. Radiat. Biol. Relat. Stud. Phys. Chem. Med. 31 (1977) 191-195]. Furthermore, the cytological hallmark of FA, the DNA crosslink-induced radial chromosome formation, exemplifies an innate impairment in the repair of these particularly cytotoxic DNA lesions [A.D. Auerbach, Fanconi anemia diagnosis and the diepoxybutane (DEB) test, Exp. Hematol. 21 (1993) 731-733]. Precisely defining the collective role of the FA proteins in DNA repair, however, continues to be one of the most enigmatic and challenging questions in the FA field. The first six identified FA proteins (A, C, E, F, G, and D2) harbored no recognizable enzymatic features, precluding association with a specific metabolic process. Consequently, our knowledge of the role of the FA proteins in the DNA damage response has been gleaned primarily through biochemical association studies with non-FA proteins. Here, we provide a chronological discourse of the major FA protein interaction network discoveries, with particular emphasis on the DNA damage response, that have defined our current understanding of the molecular basis of FA.

6.17 Manuscript Under Review

6.17.1 Comparison of Reprogramming Methods by Single Cell Analysis Identifies Premature Viral Inactivation as a Barrier to Successful Reprogramming and Reveals a Common Reprogramming Trajectory Between Cell Types

Kolling, F. W., Chung, K.-M., Chen, K., Bar, H., Schifano, E. D., Harel, O., Mandiou, Ion I., Nelson, C. E. Comparison of Reprogramming Methods by Single Cell Analysis Identifies Premature Viral Inactivation as a Barrier to Successful Reprogramming and Reveals a Common Reprogramming Trajectory Between Cell Types. *PloS One*. (2015).

Abstract: Reprogramming terminally differentiated cells to a pluripotent state by exogenous expression of the Yamanaka factors OCT4, SOX2, KLF4 and c-MYC (OSKM) has the potential to revolutionize many aspects of modern medicine. However, despite years of research this process remains highly inefficient and produces considerable cellular heterogeneity, problems that must be overcome before this technique can be used clinically. In the years following the first reports of OSKM-mediated reprogramming, several methods have been developed to reprogram cells to pluripotency in an effort to increase the efficiency and quality of iPSC generation, including using different methods of delivering the OSKM factors as well as the use of novel reprogramming factors. While many of these approaches have increased the efficiency and/or rate of reprogramming, at present it is unclear how these results manifest at the molecular level and how these various methods impact the quality and differentiation potential of resulting iPSCs. In this study, we apply single cell transcript analysis to compare the transcriptional dynamics underlying the acquisition of pluripotency in monocistronic and polycistronic OSKM systems. These two delivery methods were tested in both MRC-5 and BJ fibroblasts. We demonstrate that polycistronic viral delivery produces significantly higher reprogramming efficiencies compared with monocistronic delivery and that this effect is due in part to premature inactivation of the individual O,S,K or M vectors in the monocistronic method. In addition, we show that the activation of key pluripotency loci such as NANOG, OCT4, LIN28 and DNMT3B occurs earlier in the polycistronic condition and that these cells progress more uniformly towards pluripotency. Finally, we compare polycistronic reprogramming between MRC-5 and BJ fibroblast cells and reveal that while the order of gene activation

is similar between cell types; MRC-5 and BJ cells take divergent paths upon factor induction, followed by convergence later in the reprogramming process.

References

1. Gadue, P., Huber, T. L., Nostro, M. C., Kattman, S. & Keller, G. M. Germ layer induction from embryonic stem cells. *Exp. Hematol.* **33,** 955–64 (2005).

2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–76 (2006).

3. Buganim, Y., Faddah, D. A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* **14,** 427–39 (2013).

4. Plath, K. & Lowry, W. E. Progress in understanding reprogramming to the induced pluripotent state. *Nat. Rev. Genet.* **12,** 253–65 (2011).

5. Hochedlinger, K. & Plath, K. Epigenetic reprogramming and induced pluripotency. *Development* **136,** 509–23 (2009).

6. Jung, L. *et al.* ONSL and OSKM cocktails act synergistically in reprogramming human somatic cells into induced pluripotent stem cells. *Mol. Hum. Reprod.* **20,** 538–49 (2014).

7. Federation, A. J., Bradner, J. E. & Meissner, A. The use of small molecules in somatic-cell reprogramming. *Trends Cell Biol.* **24,** 179–87 (2014).

8. Polo, J. M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151,** 1617–32 (2012).

9. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454,** 49–55 (2008).

10. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–60 (2007).

11. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462,** 595–601 (2009).

12. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotechnol.* **25,** 1177–81 (2007).

13. Chin, M. H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5,** 111–23 (2009).

14. Hu, B.-Y. *et al.* Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4335–40 (2010).

15. Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature* **150,** 563–565 (1942).

16. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462,** 587–94 (2009).

17. Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463,** 474–84 (2010).

18. Hemberger, M., Dean, W. & Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat. Rev. Mol. Cell Biol.* **10,** 526–37 (2009).

19. Vierbuchen, T. & Wernig, M. Molecular roadblocks for cellular reprogramming. *Mol. Cell* **47,** 827–38 (2012).

20. Huang, S. Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366,** 2247–59 (2011).

21. Gurdon, J. B. Genetic reprogramming following nuclear transplantation in Amphibia. *Semin. Cell Dev. Biol.* **10,** 239–43 (1999).

22. McGrath, J. & Solter, D. Nuclear transplantation in the mouse embryo by microsurgery and cell fusion. *Science* **220,** 1300–2 (1983).

23. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385,** 810–3 (1997).

24. Prather, R. S. *et al.* Nuclear transplantation in the bovine embryo: assessment of donor nuclei and recipient oocyte. *Biol. Reprod.* **37,** 859–66 (1987).

25. Meng, L., Ely, J. J., Stouffer, R. L. & Wolf, D. P. Rhesus monkeys produced by nuclear transfer. *Biol. Reprod.* **57,** 454–9 (1997).

26. Baguisi, A. *et al.* Production of goats by somatic cell nuclear transfer. *Nat. Biotechnol.* **17,** 456–61 (1999).

27. Tachibana, M. *et al.* Human embryonic stem cells derived by somatic cell nuclear transfer. *Cell* **153,** 1228–38 (2013).

28. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–72 (2007).

29. Aoi, T. *et al.* Generation of pluripotent stem cells from adult mouse liver and stomach cells. *Science* **321,** 699–702 (2008).

30. Winkler, T. *et al.* No evidence for clonal selection due to lentiviral integration sites in human induced pluripotent stem cells. *Stem Cells* **28,** 687–94 (2010).

31. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8,** 409–12 (2011).

32. Somers, A. *et al.* Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. *Stem Cells* **28,** 1728–40 (2010).

33. Sommer, C. A. *et al.* Excision of reprogramming transgenes improves the differentiation potential of iPS cells generated with a single excisable vector. *Stem Cells* **28,** 64–74 (2010).

34. Woltjen, K. *et al.* piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458,** 766–70 (2009).

35. Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324,** 797–801 (2009).

36. Jia, F. *et al.* A nonviral minicircle vector for deriving human iPS cells. *Nat. Methods* **7,** 197–9 (2010).

37. Narsinh, K. H. *et al.* Generation of adult human induced pluripotent stem cells using nonviral minicircle DNA vectors. *Nat. Protoc.* **6,** 78–88 (2011).

38. Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7,** 618–30 (2010).

39. Judson, R. L., Babiarz, J. E., Venere, M. & Blelloch, R. Embryonic stem cell-specific microRNAs promote induced pluripotency. *Nat. Biotechnol.* **27,** 459–61 (2009).

40. Li, Z., Yang, C.-S., Nakashima, K. & Rana, T. M. Small RNA-mediated regulation of iPS cell generation. *EMBO J.* **30,** 823–34 (2011).

41. Subramanyam, D. *et al.* Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. *Nat. Biotechnol.* **29,** 443–8 (2011).

42. Anokye-Danso, F. Reprogramming somatic cells into pluripotent stem cells using miRNAs. *Methods Mol. Biol.* **1150,** 273–81 (2014).

43. Nishino, K. *et al.* DNA methylation dynamics in human induced pluripotent stem cells over time. *PLoS Genet.* **7,** e1002085 (2011).

44. Taapken, S. M. *et al.* Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. *Nat. Biotechnol.* **29,** 313–4 (2011).

45. Sperger, J. M. *et al.* Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 13350–5 (2003).

46. Ghosh, Z. *et al.* Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PLoS One* **5,** e8975 (2010).

47. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **Chapter 19,** Unit 19.10.1–21 (2010).

48. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–70 (2008).

49. Chan, E. M. *et al.* Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nat. Biotechnol.* **27,** 1033–7 (2009).

50. Maherali, N. *et al.* Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1,** 55–70 (2007).

51. Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2,** 151–9 (2008).

52. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2,** 230–40 (2008).

53. Tsubouchi, T. & Fisher, A. G. Reprogramming and the pluripotent stem cell cycle. *Curr. Top. Dev. Biol.* **104,** 223–41 (2013).

54. Baum, B., Settleman, J. & Quinlan, M. P. Transitions between epithelial and mesenchymal states in development and disease. *Semin. Cell Dev. Biol.* **19,** 294–308 (2008).

55. Samavarchi-Tehrani, P. *et al.* Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell Stem Cell* **7,** 64–77 (2010).

56. Buganim, Y. *et al.* Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150,** 1209–22 (2012).

57. Tanabe, K., Nakamura, M., Narita, M., Takahashi, K. & Yamanaka, S. Maturation, not initiation, is the major roadblock during reprogramming toward pluripotency from human fibroblasts. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12172–9 (2013).

58. Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G. P. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell* **16,** 323–37 (2015).

59. Lujan, E. *et al.* Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* **521,** 352–356 (2015).

60. Ang, Y.-S. *et al.* Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145,** 183–97 (2011).

61. Huangfu, D. *et al.* Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat. Biotechnol.* **26,** 795–7 (2008).

62. Ichida, J. K. *et al.* A small-molecule inhibitor of tgf-Beta signaling replaces sox2 in reprogramming by inducing nanog. *Cell Stem Cell* **5,** 491–503 (2009).

63. Zhao, W. *et al.* Jmjd3 inhibits reprogramming by upregulating expression of INK4a/Arf and targeting PHF20 for ubiquitination. *Cell* **152,** 1037–50 (2013).

64. Yang, P. *et al.* RCOR2 is a subunit of the LSD1 complex that regulates ESC property and substitutes for SOX2 in reprogramming somatic cells to pluripotency. *Stem Cells* **29,** 791–801 (2011).

65. Onder, T. T. *et al.* Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **483,** 598–602 (2012).

66. Koche, R. P. *et al.* Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell* **8,** 96–105 (2011).

67. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell* (2012). doi:10.1016/j.cell.2012.09.045

68. Chen, J. *et al.* H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nat. Genet.* **45,** 34–42 (2013).

69. Mattout, A., Biran, A. & Meshorer, E. Global epigenetic changes during somatic cell reprogramming to iPS cells. *J. Mol. Cell Biol.* **3,** 341–50 (2011).

70. Deng, W. AID in reprogramming: quick and efficient: identification of a key enzyme called AID, and its activity in DNA demethylation, may help to overcome a pivotal epigenetic barrier in reprogramming somatic cells toward pluripotency. *Bioessays* **32,** 385–7 (2010).

71. Li, W. *et al.* Identification of Oct4-activating compounds that enhance reprogramming efficiency. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 20853–8 (2012).

72. Orkin, S. H. & Hochedlinger, K. Chromatin connections to pluripotency and cellular reprogramming. *Cell* **145,** 835–50 (2011).

73. Koche, R. P. *et al.* Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell* **8,** 96–105 (2011).

74. Sridharan, R. *et al.* Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136,** 364–77 (2009).

75. Marión, R. M. *et al.* A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. *Nature* **460,** 1149–53 (2009).

76. Rasmussen, M. A. *et al.* Transient p53 suppression increases reprogramming of human fibroblasts without affecting apoptosis and DNA damage. *Stem cell reports* **3,** 404–13 (2014).

77. Anokye-Danso, F. *et al.* Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell* **8,** 376–88 (2011).

78. Mallanna, S. K. & Rizzino, A. Emerging roles of microRNAs in the control of embryonic stem cells and the generation of induced pluripotent stem cells. *Dev. Biol.* **344,** 16–25 (2010).

79. Onder, T. T. & Daley, G. Q. microRNAs become macro players in somatic cell reprogramming. *Genome Med.* **3,** 40 (2011).

80. Liao, B. *et al.* MicroRNA cluster 302-367 enhances somatic cell reprogramming by accelerating a mesenchymal-to-epithelial transition. *J. Biol. Chem.* **286,** 17359–64 (2011).

81. Tiscornia, G. & Izpisúa Belmonte, J. C. MicroRNAs in embryonic stem cell function and fate. *Genes Dev.* **24,** 2732–41 (2010).

82. Melton, C., Judson, R. L. & Blelloch, R. Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature* **463,** 621–6 (2010).

83. Kim, D. H. *et al.* Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell* **16,** 88–101 (2015).

84. Benevento, M. *et al.* Proteome adaptation in cell reprogramming proceeds via distinct transcriptional networks. *Nat. Commun.* **5,** 5613 (2014).

85. Hansson, J. *et al.* Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. *Cell Rep.* **2,** 1579–92 (2012).

86. Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81,** 6813–22 (2009).

87. Ghule, P. N. *et al.* Reprogramming the pluripotent cell cycle: restoration of an abbreviated G1 phase in human induced pluripotent stem (iPS) cells. *J. Cell. Physiol.* **226,** 1149–56 (2011).

88. Becker, K. A., Stein, J. L., Lian, J. B., van Wijnen, A. J. & Stein, G. S. Establishment of histone gene regulation and cell cycle checkpoint control in human embryonic stem cells. *J. Cell. Physiol.* **210,** 517–26 (2007).

89. Becker, K. A. *et al.* Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. *J. Cell. Physiol.* **209,** 883–93 (2006).

90. Nakagawa, M. *et al.* Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat. Biotechnol.* **26,** 101–6 (2008).

91. Rais, Y. *et al.* Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* **advance on,** (2013).

92. Nelson, W. J. & Nusse, R. Convergence of Wnt, beta-catenin, and cadherin pathways. *Science* **303,** 1483–7 (2004).

93. Shirakihara, T., Saitoh, M. & Miyazono, K. Differential regulation of epithelial and mesenchymal markers by deltaEF1 proteins in epithelial mesenchymal transition induced by TGF-beta. *Mol. Biol. Cell* **18,** 3533–44 (2007).

94. Ng, V. Y., Ang, S. N., Chan, J. X. & Choo, A. B. H. Characterization of epithelial cell adhesion molecule as a surface marker on undifferentiated human embryonic stem cells. *Stem Cells* **28,** 29–35 (2010).

95. Chen, H.-F. *et al.* Surface Marker Epithelial Cell Adhesion Molecule and E-cadherin Facilitate the Identification and Selection of Induced Pluripotent Stem Cells. *Stem Cell Rev. Reports* **7,** 722–735 (2011).

96. Li, R. *et al.* A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* **7,** 51–63 (2010).

97. Hotta, A. & Ellis, J. Retroviral vector silencing during iPS cell induction: an epigenetic beacon that signals distinct pluripotent states. *J. Cell. Biochem.* **105,** 940–8 (2008).

98. Tonge, P. D. *et al.* Divergent reprogramming routes lead to alternative stem-cell states. *Nature* **516,** 192–197 (2014).

99. Hussein, S. M. I. *et al.* Genome-wide characterization of the routes to pluripotency. *Nature* **516,** 198–206 (2014).

100. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448,** 313–7 (2007).

101. Carey, B. W. *et al.* Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell Stem Cell* **9,** 588–98 (2011).

102. Hou, P. *et al.* Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. *Science* science.1239278– (2013). doi:10.1126/science.1239278

103. Chew, J. *et al.* Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4 / Sox2 Complex in Embryonic Stem Cells. **25,** 6031–6046 (2005).

104. Radzisheuskaya, A. *et al.* A defined Oct4 level governs cell state transitions of pluripotency entry and differentiation into all embryonic lineages. *Nat. Cell Biol.* **15,** 579–90 (2013).

105. Behbahaninia, M., Ramey, W. L., Sindhwani, M. K. & Kalani, M. Y. S. Differential expression of pluripotency factors Sox2 and Oct4 regulate neuronal and mesenchymal lineages. *Neurosurgery* **69,** N19 (2011).

106. Thomson, M. *et al.* Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145,** 875–89 (2011).

107. Yao, S. *et al.* Retrovirus silencing, variegation, extinction, and memory are controlled by a dynamic interplay of multiple epigenetic modifications. *Mol. Ther.* **10,** 27–36 (2004).

108. Wolf, D. & Goff, S. P. TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. *Cell* **131,** 46–57 (2007).

109. He, J., Yang, Q. & Chang, L.-J. Dynamic DNA methylation and histone modifications contribute to lentiviral transgene silencing in murine embryonic carcinoma cells. *J. Virol.* **79,** 13497–508 (2005).

110. Buganim, Y., Faddah, D. a & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* **14,** 427–39 (2013).

111. Li, W. & Ding, S. Small molecules that modulate embryonic stem cell fate and somatic cell reprogramming. *Trends Pharmacol. Sci.* **31,** 36–45 (2010).

112. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. *Nat. Methods* **7,** 53–5 (2010).

113. Carey, B. W., Markoulaki, S., Beard, C., Hanna, J. & Jaenisch, R. Single-gene transgenic mouse strains for reprogramming adult somatic cells. *Nat. Methods* **7,** 56–9 (2010).

114. Wernig, M. *et al.* A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nat. Biotechnol.* **26,** 916–24 (2008).

115. Göke, J. *et al.* Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Comput. Biol.* **7,** e1002304 (2011).

116. Yamanaka, S. Induced pluripotent stem cells: past, present, and future. *Cell Stem Cell* **10,** 678–84 (2012).

117. Chin, M. H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5,** 111–23 (2009).

118. Marchetto, M. C. N. *et al.* Transcriptional signature and memory retention of human-induced pluripotent stem cells. *PLoS One* **4,** e7076 (2009).

119. Narsinh, K. H. *et al.* Brief report Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *Gene Expr.* **121,** 1217–1221 (2011).

120. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7,** 521–31 (2010).

121. Laurent, L. C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8,** 106–18 (2011).

122. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132,** 1049–61 (2008).

123. Smith, Z. D., Nachman, I., Regev, A. & Meissner, A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nat. Biotechnol.* **28,** 521–6 (2010).

124. Golipour, A. *et al.* A Late Transition in Somatic Cell Reprogramming Requires Regulators Distinct from the Pluripotency Network. *Cell Stem Cell* **11,** 769–782 (2012).

125. Rideout, W. M., Eggan, K. & Jaenisch, R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* **293,** 1093–8 (2001).

126. Gibson, J. D. *et al.* Single-cell transcript analysis of human embryonic stem cells. *Integr. Biol. (Camb).* **1,** 540–51 (2009).

127. Lowry, W. E. *et al.* Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 2883–8 (2008).

128. Rinn, J. L., Bondre, C., Gladstone, H. B., Brown, P. O. & Chang, H. Y. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS Genet.* **2,** e119 (2006).

129. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–20 (2007).

130. Boyer, L. a *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–56 (2005).

131. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C. & Melton, D. a. 'Stemness': transcriptional profiling of embryonic and adult stem cells. *Science* **298,** 597–600 (2002).

132. Young, R. a. Control of the embryonic stem cell state. *Cell* **144,** 940–54 (2011).

133. Egli, D., Birkhoff, G. & Eggan, K. Mediators of reprogramming: transcription factors and transitions through mitosis. *Nat. Rev. Mol. Cell Biol.* **9,** 505–16 (2008).

134. Yamanaka, S. Elite and stochastic models for induced pluripotent stem cell generation. *Nature* **460,** 49–52 (2009).

135. Singh, A. M. & Dalton, S. The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. *Cell Stem Cell* **5,** 141–9 (2009).

136. Rahl, P. B. *et al.* c-Myc Regulates Transcriptional Pause Release. *Cell* **141,** 432–445 (2010).

137. Park, I.-H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451,** 141–6 (2008).

138. Singh, A. M. *et al.* Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. *Stem cell reports* **1,** 532–44 (2013).

139. Neganova, I., Zhang, X., Atkinson, S. & Lako, M. Expression and functional analysis of G1 to S regulatory components reveals an important role for CDK2 in cell cycle regulation in human embryonic stem cells. *Oncogene* **28,** 20–30 (2009).

140. Flöttmann, M., Scharp, T. & Klipp, E. A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front. Physiol.* **3,** 216 (2012).

141. Hanna, J., Carey, B. W. & Jaenisch, R. Reprogramming of somatic cell identity. *Cold Spring Harb. Symp. Quant. Biol.* **73,** 147–55 (2008).

142. Shi, Y. *et al.* Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell* **3,** 568–74 (2008).

143. Li, Y. *et al.* Generation of iPSCs from mouse fibroblasts with a single gene, Oct4, and small molecules. *Cell Res.* **21,** 196–204 (2011).

144. Shi, Y. *et al.* A combined chemical and genetic approach for the generation of induced pluripotent stem cells. *Cell Stem Cell* **2,** 525–8 (2008).

145. Zhao, Y. *et al.* Two supporting factors greatly improve the efficiency of human iPSC generation. *Cell Stem Cell* **3,** 475–9 (2008).

146. Carey, B. W. *et al.* Reprogramming of murine and human somatic cells using a single polycistronic vector. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 157–62 (2009).

147. Pan, G. & Pei, D. Order from Chaos: Single Cell Reprogramming in Two Phases. *Cell Stem Cell* **11,** 445–447 (2012).

148. Golipour, A. *et al.* A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. *Cell Stem Cell* **11,** 769–82 (2012).

149. Chung, K.-M. *et al.* Single cell analysis reveals the stochastic phase of reprogramming to pluripotency is an ordered probabilistic process. *PLoS One* **9,** e95304 (2014).

150. Aasen, T. *et al.* Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat. Biotechnol.* **26,** 1276–84 (2008).

151. Papapetrou, E. P. *et al.* Stoichiometric and temporal requirements of Oct4, Sox2, Klf4, and c-Myc expression for efficient human iPSC induction and differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12759–64 (2009).

152. Miyoshi, N. *et al.* Reprogramming of mouse and human cells to pluripotency using mature microRNAs. *Cell Stem Cell* **8,** 633–8 (2011).

153. Doege, C. A. *et al.* Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. *Nature* **488,** 652–5 (2012).

154. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).

155. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8,** 532–8 (2006).

156. Segre, J. A., Bauer, C. & Fuchs, E. Klf4 is a transcription factor required for establishing the barrier function of the skin. *Nat. Genet.* **22,** 356–60 (1999).

157. Wei, Z. *et al.* Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. *Stem Cells* **27,** 2969–78 (2009).

158. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143,** 508–25 (2010).

159. Rao, M. S. & Malik, N. Assessing iPSC reprogramming methods for their suitability in translational medicine. *J. Cell. Biochem.* **113,** 3061–8 (2012).

160. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–72 (2007).

161. Tran, K. A. *et al.* Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. *Nat. Commun.* **6,** 6188 (2015).

162. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotechnol.* **25,** 1177–81 (2007).

163. Amir, E. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31,** 545–52 (2013).

164. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157,** 714–25 (2014).

165. Henzler, C. M. *et al.* Staged miRNA re-regulation patterns during reprogramming. *Genome Biol.* **14,** R149 (2013).

166. Ribeiro, A. O., Schoof, C. R. G., Izzotti, A., Pereira, L. V & Vasques, L. R. MicroRNAs: modulators of cell identity, and their applications in tissue engineering. *MicroRNA (Shāriqah, United Arab Emirates)* **3,** 45–53 (2014).

167. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–33 (2008).

168. Kuo, C.-H., Deng, J. H., Deng, Q. & Ying, S.-Y. A novel role of miR-302/367 in reprogramming. *Biochem. Biophys. Res. Commun.* **417,** 11–6 (2012).

169. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20,** 1122–30 (2013).

170. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–6 (2014).

171. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

172. Dowell, K. G., Simons, A. K., Wang, Z. Z., Yun, K. & Hibbs, M. A. Cell-type-specific predictive network yields novel insights into mouse embryonic stem cell self-renewal and cell fate. *PLoS One* **8,** e56810 (2013).

173. Yang, K. E. *et al.* Differential expression of extracellular matrix proteins in senescent and young human fibroblasts: a comparative proteomics and microarray study. *Mol. Cells* **32,** 99–106 (2011).

174. SAS Institute, Cary, N. JMP, Version 10.