Transcriptional and Chromatin Accessibility Dynamics in the Reprogramming of Somatic Cells to Pluripotency

By:

Stefan Joseph Pietrzak

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Cellular and Molecular Biology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2024

Date of final oral examination: 1/18/2024

The dissertation is approved by the following members of the Final Oral Committee:
Rupa Sridharan, Associate Professor, Department of Cell and Regenerative Biology
Barak Blum, Associate Professor, Department of Cell and Regenerative Biology
Emery Bresnick, Professor, Department of Cell and Regenerative Biology
Phillip Newmark, Professor, Department of Integrative Biology
Sushmita Roy, Professor, Department of Biostatistics and Medical Informatics

Acknowledgments

There are several people to whom I owe a debt of gratitude for their support and dedication throughout my time in graduate school and that I need to acknowledge. First and foremost, I would like to thank my thesis advisor, Rupa Sridharan, for giving me the tremendous opportunity to work in her lab for my graduate school studies. Her mentorship has been invaluable in my success here and has helped to ensure that my work was always on the right track. I have definitely become a better scientist after getting to work with her in her lab. I would also like to thank my thesis committee, Barak Blum, Emery Bresnick, Phil Newmark, and Sushmita Roy, whose valuable feedback and insight helped to make this project a success.

I also need to especially thank all of the lab-mates that I have had the pleasure of working with for all of these years. I have received a great deal of help from all of them throughout my time here, and they have all made the lab a terrific and supportive environment to be a part of. I need to especially acknowledge Coral Wille, Khoa Tran, and Nur Zafirah Zaidan, for their incredible mentorship early on in my graduate school career.

I want to thank all of the great friends that I have made here in Madison. They have been a constant source of joy, even through difficult stretches, and have made my time here a truly wonderful and memorable experience. I must also thank all of my close friends back home, for their continued support and encouragement from afar, and frequent check-ins to see how I'm doing, which has also been a great source of motivation.

Finally, I want to thank my family, particularly my mom and dad and my three sisters Rachel, Joelle, and Erika. Their unwavering love and support was undoubtedly one of the factors that motivated me to work hard and keep persevering, especially through particularly challenging times. I truly am very lucky to have the people that I do in my corner cheering me on. I know I certainly wouldn't be where I am today if it wasn't for all of them.

Abstract

Pluripotent stem cells have the incredible ability to self-renew indefinitely and to differentiate into all the cell types of the body. Remarkably, somatic cells can transition back into a pluripotent state through ectopic expression of the transcription factors OCT4, SOX2, KLF4, and MYC. This shift towards a state of higher differentiation potential represents an extraordinary capability to alter cell fate. However, reprogramming studies are challenging due to the inefficient and heterogeneous nature of the process. In this thesis, I address these challenges by enhancing reprogramming efficiency through addition of a combination of small molecules to the reprogramming culture and utilizing single-cell analytics to capture the transcriptional and chromatin accessibility dynamics of truly reprogramming cells.

A rationally designed combination of epigenetic-modifying small molecules with signaling pathway inhibitors improved reprogramming efficiency from 3% to over 40%. Single-cell RNA-seq profiling uncovered an accelerated, more highly coordinated reprogramming pathway in efficient reprogramming, with greater suppression of somatic and greater upregulation of cell cycle and pluripotent genes. It had been widely believed that the steps of reprogramming occur temporally with downregulation of somatic genes preceding downstream events like pluripotency gene upregulation. Instead gene co-expression within individual cells revealed that the reprogramming-associated transitional events are independently regulated and need not occur stepwise, disputing the existing reprogramming dogma. Aberrant regulation of necessary transcriptional changes (e.g. transient EHF expression, upregulation of EIF4A1) further compromises

the process and leads to branching away from a trajectory towards induced pluripotency.

Single-cell analysis of chromatin accessibility changes uncovered enhanced enrichment for 3D chromatin reorganization factor (KLF4, MAZ, and PATZ1) binding in high-efficient reprogramming. Cells experience more changes upon withdrawal of the ectopic reprogramming factors, including a Tcfap2c-mediated maintenance of the pluripotent state. Development of a new computational algorithm by my computational biologist collaborators – scCISINT – allowed for prediction of long-range interactions among differentially accessible regions across reprogramming clusters. I used CRISPRi repression at candidate interacting loci to validate the essential role of the putative long-range interactions in transforming cell fate. Altogether, my work has uncovered key gene expression and chromatin-associated features that guide cells along a path towards successful pluripotency acquisition.

Table of Contents

Chapter 1

Introduction and General Background

Pluripotency and Reprogramming

Pluripotent stem cells (PSCs) are characterized by their ability to self-renew indefinitely and to differentiate into any of the multitude of cell types that make up the three primary germ layers that form during embryonic development (endoderm, ectoderm, and mesoderm) (Fig.1), and thus, represent a state of high differentiation potential. PSCs arise during early development when the embryo reaches the blastocyst stage. The inner cell mass (ICM) of the blastocyst are made up of PSCs, which can be harvested and cultured as embryonic stem cells (ESCs) (Fig.1). Alternatively, PSCs can be obtained through converting a differentiated somatic cell into a pluripotent state, deemed induced pluripotent stem cells (iPSCs), in a process called somatic cell reprogramming (Fig. 1). The ability for cells to go from a state of lower differentiation potential to a higher one represents a remarkable feat in reversing and altering cell fate.

An early indicator of the facility of somatic cells to reprogram to iPSCs were observed in studies involving somatic cell nuclear transfer in which a donor somatic nucleus is put into an enucleated unfertilized oocyte (Gurdon et al., 1958; Wilmut et al., 1997) It was found that the resulting cells are able to differentiate and produce viable offspring. Furthermore, it was shown that fusion of somatic cells with pluripotent cells results in hybrid cells that are pluripotent, with the somatic genome becoming more embryonic in nature (Cowan et al., 2005; Tada et al., 2001). These results indicate that within the environment of the unfertilized oocyte or pluripotent stem cell, the nuclei of differentiated cells are able to transition back into an undifferentiated state.

Building on this, one of the key discoveries with regards to reprogramming came when Takahashi and Yamanaka found that a select group of transcription factors (TFs) – OCT4, SOX2, KLF4, and MYC (Yamanaka factors or OSKM) – could induce pluripotency in mouse fibroblasts, which provided a more efficient and simple means of acquiring iPSCs (Takahashi & Yamanaka, 2006). Importantly, mouse iPSCs are functionally equivalent to ESCs, and pass all the stringent tests for pluripotency, which include the ability to form teratomas, ability to produce chimeric offspring, and they pass the tetraploid complementation assay (Boland et al., 2009; Okita et al., 2007; X. Zhao et al., 2009). This transcription factor-mediated reprogramming was also shown to work in human somatic cells by the same Yamanaka factors (Takahashi et al., 2007), as well as through the expression of transcription factors OCT4, SOX2, NANOG, and RNA binding protein LIN28 (Yu et al., 2007). Therefore, reprogramming somatic cells into iPSCs provides a great alternative source for PSCs that avoid the ethical barriers associated with using embryo-derived tissue for research or potential clinical applications.

Cellular and Transcriptional Changes During Reprogramming

Since the discovery of TF-mediated reprogramming of somatic cells, several studies using large-scale genomics methods, such as RNA-seq, have been conducted to further understand the mechanisms underlying the shift from a differentiated to a pluripotent state. These studies, commonly using mouse embryonic fibroblasts (MEFs) as the starting cell type, have shown that reprogramming can be broken up into distinct phases, each associated with key steps in the reprogramming process. In the early initiation phase, cells initially lose their somatic cell identity as cell type-specific genes

become downregulated (Fig. 2), indicated by the loss of the fibroblast-expressed cell surface marker THY1 (Stadtfeld et al., 2008).

This is followed by a mesenchymal-to-epithelial transition (MET), where mesenchymal-associated genes (e.g. *Snai1*, *Zeb1*, *Twist1*) are downregulated and epithelial markers, such as CDH1 and EPCAM, become upregulated (Hussein et al., 2014; R. Li et al., 2010; Mikkelsen et al., 2008; Samavarchi-Tehrani et al., 2010) (Fig. 2). In the intermediate phase, reprogramming cells undergo a metabolic switch from oxidative phosphorylation to glycolysis, accompanied by an increase in glycolytic gene expression and in cell proliferation (Panopoulos et al., 2012; Varum et al., 2011) (Fig. 2). The final stages of reprogramming are characterized by cells becoming stabilized iPSCs through the upregulation of pluripotency-associated genes (*Nanog*, *Sall4*, *Lin28*, *Dppa4*, endogenous *Oct4/Sox2*) and are able to maintain their pluripotent identity upon removal of ectopic OSKM expression (transgene independence) (Apostolou & Hochedlinger, 2013; Apostolou & Stadtfeld, 2018; Golipour et al., 2012; Mikkelsen et al., 2008) (Fig. 2).

Reprogramming studies have also provided insight into the identifying features of cells at each phase of reprogramming or those that identify cells that are more prone to reprogram successfully. These features include the upregulation and downregulation of different cell surface markers, such as the aforementioned THY1 fibroblast marker. As cells reprogram, they lose *Thy1* expression, and gain expression alkaline phosphatase (AP) followed by the surface marker SSEA1. Fluorescence-activated cell sorting (FACS) of reprogramming intermediates based on the THY1 and SSEA1 surface markers further revealed that cells who lost THY1 but failed to gain SSEA1 (THY1-/SSEA1-)

represent unstable intermediates whereas cells that remain THY1+ retained high levels of fibroblast genes and failed to upregulate MET and pluripotency-associated genes (Brambrink et al., 2008; Polo et al., 2012; Stadtfeld et al., 2008). Additional reprogramming-associated surface markers include upregulation of the MET marker *Epcam* in the early phase, cell proliferation marker *c-Kit* in the intermediate phase, and *Pecam* in the late phase (Polo et al., 2012). CD44 is lost while ICAM1 is gained during reprogramming (O'Malley et al., 2013). Using mass cytometry, CD73, CD49D, and CD200 are also acquired during reprogramming (Lujan et al., 2015), and in yet another study, more MEF, iPSC, and transiently-expressed markers were identified, such as VCAM1 in MEFs and the transient marker SCA-1 (Schwarz et al., 2018). The presence or absence of these markers can thus be informative in identifying subpopulations of cells across reprogramming and can indicate at what stage of the reprogramming path they may fall (or in some instances, if they have deviated from the path).

Large-scale knockdown screens have been used to identify key proteins that act as barriers to successful reprogramming. In an RNA interference (RNAi) screen, it was observed that chromatin assembly factor-1 (CAF-1) is important for somatic cell identity. Suppression of this protein in reprogramming led to a more permissive and relaxed chromatin structure at enhancers early on, improved Sox2 binding, and a decrease in compacted heterochromatic regions in the starting somatic cells (Cheloufi et al., 2015). Similarly, using shRNAs to screen candidate genes identified SUMO2 as another reprogramming barrier. Knockdown of SUMO2 resulted in an enhanced and faster reprogramming process, implicating sumoylation of proteins as being combative towards reprogramming (Borkent et al., 2016).

There are key differences in the epigenomic and chromatin landscape between pluripotent and somatic cells that must be altered during the transition to pluripotency. One of these differences is that ESCs have a much more relaxed, open chromatin structure, as opposed to the more closed and compacted structure in somatic cells (Gaspar-Maia et al., 2011). Another major difference between MEFs and ESCs lies in their levels of various histone modifications. For example, the repressive H3K9me2/3 mark is more highly enriched in MEFs and has been shown to be a barrier to reprogramming (J. Chen et al., 2013; Soufi et al., 2012; Sridharan et al., 2013; Tran et al., 2015).

Similarly, H3K79me2/3, which mark regions of transcriptional activity, are also more enriched in MEFs compared to ESCs (Sridharan et al., 2013). While initial reports suggest that this mark is present at mesenchymal genes and must be lost before cells undergo MET (Onder et al., 2012), our lab has recently uncovered roles for H3K79me2/3 in inhibiting reprogramming beyond just this step. This is evidenced by inhibition of H3K79 methylation promoting iPSC formation in keratinocytes, which are already epithelial and do not require MET (Wille & Sridharan, 2022). Among these additional inhibitory roles are preventing establishment of a histone acetylation-rich state and a corresponding increase in transcriptional elongation levels to those seen in ESCs (Wille, Zhang, et al., 2023). In line with this, the interaction between H3K79 methyltransferase Dot1l and its cofactor AF10 was shown to promote higher order H3K79 methylation at genes, promoting maintenance of the somatic identity(Wille, Neumann, et al., 2023). Lastly, Dot1l activity upregulated expression of reprogramming-

associated genes such as *Nfix*, a TF that is important for maintaining neural and muscle cell identities (Wille & Sridharan, 2022).

Another interesting feature of ESCs is that many genes acquire bivalency, containing both the transcriptionally active mark H3K4me2/3 and the repressive H3K27me3 mark, making them poised for either turning on or off, depending on the cell the ESC is differentiating into (Azuara et al., 2006; Bernstein et al., 2006; Pan et al., 2007). Specifically, at pluripotency-associated promoters and enhancers, the acquisition of H3K4me2/3 is accompanied with loss of H3K27me3, as their expression is necessary for pluripotency maintenance (Koche et al., 2011). These examples highlight the necessity of an altered epigenome for the transition to iPSCs.

These reprogramming studies have also identified clonal intermediates that typically form later on during reprogramming and are referred to as partially reprogrammed iPSCs (pre-iPSCs) (Fig. 3). These intermediates have gone through most early steps of the reprogramming process and even adopt PSC-like features, such as the capacity for self-renewal and stem cell maintenance; however, they have not upregulated pluripotency-related genes, have failed to downregulate cell type-specific genes, or are genomically hypermethylated (Mikkelsen et al., 2008). Moreover, pre-iPSCs do not display the same binding at gene promoters by the reprogramming factors OCT4, SOX2, and KLF4 as in ESCs and iPSCs. These binding sites were most often also targets of Nanog, whose absence in pre-iPSCs could be causing the impaired OSK binding (Sridharan et al., 2009). pre-iPSCs also have a global histone modification landscape that more closely resembles that of MEFs than PSCs, such as greater levels of the repressive H3K9me2/3 modification (Sridharan et al., 2013). When the

reprogramming factor are sustained at elevated levels, somatic cells may also form "fuzzy" NANOG-positive colonies (F-class cells) that are not yet transgene independent, but are stable and display some of the defining characteristics of pluripotency, including the ability to form teratomas (Tonge et al., 2014).

While some reprogramming cells become partially reprogrammed intermediates, others may follow alternative paths and transition into a completely different cell type. It had been previously shown in some reprogramming studies that endodermal genes like *Gata4*, *Gata6*, and *Sox17* are upregulated in fibroblast reprogramming (Hou et al., 2013; Serrano et al., 2013; Y. Zhao et al., 2015), with some reports indicating them to be markers of partially reprogrammed cells and inhibitory for successful reprogramming (Mikkelsen et al., 2008; Serrano et al., 2013). In fact, one study found that some cells form colonies of induced extraembryonic endoderm (iXEN) cells (Fig. 3), expressing endoderm-associated genes (e.g. *Gata4/6*, *Sox7/16*, *Pdgfra*) at a level comparable to a blastocyst-derived primitive endoderm cell line. These iXEN cells were also capable of differentiating along other endoderm lineages (visceral and parietal endoderm) (Parenti et al., 2016). Importantly, these iXEN cells did not come from the iPSC colonies that have also formed, but rather came from a separate reprogramming pathway entirely (Parenti et al., 2016). Thus, in addition to becoming stalled intermediates, cells could also be refractory to reprogramming via falling down a trajectory towards a competing cell fate.

Alternative Reprogramming Systems

While the majority of reprogramming studies use MEFs as the starting cell type, somatic cell reprogramming has been shown to be applicable in different cells types which follow their own similar, but slightly different corresponding pathways. For example, in a comparison of three different cell types – fibroblasts, neutrophils, and keratinocytes – it was discovered that in all three cell types, they universally undergo the same main phases of reprogramming (loss of somatic identify and activation of the pluripotency transcriptional network) and share a requirement for downregulation of *Egr1*. They differ in the transcriptional changes that occur on their way to becoming iPSCs, such as keratinocytes not needing to undergo MET as they are already an epithelial cell type, and fibroblasts undergoing a transient upregulation of primitive streak genes that does not appear to occur in neutrophil or keratinocyte reprogramming (Nefzger et al., 2017). In the reprogramming of cells from the neural lineage (neural stem cells (NSCs) and astrocytes), it was shown that they must still acquire expression of both the epithelial marker Cadherin and pluripotency marker Nanog; however, the upregulation of E-cadherin can occur simultaneously with, or even after, upregulation of Nanog expression (Jackson et al., 2016). These studies help to highlight how applicable TF-mediated reprogramming is to different cell types to produce iPSCs.

In addition to using a different cell type, reprogramming can be performed by using transcription factors other than the Yamanaka factors. For example, it was shown that the combination of pluripotency factors SALL4, NANOG, ESRRB, and LIN28 (SNEL) can produce high quality iPSCs from MEFs (Buganim et al., 2014). Yet another study found that a combination of seven non-OSKM factors (7F) (JDP2, JHDM1B, MKK6, GLIS1, NANOG, ESRRB, and SALL4) was also capable of generating highly

competent iPSCs (B. Wang et al., 2019). These results illustrate the modular nature of reprogramming, with various combinations of factors having the capability to induce this transition.

It has further been discovered that by altering the cocktail of transcription factors that somatic cells are exposed to, they can drive reprogramming towards a state that resembles other early embryonic cell types outside of just ESCs. In the previously mentioned study of iXEN cells that appear in parallel with iPSCs, the researchers also found that shRNA knockdown of *Gata4* and *Gata6* led to the number of iXEN colonies to be reduced by half. Moreover, KD of *Gata6* also led to a significant increase in iPSC colonies (Parenti et al., 2016), indicating that these factors promote the iXEN fate.

Other studies have discovered that overexpression of different TFs (e.g. GATA3, EOMES, TFAP2C; or GATA3, OCT4, KLF4 and MYC) during mouse fibroblast reprogramming causes cells to transition into induced trophoblast stem cells (iTSCs) (Fig. 3), which were similar to blastocyst-derived TSCs in their epigenome, in their ability to differentiate into trophectoderm, and in their contribution towards placental development in chimeric mice (Benchetrit et al., 2015; Naama et al., 2023). Researchers later found that a subset of the same five transcription factors – GATA3, EOMES, TFAP2C, MYC, and ESRRB – were capable of producing the three pre-implantation blastocyst-like cell types: iPSCs, iTSCs, and iXEN cells. By modifying the proportion of each of these factors, they were able to guide reprogramming fibroblasts to each of these cell fates. Greater levels of EOMES pushes the cells towards the iTSC fate, while increasing ESRRB will guide cells to the iXEN fate, which will later transition to iPSCs (Benchetrit et al., 2019). It was also discovered that the TSC trajectory was

guided, in part, by DNA methylation-mediated repression of the pluripotency program (Jaber et al., 2022).

Reprogramming with Small Molecules

Despite the promise of somatic cell reprogramming to generate iPSCs for use in regenerative therapies, there are some setbacks associated with it. The process is slow and inefficient, resulting in 0.1%-5% of MEFs converting to iPSCs after about 2 weeks after inducing OSKM expression (Fig. 3), with iPSC colonies forming at different time points (Apostolou & Hochedlinger, 2013; Buganim et al., 2013; Papp & Plath, 2013). This is the case even when using cells that have all four factors under inducible expression at the same locus. As previously discussed, some reprogramming cells can become stalled pre-iPSCs or even an alternative cell type (e.g. iXEN cells), but they can also experience reprogramming-induced senescence (Banito et al., 2009), all of which factor into an inefficient process. To address this low efficiency, we and others have set out to improve reprogramming efficiency through the addition of small molecules to the reprogramming media (Esteban et al., 2010; Huangfu, Maehr, et al., 2008; Huangfu, Osafune, et al., 2008; Ichida et al., 2009, 2014; Maherali & Hochedlinger, 2009; Mikkelsen et al., 2008; Onder et al., 2012; Shi et al., 2008; Silva et al., 2008; Tran et al., 2015).

Some of these small molecules have enhanced reprogramming efficiency through the modulation of the epigenomic landscape through the inhibition of epigenome-modifying enzymes. For example, the inhibition of histone deacetylases (HDACs) with valproic acid (VPA) improved efficiency about 100-fold and could

effectively replace MYC and KLF4, reducing the number of reprogramming factors required (Huangfu, Maehr, et al., 2008; Huangfu, Osafune, et al., 2008). Moreover, inhibition of the DNA methyltransferase DNMT1 with 5-aza-cytidine (AZA) also resulted in accelerated reprogramming (Mikkelsen et al., 2008). Inhibiting the H3K9 methyltransferase G9A could also improve reprogramming efficiency and, similar to the addition of VPA, could replace one of the reprogramming factors, in this instance SOX2 (Shi et al., 2008).

Supplementing reprogramming media with ascorbic acid (AA), or Vitamin C, has also been shown to improve reprogramming efficiency (J. Chen et al., 2011), in part, by causing cells to avoid the senescence roadblock associated with reprogramming, and has also enhanced the conversion of stalled pre-iPSCs to make the final jump to iPSCs (Esteban et al., 2010; Tran et al., 2015). Ascorbic acid works through restoring alpha-ketoglutarate (2OG) dependent dioxygenase enzymes to a catalytically active state. AA acts as an electron donor, and reduces the Fe(IV) core of the inactive 2OG to Fe(II), making the enzyme active once again (Monfort & Wutz, 2013). Members of the 2OG dependent dioxygenase family of enzymes include the TET enzymes, responsible for DNA demethylation by catalyzing the conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), thereby removing the epigenetic memory associated with DNA methylation in differentiated cells. Thus, ascorbic acid enhances TET enzyme activity, leading to an increase in 5hmC levels (Blaschke et al., 2013; Hore et al., 2016; Tran, Dillingham, et al., 2019; Tran et al., 2015; Yin et al., 2013).

Additionally, ascorbic acid activates Jumonji domain-containing H3K9 demethylase enzymes (e.g. KDM3B). As stated earlier, H3K9 methylation has greater

enrichment in somatic cells (Sridharan et al., 2013) and is an epigenetic barrier to reprogramming, removal of which allows stalled pre-iPSCs to convert to iPSCs (J. Chen et al., 2013; Tran et al., 2015). It has also been shown that the increased activity of KDM3B by AA is important for the early phase of reprogramming (Tran, Dillingham, et al., 2019). AA also reactivates the H3K36 demethylases JHDM1A/1B which enhance reprogramming through repression of the reprogramming antagonistic *Ink4/Arf*, consequently suppressing senescence; additionally, JHDM1A, in concert with OCT4, activates the pluripotency-related mir302/367 microRNA clusters, thereby enhancing reprogramming (H. Li et al., 2009; T. Wang et al., 2011).

As previously mentioned, MEFs also have greater enrichment of H3K79me2/3 than ESCs (Sridharan et al., 2013). It has been shown that inhibition of Dot1l (Dot1li), the only known H3K79 methyltransferase enzyme, also improves reprogramming. When Dot1l is inhibited, this histone modification was lost at fibroblast genes early in reprogramming as well as other genes that must be repressed in pluripotent cells (Onder et al., 2012). Additionally Dot1li led to reduced higher-order H3K79me accumulation and enhanced acquisition of ESC-like H3K9 hyperacetylation, contributing to an increase in transcriptional elongation and redistribution of RNAPII throughout the gene body (Wille, Neumann, et al., 2023; Wille, Zhang, et al., 2023). Dot1li also repressed the reprogramming-associated expression of the antagonistic gene *Nfix*, and improved reprogramming when starting with an epithelial cell type (keratinocytes), further illustrating Dot1li can affect reprogramming outside of the MET step (Wille & Sridharan, 2022). Together, these results illustrate how effectively reprogramming can

be enhanced through the modulation of the epigenome via small molecule inhibition of epigenetic-modifying machinery.

Besides epigenome-modifying small molecules, chemicals that alter signaling pathways may also improve reprogramming efficiency. For example, inhibition of TGF-beta signaling with RepSox allowed for enhanced reprogramming via the upregulation of *Nanog* expression on a population of partially-reprogrammed intermediate cells, and it was shown that it could replace MYC or SOX2 in the reprogramming factor cocktail (Ichida et al., 2009; Maherali & Hochedlinger, 2009). Furthermore, inhibition of the NOTCH signaling pathway improved reprogramming efficiency through suppression of the anti-proliferative gene *p21* (Ichida et al., 2014).

Reprogramming has also been shown to be enhanced with the addition of two kinase inhibitors. The glycogen synthase kinase 3 (GSK3) inhibitor CHIR99021(CHIR) and an inhibitor of the Erk-activating MEK1/2 enzymes, PD0325901 (PD) (together, referred to as 2i) (Silva et al., 2008). It has been reported that inhibition of both of these signaling cascades promotes self-renewal of ESCs, with GSK3 inhibition operating by way of stimulation of Wnt signaling (Ying et al., 2008; Ying & Smith, 2017). In the culture of ESCs, adding 2i to the culture media, in combination with leukemia inhibitory factor (LIF), maintain ESCs in a naïve ground state, resembling PSCs from pre-implantation blastocyst ICM, through upregulation of TET1 and downregulation of DNA methyltransferase enzymes (Sim et al., 2017). Culturing ESCs with 2i has also led to a reduction in the repressive H3K27me3 mark at gene promoters, and a downregulation of somatic genes (Marks et al., 2012). In combination with ascorbic acid, GSK3 inhibition led to more efficient iPSC colony formation in multiple cell types and allowed

partially reprogrammed clones to transition into iPSCs (Bar-Nur et al., 2014). Combining GSK3 inhibition with a TGF-beta inhibitor and ascorbic acid in reprogramming resulted in ~80% of MEFs being able to form iPSC colonies after 7 days (Vidal et al., 2014).

In our lab, we have combined AA and 2i in reprogramming to great effect, acting synergistically to significantly improve efficiency and allowing stalled pre-iPSC intermediates to convert to iPSCs, in part, due to the actions of 2i against epidermal and insulin growth factor pathways which promotes expression of the pluripotency marker *Esrrb* (Tran et al., 2015). We later combined AA and 2i with the Dot1l inhibitor SGC0936 (altogether, referred to as A2S), to improve reprogramming efficiency >10-fold (Tran, Pietrzak, et al., 2019), which will be discussed further in Chapter 2 of this thesis.

Remarkably, other studies have demonstrated that reprogramming can be induced and carried out solely using small molecule compounds, effectively replacing ectopically expressed reprogramming factors as a stimulus for reprogramming (Hou et al., 2013; Y. Zhao et al., 2015). It was found that a chemical cocktail of 7 compounds could generate iPSCs. Forskolin was found to be a viable replacement of OCT4 in SKM-induced reprogramming, while VPA, CHIR, 616452, and tranylcypromine could promote reprogramming in single-factor (OCT4) reprogramming, thus acting as SKM substitutes. Combining these 5 chemicals with the late reprogramming epigenetic modifier DZNep and 2i could produce competent chemically induced pluripotent stem cells (CiPSCs) (Hou et al., 2013).

Another study from this same lab found that chemically induced reprogramming causes cells to become extra-embryonic endoderm (XEN)-like early on before transitioning to chemically-induced iPSCs (CiPSCs) late in the process, with expression

of XEN markers such as *Gata4/6* proving to be essential in chemical reprogramming. This information allowed them to improve their protocol through including additional small molecules (AM580 and EPZ004777) that promote XEN marker expression in the early stage of chemical reprogramming (Y. Zhao et al., 2015). An optimized fast chemical reprogramming system (FCR) was developed based on a large-scale screen of various compounds. Cells reprogrammed with this combination moved through a diapause-like state, paralleling a state of dormancy and low proliferation in the blastocyst that delays its implantation during development (X. Chen et al., 2023). Collectively, these studies show that chemical compounds can both enhance TF-mediated reprogramming, but can also eliminate the requirement for ectopic expression of these factors, further illustrating the malleable nature of the reprogramming protocol.

While these studies examined chemical reprogramming in MEFs, other research has unveiled that the same chemical system (with minor adjustments) can be applied to and induce reprogramming in other cell types, such as neural stem cells and intestinal epithelial cells (which, including MEFs, altogether represent cells from each of the three germ layers) (Ye et al., 2016). A human chemical reprogramming system has also been established which mediates induction of an intermediate plastic state early on and implicates the JNK signaling pathway as one of the major barriers in chemical reprogramming of human cells (Guan et al., 2022).

ATAC-seq Analysis of Reprogramming Chromatin Dynamics

Given the dynamic nature of the cellular epigenome during reprogramming and the influence that epigenetic-modifying small molecules can have on its efficiency,

recent studies have utilized the assay for transposase-accessible chromatin with sequencing (ATAC-seq) (Buenrostro et al., 2013) to probe the changing accessibility of chromatin throughout a reprogramming time course.

Profiling cells from MEF reprogramming identified distinct patterns of accessibility, with some regions going from an open to close (OC) state, and others going from closed to open (CO); these peaks are further grouped based on the day that this shift occurs, but it was discovered that these changes primarily occur in two large waves: an early closing of open sites and a late opening of closed sites. Furthermore, disruption of the correct chromatin accessibility dynamics can affect reprogramming, with somatic TFs (e.g. c-JUN, FRA1) acting as barriers to these changes (Chronis et al., 2017; D. Li et al., 2017). The histone deacetylase (HDAC) recruiter Sap30 has been implicated in the OC shift through reduced H3K27 acetylation at somatic sites (D. Li et al., 2017), and similarly binding of the HDAC protein Hdac1 was increased at OSK-bound MEF enhancers after 2 days of reprogramming (Chronis et al., 2017). After sorting cells for that that are successfully reprogramming (SSEA1+) vs refractory (THY1+), ATAC-seq revealed that many of the THY1+ cells retain accessibility at some MEF enhancer sites and also experience inadequate OCT4/SOX2 targeting and binding (Knaupp et al., 2017).

ATAC-seq has been combined with analysis of OSKM binding to elucidate the exact mechanism of these factors in guiding chromatin opening and closing, which has led to disputing claims. For example, one thought is that OCT4 and SOX2 facilitate the opening of transiently accessible regions that somatic TFs are redistributed to, away from somatic-associated loci (Knaupp et al., 2017). Others have proposed that

OCT4/SOX2/KLF4 binding at already open MEF enhancers leads to this redistribution (Chronis et al., 2017). Lastly, it was found that KLF4 plays a key role in the reorganization of 3D chromatin structure and enhancer looping, as KLF4 depletion at pluripotent enhancers compromised enhancer-promoter interactions (Di Giammartino et al., 2019). Similarly, it was discovered that the opening of chromatin by OCT4 promotes KLF4 binding early in reprogramming which mediates MET (K. Chen et al., 2020).

The chromatin changes associated with reprogramming have been examined in non-OSKM mediated systems as well. In the chemical reprogramming system, the chromatin undergoes two stages of changes, including transition towards an intermediate state (mediated by exposure to the full chemical reprogramming cocktail), followed by a push toward a pluripotent state (mediated by a switch to 2i/LIF media). From their full chemical mix, they uncovered a particularly interesting role for the synthetic thymidine analog bromodeoxyuridine (BrdU), which they found to be important in the proper opening of sites enriched for not only KLF and SOX motifs, but GATA and FOX as well (Cao et al., 2018). Analysis of the optimized fast chemical reprogramming protocol found an upregulation of motifs associated with XEN TFs (SOX17, GATA4/6, and FOXA2) (X. Chen et al., 2023). ATAC-seq of the previously mentioned 7-factor reprogramming system found that it is distinct from OSKM with some differences in the motif enrichment patterns, such as earlier opening ESRRB motif sites compared to OSKM (B. Wang et al., 2019). In a modified reprogramming protocol, cells are transiently grown in a naïve culture medium to generate naïve iPSCs, which are akin to PSCs from the pre-implantation blastocyst. ATAC-seq revealed that there were fewer

differentially open regions between the naïve iPSCs and ESCs when compared to the more post-implantation-resembling primed iPSCs (Buckberry et al., 2023)

ATAC-seq has also been applied to reprogramming systems using cell types other than fibroblasts. It was found that exposing B cells to C/EBP-alpha makes them more elite and primed for efficient reprogramming to iPSCs. ATAC-seq of these cells, along with unaffected B cells, ESCs, and Day 1 & 2 reprogramming cells found that pulsing with C/EBP-alpha resulted in 525 newly opened regions that are also accessible in ESCs, which were also enriched for Klf4 binding sites, suggesting that C/EBP-alpha and Klf4 work together to alter chromatin accessibility in B cell reprogramming (Di Stefano et al., 2016).

Single-Cell Analysis of Reprogramming

The variability in reprogramming kinetics, combined with the presence of cells falling down alternate pathways or forming stalled intermediates results in reprogramming populations that are largely heterogeneous in their composition. Therefore, bulk population-based analyses separated by timepoint can obfuscate the changes associated with the truly reprogramming cells across a time course. To overcome this issue, recent reprogramming studies have implemented a single-cell approach to better understand elucidate the key dynamics associated with this process at the resolution of individual cells.

As an example, different studies have implemented single-cell mass cytometry of cells across a reprogramming time course. From these analyses, the cell surface markers associated with transient or partially reprogrammed cells were identified

(CD73, CD49D, CD200), and represent cells in a state between MEFs and iPSCs (Lujan et al., 2015). Another single-cell mass cytometry study from the same group identified additional markers indicative of successful reprogramming (Zunder et al., 2015). Here, they found that cells with high OCT4 and KLF4 transitioned to an intermediate state marked by high expression of CD73 and CD104, but decreased CD54. From this intermediate group, cells with low KI67 will revert to a MEF-like state, while those that are high in KI67 will progress through MET. After this, the cells will diverge yet again, with cells high in NANOG, SOX2, and CD54 becoming ESC-like, and another group (marked by elevated LIN28, CD24, PDGFR) are mesendoderm-like. Single-cells have also been arranged in a fluidigm microarray to profile a set of known pluripotency/ESC and proliferative genes across reprogramming, which ultimately revealed a stochastic change in gene expression early, followed by a hierarchical cascade of late in reprogramming mediated by SOX2 (Buganim et al., 2012).

Single-Cell RNA-seq

In order to gain a better understanding of the transcriptional dynamics associated with reprogramming, recent studies have used single-cell RNA-seq (scRNA-seq) to capture the gene expression landscape within individual cells, albeit largely in low-efficiency systems. One study applied scRNA-seq to cells at different stages of reprogramming, and discovered early activation of Ras signaling and expression of long non-coding RNAs (which downregulate somatic genes) are key reprogramming events (Kim et al., 2015). Profiling cells from the chemical reprogramming timeline found that the transition from the intermittent XEN-like state to iPSCs is due to transcription of

genes from the 2-cell stage of development as well as early pluripotency genes (T. Zhao et al., 2018). Based on scATAC-seq data, a pathway for reprogramming cells to diverge into iTSCs emerged, implicating trophectoderm-associated TFs in this alternative trajectory (Liu et al., 2020).

Additional bifurcation events were revealed by scRNA-seq such as KLF4 promoting a keratinocyte fate and IFN-gamma preventing late transition to pluripotency (D. Li et al., 2017). The scRNA-seq profiles of 315,000 reprogramming cells were used to develop a computational framework called Waddington-OT, which utilizes the mathematical principle of optimal transport to identify ancestor-descendant relationships. This analysis further identified a bifurcation event where cells become stromal-like or continue on with MET; the latter group can subsequently diverge into pluripotent, neural, or extra-embryonic cells (Schiebinger et al., 2019). The use of scRNA-seq, therefore, was able to identify gene expression patterns associated with successfully reprogramming vs refractory cells and cell fate decision points during reprogramming that might otherwise be hidden from a population-based analysis.

We employed scRNA-seq on low- and high-efficiency reprogramming, uncovering an accelerated process in the presence of small molecules and gene co-expression events that challenge the existing reprogramming dogma. These results are discussed further in Chapter 2.

Single-Cell ATAC-seq

Akin to RNA-seq, the recent development of single-cell ATAC-seq (scATAC-seq) technology has also allowed for higher resolution chromatin accessibility studies to be

performed on reprogramming cell populations, all of which have thus far been performed in human samples (Liuyang et al., 2023; Nair et al., 2023; Xing et al., n.d.). Profiling reprogramming cells from human BJ fibroblasts with scATAC-seq found a decision point where cells that transition from a regulatory network centered around the somatic transcription factor Fosl1 to Tead4 will successfully acquire pluripotency as opposed to those that fail to do so (Xing et al., 2020). scATAC-seq was used to examine a chemical reprogramming system, in this case the reprogramming of human adipose-derived stromal cells (hADSCs) to epithelial-like cells using a control old condition and a new optimized, serum-free system. hADSCs, and late reprogramming cells (just prior to XEN upregulation) in the original and optimized conditions were analyzed and they found that in the optimized protocol, the XEN-associated loci were closed, while pluripotency loci were more open, suggesting that chemically reprogrammed cells can actually bypass this XEN-like state and become pluripotent directly (Liuyang et al., 2023). Lastly, scATAC-seq analysis of human fibroblast reprogramming revealed that when OSK induces reprogramming, there is opening of transient regulatory loci, which led them to them discovering that these sites bind and sequester somatic TFs, providing further evidence that redistribution of these somatic TFs via OSK opening of transient sites (Nair et al., 2023). To date, there has been no published report of scATAC-seq in mouse reprogramming.

We combined the scRNA-seq and scATAC-seq methodologies along with enhancement of reprogramming efficiency using small molecules to elucidate the differences in gene expression and chromatin accessibility that are characteristic of a high-efficient reprogramming system and contribute to the effective transition to a

pluripotent state. This research is explored further in Chapter 2 (scRNA-seq analysis) and Chapter 3 (scATAC-seq analysis) of this thesis.

Single-Cell Data Analysis Algorithms

Given the complexity of large datasets from single-cell sequencing methods, it becomes necessary to effectively process and analyze the data by reducing its complexity without losing valuable information from the data. In the wake of single-cell studies becoming more prevalent, several labs have developed different algorithms for single-cell analysis. It is therefore imperative to choose an appropriate analytical platform that is easy to implement while also getting the most out of the data.

scRNA-seq Algorithms

Two popularly used algorithms to analyze scRNA-seq data are Monocle (Qiu, Hill, et al., 2017; Qiu, Mao, et al., 2017; Trapnell et al., 2014) and Seurat (Satija et al., 2015). Both platforms perform the same basic steps important for analysis of a complex single-cell dataset. This includes a dimensionality reduction step to make the complex single-cell data easier to analyze, clustering of cells into separate groups based on similarity in their gene expression profile, and visualization of clusters in a 2D space (Fig. 4A).

What set Monocle apart is that it introduced cellular trajectory analysis, ordering cells chronologically within a pseudotime space (Fig. 4A). To do this, Monocle utilizes DDRTree (Mao et al., 2015, 2017), which generates a tree along which individual cells are organized. A unique facet of Monocle is its ability to identify branch points along the

pseudotime trajectory, which represent populations of cells that have somehow diverged from the main pathway being studied (Fig. 4A). Applying the Monocle DDRTree tool to a dataset of differentiating myoblasts, Monocle was able to recapitulate a differentiation trajectory that had one branch point where cells could take one of two paths. This divergence was the result of differences in the expression of over 800 genes associated with muscle contraction (Qiu, Mao, et al., 2017).

Monocle was applied to datasets of blood cell differentiation, along with other trajectory inference algorithms including, at the time of this study, the older original version of Monocle (Trapnell et al., 2014), Diffusion Pseudotime (DPT) (Haghverdi et al., 2016), Wishbone (Setty et al., 2016), and SLICER (Welch et al., 2016). When compared with a reference order based on expression of marker genes, Monocle v2 was as good or better than other algorithms at being able to assign cells to the correct lineage and branches. It was also the most consistent at recapitulating accurate trajectory and pseudotime order upon downsampling of the datasets compared to the other methods (Qiu, Mao, et al., 2017).

In a large-scale benchmarking experiment comparing 45 different trajectory inference tools on >300 real and synthetic datasets, Monocle was identified as one of the top performing methods with regards to the topology metric, which is a measure of how accurately the method was able to identify and project the correct trajectory shape and bifurcation events. It also performed well at analyzing datasets with more complex topologies compared to many of the other methods (Saelens et al., 2019). Given the variety of features and its high performance metrics, Monocle is a suitable tool for

effective and comprehensive scRNA-seq data analysis, and was used for scRNA-seq data analysis in Chapter 2 of this thesis.

scATAC-seq Algorithms

Analysis of scATAC-seq data presents its own set of unique challenges in that the data is quite sparse, with only 1-10% of accessible chromatin actually being targeted for fragmentation and eventual library preparation (H. Chen et al., 2019). Furthermore, unlike with RNA-seq data, there is no pre-determined finite list of potential features, i.e. a list of all possible genes, associated with ATAC-seq, leading to large cell to cell variability in the sequenced regions and high dimensionality data.

The very recent commercial availability of scATAC-seq has led to different labs trying to take advantage of the vacuum of supported analytical tools and the development of new algorithms. A recent benchmarking test (Luo et al., 2023) compared some of these algorithms including ArchR (Granja et al., 2021), Signac (Stuart et al., 2021), SnapATAC, and SnapATAC2 (Fang et al., 2021). These methods implement different dimensionality reduction methods. Signac and ArchR use latent semantic indexing (LSI), identifying the most common features across a dataset vs those that are more unique to a particular group of cells. SnapATAC, on the other hand, uses diffusion maps, which attempt to generate a graphical representation of the data to identify the most variable features. Testing datasets from a variety of sources, the benchmarking test evaluated different metrics including cell embedding (projection of reduced dimension data onto 2D space), cell partitioning or clustering, and memory

usage. SnapATAC appeared to be better than the LSI-based methods at separating complex datasets (Luo et al., 2023).

One key aspect to factor into choice of algorithm is how scalable the algorithm is to process variable dataset sizes while efficiently using both time and memory. ArchR performed the best with regards to memory usage, while SnapATAC's memory usage increased with increasing data size, making ArchR ideal for analysis of large-scale datasets, especially with limited memory storage (Luo et al., 2023).

ArchR provides additional features that surpass or are not available in other competing methods. One such feature is the identification and removal of cell doublets, in which two different cells are partitioned in the same single-cell reaction compartment (Granja et al., 2021). Differences are also seen in the identification of peaks. Signac uses a list of pre-selected peaks from which they generate a counts matrix, tabulating instances of these peaks and removing the contribution of any lowly accessible or novel regions in rare cell types. SnapATAC counts open regions within pre-determined bins; however, their bins are quite large (5kb) which far surpass the size of genomic regulatory elements, thus not allowing for clear separation of multiple elements within that window (Granja et al., 2021). ArchR curtails this issue by counting peaks using 500bp bins across the genome (Fig. 4B). For dimensionality reduction, ArchR applies the aforementioned LSI in an iterative fashion, which first identifies the most common peaks that vary across the different major clusters/cell types. Subsequently, the most variable features identified are used for the next iteration of LSI, finding even more sources of variance, even among the established clusters while also reducing any contributions from batch effects. In this study, ArchR actually performed better

clustering of a hematopoietic cell dataset compared to other methods, as these methods were unable to overcome batch effects, preferentially clustering the cells based on their input samples rather than the variety of different cell types present in the samples (Granja et al., 2021). From there, the resulting cell clusters can be used to perform downstream analyses, including finding differentially accessible regions and enriched motifs between clusters. Due to its efficient use of time and memory, its comprehensive arsenal of analytical tools, and its use of iterative LSI for effective dimensionality reduction, I employed ArchR for analysis of scATAC-seq data in Chapter 3 of this thesis.

Given the inefficiency of somatic cell reprogramming and the inherent challenges associated with studying such a heterogeneous process, my goals for this thesis were to first, establish a high-efficiency reprogramming system via a rationally designed combination of small molecules; my next goal was to then combine this system with single-cell analytics to uncover the shifting expression and chromatin accessibility dynamics associated with an efficient pathway towards pluripotency acquisition.

Figure 1

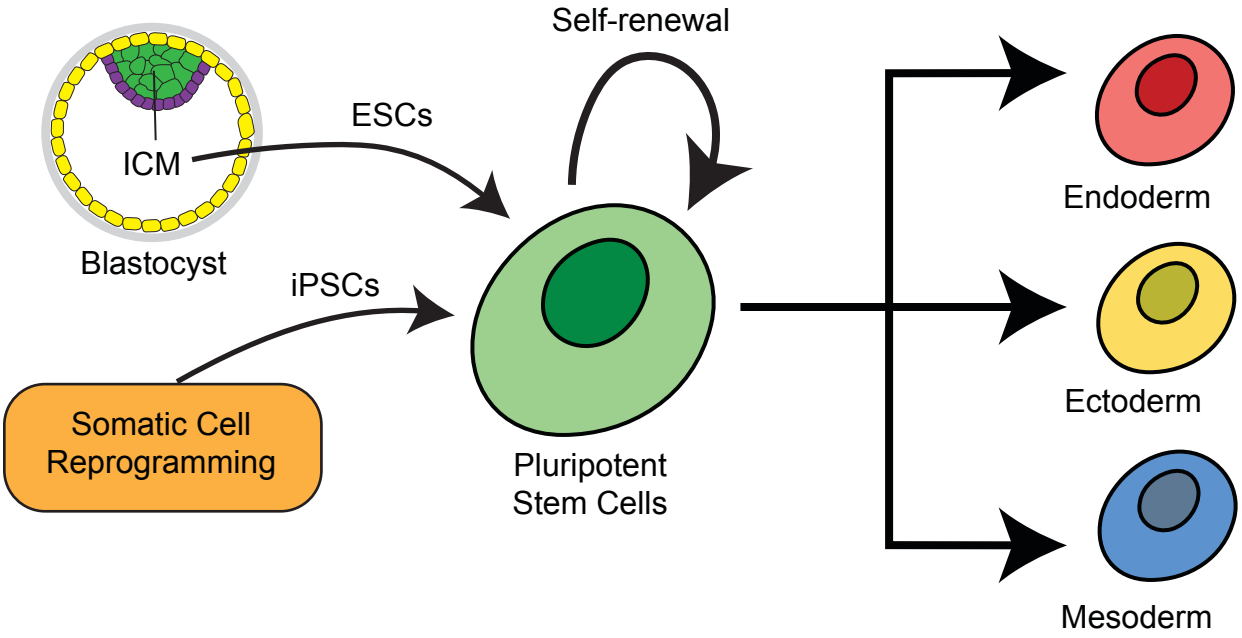


Figure 1: Pluripotent stem cell properties and sources

Pluripotent stem cells (PSCs) have ability to self-renew indefinitely and differentiate into the different cell types that make up the three germ layers. They can be derived from the inner cell mass of the blastocyst (embryonic stem cells (ESCs)) or from somatic cell reprogramming of differentiated cells (induced pluripotent stem cells (iPSCs)).

Figure 2

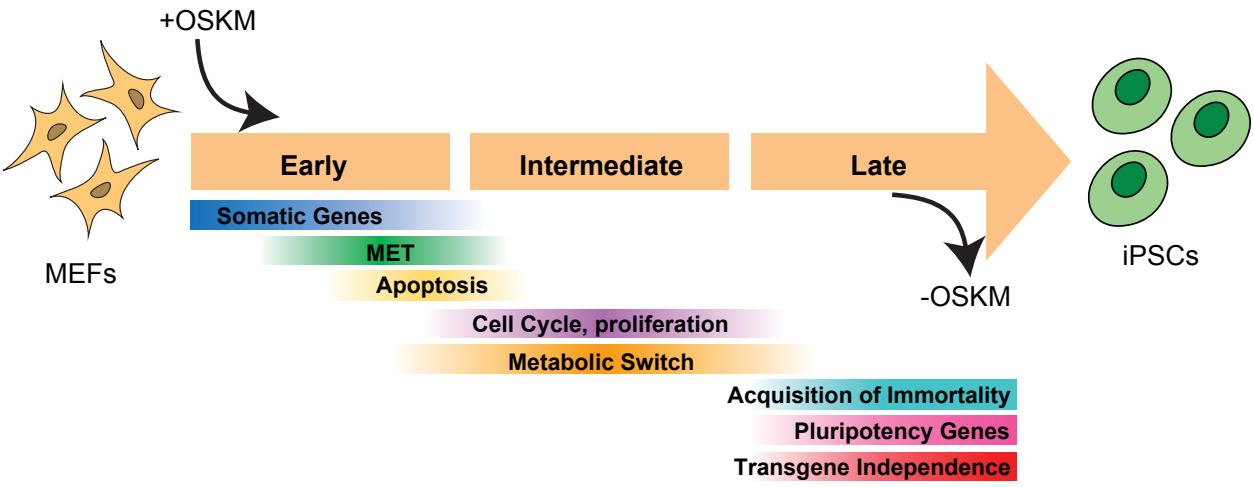


Figure 2: Cellular and transcriptional changes during reprogramming

Schematic highlighting the cellular and transcriptional changes associated with the transition from mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs). Based on population-based studies, these events have been posited to occur in a sequential and temporal manner. (Figure adapted from Apostolou & Hochedlinger, 2013).

Figure 3

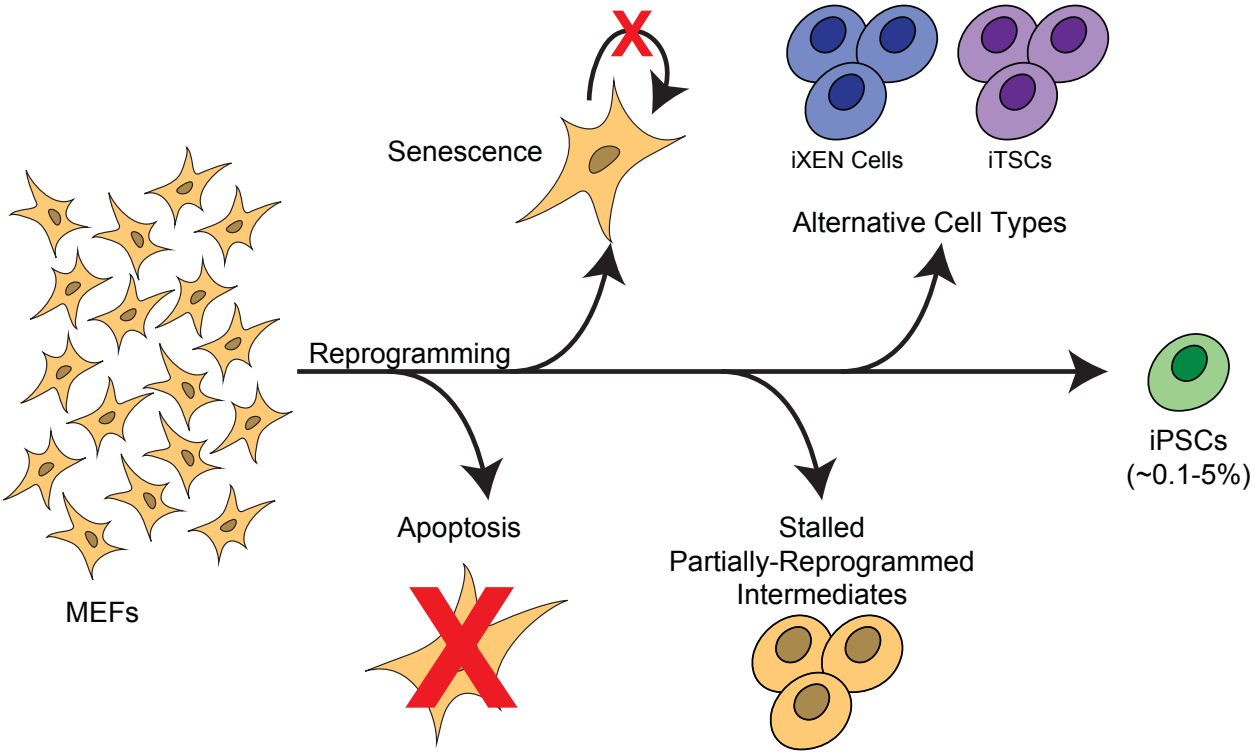


Figure 3: Reprogramming is inefficient and heterogeneous

Reprogramming is an inefficient process with only about 0.1-5% of the starting population of cells making the transition to iPSCs. Contributing to this inefficiency are cells that break away from the main reprogramming pathway, such as undergoing reprogramming-induced senescence, becoming stuck in a partially reprogrammed iPSC (pre-iPSCs) state, or transitioning to a competing cell type. These factors contribute to the prevalent heterogeneity of reprogramming populations.

Figure 4

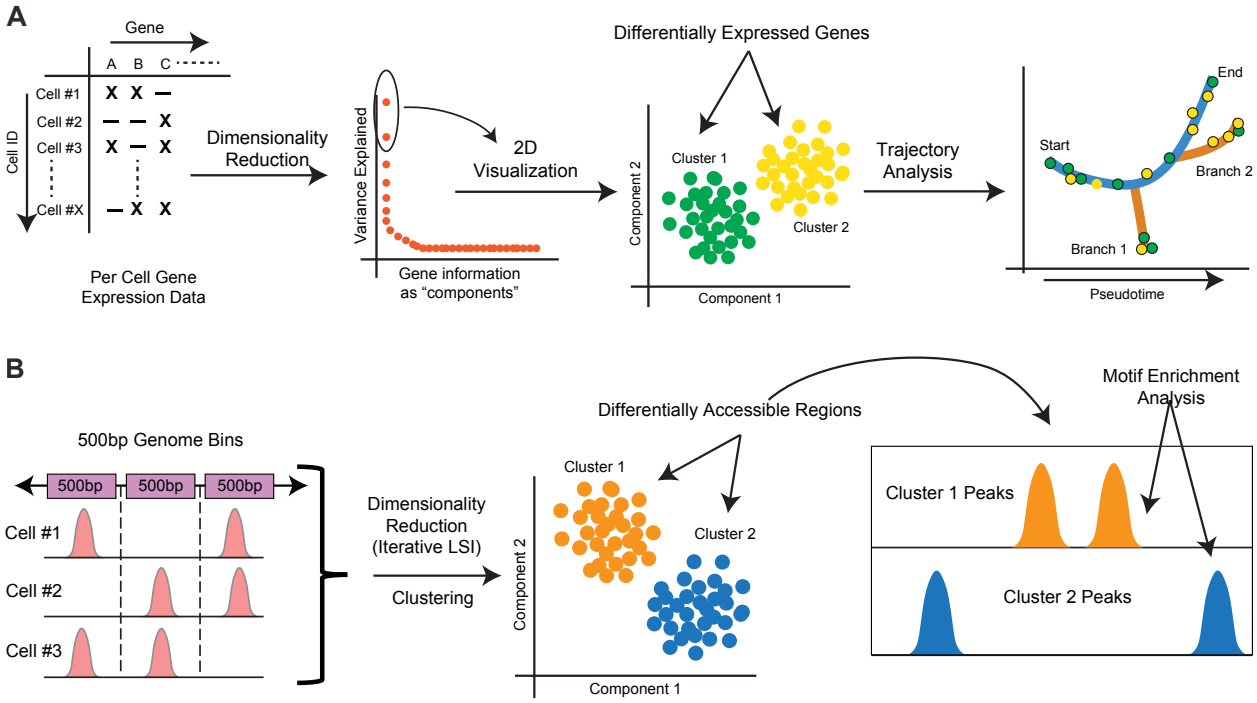


Figure 4: Workflow of scRNA-seq and scATAC-seq Data Analysis

A) Dimensionality reduction through identifying features, or genes, whose expressions contribute the most to the observed variance among the cells, which are then used to inform cell clustering. scRNA-seq analysis tools, (e.g. Monocle), can then perform downstream analyses such as identifying differentially accessible genes between clusters and constructing pseudotime trajectory.

B) In the analysis of scATAC-seq data with ArchR, instances of accessible peaks are counted using a tile matrix of 500bp bins throughout the genome. Once the counts matrix is generated, scATAC-seq data also undergoes dimensionality reduction and clustering. From there, the resulting clusters can be used to identify loci that are differentially accessible between clusters or cell populations and finding which motifs are enriched in the clusters associated with these peaks.

References

Apostolou, E., & Hochedlinger, K. (2013). Chromatin dynamics during cellular reprogramming. *Nature*, *502*(7472), Article 7472. https://doi.org/10.1038/nature12749

Apostolou, E., & Stadtfeld, M. (2018). Cellular trajectories and molecular mechanisms of iPSC reprogramming. *Current Opinion in Genetics & Development*, *52*, 77–85. https://doi.org/10.1016/j.gde.2018.06.002

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H. F., John, R. M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., & Fisher, A. G. (2006). Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, *8*(5), Article 5. https://doi.org/10.1038/ncb1403

Banito, A., Rashid, S. T., Acosta, J. C., Li, S., Pereira, C. F., Geti, I., Pinho, S., Silva, J. C., Azuara, V., Walsh, M., Vallier, L., & Gil, J. (2009). Senescence impairs successful reprogramming to pluripotent stem cells. *Genes & Development*, *23*(18), 2134–2139. https://doi.org/10.1101/gad.1811609

Bar-Nur, O., Brumbaugh, J., Verheul, C., Apostolou, E., Pruteanu-Malinici, I., Walsh, R. M., Ramaswamy, S., & Hochedlinger, K. (2014). Small molecules facilitate rapid and synchronous iPSC generation. *Nature Methods*, *11*(11), Article 11. https://doi.org/10.1038/nmeth.3142

Benchetrit, H., Herman, S., van Wietmarschen, N., Wu, T., Makedonski, K., Maoz, N., Yom Tov, N., Stave, D., Lasry, R., Zayat, V., Xiao, A., Lansdorp, P. M., Sebban, S., & Buganim, Y. (2015). Extensive Nuclear Reprogramming Underlies Lineage Conversion into Functional Trophoblast Stem-like Cells. *Cell Stem Cell*, *17*(5), 543–556. https://doi.org/10.1016/j.stem.2015.08.006

Benchetrit, H., Jaber, M., Zayat, V., Sebban, S., Pushett, A., Makedonski, K., Zakheim, Z., Radwan, A., Maoz, N., Lasry, R., Renous, N., Inbar, M., Ram, O., Kaplan, T., & Buganim, Y. (2019). Direct Induction of the Three Pre-implantation Blastocyst Cell Types from Fibroblasts. *Cell Stem Cell*, *24*(6), 983-994.e7. https://doi.org/10.1016/j.stem.2019.03.018

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., & Lander, E. S. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, *125*(2), 315–326. https://doi.org/10.1016/j.cell.2006.02.041

Blaschke, K., Ebata, K. T., Karimi, M. M., Zepeda-Martínez, J. A., Goyal, P., Mahapatra, S., Tam, A., Laird, D. J., Hirst, M., Rao, A., Lorincz, M. C., & Ramalho-Santos, M. (2013). Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-

like state in ES cells. *Nature*, *500*(7461), Article 7461. https://doi.org/10.1038/nature12362

Boland, M. J., Hazen, J. L., Nazor, K. L., Rodriguez, A. R., Gifford, W., Martin, G., Kupriyanov, S., & Baldwin, K. K. (2009). Adult mice generated from induced pluripotent stem cells. *Nature*, *461*(7260), Article 7260. https://doi.org/10.1038/nature08310

Borkent, M., Bennett, B. D., Lackford, B., Bar-Nur, O., Brumbaugh, J., Wang, L., Du, Y., Fargo, D. C., Apostolou, E., Cheloufi, S., Maherali, N., Elledge, S. J., Hu, G., & Hochedlinger, K. (2016). A Serial shRNA Screen for Roadblocks to Reprogramming Identifies the Protein Modifier SUMO2. *Stem Cell Reports*, *6*(5), 704–716. https://doi.org/10.1016/j.stemcr.2016.02.004

Brambrink, T., Foreman, R., Welstead, G. G., Lengner, C. J., Wernig, M., Suh, H., & Jaenisch, R. (2008). Sequential Expression of Pluripotency Markers during Direct Reprogramming of Mouse Somatic Cells. *Cell Stem Cell*, *2*(2), 151–159. https://doi.org/10.1016/j.stem.2008.01.004

Buckberry, S., Liu, X., Poppe, D., Tan, J. P., Sun, G., Chen, J., Nguyen, T. V., de Mendoza, A., Pflueger, J., Frazer, T., Vargas-Landín, D. B., Paynter, J. M., Smits, N., Liu, N., Ouyang, J. F., Rossello, F. J., Chy, H. S., Rackham, O. J. L., Laslett, A. L., … Lister, R. (2023). Transient naive reprogramming corrects hiPS cells functionally and epigenetically. *Nature*, *620*(7975), Article 7975. https://doi.org/10.1038/s41586-023-06424-7

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), Article 12. https://doi.org/10.1038/nmeth.2688

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., van Oudenaarden, A., & Jaenisch, R. (2012). Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell*, *150*(6), 1209–1222. https://doi.org/10.1016/j.cell.2012.08.023

Buganim, Y., Faddah, D. A., & Jaenisch, R. (2013). Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics*, *14*(6), Article 6. https://doi.org/10.1038/nrg3473

Buganim, Y., Markoulaki, S., van Wietmarschen, N., Hoke, H., Wu, T., Ganz, K., Akhtar-Zaidi, B., He, Y., Abraham, B. J., Porubsky, D., Kulenkampff, E., Faddah, D. A., Shi, L., Gao, Q., Sarkar, S., Cohen, M., Goldmann, J., Nery, J. R., Schultz, M. D., … Jaenisch, R. (2014). The Developmental Potential of iPSCs Is Greatly

Influenced by Reprogramming Factor Selection. *Cell Stem Cell*, *15*(3), 295–309. https://doi.org/10.1016/j.stem.2014.07.003

Cao, S., Yu, S., Li, D., Ye, J., Yang, X., Li, C., Wang, X., Mai, Y., Qin, Y., Wu, J., He, J., Zhou, C., Liu, H., Zhao, B., Shu, X., Wu, C., Chen, R., Chan, W., Pan, G., … Pei, D. (2018). Chromatin Accessibility Dynamics during Chemical Induction of Pluripotency. *Cell Stem Cell*, *22*(4), 529-542.e5. https://doi.org/10.1016/j.stem.2018.03.005

Cheloufi, S., Elling, U., Hopfgartner, B., Jung, Y. L., Murn, J., Ninova, M., Hubmann, M., Badeaux, A. I., Euong Ang, C., Tenen, D., Wesche, D. J., Abazova, N., Hogue, M., Tasdemir, N., Brumbaugh, J., Rathert, P., Jude, J., Ferrari, F., Blanco, A., … Hochedlinger, K. (2015). The histone chaperone CAF-1 safeguards somatic cell identity. *Nature*, *528*(7581), Article 7581. https://doi.org/10.1038/nature15749

Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., Andrade-Navarro, M. A., Buenrostro, J. D., & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, *20*(1), 241. https://doi.org/10.1186/s13059-019-1854-5

Chen, J., Liu, H., Liu, J., Qi, J., Wei, B., Yang, J., Liang, H., Chen, Y., Chen, J., Wu, Y., Guo, L., Zhu, J., Zhao, X., Peng, T., Zhang, Y., Chen, S., Li, X., Li, D., Wang, T., & Pei, D. (2013). H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature Genetics*, *45*(1), Article 1. https://doi.org/10.1038/ng.2491

Chen, J., Liu, J., Chen, Y., Yang, J., Chen, J., Liu, H., Zhao, X., Mo, K., Song, H., Guo, L., Chu, S., Wang, D., Ding, K., & Pei, D. (2011). Rational optimization of reprogramming culture conditions for the generation of induced pluripotent stem cells with ultra-high efficiency and fast kinetics. *Cell Research*, *21*(6), Article 6. https://doi.org/10.1038/cr.2011.51

Chen, K., Long, Q., Xing, G., Wang, T., Wu, Y., Li, L., Qi, J., Zhou, Y., Ma, B., Schöler, H. R., Nie, J., Pei, D., & Liu, X. (2020). Heterochromatin loosening by the Oct4 linker region facilitates Klf4 binding and iPSC reprogramming. *The EMBO Journal*, *39*(1), e99165. https://doi.org/10.15252/embj.201899165

Chen, X., Lu, Y., Wang, L., Ma, X., Pu, J., Lin, L., Deng, Q., Li, Y., Wang, W., Jin, Y., Hu, Z., Zhou, Z., Chen, G., Jiang, L., Wang, H., Zhao, X., He, X., Fu, J., Russ, H. A., … Zhu, S. (2023). A fast chemical reprogramming system promotes cell identity transition through a diapause-like state. *Nature Cell Biology*, *25*(8), Article 8. https://doi.org/10.1038/s41556-023-01193-x

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., & Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates

Reprogramming. *Cell*, *168*(3), 442-459.e20. https://doi.org/10.1016/j.cell.2016.12.016

Cowan, C. A., Atienza, J., Melton, D. A., & Eggan, K. (2005). Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells. *Science*, *309*(5739), 1369–1373. https://doi.org/10.1126/science.1116447

Di Giammartino, D. C., Kloetgen, A., Polyzos, A., Liu, Y., Kim, D., Murphy, D., Abuhashem, A., Cavaliere, P., Aronson, B., Shah, V., Dephoure, N., Stadtfeld, M., Tsirigos, A., & Apostolou, E. (2019). KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nature Cell Biology*, *21*(10), 1179–1190. https://doi.org/10.1038/s41556-019-0390-6

Di Stefano, B., Collombet, S., Jakobsen, J. S., Wierer, M., Sardina, J. L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., Limone, F., Mann, M., Porse, B., Thieffry, D., & Graf, T. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. *Nature Cell Biology*, *18*(4), Article 4. https://doi.org/10.1038/ncb3326

Esteban, M. A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S., Chen, K., Li, Y., Liu, X., Xu, J., Zhang, S., Li, F., He, W., Labuda, K., Song, Y., … Pei, D. (2010). Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells. *Cell Stem Cell*, *6*(1), 71–79. https://doi.org/10.1016/j.stem.2009.12.001

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A. K., Zhou, X., Xie, F., Mukamel, E. A., Zhang, K., Zhang, Y., Behrens, M. M., Ecker, J. R., & Ren, B. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21583-9

Gaspar-Maia, A., Alajem, A., Meshorer, E., & Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology*, *12*(1), Article 1. https://doi.org/10.1038/nrm3036

Golipour, A., David, L., Liu, Y., Jayakumaran, G., Hirsch, C. L., Trcka, D., & Wrana, J. L. (2012). A Late Transition in Somatic Cell Reprogramming Requires Regulators Distinct from the Pluripotency Network. *Cell Stem Cell*, *11*(6), 769–782. https://doi.org/10.1016/j.stem.2012.11.008

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, *53*(3), Article 3. https://doi.org/10.1038/s41588-021-00790-6

Guan, J., Wang, G., Wang, J., Zhang, Z., Fu, Y., Cheng, L., Meng, G., Lyu, Y., Zhu, J., Li, Y., Wang, Y., Liuyang, S., Liu, B., Yang, Z., He, H., Zhong, X., Chen, Q., Zhang, X., Sun, S., … Deng, H. (2022). Chemical reprogramming of human somatic cells to pluripotent stem cells. *Nature*, 1–7. https://doi.org/10.1038/s41586-022-04593-5

Gurdon, J. B., Elsdale, T. R., & Fischberg, M. (1958). Sexually Mature Individuals of Xenopus laevis from the Transplantation of Single Somatic Nuclei. *Nature*, *182*(4627), Article 4627. https://doi.org/10.1038/182064a0

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), Article 10. https://doi.org/10.1038/nmeth.3971

Hore, T. A., Meyenn, F. von, Ravichandran, M., Bachman, M., Ficz, G., Oxley, D., Santos, F., Balasubramanian, S., Jurkowski, T. P., & Reik, W. (2016). Retinol and ascorbate drive erasure of epigenetic memory and enhance reprogramming to naïve pluripotency by complementary mechanisms. *Proceedings of the National Academy of Sciences*, *113*(43), 12202–12207. https://doi.org/10.1073/pnas.1608679113

Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., Ge, J., Xu, J., Zhang, Q., Zhao, Y., & Deng, H. (2013). Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. *Science*, *341*(6146), 651–654. https://doi.org/10.1126/science.1239278

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A. E., & Melton, D. A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nature Biotechnology*, *26*(7), Article 7. https://doi.org/10.1038/nbt1418

Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W., & Melton, D. A. (2008). Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nature Biotechnology*, *26*(11), Article 11. https://doi.org/10.1038/nbt.1502

Hussein, S. M. I., Puri, M. C., Tonge, P. D., Benevento, M., Corso, A. J., Clancy, J. L., Mosbergen, R., Li, M., Lee, D.-S., Cloonan, N., Wood, D. L. A., Munoz, J., Middleton, R., Korn, O., Patel, H. R., White, C. A., Shin, J.-Y., Gauthier, M. E., Cao, K.-A. L., … Nagy, A. (2014). Genome-wide characterization of the routes to pluripotency. *Nature*, *516*(7530), Article 7530. https://doi.org/10.1038/nature14046

Ichida, J. K., Blanchard, J., Lam, K., Son, E. Y., Chung, J. E., Egli, D., Loh, K. M., Carter, A. C., Di Giorgio, F. P., Koszka, K., Huangfu, D., Akutsu, H., Liu, D. R., Rubin, L. L., & Eggan, K. (2009). A Small-Molecule Inhibitor of Tgf-β Signaling

Replaces Sox2 in Reprogramming by Inducing Nanog. *Cell Stem Cell*, *5*(5), 491–503. https://doi.org/10.1016/j.stem.2009.09.012

Ichida, J. K., Tcw, J., Williams, L. A., Carter, A. C., Shi, Y., Moura, M. T., Ziller, M., Singh, S., Amabile, G., Bock, C., Umezawa, A., Rubin, L. L., Bradner, J. E., Akutsu, H., Meissner, A., & Eggan, K. (2014). Notch inhibition allows oncogene-independent generation of iPS cells. *Nature Chemical Biology*, *10*(8), Article 8. https://doi.org/10.1038/nchembio.1552

Jaber, M., Radwan, A., Loyfer, N., Abdeen, M., Sebban, S., Khatib, A., Yassen, H., Kolb, T., Zapatka, M., Makedonski, K., Ernst, A., Kaplan, T., & Buganim, Y. (2022). Comparative parallel multi-omics analysis during the induction of pluripotent and trophectoderm states. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31131-8

Jackson, S. A., Olufs, Z. P. G., Tran, K. A., Zaidan, N. Z., & Sridharan, R. (2016). Alternative Routes to Induced Pluripotent Stem Cells Revealed by Reprogramming of the Neural Lineage. *Stem Cell Reports*, *6*(3), 302–311. https://doi.org/10.1016/j.stemcr.2016.01.009

Kim, D. H., Marinov, G. K., Pepke, S., Singer, Z. S., He, P., Williams, B., Schroth, G. P., Elowitz, M. B., & Wold, B. J. (2015). Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell*, *16*(1), 88–101. https://doi.org/10.1016/j.stem.2014.11.005

Knaupp, A. S., Buckberry, S., Pflueger, J., Lim, S. M., Ford, E., Larcombe, M. R., Rossello, F. J., Mendoza, A. de, Alaei, S., Firas, J., Holmes, M. L., Nair, S. S., Clark, S. J., Nefzger, C. M., Lister, R., & Polo, J. M. (2017). Transient and Permanent Reconfiguration of Chromatin and Transcription Factor Occupancy Drive Reprogramming. *Cell Stem Cell*, *21*(6), 834-845.e6. https://doi.org/10.1016/j.stem.2017.11.007

Koche, R. P., Smith, Z. D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B. E., & Meissner, A. (2011). Reprogramming Factor Expression Initiates Widespread Targeted Chromatin Remodeling. *Cell Stem Cell*, *8*(1), 96–105. https://doi.org/10.1016/j.stem.2010.12.001

Li, D., Liu, J., Yang, X., Zhou, C., Guo, J., Wu, C., Qin, Y., Guo, L., He, J., Yu, S., Liu, H., Wang, X., Wu, F., Kuang, J., Hutchins, A. P., Chen, J., & Pei, D. (2017). Chromatin Accessibility Dynamics during iPSC Reprogramming. *Cell Stem Cell*, *21*(6), 819-833.e6. https://doi.org/10.1016/j.stem.2017.10.012

Li, H., Collado, M., Villasante, A., Strati, K., Ortega, S., Cañamero, M., Blasco, M. A., & Serrano, M. (2009). The Ink4/Arf locus is a barrier for iPS cell reprogramming. *Nature*, *460*(7259), Article 7259. https://doi.org/10.1038/nature08290

Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., Qin, B., Xu, J., Li, W., Yang, J., Gan, Y., Qin, D., Feng, S., Song, H., Yang, D., … Pei, D. (2010). A Mesenchymal-to-Epithelial Transition Initiates and Is Required for the Nuclear Reprogramming of Mouse Fibroblasts. *Cell Stem Cell*, *7*(1), 51–63. https://doi.org/10.1016/j.stem.2010.04.014

Liu, X., Ouyang, J. F., Rossello, F. J., Tan, J. P., Davidson, K. C., Valdes, D. S., Schröder, J., Sun, Y. B. Y., Chen, J., Knaupp, A. S., Sun, G., Chy, H. S., Huang, Z., Pflueger, J., Firas, J., Tano, V., Buckberry, S., Paynter, J. M., Larcombe, M. R., … Polo, J. M. (2020). Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature*, *586*(7827), Article 7827. https://doi.org/10.1038/s41586-020-2734-6

Liuyang, S., Wang, G., Wang, Y., He, H., Lyu, Y., Cheng, L., Yang, Z., Guan, J., Fu, Y., Zhu, J., Zhong, X., Sun, S., Li, C., Wang, J., & Deng, H. (2023). Highly efficient and rapid generation of human pluripotent stem cells by chemical reprogramming. *Cell Stem Cell*, *30*(4), 450-459.e9. https://doi.org/10.1016/j.stem.2023.02.008

Lujan, E., Zunder, E. R., Ng, Y. H., Goronzy, I. N., Nolan, G. P., & Wernig, M. (2015). Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature*, *521*(7552), Article 7552. https://doi.org/10.1038/nature14274

Luo, S., Germain, P.-L., Robinson, M. D., & Meyenn, F. von. (2023). *Benchmarking computational methods for single-cell chromatin data analysis* (p. 2023.08.04.552046). bioRxiv. https://doi.org/10.1101/2023.08.04.552046

Maherali, N., & Hochedlinger, K. (2009). Tgfβ Signal Inhibition Cooperates in the Induction of iPSCs and Replaces Sox2 and cMyc. *Current Biology*, *19*(20), 1718–1723. https://doi.org/10.1016/j.cub.2009.08.025

Mao, Q., Wang, L., Goodison, S., & Sun, Y. (2015). Dimensionality Reduction Via Graph Structure Learning. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 765–774. https://doi.org/10.1145/2783258.2783309

Mao, Q., Wang, L., Tsang, I. W., & Sun, Y. (2017). Principal Graph and Structure Learning Based on Reversed Graph Embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2227–2241. https://doi.org/10.1109/TPAMI.2016.2635657

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Francis Stewart, A., Smith, A., & Stunnenberg, H. G. (2012). The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. *Cell*, *149*(3), 590–604. https://doi.org/10.1016/j.cell.2012.03.026

Mikkelsen, T. S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B. E., Jaenisch, R., Lander, E. S., & Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature*, *454*(7200), Article 7200. https://doi.org/10.1038/nature07056

Monfort, A., & Wutz, A. (2013). Breathing-in epigenetic change with vitamin C. *EMBO Reports*, *14*(4), 337–346. https://doi.org/10.1038/embor.2013.29

Naama, M., Rahamim, M., Zayat, V., Sebban, S., Radwan, A., Orzech, D., Lasry, R., Ifrah, A., Jaber, M., Sabag, O., Yassen, H., Khatib, A., Epsztejn-Litman, S., Novoselsky-Persky, M., Makedonski, K., Deri, N., Goldman-Wohl, D., Cedar, H., Yagel, S., … Buganim, Y. (2023). Pluripotency-independent induction of human trophoblast stem cells from fibroblasts. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-39104-1

Nair, S., Ameen, M., Sundaram, L., Pampari, A., Schreiber, J., Balsubramani, A., Wang, Y. X., Burns, D., Blau, H. M., Karakikes, I., Wang, K. C., & Kundaje, A. (2023). *Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency* (p. 2023.10.04.560808). bioRxiv. https://doi.org/10.1101/2023.10.04.560808

Nefzger, C. M., Rossello, F. J., Chen, J., Liu, X., Knaupp, A. S., Firas, J., Paynter, J. M., Pflueger, J., Buckberry, S., Lim, S. M., Williams, B., Alaei, S., Faye-Chauhan, K., Petretto, E., Nilsson, S. K., Lister, R., Ramialison, M., Powell, D. R., Rackham, O. J. L., & Polo, J. M. (2017). Cell Type of Origin Dictates the Route to Pluripotency. *Cell Reports*, *21*(10), 2649–2660. https://doi.org/10.1016/j.celrep.2017.11.029

Okita, K., Ichisaka, T., & Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, *448*(7151), Article 7151. https://doi.org/10.1038/nature05934

O'Malley, J., Skylaki, S., Iwabuchi, K. A., Chantzoura, E., Ruetz, T., Johnsson, A., Tomlinson, S. R., Linnarsson, S., & Kaji, K. (2013). High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature*, *499*(7456), Article 7456. https://doi.org/10.1038/nature12243

Onder, T. T., Kara, N., Cherry, A., Sinha, A. U., Zhu, N., Bernt, K. M., Cahan, P., Mancarci, B. O., Unternaehrer, J., Gupta, P. B., Lander, E. S., Armstrong, S. A., & Daley, G. Q. (2012). Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, *483*(7391), Article 7391. https://doi.org/10.1038/nature10953

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G. A., Stewart, R., & Thomson, J. A. (2007). Whole-Genome Analysis of Histone H3 Lysine 4 and

Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell*, *1*(3), 299–312. https://doi.org/10.1016/j.stem.2007.08.003

Panopoulos, A. D., Yanes, O., Ruiz, S., Kida, Y. S., Diep, D., Tautenhahn, R., Herrerías, A., Batchelder, E. M., Plongthongkum, N., Lutz, M., Berggren, W. T., Zhang, K., Evans, R. M., Siuzdak, G., & Belmonte, J. C. I. (2012). The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. *Cell Research*, *22*(1), Article 1. https://doi.org/10.1038/cr.2011.177

Papp, B., & Plath, K. (2013). Epigenetics of Reprogramming to Induced Pluripotency. *Cell*, *152*(6), 1324–1343. https://doi.org/10.1016/j.cell.2013.02.043

Parenti, A., Halbisen, M. A., Wang, K., Latham, K., & Ralston, A. (2016). OSKM Induce Extraembryonic Endoderm Stem Cells in Parallel to Induced Pluripotent Stem Cells. *Stem Cell Reports*, *6*(4), 447–455. https://doi.org/10.1016/j.stemcr.2016.02.003

Polo, J. M., Anderssen, E., Walsh, R. M., Schwarz, B. A., Nefzger, C. M., Lim, S. M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., Bar-Nur, O., Cheloufi, S., Stadtfeld, M., Figueroa, M. E., Robinton, D., Natesan, S., Melnick, A., Zhu, J., Ramaswamy, S., & Hochedlinger, K. (2012). A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell*, *151*(7), 1617–1632. https://doi.org/10.1016/j.cell.2012.11.039

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., & Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nature Methods*, *14*(3), Article 3. https://doi.org/10.1038/nmeth.4150

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, *14*(10), Article 10. https://doi.org/10.1038/nmeth.4402

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), Article 5. https://doi.org/10.1038/s41587-019-0071-9

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H., Beyer, T. A., Datti, A., Woltjen, K., Nagy, A., & Wrana, J. L. (2010). Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. *Cell Stem Cell*, *7*(1), 64–77. https://doi.org/10.1016/j.stem.2010.04.015

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), Article 5. https://doi.org/10.1038/nbt.3192

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., & Lander, E. S. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, *176*(4), 928-943.e22. https://doi.org/10.1016/j.cell.2019.01.006

Schwarz, B. A., Cetinbas, M., Clement, K., Walsh, R. M., Cheloufi, S., Gu, H., Langkabel, J., Kamiya, A., Schorle, H., Meissner, A., Sadreyev, R. I., & Hochedlinger, K. (2018). Prospective Isolation of Poised iPSC Intermediates Reveals Principles of Cellular Reprogramming. *Cell Stem Cell*, *23*(2), 289-305.e5. https://doi.org/10.1016/j.stem.2018.06.013

Serrano, F., Calatayud, C. F., Blazquez, M., Torres, J., Castell, J. V., & Bort, R. (2013). Gata4 blocks somatic cell reprogramming by directly repressing Nanog. *Stem Cells (Dayton, Ohio)*, *31*(1), 71–82. https://doi.org/10.1002/stem.1272

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., & Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, *34*(6), Article 6. https://doi.org/10.1038/nbt.3569

Shi, Y., Desponts, C., Do, J. T., Hahm, H. S., Schöler, H. R., & Ding, S. (2008). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Oct4 and Klf4 with Small-Molecule Compounds. *Cell Stem Cell*, *3*(5), 568–574. https://doi.org/10.1016/j.stem.2008.10.004

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T. W., & Smith, A. (2008). Promotion of Reprogramming to Ground State Pluripotency by Signal Inhibition. *PLOS Biology*, *6*(10), e253. https://doi.org/10.1371/journal.pbio.0060253

Sim, Y.-J., Kim, M.-S., Nayfeh, A., Yun, Y.-J., Kim, S.-J., Park, K.-T., Kim, C.-H., & Kim, K.-S. (2017). 2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms. *Stem Cell Reports*, *8*(5), 1312–1328. https://doi.org/10.1016/j.stemcr.2017.04.001

Soufi, A., Donahue, G., & Zaret, K. S. (2012). Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell*, *151*(5), 994–1004. https://doi.org/10.1016/j.cell.2012.09.045

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M., Carey, M., Garcia, B. A., & Plath, K. (2013). Proteomic and genomic approaches reveal critical functions of H3K9

methylation and heterochromatin protein-1γ in reprogramming to pluripotency. *Nature Cell Biology*, *15*(7), 872–882. https://doi.org/10.1038/ncb2768

Sridharan, R., Tchieu, J., Mason, M. J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., & Plath, K. (2009). Role of the Murine Reprogramming Factors in the Induction of Pluripotency. *Cell*, *136*(2), 364–377. https://doi.org/10.1016/j.cell.2009.01.001

Stadtfeld, M., Maherali, N., Breault, D. T., & Hochedlinger, K. (2008). Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. *Cell Stem Cell*, *2*(3), 230–240. https://doi.org/10.1016/j.stem.2008.02.001

Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., & Satija, R. (2021). Single-cell chromatin state analysis with Signac. *Nature Methods*, *18*(11), Article 11. https://doi.org/10.1038/s41592-021-01282-5

Tada, M., Takahama, Y., Abe, K., Nakatsuji, N., & Tada, T. (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. *Current Biology*, *11*(19), 1553–1558. https://doi.org/10.1016/S0960-9822(01)00459-6

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell*, *131*(5), 861–872. https://doi.org/10.1016/j.cell.2007.11.019

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676. https://doi.org/10.1016/j.cell.2006.07.024

Tonge, P. D., Corso, A. J., Monetti, C., Hussein, S. M. I., Puri, M. C., Michael, I. P., Li, M., Lee, D.-S., Mar, J. C., Cloonan, N., Wood, D. L., Gauthier, M. E., Korn, O., Clancy, J. L., Preiss, T., Grimmond, S. M., Shin, J.-Y., Seo, J.-S., Wells, C. A., … Nagy, A. (2014). Divergent reprogramming routes lead to alternative stem-cell states. *Nature*, *516*(7530), Article 7530. https://doi.org/10.1038/nature14047

Tran, K. A., Dillingham, C. M., & Sridharan, R. (2019). Coordinated removal of repressive epigenetic modifications during induced reversal of cell identity. *The EMBO Journal*, *38*(22), e101681. https://doi.org/10.15252/embj.2019101681

Tran, K. A., Jackson, S. A., Olufs, Z. P. G., Zaidan, N. Z., Leng, N., Kendziorski, C., Roy, S., & Sridharan, R. (2015). Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. *Nature Communications*, *6*(1), Article 1. https://doi.org/10.1038/ncomms7188

Tran, K. A., Pietrzak, S. J., Zaidan, N. Z., Siahpirani, A. F., McCalla, S. G., Zhou, A. S., Iyer, G., Roy, S., & Sridharan, R. (2019). Defining Reprogramming Checkpoints

from Single-Cell Analyses of Induced Pluripotency. *Cell Reports*, *27*(6), 1726-1741.e5. https://doi.org/10.1016/j.celrep.2019.04.056

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), Article 4. https://doi.org/10.1038/nbt.2859

Varum, S., Rodrigues, A. S., Moura, M. B., Momcilovic, O., Iv, C. A. E., Ramalho-Santos, J., Houten, B. V., & Schatten, G. (2011). Energy Metabolism in Human Pluripotent Stem Cells and Their Differentiated Counterparts. *PLOS ONE*, *6*(6), e20914. https://doi.org/10.1371/journal.pone.0020914

Vidal, S. E., Amlani, B., Chen, T., Tsirigos, A., & Stadtfeld, M. (2014). Combinatorial Modulation of Signaling Pathways Reveals Cell-Type-Specific Requirements for Highly Efficient and Synchronous iPSC Reprogramming. *Stem Cell Reports*, *3*(4), 574–584. https://doi.org/10.1016/j.stemcr.2014.08.003

Wang, B., Wu, L., Li, D., Liu, Y., Guo, J., Li, C., Yao, Y., Wang, Y., Zhao, G., Wang, X., Fu, M., Liu, H., Cao, S., Wu, C., Yu, S., Zhou, C., Qin, Y., Kuang, J., Ming, J., … Pei, D. (2019). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Jdp2-Jhdm1b-Mkk6-Glis1-Nanog-Essrb-Sall4. *Cell Reports*, *27*(12), 3473-3485.e5. https://doi.org/10.1016/j.celrep.2019.05.068

Wang, T., Chen, K., Zeng, X., Yang, J., Wu, Y., Shi, X., Qin, B., Zeng, L., Esteban, M. A., Pan, G., & Pei, D. (2011). The Histone Demethylases Jhdm1a/1b Enhance Somatic Cell Reprogramming in a Vitamin-C-Dependent Manner. *Cell Stem Cell*, *9*(6), 575–587. https://doi.org/10.1016/j.stem.2011.10.005

Welch, J. D., Hartemink, A. J., & Prins, J. F. (2016). SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, *17*(1), 106. https://doi.org/10.1186/s13059-016-0975-3

Wille, C. K., Neumann, E. N., Deshpande, A. J., & Sridharan, R. (2023). *DOT1L interaction partner AF10 controls patterning of H3K79 methylation and RNA polymerase II to maintain cell identity* (p. 2020.12.17.423347). bioRxiv. https://doi.org/10.1101/2020.12.17.423347

Wille, C. K., & Sridharan, R. (2022). DOT1L inhibition enhances pluripotency beyond acquisition of epithelial identity and without immediate suppression of the somatic transcriptome. *Stem Cell Reports*, *17*(2), 384–396. https://doi.org/10.1016/j.stemcr.2021.12.004

Wille, C. K., Zhang, X., Haws, S. A., Denu, J. M., & Sridharan, R. (2023). DOT1L is a barrier to histone acetylation during reprogramming to pluripotency. *Science Advances*, *9*(46), eadf3980. https://doi.org/10.1126/sciadv.adf3980

Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J., & Campbell, K. H. S. (1997). Viable offspring derived from fetal and adult mammalian cells. *Nature*, *385*(6619), Article 6619. https://doi.org/10.1038/385810a0

Xing, Q. R., El Farran, C., Gautam, P., Chuah, Y. S., Warrier, T., Toh, C.-X. D., Kang, N.-Y., Sugii, S., Chang, Y.-T., Xu, J., Collins, J. J., Daley, G. Q., Li, H., Zhang, L.-F., & Loh, Y.-H. (n.d.). Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Science Advances*, *6*(37), eaba1190. https://doi.org/10.1126/sciadv.aba1190

Xing, Q. R., El Farran, C., Gautam, P., Chuah, Y. S., Warrier, T., Toh, C.-X. D., Kang, N.-Y., Sugii, S., Chang, Y.-T., Xu, J., Collins, J. J., Daley, G. Q., Li, H., Zhang, L.-F., & Loh, Y.-H. (2020). Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Science Advances*, *6*(37), eaba1190. https://doi.org/10.1126/sciadv.aba1190

Ye, J., Ge, J., Zhang, X., Cheng, L., Zhang, Z., He, S., Wang, Y., Lin, H., Yang, W., Liu, J., Zhao, Y., & Deng, H. (2016). Pluripotent stem cells induced from mouse neural stem cells and small intestinal epithelial cells by small molecule compounds. *Cell Research*, *26*(1), Article 1. https://doi.org/10.1038/cr.2015.142

Yin, R., Mao, S.-Q., Zhao, B., Chong, Z., Yang, Y., Zhao, C., Zhang, D., Huang, H., Gao, J., Li, Z., Jiao, Y., Li, C., Liu, S., Wu, D., Gu, W., Yang, Y.-G., Xu, G.-L., & Wang, H. (2013). Ascorbic Acid Enhances Tet-Mediated 5-Methylcytosine Oxidation and Promotes DNA Demethylation in Mammals. *Journal of the American Chemical Society*, *135*(28), 10396–10403. https://doi.org/10.1021/ja4028346

Ying, Q.-L., & Smith, A. (2017). The Art of Capturing Pluripotency: Creating the Right Culture. *Stem Cell Reports*, *8*(6), 1457–1464. https://doi.org/10.1016/j.stemcr.2017.05.020

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., & Smith, A. (2008). The ground state of embryonic stem cell self-renewal. *Nature*, *453*(7194), Article 7194. https://doi.org/10.1038/nature06968

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., & Thomson, J. A. (2007). Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. *Science*, *318*(5858), 1917–1920. https://doi.org/10.1126/science.1151526

Zhao, T., Fu, Y., Zhu, J., Liu, Y., Zhang, Q., Yi, Z., Chen, S., Jiao, Z., Xu, X., Xu, J., Duo, S., Bai, Y., Tang, C., Li, C., & Deng, H. (2018). Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical

Reprogramming. *Cell Stem Cell*, *23*(1), 31-45.e7. https://doi.org/10.1016/j.stem.2018.05.025

Zhao, X., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C., Ma, Q., Wang, L., Zeng, F., & Zhou, Q. (2009). iPS cells produce viable mice through tetraploid complementation. *Nature*, *461*(7260), Article 7260. https://doi.org/10.1038/nature08267

Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., Zhu, J., Meng, G., Ge, J., Yang, S., Cheng, L., Du, Y., Zhao, C., Wang, T., Su, L., Yang, W., & Deng, H. (2015). A XEN-like State Bridges Somatic Cells to Pluripotency during Chemical Reprogramming. *Cell*, *163*(7), 1678–1691. https://doi.org/10.1016/j.cell.2015.11.017

Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M., & Nolan, G. P. (2015). A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. *Cell Stem Cell*, *16*(3), 323–337. https://doi.org/10.1016/j.stem.2015.01.015

Chapter 2

Defining reprogramming checkpoints from single-cell analyses of induced pluripotency

The work presented in this chapter is published in Cell Reports:

Tran, K.A.*, Pietrzak, S.J.*, Zaidan, N.Z.*, Siahpirani, A.F., McCalla, S.G., Zhou, A.S., Iyer, G., Roy, S., and Sridharan, R. (2019). Defining reprogramming checkpoints from single-cell analysis of induced pluripotency. *Cell Reports* 27(6):1726–1741. (* = co-first author)

Contributions: K.A.T., S.J.P., and N.Z.Z. performed data collection, analyzed and interpreted the data, and generated figures under the supervision of R.S. who designed the project and wrote the manuscript with S.R. A.F.Z. and S.G.M. performed co-expression and MERLIN analysis under the supervision of S.R. who designed the analysis. G.I. provided essential resources and critical review. A.S.Z. contributed to initial experiments and analysis.

Abstract

Elucidating the mechanism of reprogramming is confounded by heterogeneity due to the low efficiency and differential kinetics of obtaining induced pluripotent stem cells (iPSCs) from somatic cells. Therefore, we increased the efficiency with a combination of epigenomic modifiers and signaling molecules and profiled the transcriptomes of individual reprogramming cells. Contrary to the established temporal order, somatic gene inactivation and upregulation of cell cycle, epithelial, and early pluripotency genes can be triggered independently such that any combination of these events can occur in single cells. Sustained co-expression of Epcam, Nanog, and Sox2 with other genes is required to progress toward iPSCs. Ehf, Phlda2, and translation initiation factor Eif4a1 play functional roles in robust iPSC generation. Using regulatory network analysis, we identify a critical role for signaling inhibition by 2i in repressing somatic expression and synergy between the epigenomic modifiers ascorbic acid and a Dot1L inhibitor for pluripotency gene activation.

Introduction

Somatic cells can be reprogrammed to induced pluripotent stem cells (iPSCs) by the introduction of the transcription factors Oct4, Sox2, Klf4, and c-Myc (OSKM) (Takahashi and Yamanaka, 2006). Mouse iPSCs are functionally equivalent to embryonic stem cells (ESCs) because they pass all the tests of pluripotency, including tetraploid complementation (Zhao et al., 2009). The efficiency of reprogramming remains low at about 5% even when the reprogramming factors are inducibly expressed from a single locus in the mouse genome (Buganim et al., 2013). In addition, iPSC colonies appear at different times during the reprogramming process (Apostolou and Hochedlinger, 2013; Buganim et al., 2013; Papp and Plath, 2013). Identifying only those cells that successfully complete the reprogramming process versus those that fail to do so can reveal key mechanisms that make the reprogramming process inefficient. Although some markers, such as SSEA1, EPCAM, CD73, ICAM1, and CD44, enrich for successfully reprogramming cells (Lujan et al., 2015; O'Malley et al., 2013; Polo et al., 2012), it is not yet possible to prospectively identify only the cells that will become iPSCs to follow them as they reprogram.

Transcriptional profiling of bulk reprogramming populations over time has led to the description of a temporal series of events with early downregulation of somatic cell expression followed by metabolic and cell cycle changes that culminates in the activation of the pluripotency gene regulatory network (Apostolou and Hochedlinger, 2013; Apostolou and Stadtfeld, 2018). Mouse embryonic fibroblasts (MEFs) undergo a mesenchymal-to-epithelial transition (MET) before pluripotency gene activation during reprogramming (Hussein et al., 2014; Li et al., 2010; Mikkelsen et al., 2008;

Samavarchi-Tehrani et al., 2010). Importantly, whether all cells undergoing reprogramming have to trigger these programs in the same temporal order remains unknown. Due to the low efficiency and variable kinetics of obtaining iPSCs, reprogramming cultures will have heterogeneous expression profiles. Therefore, in population-based analyses of unsorted cells, expression signatures from cells that will successfully reprogram are obscured.

To overcome these issues with ensemble profiling, single-cell analysis of candidate factors in reprogramming MEFs has been performed both at the RNA and protein level. These studies have uncovered intermediate markers, a role for Ras-signaling, and a role for Sox2 in the deterministic activation of the pluripotency network. (Buganim et al., 2012; Kim et al., 2015; Lujan et al., 2015; Zunder et al., 2015). More recent experiments have focused on profiling cells during reprogramming in low-efficiency systems, including non-transgenic chemical reprogramming (Zhao et al., 2018; Guo et al., 2019; Schiebinger et al., 2019).

Reprogramming efficiency can be increased by the modulation of regulators that decrease chromatin compaction and those that perturb signaling pathways (Esteban et al., 2010; Huangfu et al., 2008; Ichida et al., 2009; 2014; Maherali and Hochedlinger, 2009; Mikkelsen et al., 2008; Onder et al., 2012; Shi et al., 2008; Silva et al., 2008; Tran et al., 2015). We and others have combined such epigenomic and signaling modulators and found that they synergistically increase reprogramming efficiency from OSKM-expressing cells (Bar-Nur et al., 2014; Tran et al., 2015; Vidal et al., 2014). In this study, we added SGC0946 (inhibitor of Dot1L, a histone H3K79 methyltransferase) along with our previous cocktail of ascorbic acid (vitamin C) and 2i

(inhibitors to mitogen-activated protein [MAP] kinase and glycogen synthetase kinase), in conjunction with OSKM to reprogram MEFs to iPSCs at an efficiency of ~40% within 6 days. Although each small molecule has been used previously, to our knowledge this particular combination (called A2S [ascorbic acid, 2i, SGC] henceforth) has not been reported.

Using single-cell RNA sequencing (RNA-seq) analysis, we profiled reprogramming MEFs along a time course in both a regular serum-containing (fetal bovine serum [FBS]) and the A2S system. We found that early events, such as epithelial and cell cycle activation, are turned on independently. Surprisingly, all mesenchymal genes are not downregulated together in the same cells, and some genes, such as Twist1, can even be found expressed with early pluripotency marker Nanog. A large majority of the cells in FBS stop cycling partly due to senescence, which can be overcome by the addition of A2S. Nanog, Oct4, and even Sox2 could be activated in individual cells, but what distinguished successful reprogramming was the detectable co-expression of these genes in different modules. Nanog was found in a sub-cluster with Epcam, Sall4, and Tdgf1; Oct4 with Zfp42; and Sox2 with Utf1 and Dppa5a. The lack of detectable expression of some markers, such as Epcam, with other pluripotency genes correlated with cells reverting to an Epcam-negative state. Functional experiments provide a role for reprogramming-specific transient upregulation of transcription factors, such as Ehf; translation initiation (Eif4a1); and factors such as Phlda2 for reaching an iPSC state. By applying a network-based analytical framework to our single-cell data, we studied the effect of individual components of A2S on the acquisition of pluripotency. Our analysis identified that

specific connections of the pluripotency network can only be made when both epigenomic modifiers are present, but without the suppression of somatic expression by the signaling inhibitors reprogramming efficiency is compromised. Thus, we have uncovered that reprogramming need not progress in discrete stages but instead is the result of co-occurring modulation of various networks.

Results

Combining epigenomic and signaling modifiers leads to high-efficiency generation of bona fide iPSCs

We reprogrammed MEFs that have a doxycycline (dox)-inducible cassette containing a transgene with four reprogramming factors: Oct4, Sox2, Klf4, and c-Myc (OSKM). iPSC generation was monitored by immunofluorescence for NANOG at various time points. The NANOG+ colonies that remained after dox withdrawal are transgene-independent iPSCs (Brambrink et al., 2008; Stadtfeld et al., 2008). In FBS conditions, NANOG+ colonies emerged by day 6 and most were transgene independent by day 12 of reprogramming, yielding an efficiency of about 3.2% (Figure 1A; STAR Methods).

As very few cells successfully reprogram in FBS, we next sought to increase reprogramming efficiency to elucidate the transcriptional changes required for pluripotency acquisition. We have previously shown that the addition of ascorbic acid (AA) and 2i increases reprogramming efficiency of both embryonic and adult fibroblasts (Tran et al., 2015). A small-molecule screen of chemicals (data not shown) revealed that the addition of an inhibitor to the H3K79 methyltransferase Dot1L called

SGC0946 (Jackson et al., 2016) to the AA+2i combination boosted iPSC generation from reprogrammable MEFs. By day 6, ~1,900 Nanog+ iPSC colonies were obtained at an efficiency of ~42% (STAR Methods) (Figure 1B). Beyond this time point, the colonies started merging with each other, and therefore, it was chosen as the endpoint for analysis. The A2S system also increased the kinetics of reprogramming because the NANOG+ colonies on day 4 were already transgene independent (Figure 1B) as compared to day 9 of FBS reprogramming (Figure 1A). To avoid biases from plating efficiencies (Schwarz et al., 2018), we further verified the efficiency by reprogramming MEFs as single cells. We found that transgene-independent colonies were obtained in ~40% of the wells in the A2S system (Figure 1C). Thus, the A2S combination of small molecules yielded a great increase in reprogramming efficiency and kinetics.

To determine whether iPSCs generated from the A2S system were bona fide, colonies were picked on day 6 from an A2S reprogramming experiment and could be passaged in FBS without loss of pluripotency. These iPSCs were karyotypically normal and produced teratomas that were comprised of cells from all three germ layers (Figures S1A and S1B).

Single-cell RNA-seq time course confirms heterogeneity of reprogramming populations

To dissect the intrinsic heterogeneity during FBS reprogramming and determine whether the A2S system accelerated or overcame the FBS reprogramming barriers, we performed single-cell transcriptomics. We profiled reprogramming cells in FBS on days

3, 6, 9, and 12; A2S on days 2, 4, and 6; as well as the starting population of MEFs and endpoint of ESCs using a microfluidics-based droplet digital sequencing system (Bio-Rad ddSeq, STAR Methods). In addition, iPSCs that were generated from the A2S system were profiled to determine their similarity to ESCs. Because AA and 2i are known to change the expression profile of ESCs (Blaschke et al., 2013; Marks et al., 2012), we also sequenced ESCs and iPSCs that had been passaged in A2S.

We obtained an average of about 55,000 reads and 13,000 uniquely identified transcripts per cell, which corresponded to a total 18,005 genes detected across all cells (Figure S1C; STAR Methods). We used the Monocle2 program (Qiu et al., 2017a, 2017b) (Figures S1D and S1E) to analyze the gene expression data and identified gene regulatory networks using the MERLIN algorithm (Chasman et al., 2016) to provide insights into the different factors that influence reprogramming efficiency. A t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis (STAR Methods) revealed the iPSCs derived from A2S when passaged in FBS clustered with ESCs grown in FBS and away from ESCs or iPSCs passaged in A2S (Figure 1D). This result further confirmed that the iPSCs had reached an ESC-like transcriptional state. As expected, ESCs cultured in A2S expressed blastocyst-enriched genes, such as Dazl, while also repressing the development-associated gene Emb and showed more homogeneous expression of naive marker Tbx3 but not Rex1 (Figure S1F).

A2S accelerates FBS reprogramming

The cells profiled from the time course analysis were grouped into 14 clusters (Figure 2A). The starting MEFs were heterogenous and occupied two clusters (cluster

2 and cluster 7) (Figure 2A). For the FBS samples, the cells on day 3 occupied a single cluster (77% of cluster 3) away from days 6, 9, and 12 reprogramming cells (Figure 2A). Similarly, the day 2 of A2S samples predominated a single cluster (92% of cluster 5), whereas the cells from day 4 and day 6 belonged to several clusters (Figure 2B; Figure S2A). Therefore, at the beginning of reprogramming, the cells are more homogeneous than later time points, irrespective of the efficiency of the system. The fact that cells from different time points cluster together based on similarity in gene expression profiles suggests that average expression from previous time-point-based analysis warrants analysis by single-cell sequencing. A small fraction of cells from A2S were found in the FBS clusters and vice versa (Figure 2B). The entire reprogramming population also clustered away from ESCs and iPSCs grown in A2S (Figure S2B). Distance in the t-SNE does not necessarily reflect the most differential gene clusters. However, given that the cells in reprogramming cultures were most similar in gene expression profile to pluripotent cells grown in serum, ESCs grown in FBS were used as the endpoint for all subsequent analyses.

From previous bulk RNA-seq and mass cytometry analysis, various cell surface markers have been identified that enrich for reprogramming cells that will transition to iPSCs (Lujan et al., 2015; Nefzger et al., 2017; O'Malley et al., 2013; Polo et al., 2012), although the same markers can have heterogeneous expression in ESCs (O'Malley et al., 2013). We reasoned that if A2S reprogramming was an accelerated version of FBS reprogramming, the same markers would be found in a greater proportion. The marker CD44 is high in MEFs, whereas ICAM1 is transiently increased in reprogramming cells (O'Malley et al., 2013). The CD44-/ICAM1+

population was two-fold greater in A2S by day 6 than FBS on day 12 (Figure S3B). Similarly, the transient CD73 intermediate marker (Lujan et al., 2015) was rapidly acquired and downregulated (Figure S3A). There was a greater decrease in the MEF-specific Thy1+ or Vcam+ cells in A2S as compared to FBS reprogramming (Polo et al., 2012; Schwarz et al., 2018) (Figure S3A). The Thy1-/Fut9+ (SSEA1) (Polo et al., 2012) and the Epcam+/Sca1-/Fut9+ (Schwarz et al., 2018) populations that are more predictive of cells that will complete reprogramming were both ~4-fold higher in A2S by day 6 as compared to FBS (Figure S3B). Notably, the gene expression of Mbd3 and Gatad2a were not affected in A2S reprogramming (Figure S3A). The absence of these proteins leads to high-efficiency reprogramming (Mor et al., 2018; Rais et al., 2013). Taken together, these results indicate that A2S improves the kinetics and efficiency of the route taken by FBS reprogramming cells.

To identify the genes that distinguished the clustering of single cells in the Monocle t-SNE analysis (Figure 2A), we examined the top 10% of differentially expressed genes between all the clusters. Because this is single-cell data, we measured both the percentage of cells displaying each of the four major patterns of expression between MEFs and ESCs as well as the average expression (Figure 2C; Figure S2C; Table S1). There was a net decrease in expression (groups A–D), which included genes in categories such as cell differentiation and migration; a reprogramming-related decrease (groups F–H), mainly composed of cell cycle, DNA replication, and spliceosome-related genes; a reprogramming-related increase (groups K–L); and a net increase from MEFs to ESCs (groups M–N), which included pluripotency genes. We also observed a fifth pattern (group O), which was made of

ribosomal genes that displayed tremendous cell-cell variability but was expressed in all cells.

Mesenchymal and epithelial changes are independently regulated

From bulk sequencing experiments, it is thought that downregulation of somatic cell gene expression, including the mesenchymal genes, are early events in reprogramming (Apostolou and Hochedlinger, 2013; Apostolou and Stadtfeld, 2018; Li et al., 2010; Samavarchi-Tehrani et al., 2010). We found that not all mesenchymal genes are rapidly decreased in all cells. The majority of the cells in group A (Figure 2C) decreased expression of developmental signaling and cell migration genes, including Tgfb3, Snai1, and Twist2 (Figure 3A). Larger fractions of cells retained expression of Id1 and Id2, and the mesenchymal factors Zeb1 and Zeb2 (group B). Expression of several collagens, Egr1 and Twist 1 (group C), was retained in an even higher proportion of cells than group B (Figure 3A). Thus, there are three different trends for populations to lose mesenchymal gene expression with a large majority of cells in FBS reprogramming still retaining MEF-like gene expression even at later time points. The mesenchymal MEFs have to transition to an epithelial state indicated by the upregulation of E-cadherin (Cdh1) (Apostolou and Hochedlinger, 2013; Apostolou and Stadtfeld, 2018; Li et al., 2010; Samavarchi-Tehrani et al., 2010). Given the differential proportion of mesenchymal genes that were turned off in individual cells, we determined the co-expression of Cdh1 with several mesenchymal genes. It should be noted that because of the limit of detection of single-cell transcriptomics, such analysis may underestimate the number of co-expressing cells.

Surprisingly, Cdh1 upregulation was compatible with the expression of mesenchymal genes, albeit in different proportions, as well as the somatic marker Thy1 (Figure 3B). Instead, from our data, it is apparent that the mesenchymal gene downregulation and E-cadherin upregulation operate as different modules. For example, the downregulation of Snai1 does not automatically lead to Cdh1 expression. We orthogonally confirmed the RNA-seq results by performing immunofluorescence for Twist1 and Cdh1 and found an overlap of both markers in the proportion predicted by the transcriptional data (Figure 3C). The trends of dual mesenchymal gene+/Cdh1+ cells were similar in A2S and FBS reprogramming (Figure S2D).

By performing a pairwise comparison between the earliest time points of the FBS and A2S time course (cluster 3 versus cluster 5; Figure 2C), we found that FBS cells on day 3 still retained the expression of genes associated with system development (Col3a1) as well as signal transduction (Fgf7, Egr1, and Igfbp3) that were greatly reduced by day 2 of A2S reprogramming. Thus, the acceleration of reprogramming in A2S is partially derived from increasing the rate of downregulation of somatic genes.

Reprogramming-specific transient gene expression patterns are important for conversion to iPSCs

Because iPSCs self-renew indefinitely, mechanisms that confer an ESC-like cell cycle improve reprogramming efficiency (Hanna et al., 2009; Mario´ n et al., 2009; Ruiz et al., 2011; Utikal et al., 2009). The starting population of MEFs heterogeneously expressed cell cycle markers to segregate into two different clusters (cluster 2 and 7; Figure 2C). Interestingly, both FBS day 3 and A2S day 2 reprogramming cells also

expressed cell cycle genes, such as Mcm6, Bub1b, and Ccnb1 (groups F–H; Figure 2C; Table S1). Therefore, either the induction of the reprogramming factors upregulated these genes in the majority of MEFs or reprogramming was productively initiated only from those MEFs that were already cycling. The initial upregulation of cell cycle observed in bulk transcriptomic data may represent the selection of cycling MEFs (Mikkelsen et al., 2008) for reprogramming rather than a true upregulation in all cells.

After this time point, there was a dramatic difference in the way the two systems behaved. In the FBS clusters, the vast majority of the cells (76% of all FBS cells) downregulated cell cycle genes (clusters 4, 11, 13, and 14), whereas a minority retained expression (cluster 10) (Figure 3D). In contrast, in the A2S system, the vast majority of the cells still retain the expression of cell cycle genes and a small fraction (21% of all A2S cells, located within cluster 8) shut these genes off (Figure 3D). This result was corroborated by immunofluorescence for the cell cycle marker Ki67 with a rapid decline by day 6 of FBS reprogramming, which was not observed in A2S cells (Figure 3E).

Cell cycle gene expression upregulation was compatible with Thy1, Zeb2, and Twist1 expression, as well as Cdh1 in both FBS and A2S systems (Figure 3G; Figure S2D). This result suggests that the cell cycle can also be activated with continued somatic expression.

It is known that in FBS, most reprogramming cells experience reprogramming-induced senescence (Banito et al., 2009; Li et al., 2009; Mikkelsen et al., 2008). Corroborating this notion, the antiproliferative Cdkn1c gene was highly upregulated in FBS reprogramming cells but not in the A2S system (Figure 3E). By contrast, p53 transcription levels were maintained in the entire population. Senescence-associated

genes, such as Ink4a, were also activated during A2S reprogramming and interleukin-6 (IL-6) remained inactivated (Figure S3A). Thus, the senescence block may be overcome by the lack of activation of Cdkn1c (Figure 3E). In this aspect, the A2S system in MEFs resembles a cohort of fast-cycling granulocytes—monocyte precursors that undergo non-stochastic reprogramming due to reduced levels of Cdkn1c (Guo et al., 2014).

Besides senescent genes, this third pattern of reprogramming-related upregulation (groups K and L; Figure 2C) was without a specific gene ontology. Because cell fate transitions are often orchestrated by transcription factors, chromatin-modifying proteins, or signaling molecules, we knocked down three genes belonging to these categories—Ano1, Aldh3a1, and Ehf—during reprogramming. Among these genes, the knock down of Ehf caused a decrease in A2S reprogramming efficiency (Figure 3H; Figure S2E). This suggests that transient upregulation of some genes is, in fact, required for reprogramming to iPSCs and does not represent a different lineage-specific endpoint.

Co-expression of core pluripotency factors are independent of each other

The activation of genes highly expressed in ESCs (groups M and N; Figure 2C) was largely restricted to reprogramming clusters C9, C6, and C10 that already expressed cell cycle genes (Figure S4A). We examined the expression of known pluripotency genes within this group. Epcam, Sall1, and Gdf3 were expressed in reprogramming clusters other than the ones with the most ESC-like characteristics (Figure 4A). This suggests that they can be activated in isolated cells and may not

predict cells completing the transition to iPSCs. Surprisingly, Sox2 was also expressed in cells other than the ones most similar to cluster 1, suggesting that its activation may not be sufficient to activate a cascade of deterministic pluripotency gene activation as previously suggested (Buganim et al., 2012) (Figure 4A). We next determined which genes were most prevalently expressed with the core pluripotency factors Oct4, Sox2, and Nanog in the reprogramming populations, while acknowledging the caveat that such analysis may be limited by the detection limit of single-cell transcriptional sequencing. Nanog was detected with Sall4, Epcam, and Tdgf1 (Cripto) (Figure 4B). Within the population of Nanog-expressing cells, Sall4 was equally expressed in both cluster 6 and cluster 9 (Figure 4B). However, Tdgf1 expression was higher in cluster 9 cells, suggesting that Tdgf1 may be more important for activating the rest of this subset (Figure 4B). On the other hand, although Oct4 was activated with Zfp42 (Figure S4B), Sox2 was found with Dppa5a and Utf1 and was part of a larger cluster that included Tet1 and Zscan10 (Figure 4C). In cluster 10 that is predominantly made of cells from FBS reprogramming, this larger subset is heterogeneously activated. In contrast, in cluster 6 that mostly contains A2S reprogramming, the whole group was coordinately upregulated (Figure 4C).

The most restricted pattern of expression included Dppa4, which is known to be a marker of the ‘‘stabilization’’ phase of reprogramming that occurs after the core pluripotency genes are activated (Golipour et al., 2012). Dppa4 was detected in the subset with Lin28a and Phlda2, a gene involved in placental growth (Salas et al., 2004) (Figure 4D).

Intrigued by this finding, we depleted the levels of Phlda2 during reprogramming.

Interestingly, although the number of NANOG-expressing colonies remained similar between Phlda2 knock down and control, we found a 25% decrease in the number of DPPA4-positive colonies (Figure 4E). Therefore, the co-expression of pluripotency factors within each subgroup may functionally predict regulators of transitions to the next stage toward pluripotency.

Similar to downregulation of MEF genes and activation of cell cycle, pluripotency gene activation is increased in a greater proportion of cells, to a higher extent and more homogenously with co-expression partners in A2S as compared to FBS reprogramming.

Continued mesenchymal expression is a roadblock to high-efficiency reprogramming

From these analyses, it is clear that A2S is more efficient than FBS reprogramming in accelerating each of the four major patterns of expression. Therefore, examining the A2S system alone would help us identify genes that are bottlenecks to the completion of reprogramming in cells that are much further along the process. In fact, when we compared the differentially expressed genes that were only related to reprogramming in FBS or A2S alone, we found about 33% unique to the A2S system (Figure S5A). The ones that were solely found in A2S reprogramming were enriched for gene ontology terms, such as system development and cell differentiation, and included pluripotency genes, such as Nanog and Oct4. In contrast, the FBS-exclusive gene expression was dominated by cell cycle genes (Figure S5A). Therefore, we further examined the A2S cells by performing a trajectory analysis in which cells are arranged in pseudotime according to similarity in gene expression patterns (Trapnell et al., 2014)

(Figure 5A). As expected, a larger fraction of day 6 (63%) cells were found in the part of the trajectory toward pluripotent cells than those that were found before the branchpoint.

We performed branched expression analysis modeling (BEAM) (Qiu et al., 2017a) to identify the genes that were over-represented in cells that continued along the trajectory toward ESCs from the ones that were found in the branch. We note that ontogenically MEFs cannot convert to ESCs but use the trajectory to determine a path toward pluripotency. At the early branchpoint, the cells that continue toward ESCs have a higher expression of epithelial genes, such as Cdh1 and Epcam (Figure 5B). At the later branchpoint, cells that continue have already activated the cell cycle and present high levels of Nanog as expected (Figures 5B and 5C). Surprisingly, the mesenchymal gene Twist1 was found to be a gene that influences the branchpoint decision even at this late point in the pseudotime trajectory (Figure 5C) and was even found to be co-expressed with Nanog. Although Nanog levels were similar in cells at the beginning of branch 2, cells that stall have a higher level of Twist1 co-expression than those that continue (Figure 5D).

From population-based studies, cells that express Epcam during intermediate phases of reprogramming have a greater probability of completing the process (Polo et al., 2012). In the branchpoint analysis, several cells that exit the trajectory express high levels of Epcam but at the end of the branch have decreased expression rather than maintained levels (Figure 5C). Given that Epcam is found co-expressed with a subset of genes (Figure 4B), we wondered whether the expression of Epcam was influenced by expression levels of other genes within its subset. In fact, we found that Epcam+ cells that continue along to complete pluripotency co-expressed higher levels

of Nanog, Tdgf1, and Sall4 than those that stall at the branchpoint (Figure 5D). This result suggests that activation of all the genes within a subset is important to sustain initial expression. Because single-cell analysis destroys the cell, the cells at the end of the branchpoint could represent those that never expressed Epcam and are at the end of the trajectory due to covariance with other genes. Therefore, we sorted cells based on the level of EPCAM expression on day 3 of A2S reprogramming (Figure 5E). After allowing reprogramming to continue for an additional 3 days, we found that 7.5% of the high and 16.6% of medium-expressing EPCAM cells gave rise to an EPCAM-negative population (Figure 5E). Taken together, these analyses suggest that without co-expression of other genes within the subset, cells may revert to an Epcam-negative state, whereas with co-expression, cells persist along the trajectory toward an ESC-like state.

A reverse pattern to Epcam is observed for the branchpoint gene, translation initiation factor Eif4a1. Here, after an initial downregulation, cells that successfully remain on the trajectory upregulate gene expression (Figure 5C). Eif4a1 is a part of the translation initiation complex along with the closely related protein Eif4a2 (Modelska et al., 2015; Williams-Hill et al., 1997). To determine if Eif4a1 had a causal role in obtaining iPSCs, we depleted its levels using RNA interference during A2S reprogramming. Interestingly, depletion of Eif4a1 severely compromised the efficiency of reprogramming (Figure 5F). This decrease was not due to a change in the number of cells or increasing cell death (Figure 5F). Taken together, these data suggest that sustained expression of genes is affected by co-expression of other factors and is required for completing the process to a productive pluripotent state.

A2S concurrently enhances downregulation of MEF genes and upregulation of ESC genes

The chemicals we used for high-efficiency reprogramming include signaling inhibitors and two epigenomic modulators— AA, which is thought to regenerate 2-oxoglutarate-dependent chromatin-modifying enzymes (Hore et al., 2016), and SGC0946, an inhibitor of Dot1L-mediated histone H3K79 methylation (Jackson et al., 2016). To understand the relative contribution of each component, we subjected MEFs to every dual combination of chemicals and assessed reprogramming efficiency on day 6. We found that SGC+2i (S2) yielded approximately half the NANOG+ colonies of the A2S combination, whereas AA+2i (A2) and AA+SGC (AS) were only 6.6% and 10.4% efficient, respectively, on day 6 of reprogramming (Figure 6A). Irrespective of the dual combinations that were used, the iPSC colonies remained NANOG+ after dox withdrawal. Exposure to each individual component had lower effects on enhancing reprogramming efficiency (data not shown).

We performed single-cell RNA-seq on reprogramming MEFs that had been subjected to each dual combination on day 4 and day 6 and compared the profiles to FBS and A2S reprogramming. Because none of the dual combinations were able to achieve the high efficiency of the A2S system, we hypothesized that each dual combination likely rewires some components of the gene regulatory network controlling the transcriptional dynamics of reprogramming. Therefore, we first reconstructed the putative regulatory network by using the FBS+A2S single cell RNA-seq (scRNA-seq) dataset collected in this study (STAR Methods) using an expression-based network

inference algorithm, MERLIN (Chasman et al., 2016). We focused on the ~1,800 genes used to initially differentiate the Monocle clusters in the FBS+A2S dataset (Figure 2A) along with sufficiently expressed regulators, such as transcription factors, chromatin remodelers, and signaling proteins (Figure 6B; STAR Methods). MERLIN is based on a probabilistic framework that predicts the regulators of a target gene based on the ability of the regulator's mRNA levels to explain the variation in a target gene's expression level. Using probabilistic modeling, MERLIN allows regulators to control target genes with similar expression levels to have non-identical regulatory programs. Furthermore, target genes are grouped into modules based on their co-expression and shared regulatory program (STAR Methods). Thus, there are two outputs of MERLIN: (1) modules that represent characteristic patterns of expression of genes and (2) networks that specify the regulators of individual genes as well as modules. The MERLIN analysis produced 15 modules with 5 or more genes. There were 4,962 interactions between 1,009 regulators and 1,628 target genes at a stringent confidence of 0.8 or higher (STAR Methods). The regulatory network captures known connections among the key pluripotency regulators and target genes (e.g. Esrrb → Klf4, Sox2 → Klf4, Esrrb ← → Sox2, Esrrb ← → Nanog) and is comparable to the performance seen when using bulk RNA-seq data (STAR Methods), providing support to the relevance of the interactions.

MERLIN modules recapitulated the four patterns of expression from MEFs to ESCs (Figure 6B). We compared the expression patterns of genes in these modules in cells treated with A2S and each dual combination to identify key similarities and differences in expression pattern across these treatments to enable us to define the requirement of each component for successful reprogramming. We found that

compared to A2S, the AS combination that omitted 2i continued to have a high expression of modules M1 through M4, (Figure 6B) which included MEF-specific genes, such as Col5a1 and Tagln, even on day 6 (Figure 6C). This trend was even more obvious for genes that are aberrantly upregulated in the early days of reprogramming (module M8) and included genes such as Oasl2 and Egr1 (Figure 6C). For the cell cycle genes that are transiently downregulated in FBS reprogramming (modules M5 through M7), every dual combination could activate these genes (e.g., Mcm6 and Ccnb1) (Figure 6C). The A2 combination was compromised in activating pluripotency genes (modules M9 through M11). Contrary to earlier reports, Dot1L inhibition does not increase Cdh1 levels (Figure 6C) any more than the combinations that do not include this small molecule (Onder et al., 2012). Interestingly, the AS combination was as good at activating several genes of the pluripotency cluster as S2 but still resulted in a smaller number of iPSC colonies (Figure 6A), likely due to the continued expression of somatic genes because of the failure to downregulate the MEF program. However, neither AS nor S2 was as good as A2S at activating pluripotency, suggesting synergistic effects of the triple combination.

We next used the high-confidence inferred regulatory network as a scaffold to estimate the relative strengths of the regulatory connections in each condition in order to identify which components of the network were present in each of the combinations (STAR Methods). Briefly, we used this network structure to fit a regression model for each gene in each condition and used the regression weight to estimate the edge strength (STAR Methods). The regression weight is reflective of the strength of the regulatory connection between a regulator and a target gene and provides information

that might not be obvious from the absolute level of expression of a gene. Hence, although a gene node could be less expressed in one condition, its connections with regulators can be stronger if the expression of its regulators can explain its expression variation. We found that there were several sub-networks that had different strength in the dual combinations compared to the A2S combination. For the modules that do not turn off somatic genes or transiently upregulated gene expression, the connections between the regulators Oas2l and Trim30, or between Col5a1 and Col1a2 were retained only in the AS condition (lacking 2i) (Figure 6D; Figures S5B and S5C). For the upregulated genes, several connections surrounding Nanog (Figure 6E) were absent in the A2 condition, whereas those around Epcam and Cdh1 were maintained (Figure S5E). For the more restricted pluripotency genes, S2 and AS differ in the kinds of connections that were made; for example, Pou5f1 was better correlated with Dppa3 in S2, whereas a greater proportion of cells expressed Esrrb with Tdh in the AS condition (Figure 6E). In the A2S condition, all these connections are stronger and new connections, such as the ones between Dppa5a, Klf2, and Dppa3, emerge (Figure 6E). The network surrounding DNA replication genes, such as Mcm6, remains strong in any of the dual combinations (Figure S5D).

Taken together, these results indicate that any combination of small molecules is able to overcome the senescence block faced by cells in FBS reprogramming. 2i is required for the downregulation of both MEF genes and transiently upregulated genes. Although A2 is sufficient to activate epithelial genes, SGC is required for the activation of pluripotency genes that emerge late. However, only in the presence of both AA and SGC, the rewiring of the pluripotency network is complete.

Discussion

Reprogramming of somatic cells to iPSCs has been studied using bulk sequencing of reprogramming populations as well as those sorted on the basis of cell surface markers (Apostolou and Hochedlinger, 2013; Hussein et al., 2014; Lujan et al., 2015; Mikkelsen et al., 2008; O'Malley et al., 2013; Polo et al., 2012). These studies have led to an understanding of reprogramming trajectories taken by the majority of the cells. Here, by applying single-cell transcriptional sequencing, we find that there is overlapping expression of genes that was thought to be temporally activated (Apostolou and Stadtfeld, 2019; Brambrink et al., 2008; Stadtfeld and Hochedlinger, 2010) (Figure 7). Because most studies have focused on MEFs as the starting cell type, an important early event is the MET, a process amenable to acceleration (Liang et al., 2012; Zhou et al., 2017). Surprisingly, here we find that mesenchymal genes are not all downregulated at the same stage. The frequently used marker of the epithelial transition Cdh1 can be upregulated in cells that continue to express mesenchymal genes, such as Twist1. Thus, our study demonstrates that in order to increase the rate of reprogramming, it may be worthwhile to focus on other small molecules that can reliably and consistently shut down mesenchymal gene expression. We also find that another epithelial gene, Epcam, can be downregulated in a few cells if it is not co-expressed with other pluripotency genes. This result mirrors the recent finding that the reliability of Epcam as a marker is enhanced by co-expression with SSEA1 and without Sca1 (Schwarz et al., 2018). Such co-expression is valuable for sustaining the expression not only of Epcam but also of the pluripotency factors, which can be activated in isolated cells even in FBS reprogramming. This includes Sox2, which was

identified by candidate sequencing to start a cascade of deterministic pluripotency (Buganim et al., 2012). We find that the level of Sox2 expression is higher when found in cells also expressing Dppa5a and Utf1.

It has also been noted that somatic cell nuclear transfer tends to activate the Oct4 locus earlier than has ever been observed for reprogramming (Bhutani et al., 2010). One reason for this may be that genes such as Ehf that are transiently upregulated may have a role in restructuring the gene networks in a way that makes the next step conducive to reach the pluripotent state. Co-opting basic translational machinery (Brumbaugh et al., 2018), such as the regulation of Eif4a1, a device used by cancer cells (Modelska et al., 2015; Wolfe et al., 2014), may also be important for reprogramming, increasing the parallels between cancer and pluripotency.

The small molecules that we have used contribute differentially to the pluripotency network. One way that any combination of the small molecules works is by decreasing the number of cells that display senescence gene expression. A greater number of cycling cells increases reprogramming efficiency (Hanna et al., 2009; Mario´ n et al., 2009; Ruiz et al., 2011; Utikal et al., 2009). Previous studies have genetically modulated the levels of cell cycle control genes, such as p53, to affect this change (Hanna et al., 2009; Mario´ n et al., 2009; Utikal et al., 2009). We now provide a chemical method that can be transiently applied to overcome the senescence barrier. By applying a network analysis method, we also identify the connections of these molecules. We find that the addition of 2i suppresses some aberrantly expressed genes and allows for faster downregulation of MEF markers. AA and SGC work together to reinforce the pluripotency program. The modulation of the dose and timing

of these factors could be harnessed in the future to rationally enhance reprogramming efficiency further.

Acknowledgments

This work was supported by NIH-NIGMS R01GM113033 and the Shaw Scientist Award to R.S. and NIH-NIGMS R01GM117339 to S.R. K.A.T. was supported by the Advanced Opportunity Fellowship from UW-Madison and NSF GRFP-DGE 1256259, S.J.P. by NHGRI 5T32HG002760, N.Z.Z. by the SCRMC fellowship from UW-Madison and American Heart Association 18PRE34080337, S.G.M.C. by the Data Sciences Initiative of UW-Madison and NIGMS T32GM007133, and A.S.Z. by NIGMS T32GM008688. We thank Mike Ducat of Biorad and Alasdair Reid and Melissa Pourpak of Illumina for providing the ddSeq instrument and experimental support; Molly Zeller and Joshua Hyman of the UW-Madison Biotechnology Center for assisting with sequencing and Profs. James Thomson and Krishanu Saha and members of the Sridharan lab for critical reading of the manuscript.

Declaration of Interests

The authors declare that they have no competing interests.

Materials and Methods

Experimental Model and Subject Details

Primary MEFs

Male and female MEFs were isolated from E13.5 time-mated embryos as described in

Tran et al., (2015) from reprogrammable mice (Sridharan et al., 2013) homozygous for the Oct4-2A-Klf4-2A-IRES-Sox2-2A-c-Myc (OKSM) transgene at the Col1a1 locus and either homozygous or heterozygous for the reverse tetracycline transactivator (rtTA) allele at the Rosa26 locus. MEFs were maintained in MEF media (DMEM, 10% FBS, L-glutamine, Pen/Strep, NEAA, 2-mercaptoethanol). Mice were maintained according to protocol approved by the UW-Madison IACUC.

Mouse Embryonic Stem Cells

Murine ESCs (V6.5 line, male) were maintained in ESC media (knockout DMEM, 15% FBS, L-glutamine, Pen/Strep, NEAA, 2-mercaptoethanol, and leukemia inhibitory factor) on a feeder layer of irradiated MEFs.

Method Details

Reprogramming

MEFs were thawed and maintained in ESC media for 2 days before plating. On day -1, 5000 cells were plated onto 0.1% gelatin-coated coverslips in 6-well plates. 24 hours post-plating (day 0), cells were counted to determine the number of cells adhered to the coverslip. This number was used to calculate reprogramming efficiency (Figure 1A). On day 0, MEFs were treated with 2 mg/mL doxycycline to induce OKSM expression and irradiated MEFs were added. For A2S and dual combination reprogramming, 50 mg/mL of ascorbic acid (Sigma A8960) and 5 mM SGC0946 (ApexBio A4167) were added on Day 0. 3 mM CHIR-99021 (Stemgent 04-0004-10) and 1 mM PD-0325901 (Stemgent 04-0006-10) (2i) were added 12 hours

post-doxycycline induction. Media containing doxycycline and small molecules was changed every two days. Efficiency of reprogramming was determined by Nanog immunofluorescence either on day of fixing as indicated, or after withdrawal of doxycycline and small molecules for an additional 4 days. Two or more biological replicates were performed for each set of reprogramming experiments. iPSC colonies were isolated from reprogramming culture on day 6 and maintained in either regular ESC media or A2S-containing ESC media on irradiated MEFs for several passages. For single-cell reprogramming, MEFs were infected with pMX-tdTomato retrovirus and FACS-sorted into 96-well plates as single tdTomato+ cell per well on irradiated MEFs. FBS and A2S reprogramming were performed as above. Doxycycline and chemicals were removed on day 11 and AP-positive wells were scored on day 15.

Immunofluorescence

Immunofluorescence was performed as described in Sridharan et al., (2009). Briefly, cells were fixed with 4% paraformaldehyde-PBS, followed by permeabilization with 0.5% TritonX-PBS and stained with antibodies in blocking buffer (1X PBS with 5% normal donkey serum, 0.2% Tween-20, and 0.2% fish skin gelatin). Nanog (CosmoBio RCAB0002P), Dppa4 (ThermoFisher Scientific PA5-47530), Cdh1 (Ebioscience 14-3249-82), and Twist1 (Novus Biologicals, NBP2-37364SS) antibodies were used at 1:100 dilution, while Ki67 (Abcam ab15580) was used at 1:200. Imaging and colony counts were performed on Nikon Eclipse Ti using NIS Elements software.

Flow cytometry

MEFs were induced to reprogram in the A2S condition as above, but without irradiated MEFs. On day 3, cells were harvested with trypsin, resuspended to a single-cell suspension and stained with Epcam antibody (CD326) – PE conjugated (BD PharMingen 563477) at 1ul per $5x10^5$ cells for 1 hour before being sorted using BD FACS Aria II. Epcam+ cells were re-plated and allowed to reprogram for an additional 3 days before another FACS was performed on day 6.

siRNA Transfection

siRNA purchased from Integrated DNA Technologies or GE Life Science were transfected using Dharmafect reagent (GE Life Sciences) according to manufacturer's instructions. For Eif4a1 and Ehf knockdown experiments, siRNA was added on the day of plating at 0.5nM. siRNA was added every 48 hours and concentration was increased gradually up to 40nM to account for increasing cell numbers. For the Eif4a1 experiment, live cell counts were performed every day using Trypan Blue exclusion. For the Phlda2 experiments, siRNA was added at days 4 and 5 at 50nM and 75nM respectively. Two siRNAs were combined for Phlda2. The following siRNAs were used: Eif4a1 siRNA #1 mm.Ri.Eif4a1.13.1, Eif4a1 siRNA#2 mm.Ri.Eif4a1.13.2, Ehf siRNA #1 mm.Ri.Ehf.13.1, Ehf siRNA #2 mm.Ri.Ehf.13.2, Phlda2 siRNA#1 mm.Ri.Phlda2.13.1, Phlda2 siRNA#2 mm.Ri.Phlda2.13.2, Non-Targeting siRNA D-001810-01-50. To evaluate knockdown efficiency, qRT-PCR was performed using primers listed Table S2.

Single-Cell RNA-sequencing

To ensure optimal viability of cells during droplet formation, cells were washed once with DPBS, followed by a media change 12 hours prior to single cell isolation. On the day of single cell isolation, cells in 6-well plates were washed five times with DPBS, dislodged with 1 mL 0.25% trypsin-EDTA and neutralized with 1 mg/ml trypsin inhibitor (Sigma Aldrich T6522). Cells were filtered through a 35 um nylon mesh (Corning 352235) and centrifuged at 300xg for 3 min. Pelleted cells were gently washed with DPBS and pelleted again at 300xg, 3 min, RT. Cells were resuspended in 1 mL 0.1% BSA-DPBS (ThermoFisher 15260037) and gently pipetted 20-50 times. Single-cell suspension was confirmed under the microscope and cell concentration and viability were measured on a Bio-Rad TC20. Cells were diluted to a final concentration of about 2500 cells/uL in 0.1% BSA-DPBS.

Single-Cell Isolation and Library Preparation

Single-cell encapsulation was performed on a ddSEQ Single-Cell Isolator (BioRad 12004336), with reagents provided in the SureCell WTA 3^0 Library Prep Kit (Illumina 20014279), according to manufacturer's instructions. Briefly, approximately 12,500 cells in single-cell resuspension were mixed with Cell Enzyme Mix containing reverse transcriptase. A ddSEQ cartridge was primed with Priming Solution before Barcode Suspension Mix, Cell Suspension Mix, and encapsulation oil were loaded onto the cartridge and into the Isolator. Generated single-cell droplets were transferred to a pre-chilled plate and run on a thermal cycler to begin reverse transcription of mRNA. Droplets were subsequently disrupted, and first-strand library cDNA was used for

second strand synthesis. Quality of pre-amplified libraries was confirmed on High Sensitivity DNA Chips on the Agilent Technology 2100 Bioanalyzer. Libraries with a minimum of 1.8 ng DNA were tagmented with DNA adapters from the SureCell WTA 3^0 Library Prep Kit and amplified.

Next-Generation Sequencing and Genome Alignment

Between 7 and 9 libraries were multiplexed per lane on an Illumina HiSeq2500 Rapid Run (2x75), with a mean of over 280 million reads per lane. Fastq files (bcl2fastqv2.19) were generated, either through Illumina BaseSpace – the Illumina computing environment for sequencing data analysis – or through the University of Wisconsin-Madison Bioinformatics Resource Center, and uploaded to Illumina BaseSpace. Sequences were aligned to *Mus musculus* 10 (mm10) genome using Spliced Transcripts Alignment to a Reference (STAR), available through the SureCell RNA Single Cell App v1.1.0 on BaseSpace. On average, 85.48% reads per sample aligned to the genome, and 2.06% reads per sample aligned to abundant features (mitochondria, small non-coding RNA, ribosomal RNA). A unique molecular identifier (UMI) per cell plot was generated using BaseSpace, which indicates the total number passing filter. A drop in the knee plot indicated a transition to empty beads, in which a cell barcode contained low UMI counts. This drop serves as the threshold for calling cells that pass the sample-specific knee filter, and all subsequent analyses were performed with cells passing this filtering step. In total, we isolated 8,334 cells, and on average, 260 cells passed knee filter per sample, with a median of 53,497 genic reads, 13,100 genic UMIs and 4,274 genes detected per cell passing filter. Several libraries

were re-sequenced in order to achieve a sequencing a depth of approximately 50,000 reads per cell for each sample.

Bioinformatic analysis

t-SNE Clustering

We used Monocle2 v2.6.3 on R version 3.4.3 (Kite-Eating Tree) http://cole-trapnell-lab.github.io/monocle-release/docs/ (Qiu et al., 2017b, 2017a) to analyze the data obtained after alignment. We initially plotted the distribution of UMI counts within each cell and filtered out any cells with UMI counts outside a range determined by:

10^(mean(log10(Total number of UMIs within all cells of dataset)) +/2*standard deviation(log10(Total number of UMIs within all cells of dataset))

Out of the 4,374 cells passing filter, 4,167 cells were within the optimal UMI range and used for downstream Monocle analysis. Genes that were not expressed in at least 1 cell were excluded from analysis. Principal component analysis (PCA) was then performed to identify the variance explained by each component of the cell dataset (cds).

(1) cds < detectGenes(cds, min_expr = 0.1)

(2) fData(cds)$use_for_ordering < fData(cds)$num_cells_expressed > 0.1 *ncol(cds)

(3) plot_pc_variance_explained(cds, return_all = F)

We reduced the number of dimensions to the number of PC components that explained the most variance, before the PCA components began to level off. Together, these components explained at least 50% of the variance for each dataset.

In order to remove the irradiated feeder MEFs from our analysis, we performed a t-distributed stochastic neighbor embedding (t-SNE clustering) using data from ESCs grown on a feeder MEF layer in FBS and A2S. Irradiated MEFs formed a separate cluster and could be identified by expression of MEF markers in the ESC samples. We used the cell IDs of the FBS-ESCs that were found with these irradiated MEFs to identify where these cells are located in an initial clustering of all FBS, A2S reprogramming, MEF, and FBS-ESCs samples. The cells associated with this cluster were removed, resulting in the total of 4,374 cells that were used in the Monocle pipeline. A table with the Cell IDs that were removed from the analysis is available on GEO under entry GSE108222.

t-SNE Cluster Analysis

To identify genes important for defining clusters within the MEF reprogramming, DE analysis was performed between all 14 clusters within the t-SNE plot of Figure 2A. To determine the distribution of cells from each sample that fall into each cluster, phenotypic data (cell barcode ID, sample, pseudotime, cluster number) was extracted for each cluster and sample. The composition of a cluster or sample was then calculated by percentage or mean of the population. The top 10% of DEGs from this list were used in generating a heatmap to visualize percentage of cells within each cluster and sample that express these genes. Gene patterns were identified by k-means clustering into 15 groups using Cluster3 software and visualized by Java TreeView (de Hoon et al., 2004; Saldanha, 2004)(Figure 2C). Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang et al., 2009) was

used to functionally annotate groups of genes from heatmaps.

Generating pseudotime trajectory

To order cells by pseudotime, the gene expression of each cell has to be compared to a standard. We chose to use the top 5% of differentially expressed genes (DEG) between seven t-SNE clusters from only the A2S reprogramming samples and the MEFs and ESCs. This ensured that we were not comparing only established cell types or gene expression at specific time points. Using MEFs as the starting point, Monocle defined a pseudo-reprogramming time trajectory, termed pseudotime, where cells are linearly ordered relative to their progress or change in gene expression relative to the starting population. Lengths of the trajectory between each branchpoint were used to define state by the Monocle algorithm.

(4) diff_test_resClusterDE < differentialGeneTest(cds, fullModelFormulaStr = ''~Cluster,''

cores = detectCores())

(5) SetOrdering(cds, ordering_genes = Top 5% DE Genes

(6) reduceDimensions(cds)

(7) orderCells(cds)

(8) plot_cell_trajectory(cds, color_by = ''Phenotype Data'')

Branched expression analysis modeling (BEAM) was performed to identify genes involved in the decision-making process of progressing along the trajectory or to a branch. Genes involved in BEAM with a q-value less than 1e-40 were then plotted along pseudotime to visualize relative expression of genes as cells progress to either

branchpoint or toward the end of the trajectory.

(9) BEAM_cds < BEAM(cds, branch_point = 1/2, cores = detectCores())

(10) plot_genes_branched_heatmap(cds[row.names(subset(BEAM_cds, qval < 1e-40)),], branch_point = 1, num_clusters = 10, cores = detectCores(), use_gene_short_name = T, show_rownames = T)

(11) plot_genes_branched_pseudotime(cds_subset, branch_point = 2, color_by = ''Cluster,'' ncol = 1)

(12) diff_test_res_PseudotimeDE < differentialGeneTest(cds, fullModelFormulaStr = ''~Pseudotime,'' cores = detectCores())

Co-expression Analysis

To determine how a pair of two different genes are co-expressed within the cell population, Monocle's cell type hierarchy function was implemented. Cells expressing a particular gene were identified using the following command:

(13) GeneName_id < row.names(subset(fData(cds), gene_short_name = = ''Gene Name''))

(14) cth < newCellTypeHierarchy()

(15) cth < addCellType(cth, ''Gene Name 1 Positive'' classify_func = function(x) { x[GeneName1_id,] > 0 })

(16) cth < addCellType(cth, '' Gene Name 2 Positive,'' classify_func = function(x) { x[GeneName2 _id,] > 0 })

(17) cds < classifyCells(cds, cth, 0.1)

Visualizing the t-SNE plot based on cell type will identify cells that are positive for one of the genes of interest, those that are double positive (labeled Ambiguous), and those that are double negative (labeled Unknown). The phenotype data table contains information on cell type, which allows us to determine how prevalent each cell type is within each cluster, sample, etc. To visualize how different genes are expressed in cells that are known to be positive for a particular gene, we also generated violin plots. Note that due to the sequencing depth of single cell RNA-Seq (also known as ‘‘Dropout’’) co-expression may be underestimated. After defining a cell type using the above command (Gene 1+ cells), we use the following code to produce violin plots:

(18) Gene1_C1 < cds[,pData(cds)$CellType = = ‘‘Gene1+’’ & pData(cds)$Cluster = = ‘‘1’’]

(19) Gene1_C1_table < as.data.frame(pData(Gene1_C1))

(20) Gene1_C1_table$Identifier < row.names(Gene1_C1_table)

(21) Gene1_C1_Id < row.names(Gene1_C1_table)

(22) cds_log < log(exprs(cds)+1)

(23) t < as.data.frame(cds_log)

(24) Gene1_C1_counts < t[, colnames(t) %in% Gene1_C1_Id]

(25) Gene1_C1_counts_all < as.data.frame(t(Gene1_C1_counts))

(26) Gene1_C1_counts_all$CellType < ‘‘Gene1+ C1’’

(27) ggplot(Gene1_counts, aes(x = CellType, y = Gene2)) + geom_violin()

Constructing gene regulatory network (MERLIN)

MERLIN is based on a probabilistic graphical model representation of a regulatory network and uses a probabilistic graph structure prior to enable genes in the same module to have similar but not identical regulators. To infer networks, we used the top 10% of differentially expressed genes identified by Monocle, and added a list of 445 known transcription factors, signaling proteins, and chromatin remodelers, as well as genes known to be involved in early stem cell state specification, which resulted in 2,100 genes in 4,633 cells. We applied MERLIN in a stability selection framework. Briefly, we created 100 subsamples by randomly selecting 2,317 cells for each, and ran MERLIN independently on each subsample. As initial cluster assignments for genes, we used k-means with 10 clusters. We used the following default options for running MERLIN: -5 for sparsity, 4 for modularity prior and 0.6 for redefining modules. The outputs of MERLIN comprise a regulatory network as well as module assignments for input genes. We next obtained consensus networks and consensus modules as described in Chasman et al., (2016). Each edge in the consensus network has a confidence value that indicates the percentage of subsamples in which that edge was inferred. Consensus modules are defined by applying hierarchical clustering to a co-clustering matrix (which is the fraction of subsamples' where a pair of genes were in the same MERLIN module). We identified a total of 15 modules with at least 5 genes spanning 291 genes. We associated each consensus module with regulators based on a significant overlap (hypergeometric test, FDR < 0.05) of regulator targets from the 80% confidence network. Furthermore, we assessed the inferred modules for enrichments of Gene Ontology processes, and found 12 of the 15 consensus modules

to be enriched.

Visualizing inferred networks for each module

For a given module, we selected all incoming edges to that module from our 80% confidence network. Next, we selected cells from each condition (A2, S2, AS, and A2S, day 4 or day 6) and applied a linear regression model to predict the expression of the target gene as a function of its regulators in the 80% confidence network. We visualized these using the program Cytoscape (Shannon et al., 2003). Briefly, the edge color corresponds to regression coefficient of that regulator for the target (from -.5 (blue) to 0 (white) to 0.5 (red)). Edge width corresponds to edge confidence (from 80% (1) to 100% (5)). Node color corresponds to percentage of cells in which that gene was expressed (from 0% (white) to 100% (green)). Node border is pink if the gene is in the given module, and gray if it is not.

Dual Combinations

A Monocle cell dataset was created using single-cell data from all dual combination reprogramming experiments as well as the data from A2S days 4 and 6. Jitter plots were generated in Monocle to illustrate expression of specific genes in each different condition. The MERLIN algorithm was applied to the dual combination and A2S RNA-seq data to generate regulatory networks for the defined modules in each reprogramming condition.

Quantification and Statistical Analysis

Information on replicates for each experiment can be found in the figure legends. p and q values for differentially expressed genes of single cell RNA-sequencing data were calculated from likelihood ratio tests on the parallel arrays of models generated through monocle.

Data and Software Availability

All single-cell RNA-seq data have been submitted to the National Center for Biotechnology Information Gene Expression Omnibus database and can be accessed at GEO: GSE108222.

Figure 1

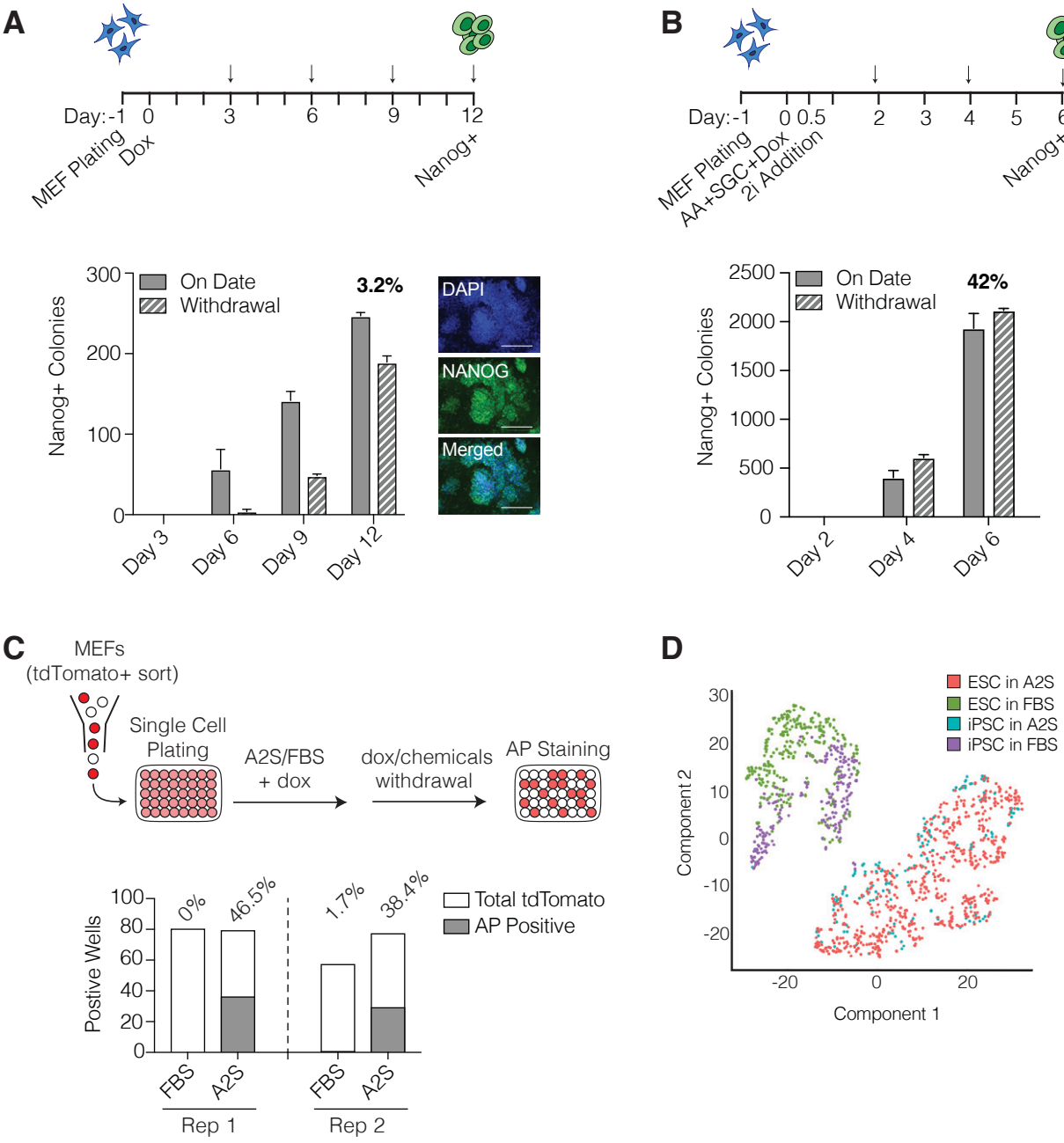

A
Day: -1 0 3 6 9 12
MEF Plating
Dox
Nanog+
Nanog+ Colonies
On Date
Withdrawal
3.2%
DAPI
NANOG
Merged
Day 3
Day 6
Day 9
Day 12
B
Day: -1 0 0.5 2 3 4 5 6
MEF Plating
AA+SGC+Dox
2i Addition
Nanog+
Nanog+ Colonies
On Date
Withdrawal
42%
Day 2
Day 4
Day 6
C
MEFs
(tdTomato+ sort)
Single Cell
Plating
A2S/FBS
+ dox
dox/chemicals
withdrawal
AP Staining
Postive Wells
0%
46.5%
1.7%
38.4%
Total tdTomato
AP Positive
FBS
A2S
Rep 1
Rep 2
D
ESC in A2S
ESC in FBS
iPSC in A2S
iPSC in FBS
Component 2
Component 1

Figure 1: Combining epigenomic and signaling modifiers leads to high-efficiency generation of bona fide iPSCs

A) Top: schematic of FBS reprogramming experiment. Cells were harvested and immunofluorescence performed on the days indicated by the arrows. Bottom: number of NANOG+ colonies counted at each indicated time point (on date) or after 4 additional days after doxycycline (dox) was removed (withdrawal). Bars represent SD between two replicate samples. Right panel – immunofluorescence images of NANOG. Scale bar, 250 mm.

B) Top: schematic of A2S reprogramming experiment. Cells were harvested and immunofluorescence performed on the days indicated by the arrows. Bottom: number of NANOG+ colonies counted at each indicated time point (On Date) or after 4 additional days after dox was removed (withdrawal). Bars represent SD between two replicate samples.

C) Top: schematic of single-cell reprogramming experiment. MEFs infected with tdTomato virus were sorted and plated in a 96-well plate. Dox-independent colonies were stained with alkaline phosphatase (AP). Bottom: number of AP+ wells observed in each condition. Percentages indicate how many of the wells were AP+ out of the total number of wells with tdTomato+ cells. Data from two independent experiments are presented.

D) Monocle clustering plot showing ESCs or iPSCs cultured in A2S or FBS media.

Figure 2

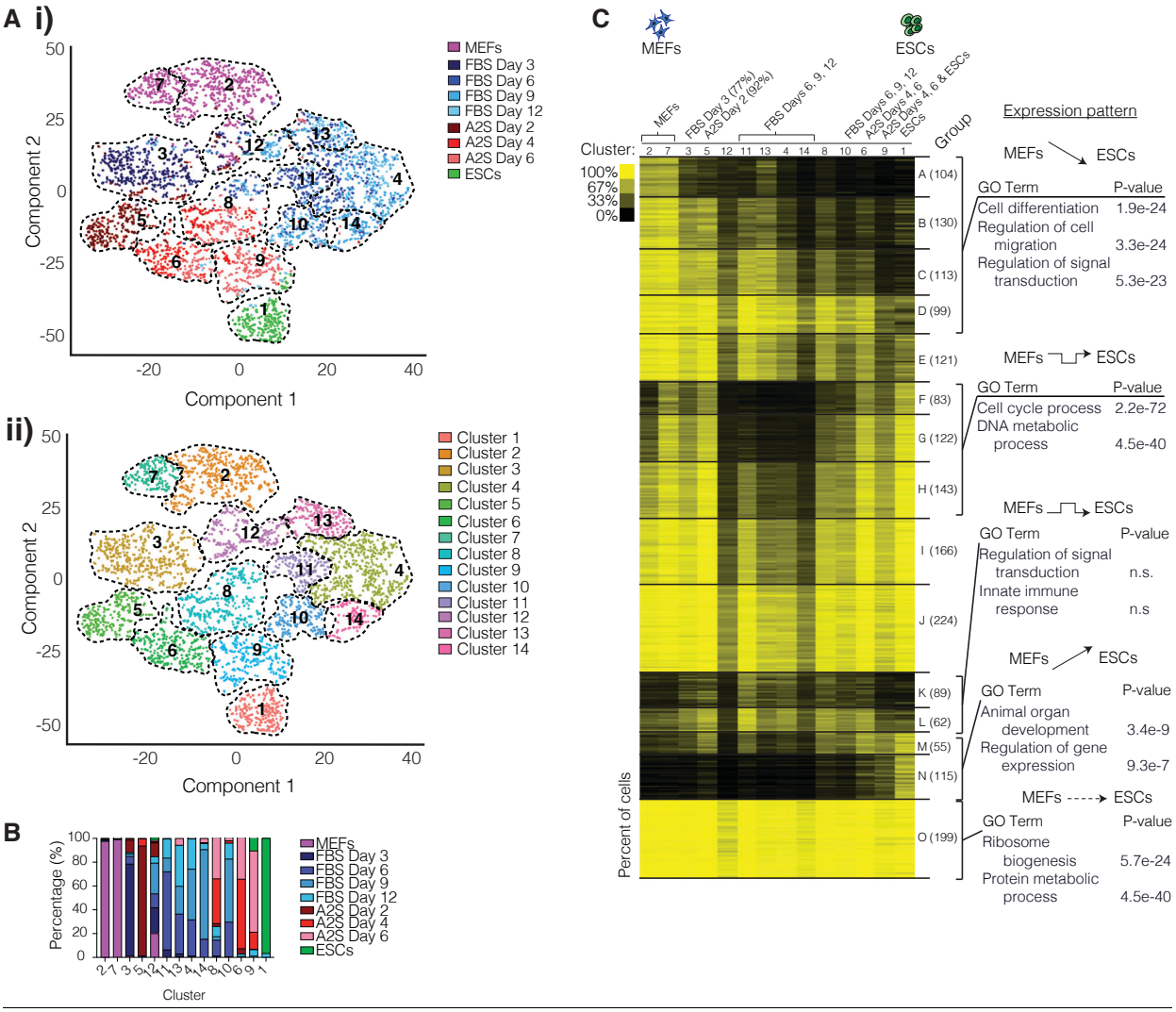

A i)
MEFs
FBS Day 3
FBS Day 6
FBS Day 9
FBS Day 12
A2S Day 2
A2S Day 4
A2S Day 6
ESCs
Component 2
Component 1
ii)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
B
Percentage (%)
Cluster
C
MEFs
ESCs
Cluster:
Group
Percent of cells
100%
67%
33%
0%
A (104)
B (130)
C (113)
D (99)
E (121)
F (83)
G (122)
H (143)
I (166)
J (224)
K (89)
L (62)
M (55)
N (115)
O (199)
Expression pattern
GO Term
P-value
Cell differentiation 1.9e-24
Regulation of cell migration 3.3e-24
Regulation of signal transduction 5.3e-23
Cell cycle process 2.2e-72
DNA metabolic process 4.5e-40
Regulation of signal transduction n.s.
Innate immune response n.s
Animal organ development 3.4e-9
Regulation of gene expression 9.3e-7
Ribosome biogenesis 5.7e-24
Protein metabolic process 4.5e-40

Figure 2: A2S accelerates FBS reprogramming

A) Monocle t-SNE plots showing clustering of reprogramming cells from FBS and A2S, MEFs, and FBS-cultured ESCs. Samples were grouped into 14 clusters. Cells colored by sample (i) and cluster (ii).

B) Graph showing the composition of each cluster from Figure 2A by sample

C) Heatmap representing the percentage of cells expressing the top 10% differentially expressed genes that define the 14 t-SNE clusters in Figure 2A. Each row represents a single gene. Genes were grouped by k-means into 15 groups labeled A to O, and the number of genes within each group are in parentheses. The 14 t-SNE clusters labeled 1–14 are presented in columns approximating their similarity to ESCs. Significant gene ontology terms associated with a specific group are labeled on the right. n.s., not significant. Arrows indicate pattern of expression change between MEFs and ESCs.

Figure 3

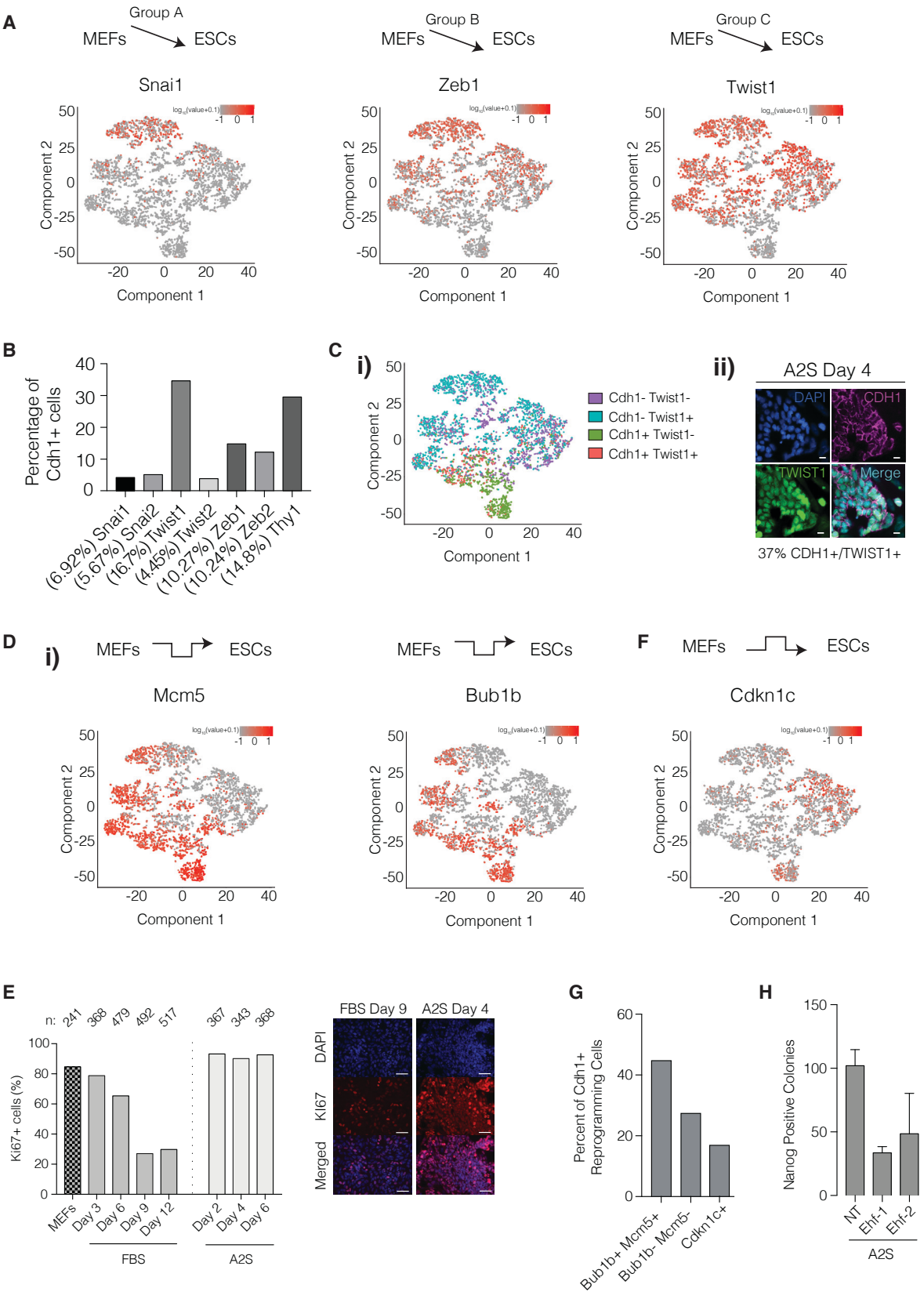

A
Group A
MEFs
ESCs
Group B
MEFs
ESCs
Group C
MEFs
ESCs
Snai1
Zeb1
Twist1
Component 2
Component 1
B
Percentage of Cdh1+ cells
(6.92%) Snai1
(5.67%) Snai2
(16.7%) Twist1
(4.45%) Twist2
(10.27%) Zeb1
(10.24%) Zeb2
(14.8%) Thy1
C
i)
Cdh1- Twist1-
Cdh1- Twist1+
Cdh1+ Twist1-
Cdh1+ Twist1+
ii)
A2S Day 4
DAPI
CDH1
TWIST1
Merge
37% CDH1+/TWIST1+
D
i)
MEFs
ESCs
Mcm5
Bub1b
F
Cdkn1c
E
n: 241 368 479 492 517 367 343 368
Ki67+ cells (%)
MEFs
Day 3
Day 6
Day 9
Day 12
Day 2
Day 4
Day 6
FBS
A2S
FBS Day 9
A2S Day 4
DAPI
KI67
Merged
G
Percent of Cdh1+ Reprogramming Cells
Bub1b+ Mcm5+
Bub1b- Mcm5-
Cdkn1c+
H
Nanog Positive Colonies
NT
Ehf-1
Ehf-2
A2S

Figure 3: Reprogramming-specific gene expression patterns are important for conversion to iPSCs

A) t-SNE plots based on Figure 2A highlighting the expression of MEF-associated mesenchymal genes that are downregulated as cells transition from MEFs to pluripotency. Top schematic indicates the pattern of expression.

B) Percentage of Cdh1+ cells that also co-express the indicated MEF genes on the x axis. The percentage of MEF gene-expressing cells that express Cdh1 is presented in brackets on the x axis. Note that because of the limit of detection of single-cell transcriptional analysis, co-expression may be underestimated.

C) (i) t-SNE plots based on Figure 2A illustrating co-expression of Cdh1 with Twist1. Note that because of the limit of detection of single-cell transcriptional analysis, co-expression may be underestimated. (ii) Immunofluorescent staining for CDH1 and TWIST1. Percentage of CDH1+/TWIST1+ colonies on A2S day4 shown below image. Scale bar, 10 mm.

D) t-SNE plots based on Figure 2A highlighting the expression of DNA replication and cell-cycle-associated genes. Top schematic indicates the pattern of expression.

E) Left: percentage of cells that are Ki67+ at each indicated reprogramming time point in FBS or A2S systems. Right: immunofluorescent staining of Ki67 during FBS and A2S reprogramming (day 9 and day 4, respectively). Scale bar, 50 mm.

F) t-SNE plot based on Figure 2A for the anti-proliferation gene Cdkn1c. Top schematic indicates the pattern of expression.

G) Percentage of Cdh1+ cells that co-express cell cycle or anti-proliferative genes. Note that because of the limit of detection of single-cell transcriptional analysis co-

expression may be underestimated.

H) Number of NANOG+ colonies on day 4 of A2S reprogramming after small interfering RNA (siRNA)-mediated knock down of Ehf. Error bars represent SD of two replicates.

Figure 4

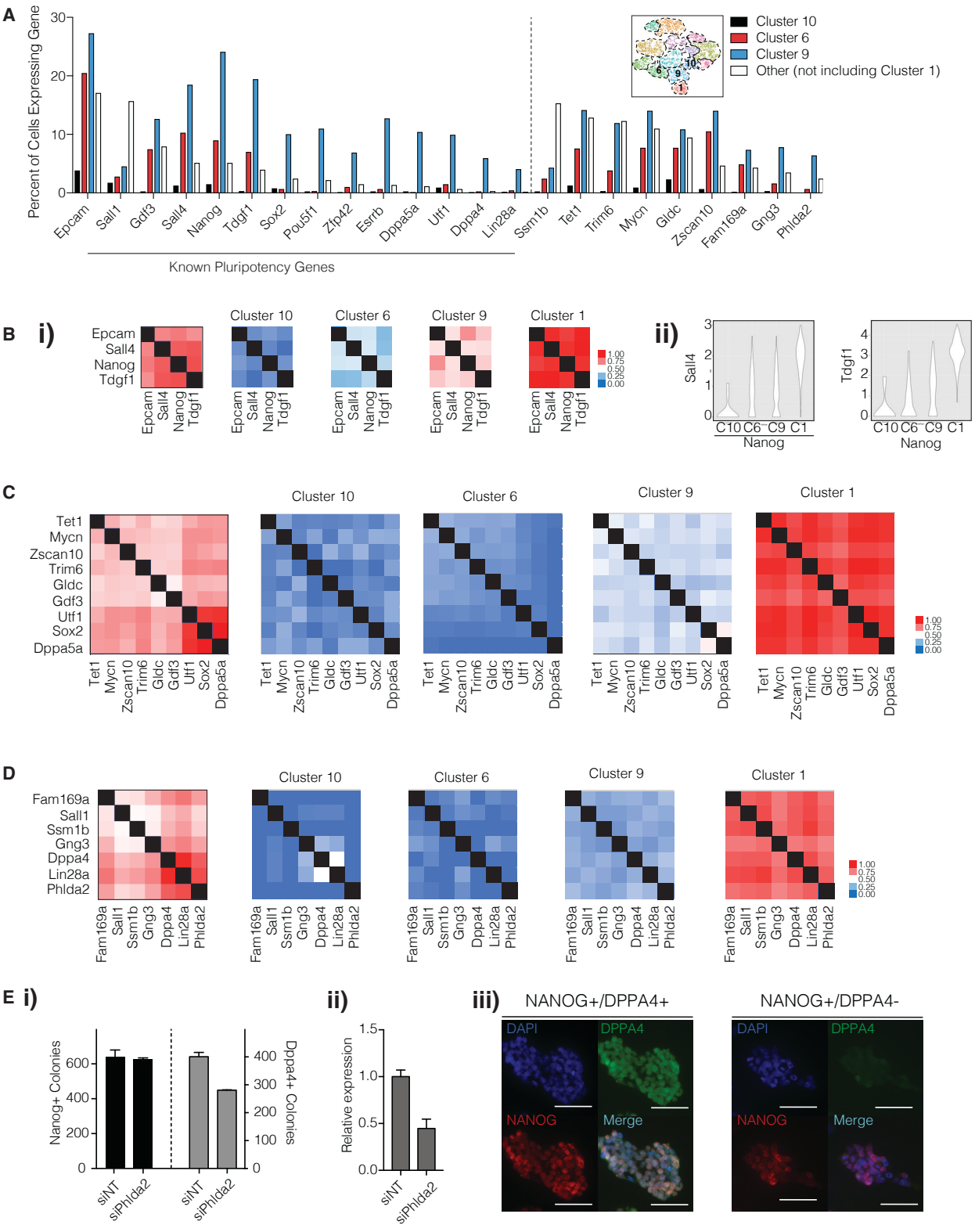

A
Percent of Cells Expressing Gene
Cluster 10
Cluster 6
Cluster 9
Other (not including Cluster 1)
Epcam Sall1 Gdf3 Sall4 Nanog Tdgf1 Sox2 Pou5f1 Zfp42 Esrrb Dppa5a Utf1 Dppa4 Lin28a Ssm1b Tet1 Trim6 Mycn Gldc Zscan10 Fam169a Gng3 Phlda2
Known Pluripotency Genes
B i)
Epcam Sall4 Nanog Tdgf1
Cluster 10
Cluster 6
Cluster 9
Cluster 1
ii)
Sall4
Tdgf1
C10 C6 C9 C1
Nanog
C
Tet1 Mycn Zscan10 Trim6 Gldc Gdf3 Utf1 Sox2 Dppa5a
Cluster 10
Cluster 6
Cluster 9
Cluster 1
D
Fam169a Sall1 Ssm1b Gng3 Dppa4 Lin28a Phlda2
Cluster 10
Cluster 6
Cluster 9
Cluster 1
E i)
Nanog+ Colonies
Dppa4+ Colonies
siNT
siPhlda2
ii)
Relative expression
iii)
NANOG+/DPPA4+
NANOG+/DPPA4-
DAPI
DPPA4
NANOG
Merge

Figure 4: Co-expression clusters of core pluripotency factors with specific subsets

A) Percentage of cells expressing each representative pluripotency-associated gene within the t-SNE clusters from Figure 2A, namely, C10, C6, C9, and in all clusters other than C1, C10, C6, and C9.

B) (i) Co-expression measured by Jaccard index clustering of genes in group N from Figure 2C for genes within Box 1 from Figure S4B in clusters C10, C6, C9, and C1. Note that because of the limit of detection of single-cell transcriptional analysis, co-expression may be underestimated. (ii) Violin plots depicting the level of expression of Sall4 and Tdgf1 in Nanog+ cells in clusters C10, C6, C9, and C1.

C) Same as (B) for genes within Box 2 of Figure S4B.

D) Same as (B) for genes within Box 3 of Figure S4B.

E) Reprogramming results upon knockdown of Phlda2 during A2S reprogramming. (i) Number of NANOG+ and DPPA4+ colonies on day 6 of A2S reprogramming after siRNA-mediated knock down of Phlda2. Error bars represent SD of two replicates. (ii) Knock down efficiency of the Phlda2 siRNAs compared to a nontargeting control. Bars represent SD between two replicate samples. (iii) Immunofluorescence images for representative NANOG+/DPPA4+ and NANOG+/ DPPA4- colonies. Scale bar, 50 mm.

Figure 5

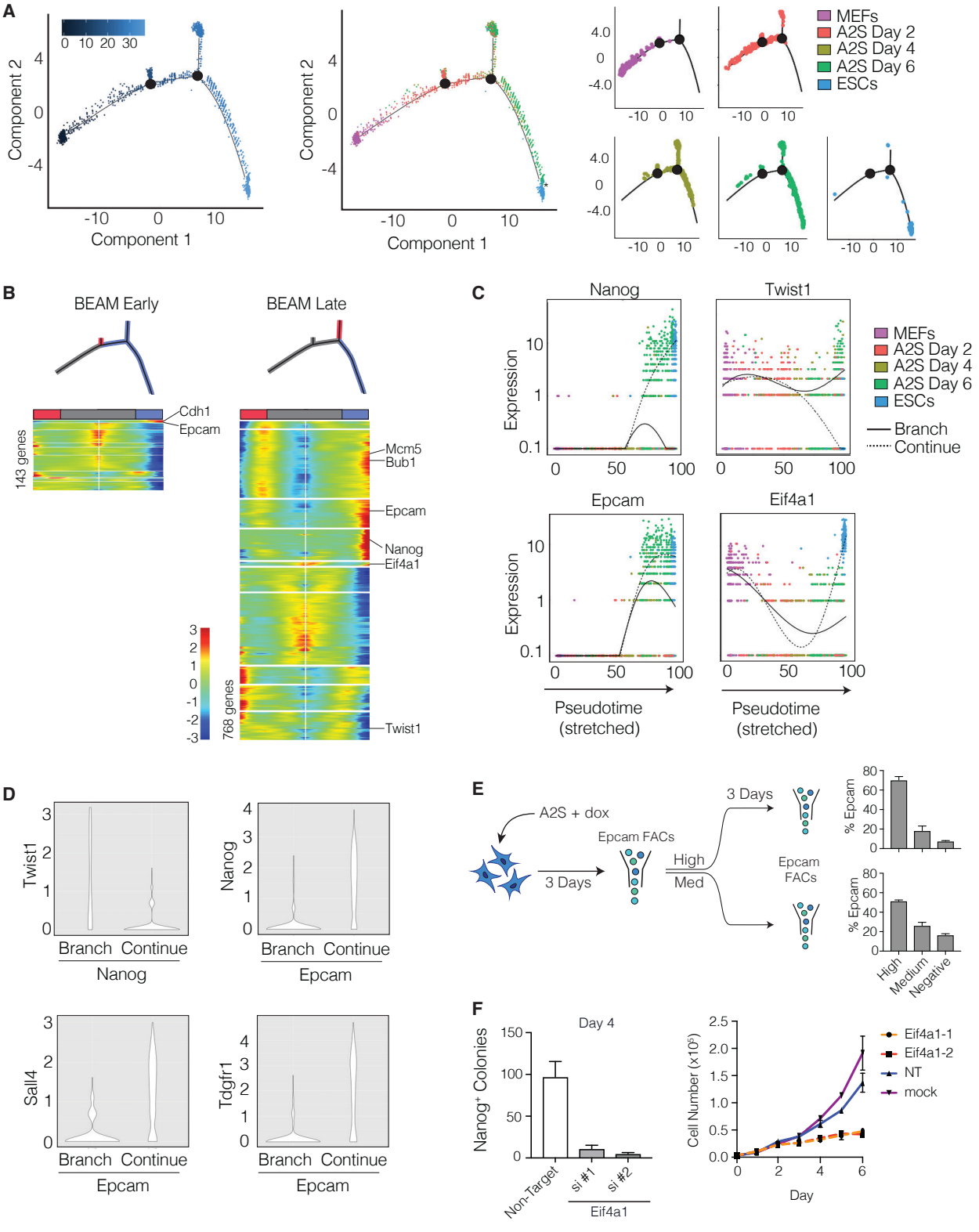

A
0 10 20 30
Component 2
Component 1
MEFs
A2S Day 2
A2S Day 4
A2S Day 6
ESCs
B
BEAM Early
BEAM Late
Cdh1
Epcam
143 genes
Mcm5
Bub1
Epcam
Nanog
Eif4a1
Twist1
768 genes
C
Nanog
Twist1
Epcam
Eif4a1
Expression
Pseudotime (stretched)
Branch
Continue
D
Twist1
Nanog
Sall4
Tdgfr1
Branch Continue
E
A2S + dox
3 Days
Epcam FACs
High
Med
% Epcam
High
Medium
Negative
F
Day 4
Nanog+ Colonies
Non-Target
si #1
si #2
Eif4a1
Cell Number (x10⁵)
Day
Eif4a1-1
Eif4a1-2
NT
mock

Figure 5: Roadblocks to high-efficiency reprogramming

A) Pseudotime trajectory generated by Monocle for the A2S reprogramming system. Left trajectory colored by pseudotime. Middle trajectory colored by sample. Asterisk indicates that MEFs cannot ontogenically convert to ESCs, but pseudotime reflects transition to a pluripotent state. Right trajectory colored by individual sample.

B) Heatmaps for clustering of genes that define the branchpoints (q-value, <1E-40) from BEAM analysis for early branch (left panel) and late branch (right panel). Center of the gray bar above heatmap is the start of the branchpoint. Red represents cells at the end of the branchpoint. Blue represents cells at the end of the continuing branch.

C) Pseudotime plots that display how the expression of the representative genes differs as cells either exit at the late branchpoint (solid line) or continue along the path toward successful reprogramming (dashed line) colored by sample.

D) Violin plots depicting the level of expression of Twist1 in Nanog+ cells (top left) and the expression of Nanog, Sall4, and Tdgf1 in Epcam+ cells in both the late branch and in the continuing segment of the trajectory.

E) Left: schematic of EPCAM sort experiment. MEFs were reprogrammed in A2S conditions for 3 days and sorted based on EPCAM expression (high or medium). These two populations underwent 3 more days of reprogramming and were sorted again based on high, medium, or no expression of EPCAM. Right: graphs depicting the percentage of the day 6 population that have high, medium, or no EPCAM expression from cells that were EPCAM-high on day 3 (top) or medium on day 3 (bottom).

F) Left: number of NANOG+ colonies on day 4 of A2S reprogramming after siRNA-mediated knock down of Eif4a1. Error bars represent SD of two replicates. Right: cell counts on each day of Eif4a1 knock down reprogramming experiment.

Figure 6

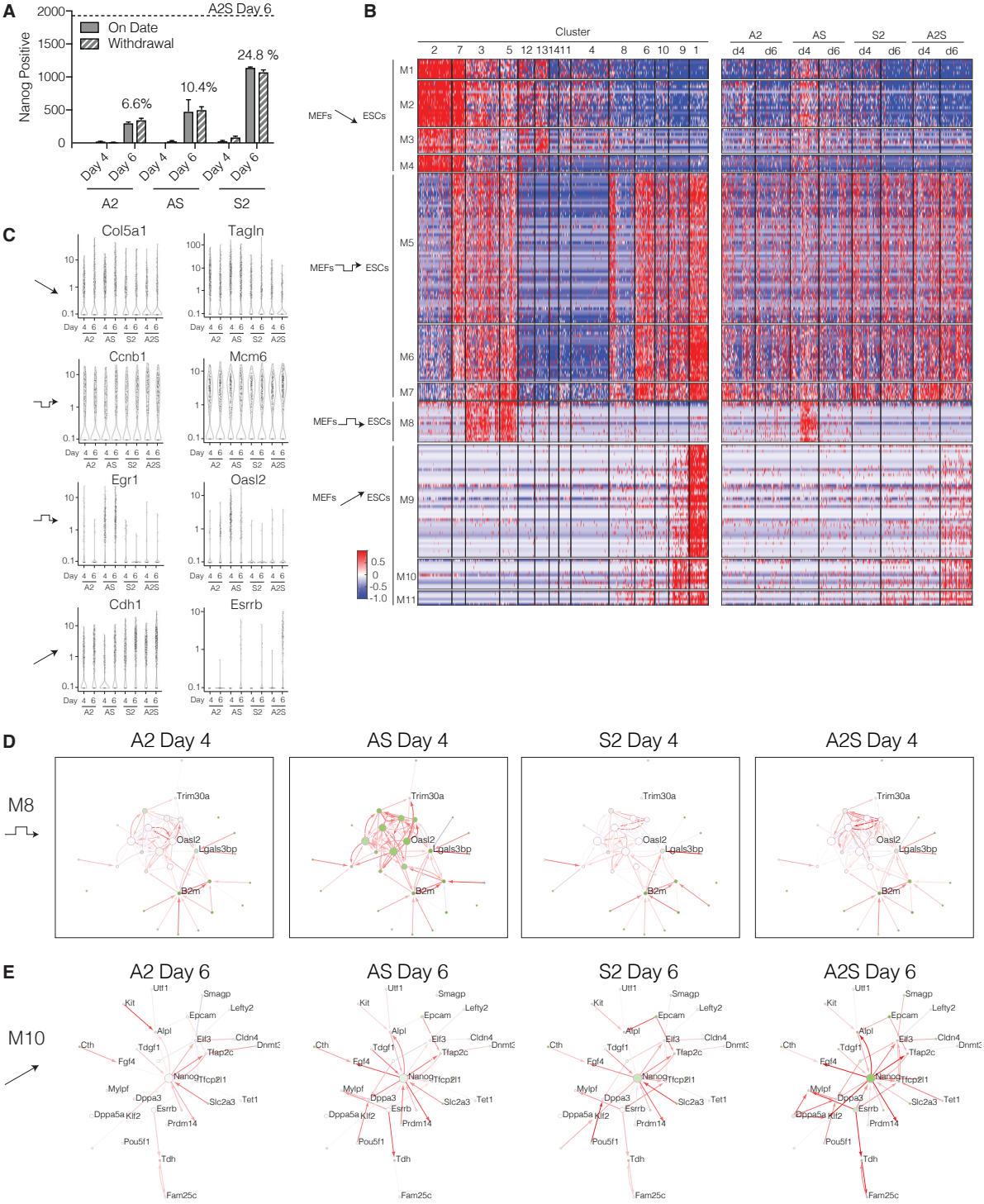

A
A2S Day 6
On Date
Withdrawal
Nanog Positive
2000
1500
1000
500
0
6.6%
10.4%
24.8 %
Day 4
Day 6
A2
AS
S2
B
Cluster
2 7 3 5 12 13 14 11 4 8 6 10 9 1
A2 AS S2 A2S
d4 d6
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
MEFs ESCs
0.5
0
-0.5
-1.0
C
Col5a1
Tagln
Ccnb1
Mcm6
Egr1
Oasl2
Cdh1
Esrrb
D
A2 Day 4
AS Day 4
S2 Day 4
A2S Day 4
M8
Trim30a
Oasl2
Lgals3bp
B2m
E
A2 Day 6
AS Day 6
S2 Day 6
A2S Day 6
M10
Utf1
Smagp
Kit
Lefty2
Epcam
Alpl
Eif3
Cldn4
Dnmt3
Cth
Tdgf1
Tfap2c
Fgf4
Nanog
Tfcp2l1
Mylpf
Dppa3
Slc2a3
Tet1
Dppa5a
Klf2
Esrrb
Prdm14
Pou5f1
Tdh
Fam25c

Figure 6: A2S concurrently enhances downregulation of MEF genes and upregulation of ESC genes

A) NANOG+ colonies on specified day or after 4 days of dox withdrawal in each dual combination (A2, AS, and S2). Dashed line: NANOG+ colonies on day 6 of A2S. Bars represent standard deviation between two replicate samples.

B) Heatmap generated from the MERLIN module analysis indicating the level of expression for the differentially expressed genes from the FBS+A2S analysis. Each row is a separate gene. Values are normalized to zero mean from the FBS and A2S reprogramming. Each column is a separate cell grouped based on the clusters in Figure 2A (left) or duration of chemical combination exposure (right). MERLIN modules are labeled as M1 through M11.

C) Violin plots of representative genes from expression patterns in Figure 6B.

D) Network wiring of regulatory connections inferred using MERLIN, colored by each reprogramming condition for the genes of a transiently expressed module. The edge color corresponds to the regression coefficient between the regulator and target connected by the edge (ranging from -0.5 (blue) to 0 (white) to 0.5 (red)) estimated using the data from the specific treatment. Edge width corresponds to edge confidence (from 80% [1] to 100% [5]). Node color corresponds to percentage of cells in a condition in which that gene was expressed (from 0% [white] to 100% [green]). Node border indicates gene membership in a module: pink if the gene is in the given module and gray if it is not. The node size is proportional to the out-degree of the node. Network corresponds to M8.

E) Same as (D) for genes in an upregulated pluripotency-associated gene module M10.

Figure 7

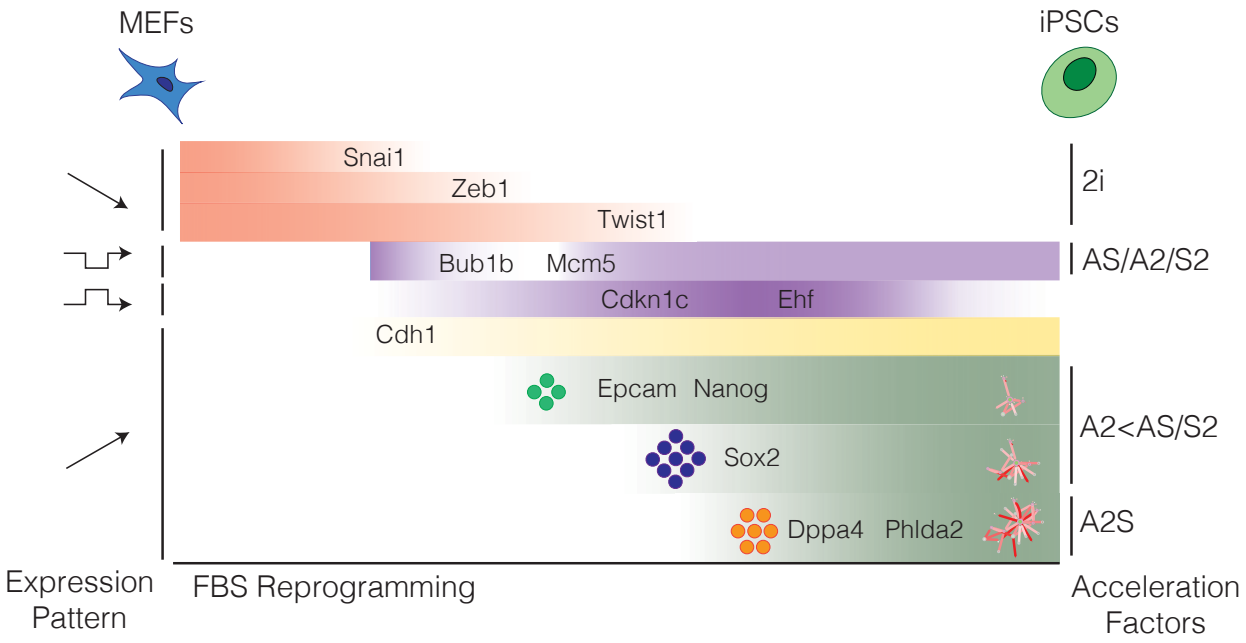
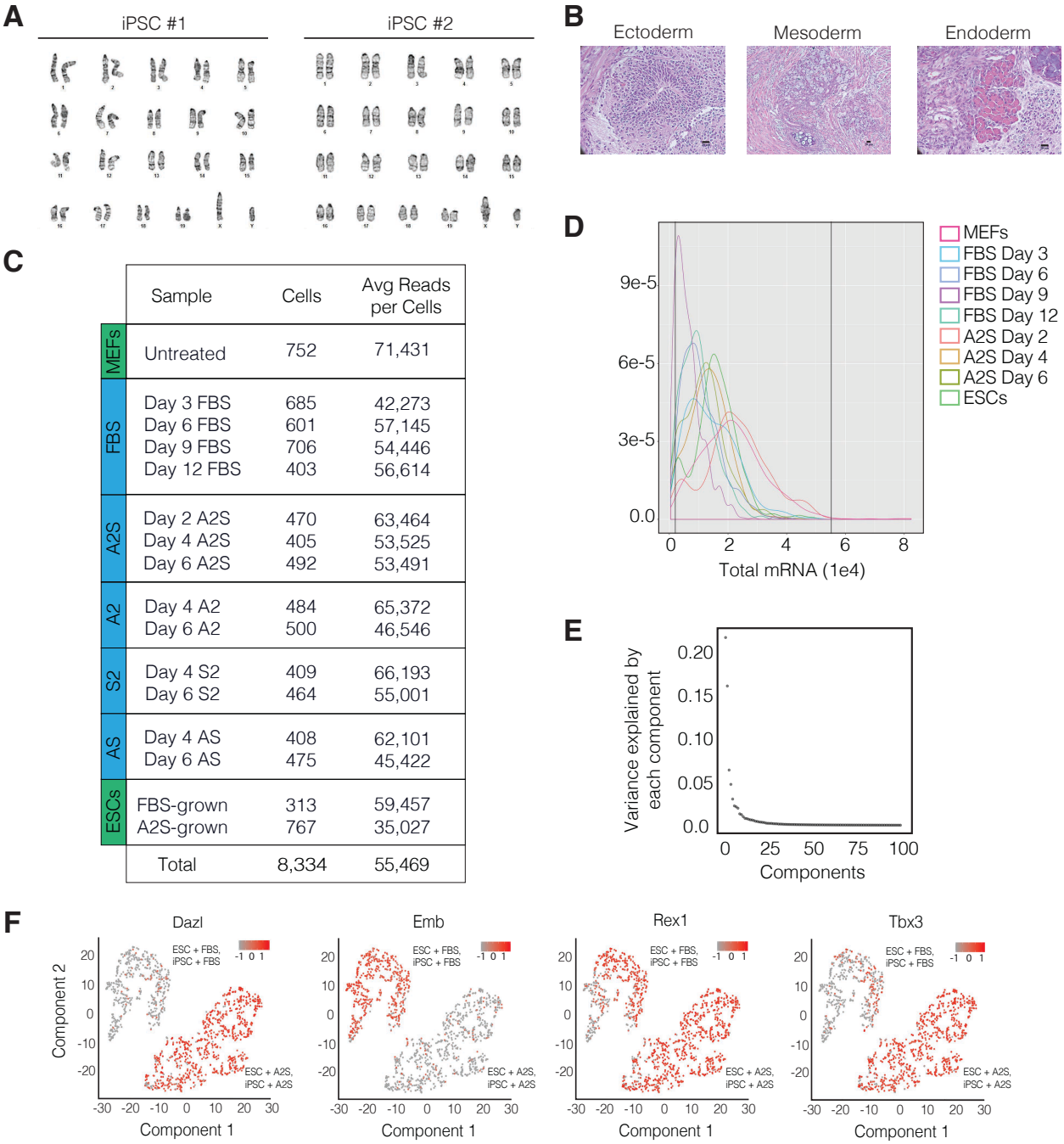


Figure 7: Model depicting regulation of key genes during MEF reprogramming

Four general gene expression patterns are observed during MEF reprogramming: downregulation, transient downregulation, transient upregulation, and gene upregulation. Mesenchymal genes are downregulated independently of each other and their expression is compatible with epithelial (Cdh1) or early pluripotency (Nanog) gene expression. Transiently regulated genes include cell cycle and anti-proliferative genes. Completion of reprogramming is enhanced by co-expression of markers, such as EpCAM with pluripotency genes (represented by colored circles), and the complete activation of the pluripotency network (represented by red and white networks). The addition of acceleration factors can impact specific gene expression patterns, whereas only the combination of A2S can lead to complete rewiring of the pluripotency network.

Figure S1



	Sample	Cells	Avg Reads per Cells
MEFs	Untreated	752	71,431
FBS	Day 3 FBS	685	42,273
	Day 6 FBS	601	57,145
	Day 9 FBS	706	54,446
	Day 12 FBS	403	56,614
A2S	Day 2 A2S	470	63,464
	Day 4 A2S	405	53,525
	Day 6 A2S	492	53,491
A2	Day 4 A2	484	65,372
	Day 6 A2	500	46,546
S2	Day 4 S2	409	66,193
	Day 6 S2	464	55,001
AS	Day 4 AS	408	62,101
	Day 6 AS	475	45,422
ESCs	FBS-grown	313	59,457
	A2S-grown	767	35,027
	Total	8,334	55,469

Figure S1 (Related to Figure 1)

A) Karyotype of two iPSC lines derived from A2S reprogramming generated from independent reprogramming experiments.

B) Cross-section images of teratoma obtained from iPSC lines displaying neuroepithelial/neuroglial cells (Ectoderm), cartilage tissue (Mesoderm), and pancreatic cells (Endoderm). Scale bar = 20μm.

C) Summary table of single-cell data generated from the Illumina BaseSpace Sequence Hub, including the number of cells passing the knee filter for each sample and the average number of genic reads per cell.

D) Monocle plot showing the upper and lower bound cutoff used to filter cells in Monocle for the FBS and A2S analysis.

E) PCA plot for variance explained from each component of the FBS and A2S Monocle analysis. We used the first 8 components for analysis.

F) Monocle plots highlighting the expression of Dazl, Emb, Rex1, and Tbx3 in the clustering pattern from Fig 1D.

Figure S2

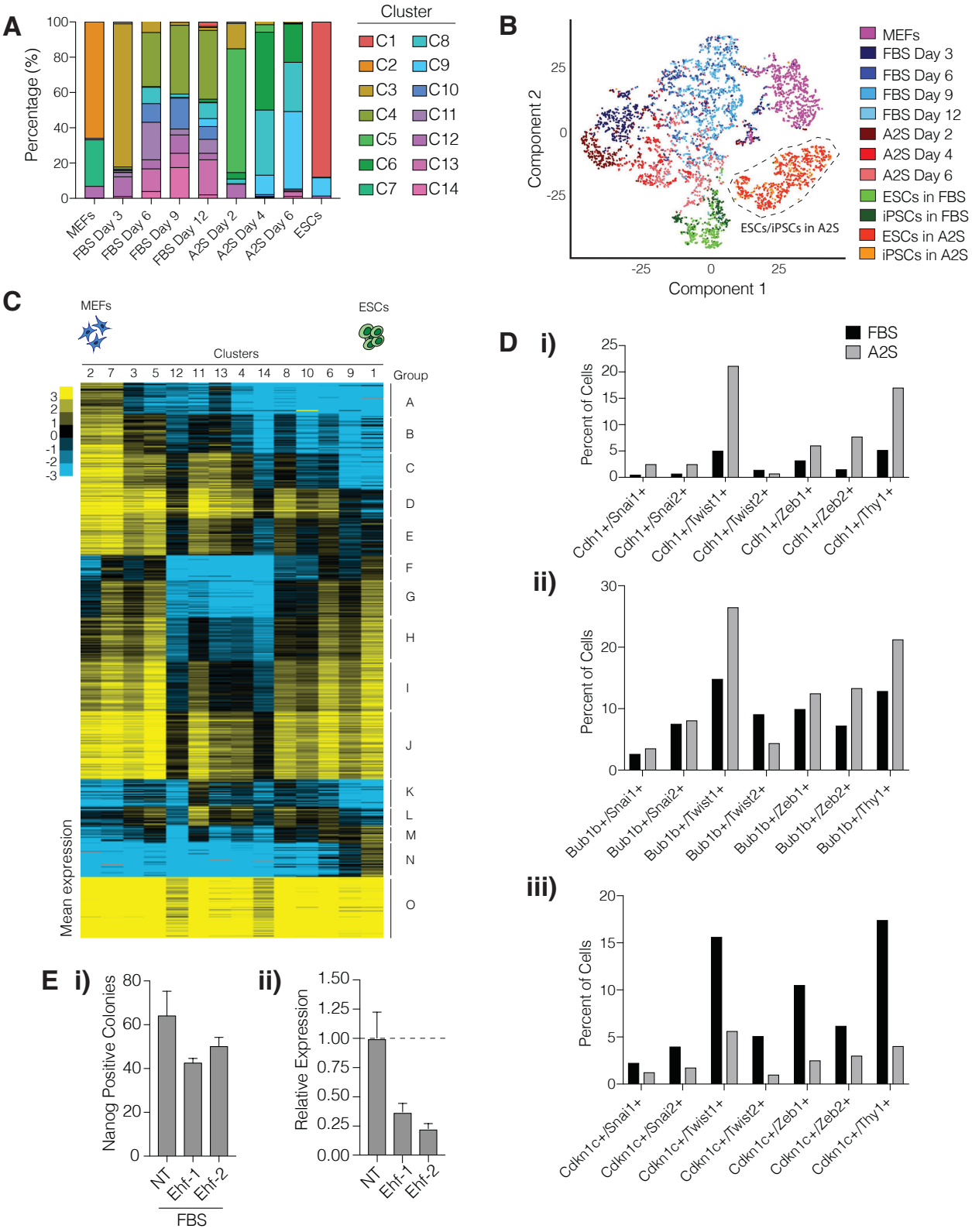

A
Percentage (%)
Cluster
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14
MEFs
FBS Day 3
FBS Day 6
FBS Day 9
FBS Day 12
A2S Day 2
A2S Day 4
A2S Day 6
ESCs
B
Component 1
Component 2
ESCs/iPSCs in A2S
MEFs
FBS Day 3
FBS Day 6
FBS Day 9
FBS Day 12
A2S Day 2
A2S Day 4
A2S Day 6
ESCs in FBS
iPSCs in FBS
ESCs in A2S
iPSCs in A2S
C
MEFs
ESCs
Clusters
Group
Mean expression
D i)
FBS
A2S
Percent of Cells
Cdh1+/Snai1+
Cdh1+/Snai2+
Cdh1+/Twist1+
Cdh1+/Twist2+
Cdh1+/Zeb1+
Cdh1+/Zeb2+
Cdh1+/Thy1+
ii)
Bub1b+/Snai1+
Bub1b+/Snai2+
Bub1b+/Twist1+
Bub1b+/Twist2+
Bub1b+/Zeb1+
Bub1b+/Zeb2+
Bub1b+/Thy1+
iii)
Cdkn1c+/Snai1+
Cdkn1c+/Snai2+
Cdkn1c+/Twist1+
Cdkn1c+/Twist2+
Cdkn1c+/Zeb1+
Cdkn1c+/Zeb2+
Cdkn1c+/Thy1+
E i)
Nanog Positive Colonies
NT
Ehf-1
Ehf-2
FBS
ii)
Relative Expression
NT
Ehf-1
Ehf-2

Figure S2 (Related to Figures 2 and 3)

A) Graph showing the composition of each sample from Fig 2A by cluster.

B) Monocle clustering of all the samples from Fig 2A along with ESCs and iPSCs cultured in A2S and FBS media.

C) Heatmap showing log-transformed mean expression data of the top 10% differentially expressed genes that define the t-SNE clustering in Fig 2A.

D) Percentage of cells in the FBS and A2S conditions that co-express Cdh1 (top), Bub1b (middle), and Cdkn1c (bottom) along with the indicated mesenchymal-associated gene.

E) Left Number of NANOG+ colonies on day 4 of FBS reprogramming after siRNA-mediated knockdown of Ehf. Error bars represent standard deviation of two replicates. Right Knockdown efficiency of each Ehf siRNA compared to a non-targeting control. Error bars represent standard deviation of two replicates.

Figure S3

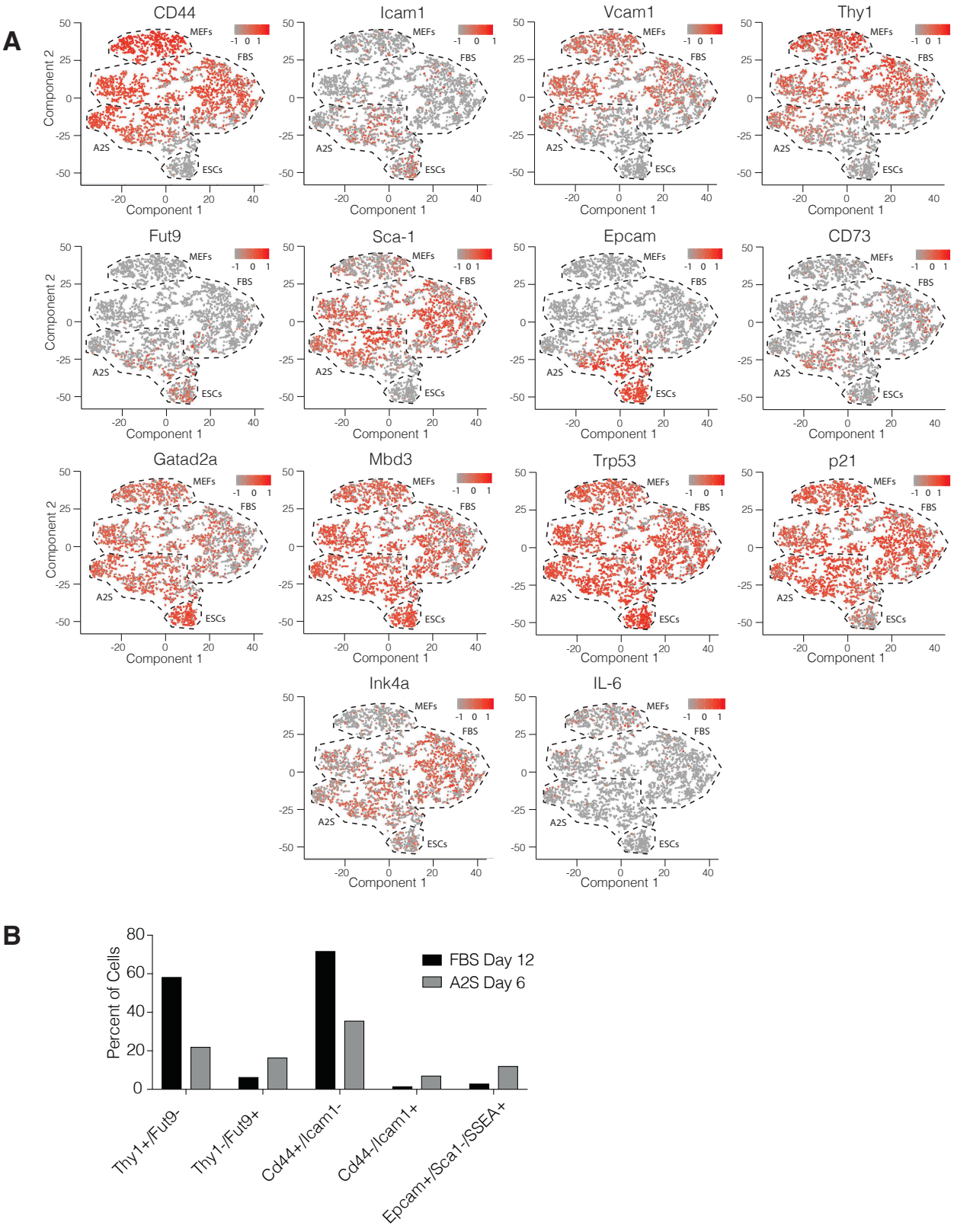

A
CD44
Icam1
Vcam1
Thy1
Fut9
Sca-1
Epcam
CD73
Gatad2a
Mbd3
Trp53
p21
Ink4a
IL-6
MEFs
FBS
A2S
ESCs
Component 1
Component 2
B
Percent of Cells
FBS Day 12
A2S Day 6
Thy1+/Fut9-
Thy1-/Fut9+
Cd44+/Icam1-
Cd44-/Icam1+
Epcam+/Sca1-/SSEA+

Figure S3 (Related to Figures 2 and 3)

A) Monocle plots highlighting the expression of selected reprogramming and senescence markers.

B) Percentage of cells on FBS Day 12 or A2S Day 6 that express each indicated combination of reprogramming markers.

Figure S4

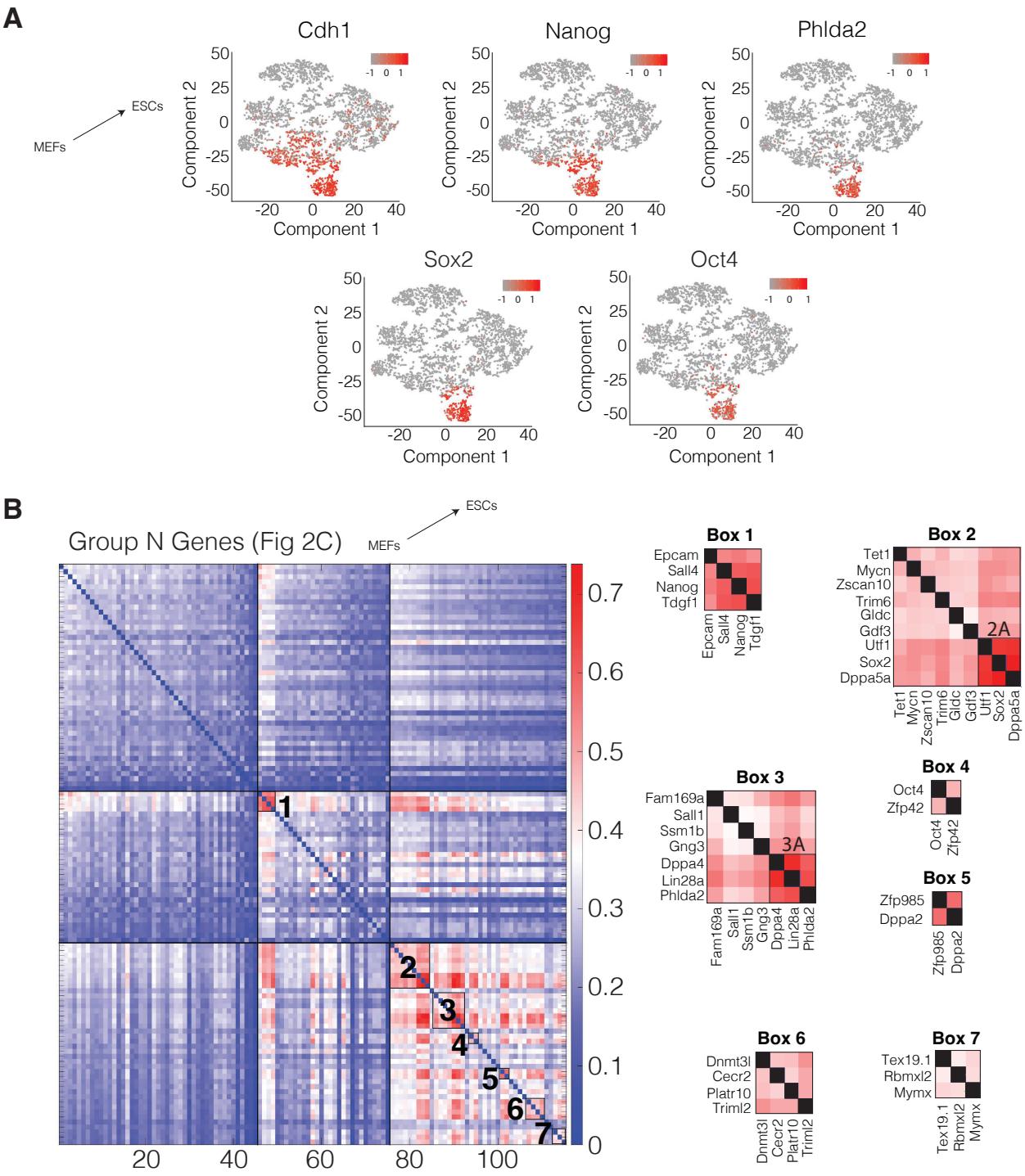

A
MEFs
ESCs
Cdh1
Nanog
Phlda2
Sox2
Oct4
Component 1
Component 2
B
Group N Genes (Fig 2C)
MEFs
ESCs
Box 1
Epcam
Sall4
Nanog
Tdgf1
Box 2
Tet1
Mycn
Zscan10
Trim6
Gldc
Gdf3
Utf1
Sox2
Dppa5a
2A
Box 3
Fam169a
Sall1
Ssm1b
Gng3
Dppa4
Lin28a
Phlda2
3A
Box 4
Oct4
Zfp42
Box 5
Zfp985
Dppa2
Box 6
Dnmt3l
Cecr2
Platr10
Triml2
Box 7
Tex19.1
Rbmxl2
Mymx

Figure S4 (Related to Figure 4)

A) Monocle plots highlighting the expression of representative pluripotency-associated genes that are upregulated as cells transition from MEFs to iPSCs.

B) Matrix showing the co-expression of pluripotency-associated genes (Group N from Fig 2C) with each other. The values correspond to the Jaccard index (number of cells positive for both genes divided by the number of cells that are positive for at least one of the two genes). All cells from the FBS and A2S analysis were used to generate this matrix. Genes of interest that displayed strong co-expression were boxed off and identified by the number shown. Zoomed in images of the numbered cohorts of genes are on the right.

Figure S5

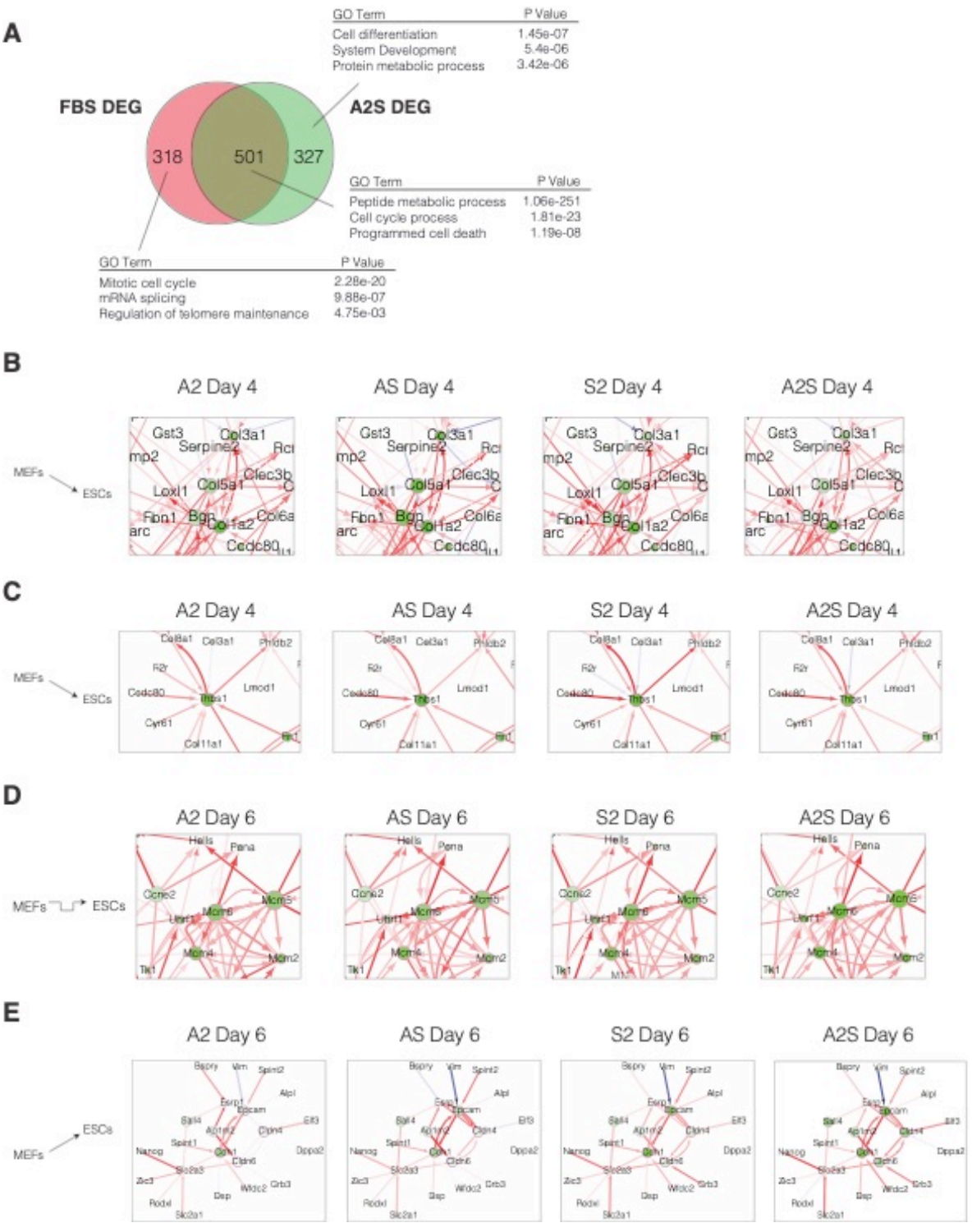

A
GO Term P Value
Cell differentiation 1.45e-07
System Development 5.4e-06
Protein metabolic process 3.42e-06
FBS DEG
A2S DEG
318
501
327
GO Term P Value
Peptide metabolic process 1.06e-251
Cell cycle process 1.81e-23
Programmed cell death 1.19e-08
GO Term P Value
Mitotic cell cycle 2.28e-20
mRNA splicing 9.88e-07
Regulation of telomere maintenance 4.75e-03
B
A2 Day 4
AS Day 4
S2 Day 4
A2S Day 4
MEFs
ESCs
C
A2 Day 4
AS Day 4
S2 Day 4
A2S Day 4
MEFs
ESCs
D
A2 Day 6
AS Day 6
S2 Day 6
A2S Day 6
MEFs
ESCs
E
A2 Day 6
AS Day 6
S2 Day 6
A2S Day 6
MEFs
ESCs

Figure S5 (Related to Figures 5 and 6)

A) Venn diagram depicting overlap of differentially expressed genes from FBS and A2S reprogramming. Clustering was performed on the A2S and FBS samples individually and differentially expressed genes were determined. This clustering and analysis were performed without the presence of MEFs or ESCs, eliminating any influence they may have on the differential gene analysis. Gene ontology terms associated with each group of genes are also displayed

B-E) Additional MERLIN network diagrams for selected patterns.

Figure S6

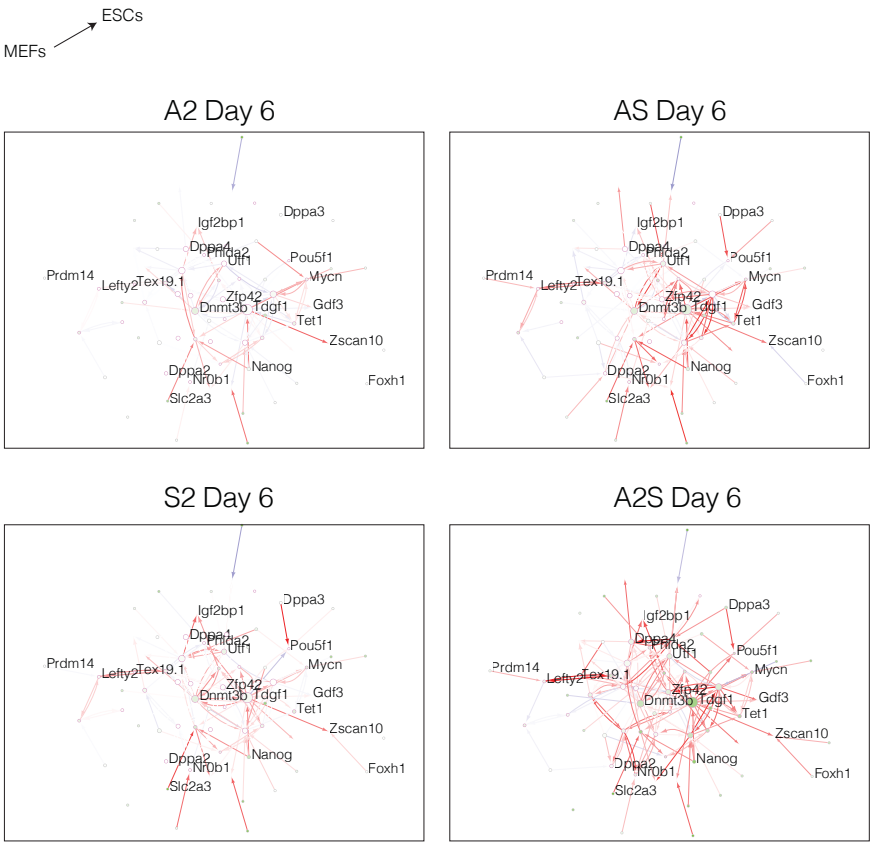


Figure S6 (Related to Figure 6)

A) Additional MERLIN network diagrams for selected pattern

References

Apostolou, E., and Hochedlinger, K. (2013). Chromatin dynamics during cellular reprogramming. Nature *502*, 462–471.

Apostolou, E., and Stadtfeld, M. (2018). Cellular trajectories and molecular mechanisms of iPSC reprogramming. Curr. Opin. Genet. Dev. *52*, 77–85.

Banito, A., Rashid, S.T., Acosta, J.C., Li, S., Pereira, C.F., Geti, I., Pinho, S., Silva, J.C., Azuara, V., Walsh, M., et al. (2009). Senescence impairs successful reprogramming to pluripotent stem cells. Genes Dev. *23*, 2134–2139.

Bar-Nur, O., Brumbaugh, J., Verheul, C., Apostolou, E., Pruteanu-Malinici, I., Walsh, R.M., Ramaswamy, S., and Hochedlinger, K. (2014). Small molecules facilitate rapid and synchronous iPSC generation. Nat. Methods *11*, 1170– 1176.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature *463*, 1042–1047.

Blaschke, K., Ebata, K.T., Karimi, M.M., Zepeda-Martı´nez, J.A., Goyal, P., Mahapatra, S., Tam, A., Laird, D.J., Hirst, M., Rao, A., et al. (2013). Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. Nature *500*, 222–226.

Brambrink, T., Foreman, R., Welstead, G.G., Lengner, C.J., Wernig, M., Suh, H., and Jaenisch, R. (2008). Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. Cell Stem Cell *2*, 151–159.

Brumbaugh, J., Di Stefano, B., Wang, X., Borkent, M., Forouzmand, E., Clowers, K.J., Ji, F., Schwarz, B.A., Kalocsay, M., Elledge, S.J., et al. (2018). Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. Cell *172*, 106–120.e21.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222.

Buganim, Y., Faddah, D.A., and Jaenisch, R. (2013). Mechanisms and models of somatic cell reprogramming. Nat. Rev. Genet. *14*, 427–439.

Chasman, D., Walters, K.B., Lopes, T.J.S., Eisfeld, A.J., Kawaoka, Y., and Roy, S. (2016). Integrating Transcriptomic and Proteomic Data Using Predictive

Regulatory Network Models of Host Response to Pathogens. PLoS Comput. Biol. *12*, e1005013.

de Hoon, M.J.L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453–1454.

Esteban, M.A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S., et al. (2010). Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. Cell Stem Cell *6*, 71–79.

Golipour, A., David, L., Liu, Y., Jayakumaran, G., Hirsch, C.L., Trcka, D., and Wrana, J.L. (2012). A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. Cell Stem Cell *11*, 769–782.

Guo, S., Zi, X., Schulz, V.P., Cheng, J., Zhong, M., Koochaki, S.H.J., Megyola, C.M., Pan, X., Heydari, K., Weissman, S.M., et al. (2014). Nonstochastic reprogramming from a privileged somatic cell state. Cell *156*, 649–662.

Guo, L., Lin, L., Wang, X., Gao, M., Cao, S., Mai, Y., Wu, F., Kuang, J., Liu, H., Yang, J., et al. (2019). Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. Mol. Cell *73*, 815–829.e7.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595–601.

Hore, T.A., von Meyenn, F., Ravichandran, M., Bachman, M., Ficz, G., Oxley, D., Santos, F., Balasubramanian, S., Jurkowski, T.P., and Reik, W. (2016). Retinol and ascorbate drive erasure of epigenetic memory and enhance reprogramming to naive pluripotency by complementary mechanisms. Proc. Natl. Acad. Sci. USA *113*, 12202–12207.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A.E., and Melton, D.A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat. Biotechnol. *26*, 795–797.

Hussein, S.M.I., Puri, M.C., Tonge, P.D., Benevento, M., Corso, A.J., Clancy, J.L., Mosbergen, R., Li, M., Lee, D.-S., Cloonan, N., et al. (2014). Genomewide characterization of the routes to pluripotency. Nature *516*, 198–206.

Ichida, J.K., Blanchard, J., Lam, K., Son, E.Y., Chung, J.E., Egli, D., Loh, K.M., Carter, A.C., Di Giorgio, F.P., Koszka, K., et al. (2009). A small-molecule inhibitor of tgf-Beta signaling replaces sox2 in reprogramming by inducing nanog. Cell Stem Cell *5*, 491–503.

Ichida, J.K., Tcw, J., Williams, L.A., Carter, A.C., Shi, Y., Moura, M.T., Ziller, M., Singh, S., Amabile, G., Bock, C., Umezawa, A., Rubin, L.L., Bradner, J.E., Akutsu, H.,

Meissner, A., and Eggan, K. (2014). Notch inhibition allows oncogene-independent generation of iPS cells. Nat. Chem. Biol. *10*, 632–639.

Jackson, S.A., Olufs, Z.P.G., Tran, K.A., Zaidan, N.Z., and Sridharan, R. (2016). Alternative Routes to Induced Pluripotent Stem Cells Revealed by Reprogramming of the Neural Lineage. Stem Cell Reports *6*, 302–311.

Kim, D.H., Marinov, G.K., Pepke, S., Singer, Z.S., He, P., Williams, B., Schroth, G.P., Elowitz, M.B., and Wold, B.J. (2015). Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. Cell Stem Cell *16*, 88–101.

Li, H., Collado, M., Villasante, A., Strati, K., Ortega, S., Can˜ amero, M., Blasco, M.A., and Serrano, M. (2009). The Ink4/Arf locus is a barrier for iPS cell reprogramming. Nature *460*, 1136–1139.

Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., et al. (2010). A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell *7*, 51–63.

Liang, G., He, J., and Zhang, Y. (2012). Kdm2b promotes induced pluripotent stem cell generation by facilitating gene activation early in reprogramming. Nat. Cell Biol. *14*, 457–466.

Lujan, E., Zunder, E.R., Ng, Y.H., Goronzy, I.N., Nolan, G.P., and Wernig, M. (2015). Early reprogramming regulators identified by prospective isolation and mass cytometry. Nature *521*, 352–356.

Maherali, N., and Hochedlinger, K. (2009). Tgf beta signal inhibition cooperates in the induction of iPSCs and replaces Sox2 and cMyc. Curr. Biol. *19*, 1718– 1723.

Mario´ n, R.M., Strati, K., Li, H., Murga, M., Blanco, R., Ortega, S., FernandezCapetillo, O., Serrano, M., and Blasco, M.A. (2009). A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. Nature *460*, 1149–1153.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A., and Stunnenberg, H.G. (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590–604.

Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature *454*, 49–55.

Modelska, A., Turro, E., Russell, R., Beaton, J., Sbarrato, T., Spriggs, K., Miller, J., Gra¨ f, S., Provenzano, E., Blows, F., et al. (2015). The malignant phenotype in breast cancer is driven by eIF4A1-mediated changes in the translational landscape. Cell Death Dis. *6*, e1603–e1612.

Mor, N., Rais, Y., Sheban, D., Peles, S., Aguilera-Castrejon, A., Zviran, A., Elinger, D., Viukov, S., Geula, S., Krupalnik, V., et al. (2018). Neutralizing Gatad2a-Chd4-Mbd3/NuRD Complex Facilitates Deterministic Induction of Naive Pluripotency. Cell Stem Cell *23*, 412–425.e10.

Nefzger, C.M., Rossello, F.J., Chen, J., Liu, X., Knaupp, A.S., Firas, J., Paynter, J.M., Pflueger, J., Buckberry, S., Lim, S.M., et al. (2017). Cell Type of Origin Dictates the Route to Pluripotency. Cell Rep. *21*, 2649–2660.

O'Malley, J., Skylaki, S., Iwabuchi, K.A., Chantzoura, E., Ruetz, T., Johnsson, A., Tomlinson, S.R., Linnarsson, S., and Kaji, K. (2013). High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. Nature *499*, 88–91.

Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Marcarci, B.O., Unternaehrer, J., Gupta, P.B., et al. (2012). Chromatin-modifying enzymes as modulators of reprogramming. Nature *483*, 598–602.

Papp, B., and Plath, K. (2013). Epigenetics of reprogramming to induced pluripotency. Cell *152*, 1324–1343.

Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., et al. (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. Cell *151*, 1617– 1632.

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

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

Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A.A., Caspi, I., Krupalnik, V., Zerbib, M., et al. (2013). Deterministic direct reprogramming of somatic cells to pluripotency. Nature *502*, 65–70.

Ruiz, S., Panopoulos, A.D., Herrerı´as, A., Bissig, K.-D., Lutz, M., Berggren, W.T., Verma, I.M., and Izpisua Belmonte, J.C. (2011). A high proliferation rate is required for cell reprogramming and maintenance of human embryonic stem cell identity. Curr. Biol. *21*, 45–52.

Salas, M., John, R., Saxena, A., Barton, S., Frank, D., Fitzpatrick, G., Higgins, M.J., and Tycko, B. (2004). Placental growth retardation due to loss of imprinting of Phlda2. Mech. Dev. *121*, 1199–1210.

Saldanha, A.J. (2004). Java Treeview–extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.-K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell *7*, 64–77.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. Cell *176*, 928–943.e22.

Schwarz, B.A., Cetinbas, M., Clement, K., Walsh, R.M., Cheloufi, S., Gu, H., Langkabel, J., Kamiya, A., Schorle, H., Meissner, A., et al. (2018). Prospective Isolation of Poised iPSC Intermediates Reveals Principles of Cellular Reprogramming. Cell Stem Cell *23*, 289–305.e5.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Shi, Y., Desponts, C., Do, J.T., Hahm, H.S., Scho¨ ler, H.R., and Ding, S. (2008). Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. Cell Stem Cell *3*, 568–574.

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T.W., and Smith, A. (2008). Promotion of reprogramming to ground state pluripotency by signal inhibition. PLoS Biol. *6*, e253.

Sridharan, R., Tchieu, J., Mason, M.J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. Cell *136*, 364–377.

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M., et al. (2013). Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1g in reprogramming to pluripotency. Nat. Cell Biol. *15*, 872–882.

Stadtfeld, M., and Hochedlinger, K. (2010). Induced pluripotency: history, mechanisms, and applications. Genes Dev. *24*, 2239–2263.

Stadtfeld, M., Maherali, N., Breault, D.T., and Hochedlinger, K. (2008). Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. Cell Stem Cell *2*, 230–240.

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

Tran, K.A., Jackson, S.A., Olufs, Z.P.G., Zaidan, N.Z., Leng, N., Kendziorski, C., Roy, S., and Sridharan, R. (2015). Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. Nat. Commun. *6*, 6188.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Utikal, J., Polo, J.M., Stadtfeld, M., Maherali, N., Kulalert, W., Walsh, R.M., Khalil, A., Rheinwald, J.G., and Hochedlinger, K. (2009). Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. Nature *460*, 1145–1148.

Vidal, S.E., Amlani, B., Chen, T., Tsirigos, A., and Stadtfeld, M. (2014). Combinatorial modulation of signaling pathways reveals cell-type-specific requirements for highly efficient and synchronous iPSC reprogramming. Stem Cell Reports *3*, 574–584.

Williams-Hill, D.M., Duncan, R.F., Nielsen, P.J., and Tahara, S.M. (1997). Differential expression of the murine eukaryotic translation initiation factor isogenes eIF4A(I) and eIF4A(II) is dependent upon cellular growth status. Arch. Biochem. Biophys. *338*, 111–120.

Wolfe, A.L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V.K., Sanghvi, V.R., Mavrakis, K.J., Jiang, M., Roderick, J.E., Van der Meulen, J., et al. (2014). RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature *513*, 65–70.

Zhao, X.-Y., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C.-L., Ma, Q.-W., Wang, L., Zeng, F., and Zhou, Q. (2009). iPS cells produce viable mice through tetraploid complementation. Nature *461*, 86–90.

Zhao, T., Fu, Y., Zhu, J., Liu, Y., Zhang, Q., Yi, Z., Chen, S., Jiao, Z., Xu, X., Xu, J., et al. (2018). Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. Cell Stem Cell *23*, 31–45.e7.

Zhou, Z., Yang, X., He, J., Liu, J., Wu, F., Yu, S., Liu, Y., Lin, R., Liu, H., Cui, Y., et al. (2017). Kdm2b Regulates Somatic Reprogramming through Variant PRC1 Complex-Dependent Function. Cell Rep. *21*, 2160–2170.

Zunder, E.R., Lujan, E., Goltsev, Y., Wernig, M., and Nolan, G.P. (2015). A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. Cell Stem Cell *16*, 323–337.

Chapter 3

Chromatin dynamics regulate somatic cell reprogramming to pluripotency

The work presented in this chapter is in preparation for publication.

Contributions: R. Sridharan designed project, and S. Pietrzak performed all experiments and data analysis with ArchR. S. Roy, S. Halberg-Spencer, and S. Zhang developed and performed analysis with scCISINT, and S. Halberg-Spencer generated all MST and PAGA graphs. All writing prepared by S. Pietrzak with critical review from R. Sridharan

Abstract

In multicellular organisms, somatic cell identity is determined during development and remains resilient to change unless challenged by disease or injury. In vitro, the ectopic expression of master transcription factors can reprogram cell fate of somatic cells even to a pluripotent embryonic stem cell-like state. This is remarkable because the properties of pluripotency are enabled by a poised chromatin state that can differentiate into any cell type**.** Here, we aimed to elucidate the chromatin-associated changes characteristic of successful reprogramming to induced pluripotent stem cells by single-cell chromatin accessibility profiling. We find that motifs for somatic transcription factors such as AP1 and RUNX gradually lose enrichment in accessible regions as those for pluripotency factors OCT4 and SOX2 become prevalent. Upon withdrawal of ectopic reprogramming factors, we uncover additional changes prior to pluripotency acquisition, including strong transient upregulation of *Tcfap2c* and its concomitant motif, which we validate to be essential in stabilization and maintenance of iPSCs. We uncover a key role for enrichment of binding sites for 3D chromatin organization-associated factors (KLF4, MAZ, PATZ1) in reprogramming success. Our results motivated the development of a versatile computational algorithm scCISINT to predict 3D interactions among differentially-accessible loci, both with a single putative enhancer controlling several genes and a single gene controlling several enhancers. Using scCISINT, we validate that a 1) TCFAP2 motif-containing peak can promote maintenance of the pluripotent state by activating the mir290 cluster, 2) a ubiquitously-accessible region that temporally controls reprogramming efficiency which is inhibitory early, but late-enhancing, and 3) a RUNX motif-containing somatic accessible region

anti-correlated with Nanog promoter opening that acts as a barrier to reprogramming. Taken together, we have uncovered key chromatin changes necessary for pluripotency acquisition and maintenance and has led to identification of loci whose 3D interactions influence reprogramming efficacy.

Introduction

Pluripotent stem cells (PSCs) are characterized by their ability to self-renew indefinitely as well as the potential to differentiate into any cell type given the right stimulus. In addition to embryonic stem cells (ESCs), that are derived by culturing cells from the inner cell mass of the blastocyst, PSCs can be obtained through the reprogramming of somatic cells, such as mouse embryonic fibroblasts (MEFs), into a pluripotent state, producing induced pluripotent stem cells (iPSCs) that are the functional equivalents of ESCs (Boland et al., 2009; Okita et al., 2007; X. Zhao et al., 2009). This remarkable facility to alter cell fate was first achieved through the ectopic expression of the Yamanaka factors OCT4, SOX2, KLF4, and MYC (OSKM) (Takahashi & Yamanaka, 2006).

One of the features that differentiates somatic from pluripotent cells is in their epigenome. ESCs have a more open chromatin structure compared to somatic cells, which have more compacted heterochromatic regions (Gaspar-Maia et al., 2011). For example, ESCs are less enriched for the repressive H3K9me2/3 histone modifications (Sridharan et al., 2013), which have been implicated as a barrier to reprogramming (J. Chen et al., 2013; Soufi et al., 2012; Tran et al., 2015). MEFs are also more enriched for H3K79me2/3, a mark of active transcription (Sridharan et al., 2013), and has been shown to be antagonistic to reprogramming. Initial studies proposed that its presence at MEF-associated genes is inhibitory for the MET (mesenchymal-to-epithelial transition) (Onder et al., 2012). However, our lab has recently unveiled roles for H3K79me2/3 and its concomitant methyltransferase Dot1l outside of MET. It was found to aberrantly upregulate expression of reprogramming-associated genes (e.g. *Nfix*) (Wille &

Sridharan, 2022), and mars the transition to ESC-like levels of H3K9 acetylation and transcriptional elongation along with proper RNAPII distribution, thus maintaining the somatic identity (Wille, Neumann, et al., 2023; Wille, Zhang, et al., 2023).

Consequently, we and others have found that supplementing reprogramming culture media with epigenetic-modifying small molecules can enhance reprogramming efficiency (J. Chen et al., 2011; Esteban et al., 2010; Huangfu, Maehr, et al., 2008; Huangfu, Osafune, et al., 2008; Jackson et al., 2016; Mikkelsen et al., 2008; Onder et al., 2012; Shi et al., 2008; Tran et al., 2015, 2019). In some studies, small molecule combinations were even shown to be able to replace the exogenous reprogramming factors to generate iPSCs (Cao et al., 2018; X. Chen et al., 2023; Guan et al., 2022; Hou et al., 2013; Ye et al., 2016; Y. Zhao et al., 2015)(Hou et al., 2013; Zhao et al., 2015; Ye et al., 2016; Cao et al., 2018; Guan et al., 2022; Chen et al., 2023). In our lab, we previously established a high-efficiency OSKM reprogramming system in MEFs wherein combining ascorbic acid (AA) - which enhances the catalytic activity of H3K9me2 demethylases (J. Chen et al., 2013; Tran et al., 2015) - , two signaling pathway inhibitors (2i), and the Dot1l inhibitor SGC0946 (A2S) enhanced reprogramming efficiency to over 40% from 3% (Tran et al., 2019).

Given the changing epigenetic landscape as cells undergo reprogramming, several studies have implemented the assay for transposase-accessible chromatin with sequencing (ATAC-seq) (Buenrostro et al., 2013) to probe accessible regions of chromatin in reprogramming populations to better understand the chromatin dynamics associated with this process. ATAC-seq of bulk populations of reprogramming cells have identified genomic regions that transition from an open-to-close or close-to-open

configuration with a major early wave of opening closed sites and second late wave of closing open sites (Li et al., 2017). Disruptions to the balance of shifting chromatin structure act as barriers to reprogramming efficiency (e.g. c-JUN, Fra1 expression) (Chronis et al., 2017; Li et al., 2017). Reprogramming-refractory cells retain accessibility at MEF enhancers (Knaupp et al., 2017). The information gleaned from ATAC-seq combined with analysis of OSK binding dynamics form ChIP-seq assays has led to further speculation into how over-expression of pluripotency factors can lead to closing of somatic enhancers. Some posit that OSK bind already open somatic regions, leading to somatic transcription factor (TF) redistribution (Chronis et al., 2017) or they open transient sites that sequester somatic transcription factors (TFs) (Knaupp et al., 2017). In addition to these activities, KLF4 was also discovered to be important for the rewiring of the 3D organization of chromatin and in promoting enhancer-promoter loop interactions (Di Giammartino et al., 2019), as well as mediating the mesenchymal-to-epithelial transition (MET) upon loosening of chromatin by OCT4 (K. Chen et al., 2020).

ATAC-seq on non-fibroblast or non-Yamanaka factor-based reprogramming systems has highlighted the different chromatin accessibility pathways that can lead to iPSCs. For example, two different chemical reprogramming systems showed that loci that gain accessibility are enriched in the TF motifs for the extraembryonic endoderm (XEN)-associated GATA and FOX TFs (Cao et al., 2018; X. Chen et al., 2023). A system that used seven factors to reprogram (JDP2, JHDM1B, MKK6, GLIS1, NANOG, ESRRB, and SALL4) found that this combination differs from OSKM in distribution of motifs at changing loci (e.g. earlier accessibility of ESRRB motifs) (B. Wang et al., 2019). In B cells, pulsing exposure to C/EBP-alpha can make them more elite and

primed to reprogram. ATAC-seq analysis of these cells found that they opened up over 500 sites that are also accessible in ESCs and were enriched for Klf4 binding sites, suggesting a cooperative relationship between C/EBP-alpha and Klf4 (Di Stefano et al., 2016).

One of the primary setbacks associated with reprogramming is that the process is slow and inefficient due to inherent cell-to-cell variability in reprogramming kinetics (Apostolou & Hochedlinger, 2013; Buganim et al., 2013; Papp & Plath, 2013). Cells may also branch off the trajectory of successful reprogramming through transitioning into alternative non-iPSC cell types (Parenti et al., 2016) or forming partially-reprogrammed stalled intermediates (Mikkelsen et al., 2008; Sridharan et al., 2009, 2013). These factors contribute to a largely heterogeneous population of reprogramming cells, making ensemble approaches to studying this process challenging, as changes associated with truly reprogramming cells are obscured by those that fail to do so. Some of the aforementioned ATAC-seq studies have tried to overcome these challenges through the use of small molecules in reprogramming media to enhance efficiency. For example, studies have used iCD1 media containing ascorbic acid and the kinase inhibitor CHIR99021 (one of the 2i inhibitors) (Li et al., 2017; B. Wang et al., 2019). Others have attempted to bypass the heterogeneity of reprogramming by analyzing ATAC-seq data on clonal partially reprogrammed iPSC (pre-iPSCs) cell lines (Chronis et al., 2017), or through sorting cells based on somatic or pluripotency-associated surface markers (Di Giammartino et al., 2019; Knaupp et al., 2017).

The recent advent of single-cell technologies, including single-cell ATAC-seq (scATAC-seq), has allowed the bypassing of the heterogeneity issue by identifying

populations of cells at comparable stages of reprogramming (or those that have branched off from the path), regardless of timepoint. As scATAC-seq is a relatively new technology, few studies have been performed applying it to reprogramming cells and all of which have been in human reprogramming systems. These studies found that in human fibroblasts, the successfully reprogramming cells transition from a FOSL1 to TEAD4-centered regulatory network (Xing et al., 2020), and also display transient opening of loci by OSK, providing further data in support of redistribution of somatic TFs at transient loci (Nair et al., 2023). In optimized chemical reprogramming of human adipose-derived stromal cells (hADSCs), cells avoid upregulation of XEN-associated sites compared to other systems (Liuyang et al., 2023). However, no study has yet performed scATAC-seq in any context of mouse reprogramming.

Here, we have applied scATAC-seq analysis to both low and high-efficiency reprogramming of MEFs using our previously described FBS and A2S systems (Tran et al., 2019), in which the combination of ascorbic acid, H3K79 methyltransferase SGC0946, and 2i improved reprogramming efficiency from ~3% to 42%. Analysis of scATAC-seq data revealed greater suppression of MEF-associated motifs and enrichment of pluripotency-associated motifs in A2S cells. Interestingly, A2S cells are also enriched for motifs of transcription factors associated with 3D chromatin organization (KLF4, MAZ, PATZ1), indicating a role for A2S in more efficiently rewiring the chromatin landscape to be more pluripotent-like. Upon withdrawal of ectopic OSKM, cells undergo even more changes, including transient accessibility of the transcription factor *Tcfap2c* gene and its associated motif, which plays a role in the transition to bona fide iPSCs, as well as enrichment of the somatic Tead4 motif in cells that are reverting

to a MEF-like state. The upregulation of 3D chromatin organization-associated motifs motivated us to delve further into the 3D interactions of scATAC-seq peaks. In a collaborative effort with computational biologist Prof. Sushmita Roy, we identified putative long-range interactions of novel enhancer peaks with strong gene interactions using a method called scCISINT. Using CRISPR-interference, we validated that an interaction of a TCFAP2 motif-containing peak in regulating a microRNA cluster that is essential for completing the transition to iPSCs. scCISINT also identified enhancer hubs that sequentially interact with somatic or pluripotency genes. We further validated the temporal activity of this putative enhancer, as reprogramming was either enhanced or suppressed upon repression of the peak either at the start or midpoint of reprogramming, respectively. This analysis has uncovered a key role for 3D chromatin modifiers in the transition to iPSCs, which is bolstered through the addition of epigenetic-modifying small molecules, leading to the discovery of novel enhancer regions that facilitate necessary transcriptional regulation in pluripotency acquisition.

Results

scATAC-seq reveals chromatin accessibility dynamics at promoter-distal regions

To investigate chromatin accessibility dynamics, we implemented single-cell ATAC-seq (scATAC-seq) on reprogramming populations. Reprogramming was performed using MEFs that contain a dox-inducible cassette with the four Yamanaka factors of OCT4, SOX2, KLF4 and MYC (OSKM) in our previously described low-efficiency FBS and high-efficiency A2S systems (Tran et al., 2019) (Fig. S1A). Cells were harvested for scATAC-seq at timepoints that match those that we have previously

profiled with scRNA-seq (FBS Days 3, 6, 9, 12; A2S Days 2, 4, and 6) (Fig. 1A). Along with unaltered starting MEFs, mouse ESCs, and iPSCs that were generated from A2S reprogramming (Tran, et al., 2019), all cell populations underwent lysis, transposition, and sequencing using the droplet-based method from 10X Genomics, averaging ~4900 cells sequenced per sample (Fig. S1B, Methods).

To analyze our sequencing data, we implemented the ArchR algorithm (Granja et al., 2021) (Fig. S1C-D). ArchR clustered cells based on the similarity in their accessibility profiles, and visualized them in 2D space via uniform manifold approximation and projection (UMAP) (Fig. 1B-D, S1E-F). The starting MEF population is found in two clusters (C6 and C4). Interestingly, cells from the earliest time point in both systems (FBS Day 3 and A2S Day 2) cluster together in C2 and C10, with FBS Day 3 making up 59% and 37% and A2S Day 2 making up 24% and 55% of C2 and C10, respectively (Fig S1E). While the proportion of FBS Day 3 and A2S Day 2 were slightly different in these two clusters, they both contained the majority of both of these samples (88%-90%) (Fig. 1D, S1E). Once the cells progress past the early timepoint, the cells from the two different reprogramming systems diverge and populate different clusters, with FBS cells primarily found in C3, C11, and C9 and A2S cells in C8 and C7 (Fig. 1D, S1E). Each of the FBS clusters contains cells from the three timepoints of Days 6, 9, and 12, again illustrating the heterogeneity associated with low-efficiency reprogramming. Conversely, the A2S samples largely separate based on timepoint in their clusters (85% of Day 4 cells in C8, 68% of Day 6 cells in C7) (Fig. 1D, S1E), highlighting that the acceleration of reprogramming by small molecules causes accessibility changes to occur in a coordinated manner at each timepoint.

Transcriptomic-based analyses provide a snapshot of the gene expression landscape of a cell population. This information can also be captured by analyzing chromatin accessibility, as accessibility at a gene promoter typically corresponds to a gene that is on and being actively transcribed. However, ATAC-seq analysis provides greater information regarding non-gene associated loci (Fig. 1E). In ArchR, a marker gene analysis identified 2,134 genes whose promoter regions are differentially accessible across all clusters (Fig. 1F). A marker peak analysis, which identifies differentially accessible regions regardless of their location in the genome, identified over 50-fold more peaks (110,871 peaks), the majority of which are located at intergenic or intronic regions and ranging from 72% to 95% of each cluster's marker peaks (with the exception of C18 and C19) (Fig. 1F). Thus, scATAC-seq provides information on reprogramming dynamics beyond a gene-centric view.

To ascertain whether the scATAC-seq data indeed matched gene expression data, we looked at the accessibility of key MEF, pluripotency, and reprogramming-associated genes in ArchR, matched to our previously published scRNA-seq data (Tran et al., 2019). As expected, MEF-associated genes (*Zeb1* and *Snai1*) were more accessible in the MEF clusters, while pluripotency genes (*Nanog* and *Sall4*) had greater accessibility in advanced reprogramming cells and ESCs/iPSCs (Fig. 1G). Additionally, the MET marker *Cdh1* becomes more accessible and *Ehf* displays a transient opening at its promoter, in line with its transient expression previously observed in scRNA-seq analysis. Examining the integrated scRNA-seq data, we observed that the gene expression profile of each marker gene generally matched the accessibility profile.

Therefore, our scATAC-seq data successfully recapitulated the dynamics of known marker genes across reprogramming.

A2S Enhances Accessibility at Key Transcription Factor Binding Sites

After validating our scATAC-seq methodology and analysis, we next sought to investigate the differences in accessible peaks that drive reprogramming in low- and high-efficiency systems. To further investigate the relationship between each of the clusters, we generated a minimum spanning tree (MST) (Fig. 2A). Briefly, MSTs are a graphical representation showing the minimum possible of directionless weighted edge connections between vertices (clusters) without forming loops. The MST represents a subset of edges from the more complex graph created with partition-based graphic abstraction (PAGA) (Wolf et al., 2019) (Fig. S2A) with the calculated edge weights between all cluster pairs. Based on the clusters, we know that the earliest time points after dox-induction from both FBS and A2S reside in C2 and C10 (Fig. 1B-D). Surprisingly, the closest connection between reprogramming cells and MEFs is between MEF cluster C4 and C3, which primarily contains FBS Day 6 (41% of C3), along with Days 9 and 12 (20% and 24% of C2, respectively) (Fig. 2A). After the C2 connection to C10, the other “early” cluster, the tree branches off either toward the FBS (C11) or A2S (C8) reprogramming clusters (Fig. 2A). The MST therefore indicates that the initial induction of reprogramming by OSKM drives drastic changes in the chromatin accessibility landscape, regardless of reprogramming system. However, the low-efficiency FBS reprogramming system has a large contingent of cells that at later

timepoints following these early changes, gain an accessibility profile that is similar to MEFs, which is largely bypassed in A2S reprogramming.

We further investigated the specific gene promoter accessibility and transcription factor motifs that were enriched in the differentially-accessible regions that informed the cell clusters. From our prior analysis of scRNA-seq data (Tran et al., 2019), we have previously found that at the transcriptome level, MEFs will separate according to their cycling status as indicated by the expression of cell cycle-associated genes, such as *Bub1b* and *Mcm5*. However, the accessibility of these genes and other cycle-related genes at their promoters did not differ between the two MEF clusters (Fig. S2B). Previous studies have shown that among a population of MEFs, there are some that are more elite and primed to reprogram more successfully (Jain et al., 2023). We investigated whether the MEF clusters represented a difference in accessibility between gene promoters that support elite cells. Genes such as *Spp1* were less accessible in primed cells (more accessible in C4). For the genes that are higher in primed cells, such as *Top2a*, the differences were less apparent, though there was a slight enrichment of *Top2a* in C6, indicating that the C6 MEFs may reprogram more successfully (Fig. S2B). It has also been posited that *Wnt1*-expressing MEFs derived from the neural crest are also more poised for reprogramming (Shakiba et al., 2019), however, *Wnt1* accessibility shows no difference between the two MEF populations (Fig. S2B).

To identify the differentiating factors between the MEF clusters based on our scATAC-seq data, we implemented a motif enrichment analysis in ArchR to find transcription factor motifs that are present in the peaks of one cluster both globally among all samples and in a pairwise fashion against the other MEF cluster. Both MEF

clusters share enrichment of key somatic motifs, such as those of AP1 TFs (Fig. 2C). Where they differ is in the enrichment of different families of differentiation and development-associated TFs. For example, C4 has greater enrichment of several HOX TF motifs that are key in posterior development, such as limb morphogenesis and reproductive tract development (e.g. *Hoxa13*, *Hoxd13*, *Hoxa11*) (Hubert & Wellik, 2023) (Fig. 2B, S2C). The uniquely enriched motifs in C6 include the GATA family (*Gata1-6*), which play roles in endodermal, ectodermal, and mesodermal differentiation, including hematopoiesis (Lentjes et al., 2016). Also enriched are the MEIS (*Meis1-3*) TF motifs, which form complexes with PBX TFs (also enriched in C6) and have been implicated in the differentiation of various organismal systems, such as neuronal development (Schulte & Geerts, 2019) (Fig. 2B, S2C). These results suggest that the main difference in the MEF clusters lies in the role that each group of MEFs would have potentially played in the developing embryo at the time of MEF acquisition. To ensure our split MEF clusters was actually from endothelial cell contamination at time of harvest, I examined promoter accessibility at *Col1a1*, which is highly expressed in fibroblast but not endothelial tissue. This gene displayed strong and comparable accessibility scores in both MEF clusters, ensuring that all of our starting cells were indeed MEFs (data not shown).

This same global and pairwise analysis of enriched motifs was applied to the reprogramming cell clusters to identify trends in motif enrichment in both FBS and A2S systems (Fig. 2C). Based on the MST results, we hypothesized that FBS cluster C3 are a group of cells that become orthogonal to reprogramming and re-acquire a more MEF-like accessibility. We found that compared to early cells (C2) and the other FBS clusters

(C11 and C9), these cells were enriched for motifs that were also shared with the MEF clusters, such as the cytokine signaling-associated STAT TFs, which can regulate proliferation and apoptosis (Awasthi et al., 2021) (Fig. 2C). Relative to the MEF clusters, C3 still has some enrichment of pluripotency motifs, such as POU5F1 (OCT4) and SOX2, and has not fully opened other somatic motifs, such as FOS and JUN, causing them to still remain separated from the MEF clusters. Thus, cells of C3 represent a large population of cells in low-efficiency reprogramming that have stalled and branched away from the successful reprogramming track.

We also observed some expected trends in motif enrichment, such as a gradient of decreasing somatic motifs (e.g. FOS, JUN, TWIST1) and increasing pluripotency motif enrichment (e.g. POU5F1, SOX2, KLF4) going from early cluster C2 to C10, and then along each reprogramming system's trajectory (FBS: C10 → C11 → C9; A2S: C10 → C8 → C7) (Fig. 2C, S2D). Between A2S clusters C8 and C7, the differences in POU5F1 and SOX2 motif enrichment were almost negligible, indicating that in A2S, most of the cells have acquired accessibility at POU5F1 and SOX2 motif-enriched locations by day 4. Applying the pairwise motif analysis between the FBS and A2S intermediate and most advanced clusters (C7 vs C9 and C8 vs C11), we observe that both FBS clusters still retain accessibility at somatic motifs, including AP1 motifs, against the corresponding A2S cluster (Fig. 2C). Specifically at the overlapping marker peaks between A2S (C7) and FBS (C9) advanced reprogramming clusters, the topmost enriched motifs among the shared peaks were KLF4 ($p = 1e\text{-}17$) and POU5F1 ($p = 1e\text{-}16$). This indicates that the pluripotency factor motif sites that are opening are the same in both conditions.

The A2S clusters have stronger motif enrichment for the reprogramming factor KLF4 over FBS cells (Fig. 2C). KLF4 binding has previously been shown to play an important role in rewiring the 3D organization of chromatin and enhancer loop formation in MEF reprogramming to iPSCs (Di Giammartino et al., 2019). In line with this, the most advanced A2S cluster C7 is also enriched for the motifs of MAZ and PATZ1 (Fig. 2C). MAZ is a co-factor of CTCF that regulates 3D chromatin organization and insulates active chromatin regions from repressive ones (Ortabozkoyun et al., 2022). PATZ1 is a chromatin remodeler that regulates gene transcription and is an important factor in maintenance of pluripotency in ESCs (Fedele et al., 2017; Rong et al., 2013). PATZ1 has also been implicated to play a role in reprogramming in a dose-dependent manner (Ma et al., 2014). Along with these, the advanced A2S cluster C7 also begins to become enriched for the pluripotency marker ESRRB motif (Fig. 2C). These results indicate that A2S reprogramming is able to surpass FBS in part because the addition of these small molecules creates a chromatin state that is more permissible for binding of proteins that can rearrange the 3D structure of chromatin to that of a pluripotent state.

We then parsed out the differences between each of three components of A2S. For this, we performed scATAC-seq on Day 4 and Day 6 reprogramming populations using each dual combination of the A2S molecules where one of them was dropped (A2, AS, and S2), and clustered these with the A2S, MEF, and ESC/iPSC cells (Fig. S2E-G). We have previously shown that S2 is more efficient than AS, followed by A2 (~25%, 10%, and 6% efficient, respectively) (Fig. S1A). We did not see any clear separation between individual cells of the dual combination or A2S samples by clustering. We observe three advanced reprogramming populations, C13, C16, and C17

in this combined analysis (Fig. S2E). The majority of C16 is comprised of A2S Day 6 cells (53%), followed by A2, S2, and AS Day 6 (21%, 17%, and 4%, respectively). C13 has a more balanced mix of all 4 samples (ranging from 14% - 32%), while the primary sample in C17 came from A2 Day 6 (42.5%) (Fig. S2G). An MST analysis of these clusters showed that there was a strong connection between the A2S-dominated C16 cluster and the ESC/iPSC clusters (Fig. S2F). Pairwise motif analysis also showed that these were enriched for the late reprogramming chromatin organization-associated KLF4, MAZ, PATZ1, and CTCF motifs compared to C13 and C17, indicating that these represent the cells that are primed to transition to iPSCs, with A2S cells more able to achieve this state.

We clustered each dual combination sample separately with A2S and FBS cells to clearly compare the advantage conferred by each dual combination compared to the low- and high-efficiency reprogramming systems (Fig. 2D). Similar to the FBS+A2S clusters in Figure 1, each of these analyses had their own respective clusters for early timepoints, FBS Days 6, 9, and 12, A2S Day 4, and A2S Day 6. Notably, in the AS analysis, 47% of AS Day 4 cells were found in the early timepoint cluster C15, compared to 8.8% for A2 Day 4, and 7.2% for S2 in their respective analyses (Fig. 2D). AS Day 6 samples were split primarily between FBS cluster C8, A2S Day 4 cluster C7, and A2S Day 6 cluster C6 (11.1%, 32.9%, and 27.8% of AS Day 6 cells, respectively). Conversely, A2 and S2 Day 4 and Day 6 consistently clustered together with their respective A2S Day 4 and 6 clusters (Fig. 2D). Thus, there are delayed chromatin changes in AS reprogramming, as indicated by the clustering of AS Day 4 and 6 cells with earlier timepoints, suggesting a role for 2i in pushing MEFs past the initial somatic

chromatin accessibility identity, likely by suppressing the expression of somatic transcription factors.

Overall, our results suggest that A2S surpasses normal FBS reprogramming through promoting enrichment of accessible sites associated with factors that mediate the reorganization of the 3D chromatin architecture.

Enrichment of Specific Motifs Mediates Stabilization of Pluripotent State Upon OSKM Withdrawal

From our analysis in Figure 1 and 2, we observe that the most advanced reprogramming cells in the A2S system begin to open locations with motifs associated with late-stage pluripotency genes (e.g. *Esrrb*) (Fig. 2C). Bona fide iPSCs are those that can remain pluripotent without the expression of exogenously provided OSKM factors with the standard in the field being about 4 days. We wondered how chromatin accessibility is altered upon withdrawal of OSKM at late stages of reprogramming. We therefore performed scATAC-seq on A2S and dual combination reprogramming cells 2 days post-OSKM withdrawal, as this is the timepoint when expression of exogenous OSKM is undetectable at the RNA level (Fig. S3A), to capture cells just as they are about to become bona fide iPSCs. We performed ArchR clustering and analysis on these withdrawal samples with all associated earlier timepoints, MEFs, ESCs, and iPSCs (Fig. 3A-B, S3B). We observed MEFs (C12 and C10) and ESCs/iPSCs (C4 and C3) occupying their own distinct clusters, while reprogramming cells follow a path from early and less advanced to more advanced cells (C8, C14, C13) (Fig. 3A). Remarkably,

we captured new clusters largely comprised of the withdrawal timepoints (C5, C7, C9, and C11) (Fig. 3B, S3B).

From the MST plot of these clusters, the new withdrawal-associated cluster C5 is spatially between the advanced reprogramming cells and PSCs (Fig. 3C). This cluster has a large proportion of A2S Day 6+2 cells (44%) followed by AS, A2, and S2 (20%, 17%, and 8%, respectively) (Fig. 3B, S3B). C9 and C11 lie between the MEF and the early reprogramming clusters with a strong edge connecting C11 and MEF cluster C12 (Fig. 3C). These clusters have a higher proportion of dual combination Day 6+2 cells (21%, 32%, and 22% of A2, AS, and S2, respectively), and about 13% of A2S Day 6+2. The proportions of each dual condition were comparable in both C9 and C11 (within 5% of each other) with the exception of S2 Day 6+2 which was noticeably higher in C9 (29%) than C11 (13%) (Fig. 3B, S3B). Directly beneath C5 lies C6, which is closest to the MEF-adjacent cluster C9 (Fig. 3C). The majority of C6 is composed of cells from A2S (29%) and S2 Day 6+2 (40%) (Fig. 3B, S3B). Thus, upon dox-withdrawal we are able to capture cells that make the final transition to pluripotency or revert back to a MEF-like state.

Cluster 5, when compared to the most advanced reprogramming cluster (C13) and the other withdrawal-associated cluster C6, had much stronger enrichment for the motifs of pluripotency marker ESRRB, as well as those for the previously described 3D chromatin organization TFs KLF4, MAZ, and PATZ1 (Fig. S3C). When compared to C6, POU5F1 and SOX2 motifs are also enriched in C5. These differential enrichments indicate that the cells of C5 reflect those cells that are on the trajectory towards becoming bona fide iPSCs (deemed the “bona fide” cluster), while C6 contains cells that

have not stably upregulated the pluripotent regulatory network, and are reverting towards a somatic state (“reverting” cluster). The reverting nature of this cluster is further illustrated by an increase in motifs for the somatic-associated TEAD factors (e.g. TEAD4) in C6 when compared to C5 (Fig. S3C). When compared to C6, the MEF-adjacent clusters of C9 and C11 have very strong enrichment of somatic motifs, including the AP1 family of TFs (e.g., FOS and JUN), though still not as strong as in the MEF cells. Additionally, in this comparison, C6 still has some residual enrichment of motifs from the late stage clusters, such as MAZ and PATZ1 compared to C9 and C11 (Fig. S3C). Based on these results, this indicates that C9 and C11 are comprised of cells that have reverted to a MEF-like state (“MEF-like” clusters). Interestingly, a pairwise comparison of C9 and C11 revealed that these clusters can be separated based on the same differences in the original MEF clusters, with one cluster (C11) enriched for posterior development HOX motifs (e.g. Hoxd13), and the other cluster (C9) more enriched for GATA and MEIS motifs (Fig. S3D). Thus, it appears that cells that come from either MEF cluster may revert back to a MEF-like state upon withdrawal, though based on the cluster compositions, S2 seems to be more effective at reprogramming the HOX motif-enriched MEFs as there were fewer S2 Day 6+2 cells in C11. Therefore, one group of MEFs, based on these scATAC-seq clusters, is not in fact more elite than the other and predictive of reprogramming success.

In an attempt to find the differences in the dual combinations and determine the role of each A2S component after OSKM withdrawal, we again performed clustering of the A2S reprogramming and withdrawal cells with those of each dual combination separately (Fig. 3D). Each UMAP contained “bona fide” clusters adjacent to the

ESCs/iPSCs, “reverting” clusters just beneath them, and “MEF-like” clusters. We previously observed that AS reprogramming was slower than A2 and S2 (Fig. 2D). In fact, of the Day 6+2 cells for each combination, AS had the highest proportion of withdrawal cells found in the MEF-like cluster (58%) over A2 and S2 (39% and 43%, respectively). Of the A2 Day 6+2 cells, the highest proportion were found in the MEF-like cluster (39%), followed by the reverting (24%) and bona fide cluster (12%) (Fig. 3D). They all had fairly low representation in the bona fide cluster (12%, 17%, and 8% of Day 6+2 cells from A2, AS, and S2, respectively), but there was a decent proportion of withdrawal cells within the reverting clusters for A2 (24%) and S2 (40%) (Fig. 3D). This suggests that cells reprogrammed with 2i are more resistant to complete reversion to a MEF-like state, however it is the combination of the epigenome-modifying chemicals ascorbic acid and SGC0946 that are required later for maintenance and stabilization of the pluripotent state.

From our motif enrichment analyses, one striking discovery we found was that some of the highest ranking enriched motifs in C5 (4 of the top 6 motifs) come from the TCFAP2 family of TFs, with TCFAP2C being the highest ranking. TCFAP2C displays similar promoter accessibility and gene expression patterns as its motif (Fig. 3E). These motifs were also highly enriched in the reverting C6 cluster, but interestingly, had little to no presence in any other clusters, including the preceding late stage reprogramming clusters or the subsequent PSC clusters (Fig. 3E). We further examined the role of TCFAP2C in reprogramming via shRNA knockdown (Fig. 3F-G). Our results show that upon knockdown of Tcfap2c, the cells are compromised in their ability to maintain iPSC

colonies. After 4 days of OSKM withdrawal, the cells treated with shTcfap2c had a 10-fold greater decrease in bona fide iPSCs than the control (Fig. 3H).

Taken together, our results show that cells undergo important changes with regards to chromatin accessibility upon OSKM withdrawal in order to establish and maintain pluripotency in the newly formed iPSCs, namely transient enrichment of promoter and motif accessibility for TCFAP2C. Failure to properly open the necessary sites enriched for important pluripotency and chromatin organization TFs contributes to cells reverting to a MEF-like state, an outcome that can be bypassed through the synergistic action of the A2S components.

Opening of OSKM Withdrawal-Associated Peak is Key in Maintenance of iPSC Colonies

In the analysis of cells from FBS and A2S reprogramming and in cells after withdrawal of OSKM, we observe that motifs associated with the 3D reorganization of chromatin become more enriched within the late stages of reprogramming (Fig. 2C), hinting at the importance for effective reorganization of chromatin structure for the successful transition to iPSCs. This motivated us to elucidate the 3D interactions of accessible sites in the genome that guide successful reprogramming, particularly those sites that open up at some point during reprogramming. To this end, we implemented a new algorithm called scCISINT (single-cell cis interactions). Briefly, scCISINT works by applying a regression-based analysis to scATAC-seq clustering data to discern which genomic regions (partitioned into 1kb bins) best predicts the accessibility of gene promoters (Fig. 4A). The output is a list of regions, each with an importance score that

is calculated based on the predicted number of interacting genes and the strength of those interactions. We used only those regions that were also found from the differentially accessible marker peak analysis from ArchR for downstream validation.

To identify regions associated with the regulation of gene expression, we overlapped the marker peaks for each cluster in ArchR with existing H3K27ac ChIP-seq data (Chronis et al., 2017; Di Giammartino et al., 2019) to identify regions of active enhancers. We took the 200 top-ranked regions based on scCISINT importance score and FDR value from the marker peak analysis in ArchR, and overlapped these with the top 200 H3K27ac enriched sites (Fig. 4A). This was performed on marker peaks from three different scATAC-seq datasets (Fig. 1C, Fig. S2E, Fig. 3A) From this narrowed down list, we identified 17 peaks of interest with distinct patterns of accessibility that came up during reprogramming or were more enriched in specific clusters (Fig. 4B, S4, 5A).

Among these 17 peaks, we pursued one peak of interest that became more open in the “bona fide” cluster of OSKM withdrawal-associated cells (Peak 1) (Fig. 4B), which was revealed via a motif analysis of the peak itself to contain a TCFAP2 motif. Among its predicted interactors from scCISINT (Fig. 4C) were the *mir290* group of microRNAs (Fig. 4B-C), which have been shown to play a variety of roles in reprogramming and stem cell maintenance. These include promoting self-renewal, upregulating core pluripotency TFs, and promoting MET (Yuan et al., 2017). The *mir290* family has also been implicated in trophoblast proliferation and placental development (Paikari et al., 2017). Another interacting locus is the promoter for protein-coding gene *AU018091*, which is located ~32kb upstream of the peak (Fig. 4B). AU018091 is also localized to

the placenta, though little is known about the protein's function. Interestingly, TCFAP2C is a known key regulator of the trophoblast lineage. Moreover, previous studies have proposed that cells can be reprogrammed to alternative cell states including trophoblast stem cells (TSCs) and even an extraembryonic endoderm (XEN) state prior to becoming iPSCs (Benchetrit et al., 2019; Jaber et al., 2022; Liu et al., 2020; Parenti et al., 2016; Y. Zhao et al., 2015). Given the enrichment of the TCFAP2C motif and our identified peak interacting with placental genes in the withdrawal cluster, we asked whether reprogramming cells have to go through a trophoblast-like state en route to becoming iPSCs. Looking at the accessibility of motifs for known trophoblast factors (TEAD4, EOMES, GATA3) revealed that unlike the specificity of TCFAP2C, EOMES and GATA3 motifs are enriched in the withdrawal as well as the PSC cluster along with known XEN factors (GATA4/6) (Fig. S3C, S3E). Conversely, the TEAD4 motif is lost, with enrichment in the reverting, MEF-like and MEF clusters (S4C). Since some of these markers are shared between these different cell states and not all are enriched in the "bona fide" cluster exclusively, we cannot definitively conclude that cells must pass through a trophoblast-like state from this data alone, and strongly implies that TCFAP2C itself must be enriched without passing through this entirely different cell state.

To assess the importance of Peak 1 in reprogramming, we used CRISPR interference (CRISPRi) to target and repress this region through the addition of H3K9me3 by dCas9-KRAB to this site prior to reprogramming induction. We cloned a guide RNA targeting this region into a plasmid with GFP and the necessary CRISPRi machinery. After transducing this plasmid into MEFs, we sorted for GFP+ cells and induced reprogramming (Fig. 4D). Compared to cells with an empty control plasmid, we

observed a significant drop in the number of NANOG-positive colonies after 4 days of OSKM withdrawal compared to the day of withdrawal in the CRISPRi-targeted cells (Fig. 4E-F). Additionally, RT-qPCR of the interactor AU018091 showed that it was significantly decreased upon OSKM withdrawal when the peak is repressed, while no significant change was observed in the other interactors, such as the constitutively expressed *Ndufa3* gene (Fig. 4G). Overall, these results highlight the facility of scCISINT in the identification of key regulatory enhancer sites, and has successfully found a region that plays an important role in the maintenance of the pluripotent state and stability of iPSC colonies.

Ubiquitously Accessible Peak is Inhibitory Early but Beneficial Late in Reprogramming

From the scCISINT output, we identified a second peak (Peak 2) of interest to investigate through CRISPRi knockdown (Fig. 5A). Peak 2 was identified from our scCISINT analysis of the clusters in Fig. 1C and is located within an intron of the non-coding RNA *Gm31735*. This peak has the highest number of interactors (42) (Fig. 5B) and highest scCISINT importance score of our 17 peaks of interest, indicating that this peak could be an enhancer hub that is capable of forming 3D interactions to regulate the expression of a large panel of different genes (Fig. 5A-C, S5). Although this peak has the highest enrichment in the most advanced A2S cluster, it has at least some level of moderate to high accessibility in nearly all clusters (Fig. 5A). Interestingly, when we look at the expression of its predicted interactors from previous scRNA-seq reprogramming data (Tran et al., 2019), they can be grouped into different categories

based on their patterns of expression across reprogramming (downregulated, upregulated or invariant expression) (Fig. 5B). The somewhat ubiquitous presence of this peak and the varying expression patterns of its associated genes suggests that this peak could play a role in regulating genes through the entirety of reprogramming, from somatic to early and late stages.

We performed CRISPRi transduction, sorting, and induction of reprogramming in the same manner as Peak 1 (Fig. 4D). Surprisingly, we found that when the peak was targeted by CRISPRi (Fig. 5D), we consistently see a significant increase in the number of iPSC colonies (IF-stained for NANOG) on both the date of OSKM withdrawal and 4 days post-withdrawal (Fig. 5E). This is contrary to the expectation that the repression of a peak that becomes more open upon reprogramming induction and is present throughout the process will compromise the process. However, since the colony numbers are increasing, this could be due to the enhancer activity of this peak on genes that are present early on but must be downregulated in the course of reprogramming. Without this enhancer, these genes' expression may be more easily downregulated, resulting in acceleration in the formation of NANOG+ colonies, contributing to the observed increase.

We therefore hypothesized that knockdown of this peak would have different effects on the reprogramming outcome if the knockdown occurred at a different timepoint along reprogramming. To address this, I induced A2S reprogramming in MEFs for three days, performing a lentiviral transduction on consecutive days (Day 3 and 4) as in our previous CRISPRi experiments (Fig. 5F). Cells were sorted on Day 6 based on expression of the GFP marker (GFP+ and GFP- cells were both collected and

plated) and reprogrammed for an additional 4 days. From this experiment, we observed that the GFP- cells (no CRISPRi KD) had ~60-65% of their cells being positive for NANOG expression. The GFP+ cells (with CRISPRi KD) had a reduction with 40-47% NANOG+ cells (Fig. 5G-H). This could be due to the fact that the knockdown of the peak will now exclusively only affect the expression of genes that come up in late reprogramming and are likely to be important for the transition to pluripotency.

Our results illustrate that Peak 2 acts as an enhancer hub that affects regulation of a large swath of genes across the reprogramming timeline, all with varying transcriptional dynamics, such that knockdown of this peak will have different effects on the reprogramming outcome. Taken together with our results from Figure 4, we were able to use computational means in combination with scATAC-seq data to identify different novel enhancer regions that affect the regulation of genes and, consequently, reprogramming success, both at specific stages (e.g. OSKM withdrawal) or throughout the duration of the reprogramming timeline.

MEF-Associated Peak is Anti-Correlated with Nearby Nanog Promoter and Antagonistic to Reprogramming

Lastly, we set out to find patterns of accessibility between correlated peaks associated with the same gene, using an added feature of scCISINT to identify the one or more marker peaks associated with each gene (Fig. 6A). In this analysis, the identified peaks can display, for example, a correlation with one another, with both peaks becoming more open or closed as reprogramming progresses; alternatively,

peaks can display an anti-correlated relationship, where as one peak closes, another opens, and vice versa (Fig. 6A).

We identified a peak of interest (Peak 3) associated with the pluripotency marker NANOG. This peak displays an accessible configuration in MEFs and early reprogramming, but is quickly closed prior to opening of the *Nanog* gene promoter (displaying an anti-correlative relationship) Fig. 6B). Of note, this peak contained within it a motif for somatic transcription factor Runx1. Analyzing ChIP-seq data for RUNX1 from Chronis et al. 2017, we found that in MEFs, RUNX1 binding was indeed enriched at this locus. We again performed reprogramming on MEFs in control and CRISPRi-mediated knockdown conditions. In so doing, we observed a significant increase in Nanog+ colonies, both on the date of OSKM-withdrawal and 4 days post-withdrawal (Fig. 6C).

These results indicate that this MEF-associated peak is inhibitory to reprogramming and illustrates another facet of scCISINT and its faculty to identify relationships between scATAC-seq peaks, including those that may be antagonistic towards reprogramming.

Discussion

Recent studies have used ATAC-seq to uncover they chromatin accessibility dynamics associated with somatic cell reprogramming, on the scale of both bulk populations (Buckberry et al., 2023; Cao et al., 2018; K. Chen et al., 2020; X. Chen et al., 2023; Chronis et al., 2017; Di Giammartino et al., 2019; Di Stefano et al., 2016; Knaupp et al., 2017; Li et al., 2017; Parenti et al., 2016) and single-cell resolution in

human reprogramming studies (Liuyang et al., 2023; Nair et al., 2023; Xing et al., 2020). These studies have led to a better understanding of the shifting chromatin landscape as different groups of loci undergo closing and/or opening. We have combined scATAC-seq technology with our A2S-mediated high-efficiency reprogramming to identify the changes associated with an enhanced reprogramming system.

Similar to these previous studies, we found a gradual extinction of accessibility of the AP1 motifs and gain of POU5F1 and SOX2 motifs. Motif analysis implicated a role for transcription factors that are known to be associated with 3D chromatin reorganization (DiGiammartino, et al., 2019; Ortabozkoyun et al., 2022; Rong Ow et al., 2013; Fedele et al., 2017; Ma et al., 2014) in successful reprogramming. This motivated us to pursue more specific reprogramming-associated 3D interactions between enhancer hubs and their predicted interacting genes using a novel algorithm scCISINT. In so doing, we were able to identify three loci that are distinct in their accessibility patterns. One peak is open primarily upon OSKM-withdrawal into pluripotency and was found to be important in maintenance of the pluripotency network. A second peak regulates transcription of genes associated with all stages of the reprogramming process, whose presence is inhibitory early in reprogramming, but its regulation of upregulated genes makes this locus important for late stage reprogramming. Finally, a *Nanog*-proximal peak that is anti-correlated with *Nanog* promoter opening was found to be inhibitory to reprogramming. While we pursued these regions to assay further, future research could involve a large-scale screen targeting the other peaks identified by our scCISINT computational analysis to more effectively identify all peaks that are crucial or

inhibitory for reprogramming, leading to more comprehensive understanding of the regulatory underpinnings of this process.

In an attempt to parse out the contribution of each individual component of A2S, we analyzed reprogramming data with the dual combination permutations of A2S (A2, AS, and S2). This revealed that 2i is key in pushing cells beyond the somatic state as loss of 2i caused cells to linger behind the progression observed in the A2S cells, which parallels previous findings from our lab from scRNA-seq analysis of reprogramming, which showed that 2i helps to suppress the somatic transcriptional network (Tran et al., 2019). Moreover, we found that the combination of AS not only promotes activation of the pluripotency network, as we also found previously (Tran et al., 2019), but it also promotes stabilization and maintenance of the pluripotency network, preventing cells from reverting to a MEF-like state upon withdrawal of exogenous OSKM expression.

An examination of reprogramming cells after withdrawal of ectopic OSKM revealed that after 2 days, these cells experience strong transient upregulation of accessibility at the trophectoderm-associated factor TCFAP2C and its associated motif. It was previously reported that TCFAP2C plays an anti-apoptotic role and promotes the mesenchymal-to-epithelial transition in reprogramming of mouse fibroblasts to iPSCs (Y. Wang et al., 2020). We have shown here that TCFAP2C activity is also important late in reprogramming, promoting maintenance and stabilization of the formed iPSC colonies. These results are similar to what was found upon CRISPRi KD of the withdrawal-associated peak, which was predicted to interact with placental and trophectoderm-associated genes. While previous reports have shown that induced trophoblast stem cells (TSCs) are an alternative cell fate during reprogramming and

form separately, but in parallel with iPSCs (Benchetrit et al., 2015, 2019; Jaber et al., 2022; Naama et al., 2023), our findings suggest that reprogramming cells may potentially need to pass through a TSC-like state upon exogenous OSKM withdrawal, prior to complete acquisition of pluripotency. Further experimentation and analysis of this population of cells is required to definitively conclude that these are indeed TSC-like cells.

Our work here has led to the development of the novel computational algorithm scCISINT. This new tool was not only able to predict the interacting partners of any given site along the entire genome, but it also allowed for discernment of these regions and their interactions at the resolution of single-cell clusters. Thus, scCISINT provides knowledge on which peaks' interactive activity is associated with which stage of the reprogramming process. Given this information, plotting the accessibility of these peaks across clusters will then help to decipher regulatory hubs whose accessibility changes in a temporal fashion. With this in mind, scCISINT has the potential to be an exceptionally useful and powerful tool that can be applied to a wide range of time-course based analyses investigating various cellular pathways.

Together, these results shed light on the intricate regulatory mechanisms that guide reprogramming. This could lead to future research uncovering combinations of peaks whose combined knockdown could additively or synergistically enhance reprogramming. This information could potentially be used to systematically improve the reprogramming process in a logical manner.

Materials and Methods

Isolation of Reprogrammable MEFs

MEFs were isolated from E13.5 embryos (as described in Tran et al., 2015) from reprogrammable mice that are homozygous for a transgene containing Oct4-2A-Klf4-2A-IRES-Sox2-2A-c-Myc (OKSM) at the Col1a1 locus (Sridharan et al., 2013). All MEFs used in this study were also homozygous for the reverse tetracycline transactivator (rtTA) allele at the Rosa26 locus. MEFs were maintained in MEF media containing DMEM, 10% FBS, L-glutamine, Pen/Strep, NEAA, and 2-mercaptoethanol. Mice were housed in agreement with our UW-Madison Institutional Animal Care and Use Committee (IACUC) approved protocol (ID M005180-R03).

Mouse Embryonic Stem Cells and Induced Pluripotent Stem Cells

The iPSCs used in this study were derived from iPSC clone line #1 from Tran et al., 2019. All iPSCs and V6.5 ESCs were maintained in ESC media containing Knockout DMEM, 15% FBS, L-glutamine, Pen/Strep, NEAA, 2-mercaptoethanol, and LIF.

Reprogramming

One day prior to reprogramming (Day -1), MEFs are plated on 0.1% gelatin-coated 6-well plates at a seeding density of 5,000 cells per well in 6-well plates. Reprogramming was induced on Day 0 with the addition of doxycycline (2 μg/ml) and a feeder layer of irradiated human neonatal foreskin fibroblasts (ATCC HFF-1 SCRC-1041). Ascorbic acid (50 μg/ml, Sigma A8960) and SGC0946 (5 μM, ApexBio A4167) are also added at this time to A2S and dual combination reprogramming samples. For shRNA KD experiments, cells were collected from extra wells and counted to determine a sitting

count which will be used to calculate efficiency via immunofluorescent staining for NANOG. The culture media is switched from MEF media to ESC media. On Day 0.5 (~12 hrs post-induction), CHIR99021 (3 μM, Stemgent 04-0004-10) and PD-0325901 (1 μM, Stemgent 04-0006-10) (2i) were added to A2S, A2, and S2 conditions. Media is changed every other day until cells are collected for analysis or until the end of experiment.

Immunofluorescence

For immunofluorescent staining, cells were fixed for 10 minutes in 4% paraformaldehyde-PBS, followed by a 10 minute permeabilization with 0.5% TritonX-PBS. Staining was performed in blocking buffer containing 1X PBS with 5% normal donkey serum, 0.2% Tween-20, and 0.2% fish skin gelatin. The NANOG antibody (CosmoBio RCAB0002P) was used to stain cells at a concentration of 1:100.

Lentiviral Packaging

Plasmids used for CRISPRi and shRNA KD experiments were transfected into 293T cells (maintained in D10 media – DMEM + 10% FBS) cells using 1 mg/ml PEI, along with packaging plasmid pspax2 (Addgene #12260) and envelop-expressing plasmid (Addgene #12259). Media was replaced 17 hrs later to D10 media with 10 mM Sodium Butyrate and 20 mM HEPES. Media was replaced again 8 hrs later to D10 media only containing 20 mM HEPEs. Supernatant was collected at 48 hrs and 72 hrs post-transfection, filtered through a 0.45 μm Steriflip filter, and virus was concentrated using 4X lentivirus concentrator (40% W/V PEG-8000, 1.2 M NaCl).

shRNA Knockdown Reprogramming

For the Tcfap2c knockdown experiments, we designed two separate shRNAs targeting Tcfap2c and cloned separately into a pLKO-Tet-on-Neo plasmid (Addgene #21916), which were packaged into lentiviral vectors. MEFs were transduced with lentivirus using 10 μg/ml polybrene on consecutive days. The following day, G418 was added at a concentration of 600 μg/ml to the culture medium. When cells from an un-transduced control had completely died, the transduced cell conditions were plated for reprogramming (Day -1).

CRISPRi Knockdown Reprogramming

For each target peak for CRISPRi KD, two guide RNAs (gRNAs) were designed and cloned into a CRISPRi machinery-containing plasmid (Addgene #71237), and packaged into lentivirus. For each target, MEFs were transduced on consecutive days with equal proportions of each plasmid containing the two gRNAs. After second transduction, cells sit for at least 36 hrs prior to FACS. Sorting was performed using a BD FACSAria cell sorter. For the Day 3 and 4 transduction experiment, cells were re-plated at a density of ~10,000 cells/well in 12-well plates.

RT-qPCR

RNA was isolated from cells using the Isolate II RNA Mini Kit (Bioline, BIO-52702). 1ug of RNA was used for reverse transcription (QuantaBio #95047). After the RT reaction, the cDNA is diluted 1:5 and 2 μl of this sample is used to set up 10 μl qPCR reactions

with SYBR Green (Bio-Rad #1725124). qPCR reactions were set up in duplicate or triplicate. Primers for 2 housekeeping genes (Gapdh, RNA Pol II) were used as control reactions.

ChIP-qPCR

Cells used for ChIP-qPCR analysis were cross-linked in 1% formaldehyde (10min, rocking). Cross-linking was stopped by 0.14 M glycine (5 min, rocking). Cells were spun down and washed 3X with 1X PBS, and the cell pellets were stored at -80°C. Cells were resuspended in 130 μl lysis buffer (1% SDS, 50 mM tris-HCl pH 8, 20 mM EDTA, protease inhibitor) and sonicated using a Covaris S220 machine with the following parameters: 15 cycles of 45 sec on/off (peak =170, duty factor = 5, cycles/burst = 200, temp = 6-8°C). The samples were spun down at max speed (21,000g) for 10 min at 6°C. The supernatant was collected and chromatin concentration was determined using the Qubit DNA HS Assay Kit (Thermo Fisher Scientific, Q32854).

The chromatin aliquots (1 μg) were diluted 1:10 in dilution buffer (16.7 mM tris-HCl pH 8, 0.01% SDS, 1.1% TritonX, 1.2 mM EDTA, and 167 mM NaCl). 1 μl of H3K9me3 ab (Active Motif #39161) was added to each aliquot which were kept at 4°C for 16 hrs, rocking. A pre-prepared mix of protein A and G Dynabeads (Thermo Fisher Scientific, 10002D and 10004D) was added to each aliquot (50 μl of beads) and rotated for 2 hrs at 4°C. Beads were washed twice for 5 min (4°C, rocking) in each of the following buffers: Buffer “A” (50 mM Hepes pH 7.9, 0.1% SDS, 1% TritonX, 0.1% deoxycholate, 1 mM EDTA pH 8, 140 mM NaCl), Buffer “B” (50 mM Hepes pH 7.9, 0.1% SDS, 1%

TritonX, 0.1% deoxycholate, 1 mM EDTA pH 8, 500 mM NaCl), LiCl Buffer (20 mM tris-HCl pH 8, 0.5% NP40, 0.5% deoxycholate, 1 mM EDTA pH 8, 250 mM LiCl), and TE Buffer (10 mM tris-HCl pH 8, 1 mM EDTA pH 8). The beads are incubated with 2 μl RNAse A in 150 μl TE buffer for 30 min at 37°C. Reverse-crosslinking was performed by adding 3 μl SDS (10%) and 5 μl Proteinase K (20 μg/μl) and incubating at 65°C overnight. The next day, the beads were removed and the supernatant was cleaned up by phenol/chloroform extraction and ethanol precipitation. Chromatin was resuspended in 40 μl water. Input samples were prepared by diluting one of the aliquots in water (1:4) and treating with 1ul RNAse A for 30 min at 37°C, then adding 1 μl proteinase K and incubating overnight at 65°C. These samples were also cleaned up with phenol-chloroform extraction and EtOH precipitation and resuspended in 40 μl water. qPCR was performed as described above (used maximum amount of chromatin 4.8 μl). qPCR reactions of 4%, 2%, 1%, 0.5%, and 0.25% of input were used to generate standard curve.

Intracellular Flow Cytometry

Cells were harvested and fixed and stained using the FOXP3/Transcription Factor Staining Buffer kit (Invitrogen #00-5523-00). The NANOG antibody (CosmoBio RCAB0002P) was used at a concentration of 1:100. Flow cytometry analysis was performed using a ThermoFisher Attune flow cytometer machine. Analysis was performed using FlowJo.

Single-Cell ATAC-seq Library Preparation

For each sample submitted for scATAC-seq, cells were harvested at the appropriate timepoint (MEFs, ESCs, and iPSCs were passaged 2X before harvesting) and a single-cell suspension was generated as described in Tran et al., 2019. Briefly, cell cultures were washed with DPBS and dissociated from plate using 0.25% Trypsin-EDTA (Gibco #25200-072). Trypsin was neutralized with soybean trypsin inhibitor (Sigma-Aldrich #T6522), and cells were filtered (40 μm) and spun down for 3min at 300g at room temperature (RT). Cells were then resuspended in 1 ml of 0.1% BSA-PBS (prepared by diluting 7.5% Bovine Albumin Fraction V solution [Gibco #15260-037] to 0.1% with DPBS) and pipetted up and down 50X. After one final spin, cells were finally resuspended in 1 ml of 0.1% BSA-PBS for nuclei isolation, and concentration was determined using an Invitrogen Countess II cell counter prior to nuclei isolation. Nuclei were isolated from the single-cell suspension in accordance with the recommended protocol from 10X Genomics (all spin steps were carried out at 300g for 3min at RT). We aimed for a targeted nuclei recovery of 4,000 and targeted read depth of 25k reads per nucleus. Nuclei concentration was determined using the Countess II prior to transposition and library preparation using the 10X Chromium platform. Sequencing was performed using the Illumina NovaSeq 6000 machine and samples were loaded onto a S1 flow cell.

scATAC-seq Data Processing

The returned sequencing data for each sample was aligned and processed using the cellranger-atac count pipeline openly available from 10X Genomics' website. Based on the number of unique barcodes (cells) in each sample, the output fragments files were

randomly downsampled to match the number of barcodes found in the sample with the lowest number. These fragments files were used for downstream analysis in ArchR.

scATAC-seq Computational Analysis

Dimensionality Reduction and Clustering

We used ArchR version 0.9.5 (https://www.archrproject.com/index.html, Granja et al., 2021) on R version 3.6.3 (“Holding the Windsock”) to analyze our single-cell ATAC-seq data post-cellranger-atac alignment and downsampling. In all ArchR analyses, mm10 was used as the default genome. ArchR Arrow files were created for each sample, which were used to create an ArchR project. The data in the ArchR project was further processed through filtering out “doublets” (i.e. single-cell droplets that encompassed more than one nucleus).

(1) ArrowFiles <- createArrowFiles(inputFiles = inputFiles, sampleNames = names(inputFiles), filterTSS = 4, filterFrags = 1000, addTileMat = TRUE, addGeneScoreMat = TRUE)

(2) proj1 <- ArchRProject(ArrowFiles = ArrowFiles, outputDirectory = "Outputs", copyArrows = FALSE)

(3) doubScores <- addDoubletScores(input = ArrowFiles, k = 10, knnMethod = "UMAP", LSIMethod = 1)

(4) proj2 <- filterDoublets(proj1)

Dimensionality reduction via iterative latent semantic indexing (LSI) was run on the ArchR project, using the ArchR default of two iterations. We then used ArchR to invoke the tool Harmony (Korsunsky et al., 2019) to perform batch correction of the different

samples, which was followed by clustering using Seurat's (Satija et al., 2015) graph clustering method.

(5) proj2 <- addIterativeLSI(ArchRProj = proj2, useMatrix = "TileMatrix", name = "IterativeLSI", iterations = 2, clusterParams = list(resolution = c(0.2), sampleCells = 10000, n.start = 10), varFeatures = 25000, dimsToUse = 1:30)

(6) proj2 <- addHarmony(ArchRProj = proj2, reducedDims = "IterativeLSI", name = "Harmony", groupBy = "Sample")

(7) proj2 <- addClusters(input = proj2, reducedDims = "IterativeLSI", method = "Seurat", name = "Clusters", resolution = 0.8)

For all analyses, the default resolution of 0.8 was used with the exception of the clusters in Fig. 1C, where we used a resolution of 0.95 to generate clusters.

Marker Gene and Marker Peak Analysis

We used the gene score matrix from the Arrow files to perform marker gene analysis. The final marker gene list was generated using a cutoff of FDR <= 0.01 & Log2FC >= 1.25)

(8) markersGS <- getMarkerFeatures(ArchRProj = proj2, useMatrix = "GeneScoreMatrix", groupBy = "Clusters", bias = c("TSSEnrichment", "log10(nFrags)"), testMethod = "wilcoxon")

Peak calling and cluster-based marker peak analysis was performed on cell groupings after initially creating pseudo-bulk replicates based on our defined clusters. The final marker peak list was generated using a cutoff of FDR <= 0.01 & Log2FC >= 1.

(9) proj3 <- addGroupCoverages(ArchRProj = proj2, groupBy = "Clusters")

(10) proj3 <- addReproduciblePeakSet(ArchRProj = proj3, groupBy = "Clusters", pathToMacs2 = "~/Path_to_macs2")

(11) proj4 <- addPeakMatrix(proj3)

(12) markersPeaks <- getMarkerFeatures(ArchRProj = proj4, useMatrix = "PeakMatrix", groupBy = "Clusters", bias = c("TSSEnrichment", "log10(nFrags)"), testMethod = "wilcoxon")

Motif Analysis

Motifs present in all accessible regions were annotated using the cisbp motif database and motif enrichment analysis was performed pairwise between clusters. To identify enriched motifs in a particular cluster against the whole population, the “bgdGroups” parameter was removed. This was followed by computing of per-cell motif deviation scores.

(13) proj4 <- addMotifAnnotations(ArchRProj = proj4, motifSet = "cisbp", name = "Motif")

(14) markerTest <- getMarkerFeatures(ArchRProj = proj4, useMatrix = "PeakMatrix", groupBy = "Sample", testMethod = "wilcoxon", bias = c("TSSEnrichment", "log10(nFrags)"), useGroups = "C1", bgdGroups = "C2")

(15) motifsUp <- peakAnnoEnrichment(seMarker = markerTest, ArchRProj = proj4, peakAnnotation = "Motif", cutOff = "FDR <= 0.1 & Log2FC >= 0.5")

(16) proj4 <- addBgdPeaks(proj4)

(17) proj4 <- addDeviationsMatrix(ArchRProj = proj4, peakAnnotation = "Motif", force = TRUE)

Integrating scRNA-seq and scATACs-seq Data

Within ArchR, we invoked Seurat to process scRNA-seq data from the corresponding samples as in our scATAC-seq data set to create a Seurat object (“seRNA”). For each cell from our scATAC-seq data, the integration finds a cell from our scRNA-seq data that is most similar and assigned that gene expression profile to the matched scATAC-seq cell

(18) proj5 <- addGeneIntegrationMatrix(ArchRProj = proj4, useMatrix = "GeneScoreMatrix", matrixName = "GeneIntegrationMatrix", reducedDims = "IterativeLSI", seRNA = seRNA, addToArrow = TRUE, force= TRUE, groupRNA = "Sample", nameCell = "predictedCell", nameGroup = "predictedGroup", nameScore = "predictedScore")

scCISINT and Additional Computational Analysis

Detailed methods regarding the scCISINT algorithm in preparation by Sushmita Roy’s lab. Marker peaks for each ArchR cluster were overlapped with sites of H3K27ac enrichment based on existing H3K27ac ChIP-seq data. ChIP-seq data obtained from Supplementary Table 2 from DiGiammartino et al., 2019 and from GEO accession number GSE90895 for Chronis et al., 2017. A marker peak was considered to have H3K27ac if it shared at least one overlapping base pair with the H3K27ac peaks. For each cluster, the overlapping peaks were sorted based on H3K27ac enrichment, and each of these were overlapped with the top 200 ranked peaks based on scCISINT importance score and marker peak q-value (ArchR). Runx1 ChIP-seq data obtained

from GEO accession number GSE90895. HOMER (http://homer.ucsd.edu/homer/motif/) used to perform motif analysis in Fig. 4B an Fig. 6B and annotation of marker peak lists in Fig. 1F(ii). All MST and PAGA graphs were created by Spencer Halberg-Spencer.

Figure 1

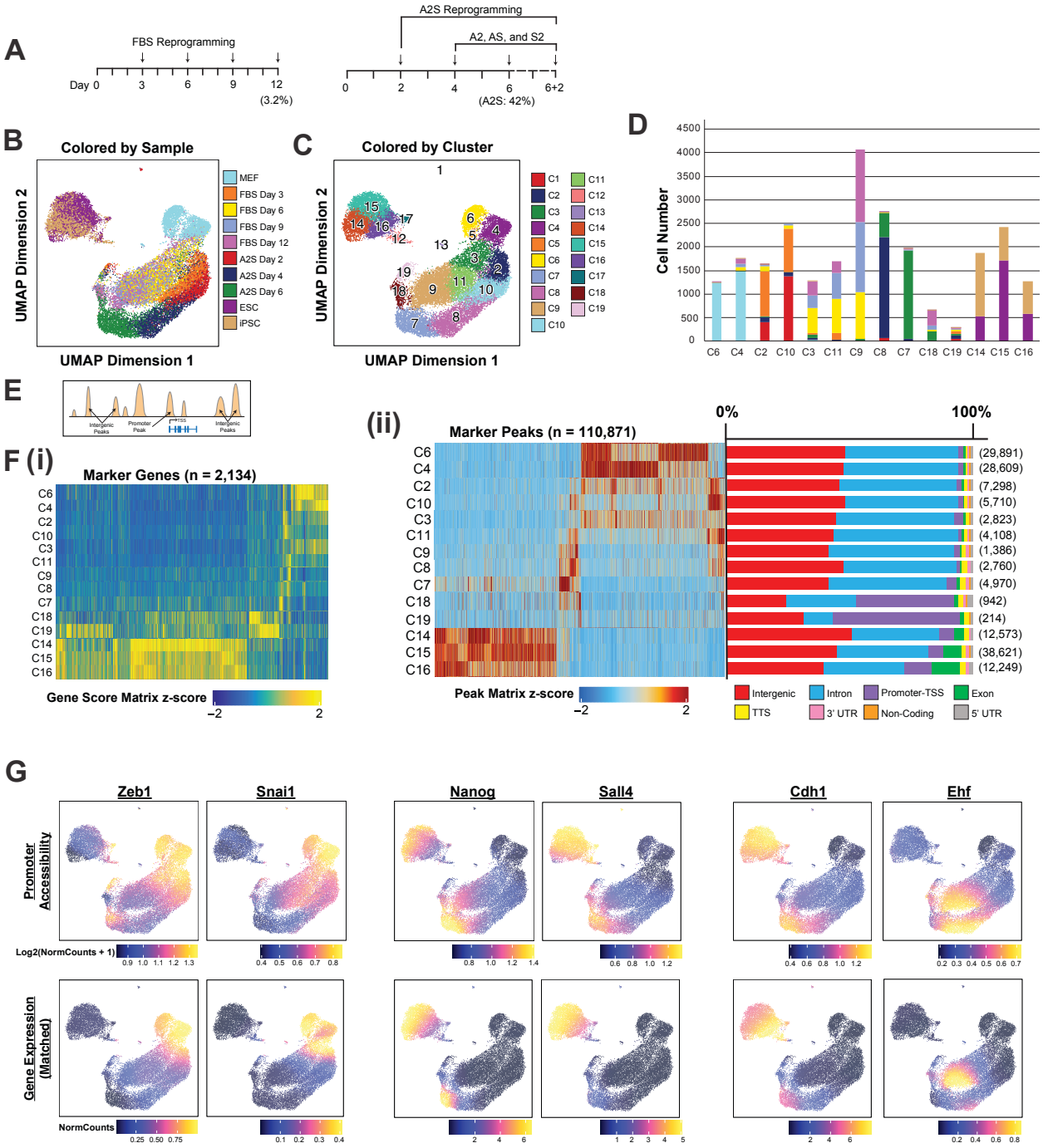

A
FBS Reprogramming
Day 0
3
6
9
12
(3.2%)
A2S Reprogramming
A2, AS, and S2
0
2
4
6
6+2
(A2S: 42%)
B
Colored by Sample
UMAP Dimension 2
UMAP Dimension 1
MEF
FBS Day 3
FBS Day 6
FBS Day 9
FBS Day 12
A2S Day 2
A2S Day 4
A2S Day 6
ESC
iPSC
C
Colored by Cluster
UMAP Dimension 2
UMAP Dimension 1
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16
C17
C18
C19
D
Cell Number
4500
4000
3500
3000
2500
2000
1500
1000
500
0
C6 C4 C2 C10 C3 C11 C9 C8 C7 C18 C19 C14 C15 C16
E
Intergenic Peaks
Promoter Peak
TSS
Intergenic Peaks
F (i)
Marker Genes (n = 2,134)
C6
C4
C2
C10
C3
C11
C9
C8
C7
C18
C19
C14
C15
C16
Gene Score Matrix z-score
−2
2
(ii)
Marker Peaks (n = 110,871)
0%
100%
C6
C4
C2
C10
C3
C11
C9
C8
C7
C18
C19
C14
C15
C16
(29,891)
(28,609)
(7,298)
(5,710)
(2,823)
(4,108)
(1,386)
(2,760)
(4,970)
(942)
(214)
(12,573)
(38,621)
(12,249)
Peak Matrix z-score
−2
2
Intergenic
Intron
Promoter-TSS
Exon
TTS
3' UTR
Non-Coding
5' UTR
G
Zeb1
Snai1
Nanog
Sall4
Cdh1
Ehf
Promoter Accessibility
Log2(NormCounts + 1)
Gene Expression (Matched)
NormCounts

Figure 1: scATAC-seq reveals chromatin accessibility dynamics at promoter-distal regions

A) Schematic of reprogramming timeline used to acquire samples for scATAC-seq. Arrows indicate timepoints at which cells were harvested in the indicated reprogramming conditions.

B) ArchR UMAP clustering plot of FBS, A2S, MEF, ESC, and iPSC samples colored by sample

C) UMAP plot from (B) colored by cluster

D) Number of cells from each sample found within the indicated clusters from Fig. 1C

E) Diagram illustrating examples of promoter-associated and intergenic peaks

F) Heatmap showing the calculated z-score for all marker genes (i) and marker peaks (ii) identified in ArchR in each of the indicated clusters. In (ii) the percent of marker peaks annotated to each of the indicated genomic regions in each cluster is shown to the right with the number of marker peaks for each cluster indicated.

G) UMAP plots of representative MEF, pluripotency, and reprogramming-associated genes, colored by promoter accessibility (gene score matrix, top row) and gene expression from integrated scRNA-seq data (Tran et al., 2019) (gene integration matrix, bottom row)

Figure 2

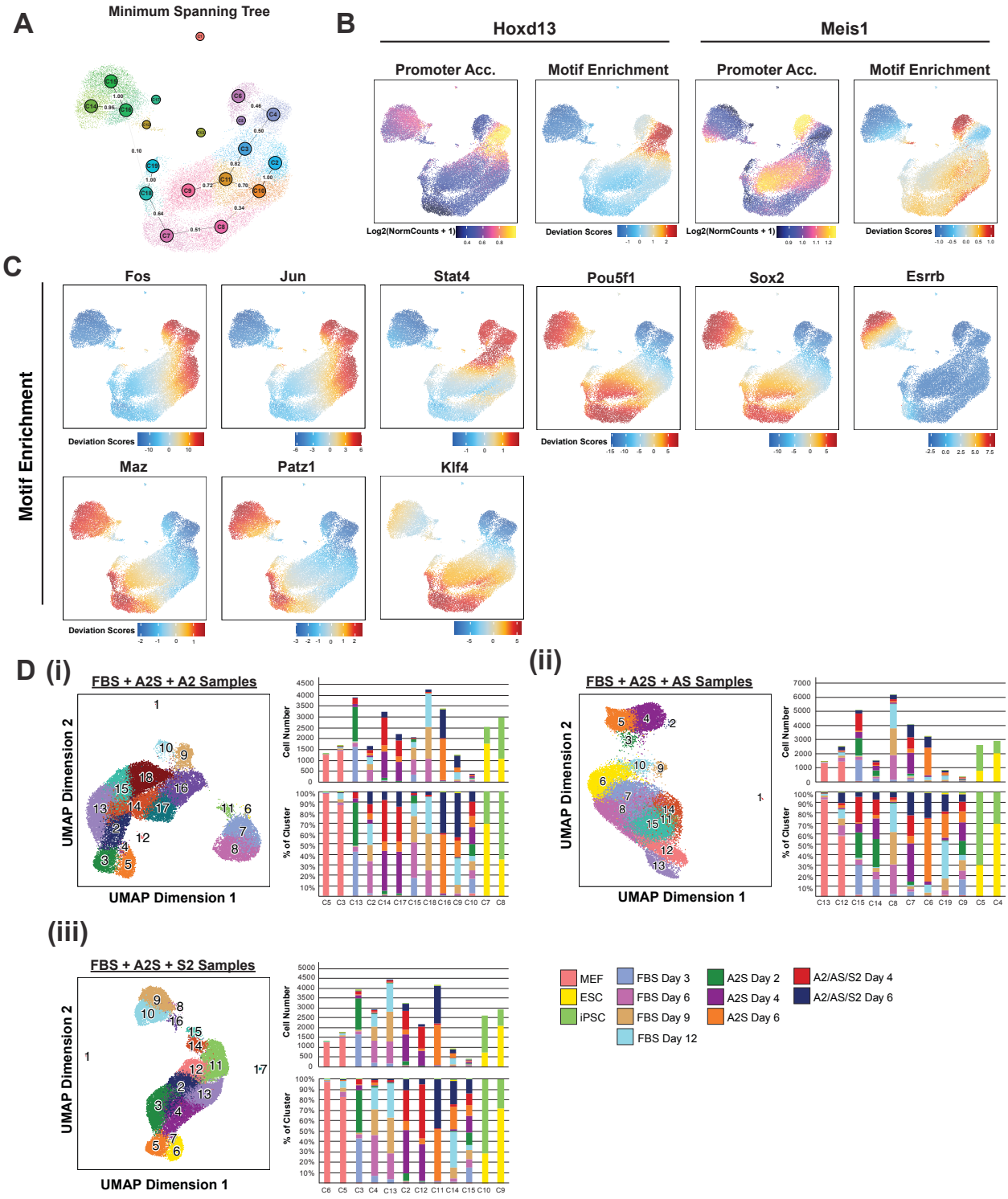

A
Minimum Spanning Tree
B
Hoxd13
Meis1
Promoter Acc.
Motif Enrichment
Promoter Acc.
Motif Enrichment
Log2(NormCounts + 1)
Deviation Scores
Log2(NormCounts + 1)
Deviation Scores
C
Motif Enrichment
Fos
Jun
Stat4
Pou5f1
Sox2
Esrrb
Maz
Patz1
Klf4
Deviation Scores
D (i)
FBS + A2S + A2 Samples
UMAP Dimension 2
UMAP Dimension 1
Cell Number
% of Cluster
(ii)
FBS + A2S + AS Samples
UMAP Dimension 2
UMAP Dimension 1
Cell Number
% of Cluster
(iii)
FBS + A2S + S2 Samples
UMAP Dimension 2
UMAP Dimension 1
Cell Number
% of Cluster
MEF
ESC
iPSC
FBS Day 3
FBS Day 6
FBS Day 9
FBS Day 12
A2S Day 2
A2S Day 4
A2S Day 6
A2/AS/S2 Day 4
A2/AS/S2 Day 6

Figure 2: A2S Enhances Accessibility at Key Transcription Factor Binding Sites

A) Minimum spanning tree calculated for the clusters in Fig. 1C. Edge weights between connected clusters are indicated

B) UMAP plots of two motifs enriched within each of the two MEF clusters, colored by promoter accessibility and motif enrichment (motif matrix)

C) UMAP plots of representative differentially enriched motifs within the reprogramming-associated clusters from Fig. 1C, colored by motif enrichment

D) UMAP plots showing clustering of FBS and A2S samples with (i) A2, (ii) AS, and (iii) S2 samples. Cluster composition is shown to the right of each UMAP plot, represented as both number of cells and percent of cells coming from each sample

Figure 3

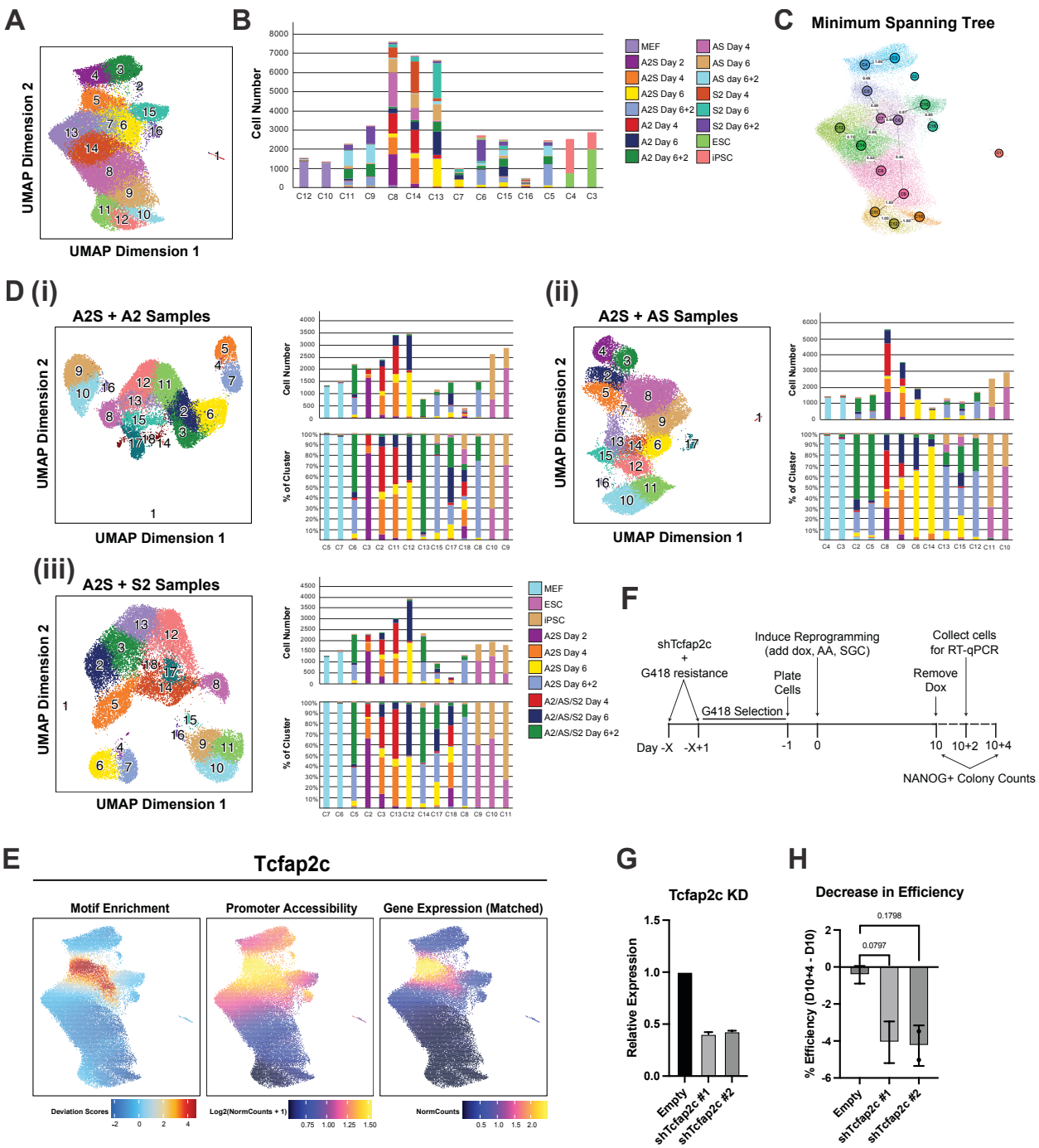

A
UMAP Dimension 2
UMAP Dimension 1
B
Cell Number
MEF
A2S Day 2
A2S Day 4
A2S Day 6
A2S Day 6+2
A2 Day 4
A2 Day 6
A2 Day 6+2
AS Day 4
AS Day 6
AS day 6+2
S2 Day 4
S2 Day 6
S2 Day 6+2
ESC
iPSC
C
Minimum Spanning Tree
D (i)
A2S + A2 Samples
(ii)
A2S + AS Samples
(iii)
A2S + S2 Samples
% of Cluster
F
shTcfap2c
+
G418 resistance
Induce Reprogramming
(add dox, AA, SGC)
Plate
Cells
Collect cells
for RT-qPCR
Remove
Dox
G418 Selection
Day -X
-X+1
-1
0
10
10+2
10+4
NANOG+ Colony Counts
E
Tcfap2c
Motif Enrichment
Promoter Accessibility
Gene Expression (Matched)
Deviation Scores
Log2(NormCounts + 1)
NormCounts
G
Tcfap2c KD
Relative Expression
Empty
shTcfap2c #1
shTcfap2c #2
H
Decrease in Efficiency
% Efficiency (D10+4 - D10)
0.1798
0.0797

Figure 3: Enrichment of Specific Motifs Mediates Stabilization of Pluripotent State Upon OSKM Withdrawal

A) UMAP clustering plot of A2S, A2, AS, and S2 samples, including their corresponding withdrawal (Day 6+2) samples, with MEFs, ESCs, and iPSCs; colored by cluster

B) Number of cells from each sample found within the indicated clusters from Fig. 3A

C) Minimum spanning tree calculated for the clusters in Fig. 3A. Edge weights between connected clusters are indicated

D) UMAP plots showing clustering of A2S samples with (i) A2, (ii) AS, and (iii) S2 samples, including their corresponding withdrawal (Day 6+2) samples. Cluster composition is shown to the right of each UMAP plot, represented as both number of cells and percent of cells coming from each sample

E) UMAP plots showing the motif enrichment (left), promoter accessibility (middle) and integrated gene expression (right) scores for Tcfap2c

F) Schematic showing experimental timeline for shRNA knockdown of Tcfap2c

G) RT-qPCR results for one replicate experiment of Tcfap2c from Day 10+2 of the shRNA knockdown experiment. Error bars indicate standard deviation between three replicates. Between two replicate experiments, the shRNAs averaged 59% and 51% knockdown for shRNA #1 and #2, respectively

H) Decrease in reprogramming efficient between Days 10 and Day 10+4 for control and shTcfap2c knockdown conditions. Error bars indicate standard deviation between two replicate experiments

Figure 4

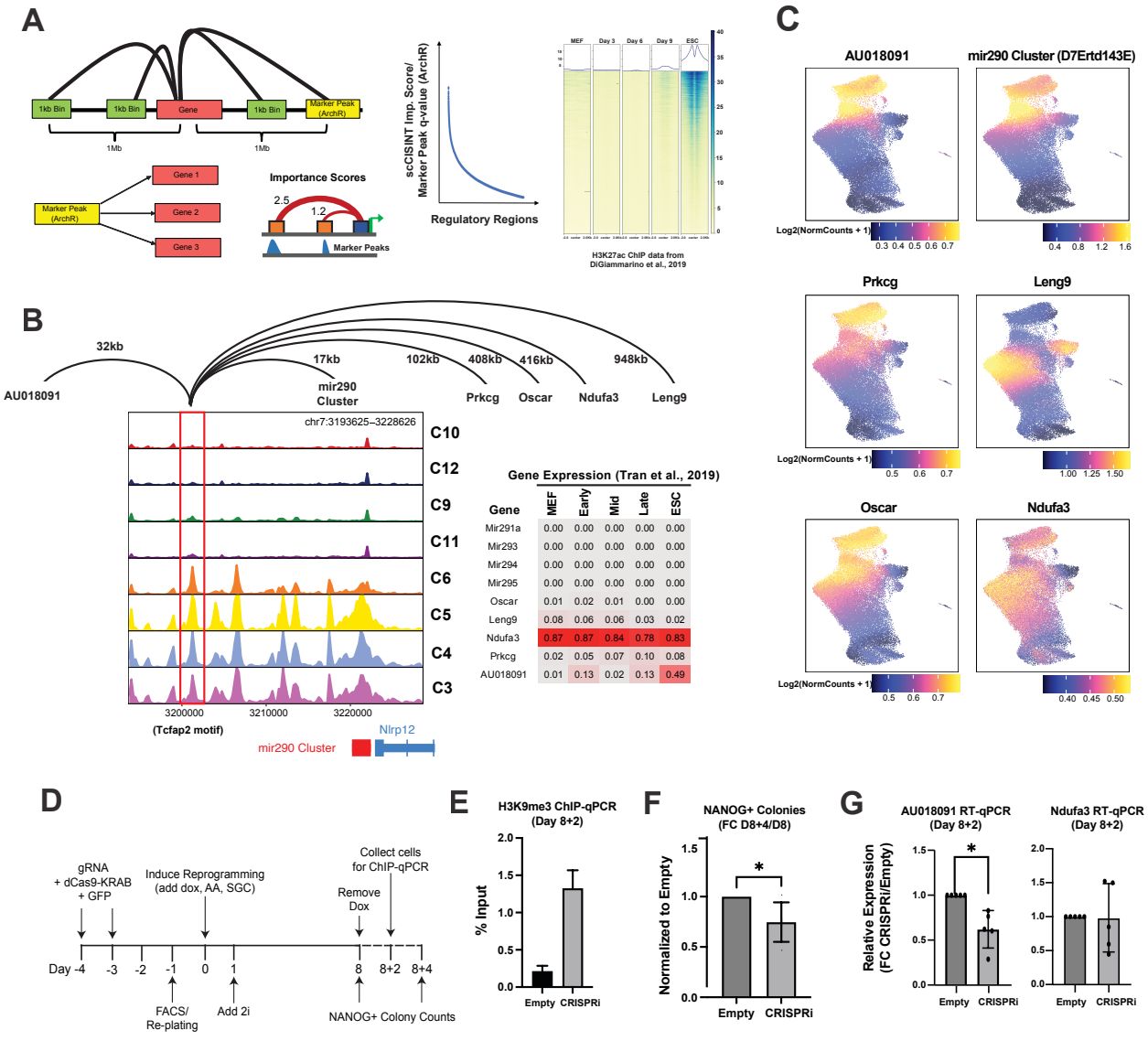


Figure 4: Opening of OSKM Withdrawal-Associated Peak is Key in Maintenance of iPSC Colonies

A) (Left) Schematic showing how scCISINT uses accessibility data to predict cis-interacting relationships between any given genomic region and gene promoters; (Middle) From scCISINT the top 200 ranked peaks by importance score and marker peak q-value were explored for downstream analysis; (Right) Example heatmap

showing H3K27ac enrichment in the marker peaks from Cluster 5 of Fig. 3A. The top 200 peaks from the sorted heatmap for each cluster were overlapped with the top 200 ranked regions from scCISINT for consideration in experimental validation

B) Track showing the accessibility of Peak 1 (boxed in red) in the indicated clusters from Fig. 3A. Distance between peak and predicted interactors is shown on top. Gene expression of predicted interactors that had associated scRNA-seq data from Tran et al., 2019 is shown on the right; numbers represent the percent of cells within indicated scRNA-seq cluster expressing that gene.

C) UMAP plots (Fig. 3A) of Peak 1 interactors colored by promoter accessibility

D) Schematic showing experimental timeline for CRISPRi repression of Peak 1

E) H3K9me3 ChIP-qPCR results for Peak1 from one replicate experiment. Error bars indicate standard deviation between 3 replicate reactions.

F) Fold change in NANOG+ colonies between D8+4 and Day 8 for control and CRISPRi samples, normalized to control. Error bars indicate standard deviation between 5 replicate experiments. (* < 0.05)

G) RT-qPCR results for AU018091 (left) and Ndufa3 (right) on reprogramming day 8+2 in control and CRISPRi conditions. Error bars indicate standard deviation between 5 replicate experiments. (* $p < 0.05$).

Figure 5

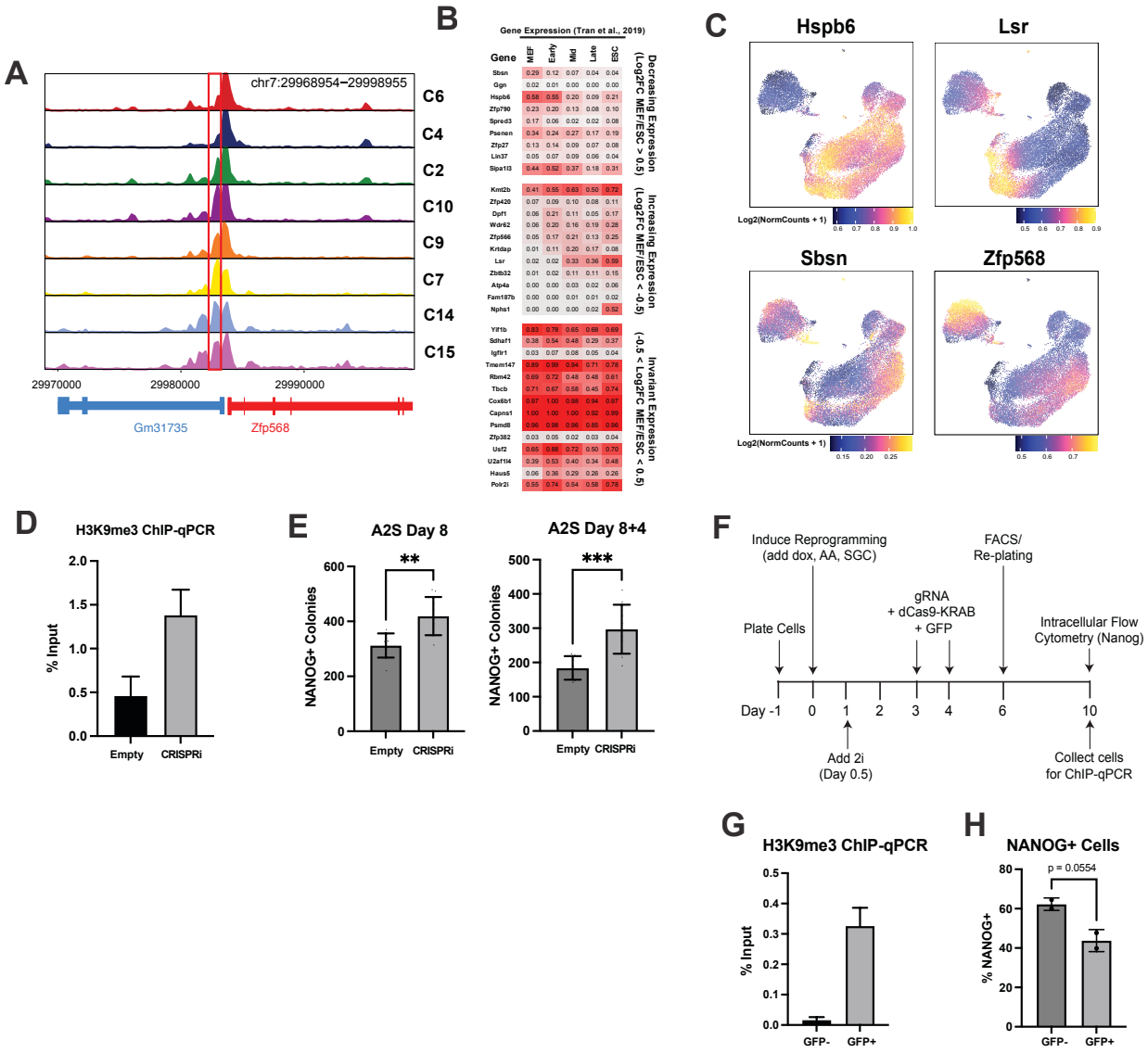

A
chr7:29968954–29998955
C6
C4
C2
C10
C9
C7
C14
C15
29970000
29980000
29990000
Gm31735
Zfp568
B
Gene Expression (Tran et al., 2019)
Gene
MEF
Early
Mid
Late
ESC
Decreasing Expression (Log2FC MEF/ESC > 0.5)
Increasing Expression (Log2FC MEF/ESC < -0.5)
Invariant Expression (-0.5 < Log2FC MEF/ESC < 0.5)
C
Hspb6
Lsr
Sbsn
Zfp568
Log2(NormCounts + 1)
D
H3K9me3 ChIP-qPCR
% Input
Empty
CRISPRi
E
A2S Day 8
A2S Day 8+4
NANOG+ Colonies
**
***
F
Induce Reprogramming (add dox, AA, SGC)
FACS/ Re-plating
gRNA + dCas9-KRAB + GFP
Plate Cells
Intracellular Flow Cytometry (Nanog)
Day -1
0
1
2
3
4
6
10
Add 2i (Day 0.5)
Collect cells for ChIP-qPCR
G
H3K9me3 ChIP-qPCR
% Input
GFP-
GFP+
H
NANOG+ Cells
p = 0.0554
% NANOG+

Figure 5: Ubiquitously Accessible Peak is Inhibitory Early but Beneficial Late in Reprogramming

A) Track showing the accessibility of Peak 2 (boxed in red) in the indicated clusters from Fig. 1C.

B) Gene expression of predicted interactors that had associated scRNA-seq data from Tran et al.; numbers represent the percent of cells within indicated scRNA-seq cluster expressing that gene.

C) UMAP plots (Fig. 1C) of Peak 2 interactors colored by promoter accessibility

D) H3K9me3 ChIP-qPCR results for Peak 2 from one replicate experiment. Error bars indicate standard deviation between 3 replicate reactions. Cells collected on Day 8.

E) Number of NANOG+ colonies on day 8 (left) and day 8+4 (right) of A2S reprogramming

F) Schematic of experimental timeline for CRISPRi suppression of Peak 2 beginning on Day 3 of A2S reprogramming

G) H3K9me3 ChIP-qPCR results for Peak 2 from one replicate experiment using the timeline from Fig. 1F. Error bars indicate standard deviation between 3 replicate reactions. Cells collected on Day 10.

H) Percent of cells that are NANOG+ by flow cytometry in cells transduced with CRISPRi machinery (GFP+) and those that were not (GFP-). Cells collected on Day 10 for flow cytometry.

Figure 6

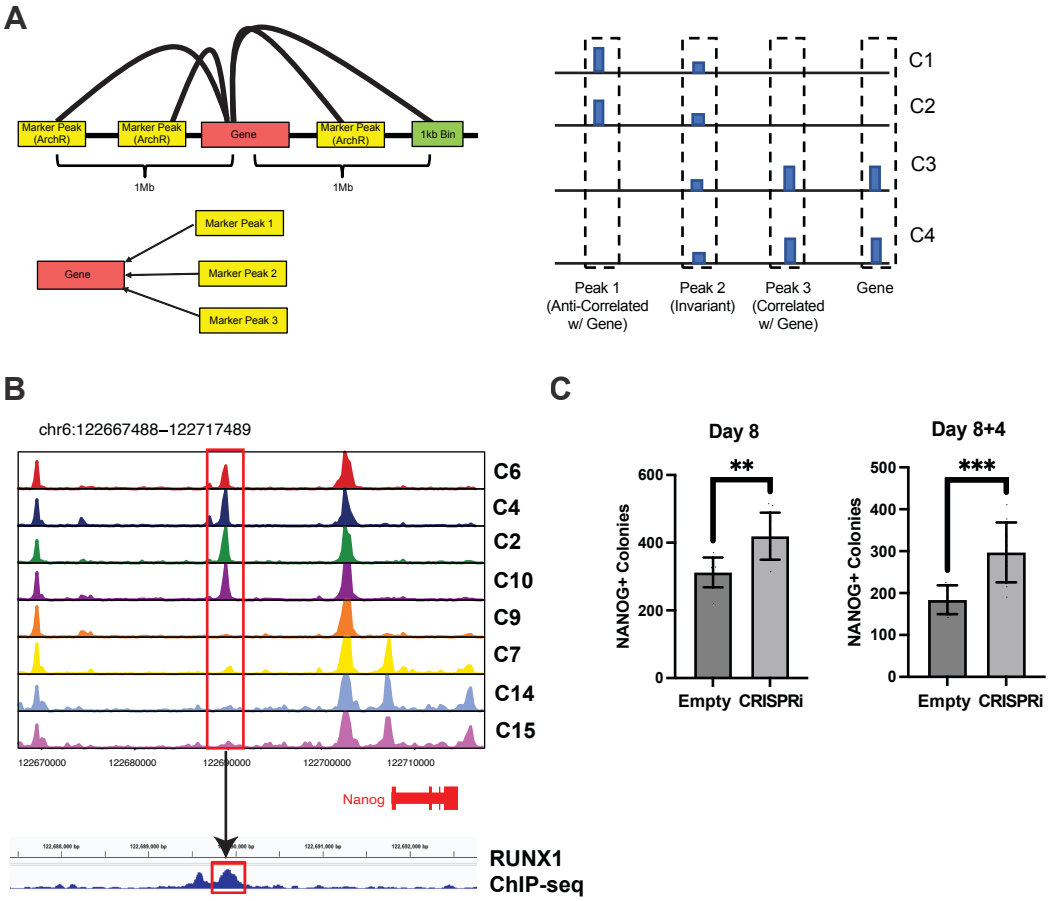


Figure 6: MEF-Associated Peak is Anti-Correlated with Nearby Nanog Promoter and Antagonistic to Reprogramming

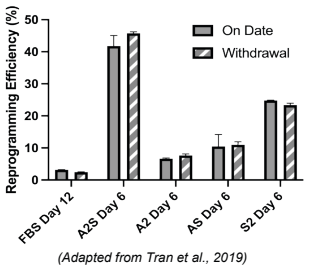
A) (Left) Schematic illustrating the capability of scCISINT to identify multiple peaks that influence a single gene. (Right) Illustration of different patterns of peaks that may occur between ATAC-seq and that scCISINT is able to identify

B) Track showing the accessibility of Peak 3 (boxed in red) in the indicated clusters from Fig. 1C. RUNX1 ChIP-seq data from Chronis et al., 2017 visualized using IGV shown below track.

C) Number of NANOG+ colonies on day 8 (left) and day 8+4 (right) of A2S reprogramming. (**p < 0.01, ***p < 0.001)

Figure S1

A



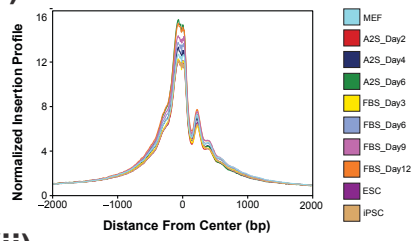
(Adapted from Tran et al., 2019)

B

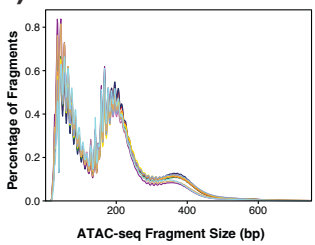
Statistics from CellRanger-ATAC processing of sequencing data

Sample	Estimated # Cells	Median Fragments/Cell	Total # of Read Pairs	# Cells Retained After ArchR Filtering	# Cells Retained After Filtering & Downsampling
A2S Day 2	4,082	2,506	71,136,435	2,792	1,962
A2S Day 4	3,818	12,556	119,785,884	3,102	2,508
A2S Day 6	4,398	9,010	150,631,033	3,875	2,775
A2S Day 6+2	3,516	21,610	232,541,590	2,705	2,990
A2 Day 4	7,771	7,271	127,559,600	6,237	2,615
A2 Day 6	5,213	6,373	121,419,830	4,636	2,804
A2 Day 6+2	3,138	18,501	211,154,590	3,347	2,993
AS Day 4	4,892	10,323	118,361,266	4,087	2,623
AS Day 6	3,631	9,486	115,247,697	3,229	2,767
AS Day 6+2	3,827	20,667	257,972,998	4,120	2,998
S2 Day 4	6,517	7,033	101,644,283	5,200	2,516
S2 Day 6	4,376	7,005	125,309,963	3,840	2,759
S2 Day 6+2	3,391	19,189	225,953,283	3,610	2,992
FBS Day 3	5,034	2,386	68,346,442	4,029	2,117
FBS Day 6	5,453	9,507	113,599,463	4,481	2,579
FBS Day 9	5,456	10,206	128,933,848	4,358	2,551
FBS Day 12	5,255	9,181	122,652,956	4,342	2,602
MEF	4,950	21,194	229,075,232	4,563	2,912
ESC	6,423	27,551	350,942,777	6,470	3,003
iPSC	7,338	18,177	354,782,998	6,720	2,944

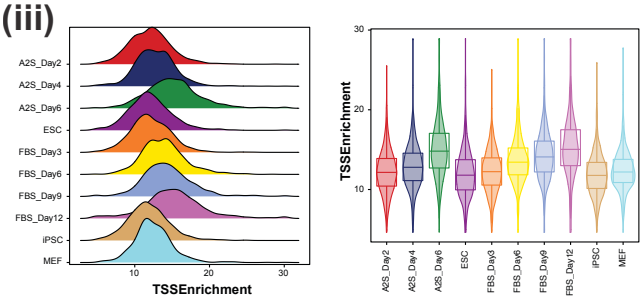
C (i)



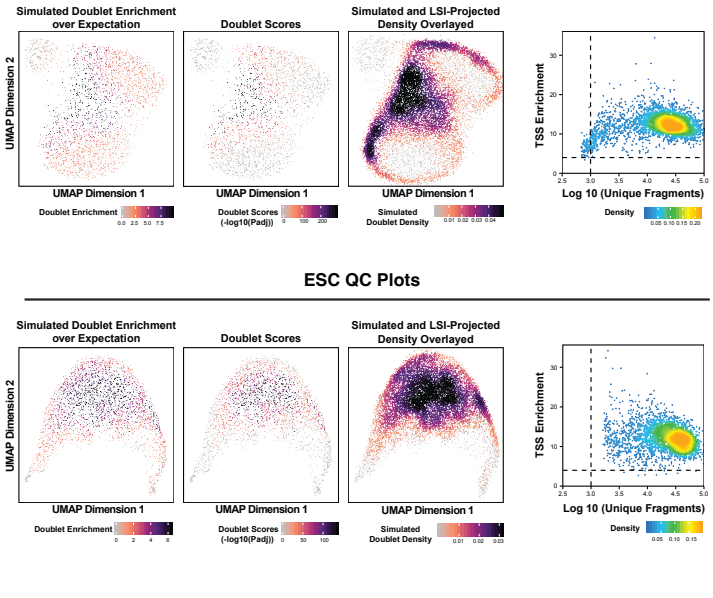
(ii)



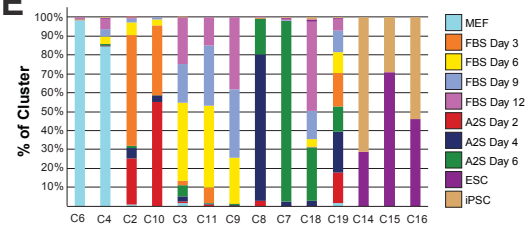
(iii)



D



E



F

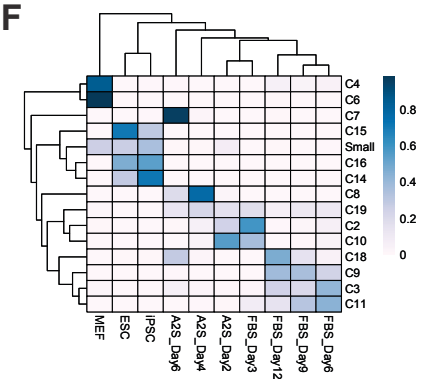


Figure S1 (Related to Figure 1)

A) Reprogramming efficiency of FBS, A2S, A2, AS, and S2 conditions on the date of OSKM withdrawal (On Date) and 4 days post-withdrawal (Withdrawal), as presented in Tran et al., 2019

B) Table summarizing the statistics of our scATAC-seq after processing with CellRanger-ATAC and the number of cells retained after ArchR QC filtering and downsampling of the data.

C) Representative QC plots output by ArchR: (i) TSS enrichment profile; (ii) Fragment size distributions; (iii) TSS enrichment scores plotted as a ridge (left) and violin (right) plot. Numbers were fairly congruent between all samples.

D) Additional ArchR QC plots from the doublet removal step in ArchR, and a filtering plot showing TSS enrichment against unique nuclear fragments per cell. Data shown for MEFs and ESCs only.

E) Composition of clusters from Fig. 1C presented as percent of cluster coming from each sample.

F) ArchR hierarchical clustering plot showing similarities between the clusters and samples along with the relative enrichment of each sample within each cluster. (Small = C1, C12, C13, and C17)

Figure S2

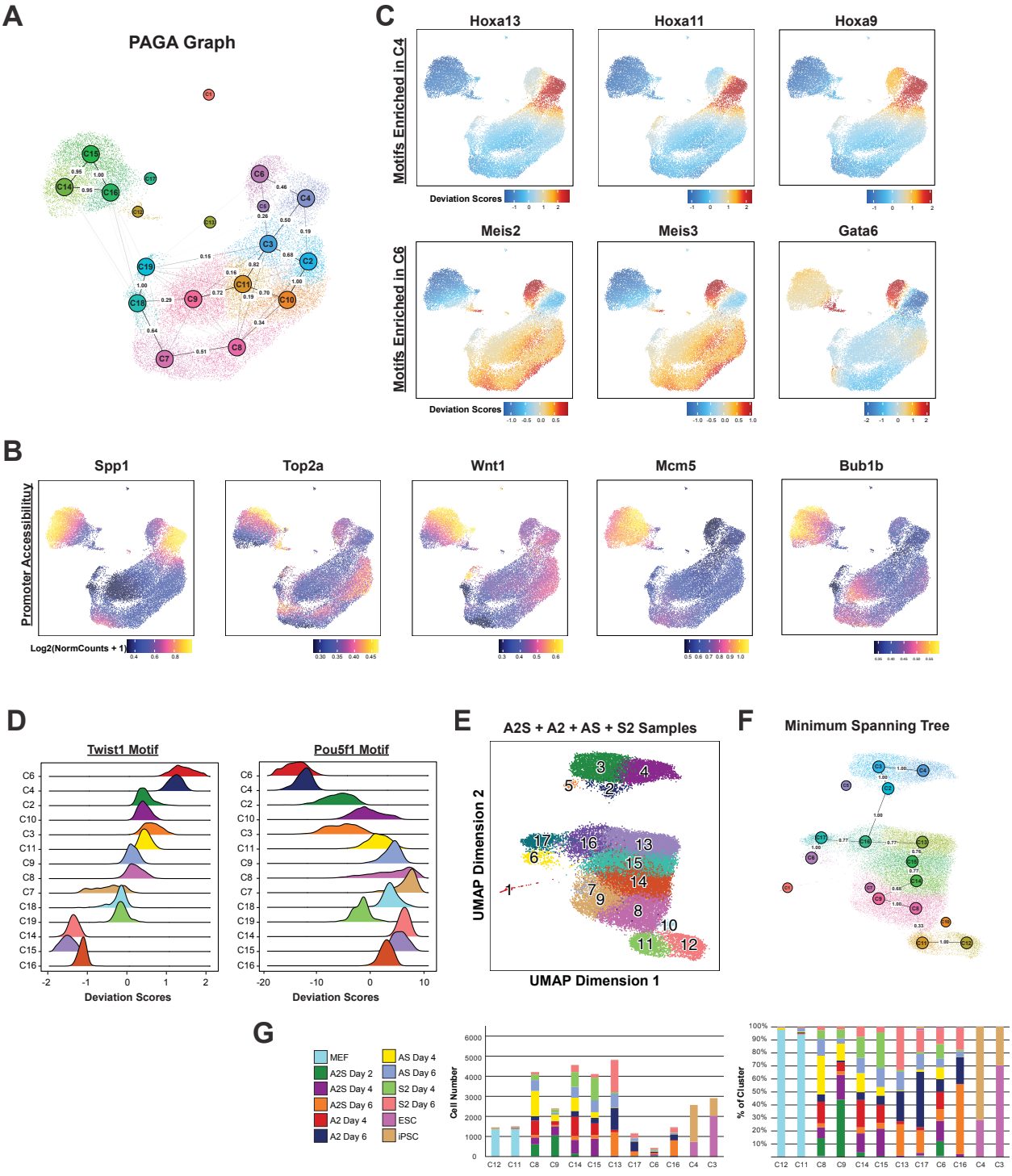

A
PAGA Graph
C
Hoxa13
Hoxa11
Hoxa9
Motifs Enriched in C4
Deviation Scores
Meis2
Meis3
Gata6
Motifs Enriched in C6
B
Spp1
Top2a
Wnt1
Mcm5
Bub1b
Promoter Accessibilituy
Log2(NormCounts + 1)
D
Twist1 Motif
Pou5f1 Motif
Deviation Scores
E
A2S + A2 + AS + S2 Samples
UMAP Dimension 2
UMAP Dimension 1
F
Minimum Spanning Tree
G
MEF
A2S Day 2
A2S Day 4
A2S Day 6
A2 Day 4
A2 Day 6
AS Day 4
AS Day 6
S2 Day 4
S2 Day 6
ESC
iPSC
Cell Number
% of Cluster

Figure S2 (Related to Figure 2)

A) PAGA graph calculated for the clusters in Fig. 1C. Edge weights between clusters are indicated

B) UMAP plots of genes previously associated with cells primed for reprogramming as well as cell cycle (Mcm5 and Bub1b), colored by promoter accessibility.

C) Additional UMAPs of motifs differentially enriched between the two MEF clusters.

D) Ridge plots showing the distribution of motif enrichment (deviation scores) for Twist1 and Pou5f1 in the indicated cluster from Fig. 1C

E) ArchR UMAP clustering plot of A2S, A2, AS, and S2 samples colored by cluster

F) Minimum spanning tree calculated for the clusters in Fig. S2E. Edge weights between clusters are indicated

G) Cluster composition of Fig. S2E, represented as both number of cells and percent of cells coming from each sample

Figure S3

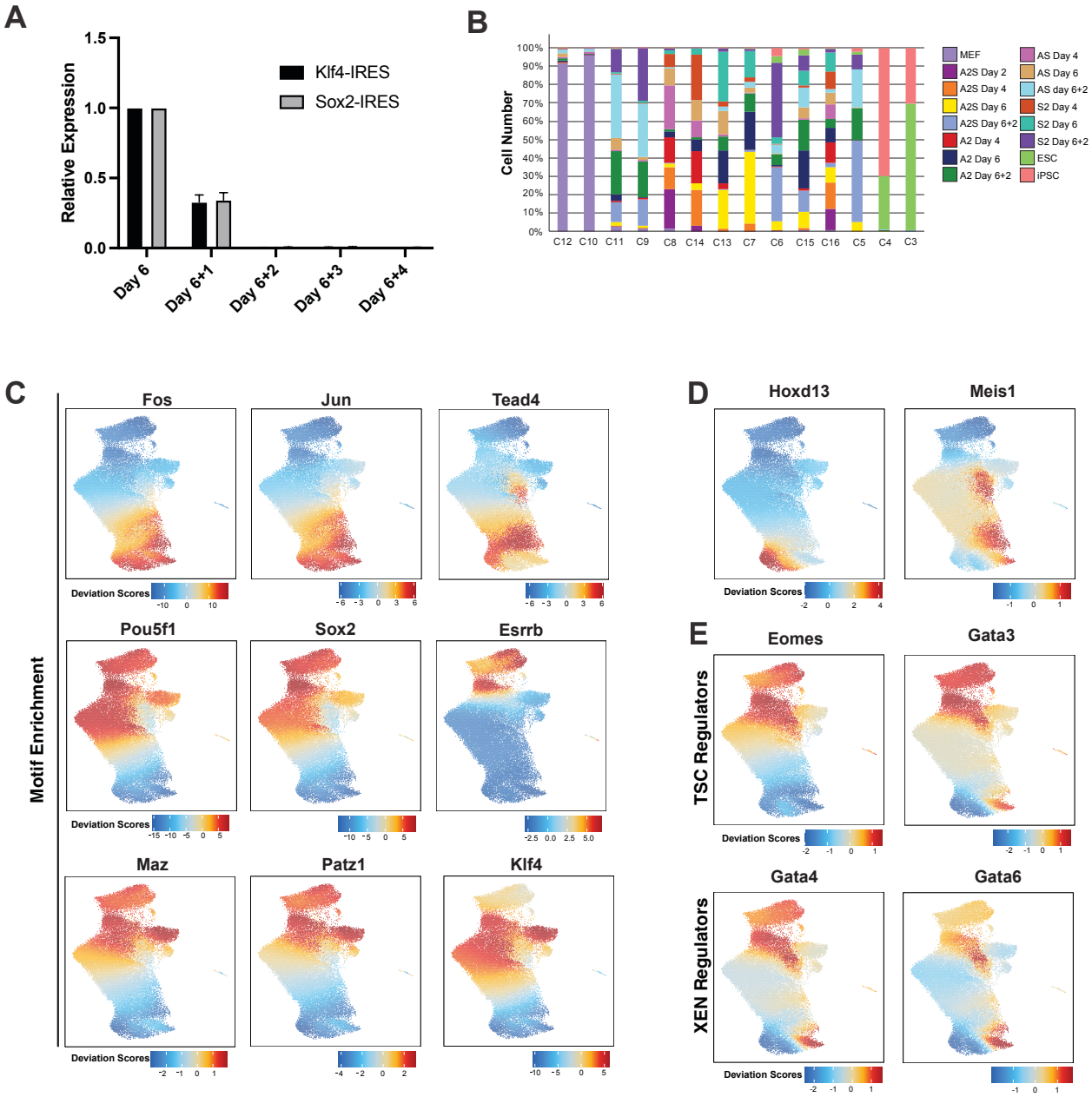

A
Relative Expression
1.5
1.0
0.5
0.0
Klf4-IRES
Sox2-IRES
Day 6
Day 6+1
Day 6+2
Day 6+3
Day 6+4
B
Cell Number
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
C12 C10 C11 C9 C8 C14 C13 C7 C6 C15 C16 C5 C4 C3
MEF
A2S Day 2
A2S Day 4
A2S Day 6
A2S Day 6+2
A2 Day 4
A2 Day 6
A2 Day 6+2
AS Day 4
AS Day 6
AS day 6+2
S2 Day 4
S2 Day 6
S2 Day 6+2
ESC
iPSC
C
Motif Enrichment
Fos
Jun
Tead4
Pou5f1
Sox2
Esrrb
Maz
Patz1
Klf4
Deviation Scores
D
Hoxd13
Meis1
E
TSC Regulators
Eomes
Gata3
XEN Regulators
Gata4
Gata6

Figure S3 (Related to Figure 3)

A) RT-qPCR results for Klf4-IRES and Sox2-IRES on A2S reprogramming Day 6 and throughout 4 days of OSKM withdrawal (Day 6+1 to Day 6+4). Error bars indicate standard deviation between 2 replicate reactions.

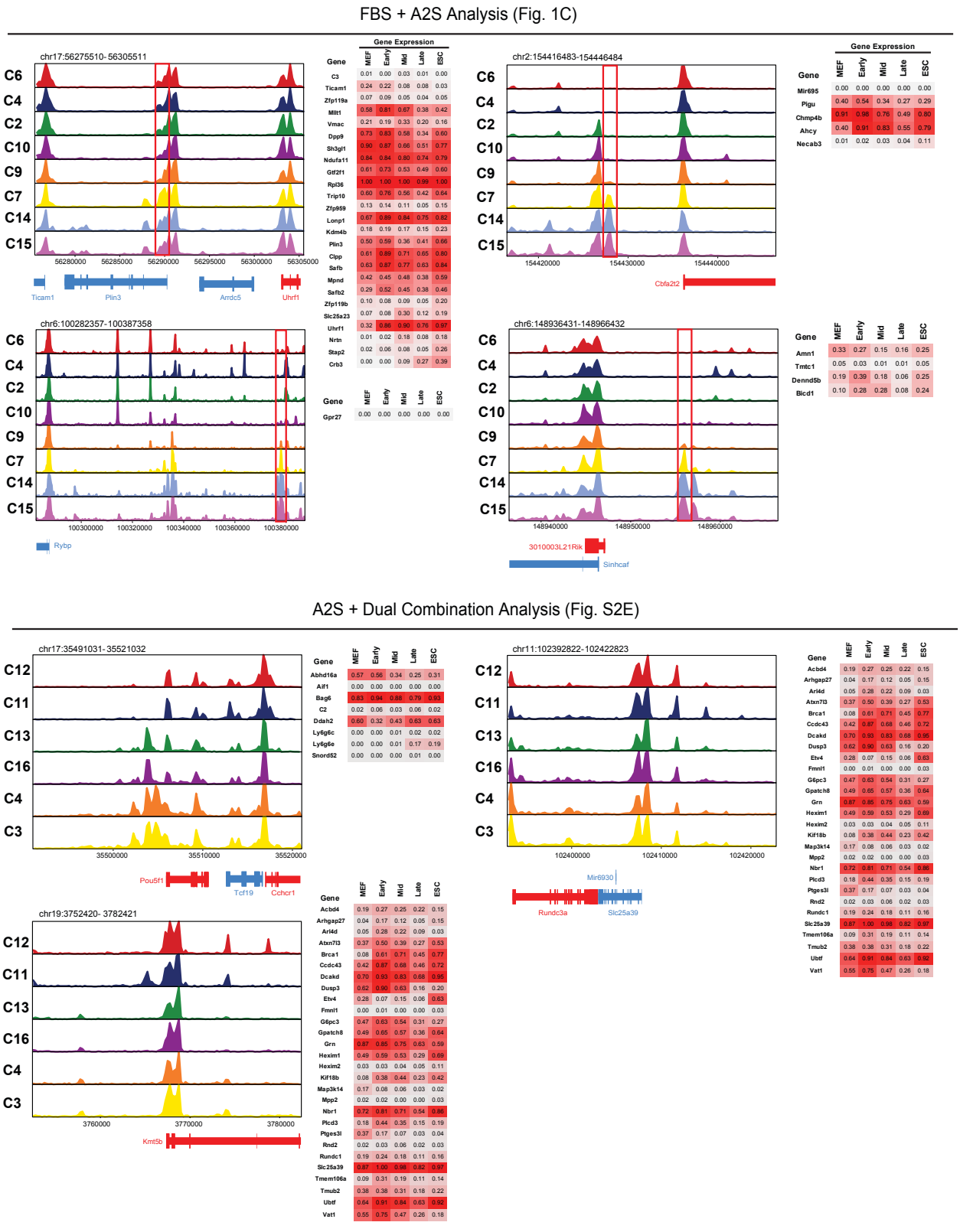
B) Composition of clusters from Fig. 3A presented as percent of cluster coming from each sample.

C) UMAP plots of representative differentially enriched motifs within the reprogramming-associated clusters from Fig. 3A, colored by motif enrichment

D) UMAP plots showing colored by motif enrichment for Hoxd13 and Meis1

E) UMAP plots of representative trophoblast stem cell (TSC) and extraembryonic endoderm (XEN) genes, colored by motif enrichment.

Figure S4


FBS + A2S Analysis (Fig. 1C)
chr17:56275510- 56305511
C6
C4
C2
C10
C9
C7
C14
C15
56280000
56285000
56290000
56295000
56300000
56305000
Ticam1
Plin3
Arrdc5
Uhrf1
Gene Expression
Gene
MEF
Early
Mid
Late
ESC
C3 0.01 0.00 0.03 0.01 0.00
Ticam1 0.24 0.22 0.08 0.08 0.03
Zfp119a 0.07 0.09 0.05 0.04 0.05
Mllt1 0.58 0.81 0.67 0.38 0.42
Vmac 0.21 0.19 0.33 0.20 0.16
Dpp9 0.73 0.83 0.58 0.34 0.60
Sh3gl1 0.90 0.87 0.66 0.51 0.77
Ndufa11 0.84 0.84 0.80 0.74 0.79
Gtf2f1 0.61 0.73 0.53 0.49 0.60
Rpl36 1.00 1.00 1.00 0.99 1.00
Trip10 0.60 0.76 0.56 0.42 0.64
Zfp969 0.13 0.14 0.11 0.05 0.15
Lonp1 0.67 0.89 0.84 0.75 0.82
Kdm4b 0.18 0.19 0.17 0.15 0.23
Plin3 0.50 0.59 0.36 0.41 0.66
Clpp 0.61 0.89 0.71 0.65 0.80
Safb 0.63 0.87 0.77 0.63 0.84
Mpnd 0.42 0.45 0.48 0.38 0.59
Safb2 0.29 0.52 0.45 0.38 0.46
Zfp119b 0.10 0.08 0.09 0.05 0.20
Slc25a23 0.07 0.08 0.30 0.12 0.19
Uhrf1 0.32 0.86 0.90 0.76 0.97
Nrtn 0.01 0.02 0.18 0.08 0.18
Stap2 0.02 0.06 0.08 0.05 0.26
Crb3 0.00 0.00 0.09 0.27 0.39
Gene
MEF
Early
Mid
Late
ESC
Gpr27 0.00 0.00 0.00 0.00 0.00
chr2:154416483-154446484
C6
C4
C2
C10
C9
C7
C14
C15
154420000
154430000
154440000
Cbfa2t2
Gene Expression
Gene
MEF
Early
Mid
Late
ESC
Mir695 0.00 0.00 0.00 0.00 0.00
Pigu 0.40 0.54 0.34 0.27 0.29
Chmp4b 0.91 0.98 0.76 0.49 0.80
Ahcy 0.40 0.91 0.83 0.55 0.79
Necab3 0.01 0.02 0.03 0.04 0.11
chr6:100282357-100387358
C6
C4
C2
C10
C9
C7
C14
C15
100300000
100320000
100340000
100360000
100380000
Rybp
chr6:148936431-148966432
C6
C4
C2
C10
C9
C7
C14
C15
148940000
148950000
148960000
3010003L21Rik
Sinhcaf
Gene
MEF
Early
Mid
Late
ESC
Amn1 0.33 0.27 0.15 0.16 0.25
Tmtc1 0.05 0.03 0.01 0.01 0.05
Denndd5b 0.19 0.39 0.18 0.06 0.25
Bicd1 0.10 0.28 0.28 0.08 0.24
A2S + Dual Combination Analysis (Fig. S2E)
chr17:35491031- 35521032
C12
C11
C13
C16
C4
C3
35500000
35510000
35520000
Pou5f1
Tcf19
Cchcr1
Gene
MEF
Early
Mid
Late
ESC
Abhd16a 0.57 0.56 0.34 0.25 0.31
Aif1 0.00 0.00 0.00 0.00 0.00
Bag6 0.83 0.94 0.88 0.79 0.93
C2 0.02 0.06 0.03 0.06 0.02
Ddah2 0.60 0.32 0.43 0.63 0.63
Ly6g6c 0.00 0.00 0.01 0.02 0.02
Ly6g6e 0.00 0.00 0.01 0.17 0.19
Snord52 0.00 0.00 0.00 0.01 0.00
chr11:102392822-102422823
C12
C11
C13
C16
C4
C3
102400000
102410000
102420000
Mir6930
Rundc3a
Slc25a39
Gene
MEF
Early
Mid
Late
ESC
Acbd4 0.19 0.27 0.25 0.22 0.15
Arhgap27 0.04 0.17 0.12 0.05 0.15
Arl4d 0.05 0.28 0.22 0.09 0.03
Atxn7l3 0.37 0.50 0.39 0.27 0.53
Brca1 0.08 0.61 0.71 0.45 0.77
Ccdc43 0.42 0.87 0.68 0.46 0.72
Dcakd 0.70 0.93 0.83 0.68 0.95
Dusp3 0.62 0.90 0.63 0.16 0.20
Etv4 0.28 0.07 0.15 0.06 0.63
Fmnl1 0.00 0.01 0.00 0.00 0.03
G6pc3 0.47 0.63 0.54 0.31 0.27
Gpatch8 0.49 0.65 0.57 0.36 0.64
Grn 0.87 0.85 0.75 0.63 0.59
Hexim1 0.49 0.59 0.53 0.29 0.69
Hexim2 0.03 0.03 0.04 0.05 0.11
Kif18b 0.08 0.38 0.44 0.23 0.42
Map3k14 0.17 0.08 0.06 0.03 0.02
Mpp2 0.02 0.02 0.00 0.00 0.03
Nbr1 0.72 0.81 0.71 0.54 0.88
Plcd3 0.18 0.44 0.35 0.15 0.19
Ptges3l 0.37 0.17 0.07 0.03 0.04
Rnd2 0.02 0.03 0.06 0.02 0.03
Rundc1 0.19 0.24 0.18 0.11 0.16
Slc25a39 0.87 1.00 0.98 0.82 0.97
Tmem106a 0.09 0.31 0.19 0.11 0.14
Tmub2 0.38 0.38 0.31 0.18 0.22
Ubtf 0.64 0.91 0.84 0.63 0.92
Vat1 0.55 0.75 0.47 0.26 0.18
chr19:3752420- 3782421
C12
C11
C13
C16
C4
C3
3760000
3770000
3780000
Kmt5b
Gene
MEF
Early
Mid
Late
ESC
Acbd4 0.19 0.27 0.25 0.22 0.15
Arhgap27 0.04 0.17 0.12 0.05 0.15
Arl4d 0.05 0.28 0.22 0.09 0.03
Atxn7l3 0.37 0.50 0.39 0.27 0.53
Brca1 0.08 0.61 0.71 0.45 0.77
Ccdc43 0.42 0.87 0.68 0.46 0.72
Dcakd 0.70 0.93 0.83 0.68 0.95
Dusp3 0.62 0.90 0.63 0.16 0.20
Etv4 0.28 0.07 0.15 0.06 0.63
Fmnl1 0.00 0.01 0.00 0.00 0.03
G6pc3 0.47 0.63 0.54 0.31 0.27
Gpatch8 0.49 0.65 0.57 0.36 0.64
Grn 0.87 0.85 0.75 0.63 0.59
Hexim1 0.49 0.59 0.53 0.29 0.69
Hexim2 0.03 0.03 0.04 0.05 0.11
Kif18b 0.08 0.38 0.44 0.23 0.42
Map3k14 0.17 0.08 0.06 0.03 0.02
Mpp2 0.02 0.02 0.00 0.00 0.03
Nbr1 0.72 0.81 0.71 0.54 0.88
Plcd3 0.18 0.44 0.35 0.15 0.19
Ptges3l 0.37 0.17 0.07 0.03 0.04
Rnd2 0.02 0.03 0.06 0.02 0.03
Rundc1 0.19 0.24 0.18 0.11 0.16
Slc25a39 0.87 1.00 0.98 0.82 0.97
Tmem106a 0.09 0.31 0.19 0.11 0.14
Tmub2 0.38 0.38 0.31 0.18 0.22
Ubtf 0.64 0.91 0.84 0.63 0.92
Vat1 0.55 0.75 0.47 0.26 0.18

A2S + Dual Combination Withdrawal Analysis (Fig. 3A)

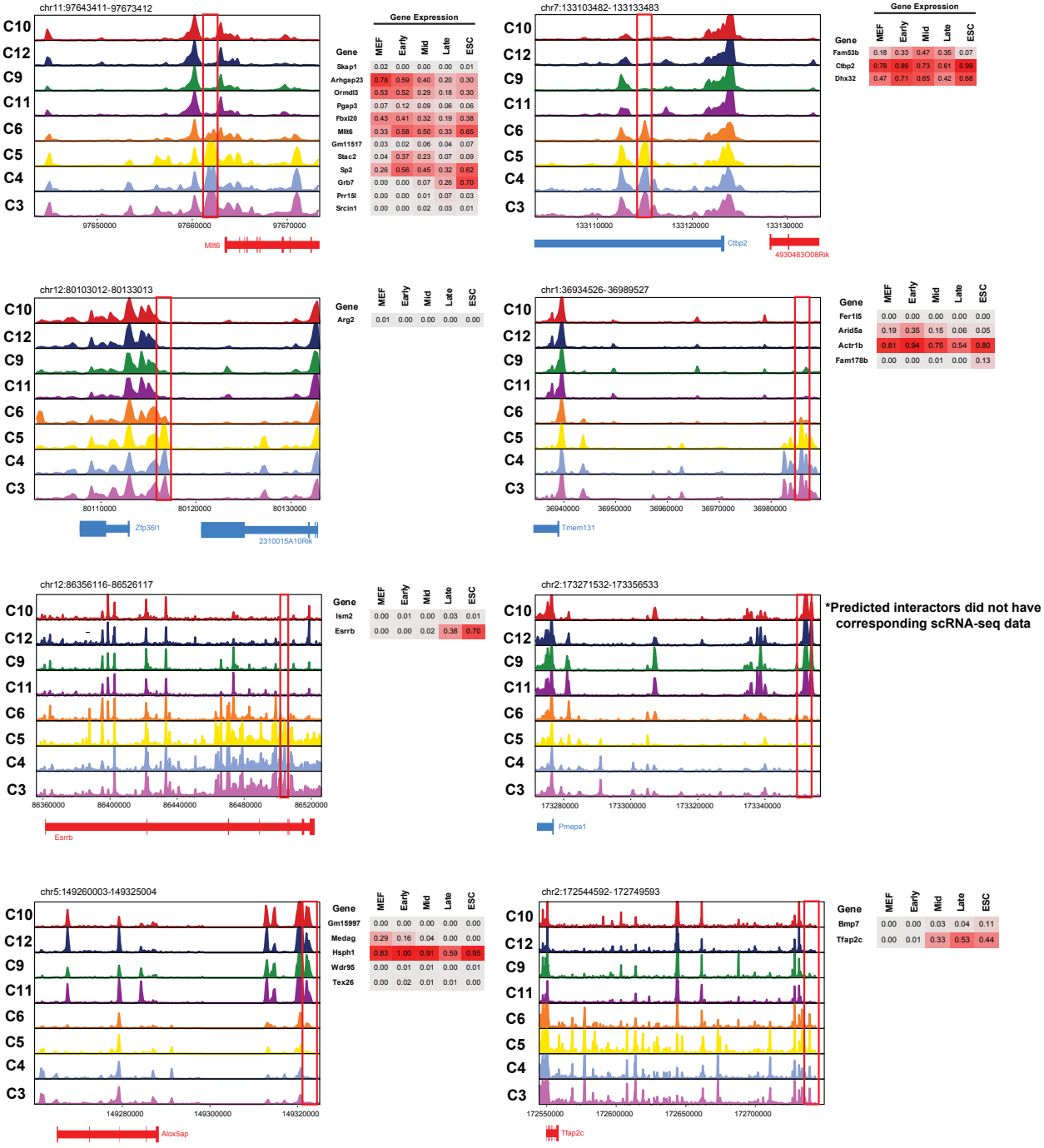
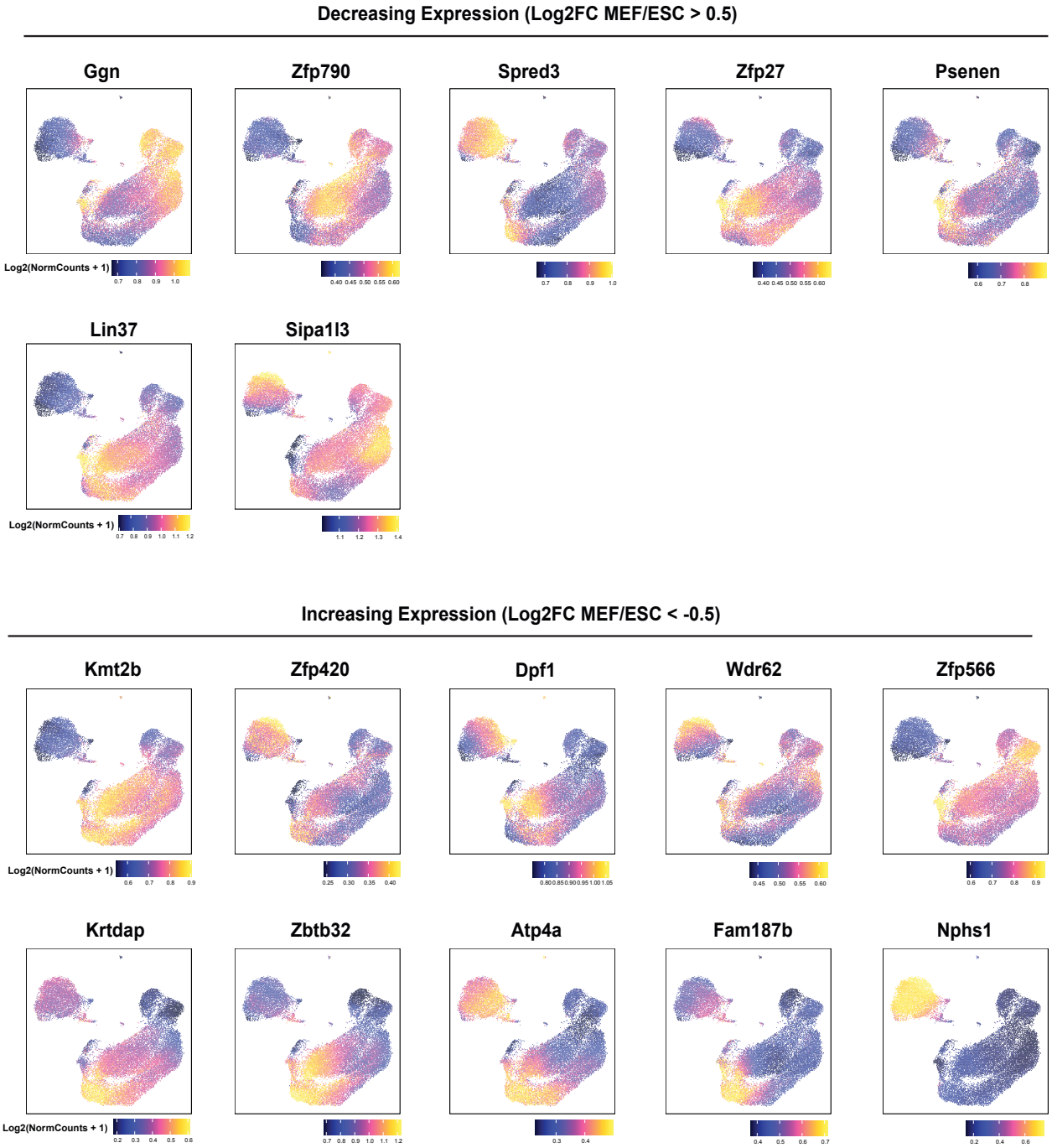


Figure S4 (Related to Figure 4)

ArchR tracks of peaks of interest identified from scCISINT analysis of the data from Fig. 1C, Fig. S2E, and Fig. 3A. Location of peaks are boxed in red. Gene expression of predicted interactors that had associated scRNA-seq data from Tran et al. are shown to the right of each peak. Numbers represent the percent of cells within indicated scRNA-seq cluster expressing that gene.

Figure S5


Decreasing Expression (Log2FC MEF/ESC > 0.5)
Ggn
Zfp790
Spred3
Zfp27
Psenen
Lin37
Sipa1l3
Log2(NormCounts + 1)
Increasing Expression (Log2FC MEF/ESC < -0.5)
Kmt2b
Zfp420
Dpf1
Wdr62
Zfp566
Krtdap
Zbtb32
Atp4a
Fam187b
Nphs1

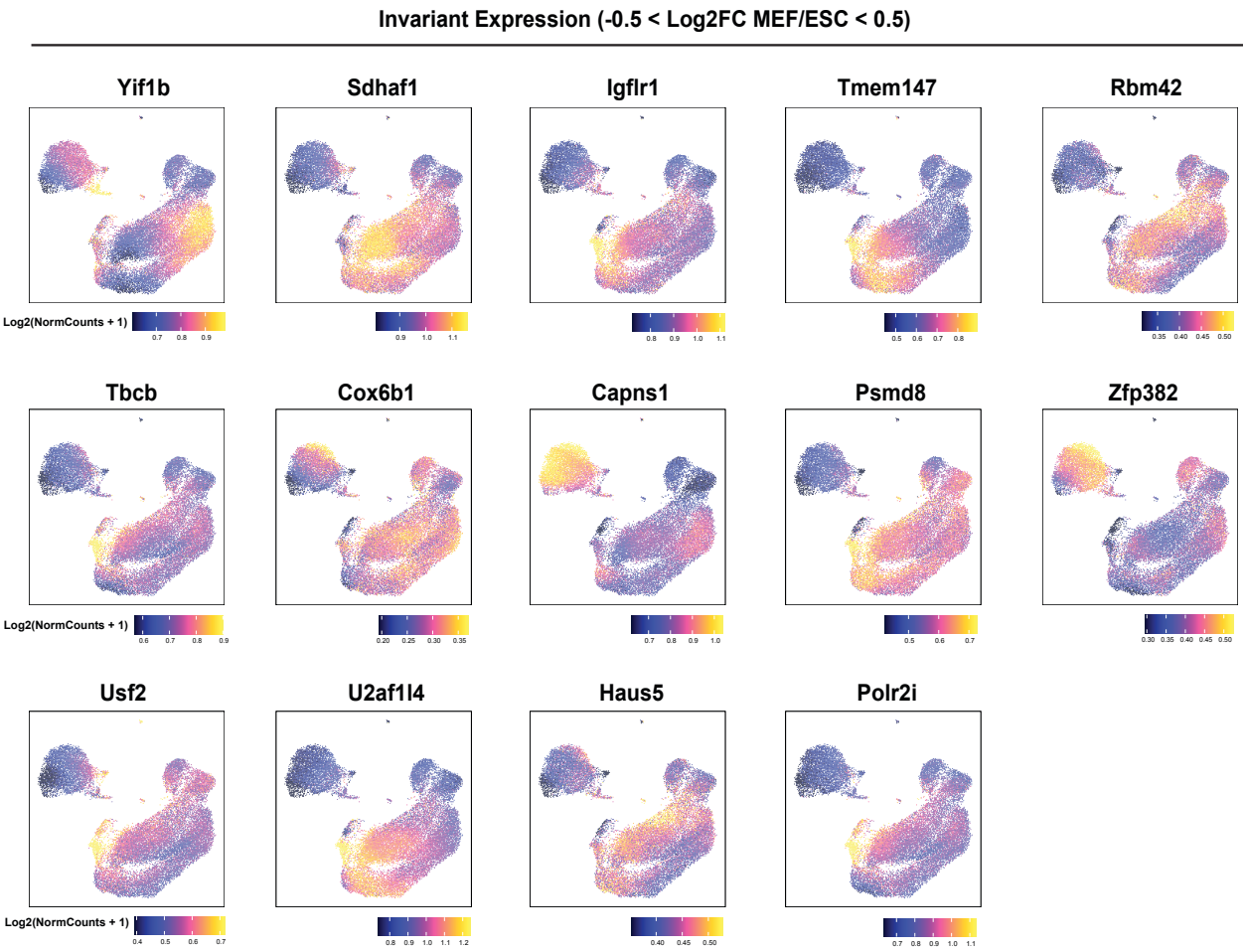


Figure S5 (Related to Figure 5)

UMAP plots (Fig. 1C) of all predicted interactors of Peak 2 listed in Fig. 5B that are not shown in Fig. 5C, colored by promoter accessibility.

References

Apostolou, E., & Hochedlinger, K. (2013). Chromatin dynamics during cellular reprogramming. *Nature*, *502*(7472), Article 7472. https://doi.org/10.1038/nature12749

Awasthi, N., Liongue, C., & Ward, A. C. (2021). STAT proteins: A kaleidoscope of canonical and non-canonical functions in immunity and cancer. *Journal of Hematology & Oncology*, *14*(1), 198. https://doi.org/10.1186/s13045-021-01214-y

Benchetrit, H., Herman, S., van Wietmarschen, N., Wu, T., Makedonski, K., Maoz, N., Yom Tov, N., Stave, D., Lasry, R., Zayat, V., Xiao, A., Lansdorp, P. M., Sebban, S., & Buganim, Y. (2015). Extensive Nuclear Reprogramming Underlies Lineage Conversion into Functional Trophoblast Stem-like Cells. *Cell Stem Cell*, *17*(5), 543–556. https://doi.org/10.1016/j.stem.2015.08.006

Benchetrit, H., Jaber, M., Zayat, V., Sebban, S., Pushett, A., Makedonski, K., Zakheim, Z., Radwan, A., Maoz, N., Lasry, R., Renous, N., Inbar, M., Ram, O., Kaplan, T., & Buganim, Y. (2019). Direct Induction of the Three Pre-implantation Blastocyst Cell Types from Fibroblasts. *Cell Stem Cell*, *24*(6), 983-994.e7. https://doi.org/10.1016/j.stem.2019.03.018

Boland, M. J., Hazen, J. L., Nazor, K. L., Rodriguez, A. R., Gifford, W., Martin, G., Kupriyanov, S., & Baldwin, K. K. (2009). Adult mice generated from induced pluripotent stem cells. *Nature*, *461*(7260), Article 7260. https://doi.org/10.1038/nature08310

Buckberry, S., Liu, X., Poppe, D., Tan, J. P., Sun, G., Chen, J., Nguyen, T. V., de Mendoza, A., Pflueger, J., Frazer, T., Vargas-Landín, D. B., Paynter, J. M., Smits, N., Liu, N., Ouyang, J. F., Rossello, F. J., Chy, H. S., Rackham, O. J. L., Laslett, A. L., … Lister, R. (2023). Transient naive reprogramming corrects hiPS cells functionally and epigenetically. *Nature*, *620*(7975), Article 7975. https://doi.org/10.1038/s41586-023-06424-7

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), Article 12. https://doi.org/10.1038/nmeth.2688

Buganim, Y., Faddah, D. A., & Jaenisch, R. (2013). Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics*, *14*(6), Article 6. https://doi.org/10.1038/nrg3473

Cao, S., Yu, S., Li, D., Ye, J., Yang, X., Li, C., Wang, X., Mai, Y., Qin, Y., Wu, J., He, J., Zhou, C., Liu, H., Zhao, B., Shu, X., Wu, C., Chen, R., Chan, W., Pan, G., … Pei, D. (2018). Chromatin Accessibility Dynamics during Chemical Induction of

Pluripotency. *Cell Stem Cell*, *22*(4), 529-542.e5. https://doi.org/10.1016/j.stem.2018.03.005

Chen, J., Liu, H., Liu, J., Qi, J., Wei, B., Yang, J., Liang, H., Chen, Y., Chen, J., Wu, Y., Guo, L., Zhu, J., Zhao, X., Peng, T., Zhang, Y., Chen, S., Li, X., Li, D., Wang, T., & Pei, D. (2013). H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature Genetics*, *45*(1), Article 1. https://doi.org/10.1038/ng.2491

Chen, J., Liu, J., Chen, Y., Yang, J., Chen, J., Liu, H., Zhao, X., Mo, K., Song, H., Guo, L., Chu, S., Wang, D., Ding, K., & Pei, D. (2011). Rational optimization of reprogramming culture conditions for the generation of induced pluripotent stem cells with ultra-high efficiency and fast kinetics. *Cell Research*, *21*(6), Article 6. https://doi.org/10.1038/cr.2011.51

Chen, K., Long, Q., Xing, G., Wang, T., Wu, Y., Li, L., Qi, J., Zhou, Y., Ma, B., Schöler, H. R., Nie, J., Pei, D., & Liu, X. (2020). Heterochromatin loosening by the Oct4 linker region facilitates Klf4 binding and iPSC reprogramming. *The EMBO Journal*, *39*(1), e99165. https://doi.org/10.15252/embj.201899165

Chen, X., Lu, Y., Wang, L., Ma, X., Pu, J., Lin, L., Deng, Q., Li, Y., Wang, W., Jin, Y., Hu, Z., Zhou, Z., Chen, G., Jiang, L., Wang, H., Zhao, X., He, X., Fu, J., Russ, H. A., … Zhu, S. (2023). A fast chemical reprogramming system promotes cell identity transition through a diapause-like state. *Nature Cell Biology*, *25*(8), Article 8. https://doi.org/10.1038/s41556-023-01193-x

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., & Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell*, *168*(3), 442-459.e20. https://doi.org/10.1016/j.cell.2016.12.016

Di Giammartino, D. C., Kloetgen, A., Polyzos, A., Liu, Y., Kim, D., Murphy, D., Abuhashem, A., Cavaliere, P., Aronson, B., Shah, V., Dephoure, N., Stadtfeld, M., Tsirigos, A., & Apostolou, E. (2019). KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nature Cell Biology*, *21*(10), 1179–1190. https://doi.org/10.1038/s41556-019-0390-6

Di Stefano, B., Collombet, S., Jakobsen, J. S., Wierer, M., Sardina, J. L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., Limone, F., Mann, M., Porse, B., Thieffry, D., & Graf, T. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. *Nature Cell Biology*, *18*(4), Article 4. https://doi.org/10.1038/ncb3326

Esteban, M. A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S., Chen, K., Li, Y., Liu, X., Xu, J., Zhang, S., Li, F., He, W., Labuda, K.,

Song, Y., … Pei, D. (2010). Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells. *Cell Stem Cell*, *6*(1), 71–79. https://doi.org/10.1016/j.stem.2009.12.001

Fedele, M., Crescenzi, E., & Cerchia, L. (2017). The POZ/BTB and AT-Hook Containing Zinc Finger 1 (PATZ1) Transcription Regulator: Physiological Functions and Disease Involvement. *International Journal of Molecular Sciences*, *18*(12), Article 12. https://doi.org/10.3390/ijms18122524

Gaspar-Maia, A., Alajem, A., Meshorer, E., & Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology*, *12*(1), Article 1. https://doi.org/10.1038/nrm3036

Guan, J., Wang, G., Wang, J., Zhang, Z., Fu, Y., Cheng, L., Meng, G., Lyu, Y., Zhu, J., Li, Y., Wang, Y., Liuyang, S., Liu, B., Yang, Z., He, H., Zhong, X., Chen, Q., Zhang, X., Sun, S., … Deng, H. (2022). Chemical reprogramming of human somatic cells to pluripotent stem cells. *Nature*, 1–7. https://doi.org/10.1038/s41586-022-04593-5

Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., Ge, J., Xu, J., Zhang, Q., Zhao, Y., & Deng, H. (2013). Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. *Science*, *341*(6146), 651–654. https://doi.org/10.1126/science.1239278

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A. E., & Melton, D. A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nature Biotechnology*, *26*(7), Article 7. https://doi.org/10.1038/nbt1418

Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W., & Melton, D. A. (2008). Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nature Biotechnology*, *26*(11), Article 11. https://doi.org/10.1038/nbt.1502

Hubert, K. A., & Wellik, D. M. (2023). Hox genes in development and beyond. *Development*, *150*(1), dev192476. https://doi.org/10.1242/dev.192476

Jaber, M., Radwan, A., Loyfer, N., Abdeen, M., Sebban, S., Khatib, A., Yassen, H., Kolb, T., Zapatka, M., Makedonski, K., Ernst, A., Kaplan, T., & Buganim, Y. (2022). Comparative parallel multi-omics analysis during the induction of pluripotent and trophectoderm states. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31131-8

Jackson, S. A., Olufs, Z. P. G., Tran, K. A., Zaidan, N. Z., & Sridharan, R. (2016). Alternative Routes to Induced Pluripotent Stem Cells Revealed by

Reprogramming of the Neural Lineage. *Stem Cell Reports*, *6*(3), 302–311. https://doi.org/10.1016/j.stemcr.2016.01.009

Jain, N., Goyal, Y., Dunagin, M. C., Cote, C. J., Mellis, I. A., Emert, B., Jiang, C. L., Dardani, I. P., Reffsin, S., & Raj, A. (2023). *Retrospective identification of intrinsic factors that mark pluripotency potential in rare somatic cells* (p. 2023.02.10.527870). bioRxiv. https://doi.org/10.1101/2023.02.10.527870

Knaupp, A. S., Buckberry, S., Pflueger, J., Lim, S. M., Ford, E., Larcombe, M. R., Rossello, F. J., Mendoza, A. de, Alaei, S., Firas, J., Holmes, M. L., Nair, S. S., Clark, S. J., Nefzger, C. M., Lister, R., & Polo, J. M. (2017). Transient and Permanent Reconfiguration of Chromatin and Transcription Factor Occupancy Drive Reprogramming. *Cell Stem Cell*, *21*(6), 834-845.e6. https://doi.org/10.1016/j.stem.2017.11.007

Lentjes, M. H., Niessen, H. E., Akiyama, Y., Bruïne, A. P. de, Melotte, V., & Engeland, M. van. (2016). The emerging role of GATA transcription factors in development and disease. *Expert Reviews in Molecular Medicine*, *18*, e3. https://doi.org/10.1017/erm.2016.2

Li, D., Liu, J., Yang, X., Zhou, C., Guo, J., Wu, C., Qin, Y., Guo, L., He, J., Yu, S., Liu, H., Wang, X., Wu, F., Kuang, J., Hutchins, A. P., Chen, J., & Pei, D. (2017). Chromatin Accessibility Dynamics during iPSC Reprogramming. *Cell Stem Cell*, *21*(6), 819-833.e6. https://doi.org/10.1016/j.stem.2017.10.012

Liu, X., Ouyang, J. F., Rossello, F. J., Tan, J. P., Davidson, K. C., Valdes, D. S., Schröder, J., Sun, Y. B. Y., Chen, J., Knaupp, A. S., Sun, G., Chy, H. S., Huang, Z., Pflueger, J., Firas, J., Tano, V., Buckberry, S., Paynter, J. M., Larcombe, M. R., … Polo, J. M. (2020). Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature*, *586*(7827), Article 7827. https://doi.org/10.1038/s41586-020-2734-6

Liuyang, S., Wang, G., Wang, Y., He, H., Lyu, Y., Cheng, L., Yang, Z., Guan, J., Fu, Y., Zhu, J., Zhong, X., Sun, S., Li, C., Wang, J., & Deng, H. (2023). Highly efficient and rapid generation of human pluripotent stem cells by chemical reprogramming. *Cell Stem Cell*, *30*(4), 450-459.e9. https://doi.org/10.1016/j.stem.2023.02.008

Ma, H., Ow, J. R., Tan, B. C. P., Goh, Z., Feng, B., Loh, Y. H., Fedele, M., Li, H., & Wu, Q. (2014). The dosage of Patz1 modulates reprogramming process. *Scientific Reports*, *4*(1), Article 1. https://doi.org/10.1038/srep07519

Mikkelsen, T. S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B. E., Jaenisch, R., Lander, E. S., & Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature*, *454*(7200), Article 7200. https://doi.org/10.1038/nature07056

Naama, M., Rahamim, M., Zayat, V., Sebban, S., Radwan, A., Orzech, D., Lasry, R., Ifrah, A., Jaber, M., Sabag, O., Yassen, H., Khatib, A., Epsztejn-Litman, S., Novoselsky-Persky, M., Makedonski, K., Deri, N., Goldman-Wohl, D., Cedar, H., Yagel, S., … Buganim, Y. (2023). Pluripotency-independent induction of human trophoblast stem cells from fibroblasts. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-39104-1

Nair, S., Ameen, M., Sundaram, L., Pampari, A., Schreiber, J., Balsubramani, A., Wang, Y. X., Burns, D., Blau, H. M., Karakikes, I., Wang, K. C., & Kundaje, A. (2023). *Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency* (p. 2023.10.04.560808). bioRxiv. https://doi.org/10.1101/2023.10.04.560808

Okita, K., Ichisaka, T., & Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, *448*(7151), Article 7151. https://doi.org/10.1038/nature05934

Onder, T. T., Kara, N., Cherry, A., Sinha, A. U., Zhu, N., Bernt, K. M., Cahan, P., Mancarci, B. O., Unternaehrer, J., Gupta, P. B., Lander, E. S., Armstrong, S. A., & Daley, G. Q. (2012). Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, *483*(7391), Article 7391. https://doi.org/10.1038/nature10953

Ortabozkoyun, H., Huang, P.-Y., Cho, H., Narendra, V., LeRoy, G., Gonzalez-Buendia, E., Skok, J. A., Tsirigos, A., Mazzoni, E. O., & Reinberg, D. (2022). CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. *Nature Genetics*, *54*(2), Article 2. https://doi.org/10.1038/s41588-021-01008-5

Paikari, A., D. Belair, C., Saw, D., & Blelloch, R. (2017). The eutheria-specific miR-290 cluster modulates placental growth and maternal-fetal transport. *Development*, *144*(20), 3731–3743. https://doi.org/10.1242/dev.151654

Papp, B., & Plath, K. (2013). Epigenetics of Reprogramming to Induced Pluripotency. *Cell*, *152*(6), 1324–1343. https://doi.org/10.1016/j.cell.2013.02.043

Parenti, A., Halbisen, M. A., Wang, K., Latham, K., & Ralston, A. (2016). OSKM Induce Extraembryonic Endoderm Stem Cells in Parallel to Induced Pluripotent Stem Cells. *Stem Cell Reports*, *6*(4), 447–455. https://doi.org/10.1016/j.stemcr.2016.02.003

Rong, O., MaHui, JeanAngela, GohZiyi, Hwa, L., Mei, C., SoongRichie, FuXin-Yuan, YangHenry, & WuQiang. (2013). Patz1 Regulates Embryonic Stem Cell Identity. *Stem Cells and Development*. https://doi.org/10.1089/scd.2013.0430

Schulte, D., & Geerts, D. (2019). MEIS transcription factors in development and disease. *Development*, *146*(16), dev174706. https://doi.org/10.1242/dev.174706

Shakiba, N., Fahmy, A., Jayakumaran, G., McGibbon, S., David, L., Trcka, D., Elbaz, J., Puri, M. C., Nagy, A., van der Kooy, D., Goyal, S., Wrana, J. L., & Zandstra, P. W. (2019). Cell competition during reprogramming gives rise to dominant clones. *Science*, *364*(6438), eaan0925. https://doi.org/10.1126/science.aan0925

Shi, Y., Desponts, C., Do, J. T., Hahm, H. S., Schöler, H. R., & Ding, S. (2008). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Oct4 and Klf4 with Small-Molecule Compounds. *Cell Stem Cell*, *3*(5), 568–574. https://doi.org/10.1016/j.stem.2008.10.004

Soufi, A., Donahue, G., & Zaret, K. S. (2012). Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell*, *151*(5), 994–1004. https://doi.org/10.1016/j.cell.2012.09.045

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M., Carey, M., Garcia, B. A., & Plath, K. (2013). Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. *Nature Cell Biology*, *15*(7), 872–882. https://doi.org/10.1038/ncb2768

Sridharan, R., Tchieu, J., Mason, M. J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., & Plath, K. (2009). Role of the Murine Reprogramming Factors in the Induction of Pluripotency. *Cell*, *136*(2), 364–377. https://doi.org/10.1016/j.cell.2009.01.001

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676. https://doi.org/10.1016/j.cell.2006.07.024

Tran, K. A., Jackson, S. A., Olufs, Z. P. G., Zaidan, N. Z., Leng, N., Kendziorski, C., Roy, S., & Sridharan, R. (2015). Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. *Nature Communications*, *6*(1), Article 1. https://doi.org/10.1038/ncomms7188

Tran, K. A., Pietrzak, S. J., Zaidan, N. Z., Siahpirani, A. F., McCalla, S. G., Zhou, A. S., Iyer, G., Roy, S., & Sridharan, R. (2019). Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. *Cell Reports*, *27*(6), 1726-1741.e5. https://doi.org/10.1016/j.celrep.2019.04.056

Wang, B., Wu, L., Li, D., Liu, Y., Guo, J., Li, C., Yao, Y., Wang, Y., Zhao, G., Wang, X., Fu, M., Liu, H., Cao, S., Wu, C., Yu, S., Zhou, C., Qin, Y., Kuang, J., Ming, J., … Pei, D. (2019). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Jdp2-Jhdm1b-Mkk6-Glis1-Nanog-Essrb-Sall4. *Cell Reports*, *27*(12), 3473-3485.e5. https://doi.org/10.1016/j.celrep.2019.05.068

Wang, Y., Chen, S., Jiang, Q., Deng, J., Cheng, F., Lin, Y., Cheng, L., Ye, Y., Chen, X., Yao, Y., Zhang, X., Shi, G., Dai, L., Su, X., Peng, Y., & Deng, H. (2020). TFAP2C facilitates somatic cell reprogramming by inhibiting c-Myc-dependent apoptosis and promoting mesenchymal-to-epithelial transition. *Cell Death & Disease*, *11*(6), 1–15. https://doi.org/10.1038/s41419-020-2684-9

Wille, C. K., Neumann, E. N., Deshpande, A. J., & Sridharan, R. (2023). *DOT1L interaction partner AF10 controls patterning of H3K79 methylation and RNA polymerase II to maintain cell identity* (p. 2020.12.17.423347). bioRxiv. https://doi.org/10.1101/2020.12.17.423347

Wille, C. K., & Sridharan, R. (2022). DOT1L inhibition enhances pluripotency beyond acquisition of epithelial identity and without immediate suppression of the somatic transcriptome. *Stem Cell Reports*, *17*(2), 384–396. https://doi.org/10.1016/j.stemcr.2021.12.004

Wille, C. K., Zhang, X., Haws, S. A., Denu, J. M., & Sridharan, R. (2023). DOT1L is a barrier to histone acetylation during reprogramming to pluripotency. *Science Advances*, *9*(46), eadf3980. https://doi.org/10.1126/sciadv.adf3980

Xing, Q. R., El Farran, C., Gautam, P., Chuah, Y. S., Warrier, T., Toh, C.-X. D., Kang, N.-Y., Sugii, S., Chang, Y.-T., Xu, J., Collins, J. J., Daley, G. Q., Li, H., Zhang, L.-F., & Loh, Y.-H. (2020). Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Science Advances*, *6*(37), eaba1190. https://doi.org/10.1126/sciadv.aba1190

Ye, J., Ge, J., Zhang, X., Cheng, L., Zhang, Z., He, S., Wang, Y., Lin, H., Yang, W., Liu, J., Zhao, Y., & Deng, H. (2016). Pluripotent stem cells induced from mouse neural stem cells and small intestinal epithelial cells by small molecule compounds. *Cell Research*, *26*(1), Article 1. https://doi.org/10.1038/cr.2015.142

Yuan, K., Ai, W.-B., Wan, L.-Y., Tan, X., & Wu, J.-F. (2017). The miR-290-295 cluster as multi-faceted players in mouse embryonic stem cells. *Cell & Bioscience*, *7*(1), 38. https://doi.org/10.1186/s13578-017-0166-2

Zhao, X., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C., Ma, Q., Wang, L., Zeng, F., & Zhou, Q. (2009). iPS cells produce viable mice through tetraploid complementation. *Nature*, *461*(7260), Article 7260. https://doi.org/10.1038/nature08267

Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., Zhu, J., Meng, G., Ge, J., Yang, S., Cheng, L., Du, Y., Zhao, C., Wang, T., Su, L., Yang, W., & Deng, H. (2015). A XEN-like State Bridges Somatic Cells to Pluripotency during Chemical Reprogramming. *Cell*, *163*(7), 1678–1691. https://doi.org/10.1016/j.cell.2015.11.017

Chapter 4

Discussion and Future Directions

Introduction

The reprogramming of differentiated somatic cells back into a pluripotent state (iPSCs) via ectopic expression of the transcription factors OCT4, SOX2, KLF4, and MYC (OSKM) has been an incredible discovery (Takahashi & Yamanaka, 2006) that represents a tremendous capacity to change cell fate. Therefore, a comprehensive understanding of the molecular underpinnings of this process is essential for translating this process into future therapeutic applications. Despite the promise of iPSCs, there remains some shortcomings and issues associated with somatic cell reprogramming. It is an inefficient process, with cell-to-cell variability in reprogramming kinetics (Apostolou & Hochedlinger, 2013; Buganim et al., 2013; Papp & Plath, 2013), leading to heterogeneous reprogramming populations.

In this thesis, my work has attempted to overcome some of these issues with reprogramming in a couple different ways. I have used a combination of epigenetic- and cell signaling-modifying compounds to improve reprogramming efficiency over 10-fold. Additionally, I have implemented single-cell technology to bypass the heterogeneity of reprogramming populations and uncover transcriptional and chromatin accessibility dynamics associated with successful reprogramming. Thus, this research has provided new insights into the regulatory mechanisms of reprogramming and the factors that influence its efficiency.

Improving Reprogramming Efficiency with Small Molecules

Previous studies have supplemented reprogramming cultures with epigenetic- (Chen et al., 2011; Esteban et al., 2010; Onder et al., 2012; Tran et al., 2015) or

signaling-modifying small molecules (Bar-Nur et al., 2014; Vidal et al., 2014) to enhance reprogramming efficiency. Our lab had previously shown that combining ascorbic acid (AA) and GSK-3 and MAPK signaling pathway inhibitors (2i) were effective at converting partially reprogrammed iPSCs (pre-iPSCs) to bona fide iPSCs (Tran et al., 2015). Here, I combined AA and 2i with an inhibitor of the only known H3K79 methyltransferase Dot1l (SGC0946), altogether referred to as A2S. Both ascorbic acid and SGC0946 act antagonistically to histone modifications that are enriched in mouse embryonic fibroblasts (MEFs) (H3K9 and H3K79 methylation, respectively) (Sridharan et al., 2013), while 2i promotes ESC self-renewal and maintains ESCs in a naïve ground state, resembling preimplantation blastocyst ICM cells (Sim et al., 2017; Ying et al., 2008; Ying & Smith, 2017). In combining these three chemicals, we synergistically improved reprogramming efficiency from about 3% at 12 days of reprogramming in normal serum-based media (FBS) to about 42% by day 6 in A2S.

To elucidate the contribution of each component of A2S on improving reprogramming efficiency, we performed reprogramming with each dual combination of the three components. Using single-cell sequencing data from these three systems, we were able to computationally identify how different gene regulatory networks are affected by each combination of chemicals. In doing so, it was revealed that 2i is primarily responsible for suppressing the somatic regulatory networks, as these networks are still strong and active in the absence of 2i (AS). Conversely, the epigenetic-modifying small molecules AA and SGC both work to coordinate the activation of pluripotency transcriptional networks. Adding on to these findings, analysis of single-cell chromatin accessibility data from dual combination reprogramming further

illustrated that 2i plays a role in , as loss of 2i causes cells to fall behind A2S in the progression towards iPSCs, paralleling what we found from our transcriptional network analysis. The combined efforts of AA and SGC not only aid in the activation of the pluripotent network, but also play key roles in maintenance and stabilization of the pluripotent network, preventing cells from reverting back to a MEF-like state upon withdrawal of exogenous OSKM expression. Further research is required to determine the exact mechanisms by which these small molecules are causing these changes.

We rationally chose ascorbic acid and SGC0946 as they influence histone modifications that are specifically enriched in MEFs. It has been previously shown that the combination of GSK-3 inhibition (one half of 2i) with TGF-beta signaling inhibition and AA was capable of enhancing iPSC colony formation in multiple cell types, including granulocyte-macrophage progenitors (GMPs) and pro-B cells (Vidal et al., 2014). However, it is unclear how our combination of A2S will affect reprogramming efficiency in other cell types or if our results may be specific to reprogramming of MEFs and other fibroblast or mesenchymal cell types. It would be interesting to test the effect of A2S or other combinations of chemicals on other cell types. Additionally, performing a screen of different combinations of small molecules to identify those that are most effective in each type of starting cell population, providing a framework for how to most efficiently generate iPSCs from a specific cell type.

Rather than using small molecules to enhance transcription factor (TF)-mediated reprogramming, as we have done, other studies have completely replaced TFs with combinations of small molecules as a means of inducing and carrying out reprogramming (Hou et al., 2013; Zhao et al., 2015; Chen 2023; Ye; Guan). Therefore,

one interesting area of research to tackle would be a direct comparison of A2S or other small molecule-mediated high-efficiency Yamanaka factor reprogramming with an all chemical reprogramming within the same cellular context (same cell type, embryo, …). This could provide new insight into the advantages or disadvantages of a combined small molecule and TF reprogramming system against a solely chemical one.

Analysis of Transcriptional Dynamics of Reprogramming

Reprogramming populations are confounded by heterogeneity, with cells experiencing varying rates at which they reprogram. Additionally, cells may fall off the main reprogramming trajectory towards alternate cell fates. In fact, surface markers such as the ESC-expressed SSEA1 and MEF-expressed THY1 genes have been used to identify cells that are more susceptible to reprogram successfully vs those that are refractory to pluripotency acquisition (Brambrink et al., 2008; Polo et al., 2012; Stadtfeld et al., 2008). These markers also identify populations of cells that have become stalled partially reprogrammed intermediate cell lines (pre-iPSCs) (Mikkelsen et al., 2008; Sridharan et al., 2009, 2013). Cells can also transition into cell types other than iPSCs. For example, it was shown that some of the reprogramming cells become induced extraembryonic endoderm (iXEN) cells in parallel with those that become iPSCs (Parenti et al., 2016), and reprogramming cells are also capable of becoming induced trophoblast stem cells (iTSCs), whose formation is enhanced through ectopic expression of TSC-associated TFs (Benchetrit et al., 2015, 2019; Naama et al., 2023).

These factors contribute to the inherent heterogeneity observed in reprogramming, and make population-based studies of reprogramming challenging, as

the features of successfully reprogramming cells are confounded by the various sub-populations of cells that arise during this process and the differential rates of reprogramming between cells. That being said, previous studies have utilized bulk population-based RNA-seq to identify the transcriptional changes that occur during reprogramming, and have used this information to develop a reprogramming dogma that describes reprogramming occurring in a step-wise temporal manner. Included in these steps are an initial downregulation of somatic genes, which is followed by a mesenchymal-to-epithelial transition (MET) (Hussein et al., 2014; R. Li et al., 2010; Mikkelsen et al., 2008; Samavarchi-Tehrani et al., 2010). The latter steps of reprogramming are characterized by an upregulation of pluripotency-associated genes and culminate in stabilization of the pluripotent network in iPSCs that can maintain their PSC state independent of exogenous OSKM expression (Apostolou & Hochedlinger, 2013; Apostolou & Stadtfeld, 2018; Golipour et al., 2012; Mikkelsen et al., 2008).

In Chapter 2 of this thesis, I found that my single-cell RNA-seq analysis of reprogramming cells disputes and challenges this established reprogramming dogma of it being a temporal series of events. Interestingly, we found that there were populations of cells that could begin the MET step prior to completely losing the somatic transcription, with cells co-expressing the MET-associated *Cdh1* with mesenchymal genes, such as *Twist1*, and the MEF marker *Thy1* to varying degrees. Similarly, upregulation of the cell cycle-associated genes *Mcm5* and *Bub1b* also occurred before complete loss of somatic gene expression. Even more strikingly, we demonstrate that cells can already upregulate expression of the pluripotency marker *Nanog* while there is still persistent expression of the mesenchymal gene *Twist1*, albeit this co-expression is

more prevalent in cells that branch off the main reprogramming trajectory. All of these findings indicate that each of these steps are independently regulated as evidenced by co-expression of somatic, MET, pluripotent, and cell cycling genes within individual cells, information that is unobtainable in population-based analyses.

Previous studies have found that some genes display a transient upregulation during reprogramming such as those associated with the primitive streak (Nefzger et al., 2017) as well as the surface marker Sca-1 (Schwarz et al., 2018). In our analysis, we also uncovered genes that display a transient pattern, including the epithelial-associated transcription factor *Ehf*. Knockdown of *Ehf* revealed that although it is not expressed in either MEFs or ESCs, its depletion actually compromises reprogramming. Further investigation is required to determine the exact mechanism by which EHF activity affects reprogramming

The cells from the scRNA-seq dataset were ordered along a pseudotime trajectory, which revealed branch points representing cells that veer away from the primary reprogramming pathway. Branching cells notably failed to upregulate or downregulate the appropriate pluripotent or somatic genes (Nanog and Twist1, respectively), but the branchpoint analysis did also find a role for the translation initiation gene Eif4a1. It's also important to note that the cells that displayed co-expression of both Twist1 and Nanog largely fell in these branches as well. Further investigation into expression and co-expression of genes highlighted how A2S accelerates normal FBS reprogramming. We observed more coordinated upregulation of groups of pluripotency-related genes, such as *Epcam*, *Sall4*, *Nanog*, and *Tdgf1*. Meanwhile in FBS, these genes were not all upregulated within the same cells. One of these subsets of genes

contained the pluripotency markers *Dppa4* and *Lin28a*, as well as the placental growth-associated *Phlda2*, whose knockdown led to a decrease in DPPA4+ colonies. Additionally, in A2S reprogramming cells, the cell cycle-associated genes *Mcm5* and *Bub1b* are more consistently expressed throughout the reprogramming process, while also repressing the antiproliferative gene *Cdkn1c*, which are downregulated and upregulated respectively in FBS reprogramming. Thus, A2S allows cells to circumvent the reprogramming-associated senescence block that normal reprogramming cells experience (Banito et al., 2009; H. Li et al., 2009; Mikkelsen et al., 2008). Together, these results suggest that A2S is able to bypass many of the transcriptional and senescence barriers that are inhibitory to reprogramming, and thereby accelerating the transition to pluripotency.

Analysis of Chromatin Accessibility Dynamics in Reprogramming

Given the shifting epigenetic landscape (J. Chen et al., 2013; Gaspar-Maia et al., 2011; Onder et al., 2012; Soufi et al., 2012; Sridharan et al., 2013) and influence of epigenome-modifying small molecules on reprogramming efficiency (Esteban et al., 2010; Huangfu, Maehr, et al., 2008; Huangfu, Osafune, et al., 2008; Mikkelsen et al., 2008; Shi et al., 2008; Tran et al., 2015), recent research has turned to the assay for transposase-accessible chromatin with sequencing (ATAC-seq) methodology (Buenrostro et al., 2013) to examine chromatin accessibility dynamics in the context of somatic cell reprogramming.

Similar to RNA-seq studies, population-based ATAC-seq analyses have been conducted to elucidate chromatin dynamics in populations of reprogramming cells

(Buckberry et al., 2023; Cao et al., 2018; K. Chen et al., 2020; X. Chen et al., 2023; Chronis et al., 2017; Di Giammartino et al., 2019; Di Stefano et al., 2016; Knaupp et al., 2017; D. Li et al., 2017; Wang et al., 2019). These studies again fail to account for the widespread heterogeneity among populations of reprogramming cells. Recently, single-cell technology has been adapted to now perform single-cell ATAC-seq (scATAC-seq), which has been applied in a few studies investigating human reprogramming systems.

Here, we have implemented scATAC-seq analysis on cells from both our low-efficiency FBS and high-efficiency A2S reprogramming systems of MEF reprogramming, again in an attempt to identify the features that distinguish successful reprogramming from an inefficient process. Many key differences were highlighted from motif enrichment analysis of accessible regions in the clusters of cells. In both our scRNA-seq and scATAC-seq analyses we saw a separation of the starting MEF population into two separate clusters. By scRNA-seq, MEFs separate based on differential expression of cell cycle markers (Mcm5, Bub1b) (Tran et al., 2019). However, scATAC-seq data separates MEFs by differential enrichment of motifs for different families of development-associated transcription factors (TFs) (MEIS, GATA, HOX).

Other studies have previously found that certain MEFs are more “elite” with a higher proclivity for successful reprogramming (Jain et al., 2023), with neural crest-derived Wnt1-expressing cells more poised for reprogramming (Shakiba et al., 2019), though these previously discovered factors do not appear to be contributing to this separation of MEFs based on scATAC-seq data. Furthermore, cells that fail to reprogram and revert back to a MEF-like state seem to come from both of these MEF

populations, suggesting that one group of MEFs is no more advantageous than the other in the transition to iPSCs. However, future research could examine subpopulations of MEFs based on these differentiating factors to see if there are any differences at all in the route that they take towards reprogramming.

One discovery that distinguishes high-efficiency from low-efficiency reprogramming is a large population of cells in the FBS system that, after an initial drastic global rewiring of the chromatin network, soon after reacquire enrichment of motifs that are also strongly represented in MEFs (e.g. the STAT motifs). Such a cluster is not observed among the A2S cells, suggesting that these chemicals prevent any cells from re-opening somatic-associated motifs and push the cells forward along the reprogramming trajectory, stably maintaining repression of MEF loci and opening of pluripotency loci as these changes happen. However, it's not clear from this data exactly what's causing the re-opening of STAT motifs, and further inquiry is required to better understand this phenomenon.

Another key difference between FBS and A2S that we observed is the enrichment of motifs associated with TFs that mediate 3D chromatin reorganization and promoter-enhancer interaction loops, including KLF4, MAZ, and PATZ1 (Fedele et al., 2017; Ma et al., 2014; Ortabozkoyun et al., 2022; Rong et al., 2013). Given this information, we decided to delve deeper into finding loci that act as enhancer hubs and whose 3D interactions are important in regulation of reprogramming. In a collaboration with Sushmita Roy's lab, this inquiry led to the development of scCISINT, which is able to use scATAC-seq data to predict interacting genes and the strength of those interactions among accessible loci.

In so doing, we were able to identify two loci with disparate accessibility patterns, that affect reprogramming along different mechanisms. One locus was exclusive to the more advanced reprogramming cells upon exogenous OSKM-withdrawal; another maintains fairly consistent accessibility throughout the duration of reprogramming. While the former predictably was found to play a role in the maintenance of iPSC colonies and independence from OSKM, the latter was found to regulate different genes at different stages of reprogramming, leading to different reprogramming outcomes when depleted at different timepoints (repression early enhanced reprogramming, while repression at a midpoint impeded it). Implementation of scCISINT also led to the discovery of a peak that has an anti-correlative accessibility pattern with the nearby Nanog gene promoter. This peak was observed to be inhibitory to reprogramming. This data serves to demonstrate the remarkable ability of scCISINT to identify regulatory peaks, and highlight the vastly different roles and accessibility dynamics of different loci at different stages of the reprogramming process. While I was only able to study these three regions in depth within the scope of this project, a great future direction for this research could involve a large-scale knockdown screen of other top-scoring loci from scCISINT. This could lead to the discovery of several peaks whose depletion can enhance or promote reprogramming. Furthermore, a systematic method of knocking down multiple loci in a combinatorial manner could help identify a more precise combination of peaks whose depletion or activation can help improve the reprogramming protocol.

We also pursued further investigation of the changes that accompany the final transition to iPSCs as they become reprogramming factor-independent. We profiled cells two days post-withdrawal. These cells experienced greater upregulation of the

pluripotency factor ESRRB, as well as strong transient regulation of TCFAP2C. Although its motif is almost exclusively enriched in these withdrawal cells, its upregulation is important for the final transition to becoming bona fide iPSC colonies. TCFAP2C is a known regulator of trophoblast cells. Moreover, our identified withdrawal-associated peak from scCISINT is predicted to interact with placental-associated miR290 microRNAs (Paikari et al., 2017; Yuan et al., 2017) and the protein-coding gene *AU018091*. As previously mentioned, it has been posited that cells can transition into iTSCs as an alternative to iPSCs. Our data, on the other hand, suggests that reprogramming cells actually pass through a trophectoderm-like state on their way to becoming iPSCs. While our scATAC-seq data is insufficient to definitively conclude this, further isolation of and investigation into this transitory population of cells could provide greater insight into the plasticity of these cells and whether they're adopting characteristics of iTSCs upon exogenous OSKM withdrawal prior to making the final transition to iPSCs.

Altogether, our single-cell analysis of the transcriptional and accessibility dynamics has helped illuminate more of the regulatory mechanisms that underlie the reprogramming process and how a high-efficiency system is better able to overcome many of the barriers associated with the transition to induced pluripotency.

Additional Potential Future Directions

While we have utilized two separate single-cell methodologies to study reprogramming dynamics, we are limited in our ability to directly examine the interplay between transcriptional and chromatin dynamics. Some methods are available that can

artificially match single-cell data across modalities (Granja et al., 2021; Stuart et al., 2019; Welch et al., 2016) obtaining this same data from the same individual cell will provide a more accurate snapshot of the relationship between these two features. Recently, multi-omics technology has become available, allowing capture of both scRNA-seq and scATAC-seq data from the same cell. While I expect there to be some correlation between the promoter and motif accessibility and expression of the corresponding gene, there could be incongruence between these two cellular features and could help further identify characteristics of reprogramming-refractory cells. Additionally, the potential combination of single-cell ATAC-seq with single-cell ChIP-seq could shed light on how TF binding influences or is influenced by chromatin accessibility dynamics, perhaps again identifying cells whose TF binding and accessibility profiles are misaligned.

The scRNA-seq and scATAC-seq analyses that we have performed here provides a snapshot of the expression and chromatin accessibility landscapes of a particular cell from any given timepoint during the reprogramming process. While we can and have used computational means to artificially generate a reprogramming trajectory in pseudotime, we don't know exactly which parent cells gave rise to the subsequent cells at subsequent timepoints during reprogramming. Therefore, another potential future study would that it would be to perform a lineage tracing experiment, in which each starting cell is tagged with an identifying sequence, to get a more accurate description of which cells from the starting population give rise to iPSCs vs those that veer away from the trajectory.

My research in this thesis has broadened our understanding of the regulatory mechanisms underlying the process of somatic cell reprogramming. The information presented here has helped to identify characteristics of successful reprogramming as well as ways in which the process can be improved, and has also set up other potential future directions for ours and other labs to pursue in the investigation of the reprogramming landscape.

References

Apostolou, E., & Hochedlinger, K. (2013). Chromatin dynamics during cellular reprogramming. *Nature*, *502*(7472), Article 7472. https://doi.org/10.1038/nature12749

Apostolou, E., & Stadtfeld, M. (2018). Cellular trajectories and molecular mechanisms of iPSC reprogramming. *Current Opinion in Genetics & Development*, *52*, 77–85. https://doi.org/10.1016/j.gde.2018.06.002

Banito, A., Rashid, S. T., Acosta, J. C., Li, S., Pereira, C. F., Geti, I., Pinho, S., Silva, J. C., Azuara, V., Walsh, M., Vallier, L., & Gil, J. (2009). Senescence impairs successful reprogramming to pluripotent stem cells. *Genes & Development*, *23*(18), 2134–2139. https://doi.org/10.1101/gad.1811609

Bar-Nur, O., Brumbaugh, J., Verheul, C., Apostolou, E., Pruteanu-Malinici, I., Walsh, R. M., Ramaswamy, S., & Hochedlinger, K. (2014). Small molecules facilitate rapid and synchronous iPSC generation. *Nature Methods*, *11*(11), Article 11. https://doi.org/10.1038/nmeth.3142

Benchetrit, H., Herman, S., van Wietmarschen, N., Wu, T., Makedonski, K., Maoz, N., Yom Tov, N., Stave, D., Lasry, R., Zayat, V., Xiao, A., Lansdorp, P. M., Sebban, S., & Buganim, Y. (2015). Extensive Nuclear Reprogramming Underlies Lineage Conversion into Functional Trophoblast Stem-like Cells. *Cell Stem Cell*, *17*(5), 543–556. https://doi.org/10.1016/j.stem.2015.08.006

Benchetrit, H., Jaber, M., Zayat, V., Sebban, S., Pushett, A., Makedonski, K., Zakheim, Z., Radwan, A., Maoz, N., Lasry, R., Renous, N., Inbar, M., Ram, O., Kaplan, T., & Buganim, Y. (2019). Direct Induction of the Three Pre-implantation Blastocyst Cell Types from Fibroblasts. *Cell Stem Cell*, *24*(6), 983-994.e7. https://doi.org/10.1016/j.stem.2019.03.018

Brambrink, T., Foreman, R., Welstead, G. G., Lengner, C. J., Wernig, M., Suh, H., & Jaenisch, R. (2008). Sequential Expression of Pluripotency Markers during Direct Reprogramming of Mouse Somatic Cells. *Cell Stem Cell*, *2*(2), 151–159. https://doi.org/10.1016/j.stem.2008.01.004

Buckberry, S., Liu, X., Poppe, D., Tan, J. P., Sun, G., Chen, J., Nguyen, T. V., de Mendoza, A., Pflueger, J., Frazer, T., Vargas-Landín, D. B., Paynter, J. M., Smits, N., Liu, N., Ouyang, J. F., Rossello, F. J., Chy, H. S., Rackham, O. J. L., Laslett, A. L., … Lister, R. (2023). Transient naive reprogramming corrects hiPS cells functionally and epigenetically. *Nature*, *620*(7975), Article 7975. https://doi.org/10.1038/s41586-023-06424-7

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of

open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), Article 12. https://doi.org/10.1038/nmeth.2688

Buganim, Y., Faddah, D. A., & Jaenisch, R. (2013). Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics*, *14*(6), Article 6. https://doi.org/10.1038/nrg3473

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, *361*(6409), 1380–1385. https://doi.org/10.1126/science.aau0730

Chen, J., Liu, H., Liu, J., Qi, J., Wei, B., Yang, J., Liang, H., Chen, Y., Chen, J., Wu, Y., Guo, L., Zhu, J., Zhao, X., Peng, T., Zhang, Y., Chen, S., Li, X., Li, D., Wang, T., & Pei, D. (2013). H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature Genetics*, *45*(1), Article 1. https://doi.org/10.1038/ng.2491

Chen, J., Liu, J., Chen, Y., Yang, J., Chen, J., Liu, H., Zhao, X., Mo, K., Song, H., Guo, L., Chu, S., Wang, D., Ding, K., & Pei, D. (2011). Rational optimization of reprogramming culture conditions for the generation of induced pluripotent stem cells with ultra-high efficiency and fast kinetics. *Cell Research*, *21*(6), Article 6. https://doi.org/10.1038/cr.2011.51

Chen, K., Long, Q., Xing, G., Wang, T., Wu, Y., Li, L., Qi, J., Zhou, Y., Ma, B., Schöler, H. R., Nie, J., Pei, D., & Liu, X. (2020). Heterochromatin loosening by the Oct4 linker region facilitates Klf4 binding and iPSC reprogramming. *The EMBO Journal*, *39*(1), e99165. https://doi.org/10.15252/embj.201899165

Chen, X., Lu, Y., Wang, L., Ma, X., Pu, J., Lin, L., Deng, Q., Li, Y., Wang, W., Jin, Y., Hu, Z., Zhou, Z., Chen, G., Jiang, L., Wang, H., Zhao, X., He, X., Fu, J., Russ, H. A., … Zhu, S. (2023). A fast chemical reprogramming system promotes cell identity transition through a diapause-like state. *Nature Cell Biology*, *25*(8), Article 8. https://doi.org/10.1038/s41556-023-01193-x

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., & Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell*, *168*(3), 442-459.e20. https://doi.org/10.1016/j.cell.2016.12.016

Di Giammartino, D. C., Kloetgen, A., Polyzos, A., Liu, Y., Kim, D., Murphy, D., Abuhashem, A., Cavaliere, P., Aronson, B., Shah, V., Dephoure, N., Stadtfeld, M., Tsirigos, A., & Apostolou, E. (2019). KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks.

Nature Cell Biology, *21*(10), 1179–1190. https://doi.org/10.1038/s41556-019-0390-6

Di Stefano, B., Collombet, S., Jakobsen, J. S., Wierer, M., Sardina, J. L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., Limone, F., Mann, M., Porse, B., Thieffry, D., & Graf, T. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. *Nature Cell Biology*, *18*(4), Article 4. https://doi.org/10.1038/ncb3326

Esteban, M. A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S., Chen, K., Li, Y., Liu, X., Xu, J., Zhang, S., Li, F., He, W., Labuda, K., Song, Y., … Pei, D. (2010). Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells. *Cell Stem Cell*, *6*(1), 71–79. https://doi.org/10.1016/j.stem.2009.12.001

Fedele, M., Crescenzi, E., & Cerchia, L. (2017). The POZ/BTB and AT-Hook Containing Zinc Finger 1 (PATZ1) Transcription Regulator: Physiological Functions and Disease Involvement. *International Journal of Molecular Sciences*, *18*(12), Article 12. https://doi.org/10.3390/ijms18122524

Gaspar-Maia, A., Alajem, A., Meshorer, E., & Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology*, *12*(1), Article 1. https://doi.org/10.1038/nrm3036

Golipour, A., David, L., Liu, Y., Jayakumaran, G., Hirsch, C. L., Trcka, D., & Wrana, J. L. (2012). A Late Transition in Somatic Cell Reprogramming Requires Regulators Distinct from the Pluripotency Network. *Cell Stem Cell*, *11*(6), 769–782. https://doi.org/10.1016/j.stem.2012.11.008

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, *53*(3), Article 3. https://doi.org/10.1038/s41588-021-00790-6

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A. E., & Melton, D. A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nature Biotechnology*, *26*(7), Article 7. https://doi.org/10.1038/nbt1418

Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W., & Melton, D. A. (2008). Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nature Biotechnology*, *26*(11), Article 11. https://doi.org/10.1038/nbt.1502

Hussein, S. M. I., Puri, M. C., Tonge, P. D., Benevento, M., Corso, A. J., Clancy, J. L., Mosbergen, R., Li, M., Lee, D.-S., Cloonan, N., Wood, D. L. A., Munoz, J.,

Middleton, R., Korn, O., Patel, H. R., White, C. A., Shin, J.-Y., Gauthier, M. E., Cao, K.-A. L., … Nagy, A. (2014). Genome-wide characterization of the routes to pluripotency. *Nature*, *516*(7530), Article 7530. https://doi.org/10.1038/nature14046

Jain, N., Goyal, Y., Dunagin, M. C., Cote, C. J., Mellis, I. A., Emert, B., Jiang, C. L., Dardani, I. P., Reffsin, S., & Raj, A. (2023). *Retrospective identification of intrinsic factors that mark pluripotency potential in rare somatic cells* (p. 2023.02.10.527870). bioRxiv. https://doi.org/10.1101/2023.02.10.527870

Knaupp, A. S., Buckberry, S., Pflueger, J., Lim, S. M., Ford, E., Larcombe, M. R., Rossello, F. J., Mendoza, A. de, Alaei, S., Firas, J., Holmes, M. L., Nair, S. S., Clark, S. J., Nefzger, C. M., Lister, R., & Polo, J. M. (2017). Transient and Permanent Reconfiguration of Chromatin and Transcription Factor Occupancy Drive Reprogramming. *Cell Stem Cell*, *21*(6), 834-845.e6. https://doi.org/10.1016/j.stem.2017.11.007

Li, D., Liu, J., Yang, X., Zhou, C., Guo, J., Wu, C., Qin, Y., Guo, L., He, J., Yu, S., Liu, H., Wang, X., Wu, F., Kuang, J., Hutchins, A. P., Chen, J., & Pei, D. (2017). Chromatin Accessibility Dynamics during iPSC Reprogramming. *Cell Stem Cell*, *21*(6), 819-833.e6. https://doi.org/10.1016/j.stem.2017.10.012

Li, H., Collado, M., Villasante, A., Strati, K., Ortega, S., Cañamero, M., Blasco, M. A., & Serrano, M. (2009). The Ink4/Arf locus is a barrier for iPS cell reprogramming. *Nature*, *460*(7259), Article 7259. https://doi.org/10.1038/nature08290

Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., Qin, B., Xu, J., Li, W., Yang, J., Gan, Y., Qin, D., Feng, S., Song, H., Yang, D., … Pei, D. (2010). A Mesenchymal-to-Epithelial Transition Initiates and Is Required for the Nuclear Reprogramming of Mouse Fibroblasts. *Cell Stem Cell*, *7*(1), 51–63. https://doi.org/10.1016/j.stem.2010.04.014

Ma, H., Ow, J. R., Tan, B. C. P., Goh, Z., Feng, B., Loh, Y. H., Fedele, M., Li, H., & Wu, Q. (2014). The dosage of Patz1 modulates reprogramming process. *Scientific Reports*, *4*(1), Article 1. https://doi.org/10.1038/srep07519

Mikkelsen, T. S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B. E., Jaenisch, R., Lander, E. S., & Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature*, *454*(7200), Article 7200. https://doi.org/10.1038/nature07056

Naama, M., Rahamim, M., Zayat, V., Sebban, S., Radwan, A., Orzech, D., Lasry, R., Ifrah, A., Jaber, M., Sabag, O., Yassen, H., Khatib, A., Epsztejn-Litman, S., Novoselsky-Persky, M., Makedonski, K., Deri, N., Goldman-Wohl, D., Cedar, H., Yagel, S., … Buganim, Y. (2023). Pluripotency-independent induction of human trophoblast stem cells from fibroblasts. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-39104-1

Nefzger, C. M., Rossello, F. J., Chen, J., Liu, X., Knaupp, A. S., Firas, J., Paynter, J. M., Pflueger, J., Buckberry, S., Lim, S. M., Williams, B., Alaei, S., Faye-Chauhan, K., Petretto, E., Nilsson, S. K., Lister, R., Ramialison, M., Powell, D. R., Rackham, O. J. L., & Polo, J. M. (2017). Cell Type of Origin Dictates the Route to Pluripotency. *Cell Reports*, *21*(10), 2649–2660. https://doi.org/10.1016/j.celrep.2017.11.029

Onder, T. T., Kara, N., Cherry, A., Sinha, A. U., Zhu, N., Bernt, K. M., Cahan, P., Mancarci, B. O., Unternaehrer, J., Gupta, P. B., Lander, E. S., Armstrong, S. A., & Daley, G. Q. (2012). Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, *483*(7391), Article 7391. https://doi.org/10.1038/nature10953

Ortabozkoyun, H., Huang, P.-Y., Cho, H., Narendra, V., LeRoy, G., Gonzalez-Buendia, E., Skok, J. A., Tsirigos, A., Mazzoni, E. O., & Reinberg, D. (2022). CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. *Nature Genetics*, *54*(2), Article 2. https://doi.org/10.1038/s41588-021-01008-5

Paikari, A., D. Belair, C., Saw, D., & Blelloch, R. (2017). The eutheria-specific miR-290 cluster modulates placental growth and maternal-fetal transport. *Development*, *144*(20), 3731–3743. https://doi.org/10.1242/dev.151654

Papp, B., & Plath, K. (2013). Epigenetics of Reprogramming to Induced Pluripotency. *Cell*, *152*(6), 1324–1343. https://doi.org/10.1016/j.cell.2013.02.043

Parenti, A., Halbisen, M. A., Wang, K., Latham, K., & Ralston, A. (2016). OSKM Induce Extraembryonic Endoderm Stem Cells in Parallel to Induced Pluripotent Stem Cells. *Stem Cell Reports*, *6*(4), 447–455. https://doi.org/10.1016/j.stemcr.2016.02.003

Polo, J. M., Anderssen, E., Walsh, R. M., Schwarz, B. A., Nefzger, C. M., Lim, S. M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., Bar-Nur, O., Cheloufi, S., Stadtfeld, M., Figueroa, M. E., Robinton, D., Natesan, S., Melnick, A., Zhu, J., Ramaswamy, S., & Hochedlinger, K. (2012). A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell*, *151*(7), 1617–1632. https://doi.org/10.1016/j.cell.2012.11.039

Rong, O., MaHui, JeanAngela, GohZiyi, Hwa, L., Mei, C., SoongRichie, FuXin-Yuan, YangHenry, & WuQiang. (2013). Patz1 Regulates Embryonic Stem Cell Identity. *Stem Cells and Development*. https://doi.org/10.1089/scd.2013.0430

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H., Beyer, T. A., Datti, A., Woltjen, K., Nagy, A., & Wrana, J. L. (2010). Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell

Reprogramming. *Cell Stem Cell*, *7*(1), 64–77. https://doi.org/10.1016/j.stem.2010.04.015

Schwarz, B. A., Cetinbas, M., Clement, K., Walsh, R. M., Cheloufi, S., Gu, H., Langkabel, J., Kamiya, A., Schorle, H., Meissner, A., Sadreyev, R. I., & Hochedlinger, K. (2018). Prospective Isolation of Poised iPSC Intermediates Reveals Principles of Cellular Reprogramming. *Cell Stem Cell*, *23*(2), 289-305.e5. https://doi.org/10.1016/j.stem.2018.06.013

Shakiba, N., Fahmy, A., Jayakumaran, G., McGibbon, S., David, L., Trcka, D., Elbaz, J., Puri, M. C., Nagy, A., van der Kooy, D., Goyal, S., Wrana, J. L., & Zandstra, P. W. (2019). Cell competition during reprogramming gives rise to dominant clones. *Science*, *364*(6438), eaan0925. https://doi.org/10.1126/science.aan0925

Shi, Y., Desponts, C., Do, J. T., Hahm, H. S., Schöler, H. R., & Ding, S. (2008). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Oct4 and Klf4 with Small-Molecule Compounds. *Cell Stem Cell*, *3*(5), 568–574. https://doi.org/10.1016/j.stem.2008.10.004

Sim, Y.-J., Kim, M.-S., Nayfeh, A., Yun, Y.-J., Kim, S.-J., Park, K.-T., Kim, C.-H., & Kim, K.-S. (2017). 2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms. *Stem Cell Reports*, *8*(5), 1312–1328. https://doi.org/10.1016/j.stemcr.2017.04.001

Soufi, A., Donahue, G., & Zaret, K. S. (2012). Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell*, *151*(5), 994–1004. https://doi.org/10.1016/j.cell.2012.09.045

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M., Carey, M., Garcia, B. A., & Plath, K. (2013). Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. *Nature Cell Biology*, *15*(7), 872–882. https://doi.org/10.1038/ncb2768

Sridharan, R., Tchieu, J., Mason, M. J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., & Plath, K. (2009). Role of the Murine Reprogramming Factors in the Induction of Pluripotency. *Cell*, *136*(2), 364–377. https://doi.org/10.1016/j.cell.2009.01.001

Stadtfeld, M., Maherali, N., Breault, D. T., & Hochedlinger, K. (2008). Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. *Cell Stem Cell*, *2*(3), 230–240. https://doi.org/10.1016/j.stem.2008.02.001

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676. https://doi.org/10.1016/j.cell.2006.07.024

Tran, K. A., Jackson, S. A., Olufs, Z. P. G., Zaidan, N. Z., Leng, N., Kendziorski, C., Roy, S., & Sridharan, R. (2015). Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. *Nature Communications*, *6*(1), Article 1. https://doi.org/10.1038/ncomms7188

Tran, K. A., Pietrzak, S. J., Zaidan, N. Z., Siahpirani, A. F., McCalla, S. G., Zhou, A. S., Iyer, G., Roy, S., & Sridharan, R. (2019). Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. *Cell Reports*, *27*(6), 1726-1741.e5. https://doi.org/10.1016/j.celrep.2019.04.056

Vidal, S. E., Amlani, B., Chen, T., Tsirigos, A., & Stadtfeld, M. (2014). Combinatorial Modulation of Signaling Pathways Reveals Cell-Type-Specific Requirements for Highly Efficient and Synchronous iPSC Reprogramming. *Stem Cell Reports*, *3*(4), 574–584. https://doi.org/10.1016/j.stemcr.2014.08.003

Wang, B., Wu, L., Li, D., Liu, Y., Guo, J., Li, C., Yao, Y., Wang, Y., Zhao, G., Wang, X., Fu, M., Liu, H., Cao, S., Wu, C., Yu, S., Zhou, C., Qin, Y., Kuang, J., Ming, J., … Pei, D. (2019). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Jdp2-Jhdm1b-Mkk6-Glis1-Nanog-Essrb-Sall4. *Cell Reports*, *27*(12), 3473-3485.e5. https://doi.org/10.1016/j.celrep.2019.05.068

Welch, J. D., Hartemink, A. J., & Prins, J. F. (2016). SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, *17*(1), 106. https://doi.org/10.1186/s13059-016-0975-3

Ying, Q.-L., & Smith, A. (2017). The Art of Capturing Pluripotency: Creating the Right Culture. *Stem Cell Reports*, *8*(6), 1457–1464. https://doi.org/10.1016/j.stemcr.2017.05.020

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., & Smith, A. (2008). The ground state of embryonic stem cell self-renewal. *Nature*, *453*(7194), Article 7194. https://doi.org/10.1038/nature06968

Yuan, K., Ai, W.-B., Wan, L.-Y., Tan, X., & Wu, J.-F. (2017). The miR-290-295 cluster as multi-faceted players in mouse embryonic stem cells. *Cell & Bioscience*, *7*(1), 38. https://doi.org/10.1186/s13578-017-0166-2

Appendix 1

Beta cell dedifferentiation induced by IRE1α deletion prevents type 1 diabetes

The work presented in this appendix is published in Cell Metabolism:

Lee, H., Lee, Y-S., Harenda, Q., Pietrzak, S.J., Oktay, H.Z., Schreiber, S., Liao, Y., Sonthalia, S., Ciecko, A.E., Chen, Y-G., Keles, S., Sridharan, R., and Engin, F. (2020). Beta dell dedifferentiation induced by IRE1α deletion prevents type 1 diabetes. *Cell Metab*, 31(4):822-836.

Contributions: H.L. designed and performed experiments, analyzed data, prepared the figures, and revised the manuscript. Y.-S.L., H.Z.O., and Q.H. contributed to in vivo experiments and analyzed data. S.P. analyzed scRNA-seq data. S. Schreiber, S. Sonthalia, Y.L., and Y.-S.L. performed immunofluorescence assays. A.E.C. performed the adoptive transfer experiment and Y.-G.C. supervised the adoptive transfer experiment, analyzed the data, and edited the manuscript. S.K. analyzed the bulk RNA-seq data and edited the manuscript. R.S. analyzed the scRNA-seq data and edited the manuscript. F.E. conceived, supervised, and supported the project; designed experiments; interpreted results; and wrote and revised the manuscript.

Abstract

Immune-mediated destruction of insulin-producing β cells causes type 1 diabetes (T1D). However, how β cells participate in their own destruction during the disease process is poorly understood. Here, we report that modulating the unfolded protein response (UPR) in β cells of non-obese diabetic (NOD) mice by deleting the UPR sensor IRE1α prior to insulitis induced a transient dedifferentiation of β cells, resulting in substantially reduced islet immune cell infiltration and β cell apoptosis. Single-cell and whole-islet transcriptomics analyses of immature β cells revealed remarkably diminished expression of β cell autoantigens and MHC class I components, and upregulation of immune inhibitory markers. IRE1α-deficient mice exhibited significantly fewer cytotoxic $CD8^+$ T cells in their pancreata, and adoptive transfer of their total T cells did not induce diabetes in $Rag1^{-/-}$ mice. Our results indicate that inducing β cell dedifferentiation, prior to insulitis, allows these cells to escape immune-mediated destruction and may be used as a novel preventive strategy for T1D in high-risk individuals.

Introduction

Type 1 diabetes (T1D) is an autoimmune disease in which insulin producing pancreatic islet β cells are targeted and destroyed by autoreactive immune cells (Atkinson, 2012; Bluestone et al., 2010; van Belle et al., 2011). Although genetic predisposition is strongly associated with T1D progression, environmental factors can also trigger T1D onset and affect its progression. Environmental factors include viral infections, toxins, reactive oxygen species (ROS), and chronic inflammation, which are well-established triggers of endoplasmic reticulum (ER) stress. ER stress initiates the unfolded protein response (UPR), which operates through inositol-requiring protein-1 (IRE1), protein kinase RNA-like ER kinase (PERK), and activating transcription factor-6 (ATF6), all of which are localized in the ER membrane and respond to stress by relaying signals from the ER to the cytoplasm and nucleus. While the UPR initially attempts to mitigate ER stress, if the stress is prolonged or severe, it switches from being a pro-adaptive to a pro-apoptotic response (Bernales et al., 2006; Walter and Ron, 2011).

Mammals have two IRE1 paralogs, IRE1α and IRE1β. IRE1α is ubiquitously expressed, whereas IRE1β is specifically expressed in digestive tissues. Dual-functioning IRE1α removes an intronic region from the transcription factor X-box binding protein 1 (XBP1) with its endoribonuclease activity, leading to its transcriptional activation (Calfon et al., 2002; Yoshida et al., 2001). Spliced XBP1 (sXBP1) then upregulates the expression of chaperones and ER-associated degradation components. IRE1α also promotes mRNA and miRNA degradation through regulated IRE1α-dependent decay (Hollien and Weissman, 2006). IRE1α, with its kinase activity, mediates the phosphorylation of the c-Jun N-terminal protein kinase (JNK) and induces

inflammatory signals and apoptosis (Urano et al., 2000). The IRE1α/XBP1 pathway is the most conserved branch of the UPR, and highly secretory pancreatic β cells have constitutively active IRE1α/ XBP1 under physiological conditions. Although IRE1α has a role in promoting cell survival under acute and mild stress conditions, it can promote cell death in the presence of unresolvable ER stress (Chen and Brandizzi, 2013). IRE1α regulates cell death by downregulating mRNAs and miRNAs involved in β cell homeostasis and survival (Han et al., 2009; Hollien and Weissman, 2006; Lipson et al., 2008; Upton et al., 2012). Hyperactivated IRE1α increases thioredoxin-interacting protein (TXNIP) mRNA stability, and in turn, elevated TXNIP activates the NLRP3 inflammasome and causes β cell death (Lerner et al., 2012; Oslowski et al., 2012). Although these studies clearly indicate that modulation of IRE1α under chronic stress versus non-stress conditions may lead to differential survival/apoptotic outcomes, the function of IRE1α in β cells in the context of autoimmune diabetes remains unclear.

Due to the autoimmune nature of T1D, there has been a longstanding interest in understanding the role of dysregulation of the immune system in the pathogenesis of T1D. However, emerging data indicate that β cells themselves can play a much more active role in the initiation and progression of autoimmunity than previously appreciated (Engin, 2016; Engin et al., 2013; Maganti et al., 2014; Soleimanpour and Stoffers, 2013; Thompson et al., 2019). The indication that the β cell UPR may play a role in pathogenesis of autoimmune diabetes is supported by the detection of dysregulated ER stress markers in inflamed islets of both mice (Tersey et al., 2012) and patients with autoimmune diabetes (Marhfour et al., 2012). We previously showed that the adaptive functions of the UPR were greatly defective in β cells of two different T1D mouse

models and human patients during the progression of T1D (Engin et al., 2013). Diabetes incidence in these mouse models was dramatically reduced upon mitigation of β cell ER stress with a chemical chaperone. Although this study provided the first direct link between a defective β cell UPR and T1D, the molecular mechanisms by which the UPR regulates pancreatic β cell death/survival in T1D still remain largely unknown.

Here, we investigated the specific role of the β cell UPR in initiation and progression of T1D and deleted IRE1α in β cells of an established T1D pre-clinical model, non-obese diabetic (NOD) mice, prior to the initiation of islet infiltration by immune cells (insulitis). Deletion of IRE1α in β cells of NOD mice led to transient mild hyperglycemia. However, unexpectedly, mice recovered from hyperglycemia within a couple of weeks and were protected from autoimmune destruction for up to a year. Single-cell transcriptional profiling in dissociated islets, bulk RNA sequencing (RNA-seq) of intact islets, as well as histological analyses demonstrated loss of mature β cell identity and remarkably increased expression of endocrine progenitor and fetal-like β cell markers, suggesting that upon losing IRE1α expression, β cells of NOD mice undergo a transient dedifferentiation. Furthermore, we identified significantly reduced MHC class I expression and defective antigen processing, notably reduced expression of β cell autoantigens, upregulation of immune inhibitory markers, as well as substantially diminished $CD8^+$ T cell population in pancreas. Taken together, we show that loss of IRE1α in β cells prior to initiation of islet immune infiltration protects against diabetes in a pre-clinical model of T1D by inducing transient dedifferentiation of β cells, which allows escape from immune-mediated destruction.

Results

IRE1α$^{β–/–}$ NOD Mice Are Protected from T1D

To examine the β cell-specific functions of IRE1α in T1D, we backcrossed IRE1α flox/flox (IRE1α$^{fl/fl}$) (Iwawaki et al., 2009) and Ins2Cre$^{ERT/+}$ mice (Dor et al., 2004) to NOD mice for more than 10–20 generations and confirmed their genetic purity on NOD background by genome scan services. To generate mice with β cell-specific IRE1α deletion, we mated IRE1α$^{fl/fl}$ and IRE1α$^{fl/fl}$; Ins2Cre$^{ERT/+}$ mice and administered tamoxifen to lactating mothers, beginning the day after delivery (Figure 1A). We reasoned that tamoxifen-mediated deletion of IRE1α in pups would allow us to elucidate its function before the onset of islet insulitis, as insulitis typically starts after weaning age in female NOD mice. The deletion of IRE1α in β cells (referred to as IRE1α$^{β-/-}$) was confirmed via immunofluorescence (IF) by staining pancreatic sections with anti-sXBP1, a direct target of IRE1α, and anti-insulin antibodies (Figure 1B), and by performing a qPCR on islets for *sXBP1* (Figure 1C). Weekly blood glucose measurements were recorded, starting from 3 weeks of age through 50 weeks (Figure 1D). Mice with a blood glucose level R250 mg/dL for two consecutive weeks were accepted as diabetic.

IRE1α deficiency in β cells of NOD female mice resulted in hyperglycemia starting after weaning age (Figure 1D), in agreement with recent reports showing that IRE1α deletion in β cells of non-stressed wild-type mice also induces hyperglycemia (Hassler et al., 2015; Tsuchiya et al., 2018). However, in striking contrast to previous reports, where deletion of IRE1α in either prenatal or adult β cells led to a diabetic phenotype (Hassler et al., 2015; Tsuchiya et al., 2018), the hyperglycemia in IRE1α$^{β-/-}$ NOD mice was temporary, with mice recovering from diabetes (94.7% of the IRE1α$^{β-/-}$

mice exhibited normoglycemia following an initial hyperglycemia) starting at 6–7 weeks of age. Surprisingly, IRE1$\alpha^{\beta-/-}$ NOD mice not only recovered from hyperglycemia but were also protected from development of T1D through 50 weeks of age (Figures 1E and 1F).

To rule out any potential artifacts arising from the expression of Cre recombinase, we generated another cohort by breeding NOD and NOD Ins2Cre$^{ERT/+}$ mice, administered tamoxifen as described above, and measured the blood glucose levels of these animals weekly up to 22 weeks of age. Unlike IRE1$\alpha^{\beta-/-}$ mice, hemizygous Ins2Cre$^{ERT/+}$ mice did not exhibit early hyperglycemia, and the diabetes incidence of Ins2Cre$^{ERT/+}$ mice was similar to that of the littermate control NOD mice as well as IRE1$\alpha^{fl/fl}$ mice (Figures S1A and S1B). In addition, the expression of Cre did not significantly differ between Ins2Cre$^{ERT/+}$ and IRE1$\alpha^{\beta-/-}$ mice (Figure S1C). Histological analyses confirmed the presence of aggressive insulitis in the islets of diabetic Ins2Cre$^{ERT/+}$ mice (Figure S1D). Finally, we did not detect any significant differences in glucose tolerance between pre-diabetic Ins2Cre$^{ERT/+}$ and NOD mice, ruling out impaired glucose homeostasis in Ins2Cre$^{ERT/+}$ mice prior to the development of diabetes (Figures S1E and S1F). Taken together, these data demonstrate that Cre transgene expression in our mouse line does not appear to alter the natural progression of diabetes in NOD mice and that the phenotype seen in IRE1$\alpha^{\beta-/-}$ mice is most likely independent of a Cre effect.

The unexpected recovery from hyperglycemia in NOD IRE1$\alpha^{\beta-/-}$ mice, as opposed to the previously demonstrated diabetic phenotype seen in non-stressed, non-autoimmune ‘‘normal wild-type’’ mice upon IRE1α deletion in β cells, led us to ask

whether autoimmunity or the intrinsic ER stress and fragility of β cells seen in NOD mice (Dooley et al., 2016) was the underlying cause of this differential phenotype. Thus, we deleted IRE1α in β cells of NOD *Rag1*$^{-/-}$ mice, which lack mature T cells and β cells. Interestingly, NOD IRE1α$^{β-/-}$; Rag1$^{-/-}$ mice phenocopied NOD IRE1α$^{β-/-}$ mice, suggesting that the transient hyperglycemia phenotype was independent of T and β cell-mediated autoimmunity but involved β cell stress (Figure S2A).

Thus, these intriguing results indicate that while inactivation of IRE1α in adult β cells under physiological conditions can lead to diabetes, under stressed conditions as seen in T1D, the loss of IRE1α at an early stage of disease (i.e., prior to initiation of insulitis) can have beneficial effects on β cell survival and function leading to protection from T1D.

Improved β Cell Function and Survival in IRE1α$^{β–/–}$ NOD Mice upon Recovery from Hyperglycemia

To investigate the cellular basis of the phenotype, we performed hematoxylin and eosin (H&E) staining on pancreatic sections obtained from 5- and 24-week-old IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (Figure 2A). Although there was no immune infiltration in the islets of IRE1α$^{fl/fl}$ mice at 5 weeks of age, we detected considerable amount of insulitis in the islets of 24-week-old IRE1α$^{fl/fl}$ mice as expected in the NOD model. In contrast, IRE1α$^{β-/-}$ mice showed improved islet morphology and substantially reduced immune infiltrates at 24 weeks of age. Additionally, immunofluorescence staining showed that insulin intensity in β cells of the 24-weekold IRE1α$^{β-/-}$ mice was comparable with the insulin intensity in β cells of normoglycemic control mice, despite initially being reduced at 5

weeks of age. (Figure 2B). CD3 (a marker of T cell lineage) staining of pancreatic sections confirmed significantly less immune infiltration in the islets of IRE1$\alpha^{\beta-/-}$ mice at 24 weeks of age (Figure 2C). To quantify the extent of islet immune cell infiltration, we generated step sections from pancreata of these mice and performed insulitis scoring. Quantification revealed a remarkable increase in the percentage of islets without any insulitis in pancreatic sections of IRE1$\alpha^{\beta-/-}$ mice in comparison with that in IRE1$\alpha^{fl/fl}$ mice (Figure 2D). Furthermore, the number of islets with aggressive insulitis was significantly decreased in IRE1$\alpha^{\beta-/-}$ mice compared with that in controls (Figure 2D). These data suggest that, in addition to the intrinsic effects of IRE1α in β cells, IRE1α could also affect the migration of inflammatory cells, as well as their infiltration, activation, and/ or cytotoxic function. We then assessed the function of β cells by measuring pancreatic insulin and proinsulin content and determining proinsulin/insulin ratio by ELISA at 7 weeks of age (Figures 2E–2G). Consistent with our histological data, both insulin and proinsulin content of the pancreata were substantially reduced in IRE1$\alpha^{\beta-/-}$ mice at this time point (Figures 2E and 2F). Proinsulin:insulin ratio was greatly increased in 7-week-old IRE1$\alpha^{\beta-/-}$ mice, suggesting a defect in processing of proinsulin (Figure 2G). At 24 weeks of age, pancreatic insulin content was still significantly less in IRE1$\alpha^{\beta-/-}$ mice in comparison with that in non-diabetic IRE1$\alpha^{fl/fl}$ mice albeit to a much lesser degree (Figure 2H), while proinsulin content (Figure 2I) and proinsulin:insulin ratio (Figure 2J) were fully restored in IRE1$\alpha^{\beta-/-}$ pancreata. These data suggest that ER secretory function and processing were significantly improved in IRE1$\alpha^{\beta-/-}$ mice at this age. Consistent with a restored ER function and secretory capacity, serum insulin levels of IRE1$\alpha^{\beta-/-}$ mice were comparable with that of control non-diabetic IRE1$\alpha^{fl/fl}$ mice (Figure

2K). Furthermore, an intraperitoneal glucose tolerance test revealed no differences in glucose tolerance between IRE1$\alpha^{\beta-/-}$ mice and non-diabetic control mice at 32 weeks of age (Figures S2B and S2C), indicating that the β cell function was substantially improved in IRE1$\alpha^{\beta-/-}$ mice.

To examine if there was an apparent difference in β cell death, we performed TUNEL assays on pancreatic sections at various time points. Although we did not detect any significant apoptosis in β cells of 3- and 5-week-old mice, there was a marked reduction in β cell apoptosis in IRE1$\alpha^{\beta-/-}$ mice at 24 weeks of age (Figures 2L and 2M). We then examined the proliferation of β cells during the hyperglycemic phase by co-staining the pancreatic sections with antibodies against insulin and proliferation marker Ki67. IRE1$\alpha^{\beta-/-}$ mice exhibited significantly less Ki67$^+$ cells in their β cells compared with those in the β cells of the control mice at 3 weeks of age, but at 5 weeks of age, there was no significant difference in Ki67$^+$ β cells between the control and IRE1$\alpha^{\beta-/-}$ mice (Figures 2N and 2O). Taken together, these data suggest that β cells of IRE1$\alpha^{\beta-/-}$ mice undergo a transient loss of both expression and content of insulin and proinsulin prior to insulitis, and that by 24 weeks of age, insulin expression and serum insulin levels are restored while insulitis and apoptosis are significantly reduced, leading to protection from T1D in these mice.

Islet Morphology and Architecture Are Altered in IRE1$\alpha^{\beta-/-}$ Mice during the Hyperglycemic Phase

To assess islet cellular composition and architecture, we performed an immunofluorescence assay on the pancreata from IRE1$\alpha^{\beta-/-}$ mice and their littermate

controls using antibodies against markers of α, β, and δ cells. At 3 weeks of age, we did not observe any obvious alterations in islet composition and architecture of IRE1α$^{β-/-}$ mice (Figure 3A). However, at 5 weeks of age, we detected a substantially increased number of glucagon-positive cells (roughly 4-fold increase in α cell area), accompanied by markedly diminished number of β cells (30% reduced β cell area) (Figures 3B and 3G) in the islets of IRE1α$^{β-/-}$ mice. Moreover, in contrast to control mice, which exhibited normal islet morphology with glucagon-positive cells residing on the islet periphery, islets of IRE1α$^{β-/-}$ mice had significantly increased proportion of a cells intermingled in the core of the islets in the hyperglycemic (5 week) phase. The area of the δ cells was increased by approximately 3-fold in the islets of IRE1α$^{β-/-}$ mice in comparison with that in their littermates (Figures 3C and 3G). This altered islet cell composition, function, and morphology were still apparent at 12 weeks of age (Figure 3D). At 24 weeks, we identified an almost 2-fold increase in α cell area, whereas β cell area had notably improved (10% reduction versus 30%) compared with that at 5 weeks of age, though it was still significantly reduced compared with that in IRE1α$^{fl/fl}$ mice (Figures 3E and 3H). The δ cell area was increased by 3-fold in the islets of IRE1α$^{β-/-}$ mice in comparison with that of their littermates (Figures 3F and 3H). Furthermore, in addition to altered α cell:β cell ratio, we identified a significantly increased number of somatostatin-positive cells in the islets of IRE1α$^{β-/-}$ mice at 5 and 24 weeks of age (Figures 3C and 3F). Although somatostatin-positive δ cells were distributed in the islet periphery in the islets of control mice, we found a relatively abundant fraction of δ cells intermingled with β cells in the islets (Figure 3C). In addition to these changes in islet morphology and hormone

expression, we detected a number of bihormonal (insulin/ glucagon co-expressing) cells in the islets of IRE1$\alpha^{\beta-/-}$ mice (Figure 3I). We also occasionally observed single islets or small islet clusters consisting of less than five to ten insulin-positive cells in IRE1$\alpha^{\beta-/-}$ pancreata (Figure 3J), suggesting that these cells might also contribute to glucose homeostasis. Finally, we examined the islet size distribution on pancreatic sections from animals at 4 and 12 weeks of age and detected a significantly smaller average islet size (<5,000 mm2) in the pancreata of 4 weeks of age IRE1$\alpha^{\beta-/-}$ mice (Figure 3K). We did not observe any differences in islet size distribution between IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice at 12 weeks of age (Figure 3L).

These data indicate that during the hyperglycemic phase, islets of IRE1$\alpha^{\beta-/-}$ mice show a strikingly disorganized architecture and altered islet composition. However, by 24 weeks of age, β cell morphology, islet architecture, and function were significantly improved in the islets of IRE1$\alpha^{\beta-/-}$ mice.

Bulk RNA-Seq on Intact Islets from Hyperglycemic IRE1$\alpha^{\beta-/-}$ Mice Indicates Changes in the Expression of Cell Survival and Differentiation Markers

To gain insight into the molecular mechanisms and consequences of this altered cellular composition, we sequenced bulk RNA from islets of IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ NOD female mice at 7 (hyperglycemic phase) and 15 weeks of age (recovery phase). Differential expression analysis of IRE1$\alpha^{\beta-/-}$ and IRE1$\alpha^{fl/fl}$ mice at 7 weeks of age (hyperglycemic phase) with edgeR (McCarthy et al., 2012; Robinson et al., 2010) at false discovery rate (FDR) of 0.05 identified 2,320 upregulated and 1,918 downregulated genes (Figures 4A and 4B). Gene set enrichment analysis (Yu et al.,

2012) of these differentially expressed genes (DEGs) revealed epithelial-mesenchymal transition, hypoxia, estrogen response, and KRAS signaling as top hits (Figure 4C). Gene ontology (GO) analysis revealed significant enrichment in several biological process, cellular components, and molecular functions. Most notable among these are upregulated extracellular matrix and downregulated vesicle organization (Figures S3A and S3B). In contrast, we identified 342 upregulated and 729 downregulated genes between the IRE1$\alpha\beta^{-/-}$ and IRE1$\alpha^{fl/fl}$ mice at 15 weeks of age (recovery phase) with major changes in sterol and cholesterol transporter activity (Figures S3C–S3F). Interestingly, markers of β cell maturation/ dedifferentiation were also significantly changed in IRE1$\alpha^{\beta-/-}$ mice during the hyperglycemic phase, suggesting a loss of mature β cell identity under stress conditions (Figure 4D). To examine if the reduced mRNA levels correlate with protein levels, we analyzed the expression of key markers of β cell maturity, Ucn3 and MafA (Blum et al., 2012; Matsuoka et al., 2004; van der Meulen et al., 2012), in control and IRE1$\alpha^{\beta-/-}$ mice via immunofluorescence assay. The protein expression levels of maturity markers were visibly diminished in β cells of 4-week-old IRE1$\alpha^{\beta-/-}$ mice (corresponding to the hyperglycemic phase) compared with those in control IRE1$\alpha^{fl/fl}$ mice (Figure 4E).

Bulk RNA-seq revealed significantly increased expression of islet hormones glucagon, somatostatin, PPY, and reduced insulin 1 and insulin 2 in IRE1$\alpha^{\beta-/-}$ mice (Figure 4F), consistent with the increases in non-b endocrine cells observed by histology (Figure 3). Interestingly, in addition to a substantially diminished expression of the β cell maturity markers, the expression of β cell ‘‘disallowed genes’’ (Pullen et al., 2010; Quintens et al., 2008; Thorrez et al., 2011), which are typically repressed in

mature adult β cells, were markedly increased in IRE1α$^{β-/-}$ mice (Figure 4G). It has been previously shown that DNA methyl transferase 3a (*Dnmt3a*) directs the methylation and repression of disallowed genes during β cell maturation (Dhawan et al., 2015). Consistent with an increased expression of disallowed genes, our RNA-seq data revealed significantly reduced *Dnmt3a* expression (p value of 7.39e-18) in IRE1α$^{β-/-}$ islets (Figure 4G). Finally, we detected markedly increased expression of the ErbB family of genes, regeneration-related genes, and growth factors in IRE1α$^{β-/-}$ islets (Figures 4H–4J). Together, bulk RNA-seq on intact islets from IRE1α$^{β-/-}$ mice indicates alterations in the expression of cell survival and differentiation markers during the hyperglycemic phase.

Single-Cell RNA-Seq Identifies Altered Proportion of Islet Cell Clusters, Hormonal Expression, and Expression of Non-b Cell Islet Cell Markers in β Cells of IRE1α$^{β–/–}$ Mice

Given that changes in the expression profile in the whole islets of IRE1α$^{β-/-}$ mice could reflect either changes in individual cells or at the population level because of the altered islet cellular composition, we performed single-cell RNA-seq (scRNA-seq) analysis in disassociated islets obtained from mice that were 5 weeks of age. Monocle analysis of the cells partitioned them into distinct clusters based on their expression profiles (Figure 5A). As expected, the major populations were α, β, and δ cells, with a small proportion of ductal, endothelial, and immune cells. We also identified a minor β cell population both in wild-type NOD and IRE1α$^{β-/-}$ mice, indicating a greater degree of heterogeneity in cell identity. We provisionally designated the minor sub-population of β

cells as ‘‘beta2 cells.’’ The proportion of each population was markedly different in the knockout mice. Cell cluster analysis indicated decreased beta1 and increased beta2 and α cell populations in IRE1α$^{β-/-}$ mice (Figure 5B). Interestingly, percentages of ductal and endothelial cells were also increased in IRE1α$^{β-/-}$ mice, whereas minimal immune cell population did not differ between IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (Figures 5A and 5B). These expression changes were not observed in bulk sample averages, indicating that our scRNA-seq method can reveal novel molecular features associated with islet cell composition.

We then examined the hormonal expression of islet cells. The β cells of IRE1α$^{β-/-}$ mice demonstrated a polyhormonal phenotype with significantly decreased *Insulin1/Insulin2* expression and markedly increased glucagon (*Gcg*), somatostatin (*Sst*), and *PPY* expression (Figure 5C). In α cells, glucagon expression levels remained similar between the wild-type and knockout mice. The minimal levels of insulin expressed in α cells were further reduced, whereas PPY expression was markedly higher in the α cells of knockout mice (Figure 5C). The δ cells of IRE1α$^{β-/-}$ mice showed significantly reduced (2-fold with a p value of 2.4e-101) Ins1 expression, whereas *Gcg* expression was markedly upregulated (greater than 3-fold with a p value of 4e-198) (Figure 5C).

Loss of IRE1α in β Cells Induces Dedifferentiation

To investigate whether the changes in the characteristics of β cells were only restricted to the hormone genes, we examined the expression of the canonical α cell (*Irx2, Ttr*) (Petri et al., 2006; Su et al., 2012) and δ cell markers (*Hhex, Rbp4*) (Artner et

al., 2010; Zhang et al., 2014) within the β cell clusters. We identified significantly increased expression of α cell (*Irx2, Ttr*) (Figure 6A) and δ cell markers (*Hhex,Rbp4*) (Figure 6B). Interestingly, the changing expression pattern was accompanied by a reduction in the number of cells expressing β cell maturity markers, such as *MafA* and *Ucn3* (Blum et al., 2012; van der Meulen et al., 2012) (Figure 6C). There were also more cells expressing ‘‘disallowed’’ genes, such as *Olfm1* and *Ndrg4* (Pullen et al., 2010) (Figures 6D and S4A). Concomitant with the lack of maturity, there was a marked increase in the expression of dedifferentiation genes *Rfx2* and *Fabp3* (Kim-Muller et al., 2016; Szabat et al., 2011). These data, along with significantly increased expression of disallowed genes and endocrine progenitor cell markers (*Aldh1a3*, *Gast*) (Cinti et al., 2016; Gittes et al., 1993) (Figure 6E), confirmed the dedifferentiation of β cells in IRE1$\alpha^{\beta-/-}$ mice. Interestingly, we also detected markedly increased expression of regeneration/proliferation-related genes in β cells of IRE1$\alpha^{\beta-/-}$ mice (Figure S4B). Finally, we demonstrated that these dedifferentiated cells indeed lost IRE1α by examining the gene expression profile of targets of *sXBP1*. Expression of many of the *sXBP1* targets (*Fkbp11*, *Erol1b*, *Fkbp2*, *Sec61g*, and *Pdia5*) was significantly reduced in β cells (Lee et al., 2003) (Figures 6F and S4C).

Next, we performed differential gene expression analysis using the Monocle package. Using this approach, we identified 469 DEGs in beta1 and 412 DEGs in beta2 cell clusters (FDR < 0.01, FC > 2) (Figures 6G and 6H). The most significantly DEGs in these β cell clusters were associated with cell differentiation, growth, survival, and immune inhibition, whereas expression of several markers associated with pancreas development and protein secretion were significantly downregulated (Figures 6G and

6H). Taken together, bulk RNA-seq from intact islets, scRNA-seq, and histological analyses identify initiation of β cell dedifferentiation in IRE1α$^{β-/-}$ mice prior to insulitis.

β Cells of IRE1α$^{β-/-}$ Mice Have Altered Expression of Genes Associated with Immune Cell Recruitment

β cells can actively participate in their autoimmune destruction by affecting the local homing of inflammatory cells, antigen presentation, and the levels of autoantigen or neoantigen expression. Thus, we examined the expression of genes that are involved in immune regulation in β cells of IRE1α$^{β-/-}$ mice. Interestingly, several immune-related genes and genes that are involved in immune cell recruitment were differentially regulated in β cells of IRE1α$^{β-/-}$ mice. Among these, significantly increased expression of genes that play a key role in the suppression of T cell, B cell, and macrophage activities was notable (Figure 7A). For example, the expression of immune inhibitory ligands Qa-2 (*H2-Q7/H2-Q9*) and Qa-1 (*H2-T23*), which were shown to play a major role in the suppression of CD4$^+$ T cell and natural killer (NK) cell responses (Carosella et al., 2008; Jiang et al., 1995; Klein et al., 1983; Robinson et al., 1989), was significantly upregulated (*H2-Q7*, p = 1.94e-20; *H2-T23*, p = 6.66e-25) in β cells of IRE1α$^{β-/-}$ mice. Interestingly, we found that the expression of tumor necrosis factor (TNF) receptor superfamily (*Tnfrs*) genes, which play a key role in the antigen presentation and in the generation of cytotoxic T cells (Ward-Kavanagh et al., 2016), was significantly downregulated (e.g., *Tnfrsf9*, p = 1.3257e-4; *Tnfrsf23*, p = 1.94e-10). The expression of several members of the tetraspanin family of genes (*Cd81*, *Cd9*, and *Cd151*), which also play important roles in the regulation of pattern recognition, antigen presentation,

and T cell proliferation (Jones et al., 2011), was also significantly altered in IRE1$\alpha^{\beta-/-}$ mice.

We demonstrated that protein expression of proinsulin and insulin, key autoantigens in triggering T1D (Arvan et al., 2012; Narendran et al., 2003; You and Chatenoud, 2006), was significantly reduced in IRE1$\alpha^{\beta-/-}$ mice (Figures 2F and 2G). We further examined whether the expression of other autoantigens was also altered in β cells of IRE1$\alpha^{\beta-/-}$ mice. In addition to proinsulin and insulin, scRNA-seq revealed substantially reduced expression of additional β cell autoantigens *Ins1/2*, *IAPP*, and *Ptprn* in IRE1$\alpha^{\beta-/-}$ mice (Figure 7B). Major histocompatibility complex (MHC) class I molecules, which present peptides derived from intracellular proteins to $CD8^+$ T cells, are assembled in the ER; ER stress was shown to affect MHC class I expression as well as processing of MHC-class-I-associated peptides (Granados et al., 2009; Ulianich et al., 2011). To elucidate whether the antigen presentation pathway was altered in IRE1$\alpha^{\beta-/-}$ mice, we examined the expression of MHC class I components and the peptide loading pathway genes. Analysis of scRNA-seq data identified significantly diminished expression of β2-microglobulin (β2m) (Figure 7C) and marked alterations in the expression of several MHC class I peptide loading pathway genes, suggesting that antigen processing was defective in IRE1$\alpha^{\beta-/-}$ mice (Figure 7D).

To evaluate the effects of IRE1α deficiency on the immune cells, we performed immunophenotyping in the pancreas, spleen, and pancreatic lymph nodes (PLNs) of these mice at 21 weeks of age. Although there were no significant alterations in the percentage of $CD4^+$ T cells, the percentage of $CD8^+$ T cells in the pancreata of IRE1$\alpha^{\beta-/-}$ mice was significantly reduced (Figures 7E and 7F). The percentage of T regulatory

cells (Tregs) (Figures S5A and S5B), B cells (Figures S5C and S5D), or macrophages (Figures S5E and S5F) did not show significant alterations in the pancreas. There were also no differences in these immune cell populations within the lymph nodes and spleen in control compared with those in IRE1$\alpha^{\beta-/-}$ mice (Figures S6 and S7). Hence, deletion of IRE1α in β cells in NOD mice significantly reduced the relative representation of $CD8^+$ T cells specifically in the pancreas, whereas other immune cell populations did not show any obvious alterations. Of note, these mice show significantly less islet infiltration by immune cells, suggesting that better islet function and survival recruited fewer immune cells to the islets, resulting in less antigen presentation. Although our immunophenotyping studies analyze the immune cells in the later stage of the disease, it is well established that macrophages and dendritic cells play a critical role during the initiation of the disease process. Thus, we examined whether the relative representation of these immune cells was altered in the islets of control and IRE1$\alpha^{\beta-/-}$ mice at 5 weeks of age. However, we did not detect any significant differences in the proportions of $Cd11c^+$ dendritic and F4/F80+ macrophage cell populations in IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice, suggesting that recruitment of these immune cells into the islets was unaltered (Figures 7G–7J). Finally, to determine the diabetogenic potential of T cells of IRE1$\alpha^{\beta-/-}$ mice, we performed adoptive transfer experiments and transferred purified total T cells of 8-week-old IRE1αfl/fl and IRE1$\alpha^{\beta-/-}$ mice into 5- to 6-week-old female immunodeficient NOD $Rag1^{-/-}$ mice and monitored them for diabetes development. Recipient mice transferred with IRE1$\alpha^{fl/fl}$ mouse T cells developed diabetes at 16 weeks after the cell transfer, and 60% of recipient mice became diabetic by 20 weeks after cell

transfer (Figure 7K). In contrast, transfer of IRE1$\alpha^{\beta-/-}$ T cells did not induce diabetes in NOD Rag1$^{-/-}$ mice, suggesting that IRE1$\alpha^{\beta-/-}$ T cells were not diabetogenic.

Taken together, our data suggest that IRE1α deletion in β cells, prior to insulitis, promotes transient β cell dedifferentiation, which significantly diminishes autoantigen expression and antigen processing, and increases the expression of immune inhibitory markers within the β cells. These phenotypic changes in β cells early in life in IRE1$\alpha^{\beta-/-}$ NOD mice most likely have effects on the autoimmune responses leading to substantially reduced CD8$^+$ T cells in the pancreas. The T cell adoptive transfer experiment further indicates that there are long-lasting effects on IRE1$\alpha^{\beta-/-}$ T cells, rendering them incapable of inducing diabetes in NOD Rag1$^{-/-}$ mice.

Discussion

To investigate the β cell-specific function of the key UPR sensor IRE1α in T1D, we generated NOD IRE1$\alpha^{\beta-/-}$ mice by exposing pups to tamoxifen via their dam's milk to achieve IRE1α deletion prior to islet infiltration of immune cells, which usually occurs later in the postnatal period. These mice, after a transient mild hyperglycemia, were protected from T1D. Recently, β cell-specific deletion of IRE1α, driven by Ins-Cre in unstressed wild-type mice (mixed C57BL/6 3 129/SvJae background), was shown to result in hyperglycemia starting from 4 weeks of age lasting up to at least 24 weeks of age (Tsuchiya et al., 2018). Similarly, deletion of IRE1α in adult β cells led to diabetes under a non-autoimmune context (C57BL/6 background) (Hassler et al., 2015). Interestingly, unlike NOD IRE1$\alpha^{\beta-/-}$ mice, no β cell dedifferentiation, bihormonal islet cells, and altered islet architecture or composition were observed in these IRE1α-

deleted mouse models, despite presence of chronic hyperglycemia (Hassler et al., 2015; Tsuchiya et al., 2018). These data indicate that loss of IRE1α in β cells has remarkably different outcomes in the context of β cell stress as seen in NOD mice.

Stress-induced dedifferentiation is well described in plants and mammalian somatic cells, such as Schwann cells, cardiac myocytes, germ cells, and β cells (Bersell et al., 2009; Chen et al., 2007; Talchai et al., 2012). Cells can use dedifferentiation as an adaptive mechanism to minimize damage (Puri et al., 2015; Shoshani and Zipori, 2011). Our scRNA-seq, bulk RNA-seq, and histological analyses demonstrate that β cells of IRE1$\alpha^{\beta-/-}$ mice have markedly increased expression of disallowed genes, and increased expression of markers of progenitor cells (*Aldh1a3*, *Gast*, and *Ngn3*), as well as reduced gene and protein expression of β cell maturity markers (*MafA* and *Ucn3*). In addition, the presence of bihormonal (Ins^+ , Glu^+) cells detected in pancreatic sections indicates that β cells of IRE1$\alpha^{\beta-/-}$ mice similarly underwent a reversible dedifferentiation process under the chronic stressed background of NOD mice. Of note, immune-independent β cell fragility, as a result of genetic variations in *Glis3* and *Xrcc4*, was shown to alter the responses of β cells to ER stress in NOD mice (Dooley et al., 2016).

Our histological and scRNA-seq data indicate increased numbers of α cells, suggesting that an α cell to β cell conversion could potentially be a mechanism for the restoration of the β cell population in islets of IRE1$\alpha^{\beta-/-}$ mice. However, conversion of a cells to β cells was reported only after extreme β cell loss (>90%) (Thorel et al., 2010), and our scRNA-seq analysis did not indicate a ‘‘β cell-like’’ signature in α cells of IRE1$\alpha^{\beta-/-}$ mice, suggesting that this may not be the main mechanism of recovery from

hyperglycemia. Successful redifferentiation of dedifferentiated β cells was reported in both mouse and human islets (Gershengorn et al., 2004; Ouziel-Yahalom et al., 2006; Wang et al., 2014). In addition, reduced insulin production was shown to promote β cell proliferation in a cell-autonomous manner (Szabat et al., 2016). Interestingly, a recent study shows that increased proliferation of β cells prior to insulitis in NOD mice is protective against T1D (Dirice et al., 2019). Thus, increased proliferation of immature β cells of IRE1$\alpha^{\beta-/-}$ mice, and/or non-recombined cells prior to insulitis, might have contributed to a diabetes-protected phenotype. However, we detected significantly less proliferation in IRE1$\alpha^{\beta-/-}$ mice at 3 weeks of age and no difference in proliferation at 5 weeks of age. Indeed, our results are consistent with previous reports indicating significantly reduced β cell proliferation upon deletion of UPR sensors Perk and Xbp1 in β cells (Lee et al., 2011; Zhang et al., 2006), ruling out the possibility that increased proliferation prior to insulitis contributes to the protection from T1D in IRE1$\alpha^{\beta-/-}$ mice. At the molecular level, Betacellulin, a ligand in ErbB signaling, was shown to play a key role in the redifferentiation and restoration of β cell gene expression and insulin content in human islets (Ouziel-Yahalom et al., 2006). Consistent with this observation, both our bulk and scRNA-seq indicate markedly increased expression of genes that are involved in β cell growth and the ErbB pathway. Indeed, Amphiregulin (Areg), an ErbB pathway ligand, has a pro-regenerative function and plays an important role in promoting the healing, and regeneration of multiple tissues, and was our top hit in bulk RNA-seq (greater than 200-fold) (Burzyn et al., 2013; Monticelli et al., 2011; Shao and Sheng, 2010). In addition to redifferentiation, neogenesis might also have contributed to the recovery from hyperglycemia in IRE1$\alpha^{\beta-/-}$ mice. Indeed, we observed small islet clusters

(<10 insulin-positive cells) and the occasional lone β cell in pancreata of IRE1α$^{β-/-}$ mice. The bulk RNA-seq demonstrated increased expression of regeneration genes in islets, and scRNA-seq revealed significantly increased ductal cell clusters in IRE1α$^{β-/-}$ mice. Unfortunately, as Cre or reporter lines on the NOD background for islet and ductal cells are not currently available, these possibilities cannot yet be explored directly with lineage tracing experiments.

How can the loss of IRE1α in β cells protect against autoimmune destruction? Could undifferentiated, immature β cells have reduced antigenicity and altered immune activating/regulating signatures that can avoid autoimmune destruction? Insulin and proinsulin have a key role in the initial triggering of the autoimmune response and driving of autoimmune β cell destruction (Arvan et al., 2012; Nakayama, 2011; Nakayama et al., 2005). Interestingly, among all the different islet cell types, only the highly secretory β cells are specifically targeted by immune cells in T1D. Thus, reducing insulin levels and allowing highly secretory β cells to rest during a critical window of the disease, together with altering ER functional capacity to disrupt assembly of MHC complex and peptide processing, may be crucial to prevent subsequent immune-mediated destruction of β cells. Recently, a specific sub-population of β cells (15%) in NOD mice was described to resist autoimmune attack. These β cells expressed reduced levels of β cell-specific genes in the face of autoimmunity, suggesting dedifferentiation of β cells (Rui et al., 2017). In addition to expressing significantly lower levels of insulin, these immune-resistant β cells exhibited substantially reduced expression of autoantigens (*Igrp*, *ZnT8*, *Gad1*, and *Ia-2*) and markedly increased expression of immune inhibitory genes (*Qa-2*, *Cd81*) compared with that in normal β

cells. Although these cells did not show altered expression of a subset of ER stress-related genes, none of the markers assessed in that study were direct targets of the IRE1α/sXBP1 branch of the UPR (Rui et al., 2017). Thus, in the same vein, in the presence of immune-independent β cell fragility in NOD mice (Dooley et al., 2016), IRE1α deletion might have caused loss of mature β cell identity. Immature β cells may have escaped autoimmune attack because of their significantly reduced expression of autoantigens, altered antigen processing, and upregulated expression of immunomodulatory genes. Consistent with this, levels of major autoantigens proinsulin, insulin, *Iapp*, and *Ptprn* were significantly reduced in dedifferentiated β cells of IRE1$\alpha^{\beta-/-}$ mice. Moreover, ER stress and the UPR can directly affect MHC class I assembly, peptide processing, and antigen presentation by regulating protein translation, degradation, decay of ER mRNAs, and ER homeostasis (Granados et al., 2009; Ulianich et al., 2011). Indeed, we identified markedly reduced expression of MHC class I component *β2m* and altered expression of MHC class I peptide loading pathway genes in IRE1$\alpha^{\beta-/-}$ mice, suggesting that protection of IRE1$\alpha^{\beta-/-}$ mice from T1D was, in part, because of UPR-dependent defects in production and processing of autoantigens, which in turn significantly reduced cytotoxic $CD8^+$ T cells and islet infiltration. Protection from autoimmune destruction is further supported by significant upregulation of immune inhibitory markers, downregulation of genes that are implicated in immune cell activation, and markedly altered expression of chemokines, cytokines, and ECM proteins, which play an important role in immune cell recruitment.

NOD IRE1$\alpha^{\beta-/-}$; Rag1$^{-/-}$ mice exhibited the same β cell functional alteration as seen in NOD IRE1$\alpha^{\beta-/-}$ mice. Thus, the phenotypic changes of the β cells in IRE1$\alpha^{\beta-/-}$

NOD mice are likely cell intrinsically regulated, but not a result of altered adaptive immune cells. In addition, two findings suggest that β cell dedifferentiation observed early in life in NOD IRE1α$^{β-/-}$ mice alters the autoimmune response of the adaptive immune cells. First, the majority of NOD IRE1α$^{β-/-}$ mice remained non-diabetic at 50 weeks of age. Second, T cells isolated from 8-week-old NOD IRE1α$^{β-/-}$ mice could not induce diabetes in NOD Rag1$^{-/-}$ recipients. Collectively, our results support the idea that dedifferentiation of β cells in young NOD IRE1α$^{β-/-}$ mice has long-lasting effects on the diabetogenic activity of T cells. There are several non-mutually exclusive mechanisms that could contribute to reduced diabetogenic activity of IRE1α$^{β-/-}$ T cells. The β cell autoreactive CD8^{+} T cells could be tolerized in the forms of anergy or deletion when they are not properly stimulated in NOD IRE1α$^{β-/-}$ mice. It is also possible that T cells with regulatory functions are enhanced in NOD IRE1α$^{β-/-}$ mice. Although we did not observe a proportional difference in FOXP3^{+} CD4^{+} Tregs in NOD IRE1α$^{β-/-}$ mice, the possibility that they are functionally enhanced cannot be ruled out. Future studies are needed to determine the mechanism underlying immune tolerance induction of β cell autoreactive T cells in NOD IRE1α$^{β-/-}$ mice.

Aberrant expression of the UPR genes was detected in β cells of mouse models of diabetes and human patients (Engin et al., 2013, 2014). Mitigation of ER stress and restoration of the UPR dysfunction with a chemical chaperone, TUDCA, prevented diabetes in pre-clinical T1D models (Engin et al., 2013). TUDCA is currently under phase I clinical trial (NCT02218619) for patients with new-onset T1D. Interestingly, the tyrosine kinase inhibitor imatinib, currently being tested in a phase II clinical trial (NCT01781975) for the treatment of new-onset T1D, was recently demonstrated to

blunt RNase activity of IRE1α and reverse T1D in the NOD mouse model (Louvet et al., 2008; Morita et al., 2017). These studies support the notion that modulating β cell UPR can be a promising therapeutic strategy for people at high risk for T1D.

Autoantibodies directed against β cell proteins are used as biomarkers for risk prediction and clinical diagnosis. The presence of multiple autoantibodies is associated with a high risk of progression to overt disease (Regnell and Lernmark, 2017). The relationship between autoantibody positivity and the presence or absence of insulitis is an actively pursued research area (Pugliese, 2016). Emerging data suggest that donors with multiple autoantibodies can have absence of insulitis (In't Veld et al., 2007; Wiberg et al., 2015). Thus, inducing a reversible dedifferentiation state for β cells to limit their antigen availability during this critical therapeutic window may provide an important non-immune-based preventive or therapeutic strategy in high-risk individuals. Interfering with antigen processing and presentation, by modulating β cell ER functional capacity and the UPR, can further support diabetes protection. Whether similar strategies can be applicable to prevent other autoimmune diseases associated with highly secretory target cells remains to be tested. Future studies identifying the function of IRE1α and the other UPR sensors during different stages of T1D progression will be necessary to fully reveal the role of β cell ER stress and the UPR in T1D.

Limitations of Study

Our current breeding scheme does not allow us to obtain the littermate control mice expressing Cre transgene alone. Thus, the mice used to identify the effects of Cre expression on diabetes progression were not obtained from the experimental group.

Although we confirmed that Cre transgene levels in these mice did not differ from the knockout mice, and Cre transgene did not alter diabetes progression and pathology in NOD mice, we still consider this a limitation. In addition, due to the lack of reporter lines on NOD background, we were not able to perform lineage tracing experiments to definitively identify the contribution of neogenesis or transdifferentiation to the recovery from hyperglycemia in IRE1$\alpha^{\beta-/-}$ mice.

Acknowledgments

We thank Dr. Mark O. Huising for his critical reading of the initial draft of the manuscript. H.L. is supported by NIH National Research Service Award T32 GM007215. Y.-G.C. is supported by NIH grants DK107541 and DK121747. A.E.C. is supported by the NIH F31 award DK118786. S.P. is supported by an NIH T32 NHGRI 5T32HG002760 award. R.S. is supported by NIH R01GM113033. F.E. is supported by grants from the JDRF-5-CDA2014-184-A-N, NIH 5K01DK102488-03, and startup funds from the University of Wisconsin-Madison School of Medicine and Public Health and Department of Biomolecular Chemistry. The flow cytometry data were collected with resources and facilities provided by the University of Wisconsin Carbone Cancer Center–Flow Cytometry Laboratory with funding from support grant P30 CA014520.

Declaration of Interests

The authors declare no competing interests

Materials and Methods

Experimental Model and Subject Details

Mouse Lines and Tamoxifen Injections

The animal care and experimental procedures were carried out in accordance with the recommendations of the National Institutes of Health Guide for the Care and Use of Laboratory Animals. The protocol (#M005064-R01-A03 by F.E. for mice) was approved by the University of Wisconsin-Madison Institutional Animal Care and Use Committee. Female NOD/ShiLtJ mice and $Rag1^{-/-}$ mice were purchased from Jackson Laboratory. All animals had *ad libitum* access to food (Envigo 2919) and water and were housed at 20-24°C on a 12 h light/12 h dark cycle. Mice were bred and maintained under specific pathogen-free conditions at University of Wisconsin-Madison under approved protocols. The IRE1α floxed mice were a gift of Dr. Takao Iwawaki (Kanazawa Medical University). Ins2-$Cre^{ERT/+}$ mice were a gift of Dr. Douglas Melton (Harvard University). Mice were backcrossed to NOD background more than 10 generations. The genetic backgrounds of all intercrossed mouse models were verified by Genome Scan Services (purity of IRE1$\alpha^{fl/fl}$ mice on NOD background >99.8% via Jackson Laboratory's Genome Scan Service, purity of Ins2Cre^{ERT} mice on NOD background >99.9% via Neogen, MiniMUGA array). To induce Cre recombinase activity, tamoxifen (T5648; Sigma-Aldrich) was dissolved in sterilized corn oil (C8267; Sigma-Aldrich) by shaking overnight in a 37°C incubator. The solution was protected from light, and 10 mg/ml tamoxifen was administrated to dams the day after delivery via intraperitoneal injection twice every 24 hours in five consecutive days. Animals were observed daily for health status, any mice that met IACUC criteria for euthanasia were immediately euthanized. Experiments were performed on female mice between 3 and 50 weeks of age.

Method Details

Histological Analyses

Pancreata from mice were fixed with 10% zinc formalin overnight and paraffin embedded. 5-μm sections of the pancreata were generated, and staining was performed after blocking with 5% normal goat serum with the following antibodies: anti-Insulin (Linco), anti-Glucagon (Cell Signaling), anti-Somatostatin (Santa Cruz), anti-Ki67 (Cell Signaling), anti-CD3 (Novus Biologicals), anti-MafA (Cell Signaling), anti-Ucn3 (Phoenix Pharmaceuticals), Alexa Fluor 488 (Invitrogen), and Alexa Fluor 568 (Invitrogen) using established protocols. After staining, slides were mounted with antifade mounting medium containing 4,6-diamidino-2-phenylindole (DAPI) (Vector Laboratories). In some cases, the harvested pancreata were fixed in 4% paraformaldehyde (VWR), embedded in OCT (Sakura), and frozen before being sectioned at 10 μm. Antigen retrieval was performed by using citrate buffer pH 6.0 (paraffin) or HistoVT (Nacalai Tesque) (frozen). Islet size was determined by manually circling insulin positive clusters in the Fiji software (Schindelin et al., 2012). Insulitis scoring was performed on step sections (three levels separated by 200-μm) of paraffin-embedded and hematoxylin-eosin stained sections. ‘‘peri-insulitis’’ is defined as focal aggregation at one pole of the islet and in contact with the islet periphery. ‘‘non-aggressive insulitis’’ refers to lesions with a clear, and often extensive, islet infiltrate occupying less than 50% of the islet area, whereas ‘‘aggressive insulitis’’ refers to an extensive infiltrate, where lymphoid cells invade the entire islet and intermingle with endocrine cells, showing extensive signs of β-cell damage. The images of the

pancreatic sections were obtained using a Nikon A1R-SI+ confocal microscope and a Nikon Storm/Tirf/Epifluorescence. The images were analyzed by ImageJ or Fiji. Two blinded individuals independently performed manual analyses and insulitis scoring.

Cell Death (TUNEL) Assay

DeadEnd Fluorometric TUNEL assay (Promega Corporation, Madison, WI) was performed on formalin-fixed, paraffin-embedded pancreatic sections according to the manufacturer's instructions.

Islet Isolation

Islets were isolated using the standard collagenase/protease digestion method. Briefly, the pancreatic duct was cannulated and distended with 4°C collagenase/protease solution using Collagenase P (Sigma-Aldrich, USA) in 1x Hank's balanced salt solution and 0.02% bovine serum albumin. The protease reaction was stopped using RPMI 1640 with 10% fetal bovine serum. Islets were separated from the exocrine tissue using Histopaque-1077 (Sigma-Aldrich, USA). Hand-picked islets were cultured overnight at 37°C in RPMI-1640 media containing 10% FBS and 1% antibiotic/antimycotic (Thermo Fisher Scientific) before use in experiments.

Insulin, Proinsulin Content

For measurement of whole pancreas insulin and proinsulin content, mice were sacrificed after a 4-hour fast and their whole pancreas insulin/proinsulin content

(μg/pancreas) was assessed by acid-ethanol extraction followed by ELISA (Alpco, Salem, NH). Samples were done in duplicate.

Glucose Tolerance Test

Glucose tolerance tests were performed on IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice simultaneously after an overnight (16 hours) fast. Blood glucose levels were measured at 0, 15, 30, 60, 90, and 120 minutes after an intraperitoneal administration of glucose at dose of 2g/kg body weight.

Immunophenotyping

Prior to organ dissection, mice were perfused with 20 ml PBS to eliminate contaminating blood leukocytes. Single-cell suspensions of the pancreata were prepared by Collagenase P (Roche) digestion. Cells from pancreatic lymph nodes and spleen were prepared by physical dissociation. The spleen was treated with ACK lysing buffer (Thermo Fisher Scientific). All stainings began with an incubation with TruStain fcX anti-mouse CD16/32. Antibodies used for subsequent stainings were: anti-CD45 (30-F11), -CD19 (6D5); -CD3 (145-2C11), -CD4 (RM4-5), -CD8 (53-6.7), -CD25 (PC61), CD11b (M1/70), -CD11c (N418), -F4/80 (BM8), and -Gr1 (RB6- 8C5) (all from BioLegend). Intracellular Foxp3 (FJK-16s) staining was performed according to eBioscience's protocol. Samples were acquired with an Attune NxT flow cytometer (Thermo Fisher Scientific) and data were analyzed with FlowJo software (Tree Star).

Adoptive Transfer

The spleens from 8-week-old IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice were physically dissociated before filtering with a 40 μM nylon mesh. After incubation with ACK lysing buffer (Thermo Fisher Scientific), total T cells were isolated by negative selection (Pan T cell isolation kit II, Miltenyi Biotec), and $4x10^6$ cells were injected intravenously into 5-6-week-old NOD.Rag1$^{-/-}$ female recipients. The recipients were followed for diabetes incidence with weekly blood glucose measurements. The purity of the transferred T-cells (>95%) was analyzed by flow cytometry using anti-TCRb, anti-CD4, and anti-CD8 antibodies (BD Biosciences).

Bulk RNA-seq

Following isolation, RNA was extracted using RNeasy Plus Mini Kit (Qiagen), including a column for elimination of genomic DNA. RNA concentration was determined using Qubit RNA HS Assay Kit (Life Technologies). RNA Integrity Number (RIN) was measured using Agilent RNA 600 Nano Kit (Agilent Technologies). RIN > 7 was used in the experiments. RNA library was generated using the TruSeq Stranded Total RNA (Human/Mouse/Rat) (Illumina). Cytoplasmic ribosomal RNA was removed from the sample using complementary probe sequences attached to magnetic beads. Subsequently, each mRNA sample was fragmented using divalent cations under elevated temperature, and purified. First strand cDNA synthesis was performed using SuperScriptII (Invitrogen, Carlsbad, California, USA), reverse transcriptase, and random primers. Second strand cDNAs were synthesized using DNA polymerase I and RNase H for removal of mRNA. Double-stranded cDNA was purified using Agencourt AMPure XP beads (Qiagen, Valencia, California, USA) as recommended in the TruSeq RNA

Sample Prep Guide. The blunt ended cDNA and the adapter-ligated products were purified using Agencourt AMPure XP beads. Quality and quantity of finished libraries were assessed using an Agilent DNA1000 series chip assay (Agilent Technologies) and Invitrogen Qubit HS cDNA Kit (Invitrogen), respectively. Cluster generation was performed using a TruSeq Paired End Cluster Kit (v4) and the Illumina cBot, with libraries multiplexed for 1x100bp sequencing using the TruSeq 250bp SBS kit (v4) on an Illumina HiSeq2500. Images were analyzed using CASAVA 1.8.2.

Single Cell RNA-seq

Following islet isolation and an overnight culture, islet cells were dissociated for 30 min at 37°C into single-cell suspensions, using a cocktail of digestive enzymes (*Accutase*; Innovative Cell Technologies, San Diego, CA). Libraries were constructed according to the Chromium Single Cell 3' Reagent Kit v2 (10x Genomics, Pleasanton, CA). Briefly, cells in single cell suspension were delivered to the University of Wisconsin-Madison Biotechnology Center on ice, where the cell concentration and viability were quantified on the Countess II (Thermo Fisher Scientific) using 0.4% Trypan Blue (Invitrogen, Carlsbad, CA). The appropriate volume of cells was loaded onto the Single Cell A Chip required for yielding a targeted cell recovery of 3000 cells. Following the completion of the Chromium run, the Gel Bead-In EMulsions (GEMs) were transferred to emulsion-safe strip tubes for GEM-RT, using an Eppendorf Master Cycler Pro thermocycler (Eppendorf, Hamburg, Germany). Following RT, GEMs were broken, and the pooled single cell cDNA was amplified. Post-cDNA amplified product was purified using SPRIselect (Beckman Coulter, Brea, CA) and quantified on a Bioanalyzer 2100 (Agilent,

Santa Clara, CA) using the High Sensitivity DNA kit. Adapters were then added to the libraries after fragmentation, end repair, A-tailing, and double-sided size selection using SPRIselect. Following adapter ligation, libraries were purified using SPRIselect, and sample-specific indexes (Chromium i7 Multiplex Kit, 10x Genomics) were added by sample index PCR. After sample index PCR, samples were double-size selected using SPRIselect, yielding final libraries compatible for Illumina sequencing. Final libraries were quantified using the Qubit High Sensitivity DNA Kit (Thermo Fisher Scientific).

t-SNE Clustering

We used Monocle2 v2.8.0 on R version 3.5.2 (Kite-Eating Tree) http://cole-trapnell-lab.github.io/monocle-release/docs/ (Qiu et al., 2017) to analyze the data obtained after alignment. Cells with unique molecular identifier (UMI) counts outside a range determined of two standard deviations were filtered, leaving 2,749 cells within the optimal UMI range for downstream Monocle analysis. Genes that were not expressed in at least 10 cells were excluded from analysis. We reduced the number of dimensions to the number of Principal components that explained the most variance (at least 50% for each dataset).

To identify genes important for defining clusters, differential gene expression (DGE) analysis in Monocle2 was performed between all 7 clusters within the t-SNE plot of Figure 5A. To determine the distribution of cells from each sample that fall into each cluster, phenotypic data (cell barcode ID, sample, cluster number) was extracted for each cluster and sample. The composition of a cluster or sample was then calculated by percentage or mean of the population.

To visualize how different genes are expressed in cells that are known to be positive for a particular gene, we also generated violin plots using ggplot function in R. DGEs between IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice were clustered using Cluster 3.0 (de Hoon et al., 2004) and visualized using Java TreeView (Saldanha, 2004).

Quantification and Statistical Analysis

For all experiments the number of biological or technical replicates (*n*), error bars, and statistical analyses have been explained in the figure legends. For each experiment where statistics were computed, we used at least *n* = 3 or more biological replicates. Sample size were not pre-determined by power analysis, but sufficiency of number of mice were estimated based on pilot experiments and previously published work (Engin et al., 2013). Samples were randomly assigned and blinded for data analysis of immunostaining and insulitis scoring. No data were excluded in this study. Data are represented as mean ± SEM and were analyzed using either the unpaired Student's t-test, or one-way ANOVA where required. $P < 0.05$ was considered statistically significant; ns = non-significant as determined by statistical analysis in GraphPad Prism v.8 (GraphPad Software, San Diego, CA).

Data and Code Availability

Data Resources

The accession number for the RNA-seq data reported in this paper are NCBI GEO: GSE144461 (bulk sequencing) and GSE144471 (single cell sequencing).

Figure 1

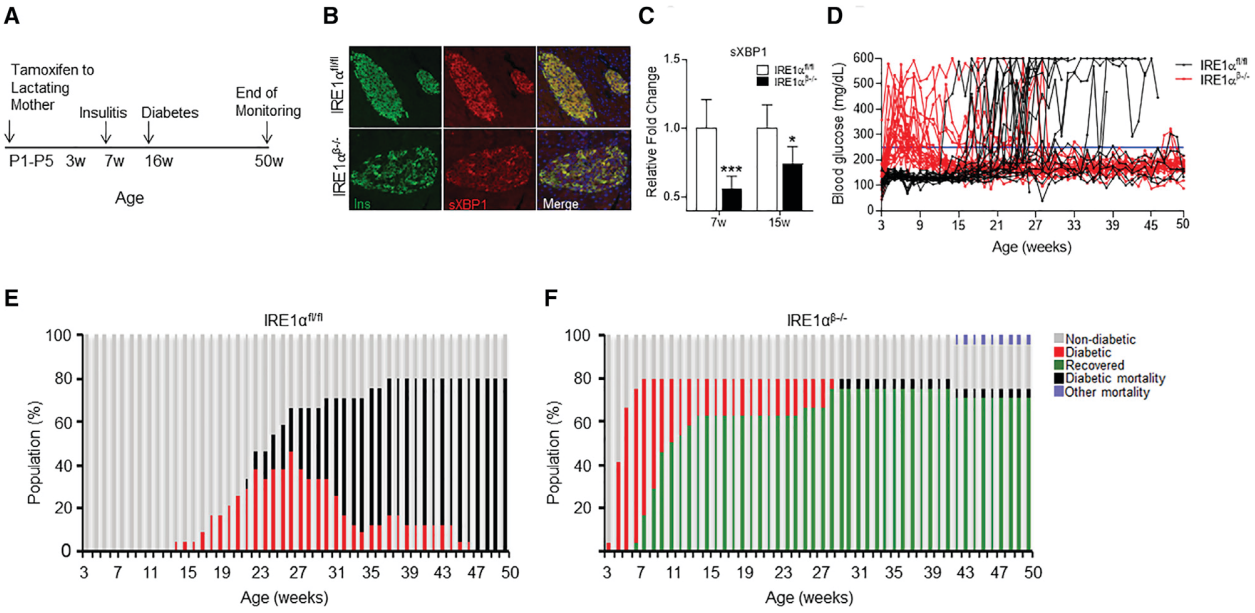


Figure 1: IRE1α$^{β-/-}$ NOD Female Mice Are Protected from T1D

A) Schematic representation of tamoxifen-induced deletion of IRE1α in β cells of NOD mice.

B) Representative image of an immunofluorescence staining on pancreatic sections from 5-week-old IRE1α$^{fl/fl}$ (upper panel) and IRE1α$^{β-/-}$ (lower panel) mice for sXBP1 expression. Sections were co-stained with anti-insulin (green) and anti-sXBP1 (red) antibodies.

C) Quantification of mRNA expression of sXBP1 in the islets of 7- and 15-week-old IRE1α$^{fl/fl}$ (7 weeks, n = 6; 15 weeks, n = 5) and IRE1α$^{β-/-}$ mice (7 weeks, n = 5; 15 weeks, n = 6) by qPCR. Data are averages of two technical replicates from a representative experiment.

D) Blood glucose levels of NOD control (IRE1$\alpha^{fl/fl}$) and IRE1$\alpha^{\beta-/-}$ mice (n = 24 per group), measured weekly upon weaning after tamoxifen administration to lactating mothers

E and F) Diabetes progression in (E) IRE1$\alpha^{fl/fl}$ and (F) IRE1$\alpha^{\beta-/-}$ mice monitored up to 50 weeks. All data are represented as mean ± SEM, with statistical analysis performed by Student's t test (***$p < 0.001$, *$p < 0.05$). w, weeks

Figure 2

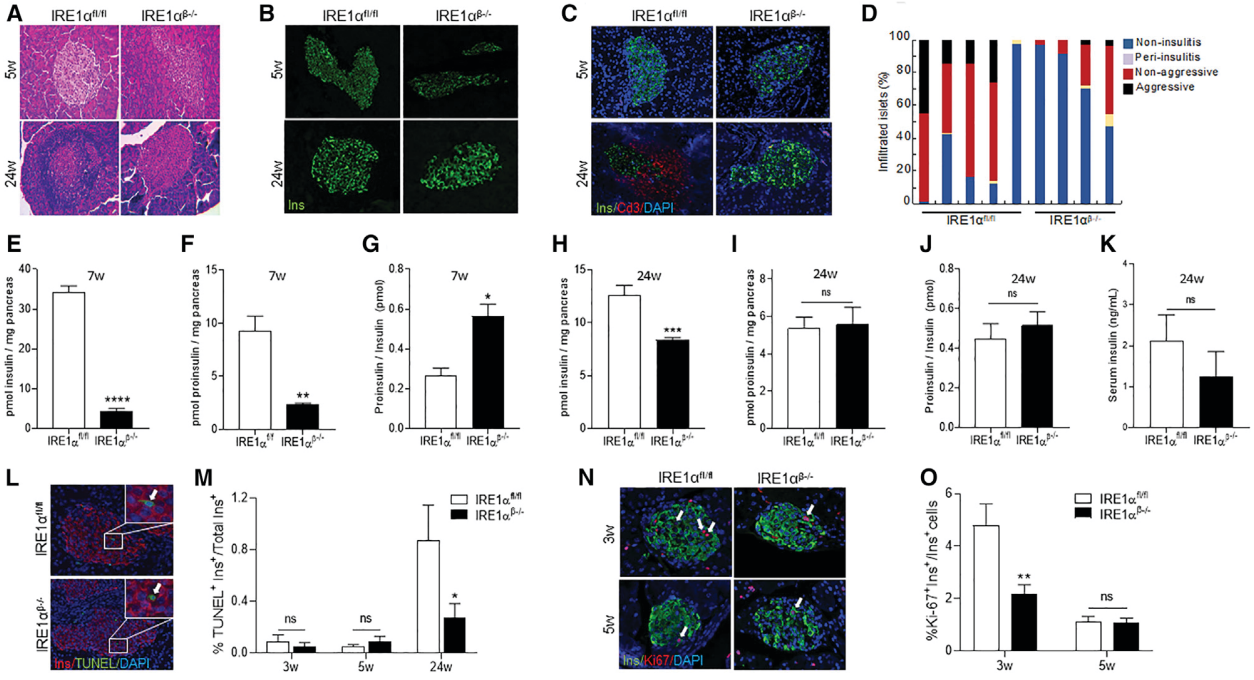


Figure 2. Improved β Cell Function and Survival in IRE1α$^{β–/–}$ NOD Mice upon Recovery from Hyperglycemia

A) Representative H&E staining of pancreatic sections from IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice at indicated time points.

B) Immunofluorescence staining on pancreatic sections from IRE1α$^{fl/fl}$ (5 weeks, n = 6; 24 weeks: n = 5) and IRE1α$^{β-/-}$ (5 weeks: n = 6; 24 weeks: n = 8) mice for insulin expression at indicated time points.

C) Immunofluorescence staining on pancreatic sections from IRE1α$^{fl/fl}$ (5 weeks: n = 6; 24 weeks, n = 5) and IRE1α$^{β-/-}$ (5 weeks, n = 6; 24 weeks, n = 8) for insulin (green) and CD3 (red) expression at indicated time points. The cell nuclei were counterstained with DAPI (blue).

D) Insulitis scoring assessed on H&E-stained step sections obtained from 24 weeks of age IRE1$\alpha^{fl/fl}$ (n = 5) and IRE1$\alpha^{\beta-/-}$ (n = 4) mice.

E and F) Insulin (E) and proinsulin (F) content of pancreata from 7-week-old mice (n = 4 per group) determined by ELISA and normalized per mg of pancreas.

G) Proinsulin-to-insulin molar ratio was calculated. Data are averages of two technical replicates from a representative experiment

H and I) Insulin (H) (n = 6 per group) and proinsulin (I) content of pancreata from 24 week-old IRE1$\alpha^{fl/fl}$ (n = 5) and IRE1$\alpha^{\beta-/-}$ (n = 7) mice determined by ELISA and normalized per mg of pancreas.

J) Proinsulin-to-insulin molar ratio was calculated. Data are averages of two technical replicates from a representative experiment.

K) Serum insulin levels of 24-week-old IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (n = 6 per group) determined by ELISA.

L) Representative images of TUNEL assay showing β cell apoptosis (arrow) on pancreatic sections from IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice. The cell nuclei were counterstained with DAPI (blue).

M) Percentage of β cell apoptosis calculated using pancreatic sections obtained from indicated time points (IRE1$\alpha^{fl/fl}$, 3, 5, and 24 weeks: n = 6, 6, and 5, respectively; IRE1$\alpha^{\beta-/-}$, 3, 5, and 24 weeks: n = 6, 6, and 8, respectively).

N) Representative fluorescence images of pancreatic sections from IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (n = 6–8 per group) for insulin (green) and Ki67 (red) expression (arrow) at indicated time points. Cells were counterstained with the nuclear dye DAPI (blue).

O) Quantification of $Ki67^{+}$ proliferating β cells in pancreatic sections from $IRE1\alpha^{fl/fl}$ (3 weeks, n = 6; 5 weeks, n = 7) and $IRE1\alpha^{\beta-/-}$ mice (3 weeks, n = 8; 5 weeks, n = 7) at indicated time points.

All data are represented as mean ± SEM, with statistical analysis performed by Student's t test (****p < 0.0001, **p < 0.01, *p < 0.05). w, weeks; ns, non-significant.

Figure 3

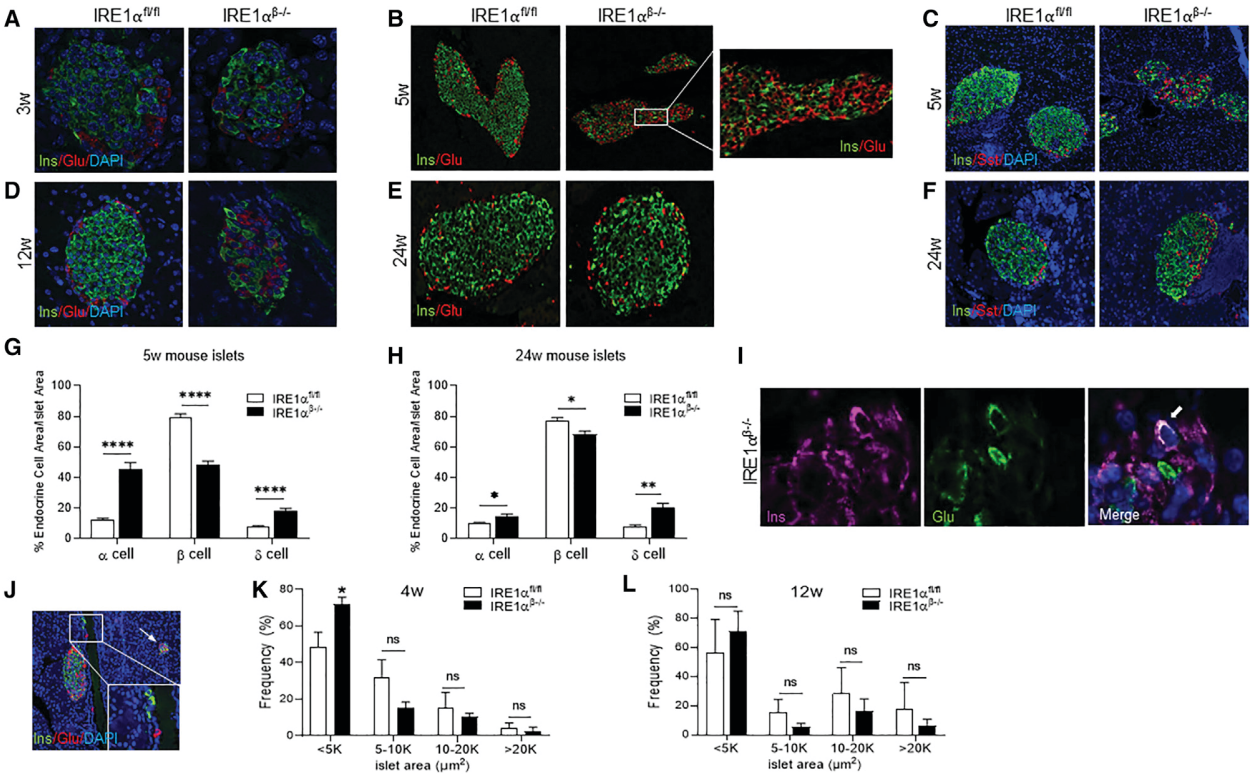


Figure 3. Islet Cell Composition Is Altered in IRE1α$^{β–/–}$ Mice during the Hyperglycemic Phase

A) Representative images of immunofluorescence staining on pancreatic sections from IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (n = 8 per group) for insulin (green) and glucagon (red) expression at 3 weeks of age. The cell nuclei were counterstained with DAPI (blue).

B) Representative images of immunofluorescence staining on pancreatic sections from IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (n = 8 per group) for insulin (green) and glucagon (red) expression at 5 weeks of age.

C) Representative images of immunofluorescence staining on pancreatic sections from IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (n = 8 per group) for insulin (green) and somatostatin (red) expression at 5 weeks of age. The cell nuclei were counterstained with DAPI.

D) Representative images of immunofluorescence staining on pancreatic sections from IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (n = 8 per group) for insulin (green) and glucagon (red) expression at 12 weeks of age. The cell nuclei were counterstained with DAPI (blue).

E) Representative images of immunofluorescence staining on pancreatic sections from IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (n = 8 per group) for insulin (green) and glucagon (red) expression at 24 weeks of age.

F) Immunofluorescence staining on pancreatic sections from IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (n = 8 per group) for insulin (green) and somatostatin (red) expression at 24 weeks of age.

G and H) Quantification of α, β, and δ cells as a percentage of total islet area at 5 weeks of age (G) and 24 weeks of age (H) using ImageJ (15–25 islets/animal/ time point).

I) Representative image of an insulin$^+$ (purple) and glucagon$^+$ (green) bihormonal cell (white) in an islet of IRE1$\alpha^{\beta-/-}$ mice. Arrow indicates the bihormonal cell.

J) Representative image of a pancreatic section from 5-week-old IRE1$\alpha^{\beta-/-}$ mice stained with anti-insulin (green) and anti-glucagon (red) antibodies, showing the presence of single β cells and small islet clusters. The arrow points to a small islet cluster. The cell nuclei were counterstained with DAPI (blue).

K and L) The quantification of islet area from H&E-stained sections at 4 weeks (K) and 12 weeks (L) of age IRE1$\alpha^{fl/fl}$ (n = 3 per time point) and IRE1$\alpha^{\beta-/-}$ (4 weeks, n = 3; 12 weeks, n = 4) mice by using ImageJ.

All data are represented as mean ± SEM, with statistical analysis performed by Student's t test (****$p < 0.0001$, **$p < 0.01$, *$p < 0.05$). w, weeks.

Figure 4

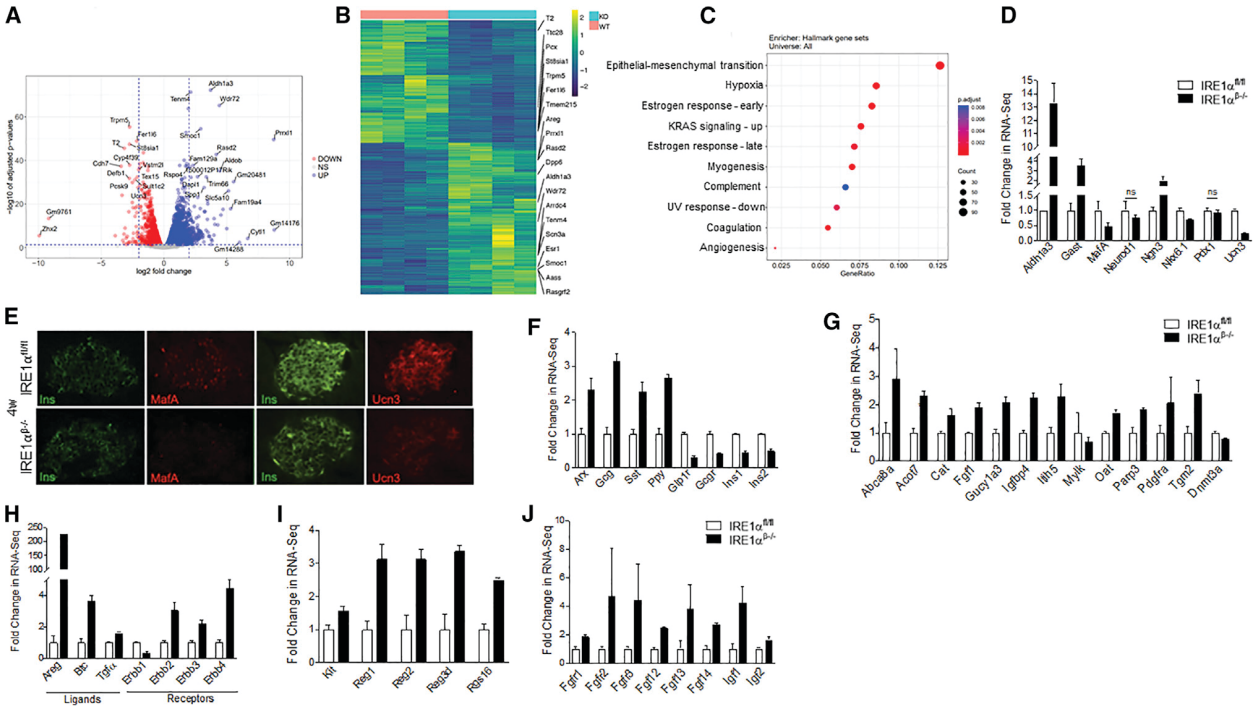


Figure 4. Bulk RNA-Seq on Intact Islets from Hyperglycemic Mice Indicates Changes in the Expression of Cell Survival and Differentiation Markers

A) Volcano plots for the edgeR analysis of the RNA-seq data from islets of 7-week-old IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ female NOD mice (n = 4 per group). Horizontal line depicts the FDR cutoff of 0.05 and the vertical lines mark log2 fold changes of -2 and 2. Genes with absolute log2 fold change larger than 5 or adjusted p value smaller than 1e-25 and absolute log2 fold change larger than 2 are labeled with their gene symbols.

B) Heatmap of expression levels for the DEGs that were identified in the IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice during the hyperglycemic phase (FDR < 0.01, FC > 2).

C) Gene set enrichment analysis, with the Molecular Signatures Database (MSigDB) Hallmark gene sets, identified ten pathways significantly associated with the upregulated genes in hyperglycemic IRE1$\alpha^{\beta-/-}$ mice in comparison with IRE1$\alpha^{fl/fl}$ mice (FDR < 0.05). GeneRatio in the x axis quantifies the proportion of the upregulated genes that are among each gene set, and Count depicts the number of upregulated genes in the gene set.

D) The mRNA expression of β cell identity and endocrine progenitor markers in the islets of 7-week-old IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (FDR < 0.05).

E) Immunofluorescence staining on frozen pancreatic sections from 4-week-old IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice for insulin (green), MafA (red), and Ucn3 (red) expression.

F–J) The mRNA expression of (F) islet cell markers, (G) disallowed genes, (H) ErbB family of genes, (I) regeneration-related genes, and (J) growth factor gene transcripts in the islets of 7-week-old IRE1$\alpha^{fl/fl}$ and IRE1$\alpha^{\beta-/-}$ mice (FDR < 0.05). ns, non-significant; w, weeks. FC, fold change.

Figure 5

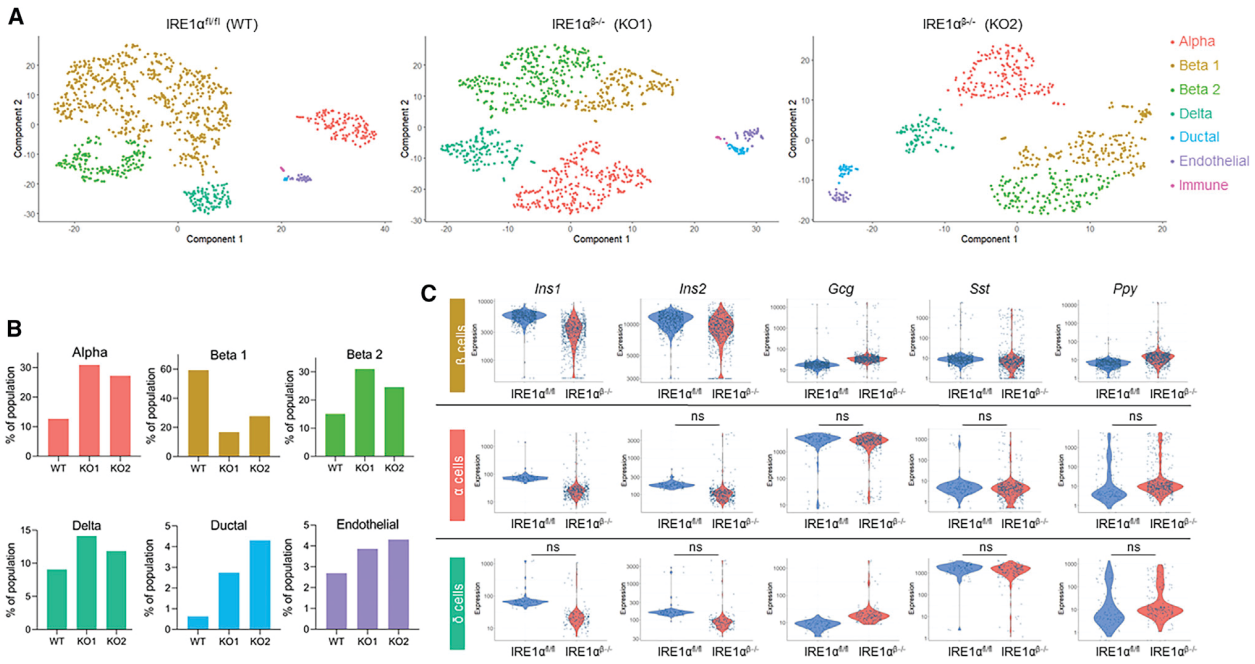


Figure 5. scRNA-Seq Identifies Altered Proportion of Islet Cell Clusters, Hormonal Expression, and Non-β Cell Islet Markers in IRE1α$^{β–/–}$ Mice

A) Distinct pancreatic islet cell clusters in IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice identified by Monocle package. Each dot represents a single cell, color-coded according to its cellular identity as defined by gene expression.

B) Percentage of population composed of cell sub-types identified in (A) in dissociated islets obtained from 5-week-old IRE1α$^{fl/fl}$ (wild-type, WT) and IRE1α$^{β-/-}$ (knockout 1 [KO1] and knockout2 [KO2]) mice.

C) Expression of islet hormones *Ins1*, *Ins2* (insulin), glucagon (*Gcg*), somatostatin (*Sst*), and pancreatic polypeptide (*Ppy*) in β cells (upper panel), α cells (middle panel), and δ cells (bottom panel) of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (FDR < 0.01). ns, non-significant.

Figure 6

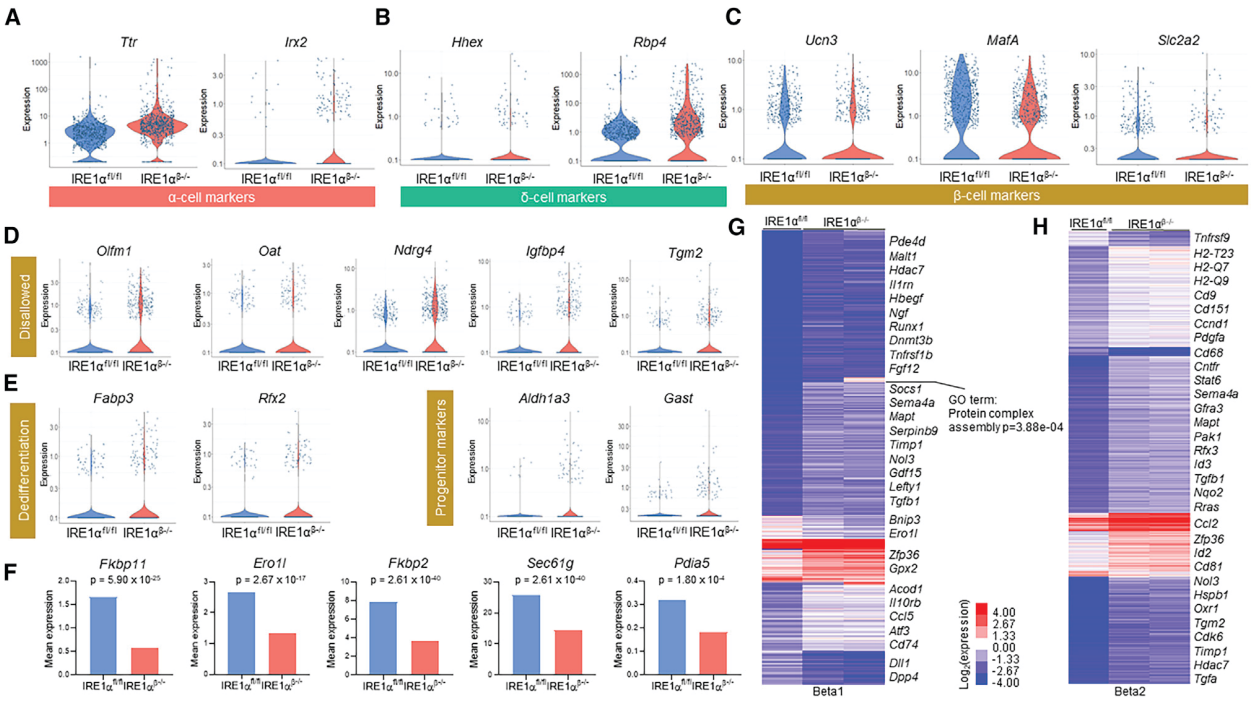


Figure 6. β Cells of IRE1α$^{β-/-}$ Mice Dedifferentiate

A–C) Expression of α cell markers (A) (*Ttr* and *Irx2*), (B) δ cell markers (*Hhex* and *Rbp4*), and (C) β cell maturity markers (*MafA*, *Ucn3*, and *Slc2a2*) in β cell clusters of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice at 5 weeks of age (FDR < 0.01).

D and E) Expression of disallowed genes (D) and dedifferentiation and endocrine progenitor markers (E) in β cell clusters of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (FDR < 0.01).

F) Mean expression of sXBP1 target genes in β cell clusters of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice.

(G and H) k-means clustering (seven clusters) of DEGs (FDR < 0.01, FC > 2) among the beta1 (G) and beta2 (H) populations (columns). Selected genes that define each cluster are displayed. Color bar represents expression changes in log2 scale. FC, fold change.

Figure 7

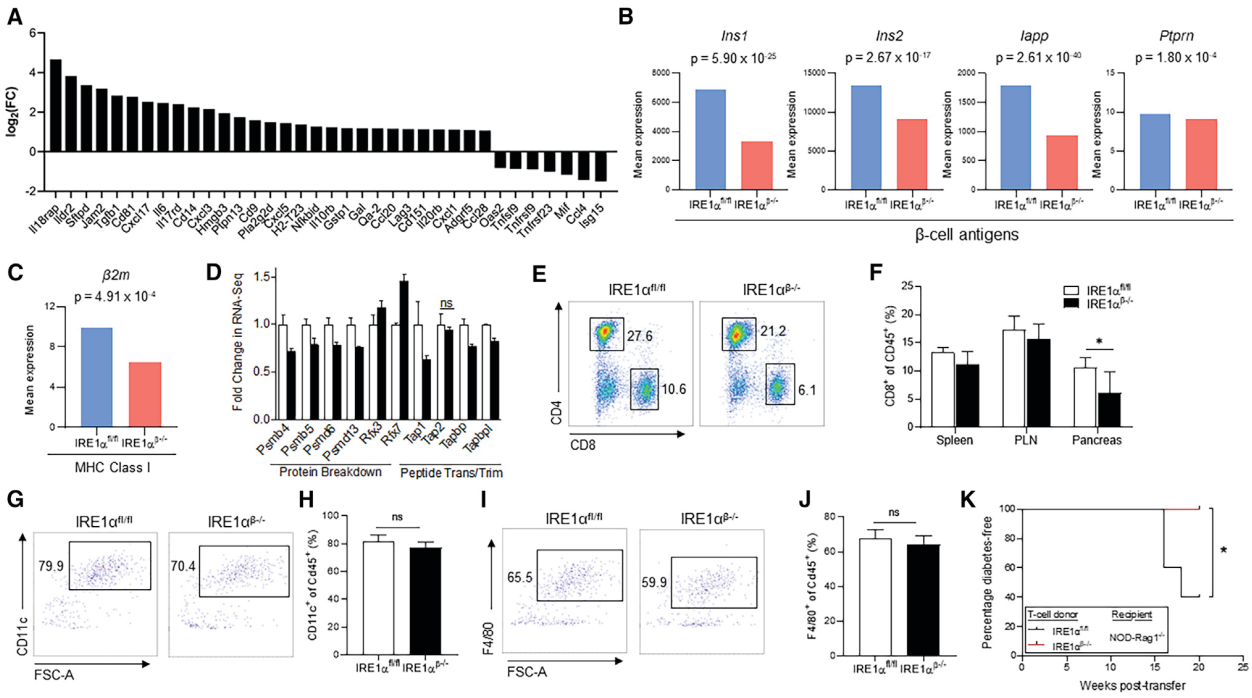


Figure 7. β Cells of IRE1α$^{β–/–}$ Mice Have Altered Expression of Genes Associated with Immune Cell Recruitment

A) The expression of genes that is key in regulation of lymphocyte activation, as well as markers of cytokine, chemokine, and ECM that are significantly altered in β cells of IRE1α$^{β-/-}$ mice compared with IRE1α$^{fl/fl}$ mice (FDR < 0.01).

B) The mRNA expression of β cell autoantigens in β cells of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice. p values are indicated.

C and D) The mRNA expression of MHC class I component β2m (C) and genes that are involved in the MHC class I loading pathway in β cells of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (D) (FDR < 0.05).

E) Fractions of $CD4^+$ and $CD8^+$ T cells in representative dot plots from pancreata of 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice after pre-gating for single, viable, and $CD45^+$ cells.

F) Immunophenotyping data showing percentage of $CD8^+$ T cells in spleen, pancreatic lymph node (PLN), and pancreas from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

G) Fractions of $CD11c^+$ dendritic cells in representative dot plots from pancreata of 5 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 7) mice after pre-gating for single, viable, and $CD45^+$ cells.

H) Quantification of percentage of $CD11c^+$ dendritic cells in pancreata from 5 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 7) mice. Data are represented as mean ± SEM, with statistical analysis performed by Student's t test.

I) Fractions of $F4/80^+$ macrophages in representative dot plots from pancreata of 5 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 7) mice after pre-gating for single, viable, and $CD45^+$ cells. (J) Quantification of percentage of $F4/80^+$ macrophages in pancreata from 5 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 7) mice. Data are represented as mean ± SEM, with statistical analysis performed by Student's t test. (K) Percentage of diabetes-free NOD $Rag^{-/-}$ mice (n = 5 per group) post-total T cell transfer from 8-week-old $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice. The incidence of diabetes was compared by log-rank (Mantel-Cox) test (*$p < 0.05$).

Figure S1

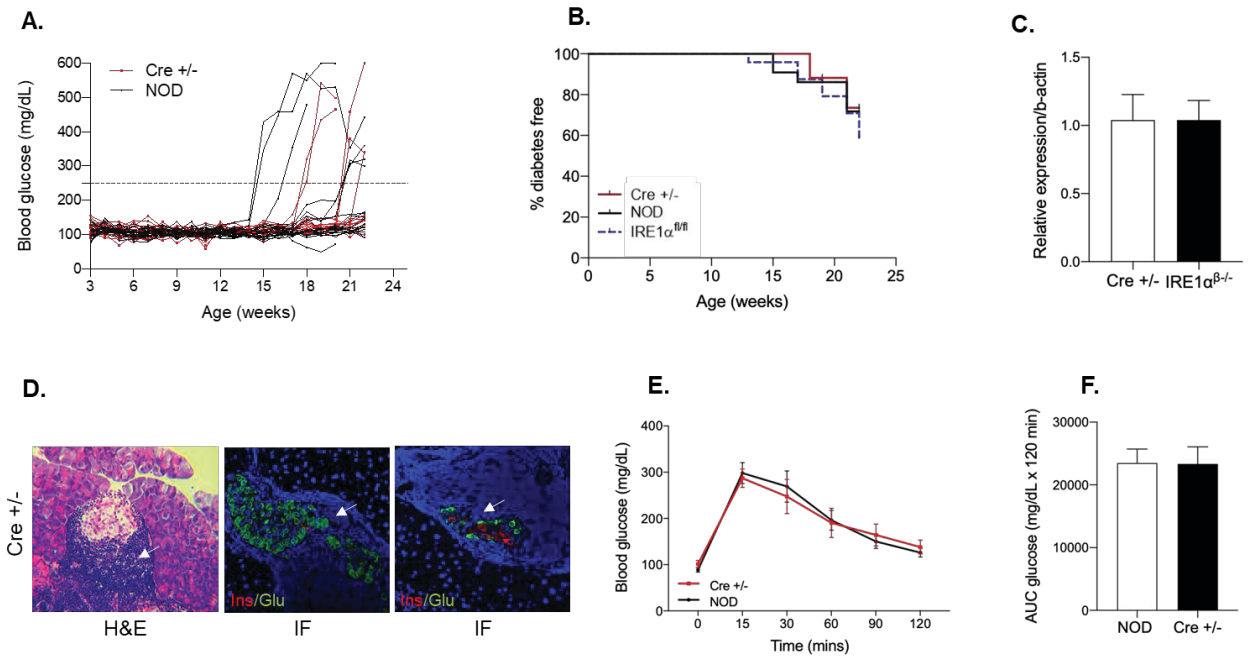


Figure S1 (Related to Figure 1): Characterization of the phenotype of Ins2Cre$^{ERT/+}$ mice

A) Blood glucose levels of female NOD(n = 22) and NOD Ins2Cre$^{ERT/+}$ (n = 17) mice.

B) The percentage of diabetes-free mice in Ins2Cre$^{ERT/+}$ (red line), NOD (black line), and IRE1$^{\alpha fl/fl}$ (blue dashed line). Cohorts plotted using a Kaplan-Meier curve. A log-rank statistical analysis was performed.

C) The expression of Cre transgene in pre-diabetic Ins2Cre$^{ERT/+}$ (n = 3) and IRE1$\alpha^{\beta-/-}$ (n = 5) mice.

D) Representative H&E and immunofluorescence images showing insulin and glucagon expression on pancreatic sections from diabetic Ins2Cre$^{ERT/+}$ mice. Arrows indicate the dense area around islets with nuclei of lymphocytes.

E) Blood glucose levels of pre-diabetic Ins2Cre$^{ERT/+}$ (red line) and NOD (black line) mice at indicated times after administration of glucose (n = 6 per group).

F) Quantification of the glucose area under the curve (AUC) of NOD and Ins2Cre$^{ERT/+}$ mice.

Data are represented as mean ± SEM and were analyzed using either the Student's t-test, or one-way ANOVA where required. $P < 0.05$ was considered statistically significant.

Figure S2

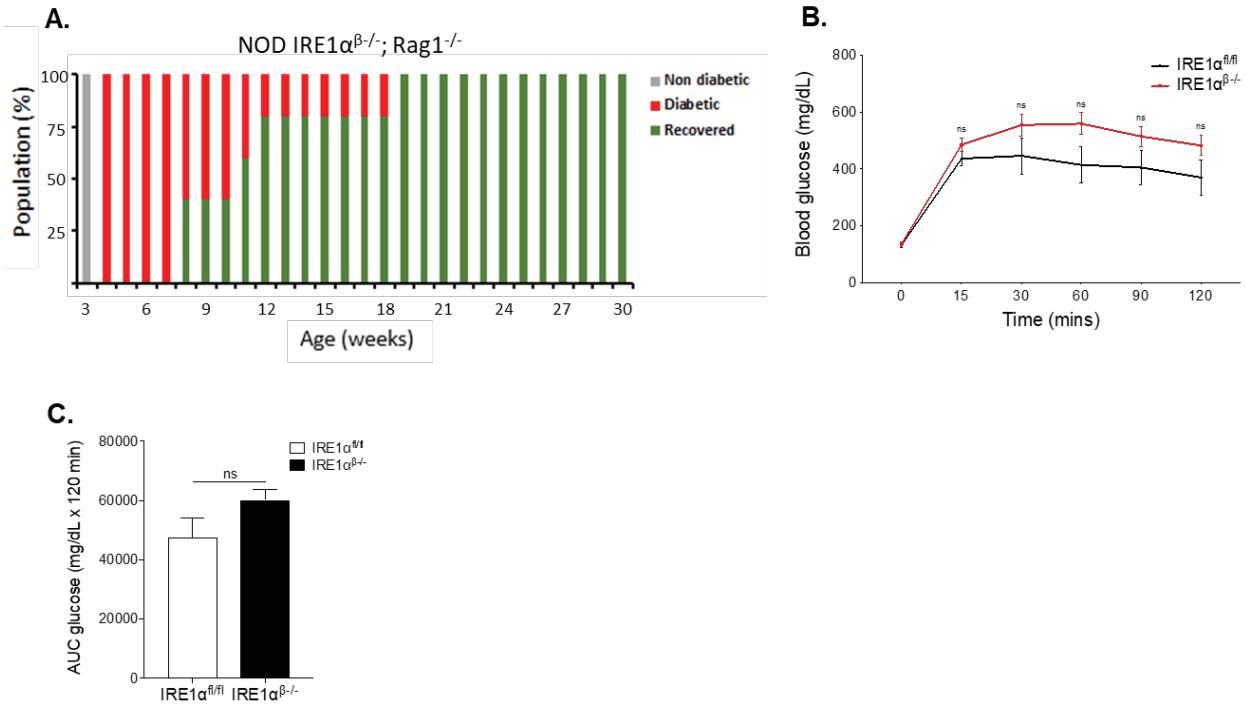


Figure S2 (Related to Figure 2): Characterization of the phenotype of NOD IRE1α$^{β-/-}$; Rag1$^{-/-}$ mice and the analysis of glucose tolerance in IRE1α$^{β-/-}$ mice

A) Blood glucose levels of female NOD IRE1α$^{β-/-}$; Rag1$^{-/-}$mice (n = 5).

B) Blood glucose levels of 32 weeks of age normoglycemic IRE1α$^{fl/fl}$ (n = 5) and IRE1α$^{β-/-}$ (n = 7) mice at indicated times after intraperitoneal injection of glucose.

C) Quantification of the glucose area under the curve (AUC) of IRE1α$^{fl/fl}$ (n = 5) and IRE1α$^{β-/-}$ (n = 7) mice.

Data are represented as mean ± SEM and were analyzed using either the Student's t-test, or one way ANOVA where required. $P < 0.05$ was considered statistically significant.

Figure S3

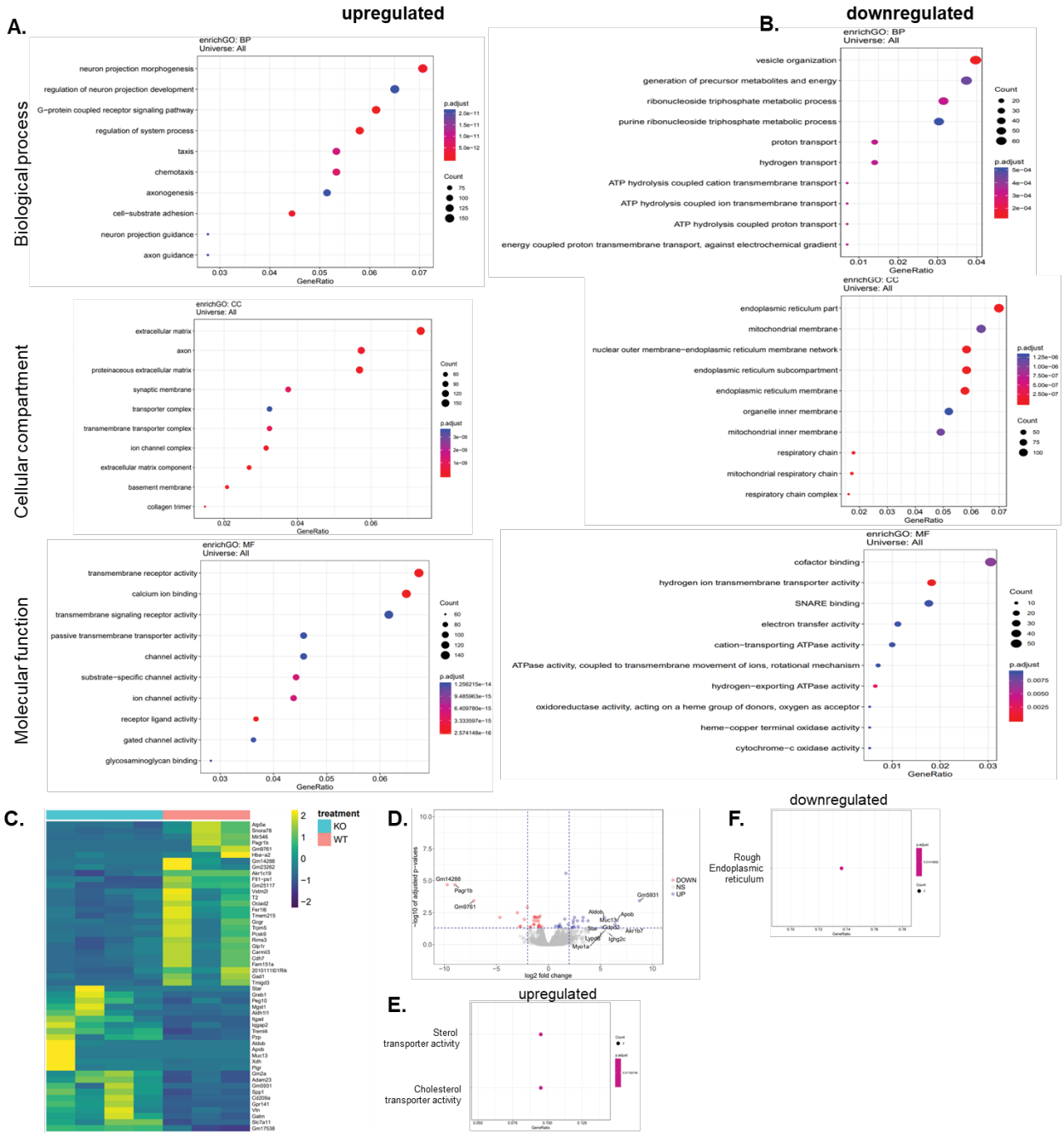

A.
upregulated
Biological process
enrichGO: BP
Universe: All
neuron projection morphogenesis
regulation of neuron projection development
G−protein coupled receptor signaling pathway
regulation of system process
taxis
chemotaxis
axonogenesis
cell−substrate adhesion
neuron projection guidance
axon guidance
GeneRatio
p.adjust
Count
B.
downregulated
enrichGO: BP
Universe: All
vesicle organization
generation of precursor metabolites and energy
ribonucleoside triphosphate metabolic process
purine ribonucleoside triphosphate metabolic process
proton transport
hydrogen transport
ATP hydrolysis coupled cation transmembrane transport
ATP hydrolysis coupled ion transmembrane transport
ATP hydrolysis coupled proton transport
energy coupled proton transmembrane transport, against electrochemical gradient
Cellular compartment
enrichGO: CC
Universe: All
extracellular matrix
axon
proteinaceous extracellular matrix
synaptic membrane
transporter complex
transmembrane transporter complex
ion channel complex
extracellular matrix component
basement membrane
collagen trimer
enrichGO: CC
Universe: All
endoplasmic reticulum part
mitochondrial membrane
nuclear outer membrane−endoplasmic reticulum membrane network
endoplasmic reticulum subcompartment
endoplasmic reticulum membrane
organelle inner membrane
mitochondrial inner membrane
respiratory chain
mitochondrial respiratory chain
respiratory chain complex
Molecular function
enrichGO: MF
Universe: All
transmembrane receptor activity
calcium ion binding
transmembrane signaling receptor activity
passive transmembrane transporter activity
channel activity
substrate−specific channel activity
ion channel activity
receptor ligand activity
gated channel activity
glycosaminoglycan binding
enrichGO: MF
Universe: All
cofactor binding
hydrogen ion transmembrane transporter activity
SNARE binding
electron transfer activity
cation−transporting ATPase activity
ATPase activity, coupled to transmembrane movement of ions, rotational mechanism
hydrogen−exporting ATPase activity
oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor
heme−copper terminal oxidase activity
cytochrome−c oxidase activity
C.
treatment
KO
WT
D.
DOWN
NS
UP
log2 fold change
F.
downregulated
Rough Endoplasmic reticulum
E.
upregulated
Sterol transporter activity
Cholesterol transporter activity

Figure S3 (Related to Figure 4): Bulk RNA-Seq in the islets of IRE1α$^{β-/-}$ mice during the hyperglycemic and recovery phases

A and B) Gene ontology (GO) enrichment analysis of up- and down-regulated genes from the bulk RNA-seq analysis of 7 weeks of age IRE1α$^{fl/fl}$ vs. IRE1α$^{β-/-}$ mouse islets (at FDR of 0.05). GeneRatio in the x-axis quantifies the proportion of the up- or down-regulated genes that are among each GO category and Count depicts the number of up- or down-regulated genes in the GO category.

C) Heatmap of expression levels for differentially expressed genes identified in 15 weeks of age IRE1α$^{β-/-}$ and IRE1α$^{fl/fl}$ normoglycemic mice.

D) Volcano plots showing differentially expressed genes in the islets of 15-week-old IRE1α$^{fl/fl}$ (n = 3) and IRE1α$^{β-/-}$ (n = 4) mice. The horizontal line depicts the FDR cutoff of 0.05 and the vertical lines mark log2 fold changes of -2 and 2. Genes with absolute log2 fold change larger than 5 or adjusted P value smaller than 1e-25 and absolute log2 fold change larger than 2 are labeled with their gene symbols.

E and F) GO enrichment of upregulated genes categorized by molecular function and cellular compartment at FDR of 0.05.

Figure S4

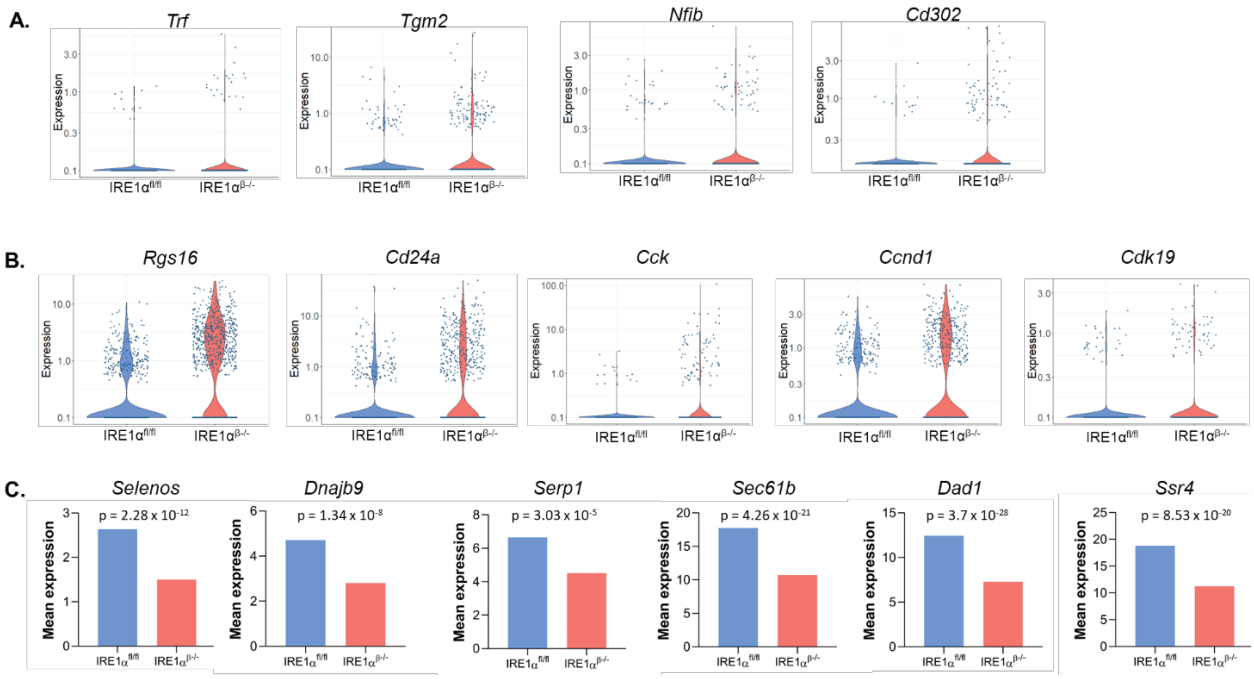


Figure S4 (Related to Figure 6): Single-cell RNA-seq identifies altered expression of disallowed, regenerative, and sXBP1 target genes in IRE1α$^{β-/-}$ mice during hyperglycemic (5 weeks of age) phase

A-C) Expression of (A) disallowed genes, (B) regenerative genes, and (C) sXBP1 target gene expression in β-cell clusters of IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice (FDR < 0.01, FC > 2)

Figure S5

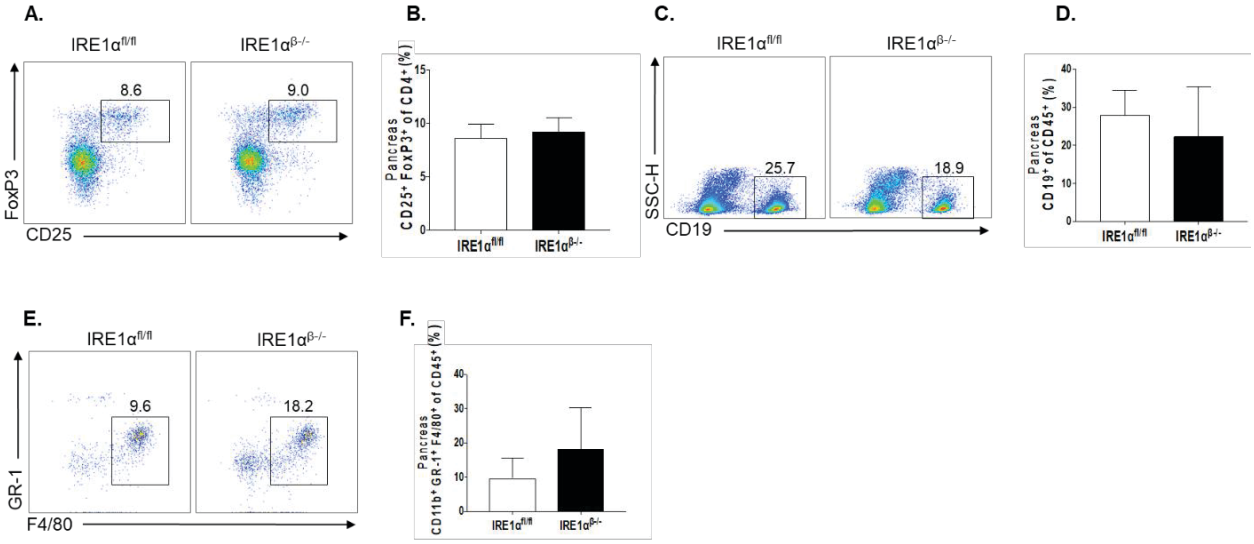


Figure S5 (Related to Figure 7): Immunophenotyping of pancreas

A) Fractions of FoxP3$^+$ CD25$^+$ cells in representative dot plots from pancreata of 21 weeks of age IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice after pre-gating for single, viable, and CD4$^+$ /CD45$^+$ cells.

B) Quantification of percentage of FoxP3$^+$ CD25$^+$ cells in pancreata from 21 weeks of age IRE1α$^{fl/fl}$ (n = 6) and IRE1α$^{β-/-}$ (n = 4) mice.

C) Fractions of CD19$^+$ B cells in representative dot plots from pancreata of 21 weeks of age IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice after pre-gating for single, viable, and CD45$^+$ cells.

D) Quantification of percentage of CD19$^+$ B cells in pancreata from 21 weeks of age IRE1α$^{fl/fl}$ (n = 6) and IRE1α$^{β-/-}$ (n = 4) mice.

E) Fractions of F4/80$^+$ macrophages in representative dot plots from pancreata of 21 weeks of age IRE1α$^{fl/fl}$ and IRE1α$^{β-/-}$ mice after pre-gating for single, viable, and CD45$^+$ cells.

F) Quantification of percentage of F4/80$^+$ macrophages in pancreata from 21 weeks of age IRE1α$^{fl/fl}$ (n = 6) and IRE1α$^{β-/-}$ (n = 4) mice.

Data are represented as mean ± SEM and were analyzed using Student's t-test.

Figure S6

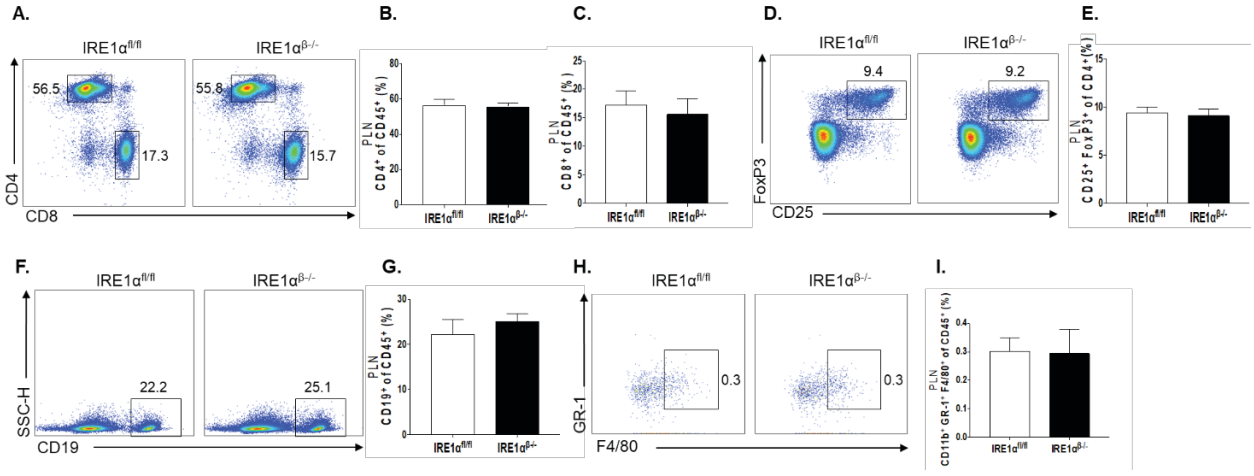


Figure S6 (Related to Figure 7): Immunophenotyping of pancreatic lymph nodes

A) Fractions of $CD4^+$ and $CD8^+$ T-cells in representative dot plots from pancreatic lymph nodes of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD45^+$ cells.

B and C) Quantification of percentage of $CD4^+$ and $CD8^+$ T-cells in pancreatic lymph nodes from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

D) Fractions of $FoxP3^+$ $CD25^+$ cells in representative dot plots from pancreatic lymph nodes of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD4^+$ /$CD45^+$ cells.

E) Quantification of percentage of $FoxP3^+$ $CD25^+$ cells in pancreatic lymph nodes from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

F) Fractions of $CD19^+$ B cells in representative dot plots from pancreatic lymph nodes of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD45^+$ cells.

G) Quantification of percentage of $CD19^+$ B cells in pancreatic lymph nodes from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

H) Fractions of $F4/80^+$ macrophages in representative dot plots from pancreatic lymph nodes of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD45^+$ cells.

I) Quantification of percentage of $F4/80^+$ macrophages in pancreatic lymph nodes from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice. Data are represented as mean ± SEM and were analyzed using Student's t-test.

Figure S7

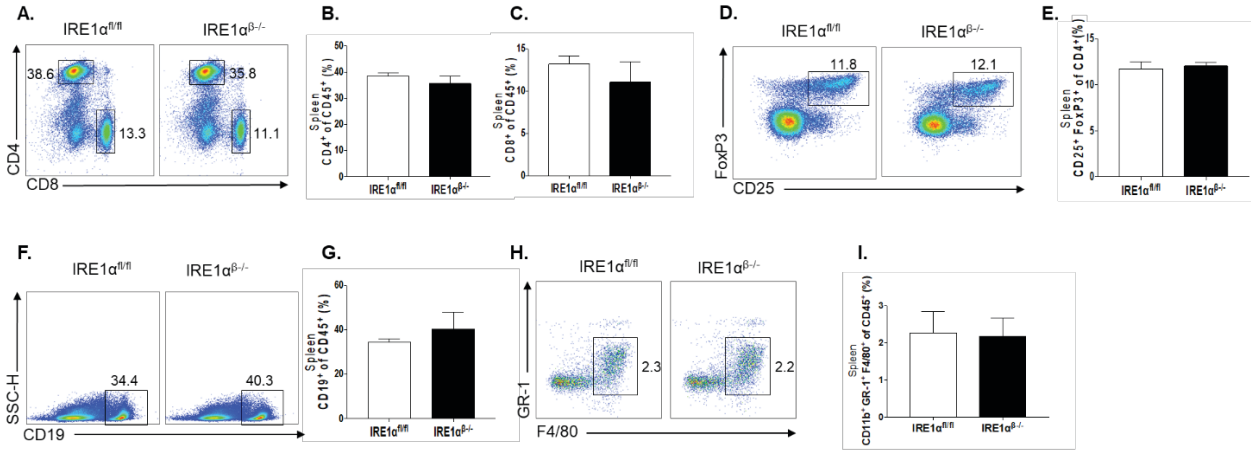


Figure S7 (Related to Figure 7): Immunophenotyping of spleen

A) Fractions of $CD4^+$ and CD8 T-cells in representative dot plots from spleen of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD45^+$ cells.

B and C) Quantification of percentage of $CD4^+$ and $CD8^+$ T-cells in spleen from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

D) Fractions of $FoxP3^+$ $CD25^+$ cells in representative dot plots from spleen of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD4^+$ /$CD45^+$ cells

E) Quantification of percentage of $FoxP3^+$ $CD25^+$ cells in spleen from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

F) Fractions of $CD19^+$ B cells in representative dot plots from spleen of 21 weeks of age $IRE1\alpha^{fl/fl}$ and $IRE1\alpha^{\beta-/-}$ mice after pre-gating for single, viable, and $CD45^+$ cells.

G) Quantification of percentage of $CD19^+$ B cells in spleen from 21 weeks of age $IRE1\alpha^{fl/fl}$ (n = 6) and $IRE1\alpha^{\beta-/-}$ (n = 4) mice.

H) Fractions of F4/80^{+} macrophages in representative dot plots from spleen of 21 weeks of age IRE1α$^{fl/fl}$ and IRE1α$^{\beta-/-}$ mice after pre-gating for single, viable, and CD45^{+} cells.

I) Quantification of percentage of F4/80^{+} macrophages in spleen from 21 weeks of age IRE1α$^{fl/fl}$ (n = 6) and IRE1α$^{\beta-/-}$ (n = 4) mice.

Data are represented as mean ± SEM and were analyzed using Student's t-test.

References

Artner, I., Hang, Y., Mazur, M., Yamamoto, T., Guo, M., Lindner, J., Magnuson, M.A., and Stein, R. (2010). MafA and MafB regulate genes critical to beta-cells in a unique temporal manner. Diabetes 59, 2530–2539.

Arvan, P., Pietropaolo, M., Ostrov, D., and Rhodes, C.J. (2012). Islet autoantigens: structure, function, localization, and regulation. Cold Spring Harb. Perspect. Med. 2, a007658.

Atkinson, M.A. (2012). The pathogenesis and natural history of type 1 diabetes. Cold Spring Harb. Perspect. Med. 2, a007641.

Bernales, S., Papa, F.R., and Walter, P. (2006). Intracellular signaling by the unfolded protein response. Annu. Rev. Cell Dev. Biol. 22, 487–508.

Bersell, K., Arab, S., Haring, B., and K€ uhn, B. (2009). Neuregulin1/ErbB4 signaling induces cardiomyocyte proliferation and repair of heart injury. Cell 138, 257–270.

Bluestone, J.A., Herold, K., and Eisenbarth, G. (2010). Genetics, pathogenesis and clinical interventions in type 1 diabetes. Nature 464, 1293–1300.

Blum, B., Hrvatin, S., Schuetz, C., Bonal, C., Rezania, A., and Melton, D.A. (2012). Functional beta-cell maturation is marked by an increased glucose threshold and by expression of urocortin 3. Nat. Biotechnol. 30, 261–264.

Burzyn, D., Kuswanto, W., Kolodin, D., Shadrach, J.L., Cerletti, M., Jang, Y., Sefik, E., Tan, T.G., Wagers, A.J., Benoist, C., and Mathis, D. (2013). A special population of regulatory T cells potentiates muscle repair. Cell 155, 1282–1295.

Calfon, M., Zeng, H., Urano, F., Till, J.H., Hubbard, S.R., Harding, H.P., Clark, S.G., and Ron, D. (2002). IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. Nature 415, 92–96.

Carosella, E.D., Moreau, P., Lemaoult, J., and Rouas-Freiss, N. (2008). HLA-G: from biology to clinical benefits. Trends Immunol 29, 125–13

Chen, Y., and Brandizzi, F. (2013). IRE1: ER stress sensor and cell fate executor. Trends Cell Biol 23, 547–555.

Chen, Z.L., Yu, W.M., and Strickland, S. (2007). Peripheral regeneration. Annu. Rev. Neurosci. 30, 209–233.

Cinti, F., Bouchi, R., Kim-Muller, J.Y., Ohmura, Y., Sandoval, P.R., Masini, M., Marselli, L., Suleiman, M., Ratner, L.E., Marchetti, P., et al. (2016). Evidence of beta-cell

dedifferentiation in human type 2 diabetes. J. Clin. Endocrinol. Metab. 101, 1044–1054.

de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics 20, 1453–1454.

Dhawan, S., Tschen, S.I., Zeng, C., Guo, T., Hebrok, M., Matveyenko, A., and Bhushan, A. (2015). DNA methylation directs functional maturation of pancreatic beta cells. J. Clin. Invest. 125, 2851–2860.

Dirice, E., Kahraman, S., De Jesus, D.F., El Ouaamari, A., Basile, G., Baker, R.L., Yigit, B., Piehowski, P.D., Kim, M.J., Dwyer, A.J., et al. (2019). Increased b-cell proliferation before immune cell invasion prevents progression of type 1 diabetes. Nat. Metab. 1, 509–518.

Dooley, J., Tian, L., Schonefeldt, S., Delghingaro-Augusto, V., Garcia-Perez, J.E., Pasciuto, E., Di Marino, D., Carr, E.J., Oskolkov, N., Lyssenko, V., et al. (2016). Genetic predisposition for beta cell fragility underlies type 1 and type 2 diabetes. Nat. Genet. 48, 519–527.

Dor, Y., Brown, J., Martinez, O.I., and Melton, D.A. (2004). Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. Nature 429, 41–46.

Engin, F. (2016). ER stress and development of type 1 diabetes. J. Investig. Med. 64, 2–6.

Engin, F., Yermalovich, A., Nguyen, T., Hummasti, S., Fu, W., Eizirik, D.L., Mathis, D., and Hotamisligil, G.S. (2013). Restoration of the unfolded protein response in pancreatic beta cells protects mice against type 1 diabetes. Sci. Transl. Med. 5, 211ra156.

Engin, F., Nguyen, T., Yermalovich, A., and Hotamisligil, G.S. (2014). Aberrant islet unfolded protein response in type 2 diabetes. Sci. Rep. 4, 4054.

Gershengorn, M.C., Hardikar, A.A., Wei, C., Geras-Raaka, E., MarcusSamuels, B., and Raaka, B.M. (2004). Epithelial-to-mesenchymal transition generates proliferative human islet precursor cells. Science 306, 2261–2264.

Gittes, G.K., Rutter, W.J., and Debas, H.T. (1993). Initiation of gastrin expression during the development of the mouse pancreas. Am. J. Surg. 165, 23–25.

Granados, D.P., Tanguay, P.L., Hardy, M.P., Caron, E., de Verteuil, D., Meloche, S., and Perreault, C. (2009). ER stress affects processing of MHC class I-associated peptides. BMC Immunol 10,10.

Han, D., Lerner, A.G., Vande Walle, L., Upton, J.P., Xu, W., Hagen, A., Backes, B.J., Oakes, S.A., and Papa, F.R. (2009). IRE1alpha kinase activation modes control alternate endoribonuclease outputs to determine divergent cell fates. Cell 138, 562–575.

Hassler, J.R., Scheuner, D.L., Wang, S., Han, J., Kodali, V.K., Li, P., Nguyen, J., George, J.S., Davis, C., Wu, S.P., et al. (2015). The IRE1alpha/XBP1s pathway is essential for the glucose response and protection of b cells. PLoS Biol 13, e1002277.

Hollien, J., and Weissman, J.S. (2006). Decay of endoplasmic reticulum-localized mRNAs during the unfolded protein response. Science 313, 104–107.

In't Veld, P., Lievens, D., De Grijse, J., Ling, Z., Van der Auwera, B., PipeleersMarichal, M., Gorus, F., and Pipeleers, D. (2007). Screening for insulitis in adult autoantibody-positive organ donors. Diabetes 56, 2400–2404.

Iwawaki, T., Akai, R., Yamanaka, S., and Kohno, K. (2009). Function of IRE1 alpha in the placenta is essential for placental development and embryonic viability. Proc. Natl. Acad. Sci. USA 106, 16657–16662.

Jiang, H., Ware, R., Stall, A., Flaherty, L., Chess, L., and Pernis, B. (1995). Murine CD8+ T cells that specifically delete autologous CD4+ T cells expressing V beta 8 TCR: a role of the Qa-1 molecule. Immunity 2, 185–194.

Jones, E.L., Demaria, M.C., and Wright, M.D. (2011). Tetraspanins in cellular immunity. Biochem. Soc. Trans. 39, 506–511.

Kim-Muller, J.Y., Fan, J., Kim, Y.J., Lee, S.A., Ishida, E., Blaner, W.S., and Accili, D. (2016). Aldehyde dehydrogenase 1a3 defines a subset of failing pancreatic beta cells in diabetic mice. Nat. Commun. 7, 12631.

Klein, J., Figueroa, F., and David, C.S. (1983). H-2 haplotypes, genes and antigens: second listing. II. The H-2 complex. Immunogenetics 17, 553–596.

Lee, A.H., Iwakoshi, N.N., and Glimcher, L.H. (2003). XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response. Mol. Cell. Biol. 23, 7448–7459.

Lee, A.H., Heidtman, K., Hotamisligil, G.S., and Glimcher, L.H. (2011). Dual and opposing roles of the unfolded protein response regulated by IRE1alpha and XBP1 in proinsulin processing and insulin secretion. Proc. Natl. Acad. Sci. USA 108, 8885–8890.

Lerner, A.G., Upton, J.P., Praveen, P.V., Ghosh, R., Nakagawa, Y., Igbaria, A., Shen, S., Nguyen, V., Backes, B.J., Heiman, M., et al. (2012). IRE1alpha induces

thioredoxin-interacting protein to activate the NLRP3 inflammasome and promote programmed cell death under irremediable ER stress. Cell Metab 16, 250–264.

Lipson, K.L., Ghosh, R., and Urano, F. (2008). The role of IRE1alpha in the degradation of insulin mRNA in pancreatic beta-cells. PLoS One 3, e1648.

Louvet, C., Szot, G.L., Lang, J., Lee, M.R., Martinier, N., Bollag, G., Zhu, S., Weiss, A., and Bluestone, J.A. (2008). Tyrosine kinase inhibitors reverse type 1 diabetes in nonobese diabetic mice. Proc. Natl. Acad. Sci. USA 105, 18895–18900.

Maganti, A., Evans-Molina, C., and Mirmira, R. (2014). From immunobiology to beta-cell biology: the changing perspective on type 1 diabetes. Islets 6, e28778.

Marhfour, I., Lopez, X.M., Lefkaditis, D., Salmon, I., Allagnat, F., Richardson, S.J., Morgan, N.G., and Eizirik, D.L. (2012). Expression of endoplasmic reticulum stress markers in the islets of patients with type 1 diabetes. Diabetologia 55, 2417–2420.

Matsuoka, T.A., Artner, I., Henderson, E., Means, A., Sander, M., and Stein, R. (2004). The MafA transcription factor appears to be responsible for tissue-specific expression of insulin. Proc. Natl. Acad. Sci. USA 101, 2930–2933.

McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res 40, 4288–4297.

Monticelli, L.A., Sonnenberg, G.F., Abt, M.C., Alenghat, T., Ziegler, C.G., Doering, T.A., Angelosanto, J.M., Laidlaw, B.J., Yang, C.Y., Sathaliyawala, T., et al. (2011). Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. Nat. Immunol. 12, 1045–1054.

Morita, S., Villalta, S.A., Feldman, H.C., Register, A.C., Rosenthal, W., Hoffmann-Petersen, I.T., Mehdizadeh, M., Ghosh, R., Wang, L., ColonNegron, K., et al. (2017). Targeting ABL-IRE1a signaling spares ER-stressed pancreatic b cells to reverse autoimmune diabetes. Cell Metab 25, 883–897.e8.

Nakayama, M. (2011). Insulin as a key autoantigen in the development of type 1 diabetes. Diabetes Metab. Res. Rev. 27, 773–777.

Nakayama, M., Abiru, N., Moriyama, H., Babaya, N., Liu, E., Miao, D., Yu, L., Wegmann, D.R., Hutton, J.C., Elliott, J.F., and Eisenbarth, G.S. (2005). Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice. Nature 435, 220–223.

Narendran, P., Mannering, S.I., and Harrison, L.C. (2003). Proinsulin-a pathogenic autoantigen in type 1 diabetes. Autoimmun. Rev. 2, 204–210.

Oslowski, C.M., Hara, T., O'Sullivan-Murphy, B., Kanekura, K., Lu, S., Hara, M., Ishigaki, S., Zhu, L.J., Hayashi, E., Hui, S.T., et al. (2012). Thioredoxin-interacting protein mediates ER stress-induced beta cell death through initiation of the inflammasome. Cell Metab 16, 265–273.

Ouziel-Yahalom, L., Zalzman, M., Anker-Kitai, L., Knoller, S., Bar, Y., Glandt, M., Herold, K., and Efrat, S. (2006). Expansion and redifferentiation of adult human pancreatic islet cells. Biochem. Biophys. Res. Commun. 341, 291–298.

Petri, A., Ahnfelt-Rønne, J., Frederiksen, K.S., Edwards, D.G., Madsen, D., Serup, P., Fleckner, J., and Heller, R.S. (2006). The effect of neurogenin3 deficiency on pancreatic gene expression in embryonic mice. J. Mol. Endocrinol. 37, 301–316.

Pugliese, A. (2016). Insulitis in the pathogenesis of type 1 diabetes. Pediatr. Diabetes 17, 31–36.

Pullen, T.J., Khan, A.M., Barton, G., Butcher, S.A., Sun, G., and Rutter, G.A. (2010). Identification of genes selectively disallowed in the pancreatic islet. Islets 2, 89–95.

Puri, S., Folias, A.E., and Hebrok, M. (2015). Plasticity and dedifferentiation within the pancreas: development, homeostasis, and disease. Cell Stem Cell 16, 18–31.

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

Quintens, R., Hendrickx, N., Lemaire, K., and Schuit, F. (2008). Why expression of some genes is disallowed in beta-cells. Biochem. Soc. Trans. 36, 300–305.

Regnell, S.E., and Lernmark, A˚. (2017). Early prediction of autoimmune (type 1) diabetes. Diabetologia 60, 1370–1381.

Robinson, P.J., Millrain, M., Antoniou, J., Simpson, E., and Mellor, A.L. (1989). A glycophospholipid anchor is required for Qa-2-mediated T cell activation. Nature 342, 85–87.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Rui, J., Deng, S., Arazi, A., Perdigoto, A.L., Liu, Z., and Herold, K.C. (2017). b cells that resist immunological attack develop during progression of autoimmune diabetes in NOD mice. Cell Metab 25, 727–738.

Saldanha, A.J. (2004). Java Treeview–extensible visualization of microarray data. Bioinformatics 20, 3246–3248.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9, 671–675.

Shao, J., and Sheng, H. (2010). Amphiregulin promotes intestinal epithelial regeneration: roles of intestinal subepithelial myofibroblasts. Endocrinology 151, 3728–3737.

Shoshani, O., and Zipori, D. (2011). Mammalian cell dedifferentiation as a possible outcome of stress. Stem Cell Rev Rep 7, 488–493.

Soleimanpour, S.A., and Stoffers, D.A. (2013). The pancreatic beta cell and type 1 diabetes: innocent bystander or active participant? Trends Endocrinol. Metab. 24, 324–331.

Su, Y., Jono, H., Misumi, Y., Senokuchi, T., Guo, J., Ueda, M., Shinriki, S., Tasaki, M., Shono, M., Obayashi, K., et al. (2012). Novel function of transthyretin in pancreatic alpha cells. FEBS Lett 586, 4215–4222.

Szabat, M., Pourghaderi, P., Soukhatcheva, G., Verchere, C.B., Warnock, G.L., Piret, J.M., and Johnson, J.D. (2011). Kinetics and genomic profiling of adult human and mouse beta-cell maturation. Islets 3, 175–187.

Szabat, M., Page, M.M., Panzhinskiy, E., Skovsø, S., Mojibian, M., FernandezTajes, J., Bruin, J.E., Bround, M.J., Lee, J.T., Xu, E.E., et al. (2016). Reduced insulin production relieves endoplasmic reticulum stress and induces b cell proliferation. Cell Metab 23, 179–193.

Talchai, C., Xuan, S., Lin, H.V., Sussel, L., and Accili, D. (2012). Pancreatic beta cell dedifferentiation as a mechanism of diabetic beta cell failure. Cell 150, 1223–1234.

Tersey, S.A., Nishiki, Y., Templin, A.T., Cabrera, S.M., Stull, N.D., Colvin, S.C., Evans-Molina, C., Rickus, J.L., Maier, B., and Mirmira, R.G. (2012). Islet betacell endoplasmic reticulum stress precedes the onset of type 1 diabetes in the nonobese diabetic mouse model. Diabetes 61, 818–827.

Thompson, P.J., Shah, A., Ntranos, V., Van Gool, F., Atkinson, M., and Bhushan, A. (2019). Targeted elimination of senescent b cells prevents type 1 diabetes. Cell Metab 29, 1045–1060.e10.

Thorel, F., Ne´ pote, V., Avril, I., Kohno, K., Desgraz, R., Chera, S., and Herrera, P.L. (2010). Conversion of adult pancreatic alpha-cells to beta-cells after extreme beta-cell loss. Nature 464, 1149–1154.

Thorrez, L., Laudadio, I., Van Deun, K., Quintens, R., Hendrickx, N., Granvik, M., Lemaire, K., Schraenen, A., Van Lommel, L., Lehnert, S., et al. (2011). Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. Genome Res 21, 95–105.

Tsuchiya, Y., Saito, M., Kadokura, H., Miyazaki, J.I., Tashiro, F., Imagawa, Y., Iwawaki, T., and Kohno, K. (2018). IRE1-XBP1 pathway regulates oxidative proinsulin folding in pancreatic beta cells. J. Cell Biol. 217, 1287–1301.

Ulianich, L., Terrazzano, G., Annunziatella, M., Ruggiero, G., Beguinot, F., and Di Jeso, B. (2011). ER stress impairs MHC class I surface expression and increases susceptibility of thyroid cells to NK-mediated cytotoxicity. Biochim. Biophys. Acta 1812, 431–438.

Upton, J.P., Wang, L., Han, D., Wang, E.S., Huskey, N.E., Lim, L., Truitt, M., McManus, M.T., Ruggero, D., Goga, A., et al. (2012). IRE1alpha cleaves select microRNAs during ER stress to derepress translation of proapoptotic caspase-2. Science 338, 818–822.

Urano, F., Wang, X., Bertolotti, A., Zhang, Y., Chung, P., Harding, H.P., and Ron, D. (2000). Coupling of stress in the ER to activation of JNK protein kinases by transmembrane protein kinase IRE1. Science 287, 664–666.

van Belle, T.L., Coppieters, K.T., and von Herrath, M.G. (2011). Type 1 diabetes: etiology, immunology, and therapeutic strategies. Physiol. Rev. 91, 79–118.

van der Meulen, T., Xie, R., Kelly, O.G., Vale, W.W., Sander, M., and Huising, M.O. (2012). Urocortin 3 marks mature human primary and embryonic stem cell-derived pancreatic alpha and beta cells. PLoS One 7, e52181.

Walter, P., and Ron, D. (2011). The unfolded protein response: from stress pathway to homeostatic regulation. Science 334, 1081–1086.

Wang, Z., York, N.W., Nichols, C.G., and Remedi, M.S. (2014). Pancreatic beta cell dedifferentiation in diabetes and redifferentiation following insulin therapy. Cell Metab 19, 872–882.

Ward-Kavanagh, L.K., Lin, W.W., Šedy , J.R., and Ware, C.F. (2016). The TNF receptor superfamily in co-stimulating and co-inhibitory responses. Immunity 44, 1005–1019.

Wiberg, A., Granstam, A., Ingvast, S., Härkönen, T., Knip, M., Korsgren, O., and Skog, O. (2015). Characterization of human organ donors testing positive for type 1 diabetes-associated autoantibodies. Clin. Exp. Immunol. 182, 278–288.

Yoshida, H., Matsui, T., Yamamoto, A., Okada, T., and Mori, K. (2001). XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. Cell 107, 881–891.

You, S., and Chatenoud, L. (2006). Proinsulin: a unique autoantigen triggering autoimmune diabetes. J. Clin. Invest. 116, 3108–3110.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16, 284–287.

Zhang, W., Feng, D., Li, Y., Iida, K., McGrath, B., and Cavener, D.R. (2006). PERK EIF2AK3 control of pancreatic beta cell differentiation and proliferation is required for postnatal glucose homeostasis. Cell Metab 4, 491–497.

Zhang, J., McKenna, L.B., Bogue, C.W., and Kaestner, K.H. (2014). The diabetes gene Hhex maintains delta-cell differentiation and islet function. Genes Dev 28, 829–834.

Appendix 2

Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets

The work presented in this appendix is published in Nature Communications:

Zhang S, Pyne S, Pietrzak S, Halberg S, McCalla SG, Siahpirani AF, Sridharan R, and Roy S. (2023). Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature Communications*. 14(3064).

Contributions: S.Z. and S.R. designed the scMTNI algorithm and experiments. S.Z. implemented the code and performed most of the experiments. S.Py. contributed towards creation of the gold standards and evaluating selected algorithms. S.G.M. and S.H. contributed toward evaluation of algorithms on the fetal hematopoiesis dataset. S.Pi. and R.S. generated the scATAC-seq data for the reprogramming experiments. A.F.S. contributed towards processing the scATAC-seq data for the reprogramming experiments and sequence-specific motifs from the Cis-BP database and assisted with collection of gold standards from the hESC and mESC cell lines. All authors contributed toward writing the manuscript.

Abstract

Cell type-specific gene expression patterns are outputs of transcriptional gene regulatory networks (GRNs) that connect transcription factors and signaling proteins to target genes. Single-cell technologies such as single cell RNA-sequencing (scRNA-seq) and single cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq), can examine cell-type specific gene regulation at unprecedented detail. However, current approaches to infer cell type-specific GRNs are limited in their ability to integrate scRNA-seq and scATAC-seq measurements and to model network dynamics on a cell lineage. To address this challenge, we have developed single-cell Multi-Task Network Inference (scMTNI), a multi-task learning framework to infer the GRN for each cell type on a lineage from scRNA-seq and scATAC-seq data. Using simulated and real datasets, we show that scMTNI is a broadly applicable framework for linear and branching lineages that accurately infers GRN dynamics and identifies key regulators of fate transitions for diverse processes such as cellular reprogramming and differentiation.

Introduction

Transcriptional gene regulatory networks (GRNs) specify connections between regulatory proteins and target genes and determine the spatial and temporal expression patterns of genes[1,2]. These networks reconfigure during dynamic processes such as development or disease progression, to specify cell type specific expression levels. Recent advances in single cell omic techniques such as single cell RNA-sequencing (scRNA-seq) and single cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq)[3] enable collecting high resolution molecular phenotypes of a developing system and offer unprecedented opportunities for the discovery of cell type-specific regulatory networks and their dynamics. However, computational methods to systematically leverage these datasets to identify regulatory networks driving cell type-specific expression patterns are limited.

Existing methods of network inference from single cell omic data[4–16] have primarily used transcriptomic measurements and have low recovery of experimentally verified interactions[17,18].Recentlya small number of methods have attempted to integrate scRNA-seq and scATAC-seq datasets[19–21] to examine gene regulation, however, many of these methods focus on defining cell clusters and the network is defined entirely based on accessible sequence-specific motif matches. This restricts the class of regulators that can be incorporated into the regulatory network to those with known motifs. Furthermore, existing methods infer a single GRN for the entire dataset or do not model the cell population structure which is important to discern dynamics and transitions in the inferred networks for cell type-specificity. To overcome the limitations of existing methods, we have developed single-cell Multi-Task Network Inference

(scMTNI), a multi-task learning framework that integrates the cell lineage structure, scRNA-seq and scATAC-seq measurements to enable joint inference of cell type-specific GRNs. scMTNI takes as input a cell lineage tree, scRNA-seq data and scATAC-seq based prior networks for each cell type. scMTNI uses a probabilistic prior to incorporate the lineage structure during network inference and outputs GRNs for each cell type on a cell lineage. We performed a comprehensive benchmarking study of multi-task learning approaches including scMTNI on simulated data and show that incorporation of multi-task learning and tree structure is beneficial for GRN inference.

We applied scMTNI to a previously unpublished scRNA-seq and scATAC-seq time course dataset for cellular reprogramming in mouse and two published scRNA-seq and scATAC-seq cell-type specific datasets for human hematopoietic differentiation. We demonstrate the advantage of scMTNI's framework to integrate scATAC-seq and scRNA-seq datasets for inferring cell type specific GRNs on linear and branching lineage topologies. We examined how the inferred networks change along the trajectory and identified regulators and network components specific to different parts of the lineage tree. Our predictions include known as well as previously uncharacterized regulators of cell populations transitioning to different lineage paths, providing insight into regulatory mechanisms associated with reprogramming efficiency and hematopoietic specification.

Results

Single-cell Multi-Task learning Network Inference (scMTNI) for defining regulatory networks on cell lineages

We developed scMTNI, a multi-task graph learning framework for inferring cell type-specific gene regulatory networks from scRNA-seq and scATAC-seq datasets (Fig. 1a), where a cell type is defined by a cluster of cells with a distinct transcriptional, and, if available, accessibility profile. scMTNI models a GRN as a Dependency network[22],a probabilistic graphical model with random variables representing genes and regulators, such as transcription factors (TFs) and signaling proteins.

scMTNI takes as input cell clusters with gene expression and accessibility profiles and a lineage structure linking the cell clusters (Fig. 1). Such inputs can be obtained from existing methods for integrative clustering[23] and lineage construction[24]. scMTNI uses the scATAC-seq data for each cell cluster to define cell type-specific sequence motif-based TF-target interactions (e.g., a motif for a particular TF, which is accessible only in specific cell types will result in a TFtarget interaction only in those cell types) which are used as a prior to guide network inference (Methods). scMTNI can also take bulk ATACseq data for corresponding cell types to generate cell type-specific prior networks or cell type-agnostic priors derived from sequence-specific motifs that in turn could be filtered with relevant ATAC-seq data. scMTNI's multi-task learning framework incorporates a probabilistic lineage tree prior, which uses the lineage tree structure to influence the similarity of gene regulatory networks on the lineage. This lineage tree prior models the change of a GRN from a start state (e.g., progenitor cell state) to an end state (e.g., more differentiated state) as a series of individual edge-level probabilistic transitions. The output of scMTNI is a set of cell type-specific GRNs one for each cell cluster in the lineage tree. scMTNI is able to incorporate both linear lineage and tree-based lineage structure. scMTNI takes known cell lineage tree structure or

computationally inferred cell lineage using, for example, a minimum spanning tree (MST[24]) approach on scRNA-seq data. While scMTNI was developed to incorporate both scRNA-seq and scATAC-seq data, it can be applied to situations where scATAC-seq, and therefore a cell type-specific prior network, is not available. We refer to the versions of our approach as scMTNI+Prior and scMTNI depending upon whether it uses prior knowledge or not. The output networks of scMTNI are analyzed using two dynamic network analysis methods: edge-based k-means clustering and topic models (Fig. 1b). These approaches identify key regulators and subnetworks associated with a particular cell cluster or a set of cell clusters on a branch.

Multi-task learning algorithms outperform single-task algorithms for single cell network inference

To evaluate scMTNI and other existing algorithms with known ground truth networks on single-cell transcriptomic data, we set up a simulation framework, which entailed creation of a cell lineage, generating synthetic networks and corresponding single-cell expression datasets for each cell type on the lineage (Fig. 2a). We used a probabilistic process of network structure evolution to generate the network structure for three cell types, each containing 15 regulators and 65 genes and between 202–239 edges (Methods). Next, we applied BoolODE17 to simulate the in silico single-cell expression data using each cell type's generated network. To mimic the sparsity in single-cell expression data, we set 80% of the values to 0. We created three datasets with different numbers of cells: 2000, 1000, and 200, referred here as datasets1, 2, and 3. We asked whether multi-task learning is beneficial compared to single-task learning

for network inference from scRNA-seq data. To this end, we compared scMTNI and four other multi-task learning algorithms, MRTLE[25], GNAT[26], Ontogenet[27], and AMuSR[28] to three single-task algorithms, LASSO regression[29], INDEP, and SCENIC[30] (Methods). Of these methods, only SCENIC uses a non-linear regression model while the others are based on linear models. INDEP is similar to scMTNI but does not incorporate the lineage prior. Each algorithm was applied within a stability selection framework and evaluated with Area under the Precision recall curve (AUPR) and F-score of top k edges, where k is the number of edges in the true network (Fig. 2b, c). On dataset 1, based on AUPR, scMTNI, MRTLE, and AMuSR are able to recover the network structure better than the other multi-task learning and single-task learning algorithms (Fig. 2b). Ontogenet performs better than the single-task learning algorithms in at least two cell types. Finally, GNAT performs comparably to the single-task learning algorithms. When comparing algorithms based on F-score of top k edges, we have similar observations that scMTNI and MRTLE have a better performance than other algorithms (Fig. 2c). Ontogenet performs better than LASSO and INDEP in at least two cell types, and comparable to SCENIC, except that Ontogenet in cell type 3 is worse than SCENIC. GNAT is comparable to the single-task learning algorithms for at least 2 of the cell types. The low F-score of AMuSR is because the inferred networks are too sparse, with fewer than 100 edges, while the other algorithms inferred similar number of edges as the true networks. These results remain consistent for datasets 2 and 3 which have fewer cells (1000 and 200, respectively); scMTNI and MRTLE remain superior in performance than other algorithms measured by both AUPR and F-score (Fig. 2b, c). We expect scMTNI to be better since the network simulation procedure is similar, but

the data generation process is different and independent from scMTNI's model. Finally, we aggregated the results across all three cell types and datasets to obtain an overall comparison of the algorithms. Here we considered algorithms across all parameter settings tested as well as the best parameter setting determined by the best F-score or AUPR. Based on the AUPR of "all parameter setting", we found that multi-task learning methods, especially scMTNI and MRTLE are generally better than single-task learning methods with higher AUPRs (Supplementary Fig. 1A, C). AMuSR also outperformed the single-task algorithms based on AUPRs, although this was not as significant as MRTLE and scMTNI. When considering the "best parameter setting", the methods were not significantly different when using AUPR, though MRTLE and scMTNI had the highest AUPR (Supplementary Fig. 1B, D). When using the Fscore, scMTNI and MRTLE remained top performing algorithms for the "all parameter setting" (Supplementary Fig. 2A, C) and the "best parameter setting" (Supplementary Fig. 2B, D). Further, GNAT and Ontogenet had a higher F-score than the single-task learning method LASSO for the "all parameter" and "best parameter" settings, respectively. AMuSR suffered on the F-score metric due to the high sparsity in the inferred networks. Across different single-task algorithms, LASSO had the worst performance. Overall, the results on the simulated networks suggest that multi-task learning algorithms have a better performance than single-task algorithms for network inference on sparse datasets such as single-cell transcriptomic data. Furthermore, scMTNI and MRTLE are able to more accurately infer networks than other multi-task learning algorithms.

Inference of gene regulatory networks of somatic cell reprogramming to induced pluripotent stem cells

Cellular reprogramming is the process of converting cells in a differentiated state to a pluripotent state and is important in regenerative medicine as well as for generating patient-specific disease models. However, this process is inefficient as a small fraction of cells get reprogrammed to the pluripotent state[31]. To gain insight into gene regulatory networks that govern the dynamics of this process, we profiled single cell accessibility (scATAC-seq) during reprogramming of mouse embryonic fibroblasts (MEFs) to the induced pluripotent state and four intermediate timepoints, day 3, day 6, day 9, and day 12, to constitute a dataset of 6 timepoints. We used LIGER to integrate the scRNA-seq and scATAC-seq datasets (Fig. 3a, b) and identified 8 clusters (Methods). Of these clusters, C4 is MEF-specific while C5 is ESC-specific (Fig.3c, d) and showed good integration of the scRNA-seq and scATAC-seq profiles (Supplementary Fig. 3). We removed C6 as it did not have scRNA-seq cells and applied a minimum spanning tree (MST[24]) approach to construct the cell lineage tree from the 7 cell clusters with both scRNA-seq and scATAC-seq (Methods, Fig. 3e). The MEF-specific cluster (C4) is at one end of the tree, while the ESC-specific cluster (C5) is at the other end. This is consistent with the starting and end state of the reprogramming process and we considered C4 to represent the root of the tree. The other clusters represented a mix of cells from different time points, which is consistent with the level of heterogeneity of the reprogramming system[32]. We further verified the identity of these intermediate clusters with a Monocle based trajectory analysis[33] which shows that C7, C2, and C3 represent

cells that might exit the trajectory towards reprogramming and C8 represents cells upstream of this point (Supplementary Fig. 4).

We applied scMTNI, scMTNI+Prior (scMTNI with prior network), INDEP, INDEP+Prior (INDEP with prior network), SCENIC and additionally CellOracle to this dataset (Fig. 3f). We included CellOracle as it combines scRNA-seq and scATAC-seq data, by using accessibility to restrict the set of edges selected based on expression. We used the matched scATAC-seq clusters to obtain TF-target prior interactions for each scRNA-seq cluster needed for INDEP+Prior, scMTNI+Prior and CellOracle (Methods). We assessed the quality of the inferred networks by comparing to multiple gold standard datasets in mouse embryonic stem cells (mESCs, Table 1): one derived from ChIP-seq experiments ("ChIP") from ESCAPE or ENCODE databases[34,35],one from regulator perturbation experiments ("Perturb")[34,36], and the third from the intersection of edges in ChIP and Perturb ("ChIP + Perturb"). We first compared the performance of the methods using F-score on the top 500, 1k, and 2k edges across methods (Fig. 3f, Supplementary Figs. 5, 6). On Perturb, CellOracle and scMTNI+Prior had the best performance, beating other algorithms significantly. On ChIP, SCENIC and CellOracle were the best performing methods. Finally, on Perturb + ChIP, CellOracle and scMTNI+Prior had the best performance. Although CellOracle had high F-scores, its inferred GRNs included a substantially smaller number of regulators (7–11) compared to SCENIC or scMTNI + Prior (29–36). In addition to F-score, we also considered the number of predictable TFs as an additional metric (Supplementary Fig. 7, Methods). This is defined as the number of individual TFs whose targets had a significant overlap with the gold standard. Higher the number of predictable TFs, the

better is a method. On ChIP, scMTNI + Prior had the highest average number of predictable TFs. scMTNI had the highest number of predictable TFs for the Perturb, Perturb + ChIP datasets followed closely by scMTNI + Prior. Overall, scMTNI+Prior had among the highest F-scores, high number of predictable TFs and a greater coverage of the gold standards compared to competing methods using expression alone (SCENIC) as well as those that either incorporated accessibility information (CellOracle, INDEP + Prior) or cell lineage information (scMTNI).

To perform an initial assessment of the network dynamics on the cell lineage, we computed F-score between each pair of inferred networks defined by the top 4k edges (Fig.3g). Both scMTNI and scMTNI + Prior networks diverged in a manner consistent with the lineage structure. scMTNI networks formed three groups of cell types, (C4, C8, C1, C7), (C2, C3) and (C5 (ESC)). scMNTI + Prior found similar groupings but placed C5 (ESC) closer to (C1, C7, C8, C4) branch. Both methods showed that C5 is closest to C1, which could be an important transitioning state of cells during reprogramming. SCENIC showed similarity among C1, C4, C7, however had lower similarity scores for most pairwise comparisons which made it difficult to discern a clear lineage structure. CellOracle topology identified the (C2, C3) group, but placed it under a subtree with (C4, C8), which, though feasible given the heterogeneity of the system, is less consistent with the gradual progression of the reprogramming process through the intermediate C7 state. The networks inferred by the other methods were very dissimilar which is biologically unrealistic given the high heterogeneity of the reprogramming system with several intermediate populations32. Overall, these results suggest that

scMTNI+Prior recovered regulatory networks of high quality and the networks exhibit a gradual rewiring of structure from the MEF to the pluripotent state.

scMTNI predicts key regulatory nodes and GRN components that are rewired during reprogramming

To gain insight into the regulatory mechanisms of cell populations that successfully reprogram versus those that do not and to further characterize these different cell clusters, we examined the rewired network components in each cell type-specific network inferred by scMTNI + Prior. We used two complementary approaches: k-means edge clustering and Latent Dirichlet Allocation (LDA, Methods). In the k-means edge clustering approach, we represented each edge in the top 4k confidence set of any cell cluster, by a vector of confidence scores in each cell cluster-specific network (if an edge is not inferred in the network it is assigned a weight of 0). Next, we clustered edges based on their edge confidence pattern into 20 clusters determined by the Silhouette Index coefficient optimization (Fig. 4a). The largest “edge clusters” exhibited interactions specific to one cell cluster (e.g., E4, E6, E7, E11, E13, E15, and E16), while smaller clusters exhibited conserved edges for more than one cell cluster (e.g., E2, E5, E12). To interpret these edge clusters, we identified the top regulators associated with each of the edge clusters (Fig.4b). E16, which was MEF-specific (C4) had Npm1, Nme2, Thy1, Ddx5, and Loxl2 as the top regulators which are known MEF-specific genes. In contrast, E11, which was ESC-specific(C5) had Klf4, Sp1, Sp3 as some of its top regulators, which have known roles in stem cell maintenance (Klf4), or are essential for early development (Sp1[37]) and post natal development (Sp3[38]). Edge clusters that

shared edges across multiple cell clusters, e.g., E5 (C4, C8, and C1), shared some of the top-ranking regulators such as Npm1 and Thy1 with the MEF-specific cluster and also identified other fibroblast-specific genes such as Col5a2 and Ybx1. Finally, E2 which comprised shared edges between cell clusters C1 and C5, contained Esrrb, as its top regulator (Fig. 4b). Esrrb plays an important role for establishing and maintaining the pluripotency network[39]. This further supports the lineage structure that C1 likely represents a population of cells that are committed to becoming pluripotent.

While the k-means analysis identified regulatory hubs specific to individual cell clusters, it was challenging to identify entire subnetworks that rewired at specific branch points because it treats each edge independently. We developed an approach by adopting Latent Dirichlet Allocation (LDA) that was recently used to study regulatory network rewiring from transcription factor ChIP-seq datasets[40] (Methods). In this approach, each TF is treated as a “document” and target genes are treated as “words” in the document. Each document (TF) is assumed to have words (genes) from a mixture of topics, each topic in turn interpreted as a pathway. TFs across cell clusters are treated as separate documents. We applied LDA with k =10 topics (Fig. 4c, d, Supplementary Figs. 8–10),and examined each of the topics based on their Gene Ontology process enrichment (Supplementary Fig. 11), and the tendency and identity of specific regulators to rewire across the cell clusters. Topics 3 and 6 are enriched for cell cycle terms (Supplementary Fig. 11). Other processes associated with these topics included immune response (topic 1), developmental processes (topics 1, 3 and 8), electron transport (topic 9), and chromosome organization (topic 10). Topic 3 networks were among the most divergent networks across the cell populations and identified

several known regulators of pluripotency (Fig. 4c). In particular, Esrrb was a hub in C5 (ESC) and C1 (closest to ESC) but absent in the other cell clusters.

We used the LDA analysis to further characterize cell populations that become pluripotent (C1-C5 branch), and those that remain stalled (C7-C3-C2 branch) by identifying regulators that gained or lost connections between these two branches. Several topics included regulators that showed a difference in connectivity between these branches including topics 2, 3, 4, 6, 8, and 9. The regulators that gained edges in the pluripotency branch compared to the stalled branch included cell cycle regulators (Top2a, Ccnb1: topic 3) and known pluripotency genes (Esrrb: topic3 and Klf4: topic 4, Fig. 4d). In contrast, regulators that gained connections in C7-C3-C2 branch relative to the C1-C5 branch (or maintained connections similar to C4), included MEF-specific genes such as Loxl2, Fosl2 (topic 2), Aebp1 (topic 6), Hoxd13 (topic 8), and Fosl1, Nme2 and Ccng1 (topic 9). Nme2 is known to regulate Myc, which is one of the four reprogramming factors[41]. Aebp1, associated with fibroblast differentiation[42], and Loxl2, associated with connective tissue[43,44], persisted in all three cell clusters in the stalled branch (C7-C3-C2). Overall, our analysis indicated that in cell populations that do not reprogram successfully, cell cycle regulators have lower connectivity while several of the MEF regulators (e.g., Nme2, Aebp1) persist or gain connections. These new predicted regulators can be perturbed to examine the impact on cellular reprogramming efficiency.

Inferring gene regulatory networks in human hematopoietic differentiation

To examine the utility of scMTNI in a different cell fate specification system, we applied scMTNI to a published scATAC-seq and scRNA-seq dataset for human adult hematopoietic differentiation[45]. This dataset profiled accessibility and transcriptomic state of immunophenotypic populations that were sorted based on cell surface markers and enabled studies of how multipotent progenitors transition into lineage-specific cell states. We considered the cell populations profiled with both scATAC-seq and scRNA-seq datasets: hematopoietic stem cell (HSC), common myeloid progenitor (CMP), granulocyte macrophage progenitors (GMP) and monocyte (Mono). These populations are known to be heterogeneous comprising multiple subpopulations[45]. To identify these sub-populations, we again applied LIGER23 and identified 10 integrated clusters of RNA and accessibility (Fig. 5a–d). Most clusters exhibited a mixed composition: C8 is mainly composed of HSCs but also included CMP0 cells; C6 and C9 are composed of GMP and CMP0 cells. C1 (73 cells) and C4 (37 cells) were mainly composed of Mono cells and were combined into C1. C5 had too few RNA cells (22 cells) and was excluded from further analysis. We next inferred a cell lineage tree from these 8 cell clusters using a minimal spanning tree approach[24] as described in the reprogramming study (Fig. 5e, Methods). As C8 is largely made up of HSC cells and HSC is the starting cell type, we treated C8 as the root of the lineage.

We applied the same set of network inference algorithms to this dataset as the reprogramming dataset: scMTNI, scMTNI+Prior, INDEP, INDEP+Prior, SCENIC and CellOracle. We assessed the quality of the inferred networks from each method by comparing them to gold standard edges from published ChIP-seq and regulator perturbation assays from several human hematopoietic cell types. This included ChIP-

seq datasets from the UniBind database (Unibind[46]), ChIP-seq (Cus_ChIP) and regulator perturbation (Cus_KO) experiments in the GM12878 lymphoblastoid cell line from Cusanovich et al.[47] and the intersection of ChIP and perturbation studies (Cus_KO+Cus_ChIP, Cus_KO+Unibind). In total, we had five gold standard networks. We used F-score and the number of predictable TFs of the top 500, 1k, 2k edges in the inferred network (Methods, Fig. 5f, Supplementary Fig. 12). The relative performance of the algorithms depended upon the gold standard. Algorithms that did not use priors (INDEP, SCENIC and scMTNI) performed comparably (with no significant difference) on three of the five gold standards. On Unibind and Cus_KO+Unibind, SCENIC is significantly better than INDEP and scMTNI (Fig. 5f, Supplementary Fig. 13). Methods that used prior knowledge, CellOracle, INDEP+Prior, scMTNI+Prior, were generally better than methods without priors for the ChIP-based datasets (Cus_ChIP, Unibind). CellOracle performs better than INDEP+Prior and scMTNI+Prior on Cus_ChIP and Unibind, but is outperformed by all methods on any of the regulator perturbation datasets. INDEP+Prior and scMTNI+Prior are comparable across the gold standard datasets with no significant difference in performance (Fig. 5f, Supplementary Fig. 13). Based on number of predictable TFs in the predicted networks (Supplementary Fig. 14), INDEP+Prior and scMTNI+Prior recovered more predictable TFs especially in KO experiments, while CellOracle recovered more predictable TFs in Cus_ChIP and UniBind. For the Unibind dataset, we had ChIP-seq based gold standard edges for different blood cell types, with 1 to 48 transcription factors (Table 1). Of the 10 cell types, methods that used priors performed significantly better than methods that did not on the GM_B-cells and Hematopoietic Stem Cells (HSCs) which had the largest number

of TFs (Supplementary Figs. 15, 16). However, CellOracle had much lower performance in other cell types and was outperformed by methods with and without priors, likely because of the smaller number of TFs in these datasets. The number of predictable TFs per dataset and method was generally low with the exception of GM_B-cells where methods with priors were better than methods without priors (Supplementary Fig. 17). However, these gold standards were much smaller and therefore can assess smaller portion of the inferred networks.

We next examined the inferred networks for the extent of change on the lineage structure (Fig. 5g). The single-task learning methods INDEP and INDEP+Prior exhibited a low overlap across each pair of cell lines and did not as such obey the lineage structure. SCENIC recovers part of the lineage structure, but placed C7 (common myeloid) close to C6 (granulocyte-macrophage progenitors (GMP)) rather than C10, which has similar sample composition as C7. In contrast, scMTNI and scMTNI+Prior were able to find two groups of cell types, one corresponding to the HSC and CMP2 branch consisting of C8, C3, and C2, and the second corresponding to the CMP0, CMP1, and GMP branch (C6, C9, C10, and C7). CellOracle also inferred a similar tree with small variations within these two groups. For this dataset, the addition of accessibility or lineage information was helpful to capture realistic extents of network level changes.

Inferring shared and lineage-specific regulators for hematopoietic differentiation

Similar to our cellular reprogramming study, we examined the scMTNI+Prior networks to identify cell type-specific regulators and network components (Fig. 6) with k-

means and LDA analysis. We applied k-means edge clustering to the union of top 5k edges in any of the cell clusters and identified 19 edge clusters (Methods). Compared to the reprogramming study, a larger portion (94% vs 86%) of the edges are specific to one cell cluster (Fig. 6a). We used these edge clusters to examine the differences and similarities at the branch between the CMP clusters (C7, C10), and the GMP clusters (C6 and C9). Edge cluster E12 was specific to C7 and C10, E18 was specific to C6 and C9, and E19 shared edges from C6, C9, C10, C7. Both E19 and E12 had YBX1 and TSC22D3 as top regulators (Fig. 6b). YBX1 is known to direct fate of HSCs with high expression in myeloid progenitor cells[48] and involved in monocyte/macrophage differentiation[49]. TSC22D3, which is a glucocorticoid leucine zipper[50], is involved in differentiation of hematopoietic stem cells[51]. E12 additionally had KLF1, FLI1, S100A4 as top regulators. KLF1 is an essential regulator for the erythroid lineage[52,53], which is derived from the myeloid progenitor cells. FLI1 also plays a role in erythroid lineage by regulating the Erythpoetin protein[54], suggesting these cells are committed to the erythroid lineage. In contrast, E18 which shared edges between C6 and C9 identified immune system-related regulators such as IRF8 and NFKBIA which have been associated with general lymphoid development (IRF8[55]) or specific lineages such as B cells (NKBIA[56]). Overall, the k-means edge clustering approach helped identify the key regulators with known or plausible roles in hematopoiesis that could explain the differences among the different lineages.

Our LDA topic analysis predicted several cell type-specific network components with different extents of conservation across the lineage (Fig. 6c, d, Supplementary Figs. 18–20). These topics were enriched in diverse biological processes such as cell

cycle (Topic 1 and 8, Supplementary Fig. 21) and blood related processes (Topic 9). Topic 2 showed a gradual rewiring of an ID2-specific network from the HSC populations (C8, C3, C2), to KLF1 and MYC centered networks for C7 and C10 which represented the CMP populations (Fig. 6c, d). ID2 which belongs to the Inhibitors of DNA family of proteins has been shown to regulate both the erythroid and lymphoid lineages[57] and is consistent with its presence in the C8, C3, C2 clusters. Furthermore, KLF1 connectivity was more pronounced in C7 compared to C10, which could indicate these cells are more committed than those in C10. Similarly, PBX1 which is a key regulator of differentiation versus self-renewal was seen in C7 and C9. Topic 3 captured additional differences between the two GMP clusters, C6 and C9, with IRF8 exhibiting more connections in C6 compared to C9 (Fig. 6d, Supplementary Fig. 18). Topics 1, 6 and 10 exhibited a conserved core around HMGB2, TSC22D3, and YBX1 respectively, across all cells clusters (Supplementary Figs. 18–20). HMGB2 is an important regulator for HSCs[58]. Both YBX1 and TSC22D3, which were also identified in our k-means analysis, have known roles in hematopoiesis[48]. Topic 8 was associated with various cell cycle and chromatin remodeling regulators such as TOP2A, CDC20, and CCNB1 (Supplementary Figs. 20, 21). Taken together, the LDA analysis identified subnetworks centered on candidate key regulators with known general roles in hematopoiesis as well as regulators involved in specific lineage decisions.

Inferring gene regulatory networks in human fetal hematopoiesis

Our applications of scMTNI so far were on cell lineages where a branching structure was computationally inferred. To examine the utility of scMTNI in a system

with known branching lineage structure, we. Applied it to a published scATAC-seq and scRNA-seq dataset of human fetal hematopoiesis[59], which captured specification to multiple blood lineages (Fig. 7a). We considered the cell populations measured with both scATAC-seq and scRNA-seq datasets at two resolutions: (1) coarse resolution comprising hematopoietic stem cell (HSC), multipotent progenitors (MPPs), lymphoid-myeloid progenitors (LMPs), MK-erythroid-mast progenitors (MEMPs), granulocytic progenitors (GPs), and (2) fine-grained resolution, which additionally included the derived cell types from these progenitor populations. We evaluated the methods that incorporate prior and their no-prior versions on this dataset: scMTNI, scMTNI+Prior, INDEP, INDEP+Prior, and CellOracle, at two levels of resolution of the cell types (Methods).

On the fine lineage, algorithms that did not use priors (INDEP and scMTNI) performed comparably based on F-score (with no significant difference) on all five gold standards (Fig. 7b, Supplementary Figs. 22, 23)). INDEP+Prior, scMTNI+Prior, which use priors were significantly better than methods without priors, while CellOracle performed the worst in all gold standards. INDEP+Prior and scMTNI+Prior are comparable across the gold standard datasets. Based on predictable TFs, scMTNI+Prior and INDEP+Prior were the best (Supplementary Fig. 24). As observed in the Buenrostro dataset, CellOracle did comparably to other methods on the ChIP-based gold standards (Unibind, Cus_ChIP), but had fewer predictable TFs in the other gold standards. The poor performance of CellOracle is likely due to its complete reliance on the prior network for determining the structure of the final inferred network. We compared scMTNI+Prior and CellOracle on the coarse lineage and observed similar

superior performance of scMTNI+Prior on both F-score and predictable TF metrics (Supplementary Fig. 30A, B).

We next examined the lineage structure by constructing an MST from pairwise distances of the inferred networks and compared it to the ground truth (Fig. 7c). The single-task learning methods INDEP and INDEP+Prior inferred networks had very low overlap for each pair of cell lines and the resulting lineage tree was different from the ground truth (Fig. 7c). In contrast, scMTNI and scMTNI+Prior were able to recover the cell lineage exactly as the input cell lineage tree. CellOracle, inferred more similarity across cell types and captured several aspects of the original lineage (e.g., MEMP deriving from HSC-MPP), but did not correctly recover several other aspects (e.g., LMPs and GPs derived from HSC, Granulocytes derived from GPs). For the coarse lineage, scMTNI+Prior and CellOracle inferred the same tree, but placed LMPs and GPs under MEMPs instead of under HSCs (Supplementary Fig. 30C). Taken together, these results show that scMTNI+Prior's framework of using lineage information and accessibility results in inference of more accurate GRN structure and dynamics during the differentiation process for known branching cell type trajectories.

Examining dynamics of GRN components for fetal hematopoiesis

We applied our k-means and LDA analysis to identify regulators associated with edge rewiring and subnetwork changes for the fine (Fig. 8a–c, Supplementary Figs. 25–28) and coarse hematopoiesis lineages (Fig. 8d, Supplementary Figs. 31–35). The k-means analysis identified edge clusters spanning multiple cell types of the lineage tree (e.g., E16, E15, E21, E14, E13, E19, E7) as well as individual lineages (E4: B cells, E3:

Granulocytes, E5: Erythrocytes, E9: Mast cells, E2: HSCMPPs, E18: MEMPs) (Fig. 8a). We examined the regulators associated with the edge clusters shared across multiple cell types and found HNRNPK and PTMA to be frequently associated with these clusters (Fig. 8b). HNRNPK has a number of regulatory functions across diverse cell types including as a regulator of hematopoiesis[60]. PTMA, which stands for prothymosin alpha is not well understood for its function but is implicated in growth and survival of cells of hematopoietic origin, and required for the filament-inducing activity of macrophage lysate[61], which would be consistent with its expression in the hematopoietic lineage[62]. E17 had edges common to the Myeloid lineage spanning HSC-MPPs, MEMPs, Mast-cells, Megakaryocytes and Erythroid populations and had ENO1, NPM1, SNRPD1 in addition to HNRNPK and PTMA as top regulators (Fig. 8b). ENO1 encodes a glycolytic enzyme which is expressed in several human tissues and has been shown to be a regulatory enzyme with links to the MYC pathway[63]. E2 had edges specific to HSC-MPPs and was associated with PTMA, SNRPD1, SOX4 and EEF1A1, which have immune-related functions. E18 which was specific to MEMPs was associated with KLF1, BRPF3 and PTMA. KLF1, which was found in the Buenrostro et al. dataset of adult hematopoiesis as well[45], is an essential regulator for the erythroid lineage[52,53], and was also found to be upregulated by Ranzoni et al. as cells transitioned from HSC/MPP to MEMPs[59]. E16and E14 are edge clusters shared across all cell types with EEF1A1, CDC20, HMGN2, NPM1, TOP2A as top regulators. HMGN2 belongs to the high mobility group of proteins, which was identified in our analysis of the Buenrostro et al. dataset as well. Other regulators implicated cell cycle (CDC20, TOP2A) or more general regulators of development and proliferation (NPM1). Cell-cycle

and cell-fate decisions are inherently tied especially in progenitor populations where the cell fate decision could be influenced by the cell cycle stage of the cells[64]. The k-means analysis of the coarse lineage exhibited much more shared network structure compared to the fine lineage, though it also identified edge sets specific to each coarse cell type (E1: HSC, E3: GPs, E2: LMPs, Supplementary Fig. 31). Several of the regulators identified in the fine lineage analysis were seen in the coarse lineage analysis showing overall consistency of our results. For example, E8 which had edges shared across all cell types had EEF1A1, FOS, HMGN2, NPM1 as the top regulators. Similarly, KLF1 was identified in the MEMP-specific edge cluster in the coarse (E4) and fine lineages (E17). The coarse lineage analysis also found additional regulators. For example, E2, which was specific to the LMP lineage was associated with IRF8, KLF3, BAG4, and MAP2K7. IRF8, which was identified in the Buenrostro et al. dataset as well plays a key role in innate immune response and is an essential for development of the lymphoid lineage including B cells[55], monocytes and pDCs[65].

Our LDA analysis identified topics representing subnetworks that rewire from the HSC state to different lineages (Methods). The topic genes were enriched in immune response (topic 1), cell-cycle (topics 2, 3 and 5), cellular respiration (topic 4) and general metabolic processes (topic 7, Supplementary Fig. 29A). LDA topic 3 identified a regulatory subnetwork that gained connections in B cells for regulators like FOXP4 and PPR2R5B (Fig. 8c, Supplementary Fig. 26) and was enriched for cell cycle processes (Supplementary Fig. 29A). In contrast, topic 1 represented an opposite pattern of gradual loss of edges connected to FOS from HSC-MPP to downstream lineages (Supplementary Fig. 25). FOS was found to be upregulated in Ranzoni et al. in the

HSCs/MPPs population[59]. Other topics exhibited conserved hubs like PTMA (topic 4, Supplementary Fig. 26), HNRNPK (topic 8, Supplementary Fig. 27)), and NPM1 (topic 5, Supplementary Fig. 26) across multiple lineages and several cell cycle regulators such as TOP2A and CDC20 (topic 2, Fig. 8c, Supplementary Fig. 25). On the coarse lineage, the LDA analysis revealed more hubs in HSC-MPPs which were lost when differentiating to the other lineages (Fig. 8d, Supplementary Figs. 31–35). The exceptions were ENO1 (topic 7, Supplementary Fig. 34), HMGN2 and NPM1 (topic 4, Supplementary Figs. 31, 33) and PTMA (topic 3, Supplementary Fig. 31), which persisted at all lineages. NPM1, which was found both in fine and coarse tree, plays an important role in hematopoietic progenitors, especially in early myeloid differentiation[66].A few regulators also gained connections in specific lineages, for example, LGALS1 (topic 3), JAG1 (topic 7), CDK1 (topic 4) had more edges in the LMP lineage and PLEK in the MEMP lineage (Supplementary Fig. 31). Both LGALS1[67] and JAG1[68] have been shown to be involved in hematopoiesis, however, the specific roles in this process is not as well-characterized. In topic 5, we observed the persistence of an IRF8-specific network from the HSCs/MPPs to LMPs populations, which was lost in MEMPs/GPs lineage and is consistent with our k-means analysis and our results from Buenrostro et al. (Supplementary Fig. 33). Taken together, the k-means and LDA analysis identified several components of fetal hematopoiesis GRNs that changed as cells differentiated from HSC-MPP to differentiated cell types. While many of the regulators have well-characterized roles in hematopoiesis, several are previously uncharacterized that can be followed up with targeted functional studies.

Discussion

Single-cell technologies have transformed our ability to study cellular heterogeneity and cell-type specific gene regulation of known and novel cell populations. Defining gene regulatory networks from scRNA-seq data of developmental systems has remained challenging as most existing methods have assumed a static view of the GRN and do not leverage accessibility to inform the GRN structure. To address this need, we developed single-cell Multi-Task Network Inference (scMTNI), a probabilistic graphical model-based approach that uses multi-task learning to infer cell type-specific GRNs on a cell lineage tree by integrating scRNA-seq and scATAC-seq data and model the dynamics of these regulatory interactions on a lineage. A major benefit of the scMTNI framework is its flexibility in incorporating different sources of accessibility information as well as the ability to model dynamics on cell lineages of different topologies. The probabilistic prior-based framework makes scMTNI more robust to noisy or incomplete accessibility data and allows the incorporation of additional regulators such as signaling proteins and TFs with no binding information. Guided by the cell lineage structure, scMTNI's inferred networks exhibit meaningful changes along the trajectory and identify regulators and network components specific to cell populations transitioning to different lineage paths.

Multi-task learning is well-suited for the inference of cell type-specific GRNs. However, a key question is how to implement multi-task learning for GRN inference. A number of multi-task learning algorithms were developed for inferring GRNs and functional networks from bulk transcriptomic data but have not been systematically compared for their effectiveness on single-cell transcriptomic data. Some approaches,

such as AMuSR[28] have used a flat hierarchy where all the tasks are considered equally related. For heterogeneously related datasets, a hierarchy or a tree is well-suited to model the dependence across datasets. Such hierarchies can be implemented as a phylogenetic tree with observed data at the tips of the tree as in GNAT[26] and MRTLE[25],or as a cell-lineage tree with observations at all nodes in the tree. scMTNI and MRTLE both use a tree-based structure prior, whereas AMuSR, GNAT, and Ontogenet used a regularized regression parameter to implement multi-task learning. scMTNI and MRTLE have better performance in predicting the gene regulatory relationships than single-task learning algorithms. The performance of Ontogenet is better than the single-task learning algorithms LASSO and INDEP in at least two cell types, and comparable to SCENIC. A prominent factor contributing to the difference in the performance of the algorithms was whether the models inferred a directed graph versus an undirected graph, with GNAT generally suffering likely due to this reason. Performance of GNAT is worst among multi-task learning algorithms and comparable to the single-task learning algorithms. We speculate that the undirected graphical models learned by GNAT might be a reason that the performance is not as good as other multi-task learning algorithms. We also examined the performance of algorithms across different parameter settings that control for sparsity as well as for sharing information. We found that the algorithms were generally robust to the setting of sharing and more sensitive to the extent of sparsity. However, multi-task learning algorithms generally outperformed single-task learning algorithms indicating that this is a useful direction for methodological development for GRN inference from single cell omic datasets. Importantly, single-task

learning infers very different networks that makes it challenging to study transitions across the networks.

Once GRNs are inferred across multiple cell types, the next challenge is to examine which components of the GRNs change along the lineage. We developed two complementary techniques to study dynamics. Our k-means edge clustering method was able to find regulatory connections that were unique to each cell cluster, while our LDA topic model-based dynamic network analysis highlighted subnetworks that were activated or deactivated along the lineage. We applied our tools to study GRN dynamics in adult and fetal hematopoietic cell differentiation and reprogramming from mouse embryonic fibroblasts to embryonic stem cells. We found that these systems exhibited different dynamics, with the reprogramming system exhibiting more edges shared across populations compared to the adult hematopoietic system which identified most edges as cell cluster-specific. In all three systems, our analysis identified known and previously uncharacterized regulators. For example, in the reprogramming system, we found that cells that were closer to the end point pluripotent state already had an Esrrb-centered GRN component active. In contrast, cells that were on an alternate trajectory exhibited persistence of the MEF regulatory program including regulators such as Aebp1. Between adult and fetal hematopoiesis we found several shared regulators that were known lineage-specific regulators (e.g., IRF8 in the lymphoid lineage), but also identified regulators unique to each system which could be followed up with future validation studies.

scMTNI currently assumes that the input lineage structure is accurate. However, lineage construction, especially from integrated scRNA-seq and scATAC-seq datasets

is a challenging problem. One direction of future work is to assume the initial lineage structure is inaccurate and incorporate the refinement of the lineage structure as part of the GRN inference procedure. A second direction of work is to model more fine-grained transitions within each cell population, for example using RNA velocity or pseudotime[69], which will complement the coarse-grained dynamics that scMTNI currently handles. Studies from bulk RNA-seq data have shown that estimating hidden transcription factor activity (TFA)[70] can further improve the performance of network inference. Thus, another direction of future work is to estimate hidden TFA and incorporate these to improve the accuracy of the inferred networks. Finally, SCENIC generally outperforms the single-task learning algorithms which do not use prior, which is likely because of its regression-tree based model that captures non-linear dependencies and is less prone to the sparsity of the dataset. While scMTNI's stability selection framework can capture some non-linearities, another direction of future work is to extend scMTNI to model more non-linear dependencies.

In summary, scMTNI is a tool to infer cell type-specific regulatory networks and their dynamics on a cell lineage which combines scRNA-seq and scATAC-seq data. As single cell multi-omic datasets become increasingly available, we expect scMTNI to be broadly applicable to predict GRNs and prioritize regulators associated with regulatory network dynamics across cell types in diverse cell-fate specification processes.

Acknowledgements

We thank the Center for High Throughput Computing at University of Wisconsin-Madison for computational resources. This work is supported by the National Institutes

of Health NIGMS grant 1R01GM117339 (S.R., S.Z., A.F.S.) and 1R01GM144708-01A1 (S.Z., S.G.M, S.R., S.H.), the Department of Energy grant DE-SC0021052 (S.Py.), and grant 2R01GM113033 (R.S., S.Pi.). We thank Dr. Jason Buenrostro for help with accessing the scRNA-seq data for the adult hematopoiesis dataset from Buenrostro et al.

Competing interests

The authors declare no competing interests.

Materials and Methods

This research complies with all relevant ethical regulations. Mice used in the reprogramming study were maintained in agreement with our UW-Madison Institutional Animal Care and Use Committee (IACUC) approved protocol (ID M005180-R03).

Single-cell Multi-Task Network Inference (scMTNI)

Single-cell Multi-Task Network Inference (scMTNI) is a probabilistic graphical model-based approach that uses multi-task learning to infer gene regulatory networks for cell types related by a cell lineage tree (Fig. 1). We define a cell type to be a group of cells with similar transcriptome and accessibility levels as defined by existing cell clustering methods. Each task learns the gene regulatory network (GRN), $G^{(d)}$ for each cell type or cell cluster d. Given cell type-specific datasets for M cell types, $D = \{D^{(1)}, \cdots, D^{(M)}\}$, our task is to find the set of graphs $G = \{G^{(1)}, \cdots, G^{(M)}\}$ and parameters $\Theta = \{\theta^{(1)}, \cdots, \theta^{(M)}\}$ for each of the cell types. $G^{(d)}$ is modeled as a dependency network[22], a class of

probabilistic graphical models for inferring directed, predictive relationships among random variables (regulators and genes). Each gene is modeled as a random variable $X_i^{(d)}$ which encodes the expression level of gene i in each cell. A conditional probability distribution $P(X_i^{(d)}|R_i^{(d)})$ models the relationship between gene i and its set of regulators, $R_i^{(d)}$ in cell type d. In a dependency network, GRN inference entails estimating the regulators $R_i^{(d)}$ for each gene i in each cell type d. To enable joint learning of these cell type-specific networks, our goal is to find the set G ={$G^{(1)}$, ⋯ , $G^{(M)}$} and parameters Θ ={$\theta^{(1)}$, ⋯ , $\theta^{(M)}$} by estimating the posterior distribution of these two sets and finding their maximum a posteriori values:

$$\mathbf{P(G,\Theta|D)} \propto \mathbf{P(D|G,\Theta)P(\Theta|G)P(G)} \quad (1)$$

P(D|G, Θ) is the data likelihood, expanded as $\prod_d P(D^{(d)}|G^{(d)}, \theta^{(d)})$. In a dependency network, pseudo likelihood[22] is used to approximate the data likelihood for each cell type, defined as the products of the conditional distribution of each random variable $X_i^{(d)}$given its neighbor set $R_i^{(d)}$ in cell type d, $P(X_i^{(d)}|R_i^{(d)}, \theta_i^{(d)})$. Thus, the likelihood can be written as:

$$\mathbf{P(D|G,\Theta)} \propto \prod_{d\in\{1,\ldots,M\}} \prod_{i\in\{1,\ldots,N\}} \mathbf{P}(X_i^{(d)}|\mathbf{R}_i^{(d)},\boldsymbol{\theta}_i^{(d)}) \quad (2)$$

Given the neighbor set $R_i^{(d)}$, the above quantity can be computed efficiently. We assume that each variable $X_i^{(d)}$ and its neighbor set $R_i^{(d)}$ in cell type d are from a multi-variate Gaussian distribution. Thus, $P(X_i^{(d)}|R_i^{(d)}, \theta_i^{(d)})$ can be modeled using a conditional Gaussian distribution with mean $\mu_{X_i^{(d)}|R_i^{(d)}}$ and variance $\sigma_{X_i^{(d)}|R_i^{(d)}}$which can be

estimated in closed form. $R_i^{(d)}$ is selected from the input list of regulators using a greedy search algorithm, executed in parallel across all cell types (See Supplementary Methods). The second term P(Θ|G) in Equation (1) is estimated using the maximum likelihood settings of the parameters. The third term P(G)=P($G^{(1)}$, ⋯ , $G^{(M)}$) in the objective function is the structure prior and is defined in a way to capture the state of an edge across all cell types modeled, where G ={$G^{(1)}$, ⋯ , $G^{(M)}$}. We assume that P(G) is composed of two priors, one is the cell-type specific prior P(T), where T ={$T^{(1)}$,...,$T^{(M)}$}, and the other one is a cell lineage structure prior P(S) which captures the similarity between related cell types along the cell lineage tree, where S ={$S^{(1)}$,...,$S^{(M)}$}.

P(T) is the cell-type specific prior, which decomposes over a product of cell-type specific graphs: $P(T^1, \dots, T^M) = \prod_{d=1}^{M} P(T^d)$. The $P(T^d)$ decomposes over a product of individual edge configurations, $P(I_{u,v}^{(d)})$, where $I_{u,v}^{(d)}$is an indicator function that represents whether there exists an edge between regulator u to target gene v in cell type d, $X_u \rightarrow X_v$ as follows:

$$I_{u,v}^{(d)} = \begin{cases} 1, \text{if there is an edge from } u \text{ to } v \text{ in cell type } d, \\ 0, \text{otherwise.} \end{cases} \quad (3)$$

As in Roy et al.[71], we model the prior probability using a logistic function:

$$\mathrm{P}\left(I_{u,v}^{(d)} = 1\right) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * m_{uv}^{(d)})}} \quad (4)$$

The β_0 parameter is a sparsity prior that controls the penalty of adding of a new edge to the network, which takes a negative value ($\beta_0 < 0$). A smaller value of β_0 will result in a higher penalty on adding new edges and will therefore infer sparser networks. The β_1

parameter controls how strongly motifs are incorporated as prior ($\beta_1 \geq 0$). A higher value of β_1 will result in motif presence being valued more strongly to select an edge. β_1 is set to 0 when there is no cell type-specific motif information available. $m_{u,v}^{(d)}$ is the weight of the edge from regulator u to target v in the prior network and is computed based on the motif instance score if gene v has a motif instance of regulator u in its promoter region, additionally filtered by available bulk or single cell ATAC-seq peaks. Thus, we have

$$\mathrm{P}(\mathbf{T}) = \prod_{d=1}^{M} \mathrm{P}(T^{(d)}) = \prod_{d=1}^{M} \prod_{u,v;u\neq v} \mathrm{P}(I_{u,v}^{(d)}) \qquad (5)$$

The cell lineage structure prior P(S) is constructed to make use of multi-task learning. We define $\mathrm{P}(S^{(1)},...,S^{(M)})$ as a product over a set of edges between regulators and target genes: $\prod_{u,v;u\neq v} P(I_{u,v}^{(1)}, \ldots, I_{u,v}^{(M)})$. Under the assumption that the prior probability of the edge state in one cell type is only dependent upon its state in the predecessor cell type, we have:

$$\mathrm{P}(\mathbf{S}) = \prod_{u,v;u\neq v} \mathrm{P}(I_{u,v}^{(1)},\ldots,I_{u,v}^{(M)}) = \prod_{u,v;u\neq v} \prod_{d\in\{1,\ldots,M\}} \mathrm{P}(I_{u,v}^{(d)}|I_{u,v}^{pa(d)})\mathrm{P}(I_{u,v}^{(r)}), \qquad (6)$$

where pa(d) denotes the predecessor cell type of cell type d on the cell lineage tree and r denotes the starting root cell type. $P(I_{u,v}^{(d)}|I_{u,v}^{pa(d)})$ is a measure of overall gain and loss of regulatory connections between related cell types and is assumed to be the same across the set of edges. Thus, it can specified by three parameters: the probability of gaining a regulatory edge in the root cell type, $p_r = P(I_{u,v}^{(r)})$, the probability of gaining a regulatory edge in cell type d given that the edge does not exist in its predecessor cell type, $p_g^d = P(I_{u,v}^{(d)} = 1|I_{u,v}^{pa(d)} = 0)$, and the probability of maintaining a regulatory edge in

cell type d, given it is present in its predecessor cell type $p_m^d = P(I_{u,v}^{(d)} = 1 | I_{u,v}^{pa(d)} = 1)$. These parameters of the priors can be set by the user or estimated empirically by analyzing different configurations and selecting those values with the best agreement with existing biological knowledge of the system. scMTNI uses a greedy score-based structure learning algorithm. Please refer to Supplementary Methods for details.

Input datasets

Simulated datasets

To benchmark the performance of different multi-task and single-task learning algorithms, we simulated single cell expression data from a lineage resembling a linear differentiation process for three cell types (Fig. 2a). We simulated network dynamics on the lineage while controlling the extent of similarity with the three prior parameters: p_r, the probability of having an edge in the starting/ root cell type; p_g^d, the probability of gaining an edge in cell type d that is not in the predecessor cell type; p_m^d, the probability of maintaining an edge in cell type d from the predecessor cell type. We set $p_r = 0.5$, $p_g^d = 0.4$, and $p_m^d = 0.7\ or\ 0.8$ and simulated three networks from a linear lineage tree for each of the three cell types, each with 15 regulators and 65 genes. Next, we applied BoolODE on the simulated gene regulatory networks and generated single cell expression data for 2000 cells for each cell type. To mimic the dropouts in the scRNA-seq data, we added 80% sparsity uniformly to all genes on the simulation data. We refer to this simulated dataset as dataset 1, consisting of 65 genes and 2000 cells for three cell types. We generated smaller sample sizes of these datasets, dataset 2 and dataset 3 by downsampling dataset 1 to 1000 cells (dataset 2) and 200 cells (dataset 3). We

applied each of the algorithms on these three datasets within a stability selection framework and evaluated their performance based on AUPR and F-score as described in the Evaluation section.

Human hematopoietic differentiation data

Buenrostro et al.[45] measured single-cell accessibility (scATAC-seq) and single-cell RNA sequencing (scRNA-seq) data to study the regulatory dynamics during human hematopoietic differentiation for multiple immunophenotypic cell types: hematopoietic stem cells (HSCs), common myeloid progenitors (CMPs) and granulocyte-macrophage progenitors (GMPs) and Monocytes (Mono). We downloaded processed scRNA-seq data for each cell type from Data S2 of Buenrostro et al. (https://ars.els-cdn.com/content/image/1-s2.0-S009286741830446Xmmc4.zip) and fragment files for the scATAC-seq data from Chen et al.[72] (https://github.com/pinellolab/scATAC-benchmarking/tree/ master/Real_Data/Buenrostro_2018). For the scATAC-seq data, we mapped the fragments into 23,347,540 bins with length of 1000bp. Next, we mapped 1 kb bins to the nearest gene and extracted cells with cell barcodes labeled as HSC, CMP, GMP, and Mono. Next, we filtered out genes with sum of counts in all samples less than 100, producing a processed scATAC-seq dataset with 54,344 genes and 1315 cells across the four cell types. We extracted the count matrix of scRNA-seq from these four cell types; note that CMP cells were in three different clusters: CMP0, CMP1, and CMP2. After filtering out genes with non-zero expression in less than 5 cells, the scRNA-seq data had 12,558 genes and 4165 cells. We normalized the count matrix for depth and variance stabilization based on the pagoda pipeline[73]. We kept 12,393

common genes between scATAC-seq and scRNA-seq data and applied LIGER23 to define integrated cell populations. We applied LIGER with k ∈ 8, 10, 12, 15, 20 factors and found k = 10 to be most appropriate. Cluster C8 was mainly composed of HSCs, C6 was mainly composed of GMP cells, C7 was mainly CMP0 cells, C1 was composed of Monocyte cells, and the rest of the clusters were a combination of several cell types. C5 had too few RNA cells (22 cells) so we excluded it from further analysis. Since the composition of C1 (73 cells) and C4 (37 cells) are very similar, mainly GMP and Mono cells, we combined these two clusters as C1. We inferred a cell lineage tree from the 8 cell clusters using a minimal spanning tree (MST) approach using the python package scipy.sparse.csgraph. Briefly, we used the mean expression profiles across samples of these cell clusters and computed the Euclidean distance between every pair of cell clusters. Then, we inferred the MST from the distance matrix using scipy.sparse.csgraph.

To derive the prior network for each cell cluster we created cluster-specific bam files from the scATAC-seq data using the LIGER clusters. We pooled these bam files to generate pseudo bulk accessibility coverage and applied MACS2 (v2.1.0) to identify scATAC-seq peaks for each cell cluster[74]. We obtained sequence-specific motifs from the Cis-BP database (http://cisbp.ccbr.utoronto.ca/)[75] and used the script pwmmatch.exact.r available from the PIQ toolkit76 to identify significant motif instances genome-wide using the human genome assembly of hg19. We mapped motifs to each scATAC-seq peak and mapped the peak to a gene if it was within ± 5000 bp of the transcription start site (TSS) of a gene. In this case, we connect all motifs to a TSS that are mapped to the same scATAC-seq peak. We used the maximum motif score from

pwmmatch.exact.r for each motif-TSS pair and took the maximum value among all TSSs of a gene as the value for each motif-gene pair. The motif instance score is the log ratio of the Position Weight Matrix (PWM) match score to a uniform background. Finally, to generate the edge weight for each TF gene pair, we used the max score among all motifs mapped to the same TF. To normalize the edge weights across TFs, we converted these weights into percentile scores and selected the top 20% of edges as prior edges.

Mouse cellular reprogramming data

We generated an scATAC-seq time course dataset for cellular reprogramming from mouse embryonic fibroblast (MEFs) to induced pluripotent cells (iPSCs). The dataset contains a total of 6 time points corresponding to the starting MEF, the end pluripotent state (mESC), and four intermediate timepoints of day 3, day 6, day 9 and day 12. The mice used to generate the MEFs used for reprogramming were housed in a facility that ran a 12 h light/12 h dark cycle, had an ambient temperature 72 ◦F and maintained humidity between 20–50%. Mice were maintained in agreement with our UW-Madison Institutional Animal Care and Use Committee (IACUC) approved protocol (ID M005180-R03). Male and female mice of breeding age (at least 6–8 weeks old) from a mixed 129/Bl6 background that are homozygous for the Oct4-2A-Klf4-2A-IRES-Sox2-2A-cMyc (OKSM) transgene at the Col1a1 locus and heterozygous for the reverse tetracycline transactivator (rtTA) allele at the Rosa26 locus were time-mated, from which MEFs were isolated at E13.5. On E13.5, the pregnant female mouse is carefully dissected and all embryos are removed. The head and neck region of the embryo is

separated from the rest of the body and any organ tissues present are also removed, leaving only the fibroblasts. The remaining fibroblast tissue is emulsified and plated onto a 15 cm. The cells are passaged 1–2 additional times before being collected and stored in liquid nitrogen until the start of the experiment. In this study, MEFs with a homozygous genotype for the OSKM transgene and rtTA allele were used for reprogramming experiments. Male neonatal human foreskin fibroblasts (HFFs) from American Type Culture Collection (HFF-1 SCRC-1041) were used as feeders for our reprogramming cells. HFFs were passaged and expanded ~5 times prior to being irradiated. HFFs were irradiated at a level of 80 Gray prior to being used as feeders for the reprogramming MEFs. The process of somatic cell reprogramming is unaffected and is not influenced by the sex of the starting cell population, so the sex of the MEFs used in this experiment is unknown as it is irrelevant to the observed results.

On Day -2, E13.5 reprogrammable MEFs were thawed and on Day -1, they were plated in gelatinized 6-well plates at a seeding density of 5000 cells per well. Reprogramming was induced on Day 0 by adding 2 ug/ml doxycycline (Sigma-Aldrich D9891) to each well, which induced OKSM expression, as well as irradiated DR4 feeder MEFs. Reprogramming cells were maintained in ESC media (knockout DMEM (Gibco #10829-018), 15% FBS (Biowest S1620), L-glutamine (Gibco #15140-122), Pen/Strep (Gibco #33050-061), NEAA (Gibco #11140-050), 2-mercaptoethanol (Sigma-Aldrich #M6250) and leukemia inhibitory factor (Sigma-Aldrich #L5158)). Media was changed every two days. Cells were collected and prepared in a single-cell suspension on days 3, 6, 9, and 12. To generate single-cell suspensions, cells in the wells were washed 5X with DPBS (Gibco #14190-144) and dissociated from plate using 0.25% Trypsin-EDTA

(Gibco #25200-072). Trypsin was neutralized with soybean trypsin inhibitor (Sigma-Aldrich #T6522), cells were filtered through a 40um filter, and spun down for 3min at 300xg (RT). Cells were then resuspended in 1ml of 0.1% BSA-PBS (prepared by diluting 7.5% Bovine Albumin Fraction V solution (Gibco #15260-037) to 0.1% with DPBS) and pipetted up and down 50X. 6 ml of 0.1% BSAPBS were added to cells and spun down again at 300 × g for 3 min. Cells were finally resuspended in 1 ml of 0.1% BSA-PBS. Cell concentration was determined using an Invitrogen Countess II cell counter prior to nuclei isolation, transposition, and single-cell ATAC-sequencing.

scATAC-seq data were generated using the 10x Genomics platform with a targeted nuclei recovery of 4000 and targeted read depth of 25k reads per nucleus. Sequencing was performed using the Illumina NovaSeq 6000 machine and samples were loaded onto a S1 flow cell. The scATAC-seq data was first processed through CellRanger ATAC pipeline (version 1.1.0) to provide the fragments file. We binned the genome at non-overlapping 1 kb bin and computed the number of fragments mapped to each 1 kb bin. Next, we mapped 1 kb bins to the nearest gene for all of the samples. The processed scATAC-seq data contains 25,824 genes and 30,344 cells.

We downloaded scRNA-seq datasets (GEO: GSE108222) for the same time points from ref. 32. We concatenated the expression data from two replicates at each time point and normalized the concatenated matrix for depth and variance stabilization based on a simplified implementation of the pagoda pipeline[73]. Next, for each time point, we removed genes with expression in less than 5 cells. We took the union of genes among all time points and concatenated the expression data across all time points as our final scRNA-seq data matrix. The processed scRNA-seq dataset contains 14,953

genes and 3460 cells. We had a total of 11,926 genes in common between the two datasets, which were used for downstream analysis. We applied LIGER with k ∈ 8, 10, 12, 15, 20 and found k = 8 to provide the optimal clustering of the scRNA-seq and scATAC-seq data determined based on the clustering of the accessibility and transcriptome of the MEF and ESC time points. We inferred a minimal spanning tree from the distance matrix of the pseudobulk expression profiles of each cluster using scipy.sparse.csgraph, similar to the Buenrostro et al. hematopoiesis dataset, and used it as the cell lineage tree. The prior motif was generated in the same way as for the hematopoiesis differentiation dataset using motifs for mouse from the CisBP database[75]. We used the 10 mm mouse genome assembly for this analysis.

Human fetal hematopoietic differentiation data

Ranzoni et al.[77] measured scRNA-seq and scATAC-seq data to study the regulatory dynamics during human developmental hematopoiesis for multiple immunophenotypic blood cell types from fetal liver and bone marrow. We obtained the scRNA-seq (gene by cell) and scATAC-seq data (peak by cell) matrices from https://gitlab.com/cvejic-group/integrativescrna-scatac-human-foetal. We used the annotated cell clusters in ref. 77 for the scRNA-seq data: HSCs/MPPs combined with cycling HSCs/MPPs (HSCs-MPPs), lymphoid-myeloid progenitors (LMPs), MK-erythroid-mast progenitors combined with cycling MEMPs (MEMPs), granulocytic progenitors (GPs), granulocytes, erythroid cells, megakaryocytes, mast cells, monocytes, plasmacytoid dendritic cells (pDCs) and B cells. We took the union of genes among all cell types and concatenated the expression data as our final scRNA-seq data

matrix. We normalized this concatenated matrix for depth and performed variance stabilization based on the pagoda pipeline[73] and removed genes with expression in less than 20 cells. The labeling provided by Ranzoni et al. for the scATAC-seq data omitted many of these cell types making it challenging to determine cell-type specific priors. To overcome this challenge we utilized a label transfer technique based on the method provided in the Seurat v3 package[78]. Briefly, we embedded the scRNA-seq and scATAC-seq cells (after mapping peaks to gene promoters) into a shared lower dimensional embedding (k = 10) utilizing LIGER[23]. We next defined “anchors”, which are pairs of cells that provide a correspondence between the scRNA-seq and scATAC-seq modalities. Each anchor is defined as a mutual nearest neighbor in the lower dimensional space and has an anchor score computed based on the overlap of within and between dataset neighborhoods as specified in the Seurat v3 package. Once the anchor scores are established, we computed the anchor weights for each cell in the scATAC-seq data and transferred labels based on a linear combination of the anchor weights and labels associated with the scRNA-seq cells. Each scATACseq cell with a label score greater than 0.3 was assigned the maximally scoring label. Cells with score below 0.3 were not used to generate the prior network.

To derive the prior network for each cell type, we extracted scATAC-seq peaks present in each cell type derived from our label transfer method. For LMPs, as there are no cells in the scATAC-seq data labeled as LMPs, we took the union of peaks across LMP’s derived cell types (monocytes, pDCs, and B cells) as the scATAC-seq peaks for LMPs. We used a similar strategy as the Buenrostro et al. dataset to generate the prior network. Briefly, we used the same sequence-specific motifs from the Cis-BP

database[75] as the Buenrostro et al. data, mapped motifs to each scATAC-seq peak and mapped the peak to a gene if it was within ± 5000 bp of the gene TSS. For the coarse cell lineage tree, we merged all derived cell types from each parent cell type to produce four cell populations as follows: monocytes, pDCs, NK cells and B cells were merged with the LMP cells; erythroid cells, megakaryocytes, and mast cells were merged with MEMPs; and Granulocytes were merged with GPs. We applied the same approach as the fine tree to prepare the scRNA-seq expression data and prior networks for each cell type using union of scATAC-seq peaks in each cell type and its derived cell types.

Application of network inference algorithms on simulated datasets

We used the simulated datasets to perform benchmarking of the different network inference algorithms. We also used this dataset to study the sensitivity of the algorithms to the different parameter settings. Below we describe each of the algorithms as well as the parameters used for each of the algorithms for the simulated datasets. For all three simulation datasets, we applied all algorithms other than SCENIC within a stability selection framework to estimate the confidence score for each edge in the predicted networks. For stability selection, we subsampled each dataset 20 times randomly using half of the cells and all genes. SCENIC has its own internal sub-sampling and directly outputs the edge importance. scMTNI and baseline methods require list of regulators and target genes information as input. This information is provided to all methods under comparison.

scMTNI: scMTNI has five hyper-parameters: p_r, probability of having an edge in the starting cell type; p_g^d , probability of gaining an edge in a child cell type d; p_m^d the

probability of maintaining an edge in d from its immediate predecessor cell type; a sparsity penalty β_0,that controls penalty for adding edges; β_1, that controls the strength of incorporating prior network. We tested different configurations of the hyper-parameters: $p_r \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$, and $p_g^d \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$, and $p_m^d \in \{0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9\}$, $\beta_0 \in \{-0.005, -0.01, -0.05, -0.1, -0.5\}$. β_1 was set to 0 as there is no prior network in the simulations. If the size of the predicted network for a parameter setting was smaller than the size of the simulated network, we disregarded this parameter setting for comparison. We used the area under the precision-recall curve(AUPR) to compare the scMTNI inferred networks to simulated networks. We also computed F-score on top K edges ranked by the confidence score (where K is the number of edges in the simulated network, C1: K =202, C2:K =217, C3:K = 239). Overall performance of scMTNI was stable across different parameter configurations (Supplementary Fig. 36, Supplementary Methods). To compare against methods, we used values from the best parameter settings for each dataset and cell type as well as all parameter settings (Supplementary Figs. 1, 2).

MRTLE: Multi-species regulatory network learning (MRTLE)[25] is a probabilistic graphical model-based algorithm that uses phylogenetic structure, transcriptomic data for multiple species, and sequence-specific motifs to infer the genome-scale regulatory networks across these species simultaneously. It was developed for bulk transcriptomic data and uses a dependency network model to specify the directed relationship among regulators to target genes. Sequence-specific motif instances can be incorporated as prior knowledge to favor edges supported with the presence of motifs. The multi-task

learning framework is embedded in the phylogenetic prior, which captures the evolutionary dynamics of regulatory edge gain and loss guided by the phylogenetic structure. The MRTLE algorithm has four parameters: p_g the probability of gaining an edge in a child species s that is not in the ancestor species; p_m, the probability of maintaining an edge in a species s given it is also in s's immediate ancestor of s; β_0, a sparsity penalty that controls penalty for adding edges, and a penalty β_1 that controls the strength of motif prior. In the simulation case, we examined different parameter configurations: $p_g \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}$, $p_m \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85\}$, $\beta_0 \in \{-0.005, -0.01, -0.05, -0.1, -0.5, -1\}$. β_1 was set to 0. The overall performance of MRTLE was stable across different parameter configurations (Supplementary Fig. 37). Similar to scMTNI, we used the AUPR and F-score of top K edges to select the best parameter setting. The best parameter setting and all parameter settings were used to compare against other algorithms.

GNAT: The GNAT[26] algorithm uses a hierarchy of tissues to share information between related tissue and infers tissue-specific gene co-expression networks. It was developed for bulk transcriptomic data. GNAT models each network using a Gaussian Markov Random Field (GMRF). It has two parameters: the L_1 penalty λ_s that controls the sparsity of the network, and the L_2 penalty λ_p that encourages the precision matrix of children to be similar to its parent precision matrix. It initially learns a co-expression network for each leaf tissue. Then it infers the networks in internal nodes using the networks in the leaf nodes and updates the networks in leaf nodes iteratively until convergence. Since GNAT learns undirected networks, we transformed them to directed networks by adding edges from a regulator to a target. If the nodes of an edge are both

candidate regulators, we output the edge in both directions. We tested different parameter configurations of λ_s and λ_p. For data 1(n = 2000), λ_s were set to {30, 31, 32,..., 37}, and λ_p were set to {30, 31, 32,..., 40}. For data 2 (n = 1000), λ_s were set to {18, 19,..., 22}, and λ_p were set to {18, 19,..., 25}. For data 3 (n = 200), λ_s were set to {5, 6, 7, 8}, and λ_p were set to {5, 6, 7, 8}. We found that λ_s dominates the performance and under the same λ_s, changing λ_p does not change the performance substantially (Supplementary Fig. 38). If the size of the predicted network for a parameter setting is smaller than the size of the simulated network, we removed this parameter setting. The ranges of λ_s and λ_p are slightly different and varying across different datasets. We used AUPR and F-score of top K edges to select the best parameter settings. We compared the algorithms using the best and all parameter settings.

Ontogenet: The Ontogenet[27] algorithm was developed to reconstruct lineage-specific regulatory networks using cell type-specific gene expression data across cell lineages. It was developed for bulk transcriptomic data. To infer the regulatory networks for each cell type, Ontogenet uses a fused LASSO framework combined with an additional L_2 penalty. The L_1 penalty is introduced to control the sparsity of regulators, while the L_2 penalty is used to select correlated predictors. The multi-task learning uses a fused LASSO framework with an additional L_1 penalty on the difference of the regression weight of related cell types, which encourage the consistency of regulatory programs between related cell types. The Ontogenet algorithm has three parameters: the L_1 penalty λ that controls the sparsity of the network, the L_2 penalty κ that handles correlated predictors, and γ that encourages the similarity of regulatory programs between related cell types. We tested different parameter configurations of λ, γ and κ.

For data 1 (n = 2000), λ were set to {1000, 1250, 1500, 1750, 2000, 2250, 2500}, and γ were set to {1000, 1250, 1500, 1750, 2000, 2250, 2500}. For data 2 (n = 1000), λ were set to {500, 1000, 2000, 3000}, and γ were set to {500, 1000, 2000, 3000}. For data 3 (n =200), λ were set to {475, 500, 525}, and γ were set to {475, 500, 525}. κ was set to {1, 5, 10} for each of the datasets. We found that λ and γ dominate the performance, while changing κ does not change the performance significantly (Supplementary Fig. 39). If the size of the predicted network for a parameter setting is smaller than the size of the simulated network, we removed this parameter setting. The ranges of λ and γ are slightly different and vary across different datasets in order to infer similarly sized networks for different datasets. We used AUPR and F-score of top K edges to select the best parameter settings. We compared the algorithms using the best and all parameter settings.

AMuSR: The Inferelator-AMuSR[28] algorithm uses sparse block-sparse regression to estimate the activities of transcription factors and infer gene regulatory networks from expression datasets. The multi-task learning approach decomposes the model coefficients matrix into a dataset-specific component using a sparse penalty and a conserved component using a block-sparse penalty to capture both conserved interactions and dataset-unique interactions. It is able to incorporate prior knowledge from multiple resources and robust to false interactions in the prior network. For our simulation setting, we applied AMuSR without TFA estimation by setting worker.set_tfa(tfa_driver = False) in the SingleCellWorkflow from Inferelator 3.0 package. To be comparable across different algorithms, AMuSR was applied on the same subsample of the three simulation datasets within a stability selection framework

to estimate the confidence score for each edge in the AMuSR networks. The AMuSR algorithm has two sparsity parameters: λ_s that controls the sparsity of the network for each dataset, the block-sparse penalty λ_b that controls the sparsity of the conserved network across all datasets. AMuSR has its own parameter selection framework (see ref. 28 for details) and uses extended Bayesian information criterion (EBIC) to select the optimal (λ_s, λ_b). We additionally externally tuned the parameters by setting c to {0.01, 0.02154435, 0.04641589, 0.1, 0.21544347, 0.46415888, 1, 2.15443469, 4.64158883,10} and set $\lambda_b = c * \sqrt{\frac{d*\log(p)}{n}}$ as suggested in the paper, where d is the number of cell types, n is the number of samples and p is the number of genes. However, by setting λ_b to 0 and λ_s to 0 (the lowest sparsity settings), we found that the inferred networks are too sparse with 7–100 edges for data 1, and 71–129 edges for data 2. We kept two settings for AMuSR, one using our criteria to select the best setting based on AUPR and F-scores among different c settings (AMuSR_tuned) and another version using AMuSR's default optimal parameter selection (AMuSR_default). We computed AUPR and F-score of top K edges (where K is the number of edges in the simulated network) for AMuSR inferred networks with optimal parameter settings for comparison with other algorithms. We compared the algorithms using the optimal and all parameter settings.

INDEP: The INDEP algorithm is the single-task framework of scMTNI which does not have the prior for sharing information across cell types and infers a regulatory network for each cell type independently. Similar to scMTNI, it models each network using a dependency network. INDEP learns the graphs for each cell type using a greedy graph learning algorithm with a score-based search, where the score contains

only the data likelihood. At each iteration, the algorithm computes the change in data likelihood score[22] for all candidate regulators for each target gene, selects the best regulator for the target gene and adds this (regulator, target) edge to the current graph. INDEP has two parameters in the model: a sparsity penalty β_0 that controls penalty for adding edges, and a penalty β_1 that controls the strength of motif prior. In the simulation case, β_0 were set to {−0.005, −0.01, −0.05, −0.1, −0.5, −1}, and β_1 were set to 0. AUPR and F-score of top K edges were used to select the best parameter settings (Supplementary Fig. 40). If the size of the predicted network for a parameter setting is smaller than the size of the simulated network, we removed this parameter setting. As mentioned above, we compared INDEP to other algorithms using best and all parameter settings for a dataset.

LASSO: The LASSO method uses linear regression with L_1 regularization. For each gene, we use the expression profiles of candidate regulators to predict the expression profiles of this gene. The regulators with non-zero coefficients are inferred as the regulators for this gene and these edges are added to the gene regulatory network. We used MATLAB implementation of LASSO regression. Similar to scMTNI, GNAT, INDEP, Ontogenet, AMuSR, LASSO was run on the same subsample of the three simulation datasets within a stability selection framework to estimate the confidence score for each edge in the networks. LASSO has only the L_1 penalty λ that controls the sparsity of the network. In the simulation case, λ were set to {0.01, 0.02, 0.03, 0.04, 0.05, 0.06}. AUPR and F-score of top K edges were used to select the best parameter settings (Supplementary Fig. 41). If the size of the predicted network for a parameter setting is smaller than the size of the simulated network, we removed this

parameter setting. We compared LASSO to other algorithms using the best and all parameter settings.

SCENIC: The SCENIC[30] algorithm uses GENIE3 or GRNBoost2 to infer TF-target relationships available as part of the Arboreto framework[79]. We used the GRNBoost2 algorithm with default parameters for network inference. SCENIC is based on an ensemble of trees with its own bootstrapping and hence was directly applied to each cell type-specific dataset in the simulation. SCENIC uses the feature importance score of each edge to rank the edges in the inferred network. We computed AUPR and F-score of top K edges (where K is the number of edges in the simulated network) for SCENIC inferred networks for comparison with other algorithms.

Application of network inference algorithms to cellular reprogramming data

We applied scMTNI, scMTNI+Prior, INDEP, INDEP+Prior, SCENIC, and CellOracle to the cellular reprogramming data, which contains 12,216 genes and 2036 potential regulators (Table 2). All of these methods require list of regulators and target genes information provided as input, and the same information is provided to all methods under comparison. The CellOracle algorithm is a new method that can integrate scRNA-seq profiles with non-transcriptomic data (such as bulk ATAC-seq and scATAC-seq profiles) to infer cell type-specific GRNs[21]. The algorithm is based on a regularized linear regression model and implemented in a Bayesian Ridge or Bagging Ridge framework to improve stability and reproducibility. CellOracle uses scATAC-seq data or bulk ATAC-seq data to identify accessible promoters and enhancers, and then scans TF motifs to construct a context-independent “base GRN”. Subsequently, for each context,

CellOracle assigns edge weights to the edges of the base GRN with the help of the context-specific scRNA-seq profiles. To infer the edge weights, CellOracle builds a regularized linear regression model to predict the expression of target gene using expression of candidate regulators. The inferred GRNs are context-specific weighted directed graphs with regression coefficients corresponding to the strength of the connections.

scMTNI and INDEP algorithms were applied within a stability selection framework to estimate edge confidence. In the stability selection framework, we subsampled the data 50 times, each with 12,216 genes and 2/3 of the cells, applied the algorithms to each subsample and used the inferred networks to estimate the confidence score for each TF-target edge in the predicted networks. In both scMTNI and scMTNI+Prior, we used the following hyper-parameter settings for the lineage structure prior $p_r = 0.2$, $p_g^d = 0.2$ and $p_g^d = 0.8$. For the sparsity prior we set $\beta_0 = -0.9$ for scMTNI, and $\beta_0 \in \{-0.9, -2, -3, -4\}$ for scMTNI+Prior. To generate the prior network, we used the matched scATAC-seq clusters to obtain TF-target prior interactions for each scRNA-seq cluster. For scMTNI+Prior which uses the scATAC-seq prior, we set $\beta 1 \in \{2, 4\}$. INDEP and INDEP+Prior were applied on the same subsampled data followed by edge confidence estimation. We used the same settings for β_0 and β_1 for INDEP as scMTNI. Final results of scMTNI+Prior used $\beta_0 = -4$ and $\beta_1 = 4$, which was determined by the distribution of edges at different confidences. Final results for INDEP +Prior used $\beta_0 = -4$ and $\beta_1 = 4$. scMTNI and INDEP were run in parallel by splitting the target gene set into subsets, e.g., of 50 genes while keeping the regulator list and other settings the same. The inferred networks of each subset target genes were

concatenated as the final inferred network. The average runtime and memory usage of scMTNI and scMTNI+Prior for this dataset are reported in Supplementary Table 2. SCENIC has its own subsampling framework which can estimate an edge importance, and was applied to the entire dataset with default parameter settings. CellOracle was applied using the Bagging Ridge regression model, which has its own bootstrapping to estimate edge importance. CellOracle was applied to the entire dataset with default parameter settings and the same prior networks as for INDEP+Prior and scMTNI+Prior to enable a fair comparison of their GRN inference capabilities.

Application of network inference algorithms to human adult hematopoietic differentiation data

We used a similar workflow for the human hematopoietic differentiation dataset as the reprogramming system. This dataset had 11,994 genes and 1999 potential regulators (Table 2). We subsampled the scRNA-seq data for each cell cluster 50 times, each with 11,994 genes and 2 3 of the cells, and applied scMTNI, scMTNI+Prior, INDEP, INDEP+Prior on each subsample to estimate the edge confidence of the GRNs. For scMTNI and scMTNI+Prior, the lineage structure prior parameters were set as follows: $p_r = 0.2$, $p_g^d = 0.2$ and $p_g^d = 0.8$. The sparsity prior β_0 was set to −0.9 for scMTNI. For scMTNI+Prior, the sparsity prior was set $\beta_0 \in \{-0.9, -2, -3, -4\}$ and $\beta 1 \in \{2, 4\}$. For INDEP and INDEP+Prior, we used the same settings for β_0 and β_1 as scMTNI and scMTNI+Prior respectively. Final results of scMTNI+Prior are with $\beta 0 = -4$ and $\beta_1 = 4$ and final results for INDEP+Prior are using $\beta_0 = -4$ and $\beta_1 = 4$. The runtime and memory usage of scMTNI and scMTNI+Prior for this dataset are reported

Supplementary Table 2. SCENIC was applied to the entire dataset with default parameter settings. CellOracle was applied to the entire dataset with default parameter settings using the same prior networks as for scMTNI+Prior and INDEP+Prior. The same list of regulators and target genes are provided to all methods under comparison.

Application of network inference algorithms to human fetal hematopoiesis data

We applied scMTNI, scMTNI+Prior, INDEP, INDEP+Prior and CellOracle to the fine-grained lineage version of this dataset using a similar workflow as the other datasets. We applied scMTNI+Prior and CellOracle to this dataset when using the coarse lineage structure. For the fine-grained lineage, there are 16,737 genes and 2195 potential regulators. For the coarse lineage, there are 17,425 genes and 2227 potential regulators (Table2). We subsampled the scRNA-seq data for each cell cluster 50 times, each with all genes and 2 3 of the cells, and applied scMTNI, scMTNI+Prior, INDEP, INDEP+Prior on each subsample to estimate the edge confidence of the GRNs. For scMTNI and scMTNI+Prior, the lineage structure prior parameters were set as follows: $p_r = 0.2$, $p_g^d = 0.2$ and $p_g^d = 0.8$. The sparsity prior β_0 was set to −0.9 for scMTNI. Final results of scMTNI+Prior are with β_0 = −4 and β_1 =4 and final results for INDEP+Prior are using β_0 = −4 and β_1 =4.INDEPand INDEP+Prior used the same settings for β_0 and β_1 for as scMTNI and scMTNI+Prior, respectively. The runtime performance and memory usage of scMTNI and scMTNI+Prior are reported in Supplementary Table 2. CellOracle was applied to the entire dataset with default parameter settings with the same prior networks as scMTNI+Prior and INDEP+Prior. The same list of regulators and target genes are provided to all methods under comparison.

Evaluation

Gold standard datasets

To evaluate the predicted networks of different inference algorithms on real data, we downloaded and processed several gold standard datasets (Table 1). For mouse reprogramming study, we curated multiple experimentally derived networks of regulatory interactions from the literature and existing databases. The statistics of the gold standard datasets are provided in Table 1. One of these datasets is ChIP-chip or ChIP-seq based gold standard (referred to as “ChIP”) from ESCAPE (http://www.maayanlab. net/ESCAPE/) or ENCODE databases[34,35] (https://www.encodeproject. org/), which contains ChIP-chip or ChIP-seq experiments in mouse ESCs. The second dataset is a knock down-based gold standard (referred to as “Perturb”), which is derived from regulator perturbation followed by global transcriptome profiling[34,36]. We took a union of the networks from LOGOF (loss or gain of function) based gold standard networks from the ESCAPE database[34] and the networks from Nishiyama et al.[36] as the perturbation interactions. Finally, we took the intersection of the interactions between ChIP and knock-down based gold standards to create the third gold standard network referred to as “ChIP+Perturb”.

For human hematopoietic cell types, we have five gold standard datasets. Two gold standard datasets were a ChIP-based (Cus_ChIP) and a regulator knock down-based (Cus_KO) dataset in GM12878 lymphoblastoid cell line downloaded from Cusanovich et al.[47].Forthe knock down dataset, we had TF-target relationships at two p-value thresholds, 0.01 and 0.05. We used the TF-target relationships at 0.01 to have a

more stringent gold standard. The third gold standard was from human hematopoietic cell types from the UniBind database (https://unibind.uio.no/)[46], which has high confidence TF binding site predictions from ChIP-seq experiments. To obtain the TF-gene network, we mapped TF binding sites to the nearest gene if there is overlap between the TF binding sites and the promoter of the gene defined by ±5000 bp of the gene TSS. If multiple ChIP-seq datasets were available for the same TF in a given cell type, we took the union of TF-gene edges for the same cell type. We took the union of these individual cell type-specific gold standards to create our Unibind gold standard (UniBind). Finally, we took the intersection of the ChIP-based gold standards with the knock down based gold standards, to produce the fourth and fifth gold standards, Unibind+Cus_KO and CusChIP+Cus_KO. The statistics of the gold standard datasets are provided in Table 1.

Area under the precision recall curve

To evaluate the performance of scMTNI and other algorithms, we compared the inferred networks to the simulated networks or interactions from the gold standard datasets based on Area under the precision recall curve (AUPR). Edge weights for all but the SCENIC and CellOracle algorithms were obtained using stability selection. Both SCENIC and CellOracle have internal bootstrapping or bagging approaches to estimate confidence in the inferred edges. In our stability selection framework, we generated N random subsamples of the data, inferred a network for each subsample, and calculated a confidence score for each edge as the fraction of how many times this edge was present in the inferred networks across all subsamples. Next, we ranked the edges by

the confidence score and estimated precision and recall as a function of edge confidence. Precision P is defined as the fraction of the number of edges that are true positives among the total number of predicted edges. Recall R is defined as the fraction of the number of edges that are true positives among the total number of true edges. Then, we plotted the precision recall curve and estimated the area under this curve using the AUC Calculator package developed by Davis et al.[80]. The area under the precision recall curve is computed as an overall assessment of the inferred networks compared to "true" networks. The higher AUPR, the better the performance. For the real scRNA-seq datasets, we filtered the inferred networks to include TFs and targets that were in the gold standard.

F-score

While AUPR uses a ranking of the edges, F-score is a metric to compare a set of predicted edges to a set of "true" edges. F-score is defined as the harmonic mean of the precision (P) and recall (R), F-score = (2*P*R)/(P+R). F-score enables us to control for the number of edges across network inference algorithms as these can vary significantly across algorithms. To control for number of edges in the predicted networks, we ranked the predicted network by the confidence score or edge weight, selected top K edges and computed F-score compared to simulated networks or gold standard networks. K in the simulated datasets corresponded to the size of the simulated networks. For the real datasets, we considered top 500, 1000, 2000 edges. We obtained the top K edges after filtering the inferred networks based on the TFs and

targets in the gold standard networks. The higher the F-score, the better the performance.

Predictable transcription factors (TFs)

Predictable TFs was defined based on the gold standard datasets similar to McCalla et al.[18]. For each TF's target set in the gold standard network, we computed its overlap with the predicted targets in the inferred network and used the hypergeometric test to assess the significance of overlap. We consider a TF to be predictable if the P-value < 0.05. We count the total number of predictable TFs for each algorithm as a metric of evaluation. The higher the number of predictable TFs, the better the performance.

Examining network dynamics on cell lineages

We used several global and subnetwork-level methods to examine how regulatory networks change on a cell lineage. These include F-score based comparison of all pairs of networks on the lineage, k-means based edge clustering and Latent Dirichlet Allocation (LDA) model.

F-score based analysis of inferred network change along cell lineage tree

To examine the overall conservation and divergence between the inferred cell type-specific networks along the cell lineage tree, we computed F-score on the predicted networks between each pair of cell types and applied hierarchical clustering on the inferred networks based on the F-score. To compute F-score, we selected top X edges ranked by confidence score to obtain a reliable network for each cell type. This was 4k

in the mouse reprogramming dataset and 5k for the hematopoietic differentiation datasets. We visualized the dendrogram obtained from the hierarchical clustering and compared this to the original cell lineage tree.

Latent Dirichlet Allocation (LDA) model for regulatory network rewiring

We adopted Latent Dirichlet Allocation (LDA) to examine subnetwork level rewiring as described in TopicNet[40]. LDA was originally developed to cluster documents based on their word distributions. Each document, i is assumed to have a certain composition of topics, as captured by a θ_i parameter and each topic, k, is assumed to have a specific distribution of words denoted by a φ_k parameter. In the application of LDA to a regulatory network, we first concatenated the TF by target network across cell types to have as many rows as there are TFs times the number of cell types. Each TF in a cell type is treated as a document and its targets are treated as words in the document. The topic distribution for all documents constitutes a M × K matrix for document-topic distribution, where M is the total number of TFs in any of the networks and K is the total number of topics. The distribution of words (genes) in each topic is captured by a K × V matrix for V genes. Each gene can be assigned to a topic based on its maximum probability across topics. We applied LDA to the 80% confidence networks of all cell clusters or types inferred from scMTNI+Prior with 10 or 15 topics and found 10 topics to be suitable for all three datasets. We extracted the subnetworks in each cell type associated with each topic by obtaining the induced graph for the genes and regulators associated with each topic and visualized the giant components of each network to identify change across cell clusters within the same topic. To interpret the topics in each

cell type, we tested the genes in the cell type-specific subnetwork for each topic for enrichment of gene ontology (GO)[81] processes using a hypergeometric test with FDR correction. We define the gene set for each topic to include the cell-type specific regulators and targets per cell type. We used an FDR < 0.01 to determine significant enrichment (Supplementary Figs. 11, 21, 29). These results are described in Supplementary Figs. 8-10 for mouse cellular reprogramming, in Supplementary Figs. 18–20 for the hematopoietic differentiation data from Buenrostro et al., in Supplementary Figs. 25–28 for the fetal hematopoiesis fine-grained lineage and in Supplementary Figs. 31–35 for the fetal hematopoiesis coarse lineage data.

Statistics and reproducibility

In the scATAC-seq reprogramming experiment, six samples representing different time points of the reprogramming study were used. The sample size is the number of biological samples. We chose six samples to analyze because these specific timepoints, along with MEFs and ESCs, provide sufficient coverage on the various states and progression of cells during the reprogramming process. One biological replicate for each sample data was used for analysis. Previous experiments were conducted in which cells were reprogrammed using identical conditions and reagents (see Tran et al.32). The setup of experiments in this paper assume that one experimental replicate and one scATAC-seq submission for each sample reflects the same reprogramming time course observed in our previous experiments. For randomization, MEFs from a single embryo were randomly seeded at a density of 5000 cells per well in 6-well plates. Blinding was not applicable to this study as no portion of

this data can be skewed based on participant's knowledge of the experiment. All cells from the reprogramming plates were collected during scATAC-seq submission and the scATAC library prep and sequencing portions were performed by unbiased third parties who have no knowledge of any experimental details.

Network inference was done in a stability selection mode where we drew multiple subsamples from the original data. Each subsample's size was set to 2/3 of the number of cells in the dataset. This number was determined to enable sufficient number of cells for each subsample. Subsamples were generated by selecting uniformly at random samples from our full dataset. We have provided code, scripts, inputs and outputs from our experiments to enable replication of our study. For data exclusion, cells with low read depth and genes with fewer than 5 or 20 measurements were filtered from downstream analysis. Some cell clusters were excluded if they had either no or too few scRNA-seq cells. Cluster C1 for the hematopoietic differentiation data from Buenrostro et al. was removed from evaluation using the gold standards due to very few TFs overlapping the gold standards compared to the other cell clusters.

Data Availability

The reprogramming scATAC-seq dataset generated in this study has been deposited to Gene Expression Omnibus (GEO) with accession ID GSE208620. The scRNA-seq datasets for the same time points from Tran et al.[32] were downloaded from Gene Expression Omnibus (GEO) with accession ID GSE108222. The processed cluster-specific scRNA-seq matrices and the prior networks for reprogramming study are available at Zenodo https://zenodo.org/record/7879228[82].

The scRNA-seq data for human hematopoietic differentiation from Buenrostro et al. were downloaded from Data S2 of Buenrostro et al. (https://ars.els-cdn.com/content/image/1-s2.0-S009286741830446Xmmc4.zip) and the scATAC-seq data were downloaded from Chen et al.[72] (https://github.com/pinellolab/scATAC-benchmarking/tree/ master/Real_Data/Buenrostro_2018). The scATAC-seq data are also available from GEO accession GSE96772.ThescRNA-seqdata(DataS2 from Buenrostro et al.,) and the scATAC-seq data have been additionally uploaded to Zenodo https://zenodo.org/record/7879228. The processed datasets for human hematopoietic differentiation are available at Zenodo https://zenodo.org/record/7879228.

The scRNA-seq (gene by cell) and scATAC-seq (peak by cell) data matrices for the human fetal hematopoietic differentiation data from Ranzoni et al. were obtained from https://gitlab.com/cvejic-group/ integrative-scrna-scatac-human-foetal. These are also available at ArrayExpress: E-MTAB-9067 for scRNA-seq and E-MTAB-9068 for scATAC-seq. The cluster-specific scRNA-seq matrices and the prior networks are available at Zenodo https://zenodo.org/record/7879228.

For the mouse reprogramming study, the ChIP-based gold standard datasets were downloaded from ESCAPE (http://www.maayanlab. net/ESCAPE/) and ENCODE databases[34,35] (https://www.encodeproject. org/). The Perturbation-based gold standard networks were constructed from a union of the networks from LOGOF (loss or gain of function) based gold standard networks from ESCAPE database[34] and the networks from Nishiyama et al.[36]. The mouse gold standard datasets are available at Zenodo https://zenodo.org/record/7879228.

For the human hematopoietic data, two gold standard datasets were a ChIP-based (Cus_ChIP) and a regulator knock down-based (Cus_KO) dataset in GM12878 lymphoblastoid cell line downloaded from Cusanovich et al.[47]. The third gold standard from ChIP-seq experiments in human hematopoietic cell types was downloaded from the UniBind database (https://unibind.uio.no/)[46].Thehumangold standard datasets are available at Zenodo https://zenodo.org/record/ 7879228.

The source data underlying Figs. 2–8, Supplementary Figs. 2, 3, 5, 7–10, 12, 14, 15, 17–20, 22, 24–28, 20–29, 30–49, the cluster-specific scRNA-seq matrices and the prior networks for all datasets and scMTNI inferred consensus networks are available at Zenodo https://zenodo. org/record/7879228[82]. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. Source data are provided with this paper.

Code Availability

The scMTNI code and custom scripts to process data and compute various validation metrics and perform dynamic network analysis are available at https://github.com/Roy-lab/scMTNI and Zenodo https:// doi.org/10.5281/zenodo.785453583. Custom scripts include shell scripts, python scripts, R scripts and MATLAB scripts and we used R version 3.5.1, MATLAB version R2014b, and Python version 3.6.12 to perform data analysis. The scATAC-seq data was processed through CellRanger ATAC pipeline (Version 1.1.0). The simplified implementation of the pagoda pipeline for normalizing scRNA-seq data for depth and variance stabilization is available at

https://github.com/Roy-lab/scMTNI/blob/master/Scripts/Integration/adjustVariance_depth_ Generic.R. R package rliger version 1.0.0 was used to integrate scRNAseq and scATAC-seq data, and the R script is available at https://github.com/Roy-lab/scMTNI/tree/master/Scripts/Integration/. To generate prior networks, we used MACS v2.1.0 to call ATAC-seq peaks to generate prior networks and used custom code for mapping TF binding peaks to genes, which is available at https://github.com/Roy-lab/ scMTNI/tree/master/Scripts/genPriorNetwork/. The scripts for evaluation based on AUPR and F-score are available at https://github.com/ Roy-lab/scMTNI/tree/master/Evaluation/. The scripts for dynamic network analysis are available at https://github.com/Roy-lab/scMTNI/ tree/master/Scripts/Network_Analysis/.

Figure 1

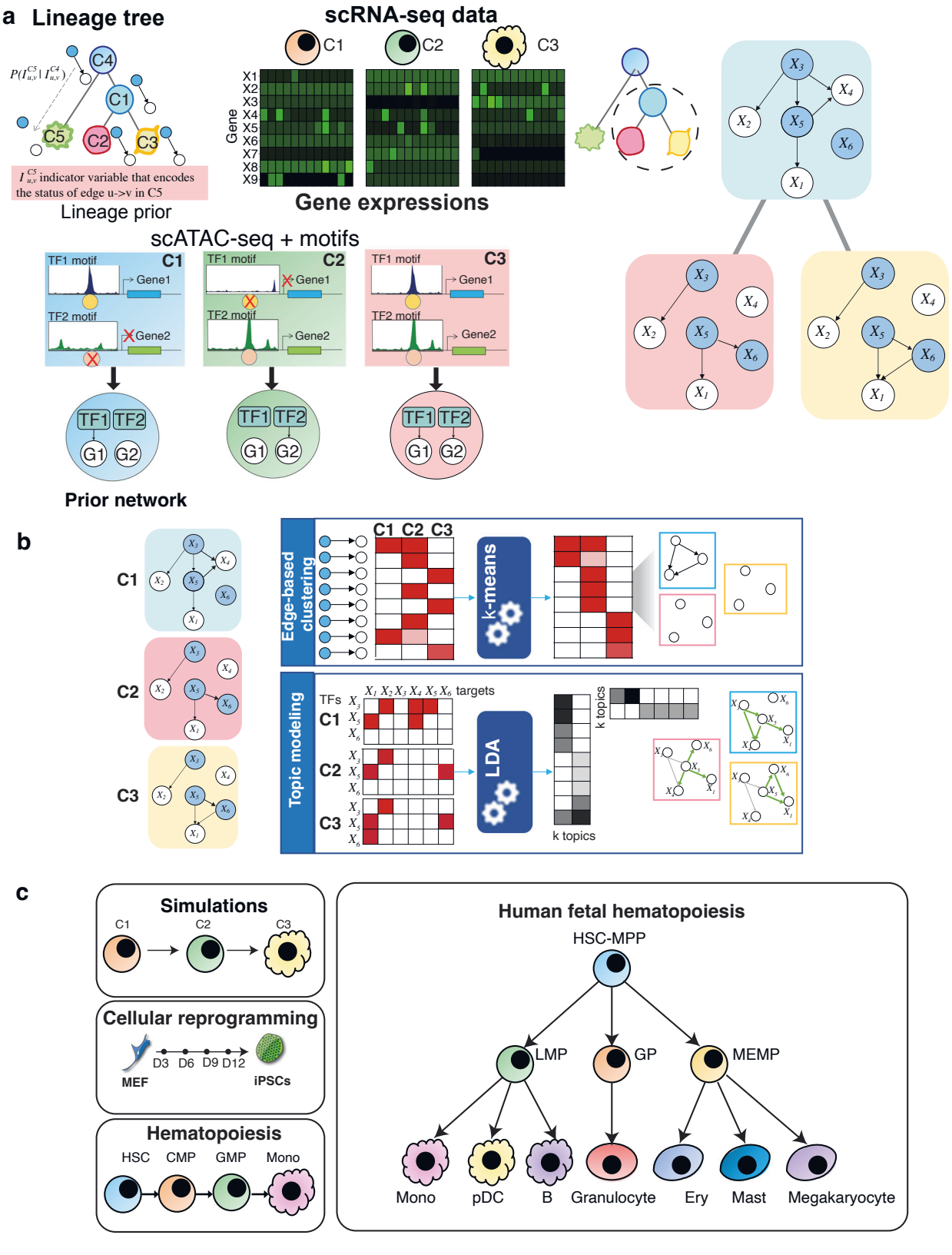

a Lineage tree
scRNA-seq data
C1
C2
C3
C4
C5
Gene
X1
X2
X3
X4
X5
X6
X7
X8
X9
Gene expressions
Lineage prior
scATAC-seq + motifs
TF1 motif
TF2 motif
Gene1
Gene2
TF1
TF2
G1
G2
Prior network
b
Edge-based clustering
k-means
Topic modeling
LDA
TFs
targets
k topics
c
Simulations
Cellular reprogramming
D3 D6 D9 D12
MEF
iPSCs
Hematopoiesis
HSC CMP GMP Mono
Human fetal hematopoiesis
HSC-MPP
LMP
GP
MEMP
Mono
pDC
B
Granulocyte
Ery
Mast
Megakaryocyte

Fig. 1 | An overview of the scMTNI framework

a. scMTNI takes as input a cell lineage tree and cell type-specific scRNA-seq data and cell type-specific prior networks derived from scATAC-seq datasets. If scATAC-seq data is not available, bulk or sequence-based prior networks can be used for the cell types. The output of scMTNI is a set of cell type-specific gene regulatory networks for each cell type on the cell lineage tree.

b. The output networks of scMTNI are analyzed using two dynamic network analysis methods: edge-based k-means clustering and Latent Dirichlet Allocation (LDA) based topic models to identify key regulators and subnetworks associated with a particular cell cluster or a set of clusters on a branch.

c. Datasets used with scMTNI. The simulation data comprised a linear trajectory of three cell types, while the three real datasets came from a reprogramming timeseries process, immunophenotypic cell types identified during human adult hematopoietic differentiation, and immunophenotypic blood cells during human fetal hematopoiesis. MEF mouse embryonic fibroblast, iPSCs induced pluripotent cells, HSC hematopoietic stem cell, CMP common myeloid progenitor, GMP granulocyte-macrophage progenitors, Mono monocyte, HSC-MPP hematopoietic stem cells and multipotent progenitors, LMP lymphoid-myeloid progenitors, MEMP MK-erythroid-mast progenitors combined with cycling MEMPs, GP granulocytic progenitors, Ery erythroid cells, pDC plasmacytoid dendritic cells.

Figure 2

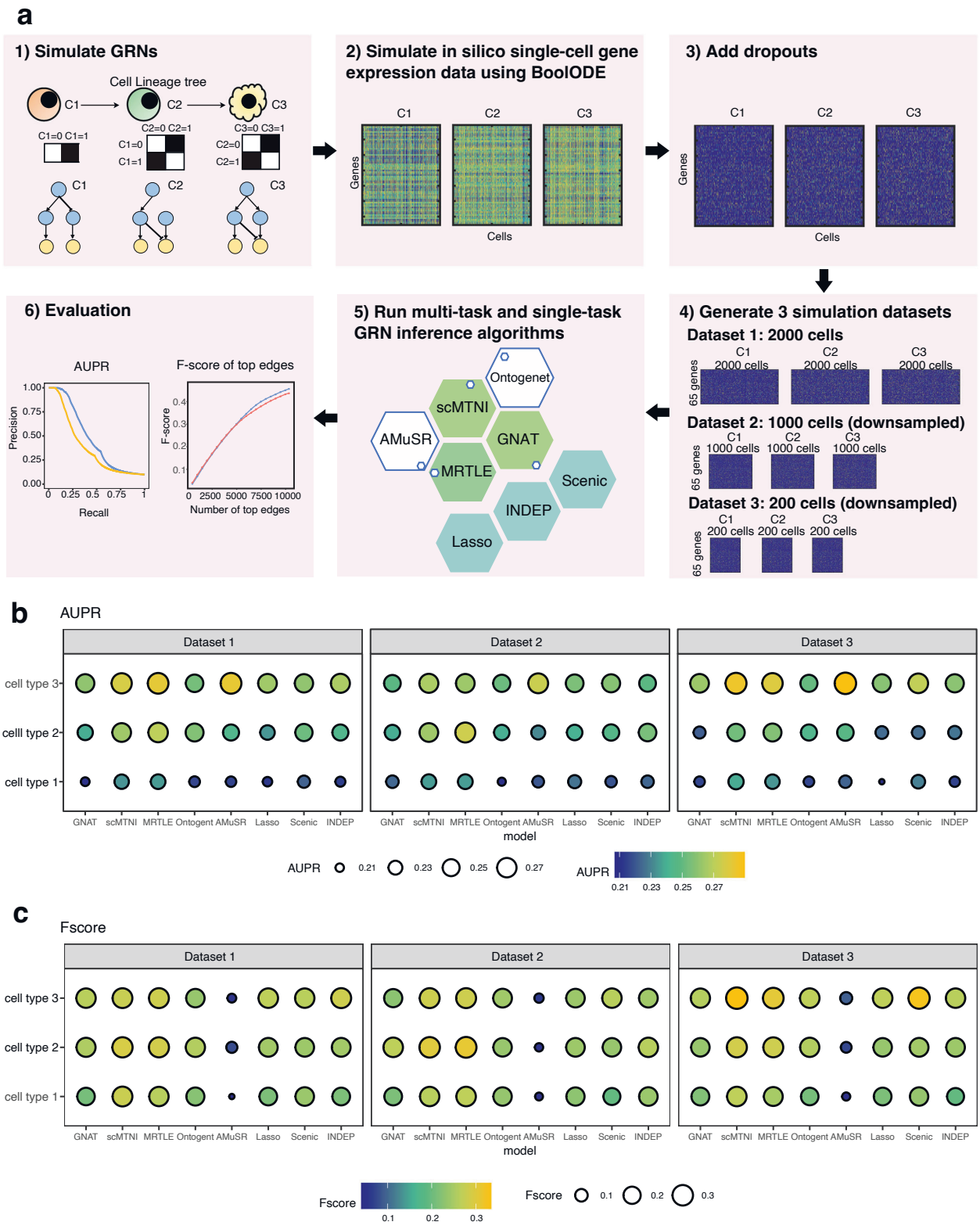

a
1) Simulate GRNs
Cell Lineage tree
C1
C2
C3
C1=0 C1=1
C2=0 C2=1
C3=0 C3=1
2) Simulate in silico single-cell gene expression data using BoolODE
Genes
Cells
3) Add dropouts
4) Generate 3 simulation datasets
Dataset 1: 2000 cells
2000 cells
65 genes
Dataset 2: 1000 cells (downsampled)
1000 cells
Dataset 3: 200 cells (downsampled)
200 cells
5) Run multi-task and single-task GRN inference algorithms
Ontogenet
scMTNI
AMuSR
GNAT
MRTLE
Scenic
INDEP
Lasso
6) Evaluation
AUPR
Precision
Recall
F-score of top edges
F-score
Number of top edges
b
AUPR
Dataset 1
Dataset 2
Dataset 3
cell type 3
celll type 2
cell type 1
GNAT scMTNI MRTLE Ontogent AMuSR Lasso Scenic INDEP
model
AUPR 0.21 0.23 0.25 0.27
c
Fscore
Fscore 0.1 0.2 0.3

Fig. 2 | Benchmarking algorithms on simulated data

a. Simulation framework for scMTNI. We first simulate GRNs for cell types across a cell lineage tree. Next, we generate in silico single-cell gene expression data for each cell type using BoolODE using the simulated GRNs and add 80% zeros in the simulation data. Then, we apply five multi-task learning algorithms and three single-task learning algorithms for GRN inference to the simulated datasets and predict networks in stability selection framework. We compare the performance of these algorithms based on area under precision and recall curve (AUPR) and F-score of top edges.
b. AUPR comparing inferred networks to ground truth networks of simulated datasets 1, 2, 3.
c. F-score comparing top K edges in the inferred networks to those in the ground truth networks of simulated datasets 1, 2, 3 (cell type 1: K =202, cell type 2: K = 217, cell type 3: K = 239). The brighter and larger the circle the better the performance of the algorithm. Source data are provided as a Source Data file.

Figure 3

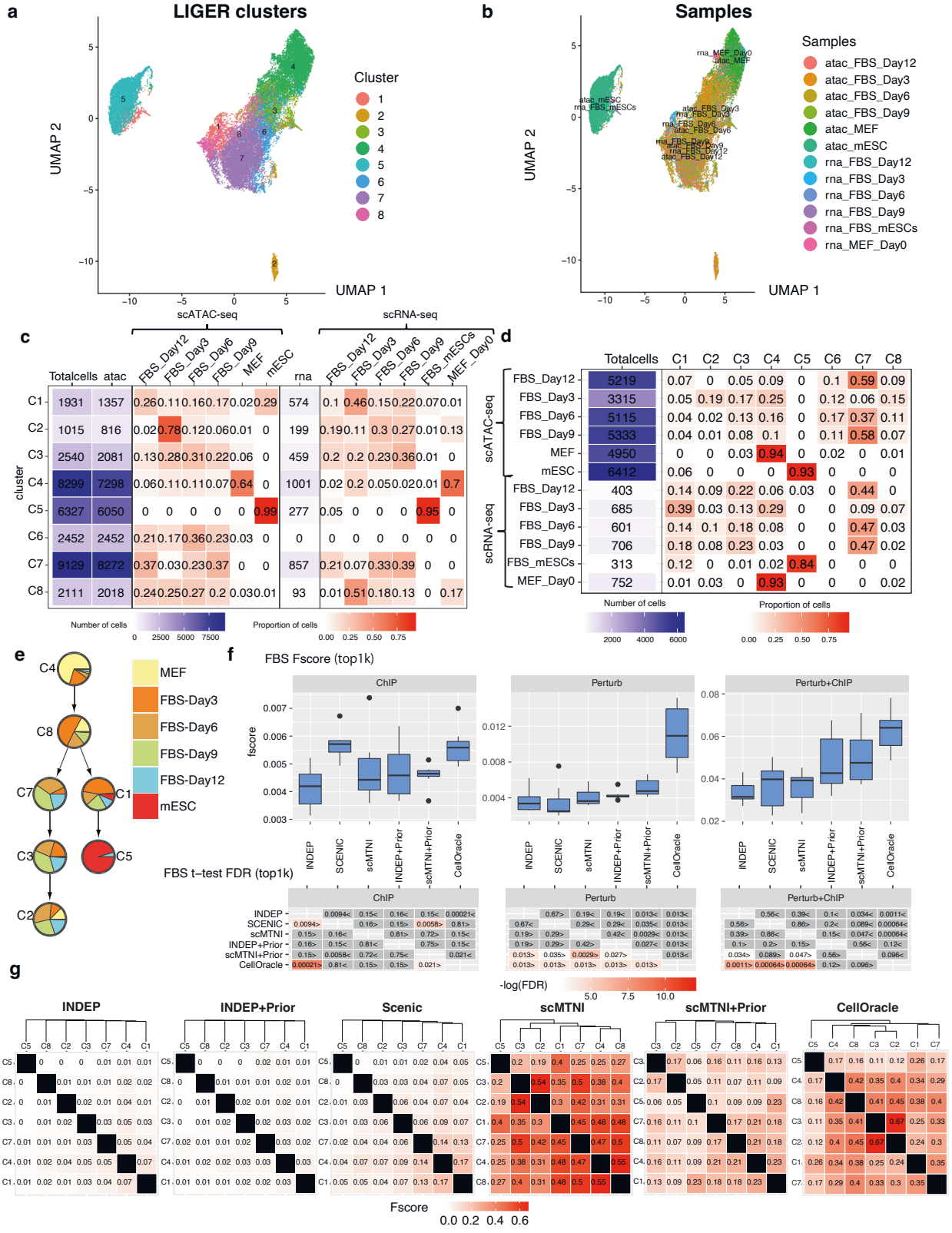

a
LIGER clusters
UMAP 2
UMAP 1
Cluster
1
2
3
4
5
6
7
8
b
Samples
UMAP 2
UMAP 1
Samples
atac_FBS_Day12
atac_FBS_Day3
atac_FBS_Day6
atac_FBS_Day9
atac_MEF
atac_mESC
rna_FBS_Day12
rna_FBS_Day3
rna_FBS_Day6
rna_FBS_Day9
rna_FBS_mESCs
rna_MEF_Day0
c
scATAC-seq
scRNA-seq
Totalcells
atac
rna
cluster
Number of cells
Proportion of cells
d
Totalcells
scATAC-seq
scRNA-seq
Number of cells
Proportion of cells
e
MEF
FBS-Day3
FBS-Day6
FBS-Day9
FBS-Day12
mESC
f
FBS Fscore (top1k)
ChIP
Perturb
Perturb+ChIP
fscore
INDEP
SCENIC
scMTNI
INDEP+Prior
scMTNI+Prior
CellOracle
FBS t−test FDR (top1k)
-log(FDR)
g
INDEP
INDEP+Prior
Scenic
scMTNI
scMTNI+Prior
CellOracle
Fscore

Fig. 3 | Inference of cell-type specific networks of mouse cellular reprogramming data

a. UMAP of LIGER cell clusters on the scATAC-seq data and scRNA-seq data.
b. UMAP depicting the sample labels of the scATAC-seq and scRNA-seq data from mouse cellular reprogramming.
c. The distribution of samples in each LIGER cluster.
d. The distribution of LIGER clusters in each sample.
e. Inferred lineage structure for scMTNI linking the 7 cell clusters with scRNA-seq measurements.
f. F-score of top 1k edges in predicted networks of scMTNI, scMTNI+Prior, INDEP, INDEP+Prior, SCENIC and CellOracle compared to three gold standard datasets: ChIP, Perturb and Perturb+ChIP. The top boxplots show the F-scores of n = 7 cell clusters, while the bottom heatmaps show FDR corrected t-test comparing the F-scores of the row algorithm to that of the column algorithm. The two-sided paired t-test is conducted on F-scores of n = 7 cell clusters for every pair of algorithms. A FDR < 0.05 was considered significantly better. The sign < or > specifies whether the row algorithm's F-scores were worse or better than the column algorithm's F-scores. The color scale is specified by −log(FDR), with the red color proportional to significance. Non-significance is colored in gray. In the boxplot, the horizontal middle line of each plot is the median. The bounds of the box are 0.25 quantile (Q1) and 0.75 quantile (Q3). The upper whisker is the minimum of the maximum value and Q3 +1.5*IQR,whereIQR = Q3 − Q1. The lower whisker is the maximum of the minimum value and Q1 − 1.5*IQR.

g. Pairwise similarity of networks from each cell cluster using F-score on the top 4k edges. Rows and columns are ordered based on the dendrogram created using the F-score similarity. Source data are provided as a Source Data file.

Figure 4

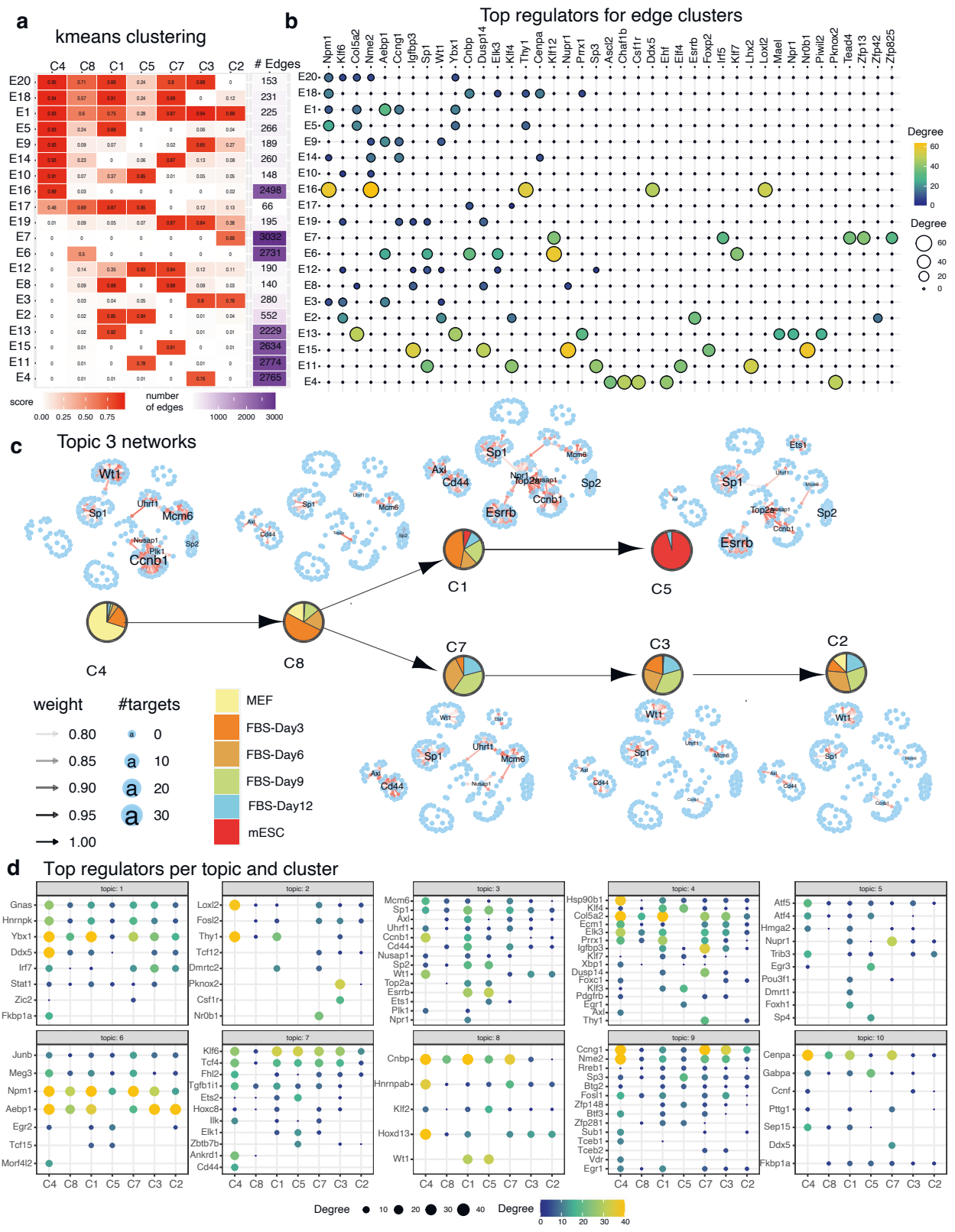

a
kmeans clustering
C4 C8 C1 C5 C7 C3 C2 # Edges
E20 153
E18 231
E1 225
E5 266
E9 189
E14 260
E10 148
E16 2498
E17 66
E19 195
E7 3032
E6 2731
E12 190
E8 140
E3 280
E2 552
E13 2229
E15 2634
E11 2774
E4 2765
score
0.00 0.25 0.50 0.75
number of edges
1000 2000 3000
b
Top regulators for edge clusters
Npm1 Klf6 Col5a2 Nme2 Aebp1 Ccng1 Igfbp3 Sp1 Wt1 Ybx1 Cnbp Dusp14 Elk3 Klf4 Thy1 Cenpa Klf12 Nupr1 Prrx1 Sp3 Asl2 Chaf1b Csf1r Ddx5 Ehf Elf4 Esrrb Foxp2 Irf5 Klf7 Lhx2 Loxl2 Mael Npr1 Nr0b1 Piwil2 Pknox2 Tead4 Zfp13 Zfp42 Zfp825
Degree
60
40
20
0
Degree
60
40
20
0
c
Topic 3 networks
Wt1 Sp1 Uhrf1 Mcm6 Nusap1 Plk1 Ccnb1 Sp2
Axl Cd44 Sp1 Mcm6 Nusap1 Top2a Ccnb1 Sp2 Esrrb
Ets1 Sp1 Top2a Ccnb1 Sp2 Esrrb
C4
C8
C1
C5
C7
C3
C2
weight
0.80
0.85
0.90
0.95
1.00
#targets
0
10
20
30
MEF
FBS-Day3
FBS-Day6
FBS-Day9
FBS-Day12
mESC
d
Top regulators per topic and cluster
topic: 1
Gnas Hnrnpk Ybx1 Ddx5 Irf7 Stat1 Zic2 Fkbp1a
topic: 2
Loxl2 Fosl2 Thy1 Tcf12 Dmrtc2 Pknox2 Csf1r Nr0b1
topic: 3
Mcm6 Sp1 Axl Uhrf1 Ccnb1 Cd44 Nusap1 Sp2 Wt1 Top2a Esrrb Ets1 Plk1 Npr1
topic: 4
Hsp90b1 Klf4 Col5a2 Ecm1 Elk3 Prrx1 Igfbp3 Klf7 Xbp1 Dusp14 Foxc1 Klf3 Pdgfrb Egr1 Axl Thy1
topic: 5
Atf5 Atf4 Hmga2 Nupr1 Trib3 Egr3 Pou3f1 Dmrt1 Foxh1 Sp4
topic: 6
Junb Meg3 Npm1 Aebp1 Egr2 Tcf15 Morf4l2
topic: 7
Klf6 Tcf4 Fhl2 Tgfb1i1 Ets2 Hoxc8 Ilk Elk1 Zbtb7b Ankrd1 Cd44
topic: 8
Cnbp Hnrnpab Klf2 Hoxd13 Wt1
topic: 9
Ccng1 Nme2 Rreb1 Sp3 Btg2 Fosl1 Zfp148 Btf3 Zfp281 Sub1 Tceb1 Tceb2 Vdr Egr1
topic: 10
Cenpa Gabpa Ccnf Pttg1 Sep15 Ddx5 Fkbp1a
C4 C8 C1 C5 C7 C3 C2
Degree
10 20 30 40
Degree
0 10 20 30 40

Fig. 4 | Network dynamics analysis of GRNs from cellular reprogramming

a. k-means clustering analysis of top 4k edges in inferred networks. Shown are the mean profiles of edge confidence of 20 edge clusters. Each row corresponds to an edge cluster and each column corresponds to a cell cluster. The red intensity corresponds to the average confidence of edges in that cluster. Shown also are the number of edges in the edge cluster.
b. Top 5 regulators for each edge cluster. Shown are only regulators that have at least 10 targets in any edge cluster. The size and brightness of the circle is proportional to the number of targets.
c. LDA topic 3 networks along the cell lineage. The layout of each network is the same, edges present in a particular cell cluster are shown in red. Labeled nodes correspond to regulators with degree larger than 10.
d. Top cell cluster-specific regulators for each topic. Shown are only regulators that have at least 10 targets in any cell cluster. The more yellow and larger the circle, the greater are the number of targets for the regulator. Source data are provided as a Source Data file.

Figure 5

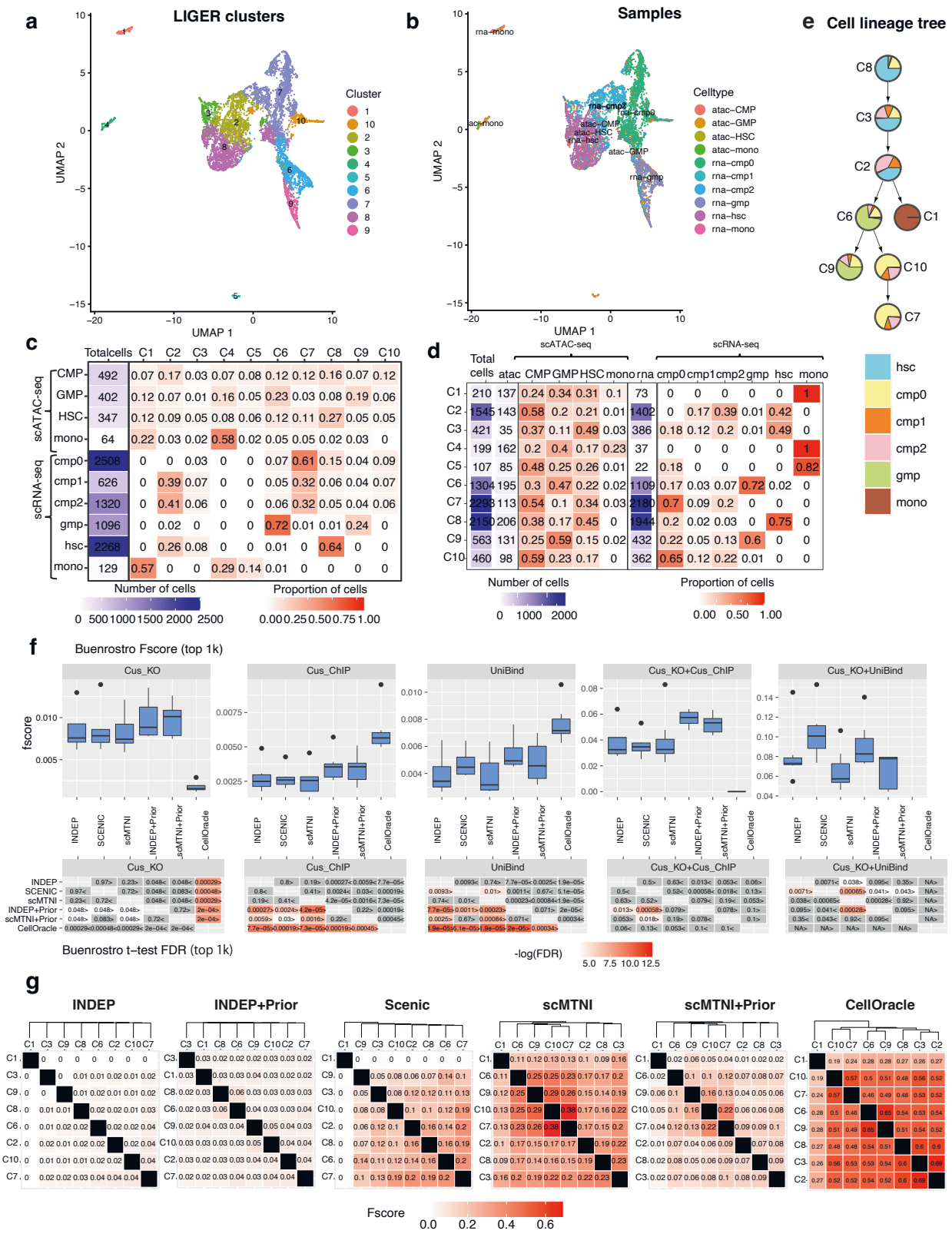

a
LIGER clusters
UMAP 1
UMAP 2
Cluster
b
Samples
Celltype
atac–CMP
atac–GMP
atac–HSC
atac–mono
rna–cmp0
rna–cmp1
rna–cmp2
rna–gmp
rna–hsc
rna–mono
e
Cell lineage tree
c
Totalcells
scATAC-seq
scRNA-seq
Number of cells
Proportion of cells
d
scATAC-seq
scRNA-seq
Total cells
Number of cells
Proportion of cells
f
Buenrostro Fscore (top 1k)
Cus_KO
Cus_ChIP
UniBind
Cus_KO+Cus_ChIP
Cus_KO+UniBind
fscore
INDEP
SCENIC
scMTNI
INDEP+Prior
scMTNI+Prior
CellOracle
Buenrostro t–test FDR (top 1k)
-log(FDR)
g
INDEP
INDEP+Prior
Scenic
scMTNI
scMTNI+Prior
CellOracle
Fscore

Fig. 5 | Inference of cell type-specific networks for human hematopoietic differentiation data

a. UMAP of LIGER cell clusters of the scATAC-seq and scRNA-seq data.
b. UMAP depicting the original cell types (samples) with scATAC-seq and scRNA-seq data.
c. The distribution of cell clusters in each sample.
d. The distribution of samples in each LIGER cluster.
e. Inferred lineage structure linking the eight cell clusters with scRNA-seq data.
f. Boxplots showing F-scores of n = 7 cell clusters (all cell clusters excluding C1) for top 1k edges in predicted networks from scMTNI, scMTNI+Prior, INDEP, INDEP+Prior, SCENIC and CellOracle compared to gold standard datasets (top). FDR-corrected t-test to compare the F-score of the row algorithm to the F-score of the column algorithm (bottom). The two-sided paired t-test is conducted on F-scores of n = 7 cell clusters for every pair of algorithms. A FDR < 0.05 was considered significantly better. The sign < or > specifies whether the row algorithm's F-scores were worse or better than the column algorithm's F-scores. The color scale is specified by − log(FDR), with the red color proportional to significance. Non-significance is colored in gray. In the boxplot, the horizontal middle line of each plot is the median. The bounds of the box are 0.25 quantile (Q1) and 0.75 quantile (Q3). The upper whisker is the minimum of the maximum value and Q3 +1.5*IQR,whereIQR = Q3 − Q1. The lower whisker is the maximum of the minimum value and Q1 − 1.5*IQR.

g. Pairwise similarity of networks from each cell cluster using F-score on the top 5k edges. Rows and columns ordered by hierarchical clustering using F-score as the similarity measure. Source data are provided as a Source Data file.

Figure 6

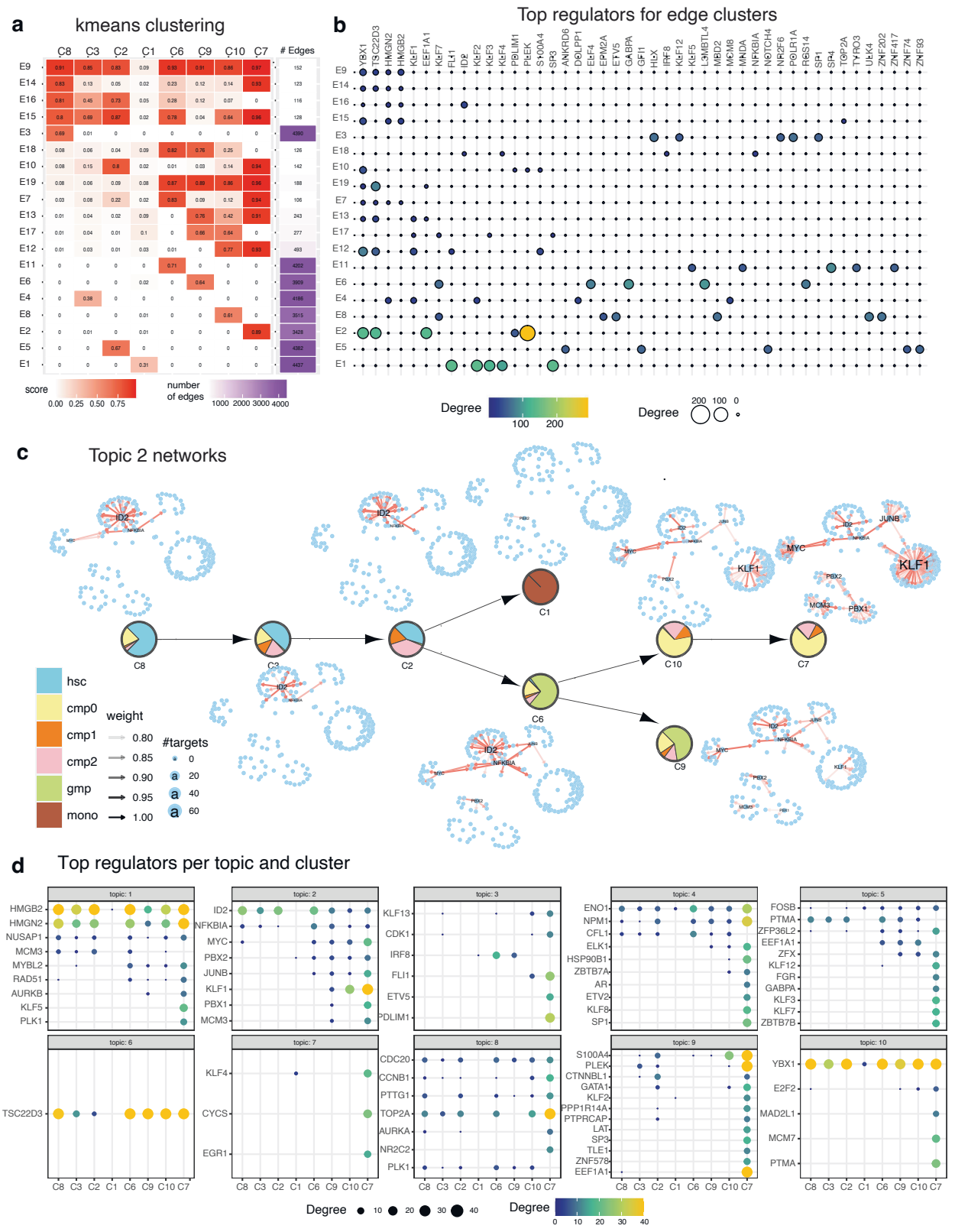

a
kmeans clustering
C8 C3 C2 C1 C6 C9 C10 C7
Edges
score
number of edges
b
Top regulators for edge clusters
Degree
Degree
c
Topic 2 networks
hsc
cmp0
cmp1
cmp2
gmp
mono
weight
#targets
C8
C3
C2
C1
C6
C10
C9
C7
d
Top regulators per topic and cluster
topic: 1
topic: 2
topic: 3
topic: 4
topic: 5
topic: 6
topic: 7
topic: 8
topic: 9
topic: 10
Degree
Degree

Fig. 6 | Network rewiring during hematopoietic differentiation

a. k-means edge clusters of the top 5k edges (rows) across 8 cell clusters (columns). The edge confidence matrix was clustered into 19 clusters to identify common and divergent networks. The red intensity corresponds to the average confidence of edges in that cluster. Shown also are the number of edges in the edge cluster.
b. Top 5 regulators of each edge cluster. Shown are only regulators with at least 10 targets in a given edge cluster. The size and brightness (yellow) of the circle is proportional to the number of targets.
c. Topic-specific networks across each cell cluster for topic 2. The layout of each network is the same, edges present in a particular cell cluster are shown in red. Labeled nodes correspond to regulators with degree larger than 10.
d. Top regulators associated with each cell cluster's network in each topic. Shown are only regulators that have at least 10 targets in any cell cluster. The more yellow and larger the circle, the greater are the number of targets for the regulator. Source data are provided as a Source Data file.

Figure 7

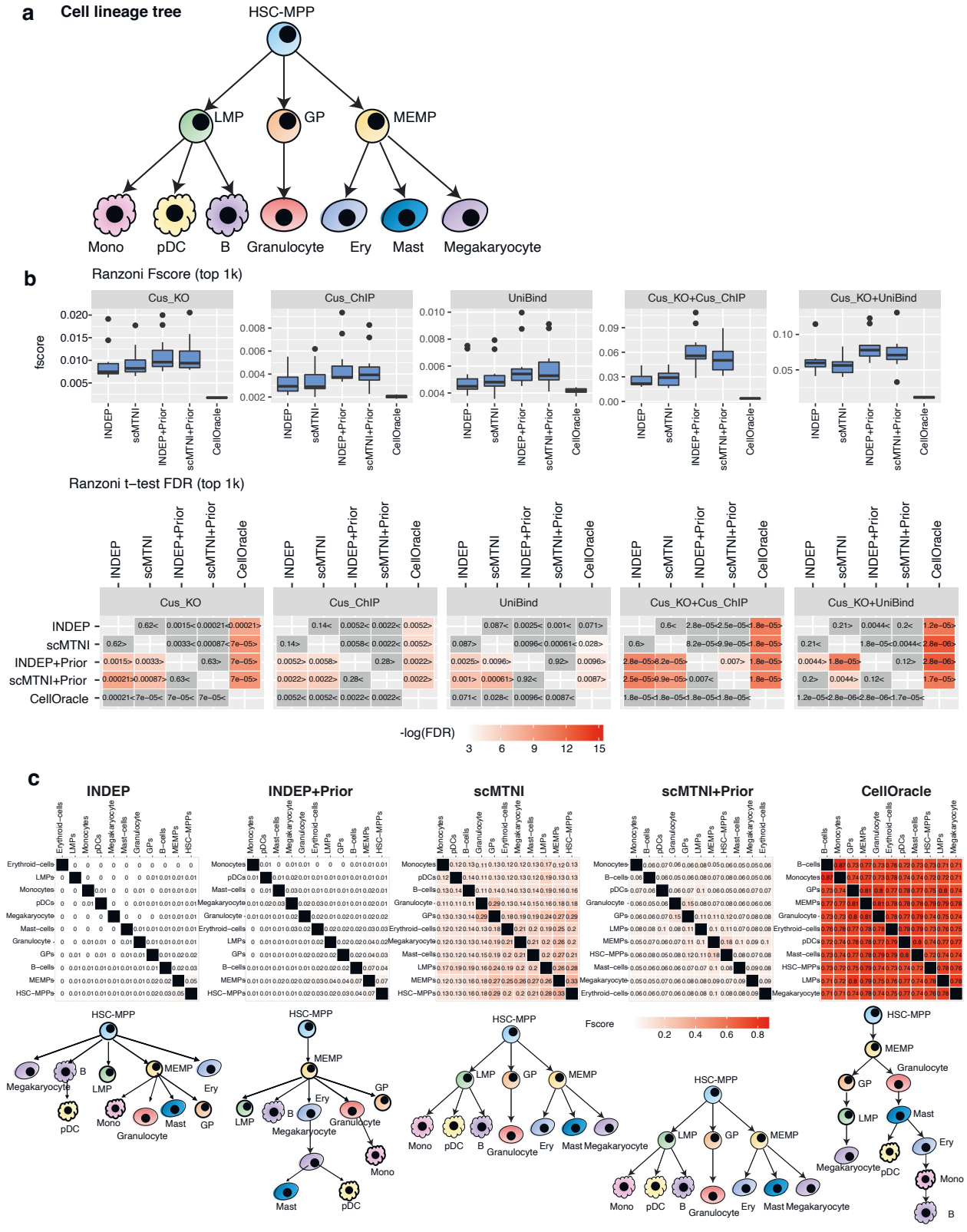

a
Cell lineage tree
HSC-MPP
LMP
GP
MEMP
Mono
pDC
B
Granulocyte
Ery
Mast
Megakaryocyte
b
Ranzoni Fscore (top 1k)
Cus_KO
Cus_ChIP
UniBind
Cus_KO+Cus_ChIP
Cus_KO+UniBind
fscore
INDEP
scMTNI
INDEP+Prior
scMTNI+Prior
CellOracle
Ranzoni t–test FDR (top 1k)
-log(FDR)
c
INDEP
INDEP+Prior
scMTNI
scMTNI+Prior
CellOracle
Fscore

Fig. 7 | Inference of cell type-specific networks for human fetal hematopoiesis data

a. Cell lineage structure linking the cell clusters from scRNA-seq.

b. Boxplots showing F-scores of n = 11 cell clusters for top 1k edges in predicted networks from scMTNI, scMTNI+Prior, INDEP, INDEP+Prior, and CellOracle compared to gold standard datasets (top). FDR-corrected t-test to compare the F-score of the row algorithm to the F-score of the column algorithm (bottom). The two-sided paired t-test is conducted on F-scores of n = 11 cell clusters for every pair of algorithms. A FDR < 0.05 was considered significantly better. The sign < or > specifies whether the row algorithm's F-scores were worse or better than the column algorithm's F-scores. The color scale is specified for − log(FDR), with the red color proportional to significance. Non-significance is colored in gray. In the boxplot, the horizontal middle line of each plot is the median. The bounds of the box are 0.25 quantile (Q1) and 0.75 quantile (Q3). The upper whisker is the minimum of the maximum value and Q3 +1.5*IQR,whereIQR = Q3 − Q1. The lower whisker is the maximum of the minimum value and Q1 − 1.5*IQR.

c. Pairwise similarity of networks from each cell cluster using F-score on the top 5k edges. Rows and columns ordered by hierarchical clustering using F-score as the similarity measure. Reconstructed cell lineage trees are shown at the bottom of the pairwise F-score similarity matrix and are constructed using the MST algorithm on the F-score matrix. HSC-MPP hematopoietic stem cells and multipotent progenitors, LMP lymphoid-myeloid progenitors, MEMP MK-erythroid-mast progenitors combined with cycling MEMPs, GP granulocytic progenitors, Ery erythroid cells, Mono

monocyte, pDC plasmacytoid dendritic cells. Source data are provided as a Source Data file.

Figure 8

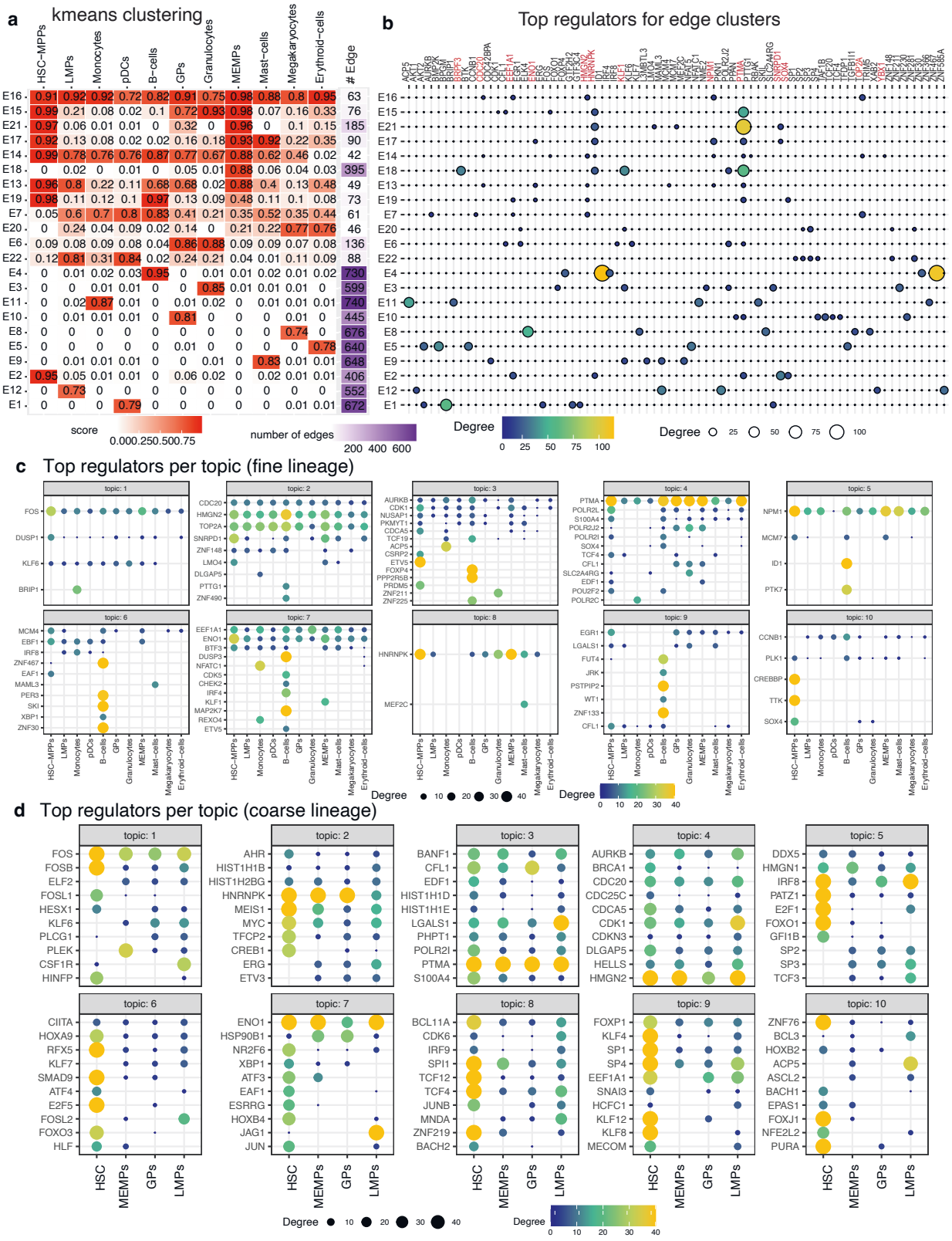

a
kmeans clustering
HSC−MPPs
LMPs
Monocytes
pDCs
B−cells
GPs
Granulocytes
MEMPs
Mast−cells
Megakaryocytes
Erythroid−cells
Edge
score
number of edges
b
Top regulators for edge clusters
Degree
c
Top regulators per topic (fine lineage)
d
Top regulators per topic (coarse lineage)
HSC
MEMPs
GPs
LMPs

Fig. 8 | Network rewiring during human fetal hematopoiesis

a. k-means edge clusters of the top 1k edges (rows) across 11 cell clusters (columns). The edge confidence matrix was clustered into 21 clusters to identify common and divergent networks. The red intensity corresponds to the average confidence of edges in that cluster. Shown also are the number of edges in the edge cluster.
b. Top 5 regulators of each edge cluster. The size and brightness of the circle is proportional to the number of targets. Regulators mentioned in text are in red.
c. Top regulators associated with each cell cluster's network in each topic for fine-grained lineage tree. Shown are only regulators that have at least 10 targets in any cell cluster. The brighter and larger the circle, the greater are the number of targets for the regulator.
d. Top regulators associated with each cell cluster's network in each topic for coarse lineage tree. Shown are only regulators that have at least 10 targets in any cell cluster. The brighter and larger the circle, the greater are the number of targets for the regulator. For ease of interpretation only the top 10 regulators per topic are shown. The full list of regulators per topic are shown in Supplementary Fig. 31. Source data are provided as a Source Data file.

Additional supplementary information and figures are available at:

https://doi.org/10.1038/s41467-023-38637-9

Table 1 | The statistics of the gold standard datasets used for the mouse reprogramming and human hematopoiesis studies

Dataset	Gold standards	Number of TFs	Number of targets
Mouse reprogramming	ChIP	54	31,367
	Perturb	179	21,019
	Perturb + ChIP	47	6109
Human hematopoiesis	Hematopoietic stem cells (HSC)	6	9173
	CD14_monocytes	1	6523
	megakaryocytes	4	8733
	erythroid_progenitors	1	7955
	R3R4_erythroid_cells	1	8494
	macrophages	1	163
	CD34_hematopoietic_stem_cells-derived_proerythroblasts	3	5847
	T-cells	3	6189
	B-cells	1	7036
	GM_B-cells	48	10,597
Human hematopoiesis	UniBind	56	10,621
	Cus_ChIP	149	6179
	Cus_KO	50	6108
	Cus_KO + Cus_ChIP	26	2124
	Cus_KO + UniBind	12	2020

Table 2 | The statistics of the real datasets and the size of the prior networks in mouse cellular reprogramming data, human hematopoietic data from Buenrostro et al., and human fetal hematopoiesis data from Ranzoni et al.

	Real dataset		Prior network		
Dataset	# regulators	# genes	avg. # of regulators	avg. # of genes	avg. # of edges
Cellular reprogramming	2036	12216	397	11290	892666
Adult hematopoiesis	1999	11994	324	10283	665931
Fetal hematopoiesis (fine tree)	2195	16737	255	9403	541813
Fetal hematopoiesis (coarse tree)	2227	17425	328	12308	865983

The averages are computed across the cell clusters or cell types for each dataset (cellular reprogramming data: n = 7, adult hematopoiesis data: n = 8, fetal hematopoiesis (fine tree): n = 11, fetal hematopoiesis (coarse tree): n =4).

References

1. Chronis, C. et al. Cooperative binding of transcription factors orchestrates reprogramming. Cell 168, 442–459.e20 (2017).

2. Smith, Z. D., Sindhu, C. & Meissner, A. Molecular features of cellular reprogramming and development. Nat. Rev. Mol. Cell Biol. 17, 139–154 (2016).

3. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. Nature 541, 331–338 (2017).

4. McDavid, A., Gottardo, R., Simon, N. & Drton, M. Graphical models for zero-inflated single cell gene expression. Ann. Appl. Stat. 13, 848–873 (2019).

5. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267.e3 (2017).

6. Matsumoto, H. et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics 33, 2314–2321 (2017).

7. Lim, C. Y. et al. BTR: training asynchronous Boolean models using single-cell expression data. BMC Bioinform. 17, 355 (2016).

8. Qiu, X. et al. Towards inferring causal gene regulatory networks from single cell expression Measurements. bioRxiv426981 https:// www.biorxiv.org/content/10.1101/426981v1 (2018).

9. Intosalmi, J., Mannerström, H., Hiltunen, S. & Lähdesmäki, H. SCHiRM: single cell hierarchical regression model to detect dependencies in read count data. bioRxiv335695 https://www. biorxiv.org/content/10.1101/335695v1 (2018).

10. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics 33, 764–766 (2017).

11. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086 (2017).

12. Zhang, R., Ren, Z. & Chen, W. SILGGM: an extensive R package for efficient statistical inference in large-scale gene networks. PLoS Comput. Biol. 14, e1006369 (2018).

13. Ocone, A., Haghverdi, L., Mueller, N. S. & Theis, F. J. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics 31, i89–i96 (2015).

14. Lim, C. Y. et al. Btr: training asynchronous boolean models using single-cell expression data. BMC Bioinform. 17, 1–18 (2016).

15. Chan, T. E., Stumpf, M. P. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).

16. Matsumoto, H. et al. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. Bioinformatics 33, 2314–2321 (2017).

17. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods 17,147–154 (2020).

18. McCalla, S. G. et al. Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. G3 Genes|Genomes|Genetics https://doi.org/10.1093/ g3journal/jkad004 (2023).

19. Jansen, C. et al. Building gene regulatory networks from scatac-seq and scrna-seq using linked self organizing maps. PLoS Comput. Biol. 15, e1006555 (2019).

20. Zeng, W. et al. Dc3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. Nat. Commun. 10, 1–11 (2019).

21. Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. Nature 614, 742–751 (2023).

22. Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R. & Kadie, C. Dependency networks for inference, collaborative filtering, and data visualization. J. Mach. Learn. Res. 1, 49–75 (2000).

23. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 177, 1873–1887 (2019).

24. Kester, L. & van Oudenaarden, A. Single-cell transcriptomics meets lineage tracing. Cell Stem Cell 23,166–179 (2018).

25. Koch, C. et al. Inference and evolutionary analysis of genome-scale regulatory networks in large phylogenies. Cell Syst. 4, 543–558 (2017).

26. Pierson, E. et al. Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput. Biol. 11, e1004220 (2015).

27. Jojic, V. et al. Identification of transcriptional regulators in the mouse immune system. Nat. Immunol. 14,633–643 (2013).

28. Castro, D. M., De Veaux, N. R., Miraldi, E. R. & Bonneau, R. Multistudy inference of regulatory networks for more accurate models of gene regulation. PLoS Comput. Biol. 15, e1006591 (2019).

29. Tibshirani, R. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B (Methodological) 58,267–288 (1996).

30. Aibar, S. et al. Scenic: single-cell regulatory network inference and clustering. Nat. Methods 14,1083–1086 (2017).

31. Sridharan, R. & Plath, K. Illuminating the black box of reprogramming. Cell Stem Cell 2,295–297 (2008).

32. Tran, K. A. et al. Defining reprogramming checkpoints from single-cell analyses of induced pluripotency. Cell Rep. 27, 1726–1741 (2019).

33. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods 14,979–982 (2017).

34. Xu, H. et al. Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford) 2013,bat045(2013).

35. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489,57–74 (2012).

36. Nishiyama, A. et al. Uncovering early response of gene regulatory networks in escs by systematic induction of transcription factors. Cell Stem cell 5,420–433 (2009).

37. Marin,M.,Karis,A.,Visser,P.,Grosveld,F.&Philipsen,S.Transcription factor sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. Cell 89, 619–628 (1997).

38. Bouwman, P. et al. Transcription factor sp3 is essential for postnatal survival and late tooth development. EMBO J. 19, 655–661 (2000).

39. Festuccia, N., Owens, N. & Navarro, P. Esrrb, an estrogen-related receptor involved in early development, pluripotency, and reprogramming. FEBS Lett. 592,852–877 (2018).

40. Lou, S. et al. Topicnet: a framework for measuring transcriptional regulatory network change. Bioinformatics 36,i474–i481 (2020).

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

42. Wang,D.,Rabhi,N.,Yet,S.-F.,Farmer,S.R.&Layne,M.D.Aortic carboxypeptidase-like protein regulates vascular adventitial progenitor and fibroblast differentiation through myocardin related transcription factor a. Sci. Rep. 11, 3948 (2021).

43. González, A., López, B., Ravassa, S., San José, G. & Díez, J. The complex dynamics of myocardial interstitial fibrosis in heart failure. Focus on collagen cross-linking. Biochim. Biophy. Acta (BBA)-Mol. Cell Res. 1866,1421–1432 (2019).

44. Rao, M. et al. Resolving the intertwining of inflammation and fibrosis in human heart failure at single-cell level. Basic Res. Cardiol. 116, 1–19 (2021).

45. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell 173,1535–1548 (2018).

46. Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A. & Mathelier, A. Unibind: maps of high-confidence direct tf-dna interactions across nine species. bioRxiv2020-11 (2021).

47. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. PLoS Genet 10, e1004226+ (2014).

48. Bhullar, J. & Sollars, V. E. Ybx1 expression and function in early hematopoiesis and leukemic cells. Immunogenetics 63, 337–350 (2011).

49. Alidousty, C. et al. Calcineurin-mediated yb-1 dephosphorylation regulates ccl5 expression during monocyte differentiation. J. Biol. Chem. 289,21401–21412 (2014).

50. de Barros, Z. V. et al. 3142–glucocorticoid-induced leucine zipper (gilz) intrinsically regulates hematopoietic stem cell function. Exp. Hematol. 88,S82(2020).

51. Delgado, M. D. & León, J. Myc roles in hematopoiesis and leukemia. Genes Cancer 1,605–616 (2010).

52. Doré, L. C. & Crispino, J. D. Transcription factor networks in erythroid cell and megakaryocyte development. Blood J. Am. Soc. Hematol. 118,231–239 (2011).

53. Siatecka, M. & Bieker, J. J. The multifunctional role of eklf/klf1 during erythropoiesis. Blood J. Am. Soc. Hematol. 118, 2044–2054 (2011).

54. Tamir, A. et al. Fli-1, an ets-related transcription factor, regulates erythropoietin-induced erythroid proliferation and differentiation: evidence for direct transcriptional repression of the rb gene during differentiation. Mol. Cell. Biol. 19, 4452–4464 (1999).

55. Wang, H. & Morse, H. C. Irf8 regulates myeloid and b lymphoid lineage diversification. Immunol. Res. 43,109–117 (2009).

56. Wuerzberger-Davis, S. M. et al. Nuclear export of the nf-κb inhibitor iκbα is required for proper b cell and secondary lymphoid tissue formation. Immunity 34,188–200 (2011).

57. Ji, M. et al. Id2 intrinsically regulates lymphoid and erythroid development via interaction with different target proteins. Blood J. Am.Soc.Hematol.112,1068–1077 (2008).

58. Zhang, C. et al. Latexin regulation by HMGB2 is required for hematopoietic stem cell maintenance. Haematologica 105, 573–584 (2020).

59. Ranzoni, A.M. et al. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell 28,472–487.e7 (2021).

60. Gallardo, M. et al. hnrnp k: a novel regulator of hematopoiesis and a potential predictive biomarker in acute myeloid leukemia. Blood 122, 226 (2013).

61. Case, N. T. et al. The macrophage-derived protein ptma induces filamentation of the human fungal pathogen candida albicans. Cell Rep. 36, 109584 (2021).

62. Samara, P., Ioannou, K. & Tsitsilonis, O. Prothymosin alpha and immune responses: are we close to potential clinical applications? Vitam. Horm. 102,179–207 (2016).

63. Lopez de Lapuente Portilla, A. et al. Genome-wide association study on 13,167 individuals identifies regulators of blood cd34+ cell levels. Blood 139,1659–1669 (2022).

64. Soufi, A. & Dalton, S. Cycling through developmental decisions: how cell cycle dynamics control pluripotency, differentiation and reprogramming. Development 143,4301–4311 (2016).

65. Sichien, D. et al. Irf8 transcription factor controls survival and function of terminally differentiated conventional and plasmacytoid dendritic cells, respectively. Immunity 45,626–640 (2016).

66. Raval, A. et al. Npm1 haploinsufficiency results in increased numbers of hematopoietic stem cells and progenitor cells. Blood 114, 738 (2009).

67. Ruvolo, P. P. et al. Lgals1 acts as a pro-survival molecule in aml. Biochim. Biophys. Acta (BBA) Mol. Cell Res. 1867, 118785 (2020).

68. Shao, L., Paik, N. Y. & Pajcini, K. V. Hematopoietic jagged1 is required for the transition of hematopoietic stem cells from the fetal liver to the adult bone marrow niche. Blood 136,10–11 (2020).

69. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. Rna velocity-current challenges and future perspectives. Mol. Syst. Biol. 17, e10282 (2021).

70. Miraldi, E. R. et al. Leveraging chromatin accessibility for transcriptional regulatory network inference in t helper 17 cells. Genome Res. https://doi.org/10.1101/gr.238253.118 (2019).

71. Roy, S. et al. Integrated module and Gene-Specific regulatory inference implicates upstream signaling networks. PLoS Comput. Biol. 9, e1003252+ (2013).

72. Chen, H. et al. Assessment of computational methods for the analysis of single-cell atac-seq data. Genome Biol. 20,1–25 (2019).

73. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods 13, 241–244 (2016).

74. Zhang, Y. et al. Model-based analysis of chip-seq (macs). Genome Biol. 9,1–9(2008).

75. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158,1431–1443 (2014).

76. Sherwood, R. I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling dnase profile magnitude and shape. Nat. Biotechnol. 32,171–178 (2014).

77. Ranzoni, A.M. et al. Integrative single-cell rna-seq and atac-seq analysis of human developmental hematopoiesis. Cell Stem Cell 28,472–487 (2021).

78. Stuart, T. et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902 (2019).

79. Moerman, T. et al. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics 35, 2159–2161 (2019).

80. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. In Proc. 23rd International Conference on Machine Learning,233–240 (2006).

81. Ashburner, M. et al. Gene ontology: tool for the unification of biology. Nat. Genetics 25,25–29 (2000).

82. Zhang, S. et al. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. Zenodo https://doi.org/10.5281/zenodo.7834742 (2023).

83. Zhang, S. et al. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. scMTNI https://doi.org/10.5281/zenodo.7854535 (2023). https://github. com/Roy-lab/scMTNI.