Title
Deconstructing Cell Fate Transition Dynamics and Epigenetic Heterogeneity using Single Cell Technology

Author
Sabri, Shan

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Deconstructing Cell Fate Transition Dynamics and Epigenetic Heterogeneity using Single Cell Technology

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Bioinformatics

by

Shan Sabri

2020

ABSTRACT OF THE DISSERTATION

Deconstructing Cell Fate Transition Dynamics and

Epigenetic Heterogeneity using Single Cell Technology

by

Shan Sabri

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2020

Professor Kathrin Plath, Co-Chair

Professor Jason Ernst, Co-Chair

The ability to create pluripotent stem cells (PSCs) from any tissue by a process called reprogramming (yielding induced pluripotent stem cells, iPSCs) has ushered in an era of personalized medicine. However, reprogramming protocols are not trivial and are nearly always inefficient, often yielding an efficiency of less than 0.1%. Similar low efficiencies occur for many of the forward differentiation protocols, where it is also a major question which cell types are made and how they compare to the cells present *in vivo*. In this work, we applied single cell RNA-sequencing on iPSC reprogramming from different somatic cells to define the transcriptional changes in the process and the role of the reprogramming factors and somatic TFs in the reorganization of cell identity, revealing the critical role of intermediate ectopic gene expression. Changes to the reprogramming transcription factor complement results in similar intermediates while skewing the number of cells that reached particular cell stages. Intriguingly, distinct transcription factors induced unique novel ectopic transient gene networks, the character of which

influenced the efficiency of reprogramming. This work thoroughly describes the processes of cell-fate decision-making, and uncovers the nature of the ectopic gene expression state as a gate keeper of reprogramming progression.

Building on this, we also establish a novel computational method to deconvolve the epigenetic control of heterogeneous processes, such as reprogramming and differentiation, thereby uncovering mechanisms underlying cell type specifications and transitions. Using techniques from machine learning, we train models to learn the relationship between the transcriptome and epigenome from an atlas of homogeneous cell populations, then apply these models to single cell populations. Our results illustrate accurate deconvolution of a human fetal brain organoid for which we have predicted the H3K27ac epigenomic landscape, a histone modification mark that is nearly impossible to profile at single cell level.

Together, my graduate work has focused on developing novel computational methods and analysis techniques that leverage single cell genomics for studying gene regulation while gaining insight into the mechanisms underlying cell fate change processes, as well as how to effectively derive single cell type chromatin state data.

The dissertation of Shan Sabri is approved.

Sriram Sankararaman

Xinshu (Grace) Xiao

Jason Ernst, Committee Co-Chair

Kathrin Plath, Committee Co-Chair

University of California, Los Angeles

2020

DEDICATION

This dissertation is dedicated to each of my parents, for their support and encouragement.

Table of Contents

LIST OF FIGURES AND TABLES

CHAPTER 2

ACKNOWLEDGEMENTS

I am forever grateful to my mentors, Dr. Kathrin Plath and Dr. Jason Ernst, for their support and guidance throughout my doctorate work in their labs. They placed me in a unique, cross-functional environment that has helped me grow as a scientist and allowed me to appreciate the rigor required for pursuing high impact scientific research. I will always value my time in their labs and throughout my professional career.

I would like to thank my thesis committee, Sriram Sankararaman and Xinshu (Grace) Xiao, for their thoughtful discussions regarding my research projects and coursework. I would also like to thank all faculty, staff and students within the UCLA Bioinformatics Interdepartmental Program for their contribution to an invaluable education and student experience.

To the current and former members of the Ernst and Plath labs, I will always be thankful for your mentorship, friendship, scientific support and constructive criticism. Many thanks to Drs. Justin Langerman and Constantinos Chronis for their hard work and contributions to the work described in Chapter 1. Additional thanks Drs. Haibin Xi, April Pyle and Brigitte Gomperts for allowing me to be a part of their exciting collaborations. I will remember you all throughout my career.

I would also like to acknowledge UCLA's Graduate Division for providing me with the Dissertation Year Fellowship, as well as the Rose Hills Foundation and the Eli and Edythe Broad Foundation for supporting three consecutive years of my pre-doctoral training.

Lastly, to my family and friends, thank you for listening, offering me advice and supporting me through this entire process, which shaped me to being the person I am today. I am extremely grateful for the sacrifices my parents have made to give me the opportunity and encouragement to pursue an education. I have been fortunate enough to share this journey, the good and the bad, with my lovely girlfriend, Sara Harris – thank you for being there when I need you most.

Chapter 2 is a version of a manuscript in review reporting findings from a collaborative project with equal effort from Justin Langerman and me. Its authors are Langerman J.*, Sabri S.*, Chronis C., Ernst J., and Plath K.

Chapter 3 is a version of a manuscript in preparation for peer-reviewed publication. It is authored by Sabri S., Allison T., Chronis C., Jacobson E., Deng W., Chovanec P., Plath K., and Ernst J.

Chapter 4 is comprised of selected manuscripts that are the result of successful collaborations. These include (1) a peer-reviewed manuscript with Drs. April Pyle and Haibin Xi recently published in *Cell Stem Cell*, and (2) a manuscript in review authored with Drs. Brigitte Gomperts, Gianni Carraro and Justin Langerman.

VITA

EDUICATION

2020	Ph.D. Candidate, Bioinformatics University of California, Los Angeles
2014	M.Sc. Bioinformatics Johns Hopkins University
2012	B.Sc. Bioinformatics, minor Biostatistics Loyola University Chicago

RESEARCH EXPERIENCE

2014 – Present	Ph.D. trainee Department of Biological Chemistry University of California, Los Angeles
2018	Bioinformatics Research Intern Department of Computational Biology and Bioinformatics Genentech

ACADEMIC FELLOWSHIPS

2019	UCLA Dissertation Year Fellowship
2017	Philip J. Whitcome Fellowship (awarded but declined)
2016 – 2018	Broad Stem Cell Research Center Fellowship

PUBLICATIONS AS A UCLA GRADUATE STUDENT

Xi H., *et al*. A Human Skeletal Muscle Atlas Identifies the Trajectories of Stem and Progenitor Cells across Development and from Human Pluripotent Stem Cells. *Cell Stem Cell*. 1934-5909 (2020). Epub ahead of print.

Polioudakis D., *et al.* A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron **103**, 785-801.* (2019).

Allison, T. *et al.* Defining Transcriptional Signatures of Human Hair Follicle Cell States. *J. Investig. Dermatol **140**, 764-733. (2019)*

Stefano, BD., *et al.* Reduced MEK Inhibition Preserves Genomic Stability in Naïve Human ES Cells. *Nature Methods **15***, 732-740. (2018).

Allison T., *et al*. Identification and Single-Cell Functional Characterization of an Endodermally Biased Pluripotent Substate in Human Embryonic Stem Cells. *Stem Cell Reports **10**, 1895-1907* (2018).

Sereti KI., *et al.* Analysis of cardiomyocyte clonal expansion during mouse heart development and injury. *Nature Communications **9**, 1-13 (2018).*

Sahakyan, A., *et al*. Human Naive Pluripotent Stem Cells Model X Chromosome Dampening and X Inactivation. *Cell Stem Cell* ***20,*** 87-101. (2017).

Chronis C.*, Fiziev P.*, *et al*. Cooperative binding of transcription factors orchestrates reprogramming. *Cell* ***168***, 442-459. (2017).

MANUSCRIPTS IN REVIEW

Langerman J.*, Sabri S.*, et al. Identification of Conserved Mechanisms Underlying the Gene Expression Changes During Reprogramming Processes. *In review* (2020).

Gianni C.*, Langerman J.*, et al. A molecular atlas of proximal airway identifies subsets of known airway cell types revealing details of the unique molecular pathogenesis of Cystic Fibrosis. *In review* (2020).

MANUSCRIPTS IN PREPARATION

Sabri S.*, *et al*. Genome-wide deconvolution of single cell type chromatin features using single cell gene expression data. (2020).

Chapter 1. Introduction

Pluripotent stem cells hold great promise in the field of personalized and regenerative medicine. Though for realizing their potential, it is essential to produce and forward differentiate these stem cells to desired cell fates, yet protocols for doing so are not trivial and often inefficient. In 2006, Kazutoshi Takahashi and Shinya Yamanaka made the groundbreaking observation that a cocktail of transcription factors, Oct4, Sox2, Klf4, and cMyc (commonly referred to as OSKM), can induce pluripotency in somatic cells resulting in induced pluripotent stem cells (iPSCs). Since this discovery, the amount of data gathered on iPSCs has been astonishing yet little is known molecularly about the inefficiencies of this process. The understanding of what regulatory mechanisms drive cellular reprogramming and differentiation is a fundamental question in the field of stem cell research. Knowing the cellular decision-making underlying the reprogramming of cells to induced pluripotency will allow researchers to deconstruct the process thereby identifying and overcoming bottle necks that hinder efficiency. Reprogramming to iPSCs requires the silencing of the somatic program and activation of the pluripotency program while preventing alternative fates, but it still remains unclear how these events are regulated and related to each other, and what aspects are conserved when iPSCs are derived from different cell types.

Various studies have employed genome-wide techniques to reveal changes in gene expression and chromatin state that occur during iPSC reprogramming at the population level. However, population-based studies cannot be used for a full understanding of the reprogramming process, because they produce an ensemble response with no knowledge of the cellular variance of that response. To overcome this, we have adapted and applied the droplet-based Drop-seq method to characterize the transcriptomes of individual cells in a high-throughput manner. We profiled 80,000+ individual cells at many time points and conditions to characterize transcriptional identities and dynamic trajectories describing paths from the starting to end states along reprogramming. My work has aimed to understand the fundamental question of why individual

cells follow distinct trajectories in differentiation and reprogramming processes, establish different cell states, and/or fail to reprogram or differentiate, even when starting from homogeneous cell populations grown in defined culture conditions. Doing so has provided a stepping stone towards the mechanistic dissection of cell fate change processes and the rational manipulation of cell fates for diverse applications.

In **Chapter 2** we investigate cell fate transition dynamics along the path toward pluripotency using single cell technology. We profiled iPSC reprogramming from different somatic cells to define the transcriptional changes in the process and the role of the reprogramming factors and somatic TFs in the reorganization of cell identity. To address this, we develop a computational method, Gene Expression Network Discovery (GEND), for the discovery of co-regulated gene networks allowing us to captures both gradual transcriptional changes and rare populations with specific gene signatures. Using GEND, we reveal the critical role of intermediate ectopic gene expression. The ontology of genes within these ectopic networks are related to various developmental processes unrelated to iPSC reprogramming, indicating that genes normally expressed in different lineages are transiently upregulated during reprogramming, and that often in the same cell. We found that among those different lineage markers, are genes that facilitate the mesenchymal-to-epithelial transition, which is critical for pushing the cells towards the pluripotent state. Intriguingly, we show that changes to the reprogramming transcription factor complement results in the intermediate state being expanded or depleted, respectively, when the reprogramming process is enhanced or blocked. Our work provides a general framework for how one cell identity emerges from another in reprogramming processes, and uncovers mechanistic insight to enhancing the process.

In **Chapter 3** we develop a computational method to deconvolve a bulk, population-level chromatin profile into the underlying heterogeneous cell subtypes using single cell gene expression information. To achieve this, we train machine learning models to learn the

relationship between the transcriptome and epigenome from an atlas of purified cell types, then apply these models to single cell subtypes. We show that a properly trained model can computationally predict enhancers in single cells from solely using gene expression information as input. Our results show accurate deconvolution of histone modification H3K27ac profile for a human fetal brain organoid into six underlying cell types defined by single cell sequencing. Our method alleviates the need for an additional chromatin assay or cells sorting to profile heterogeneous samples and provides valuable information to guide hypothesis generation, target prioritization, and design of follow-up experiments.

In **Chapter 4** I present the works of two collaborations that have resulted in manuscripts. The first peer-reviewed manuscript is published in *Cell Stem Cell* and was in collaboration with Dr. April Pyle's lab at UCLA. In this work, we develop a never-before-seen roadmap tracking how human skeletal muscle develops through a process called myogenesis, from an early embryonic state to adulthood, including the formation of muscle stem cells, which are essential to treat muscle disorders. We established a computational method, similar to that described in Chapter 2, to map gene networks related to muscle progenitor and stem cells to a developmental continuum thereby allowing us to match genetic signatures found in pluripotent stem cell-derived muscle cells with their corresponding locations on the roadmap of human muscle development. Interestingly, we find some networks are enriched for cell types that support muscle cells. This developmental roadmap has laid a foundation for developing muscle stem cells in the lab that can be used for regenerative cell therapies for a variety of muscle diseases, including muscular dystrophies.

Here, I also present a manuscript in peer-review for publication from a collaboration with three institutions including Drs. Kathrin Plath and Brigitte Gomperts labs at UCLA, Dr. John Mahoney's group at the Cystic Fibrosis Foundation and Dr. Barry Stripp's lab at the Lung and Regenerative Medicine Institute of Cedars-Sinai Medical Center. We present a molecular atlas

defined by 38 patient samples of the proximal airway epithelium to define disease-related changes to the proximal airway of cystic fibrosis (CF) donors undergoing transplantation for end-stage lung disease as compared to the proximal airway of previously healthy lung donors. For the first time, we report novel identification of proximal airway epithelial basal, secretory, and ciliated molecular subtypes. We show that the secretory subtype is associated with mucosal immunity in CF donors while the ciliated subtype was found to have a larger number of transitioning precursors in these same donors suggesting greater plasticity than controls. Our results provide insights for the development of new targeted therapies for CF-related airway dysfunction.

All studies in these chapters leverage high-throughput droplet-based single cell technology to profile thousands of cells in a variety of conditions to identify key genetic signatures and differences.

Chapter 2. Identification of conserved mechanisms underlying the gene expression changes during reprogramming processes

Abstract.

Reprogramming to iPSCs requires the silencing of the starting and activation of the pluripotency program while preventing alternative fates, but it still remains unclear how these events are regulated and related to each other, and what aspects are conserved when iPSCs are derived from different cell types. We applied single cell RNA-sequencing to of MEF-, Keratinocyte- and NSC-to-iPSC reprogramming and created a method to define gene networks over reprogramming time.

For all three systems, cell fate transitions proceed via the silencing of the most tissue-specific starting cell type genes, followed by stepwise silencing of more broadly expressed somatic genes, while upregulation of target cell genes starts with more broadly expressed genes and ending with most iPSC-specific genes. The most tissue specific somatic and iPSC genes are rarely expressed together, while the broader gene networks coexist. The silencing of the most somatic-specific genes enables expression of ectopic networks, including co-expression of genes from diverse lineages, not expressed in the starting and end states, which are unique in each reprogramming system. The ectopic networks are always silenced together with the broadly expressed somatic genes during induction of the most iPSC-specific genes.

Mechanistically, the timing of these transitions is explained by the regulatory input from the reprogramming and somatic TFs and genomic architecture (CpG content, DNA accessibility, enhancer location). Stalling can occur by failure to silence somatic networks, via deviation through the induction of stress responses, and in an ectopic network expressing intermediate state.

Changes to the reprogramming transcription factor complement results in similar intermediates while skewing the amount of cells that reached particular cell stages. Intriguingly, distinct transcription factors induced unique novel ectopic transient gene networks, the character of which influenced the efficiency of reprogramming. Thus, our work provides a general framework

for how one cell identity emerges from another in reprogramming processes, and uncovers the nature of the ectopic gene expression state as a gate keeper of reprogramming progression.

Introduction.

The conversion of somatic cells to induced pluripotent stem cells (iPSCs) by the overexpression of the transcription factors (TFs) Oct4, Sox2, Klf4, and cMyc (OSKM)[1] holds great promise for the treatment and understanding of diseases, and provides a powerful tool for studying mechanisms underlying cell fate changes[1–3]. To this end, profiling of reprogramming cultures with a multitude of genomics[4–10], imaging and proteomics[11] approaches has shown that reprogramming to iPSCs requires proliferation and entails the silencing of the starting somatic gene expression program, in the case of mesenchymal starting cells the mesenchymal to epithelial transition (MET), a process in which cells lose their mesenchymal characteristics and acquire epithelial features, and the induction of the pluripotency program.

However, as most cells fail to reprogram due to the inefficiency and asynchrony of reprogramming events, the relationship between these expression programs has remained unclear. Consequently, single cell RNA-sequencing (scRNA-seq) approaches[12-15] have recently been applied to the reprogramming process, starting cell population that correlates with reprogramming potential, defined a continuum of cell states during reprogramming, and defined the role of signaling modulation.

Despite these advances, it is still not well defined how the reprogramming transcriptomes changes, whether it occurs through binary switches, stochastic expression, coordinated gradual cascades, or through novel intermediates with co-expression. Moreover, which transcriptional changes are conserved when different starting cell types are reprogrammed, is still little understood. Insights into these questions are critical for elucidating the mechanisms underlying cell fate change during reprogramming.

To address these issues, we applied single cell RNA-sequencing[16] to the induction of iPSCs from mouse embryonic fibroblasts (MEFs), keratinocytes, and neural progenitors (NPCs), respectively. In these experiments, we also altered culture conditions, the stoichiometry of

reprogramming factors, and expressed somatic or pluripotency TFs together with OSKM, to define mechanisms associated with enhancement and inhibition of reprogramming, capturing 80,690 single cell transcriptomes (median detection of 3884 UMIs and 1799 genes per cells) in total (Table 2-1). These data allowed us to define the steps towards iPSCs that are conserved across different cells of origin and underlying mechanisms. We also show that a similar transcription logic applies to trans-differentiation processes. Taken together, our work therefore uncovers general principles underlying transcription factor-induced cell fate changes.

Results.

We applied single cell transcriptomics on the MEF-to-iPSC conversion, to first define transcription changes along the reprogramming trajectory in a well-studied system. We profiled a reprogramming time course of female MEFs carrying a doxycycline (dox)-inducible polycistronic OSKM cassette[16,17] (Fig 1A, Supp Fig 1A, Table 2-1). Upon OSKM expression (Supp Fig S2-1B), virtually all cells shifted away from the starting MEF state, and shifted towards the iPSC state over time (Fig 2-1B, Supp Fig S2-1), which was confirmed with additional dimensionality reduction methods and in three additional time courses (Supp Fig S2-2). iPSCs were defined by well-established molecular features such as silencing of somatic genes like TWIST and COL1A1, silencing of Xist and inactive X reactivation, and expression of the pluripotency markers Nanog, Esrrb, endogenously-encoded Pou5f1 and Sox2, and were first detected by day 9 consistent with the latency of iPSC induction (Fig 2-1C, Supp Fig S2-1C-F)[1,5,,11,19–23].

To delineate the molecular differences that arise as cells convert from MEFs into iPSCs, we developed a new method, Gene Expression Network Discovery, or GEND (Fig 2-1D). GEND is a fast, sensitive approach to discover sets of co-regulated genes which uniquely captures both gradual transcriptional changes and rare populations with specific gene signature, which we used to detect all cell identities. We identified networks with higher expression in MEFs than in iPSCs (somatic networks DOWN1-4) (Fig 2-1Ei), which differed in the number of cells expressing them from few (DOWN1) to nearly all cells (DOWN4), in the time points at which they were detected, and in their associated with Gene Ontology terms from extracellular matrix (DOWN1-3) to cytoskeleton regulation (DOWN4). (Fig 2-1Ei/1F, Supp Fig S2-3). DOWN1 and DOWN2 were upregulated early in reprogramming (days 2-4), partially due to KSR addition (Supp Fig S2-4A-E).

We also identified eight programs that were more highly expressed in iPSCs than in MEFs (Fig 2-1Eii, pluripotency networks UP 1-8), and contained genes associated with mRNA

metabolism (UP1/2/5), chromatin organization (UP2-6), cell division (UP1-4,6), DNA repair (UP1,3-6), and embryonic development (UP6-8), based on GO analysis (Supp Fig S2-3). The UP6-8 networks contain many TFs including key regulators of the pluripotent state including Tcfcp2l1 and Tfap2c (UP6), endogenously expressed Sox2 and Pou5f1, Sall4, Nr5a2, Nanog, MycN, Lin28 (UP7), and Zfp42, Esrrb, Nr0b1 (UP8) (Fig 2-1Eii, Supp Fig S2-5). Similar to the DOWN programs, the UP programs differed in the number of cells expressing them from many cells (UP1) to few (UP8) and time points (Fig 2-1Eii/F, Supp Fig S2-3B).

In addition to up- and downregulated genes, we reproducibly identified several additional networks (Fig 2-1E, Supp Fig S2-6), including five containing ectopically expressed genes, that are expressed in d2-24 cultures, but not in MEFs and iPSCs. The ectopic networks include the Osteogenic and Interferon response networks, the Neural and post-implantation epiblast networks, which had unique TF profiles (Neural: Nkx6-1, Ascl1 and Pou3f2, - Epiblast: FoxA1/A2, Gata4, Sox17, Lhx1, and Eomes), and the Transient network (Fig 2-1Eiii, Supp Fig S2-3). An epiblast-like cell stage had been proposed as an intermediate during reprogramming, however we performed a lineage tracing experiment using X chromosome reactivation and revealed that the post-implantation epiblast cells differentiated from pluripotent iPSCs (Supp Fig S2-7), indicating that they do not arise along the reprogramming path.

The Transient program warranted further consideration as it contained genes previously found to be upregulated in the middle of reprogramming, such as Itgb4 and Ehf, and was present in cells which were more competent to finish reprogramming (Supp Fig S2-8A/B). The Transient network included genes that are normally in unrelated tissues, including Avil (highly expressed in stomach, intestine, brain), Cacna2d2 (brain and heart), Chst3 (most highly expressed in the spleen controlling T cell maintenance), Ehf (an epithelium-specific Ets transcription factor), and Prx (expressed highly in the brain with a function in Schwann cells), and Crym (a thyroid hormone binding protein in muscle, neurons, and prostate), which we confirmed to be translated mid-

reprogramming (Supp Fig S2-8C-E). This network captures the previously reported upregulation of lineage-specific markers from unrelated lineages during reprogramming[8], and show that these genes are co-expressed in individual cells (Supp Fig S2-8G). Despite the heterogeneity, these data show that the induction of coordinated cell fate programs is a rare event during reprogramming, limited to a small Neural network in this system.

We next sought to find what other properties were correlated to the different proportions of cells that the various DOWN and UP networks are expressed in, from few cells (narrow) to many (broad). Calculating the ESC to MEF expression ratio of each network using population RNA-seq data, we found that the narrow DOWN1/2 networks were more fibroblast specific compared to the broader DOWN3/4 (6-10 fold vs 2-4 fold respectively), and upregulated genes become more ESC specific from the broadest UP1 to the narrowest UP8 network (from 2 fold to 17 fold) (Fig 2-1G, Supp Fig S2-9A). We asked whether the narrower networks were also more restricted in their general tissue specificity. We determined the number of tissues where each gene within the networks was expressed, using ENCODE expression data from eighteen different tissues. DOWN1/2 were more restricted in their tissue expression than DOWN3/4, and UP 1-5 were more broadly expressed than UP6-8 (Fig 2-1H). These findings show that the tissue specificity of genes is tied to the regulation of gene expression during reprograming. Pairwise expression scatterplots demonstrate the more broadly expressed networks such as DOWN3/4 and UP1-5 are co-expressed in a large number of cells, however the narrow programs DOWN1/2 and UP6-8 do not overlap (Fig 2-1I). This was confirmed by examining the expression of the most cell type specific MEF-and ESC genes (defined based on population RNA-seq data up or down 5 fold (Table 2-2, Supp Fig S2-9). These data indicated that the most MEF-specific genes are downregulated before the most ESC-specific genes are upregulated.

We therefore hypothesized that by ordering the reprogramming cells via a trajectory based on tissue specific gene antagonism, we could define the changes and overlaps in gene network

expression which are occurring during reprogramming time. Indeed, a trajectory based on MEF- and ESC- specific gene antagonism, coined the trajectory score, was in close agreement with algorithmic progression modeling by Monocle2[24,25] (r=0.96, Fig 2-1J, Supp Fig S2-9K). This illustrates the existence of cells with an intermediate trajectory score exist which express only low levels of both MEF and ESC-specific genes, marking cells which capture the transition to the pluripotent state in the reduce dimensionality plots (Fig 2-1J). Along the reprogramming trajectory, DOWN1-4 and UP1-8 change in a gradual and stepwise manner (Fig 2-1K). Gene networks with higher tissue specificity are more rapidly downregulated/ harder to upregulate, during reprogramming, than broader programs. In contrast, the Transient gene network was most strongly upregulated when the most tissue specific UP and DOWN genes were lowly expressed (Fig 2-1K, Supp Fig S2-8F).

To define cell states that arise over reprogramming time, we developed a method, Combinatorial Expression of Networks Into Cell States (CENICS), to quantify the combinatorial expression of the DOWN, UP, Transient, Osteogenic, and Interferon networks. We designated cells as 'On' or 'Off' for the networks and determined the percentage of cells with combinations of these 'On' networks. By doing this for each reprogramming time point, we can precisely determine the appearance of novel cell states and stalling points over time (Fig 2-1L, Supp Fig S2-10). The majority of starting MEFs were positive for DOWN2/3/4 and were also positive for UP 1/2/5 consistent with their broad expression character. Upon OSKM induction (d2-d4), many cells populate a surprising variety of new cell states (Fig 2-1Lii). For instance, some cells express DOWN1 and/or the Interferon or Osteogenic networks, which correlates with an increase in somatic identity along the trajectory axis (up to a score of -1.0). The DOWN1/Interferon networks are early responses which are lost from the culture at late time points, while the Osteogenic network is enriched in states which accumulate over d9-d15, consistent with the idea that these networks represent non-productive events. Other cells have downregulated DOWN2, induced the Transient network in the presence of DOWN3, and are closer to the iPSC state, consistent with

the transient network being expressed in productive intermediates. However, some cells also maintain the initial MEF state, indicating a highly heterogeneous penetrance of reprogramming efficacy. By day 9, the number of cells expressing the Transient network increases and 6% of cells express UP6 along with DOWN3 and Transient network expression, while another 7% express UP6-8 together, and lacked DOWN2 /3 and Transient gene expression, capturing the fully reprogrammed state. After several rounds of splitting (day 24/34), pluripotent cells with all UP programs have outcompeted the other cell states.

The first cells along the trajectory which induced UP6 had lower expression of DOWN3 and Transient networks compared to the prior cell states, and high expression of UP6 was correlated with loss of DOWN3 and Transient expression and gain of UP7/8 (Fig 2-1Liii). Only a small proportion of cells with incomplete UP6-8 combinations accumulated over time, indicating that cells convert quickly through pre-pluripotent states after inducing UP6. Thus, the gradual upregulation of UP6 and repression of DOWN3 and the Transient networks is a critical for transition into the iPSC state, and only happens late in reprogramming (d9) with low efficiency. Intriguingly, UP6 transcription factors are sporadically expressed in an uncoordinated manner prior to this transition, indicating that co-expression in the same cell is related to suppression of DOWN3 (Fig 2-1Liii). Late intermediates that do not induce UP6 accumulate in a state expressing DOWN3 and Transient networks without the cell division networks (UP3/4) suggesting that cells in this state require proliferation to upregulate UP6, consistent with prior reports that cell proliferation is critical for iPSC generation. Based on correlation of the cell states, this induction of UP6 reflects a major inflection point in gene expression wherein the MEF character of the cell is finally lost (Fig 2-1M). This represents the largest change in cell identity, greater than the initial loss of DOWN2 and induction of the Transient network. Our data uncover that first the downregulation of DOWN2 and induction of the Transient network, and later of the DOWN3/Transient networks together with the induction of the UP6 program are major bottlenecks.

The loss of MEF identity at the DOWN3 to UP6 transition raised the question of how these events relate to the mesenchymal to epithelial transition (MET), a previously described key bottleneck of reprogramming. Taking well established mesenchymal and epithelial genes of MET, we surprisingly find that mesenchymal genes fall gradual beginning with the repression of DOWN2 through to the induction of UP7/8, while epithelial genes become expressed at high levels together only after DOWN3 is repressed, in UP6 positive cells. Consistent with this, sorting for E-Cadherin (Cdh1, the surrogate marker for the MET in the reprogramming field) expressing cells, at day 6 post OSKM induction (Supp Fig S2-11), we more strongly enriched for cells which expressed UP6 without DOWN3, and depleted for non-productive Interferon and Osteogenic states (Fig 2-1Ni/ii). Notably, at this d6 timepoint we detect a large latency of induction of UP7/8, as both sorted and unsorted cells accumulate in UP1-6 states. Cdh1+ sorted cells also existed in many DOWN3 positive states, indicating that CDH1 protein is produced stochastically prior to other epithelial genes, similar to other UP6 TFs (Supp Fig S2-11F/G). The co-expression of epithelial markers of MET happens concomitantly with the downregulation of DOWN3 alongside increasing UP6 expression, whereas the mesenchymal markers of MET are gradually lost at differing rates, suggesting these two processes are not strictly coordinated (Fig 2-1Niii).

Together our data give rise to the following model (Fig 2-1O). Upon OSKM induction, cells acquire ectopic gene expression, either gaining Transient network expression or inducing differentiation/inflammatory responses (Osteogenic/DOWN1/Interferon), events have been previously associated with bifurcation into non-productive reprogramming paths. The next major step is loss of DOWN2, associated with lower DOWN3 expression, lower mesenchymal marker gene expression, higher Transient network expression, and a general increase in UP1-5 network genes including chromatin and replication regulators. This state potentiates the induction of UP6, which increases while DOWN3 is still expressed at low levels. High UP6 triggers the final loss of DOWN3 and Transient expression, and the simultaneous expression of epithelial genes, which correlates with a major shift in cell identity. Eventually, these cells express the final UP7/8 network

genes to establish to iPSC state. At different points along this path arise cells expressing the Osteogenic, Neural, and Epiblast gene signatures, and stalled non- dividing intermediates. These results were reproduced in 3 other experiments (Supp Figs S2-2, S2-6, S2-10).

Next, we wanted to understand the regulatory mechanisms underlying the step-wise changes of the DOWN, UP, and Transient networks along the productive reprogramming path by analyzing CpG content, DNA methylation, chromatin accessibility, histone modifications, and transcription factor occupancy, properties which have been extensively examined during reprogramming. We found that the promoters of more cell type-specific DOWN1/2 genes often lack CpG islands and represent low and intermediate CpG density promoters (LCPs and ICPs), whereas those of DOWN3/4 genes that become repressed later in reprogramming more often contain CpG islands (HCPs, high CpG density promoters) (Fig 2-2A). The converse applies to upregulated genes, where the most broadly expressed genes (UP1-5) often carry promoters with CpG islands and those induced last (UP7/8) largely lack CpG islands (Fig 2-2A). In line with these observations, the promoter regions of UP7/8 genes are on average highly methylated in MEFs and reprogramming intermediates but not in the pluripotent state, in agreement with the control of LCP-containing genes by DNA methylation (Fig 2-2B). Similarly, promoters of DOWN1/2 genes are more methylated at late reprogramming stages (Fig 2-2B). These results are consistent with previous studies showing that DNA methylation is more dynamic at tissue specific genes, linking promoter architecture to timing of expression during reprogramming.

Next we found that broad network genes have more open chromatin sites throughout reprogramming while the most-narrow networks (DOWN1, UP7/8) only have open chromatin sites when they are expressed (Fig 2-2C), suggesting that enhancer usage may differ between networks. We next analyzed what types of enhancers are associated with different gene networks, using previously defined MEF (ME), transient (TE), and ESC (PE) cell enhancer annotations[4] (Fig 2-2D). Approximately 70% of enhancers located within +/- 20kb of DOWN1/2 transcription start

sites were MEF enhancers. DOWN3 genes had a lower proportion of MEF enhancers but more transient and pluripotency enhancers, and DOWN4 was even more skewed towards pluripotency enhancers. This suggests that broadness of DOWN3 and 4 expression is due to the use of new enhancer sites which were not active in MEFs. Of all network genes, transient genes were most strongly enriched for transient enhancers. UP1-8 genes are associated with 70+% of pluripotency enhancers, and all are depleted for MEF enhancers, particularly UP6-8.

Since UP genes have few differences in the proportion of PE enhancers nearby, we next considered whether the subclass of pluripotent enhancer was differentially distributed; gradually induced (E13) vs late induced(E16/17/18), and intergenic vs intragenic(E14-15) (Fig 2E). UP (1-5) gene sets were enriched for intragenic enhancers (E14/15) situated in constitutively transcribed genes[4], (Fig 2-2E/F), suggesting that the transcribed state of these enhancer regions promotes the early induction of these genes. UP4-8 genes are enriched for the gradually induced E13 enhancers, while UP6/7/8 genes are depleted of E14 enhancers, consistent with their late induction, and are instead enriched for E16, E17, and E18, which open late in reprogramming. This suggests that the difficulty of upregulating UP6-8 networks is linked to a major shift in pluripotency enhancer usage, including a change from intragenic to intergenic enhancer usage. A similar observation applies to the DOWN programs, where the DOWN1 network is enriched only for intergenic MEF enhancers (E5,6,9,10), while DOWN2/3/4 are enriched for all MEF enhancers including intragenic E7/8. Intriguingly, DOWN3, DOWN4, and Transient are most enriched across various MEs, TEs, and PEs, suggesting these gene networks can be regulated by different enhancers over time.

To understand if specific transcription factors control distinct enhancer elements, we scanned for motifs in each of the enhancer subclasses. We found that MEF enhancers were more strongly enriched for Jun, Fos, Fra, Runx, and Tead motifs, and that Pluripotent enhancers, particularly E13, E15 and E17, were most enriched for Klf, Oct, Sox, and the Oct-Sox-Tcf-Nanog

compound motif (Fig 2-2G, Supp Fig S2-12). The Oct, Sox, Klf motifs, and the compound motif had the highest relative motif density among PEs E13, E15, and E17; we additionally found the pluripotent TF Esrrb had specifically high motif density at these enhancers, consistent with the idea that TFs expressed late in reprogramming contribute to the opening of certain enhancers. Surprisingly, the transient E11/12 and pluripotent E13/14 enhancers contained strong motif enrichment for both the somatic and reprograming transcription factor motifs, whereas E15-E18 have reduced AP-1 and Runx motif density. This suggested that these transient and less cell type specific pluripotent enhancers, and thereby broad UP/DOWN and transient genes, were regulated by both types of TFs. Indeed, many somatic TFs are DOWN3 network genes, such as Jun and CEBPB, are broadly expressed over much of reprogramming (Fig 2-2J).

We took advantage of previously published bulk ChIP-seq profiles for the reprogramming factors generated at 48hours of MEF reprogramming culture, a stalled pre-iPSC intermediate, and ESCs[4]. We found that genes were increasingly enriched for binding of the reprogramming factors from the most cell type-specific DOWN1 to the broadest program DOWN4 (Fig 2-2G). A similar result was obtained for UP programs, with broadly expressed up-regulated genes enriched for binding by all reprogramming factors across all reprogramming stages, and the most specific UP6/7/8 genes are specifically lacking early OSKM binding events (Fig 2-2G). Transient genes were enriched for O,S and K binding throughout reprogramming, except in ESC.

We next looked at how somatic TFs may be contributing to the dynamic chromatin remodeling during reprogramming (Fig 2-2H). As with the OSKM TFs, somatic TFs were enriched at the broad networks (DOWN3/4, UP 1-5) and more depleted in the narrow networks at the early time point. The definition of the step-wise induction and repression of genes based on single cell data allowed us to uncover features of genomic architecture (CpG content, enhancer location, motif content) that define the complex transcriptional changes underlying reprogramming.

Given that our data indicated differential regulation of our gene expression networks by binding of reprogramming factors, somatic factors and pluripotent factors, we asked what effect an increase in the expression level of certain reprogramming factors, the ectopic expression of a somatic TF, or the precocious induction of an ES specific TF, would have on the reprogramming path. To test this, we infected MEFs carrying the OSKM-cassette 2 days prior to induction of OSKM with either the somatic factor Jun, or the reprogramming factors Oct4 or Klf4, or the specifically PE linked transcription factor Esrrb, and analyzed the ectopically expressing MEFs and the reprogramming process over time with a single cell RNA-seq time course (Fig 2-3A, Supp Fig S2-13). Each time course was compared to control OSKM-cassette cells infected only with RFP expressing virus. As expected, Jun blocked reprogramming (no cells past the trajectory score of 0.067), Klf4 and Oct4 had moderate effects on reprogramming, and Esrrb strongly enhanced iPSC colony formation (Supp Fig S2-13C-E).

We found that all ectopic TFs allowed the downregulation of the most MEF specific genes, and, except in the JUN time course, that enabled the subsequent upregulation of ESC specific genes (Fig 2-3B). This demonstrates that even when a somatic transcription factor is overexpressed, cells move along the trajectory score defined by the most cell type specific genes (Supp Fig S2-13F) and that Jun expression does not especially enforce the starting state.

Before defining new expression changes linked to ectopic TFs, we first analyzed the average expression of our previously established gene networks along the reprogramming trajectory, to compare changes to the order of transcriptional changes established in normal reprogramming (Fig 2-1). The average expression of the DOWN1-4, Transient and UP1-8 networks was similar among all time courses (Fig 2-3C, Supp Fig S2-14). Thus, regardless of whether reprogramming was enhanced or blocked, the Transient network was expressed and the general gradual stepwise regulation of the UP and DOWN networks was preserved. However,

the proportion of cells at a given trajectory score differed between experiments (Fig 2-3D), suggesting the proportion of cells in reprogramming intermediates differs by experiment.

Indeed, applying CENICS to perform cell state classification, using the same networks as time course 1, we found all time courses established the same intermediates as the control time course but in different proportions (Fig 3E). Expression of the previously described non-productive path networks (DOWN1, Interferon, and Osteogenic) were generally found in fewer cells in the Jun, Klf4, Oct4, and Esrrb time courses. We found that ectopic expression at d0 only led to major cells state changes in Klf4 expressing MEFs, with Transient gene expression in 7.5% of starting cells, indicating that Klf4 is sufficient to strongly induce these genes. By d3, Klf4 cells were found overrepresented in Transient expressing states and the earliest arising UP6 cells (still co-expressing DOWN3) were depleted in the presence of Jun overexpression, increased in the Oct and Klf4 time courses, and dramatically amplified in the Esrrb time course. Klf4 and Esrrb overexpression both enhanced the downregulation of the strongly MEF-specific DOWN2. Surprisingly, although less efficiently than in control reprogramming, Jun expressing cells can still upregulation the Transient network and downregulate DOWN2. After d3 all time courses except Jun reach UP7/8 expressing states, but more Esrrb cells reached it sooner.

Over time, cells accumulate in stall points including the typical DOWN3 and Transient expressing state: In Oct4 and Klf4, cells heavily accumulated in a DOWN3, UP6, and Transient expressing state, while cells from the Oct4 and Esrrb time course also were overrepresented in UP1-6 expressing states without DOWN3, beyond the typical stalling points. A very small proportion of Jun cells manage to induce low UP6 expression, but the majority of cells stall before reaching those advanced states, indicating that Jun overexpression impairs multiple transitions: the exit from the starting somatic states, early reprogramming intermediates and blocks the silencing of DOWN3 and Transient networks. Klf4, Oct4, and Esrrb all strongly enhance the generation of late reprogramming intermediates but surprisingly these intermediates do not

immediately resolve into fully established iPSCs. Specifically, Klf4 and some Oct4 cells remain stalled in DOWN3/Transient expressing states, while Esrrb and some Oct4 cells are delayed in UP1-6 states awaiting UP7/8 upregulation, consistent with the distinct genomic features of UP7/8 genes (Fig 2-2). Despite this delay, Esrrb makes more UP6 intermediates earlier, which resolve more efficiently than in other time courses. Consistent with early and efficient UP6 expression, expression of ectopic ESRRB and OSKM for only three days results in iPSC colony formation (Supp Fig S2-13E). Taken together, these data show that the cell states and stepwise order of events are basically the same in all conditions, when considering previously defined networks. However, the proportion of intermediates and the intermediates where cells accumulate over time (stall) differ.

Despite the conservation of average network gene expression along the trajectory, specific genes were expressed or downregulated precociously with respect to the trajectory score, such as the UP6 TF Tfcp2l1 upon Esrrb overexpression (Fig 2-3F, Supp Fig S2-15). Another example is Klf4 induced precocious induction of Cdh1 uniquely among MET signature genes (Supp Fig S2-16). These divergent targets raised the possibility that new gene patterns arose in the ectopic time courses. Therefore, we applied GEND and identified several unique networks specific to each overexpression experiment, some of which included over one thousand genes, consistent with control and overexpressing cells mixed poorly in tSNE space (Fig 2-3G-J). Some of these networks were induced by the ectopic TF in starting MEFs, such as Klf4-21, Klf4-26, and Oct4-37, while others were induced in intermediate reprogramming states (for instance Esrrb-23, Oct4-8, and Jun-11) (Fig 2-3I).

Intriguingly, the two late stalling cell states (DOWN3, Transient and UP6, and UP1-6) often have the highest expression level of intermediate ectopic networks, and transitions out of these states towards completion of reprogramming, or the block of Jun reprogramming, are associated with their down regulation (Fig 2-3J). Thus, we conclude that at least some of these ectopically

expressed programs block the efficient induction of the pluripotency networks UP6-8. Many of these ectopic programs are enriched for metabolism related gene ontologies, but some of the Klf4 and Oct4 ectopic network genes are related to differentiation processes and typically expressed in few tissues (Fig 2-3H, Supp Fig S2-17A). Intriguingly, they include many transcription factors linked to various differentiation processes, in particular the broad ectopic programs Jun-6 and Oct4-13 which have over 40 TFs each, in contrast to Esrrb-19, which includes precocious expression of many pluripotent TFs linked to stem cell maintenance (Supp Fig S2-17B/C).

Taken together, in all instances of MEF reprogramming, ectopic expression networks can be observed. In normal OSKM reprogramming, we observe the Transient network, while when the reprogramming factor stoichiometry is altered or other transcription factors are added we uncovered additional ectopic networks. Depending on their nature, these networks enhance, stall, or block reprogramming (Fig 2-3K). Consistent with this idea, overexpression of MyoD in addition to OSKM induces skeletal muscle genes and completely blocks iPSC reprogramming (Supp Fig S2-18).

We have shown all examined MEF reprogramming processes proceed via the stepwise downregulation of the somatic program and the overlapping stepwise upregulation of the target program, induction of key pluripotent genes after shutoff of key MEF programs, and induction of transient/ectopic gene signatures linked to reprogramming efficiency that are dependent on the reprogramming factor cocktail. Yet it remains unclear whether this logic applies when other somatic cells are reprogrammed by OSKM. To uncover the dynamics of different reprogramming systems, we obtained ssRNA-seq data from neural stem cell (NSC)- and keratinocyte- to-iPSC reprogramming, enabling the first single cell based comparison of three systems (MEF, NSC, Keratinocyte) towards a common iPSC target.

We found that iPSC induction from NSCs was extremely inefficient (NSC+Dox), and the overexpression of Esrrb (NSC+Esrrb), as seen in the MEFs, enabled efficient reprogramming,

and a simple split of the culture enabled reprogramming in a few cells but with high latency (NSC+Split) (Fig 2-4A-C, Supp Fig S2-19). Similar to MEF reprogramming, the most NSC and ESC specific genes were mutually exclusive in all cases (Fig 2-4D), indicating that reprogramming NSCs first shut down their most NSC specific genes and later induce the most ESC specific genes and allowing us to order cells along the reprogramming axis (Fig 2-4E).

Using GEND to determine the gene expression networks for NSC reprogramming, we detected several somatic ($DOWN^{NSC}$ 1-4) and pluripotent (UP^{NSC} 1-5) networks that gradually change along the trajectory axis, and six ectopic programs (Fig 2-4F, Supp Fig S2-19D-E). Distinct GO terms were associated with the ectopic programs: NSC- E2/E3 (Neural), E4 (Metabolism), E5/ E401 (Cell adhesion and vasculature) and E9/ E13 (Wound/Immune response) (Supp Fig S2-19D), and E2 contained a large number of differentiation-associated TFs. E5/401 most resembled the expression pattern of the MEF Transient network (Fig 2-4F).

The UP^{NSC} and $DOWN^{NSC}$ networks could be divided into more and less cell type specific groups (Fig 2-4G, H). Specifically, $DOWN^{NSC}$1-2 contained the most NSC-specific genes (such as Npas3 and Olig2) and displayed the strongest neural ontology terms, while $DOWN^{NSC}$3-4 genes were broadly expressed in many tissues; and UP^{NSC}1-3 contained many ubiquitously expressed genes (including cell division genes in UP1) and UP^{NSC}4-6 the pluripotency regulators, such as Zic3, Nanog, Sall4, Tfap2c, MycN. The NSC-specific ($DOWN^{NSC}$1/2) are downregulated before the more broadly expressed NSC networks ($DOWN^{NSC}$3/4), and similarly the less tissue specific UP genes (UP^{NSC}1-3) are upregulated before the pluripotent-specific genes (UP^{NSC}4-6) (Fig 2-4F). The broad UP and DOWN programs overlap expression, and have lower DNA methylation and higher CpG content at promoters (Fig 2-4I).

Using CENICS to perform cell state classification with these NSC-specific networks (Fig 2-4J), we find that over time, the NSC control time course cells increase in somatic identity by not shutting off the DOWN networks, strongly inducing NSC-E2/E3, and loses cell division signatures

(UP1). In contrast, NSC+Esrrb cells quickly downregulate the NSC specific narrow DOWN1/2 networks, and induce the transient programs E4 and E401, moving towards the iPSC state. Yet the induction of UP^{NSC}4-6, combined with the downregulation of DOWN3/4 and E401, only occurs with delay. NSC+Split cells can read advanced intermediate states, but are unable to efficiently downregulate the DOWN3/4 and E401 networks and upregulate the UP5 genes, indicating that reprogramming NSCs have trouble escaping the broad neural programs even after prolonged time.

Taken together, similarly to MEF reprogramming, upon OSKM induction NSCs can either progress toward non-productive cell states in which the neural cell identity is enhanced, or they may downregulate the NSC specific narrow networks and induce transient networks to reach intermediate cell states. Cells reaching these intermediates then stall before they can downregulate the broad neural and transient networks, and upregulate the pluripotency specific programs UP4-6.

Regarding Keratinocyte reprogramming, we again compared high and low efficiency systems (with and without the Keratinocyte stimulating substrate fibronectin), with both reaching the final iPSC stage but the +Fibro sample reaching it substantially faster and more efficiently (Fig 2-4K-M, Supp Fig S2-20).

The overall characteristics of MEF-iPSC and NSC-iPSC reprogramming were also shared by Keratinocyte to iPSC reprogramming, with antagonistic expression of the most cell type specific genes (Fig 2-4N), gradual movement of cells along a trajectory score towards iPSC states (Fig 2-4O), and unique UP and DOWN gene networks which stratified into narrowly expressed cell type specific genes and broadly generally expressed genes (Fig 2-4P). These properties were again correlated with tissue specificity (Fig 2-4Q), start and end state specificity (Fig 2-4R) and general underlying promoter architecture and DNA methylation (Fig 2-4S). Strikingly, DOWN1/2

had only 1 TF, while DOWN4/5 had the most TFs, with those in DOWN5 having much broader tissue expression than those in DOWN4 (Supp Fig S2-20).

We also detect a range of ectopic programs peaking in intermediate states (Fig 2-4P) and with various GO terms, including E1 and E10 which interestingly possessed enhanced mesenchymal characteristics, including mesenchymal TFs and collagens.

Upon OSKM induction (d3), CENECs revealed that a subset of +Fibro cells induce ectopic programs (E1,E6,E7,E10), some cells shift from DOWN1 expression to DOWN2/3, and very few cells reach states with low DOWN5 and with UP4-UP7 expression (Fig 2-4T), again showing the upregulation of later pluripotency genes is associated with the downregulation of the broadest somatic program. By d6, some +Fibro cells reach UP4-UP9 without DOWN network expression, but the majority of the most advanced cells are stalled prior to the UP8/9 induction, which is accomplished efficiently by d13. Non-productive states expressing mesenchymal ectopic programs emerge accumulate late in +Fibro. In the less efficient control, the majority of cells do not induce UP7 efficiently even by d17 and accumulate in intermediate states with DOWN4/5 and the transient E7 network expression. Thus, reaching a high UP7 low DOWN5 state, and UP8/9 induction represent the key bottlenecks of keratinocyte reprogramming. Overall, NSC and Keratinocyte iPSC reprogramming proceed via similar stepwise transition through intermediate stages, but have, as in the MEF reprogramming perturbation experiments, context specific differences based on their starting state in the stalling points encountered, and again transiently expressed ectopic programs are associated with reprogramming blocks. NSCs have difficulty shutting down the broad general identity and neural ectopic networks, while Keratinocytes are particularly inhibited at upregulation of the narrow UP genes.

We next investigated which genes all three systems had in common, either as shared-repressed, shared-transient, or shared-induced during reprogramming using general induction/repression criteria (Fig 2-4U). Approximately 2100 downregulated genes are shared in

all three systems, including Jun and Fos family members, Runx1, and Cebpb, all of which have enriched motifs underlying regulatory sites during MEF reprogramming and interfere with the progression to the pluripotent state (Fig 2-2). This finding suggests that OSKM counteracts a conserved set of broadly expressed TFs in each of the three reprogramming systems, hinting at a mechanism through which OSKM can disassemble distinct cell identities. Ontology analysis shows that the downregulation of shared genes proceeds from differentiation/development terms to the regulation of organelles and membrane trafficking (Fig 2-4V, Supp Fig S2-21).

Over 1500 genes are upregulated in all three systems, sharing many of the pluripotency transcription factors. Ordering these genes along reprogramming time revealed that genes linked to general processes related to RNA and chromatin biology (Ncor1, Pcgf3, Tbl1xr1, Nipbl, Ctcf, Kdm5c, REST, and Jarid2) are the first to be upregulated, followed by cell cycle regulators, and culminating in the induction of stem and pluripotency cell regulators (Tfcp2l1, Tead4, Tfap2c, Nr5a2, Nanog, Mycn, and Zic3) (Fig 2-4V, Supp Fig S2-21).

A very small proportion of transiently expressed genes was shared across the three reprogramming processes (including Ceacam1/2, Myb, Crym, and Chst3) (Fig 2-4U), consistent with the fact that ectopic genes were specific to each starting cell type. This indicates that each reprogramming cell type will generate a unique ectopic signature which will have differential effects on reprogramming efficiency. We hypothesized that divergent ectopic expression is a conserved feature of other reprogramming processes. To test this, we analyzed data from three published direct reprogramming systems; MEF-to-iNeuron and MEF-to-iCardiomyocyte (iCM), and MEF-to-Endoderm Progenitor (Supp Figure 2-22). Similar to iPSC reprogramming, we found that the most cell type specific networks were exclusively expressed, that the downregulation of the starting cell program was associated with induction of ectopic genes, and that each ectopic program had distinct GO terms, in all three systems.

Based on our analysis of multiple reprogramming systems, we conclude that these general principles define cell fate transition; 1) the most tissue-specific identity genes must first be silenced, 2) a broad general gene expression program is always maintained longer and together with reprogramming factors create novel ectopic programs, accompanied by induction of the broader target cell programs. Over reprogramming time, progressing cells eventually 3) silence the general somatic identity and ectopic program, and finally allow for 4) induction of the most target cell specific identity genes (Fig 2-4W). The nature of the ectopic programs has a large influence on the progression of reprogramming events. This represents a new framework for our understanding of the reprogramming process, and raises the question of whether differentiation processes occur via a similar logic.

Conclusion.

A major question in the reprogramming field has been how somatic cells progress to the pluripotent state. Our analyses show that when reprogramming a cell to an end state, cells first need to suppress the narrow, tissue-specific gene expression programs. Following this, the gradual decommissioning of the broad general cell identity will allow for the establishment of various ectopic states that allow for the expression of genes normally expressed in different lineages. We show that the nature of the ectopic programs has a large influence on the progression of reprogramming events. As the poised cell state is established, various cell cycle and chromatin regulators are induced, likely leading to an increase in proliferation that has been associated with faithful reprogramming[26–30]. The induction of MET, probably favored due to the direct targeting of MET regulators by several reprogramming factors[31], is followed by effective pluripotency gene upregulation, which in turn restricts the expression of other developmental targets. At this stage in reprogramming, cells must neutralize their ectopic networks and concurrently begin the induction of target cell programs in a stepwise manner in order to fully reprogram.

We show that despite gene expression differences, the enhancement or suppression of OSKM reprogramming follows the same path to pluripotency but enables variation in the latency. For example, the overexpression of Esrrb enables the progression along the path to pluripotency for a vast fraction of the cells by dramatically enhancing the transition to the poised expression state, emphasizing the importance of MEF-program silencing for reprogramming progression. By defining distinct steps on the path to pluripotency and insight underlying these progressions, our single cell analysis of the reprogramming process will allow for a targeted dissection of the mechanisms underlying each of these steps.

Figure Legends

Figure 2-1. Single cell sequencing data of thousands of cells capture the transcriptional heterogeneity of reprogramming, latency of pluripotency induction and a number of etopic networks arising at an intermediate state. (A) Schematic of MEF-to-iPSC reprogramming time course 1 with time points taken for single cell sequencing indicated. The reprogramming culture was passaged four times post d15. Single cell transcriptomes are visualized on a tSNE embedding and colored by time point. (B) tSNE embedding of single cell transcriptomes with cells for each time point shown separately. (C) tSNE plots showing the normalized expression of selected marker genes in individual cells for all cells of time course 1. (D) Schematic of Gene Expression Network Discovery (GEND) analysis. (E) Average expression plots of genes assigned to the various Down (i), Up (ii), and Ectopic (iii) networks defined by GEND. Selected transcription factors associated with each network are labeled within each panel. (F) Barplot of the proportion of cells that express networks above 30% of their max expression. (G) Box plots showing the log ratio distribution of normalized expression for the network genes (Log2(RPKM+1)) from MEF- vs ESC- whole cell RNA Seq data. (H) Violin plots illustrating the distribution for each network of the number of tissues that are moderately expressing each gene (RPKM>1) from a tissue compendium. (I) Density scatter plots showing the pairwise relationship between the average expression of selected networks for individual cells.(J) Visualization of difference in average normalized MEF- and ESC-signature genes (log2 fold>5, p-value>0.01) expression levels, or the 'trajectory score' (i), for each cell of time course 1. (ii) Line graph showing the smoothed average normalized ESC- and MEF- signature gene expression values over the trajectory score. (K) Graph of the smoothed average expression of individual networks against trajectory score, separated by class. Individual networks are colored according to the key at top. (L) Heatmap of the top 30 cell state combinations derived by CENICS, showing the combination of programs expressed over 30% of maximum program expression (i), the percent of cells in each combination state (ii), and the average normalized expression for select networks in each state (iii). (M) Pairwise Pearson

correlation heatmap of cells expressing program combinations from (L). (N) As in (L), but for cells sorted for E-Cadherin on day 6. (O) Illustrated model showing the productive/non-productive reprogramming cell states during progression.

Figure 2-2. Features of genomic architecture define the regulatory mechanisms underlying the step-wise changes of the DOWN, UP, and Transient networks along the productive reprogramming path. (A)Stacked bar graph showing the CpG content of promoters for genes in selected networks, classified as High CpG (HCP), Intermediate CpG (ICP), or Low CpG (LCP) promoters. (B) Heatmap illustrating the average percent of CpG methylation at gene promoters for selected networks, at four distinct stages of iPSC reprogramming. (C) As in (B), but showing the average number of ATAC peaks per program. Peaks are assigned to genes by taking genes assigned to a given network and tallying peaks that overlap +/- 20Kb TSS windows, which is then normalized and scaled. (D) Stacked bar graph showing the enrichment of chromatin states along selected networks. Chromatin states are computed using ChromHMM by learning a stacked 35-state model. States are assigned to specific enhancers trajectories (E5-E10 = MEF enhancers (ME), E11-E12 = transient enhancers (TE) and E13-E18 = pluripotent enhancers (PE)) based on learned state enrichments.

(E) Stacked bar graph showing the relative fraction of selected chromatin states that are intergenic, intragenic, or within +/- 2kb of the start site (TSS) of a gene. (F) Heatmap showing the enrichment of selected networks against chromatin states. Enrichment is computed by taking genes assigned to a given network and computing states that overlap +/- 20Kb TSS windows, and normalizing by a foreground containing all genes. (G) Density distribution of selected motifs centered at specific enhancer trajectories defined in (D). (H) Density distribution of the Jun/AP-1 motif near genes within selected networks centered at ATAC peaks and overlapping with different enhancer classes as defined in (D). (I) Normalized expression is shown for selected Jun, Fos, Fra, Runx, and Tead genes. (J) Enrichment for ChIP-seq peaks for selected transcription factors

at four distinct stages (MEF, 48h, pre-iPSC, ESC) of iPSC reprogramming, within +/- 20kb of gene promoters for the given networks.

Figure 2-3. Cell states and the stepwise order of events are conserved when the reprogramming process is modified through the overexpression of certain reprogramming factors. (A) Schematic showing viral overexpression of selected TFs in MEFs carrying the tet-inducible OSKM cassette, followed by scRNA-seq across various time points. (B) Scatter plot of average normalized MEF and ESC- signature gene expression across all individual cells from each overexpression experiment. The key at top indicates the density of cells. (C) Graph of the smoothed average expression of individual networks from time course 1 (Fig 2-1) against trajectory score for each experiment described in (A). Individual networks are colored according to the key at top and separated by network class. (D) Density distribution of cells along the trajectory score split by control for each experiment in (B). (E) Heatmap showing cell states derived by CENICS, key (black, white; top) followed by heatmaps (bottom) for each control and overexpression experiment with the percent of cell in each of the top 30 more frequent states is shown by time point. (F) Average expression of Tcfp2l1 by cell state per experiment. (G) tSNE plots divided by control and (i) Jun, (ii) Klf4, (iii) Oct4 and (iv) Esrrb overexpression for the average expression of novel ectopic network programs with corresponding violin plots illustrating the distribution of expression level. (H) Violin plots showing the number of expressed tissues (>1 RPKM) per gene in each network from a tissue compendium, for the novel ectopic networks. (I) Graph of the smoothed average expression of (i) Jun, (ii) Klf4, (iii) Oct4 and (iv) Esrrb ectopic networks against the trajectory score. (J) Heatmap showing the average normalized expression for the novel ectopic networks in each cell state from (Fig 2-3E). (K) Illustrated model showing the expression of various ectopic networks along the reprogramming trajectory.

Figure 2-4. The analysis of various reprogramming systems defines general principals of cell fate transitions. (A) Schematic of neural stem cells (NSCs)-to-iPSCs reprogramming time course via the induction of OSKM, OSKM+ overexpression of Esrrb or OSKM+ culture splits, with time points taken for single cell sequencing. (B) tSNE plots of NSC-to-iPSCs for each condition with NSC and iPSC populations outlined. (C) Violin distribution plots showing the average normalized expression of the UP7 network (from Fig 2-1) by timepoint and experiment. (D) Scatter plot of average normalized NSC- and ESC- signature gene expression across all individual cells from each experiment. (E) Density distribution of cells along the trajectory score for each experiment which day 0 cells having their own distribution. (F) Graph of the smoothed average expression of individual NSC-specific networks against trajectory score for each experiment described in (A). Individual networks are colored according to the key at top and separated by network class and experiment. (G) Violin plots showing the distribution of the number of expressed tissues from a tissue compendium, for genes in NSC-specific networks in (F). (H) Box plots showing the log ratio distribution of normalized expression (Log2(RPKM+1)) for NSC- and ESC- signature genes along the NSC-specific networks. (I) Heatmap illustrating the average percent of CpG methylation at promoters of NSC network genes within four distinct tissues (top) and a stacked bar graph showing the relative fraction of each CpG promoter class(bottom) for NSC-specific networks. (J) Heatmap of the top 30 cell states found by CENICS, with key (black, white; top) followed by heatmaps (bottom) for each NSC experiment with the percent of cell in a given time point in each cell state. (K) Schematic of Keratinocytes-to-iPSCs via, either with control plated-gelatin or with fibronectin-s reprogramming time courses with time points taken for single cell sequencing indicated. (L-T) as in (B-J), but with Keratinocyte-specific networks. (U) Venn diagrams illustrating the shared genes in broadly defined Down, Up and Transient classes (>2 fold expression in beginning, middle, or end of trajectory) across time course 1, the Keratinocyte experiments and the NSC experiments. (V) Gene Ontology (GO) terms associated with shared genes in (U). The top three GO terms for early to late gene bins along the trajectory were selected to populate the

list of terms. A dot represents significant presence of a GO term in that network, colored and sized by its significance, as shown by the key on the right. (W) Model showing the significant gene changes that are coordinated with all observed iPSC reprogramming progression events.

Figure S2-1. Expression of canonical marker genes confirm latency of pluripotency. (A) Total fraction of cells captured by time point. tSNE embedding showing the (B) expression of the doxycycline (dox)-inducible polycistronic OSKM cassette, (C) simultaneous expression of endogenously-encoded Pou5f1 and Sox2, (D) silencing of Xist and inactive X reactivation, and (E) known somatic and pluripotent marker genes. (F) Total fraction of cells captured from each timepoint expressing marker genes in (E).

Figure S2-2. Various dimensionality reduction techniques recapitulate transcriptional heterogeneity of reprogramming, latency of iPSC induction. (A) Supplemental embedding of single cell data using (i) UMAP, (ii) Spring, and (iii) Palantir methods of dimensionality reduction on Time course 1 colored by time point. (B) As in (A), but facet by time point. (C) UMAP embedding of (i) full and facet (iii) Time courses 2, 3, and 4, containing timepoints described in (ii) schematic. (D) As in (C), but facet by time point. (E, F) Expression of marker genes in Figure 2-1C on embedding described in (A) and (C). As in Figure S2-1C, D but on embeddings described in (A) and (C).

Figure S2-3. Distribution of expression and Gene Ontology terms associated with Gene Expression Networks defined by GEND. (A) Violin plots illustrating the distribution of expression by timepoint for gene expression networks defined in Figure 2-1E. (B) Proportion of cells within a given timepoint with greater than 30% of the max expression of each gene expression network. Gene Ontology terms associated with expression (C) networks in Figure 2-1E and (D) Osteogenic ectopic network. Selected marker genes enriched in (E) neuronal and (F) post-implantation

epiblast networks. (G) Scatter plot of the per-cell average normalized expression of the post-implantation epiblast state as a function of late UP networks where cells are colored according to their Trajectory Score.

Figure S2-4. Culture conditions influence the gene expression changes in the first 48hrs of reprogramming. (A) Summary of the experimental conditions to assess the influence of culture conditions on gene expression at 48hrs of reprogramming. Briefly, wildtype (wt) or tetO OSKM MEFs were split and put under the various indicated culture media for 48hrs, containing either knockout serum replacement (KSR) or fetal bovine serum (FBS), with and without doxycycline, as indicated. Cells were isolated at 48 hrs for single cell RNA-seq. For the following panels in this figure, the different conditions are referred to as samples 1-5. (B) Heatmap of the expression of all genes that were at least two-fold more highly expressed in cells from a particular sample, with their names listed on the right. Notably, few genes were detected with this cut-off indicating that most gene expression changes at 48 hrs are relatively small, consistent with prior findings 35. We observed six categories of gene expression differences (noted with by the color code on the left), with genes up in FBS+OSKM (5 UP), up in KSR with dox (regardless of OSKM induction) and down in OSKM+FBS ("2 and 4 Up/5 DOWN"), up in KSR with dox (regardless of OSKM induction) ("2 and 4 UP"), down in KSR with dox (regardless of OSKM induction) ("2 and 4 DOWN"), down after dox addition (regardless of OSKM induction ("2, 4 and 5 DOWN"), and up after OSKM induction ("4 and 5 UP"). (C) Up to four most significantly enriched GO terms for each gene group in (B). Consistent with the shift observed in MEF signature gene expression in samples which were under KSR, the majority of ontology terms enriched in samples under KSR at 48hrs were fibroblast-related. (D) Normalized expression for several example genes from the KSR-induced genes ("2 and 4 UP") and the OSKM-induced genes ("4 and 5 UP") from (E), displayed on the tSNE map of time course 1. "2 and 4 UP" genes (Dcn, Itm2a, and Cxcl5) showed preferential induction near the top of the tSNE and reduced expression in intermediate and pluripotency cell

states. “4 and 5 UP” genes (Il1rn, Apod, and Tek) were also induced by OSKM induction and repressed in pluripotency-like cells. Thus, the first wave of gene response to OSKM at 48 hrs did not include pluripotency gene targets. (E) Stacked bar graph showing the number of genes in each network for each labeled sample in (A). (F) Normalized expression for several example genes from the KSR-induced genes (“2 and 4 UP“) and the OSKM-induced genes (“4 and 5 UP”) from (E), displayed on the tSNE map of time course 1 (Fig 1B). “2 and 4 UP” genes (Dcn, Itm2a, and Cxcl5) showed preferential induction near the top of the tSNE and reduced expression in intermediate and pluripotency cell states defined in Figure 2. “4 and 5 UP” genes (Il1rn, Apod, and Tek) were also induced by OSKM induction and repressed in pluripotency-like cells. Thus, the first wave of gene response to OSKM at 48 hrs did not include pluripotency gene targets.

Figure S2-5. Relative timing of pluripotency regulator expression. (A) Table indicating the UP network assignment for pluripotency regulators. (B) tSNE plots of all cells in time course 1, showing the transcript level of each of the pluripotency regulator in (A). Notably, some pluripotency regulators are expressed in cells outside of the late pluripotent states (for instance Gdf3, Dppa5a, Sall1, Sall4, Nanog). However, the early induced pluripotency regulators were often not co-expressed in the same cells. (C) tSNE plots of the simultaneous expression of late UP network genes indicated in (A).

Figure S2-6. Supplemental networks defined by GEND. (A) tSNE plots showing the expression of various supplemental gene networks defined by GEND. For the same networks in (A), we show the (B) distribution of expression of by time point, (C) proportion of cells expressing these networks with a threshold of greater than 30% of max expression, and (D) Gene ontology terms. (E) UMAP expression plots on the full time courses 2, 3, 4 embedding of the average normalized expression of main networks defined in Figure 2-1E. (F) As in (E), but with networks defined in (A). (G) Fraction of cells expressing each network defined in Figure 2-1E with a threshold of

greater than 30% of max expression. (H) As in (G), but with network in (A). (I) Pairwise density scatter plots of main DOWN gene networks as a function of UP gene networks, further labeled by Broad and Narrow categories. (J) Average smoothed line plots of main networks with selected ectopic networks on full time course 2, 3, 4 embedding.

Figure S2-7. Lineage tracing experiment reveal the post-implantation epiblast cells differentiated from pluripotent iPSCs. (A) Schematic diagram of lineage tracing experiment where cells are sorted based on GFP. (B) tSNE embedding of cells from experiment in (A), colored by (i) timepoint, (ii) trajectory score, and (iii) selected somatic and pluripotency genes. (C) Average smoothed line plots showing the transcriptional dynamics of programs defined in Figure 2-1E. tSNE expression of (D) GFP construct and (E) pMX Xist transcript, also the (F) distribution of expression as a function of the trajectory score. (G) Density scatter plot of expression of (D) against (E). tSNE plots showing the expression of (H) late UP programs, (I) the post-implantation epiblast state and (J) two supplemental networks with associated (K) Gene Ontology terms. (L) K-means clustering of embedding in (B) and the X-linked GFP expression distribution showing enrichment in late pluripotent-like cells.

Figure S2-8. Transient network captures the upregulation of lineage-specific markers from unrelated lineages during reprogramming. Bar plots showing the enrichment of (A) Nanog+ colonies and (B) percent of colony forming cells or Itbg4+ sorted cells. (C) tSNE and (D) line plots showing the enrichment of selected Transient network genes. (E) DAPI and CRYM staining for colonies as a function of selected time points ranging from Day 0 to Day 12. (F) Scatter plot of average normalized MEF- and ESC-signature genes colored by average Transient network normalized expression. (G) tSNE plot of simultaneous expression of Transient network genes illustrating expression at an intermediate stage of reprogramming.

Figure S2-9. MEF-and ESC- specific gene antagonism is used to define an axis of reprogramming progression. (A) Boxplots showing the distribution of expression within population-level MEF and ESC samples as a function of gene networks defined in Figure 2-1E. (B) tSNE plots showing the average normalized expression of (i) MEF- and (ii) ESC-specific gene sets. (C) Time course 1 single cell density scatter plot of the average MEF identity against the average ESC identity facet by time point. Reduced dimensionality plots showing the trajectory score on Time course 1 (E) tSNE facet by timepoint, (E) UMAP, (F) Spring, (G) Palantir, and (H) fully-embedded tSNE. (I) Monocle2 spanning tree colored by trajectory score. (J) tSNE overlaying Monocle2 Pseudotime showing (K) strong agreement with the trajectory score. (L) Scatterplot showing the trajectory score against an imputed trajectory score (Imputing dropout on MEF- and ESC-signature genes). (M) Scatter plot with a smoothed regression line showing the relationship between the number of expressed genes and transcripts for each cell as a function of the trajectory score. (N) Violin plots showing the distribution of the Euclidean distance between cells that have been stratified into 8 bins by position along the trajectory score. (O) Full UMAP embedding with trajectory score overlay on Time course 2, 3, and 4.

Figure S2-10. Utilization of CENICS to pinpoint bottle necks along iPSC reprogramming. Heatmap of the top 100 cell state combinations derived by CENICS for (A) Time course 1, showing the combination of programs expressed over 30% of maximum program expression, the percent of cells in each combination state. (B) As in (A), but for the top 30 cell state combinations for the full time course 2, 3, 4. (D) The average normalized expression for MET UP/DOWN networks in each state combination from (B).

Figure S2-11. Sorting for E-Cadherin expressing cells, at day 6 post OSKM induction. (A) Diagram of the experimental design of an independent reprogramming experiment in which cells were analyzed at d6 of reprogramming with and without sorting for the surface marker CDH1. (B) tSNE

map for all cells from the CDH1+ sorted and unsorted cell populations. The full tSNE embedding is colored light grey, and cells from the specific sample are colored by their expression (dark grey to purple, color scale shown at top). (C) As in (B), except that the trajectory score is shown for each cell. (D) tSNE plots showing the average normalized expression of main gene networks from Figure 2-1E. (E) Smoothed regression lines illustrating the expression trends of networks in (D) for CDH1+ cells. tSNE plots showing the simultaneous expression of (F) UP8 and (G) a set of genes from UP6-8, normalized expression of selected (H, I) MET signature genes, (J) the simultaneous expression of those MET genes, and finally the (K, L) simultaneous expression trend of these MET gene sets as a function of the trajectory score.

Figure S2-12. Transcription factor control over enhancer elements. (A) As in Figure 2-2A but with additional ectopic networks. (B) The distribution of various histone modification profiles from 4 distinct stages of iPSC reprogramming profiled at the population level around enhancer chromatin states. (C) As in Figure 2-2B but with programs in (A). (D) Motif p-value enrichment heatmap for chromatin states. (E) Selected Motif descriptions used for downstream panels. Motif density scan over TSS of genes within (F) selected intermediate DOWN, Transient, and UP networks; and (G) Early DOWN and Late UP networks.

Figure S2-13. Overexpression of ectopic TFs to initiate iPSC reprogramming. tSNE embedding for the overexpression of each ectopic TF, facet by respective control experiment and colored by (A) timepoint and (B) normalized expression of respective ectopic TF. Bar charts showing Nanog+ colony counts for (C) Jun (D) Klf4 and Oct4 overexpression relative to their respective control experiments. (E) Bar chart of Dppa4+ colony counts for Esrrb overexpression. (F) tSNE plots showing trajectory scores for each over expression experiment described in (A).

Figure S2-14. Expression changes linked to ectopic TF overexpression among main gene networks. Pairwise average normalized expression plots between UP, Transient, and DOWN programs for (A) Jun, (B) Klf4, (C) Oct4, and (D) Esrrb.

Figure S2-15. Relative timing of pluripotency regulator expression upon ectopic TF overexpression. (A) tSNE plots of all cells in each overexpression experiment, showing the normalized expression level of the selected pluripotency regulators.

Figure S2-16. The effect of ectopic TF overexpression on MET signature genes. tSNE plots showing the normalized expression of selected (A) MET Down and (B) MET Up genes for each over expression experiment, facet by their respective controls. (C) Smoothed expression trend line illustrating the relationship between MET Up/Down genes as a function of each experiment's trajectory score.

Figure S2-17. Ectopic programs are enriched for metabolism and various differentiation processes. Dot plot showing Gene Ontology terms enriched in ectopically expression programs utilizing (A) all expressed genes and (B) transcription factor. (C) A table listing the TFs expressed in each ectopic program. (D) Pairwise scatter plots of main gene networks DOWN2-3, Transient and UP 6 against selected ectopic programs where each cell is colored by its trajectory score.

Figure S2-18. Overexpression of MyoD to OSKM induces skeletal muscle genes and completely blocks iPSC reprogramming. (A) Schematic of MyoD experiment. (B) tSNE embedding of MyoD experiment where each cell is colored by its captured time point and facet by control experiment with matching time points. (C) tSNE plots of the average normalized expression of DOWN2-3, Transient and UP7 gene networks, also shown is the distribution of each network's expression by captured time point. (D) tSNE of Tnnt2 normalized expression.

Figure S2-19. Enabling efficient NSC-to-iPSC reprogramming through the overexpression of Esrrb. (A) tSNE plots of gene networks UP7-8 and (B) selected pluripotency regulators for each NSC experiment with respective control. (C) Smoothed line plots showing the relationship between the average normalized MEF- and NSC-signature genes as a function of trajectory score for each experiment in (A, B). (D) Gene ontology and (E) tSNEs of the average normalized expression of the UP^{NSC}, MID^{NSC}, and $DOWN^{NSC}$ expression networks for each experiment in (A, B). (F) A table listing the TFs included within programs defined in (D, E).

Figure S2-20. Enabling efficient Keratinocyte-to-iPSC reprogramming though Keratinocyte stimulating substrate fibronectin. As in Figure S2-19, but with Keratinocyte-to-iPSC reprogramming experiments.

Figure S2-21. Expression and Gene ontology analysis of shared sets of genes among 3 reprogramming systems. (A) Dot plot showing the top 5 gene ontology terms for shared-repressed and shared-induced gene sets of UP and DOWN gene networks. (B) Smoothed line plots over the trajectory score for the shared-repressed and shared-induced gene sets for time course 1, NSC- and Keratinocyte-to-iPSC reprogramming.

Figure S2-22. Cell fate transition paradigm is conserved in other TF-induced direct reprogramming systems. For three published datasets containing MEF- (A) to-Induced neurons, (B) to-induce cardiomyocytes and (C) to-induced endoderm progenitors, we show smoothed line plots of the relationship between the start state program and the respective end state program along the trajectory score. We also include three classifications of gene sets designated to DOWN, Transient, and UP networks along the trajectory score, as well as, Gene Ontology terms of the Transient state program.

Figures

Figure 2-1 – Single cell sequencing data of thousands of cells capture the transcriptional heterogeneity of reprogramming, latency of pluripotency induction and a number of ectopic networks arising at an intermediate state

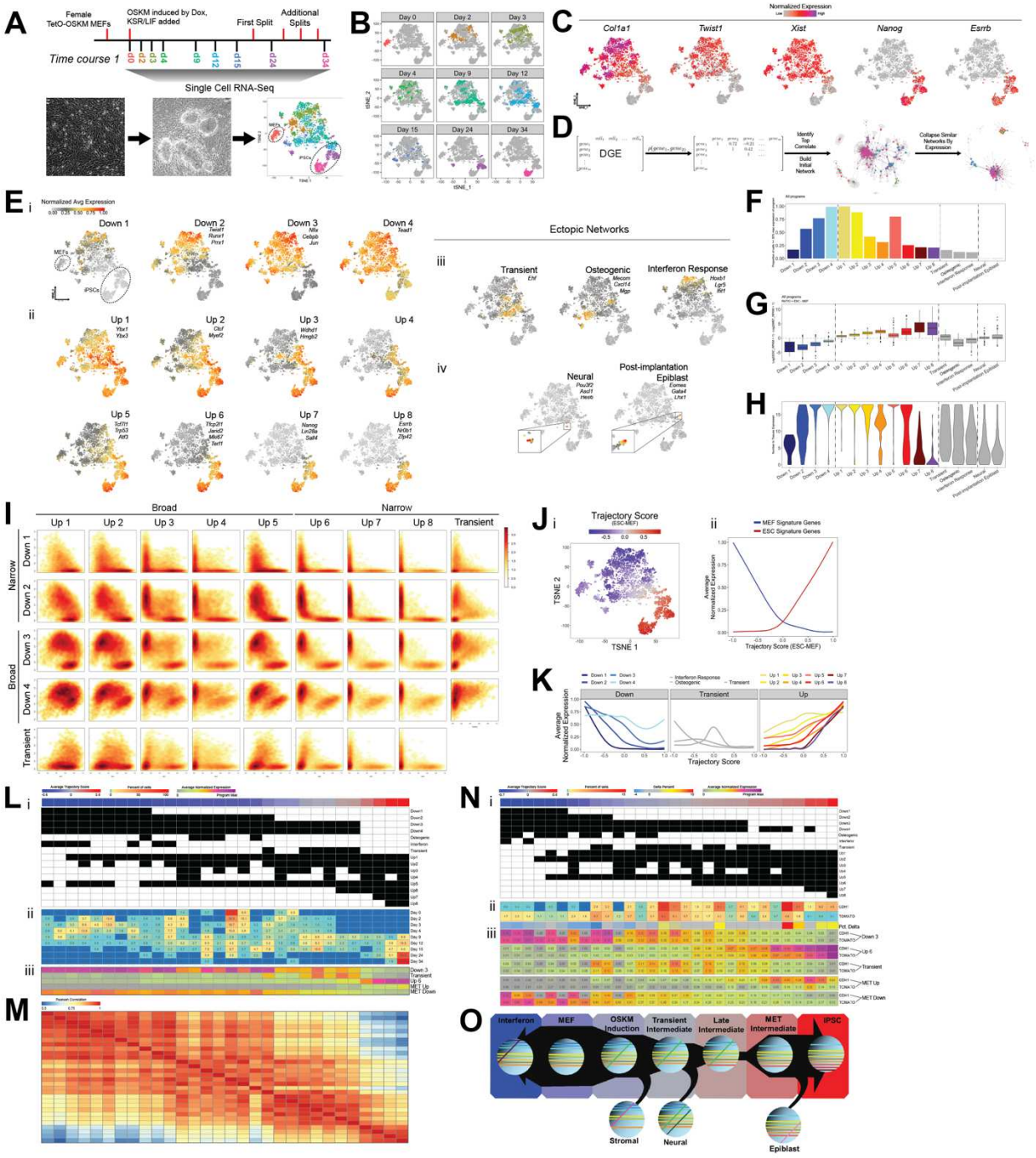


Figure 2-2 – Features of genomic architecture define the regulatory mechanisms underlying the step-wise changes of the DOWN, UP, and Transient networks along the productive reprogramming path

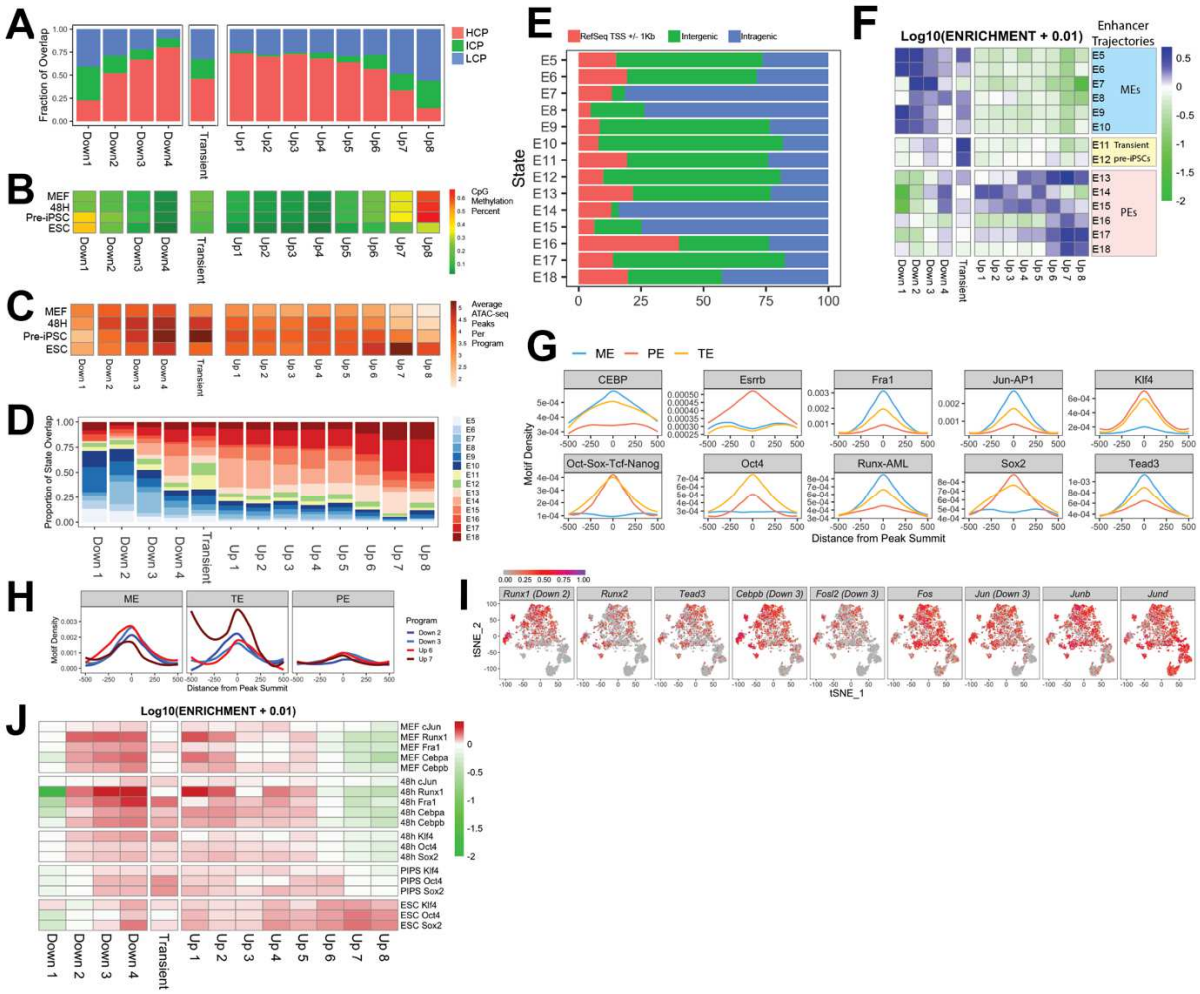


Figure 2-3 – Cell states and the stepwise order of events are conserved when the reprogramming process is modified through the overexpression of certain reprogramming factors

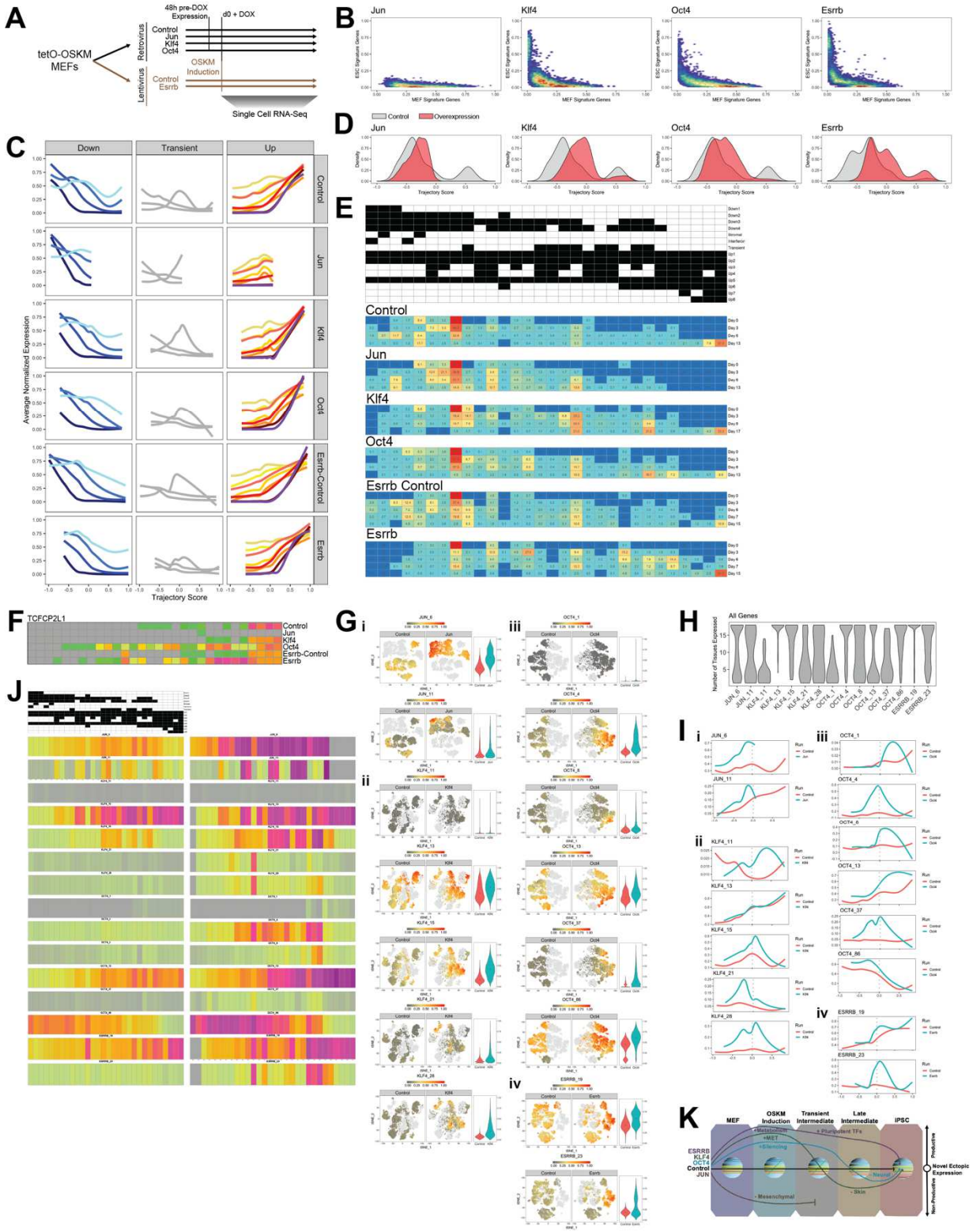


Figure 2-4 – The analysis of various reprogramming systems defines general principals of cell fate transitions

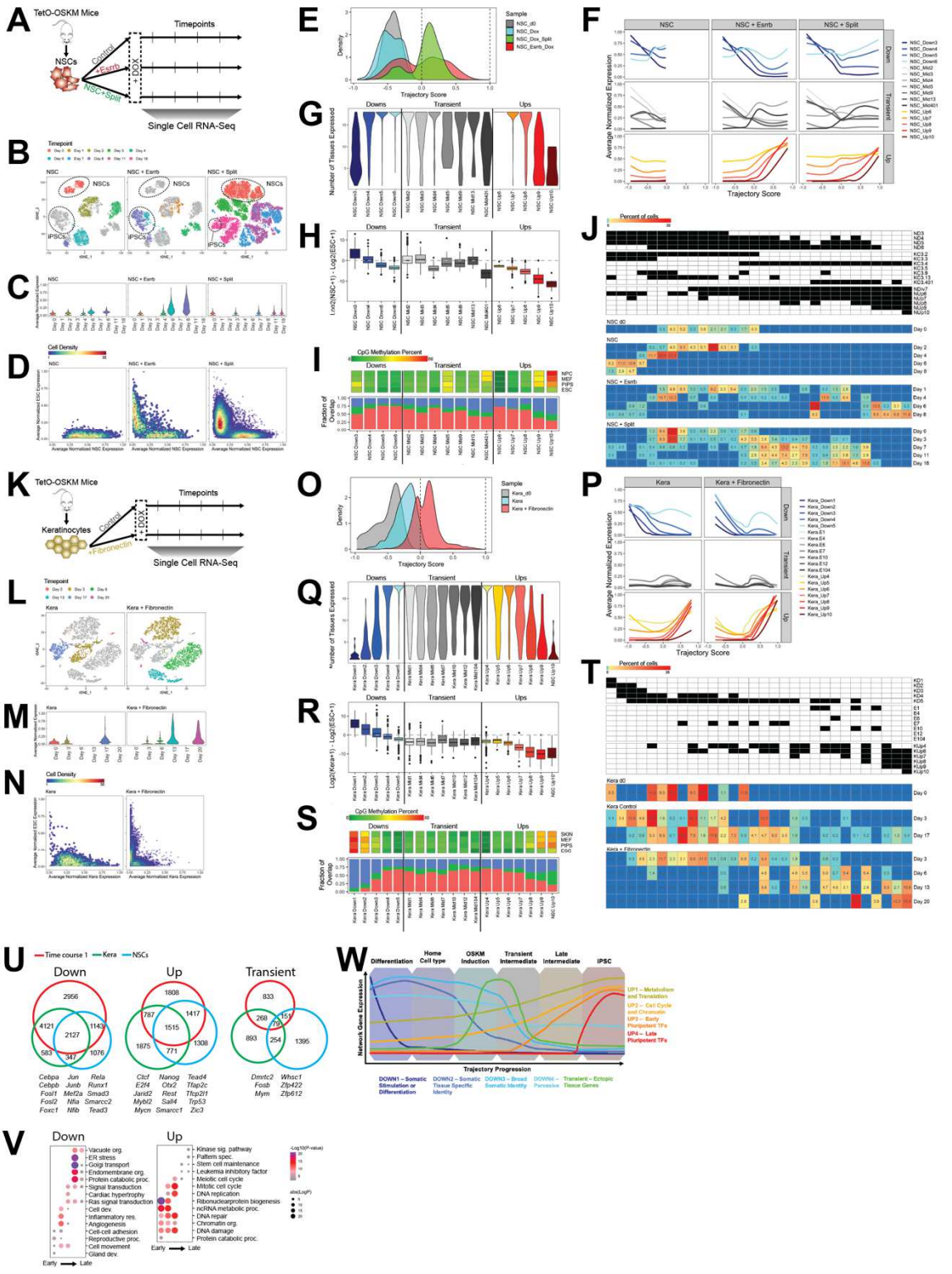


Figure S2-1 – Expression of canonical marker genes confirm latency of pluripotency

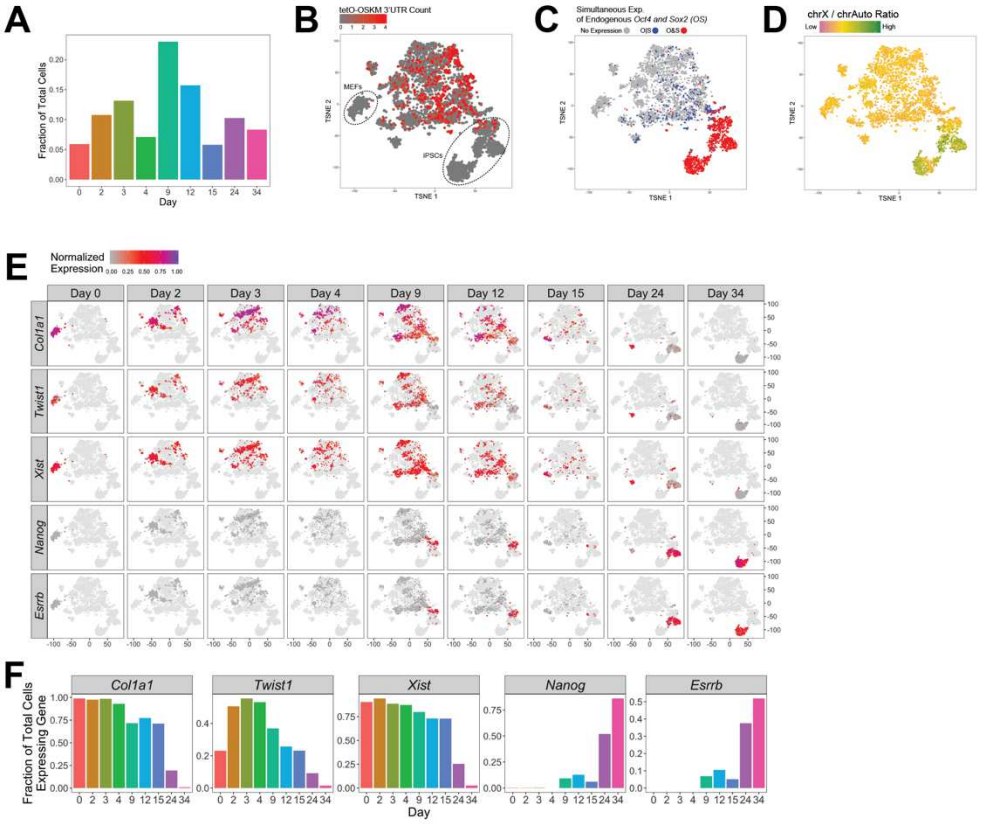


Figure S2-2 – Various dimensionality reduction techniques recapitulate transcriptional heterogeneity of reprogramming, latency of iPSC induction

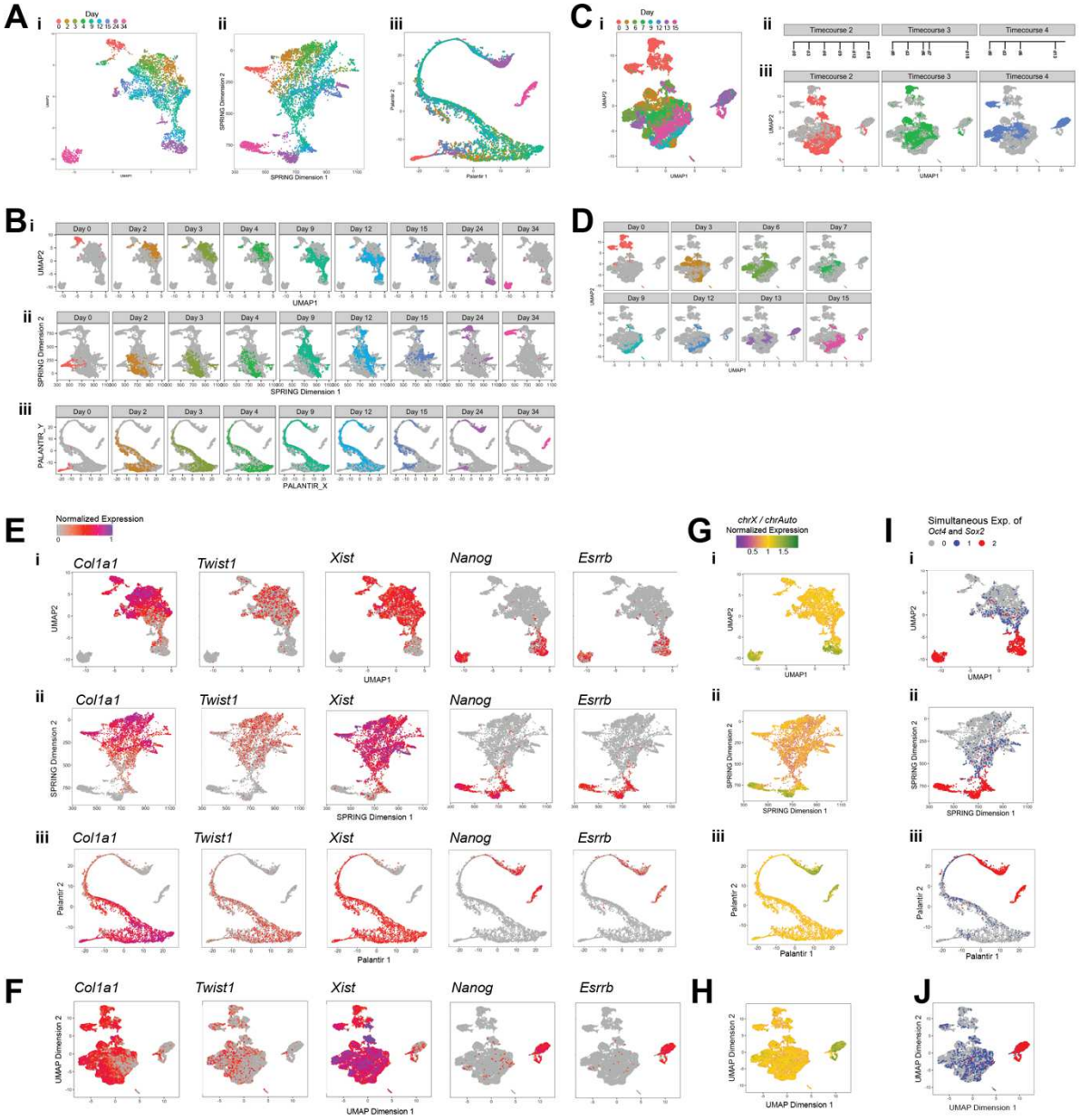


Figure S2-3 – Distribution of expression and Gene Ontology terms associated with Gene Expression Networks defined by GEND

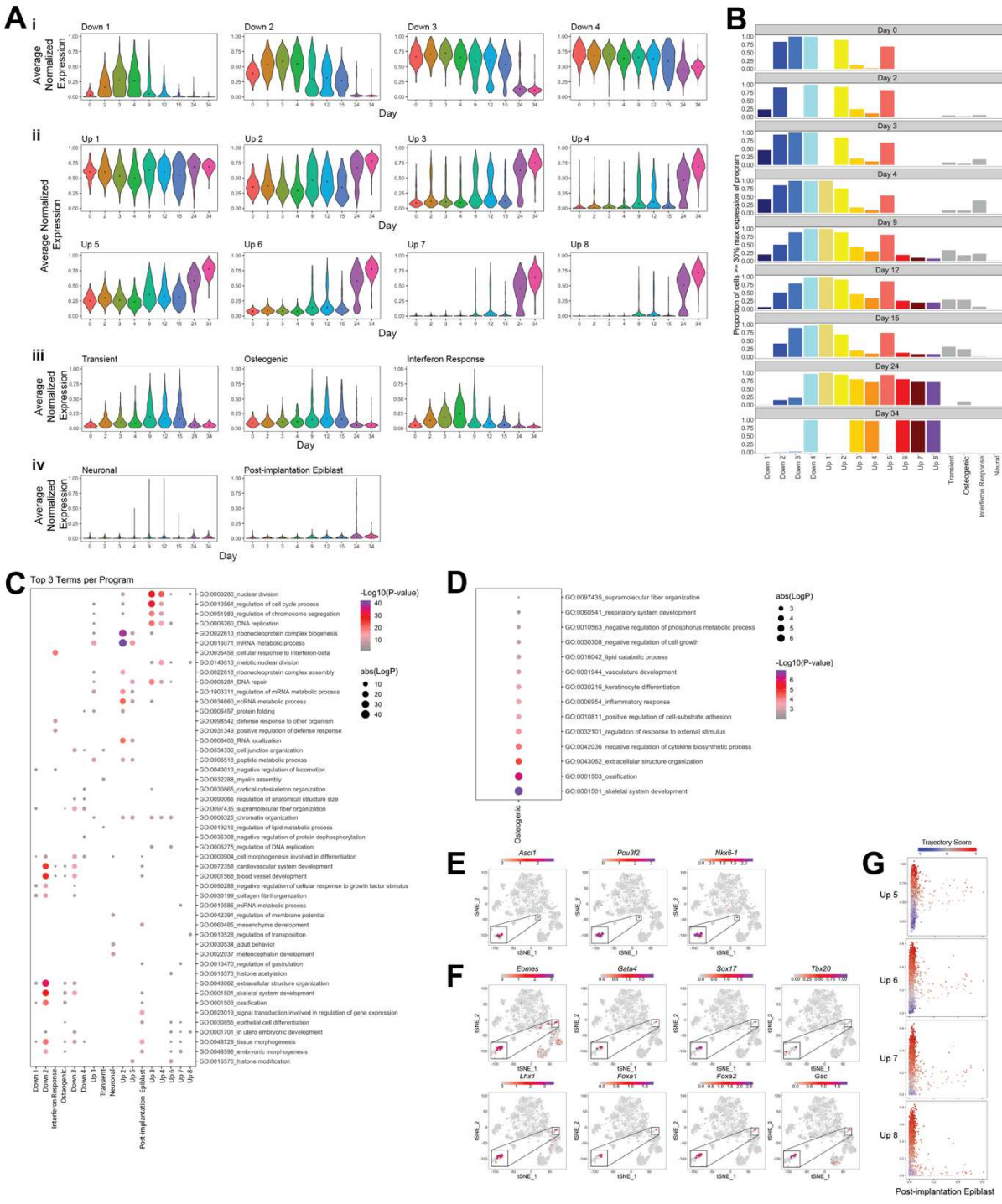


Figure S2-4 – Culture conditions influence the gene expression changes in the first 48hrs of reprogramming

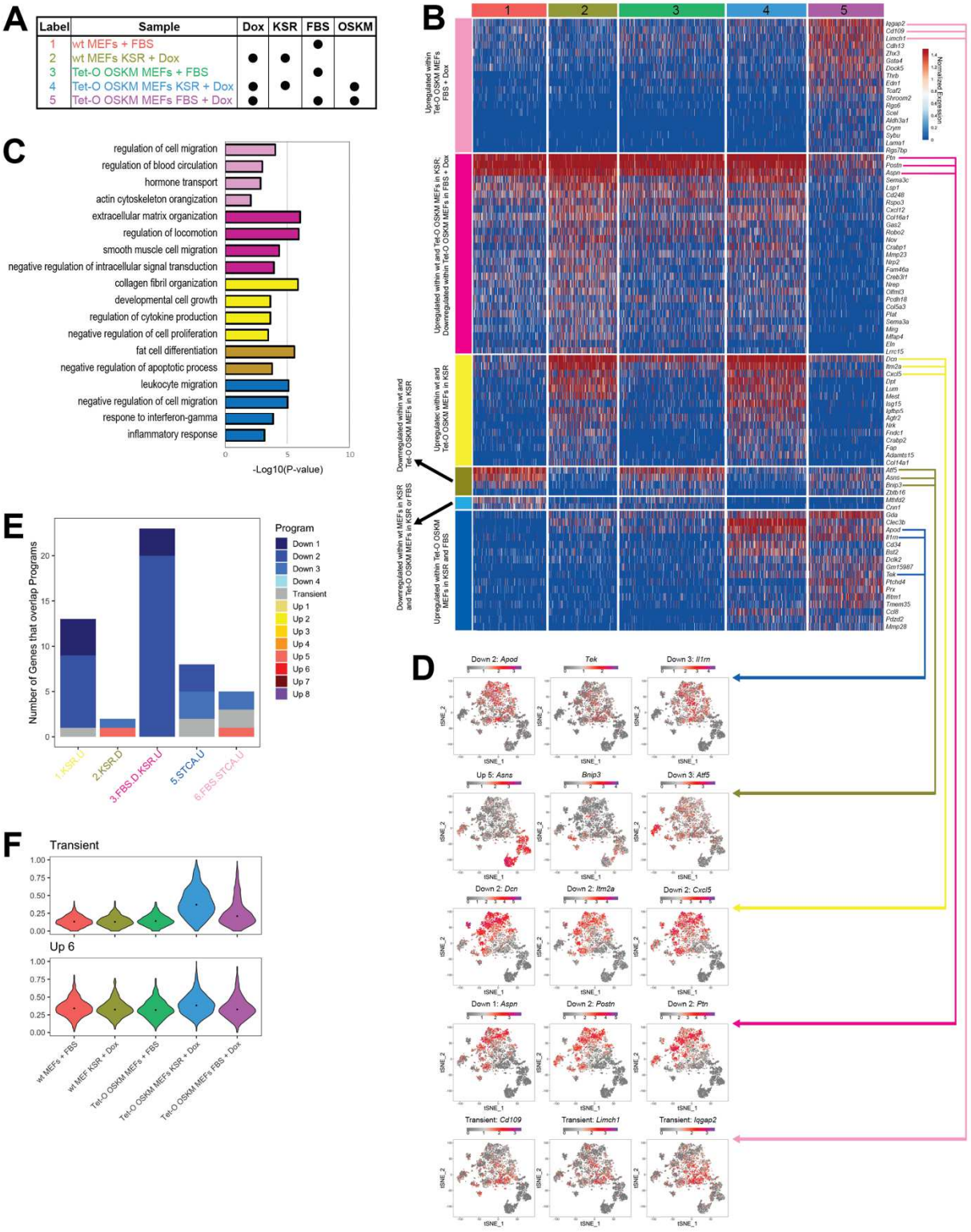


Figure S2-5 – Relative timing of pluripotency regulator expression

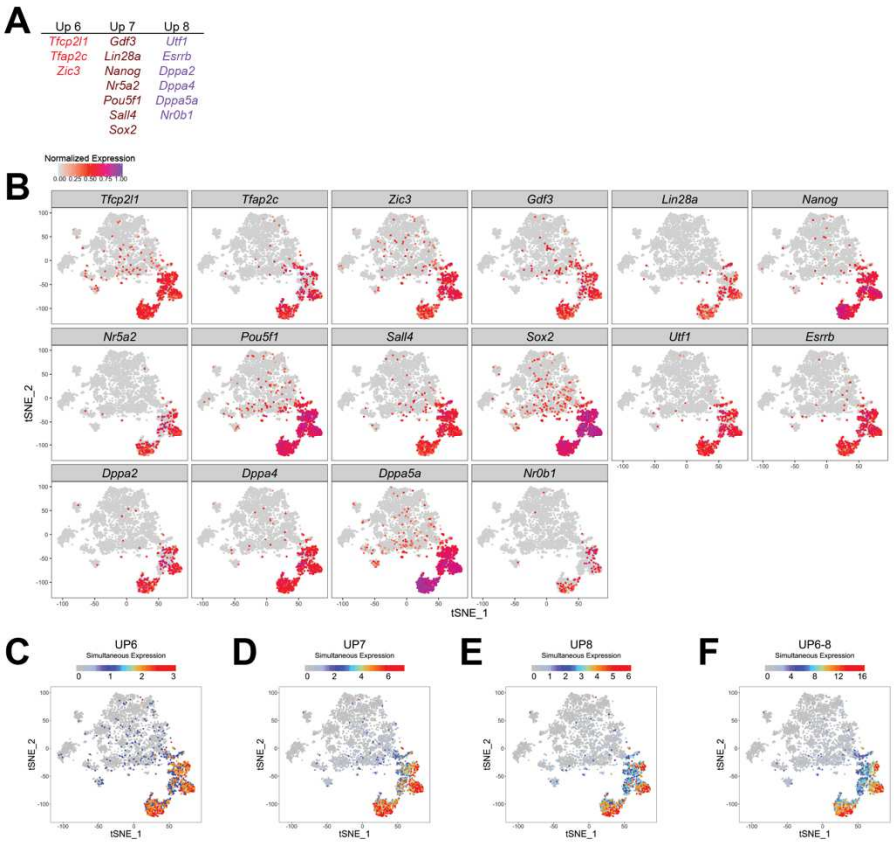


Figure S2-6 – Supplemental networks defined by GEND

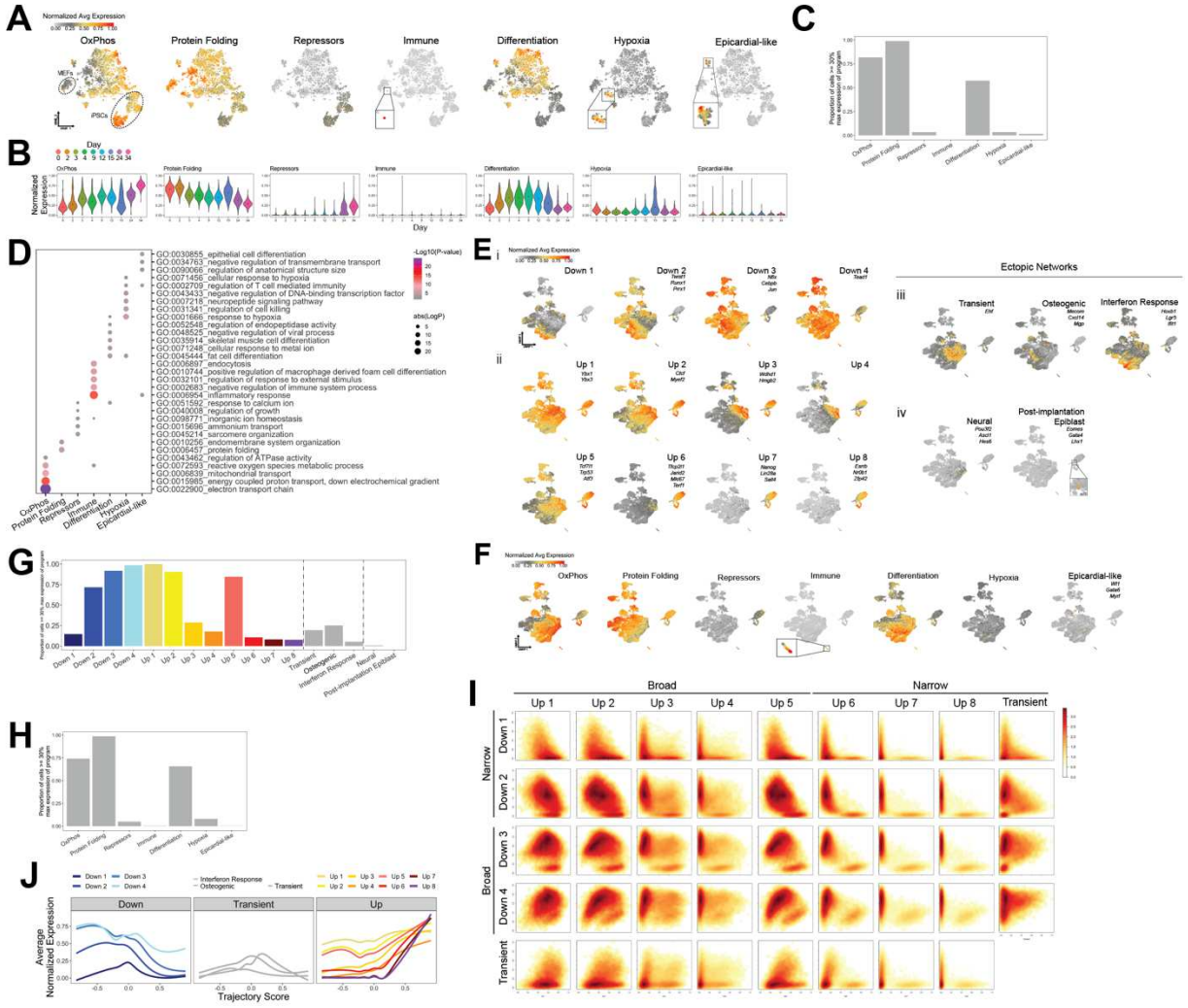


Figure S2-7 – Lineage tracing experiment reveal the post-implantation epiblast cells differentiated from pluripotent iPSCs

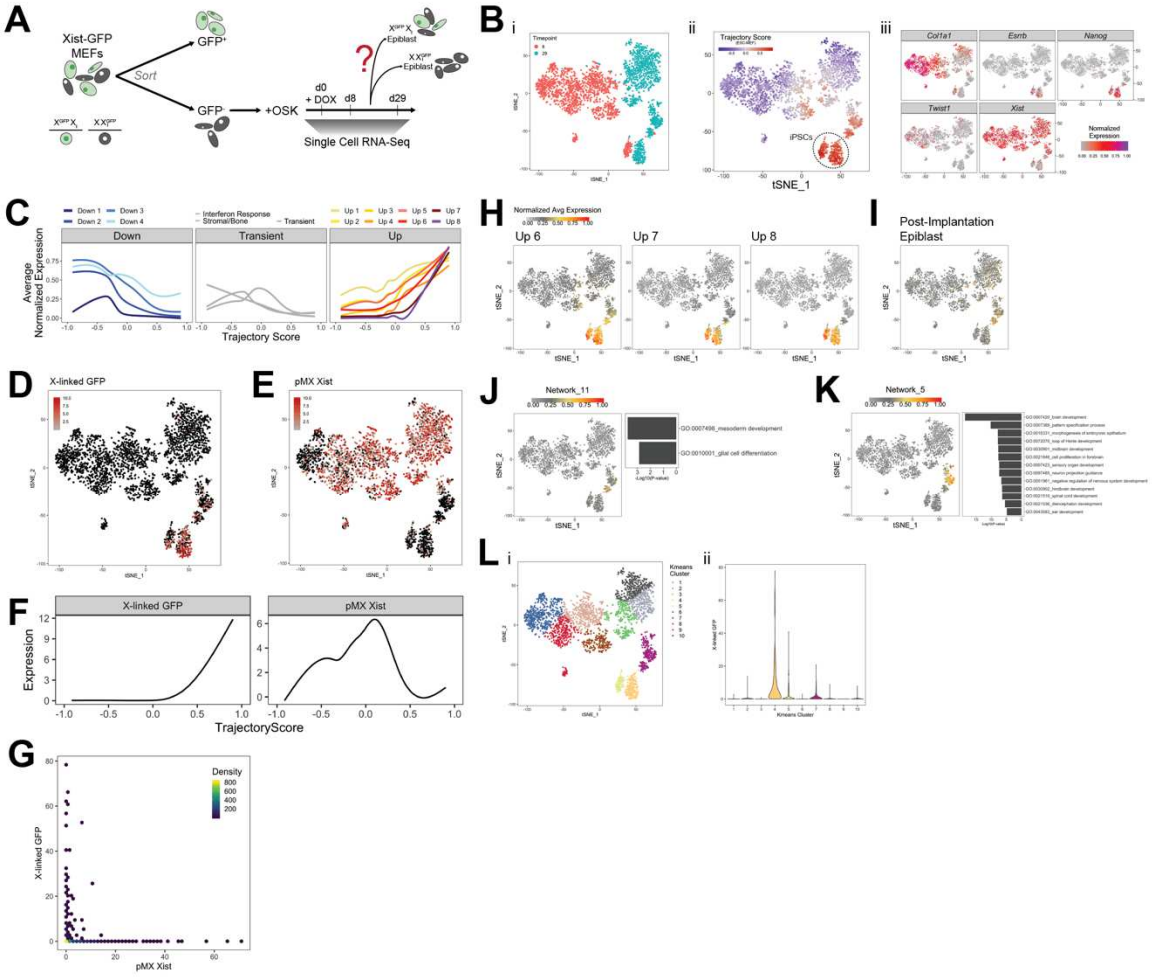


Figure S2-8 – Transient network captures the upregulation of lineage-specific markers from unrelated lineages during reprogramming

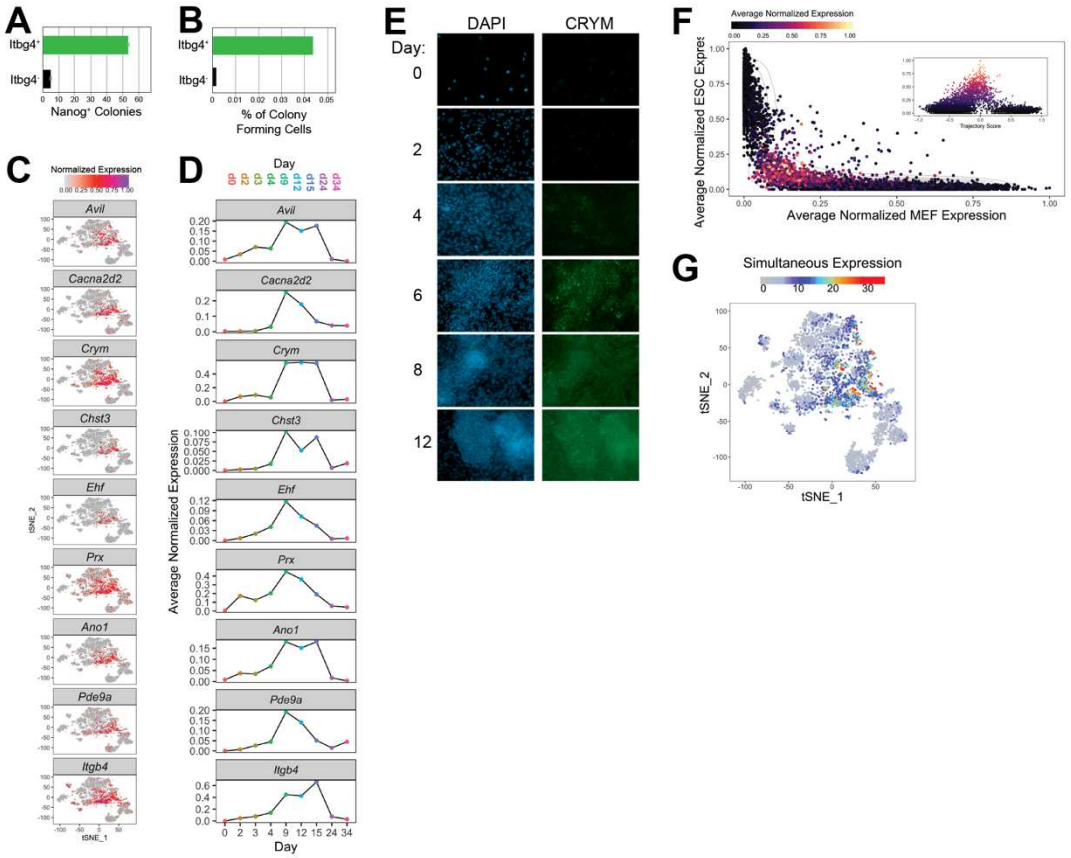


Figure S2-9 – MEF-and ESC-specific gene antagonism is used to define an axis of reprogramming progression

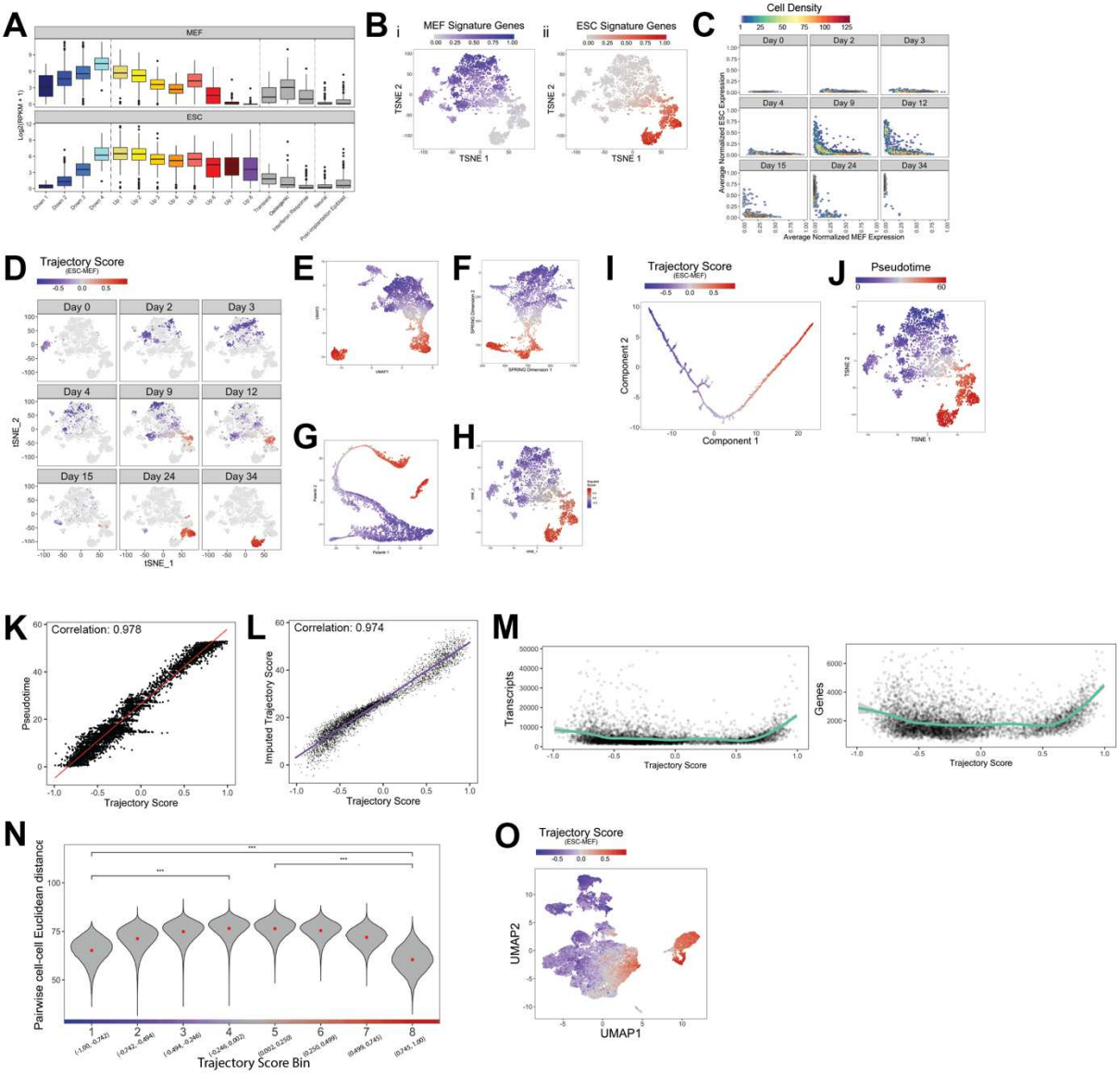


Figure S2-10 – Utilization of CENICS to pinpoint bottle necks along iPSC reprogramming

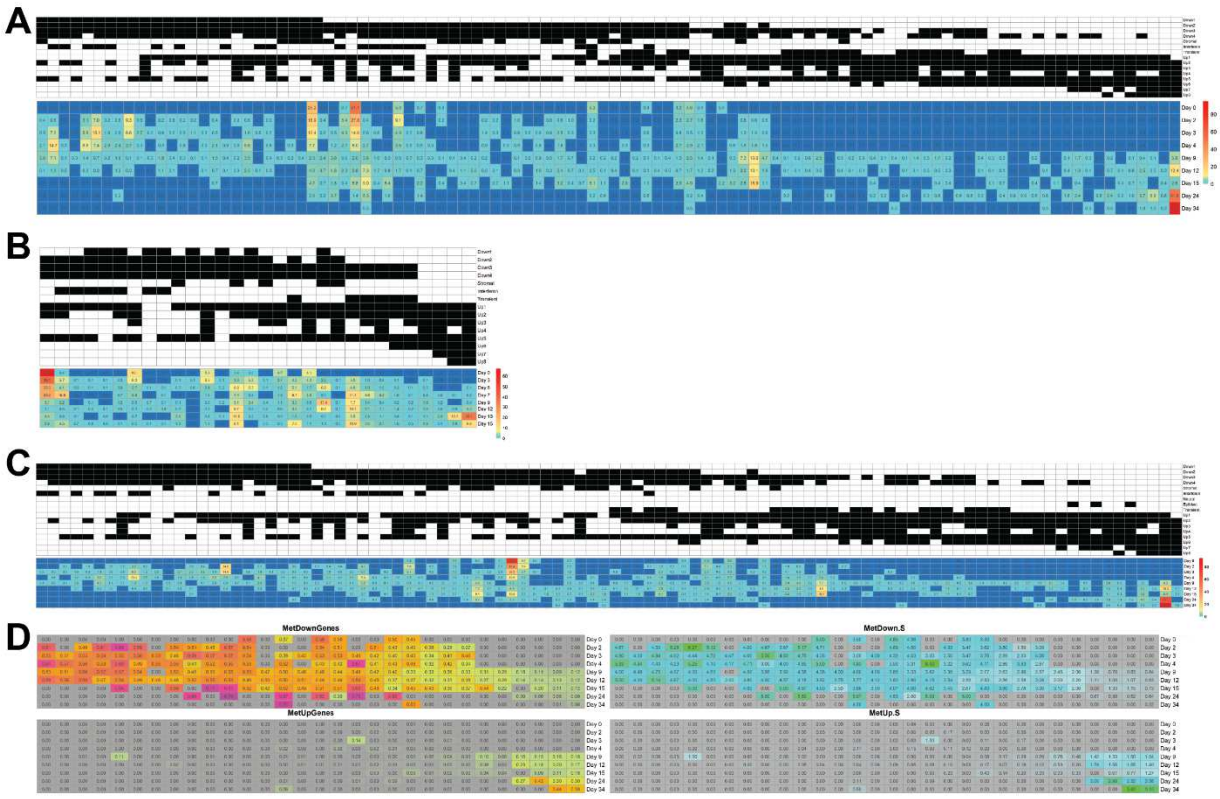


Figure S2-11 – Sorting for E-Cadherin expressing cells, at day 6 post OSKM induction

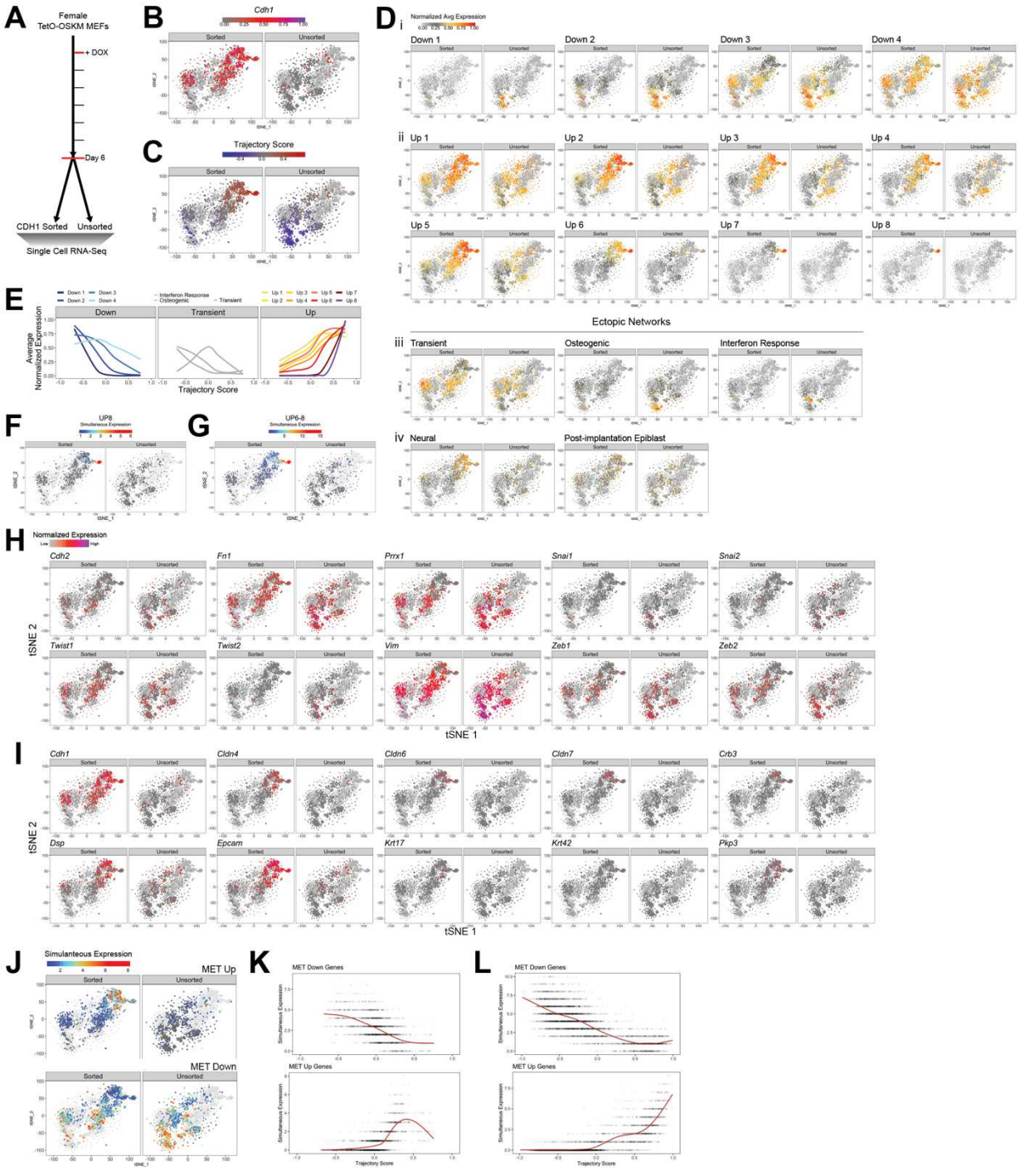


Figure S2-12 – Transcription factor control over enhancer elements

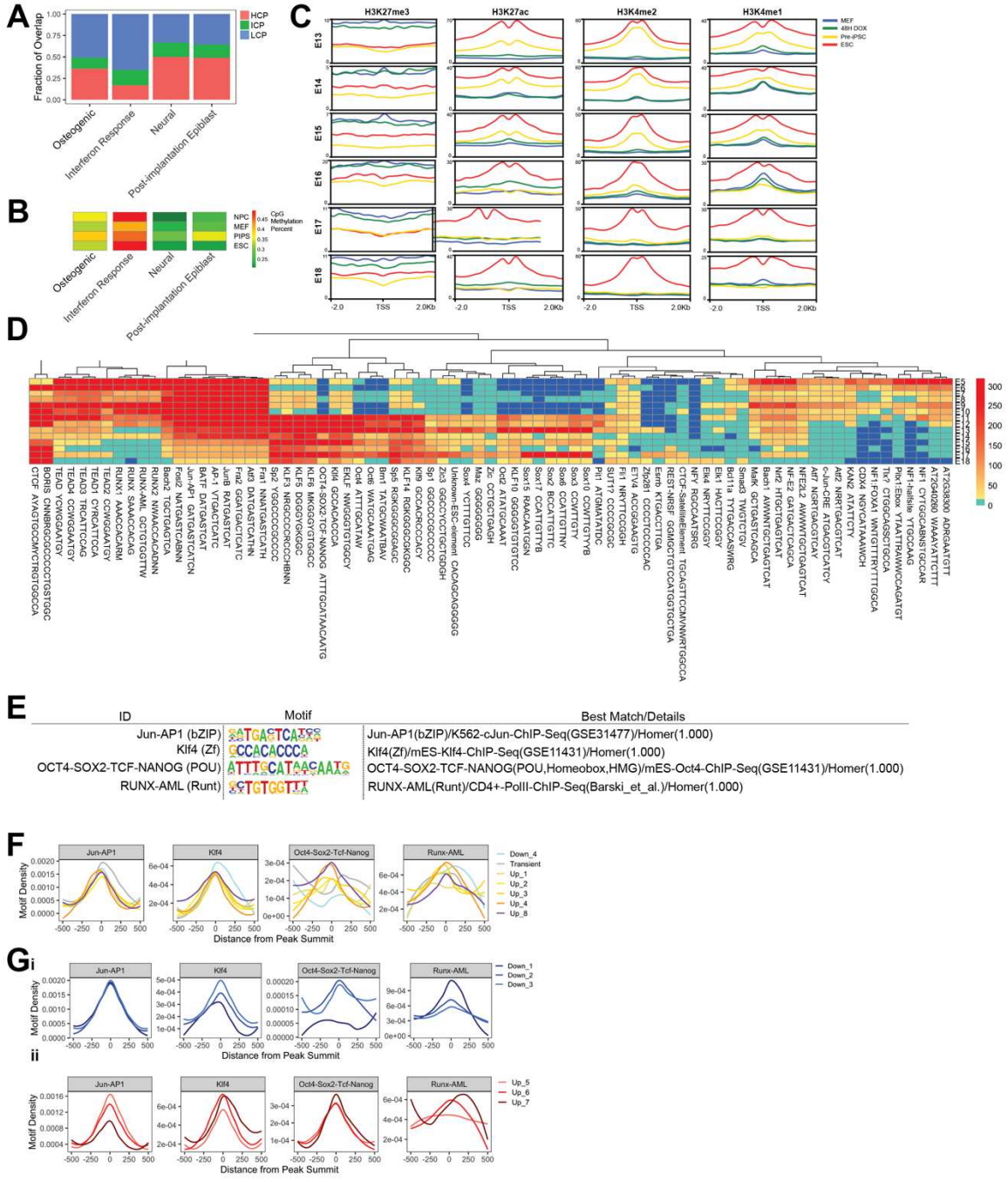


Figure S2-13 – Overexpression of ectopic TFs to initiate iPSC reprogramming

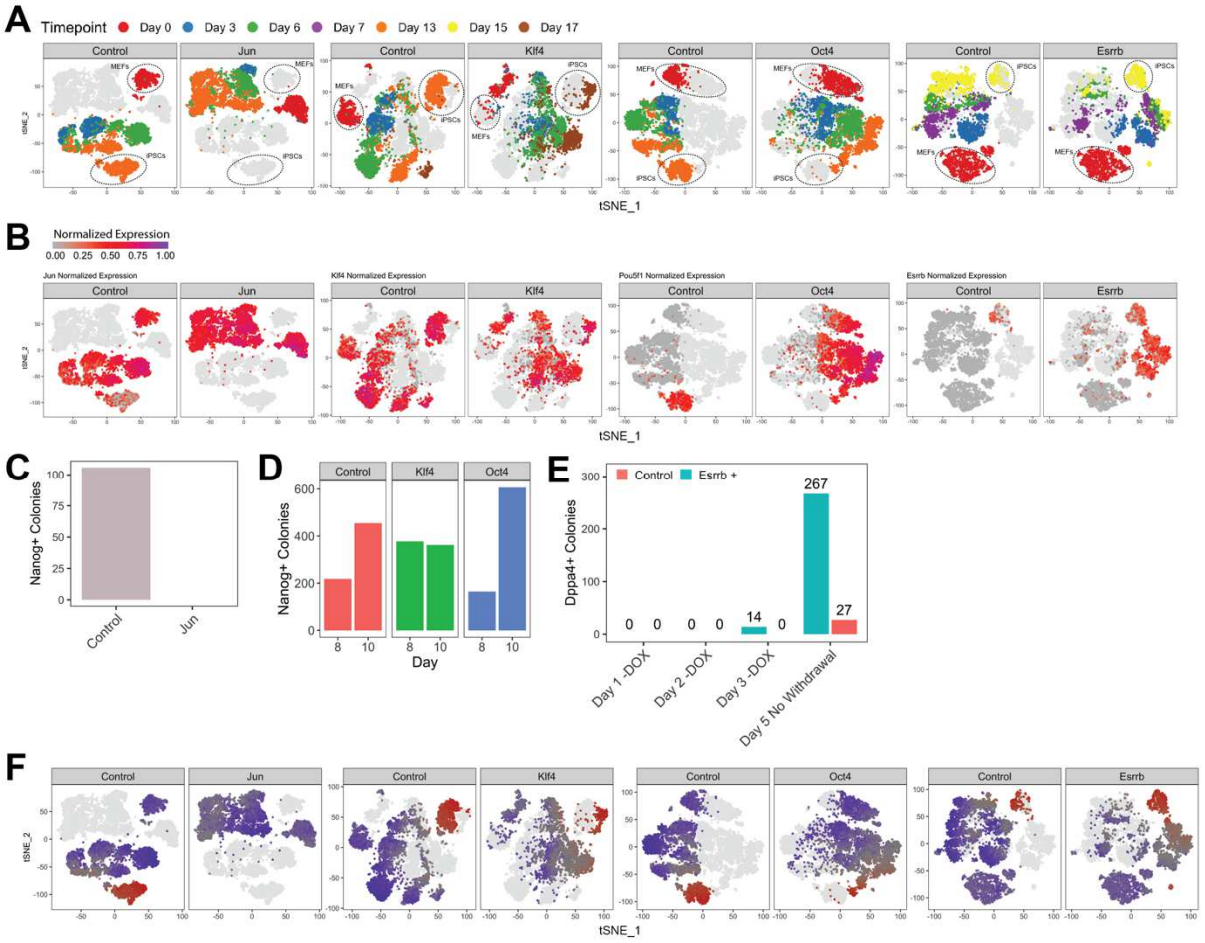


Figure S2-14 – Expression changes linked to ectopic TF overexpression among main gene networks

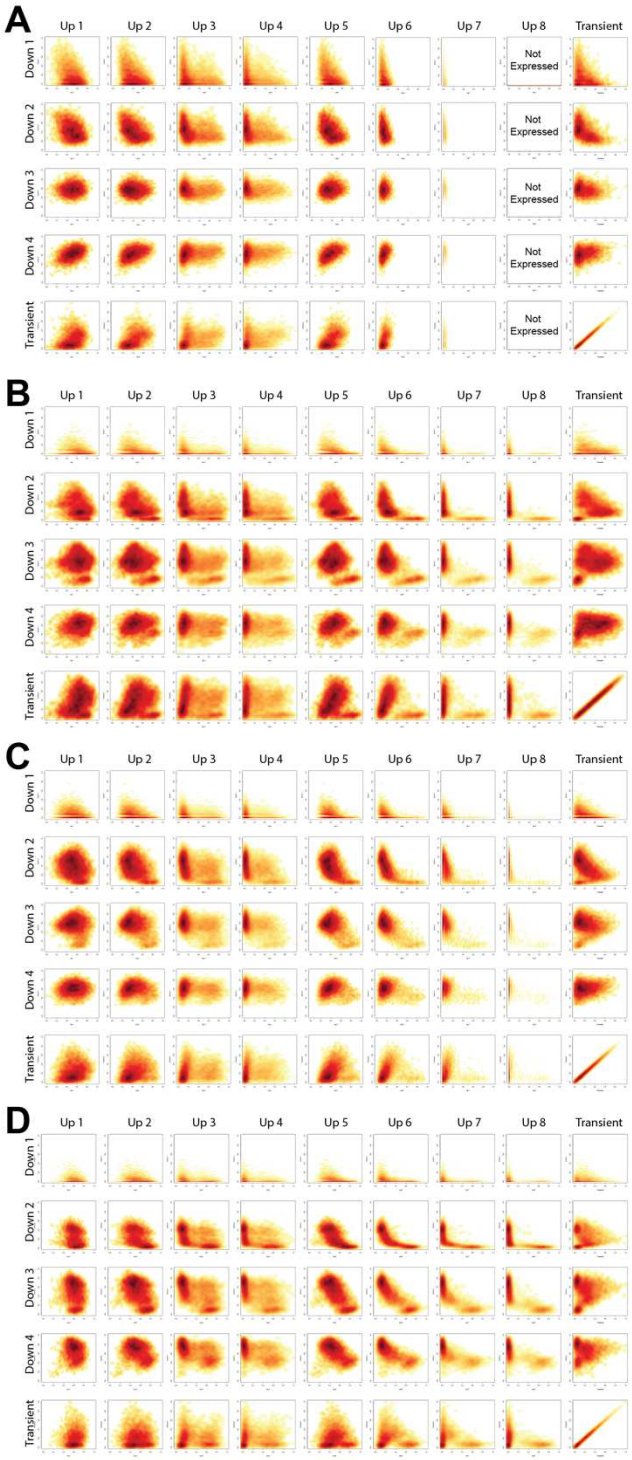


Figure S2-15 – Relative timing of pluripotency regulator expression upon ectopic TF overexpression

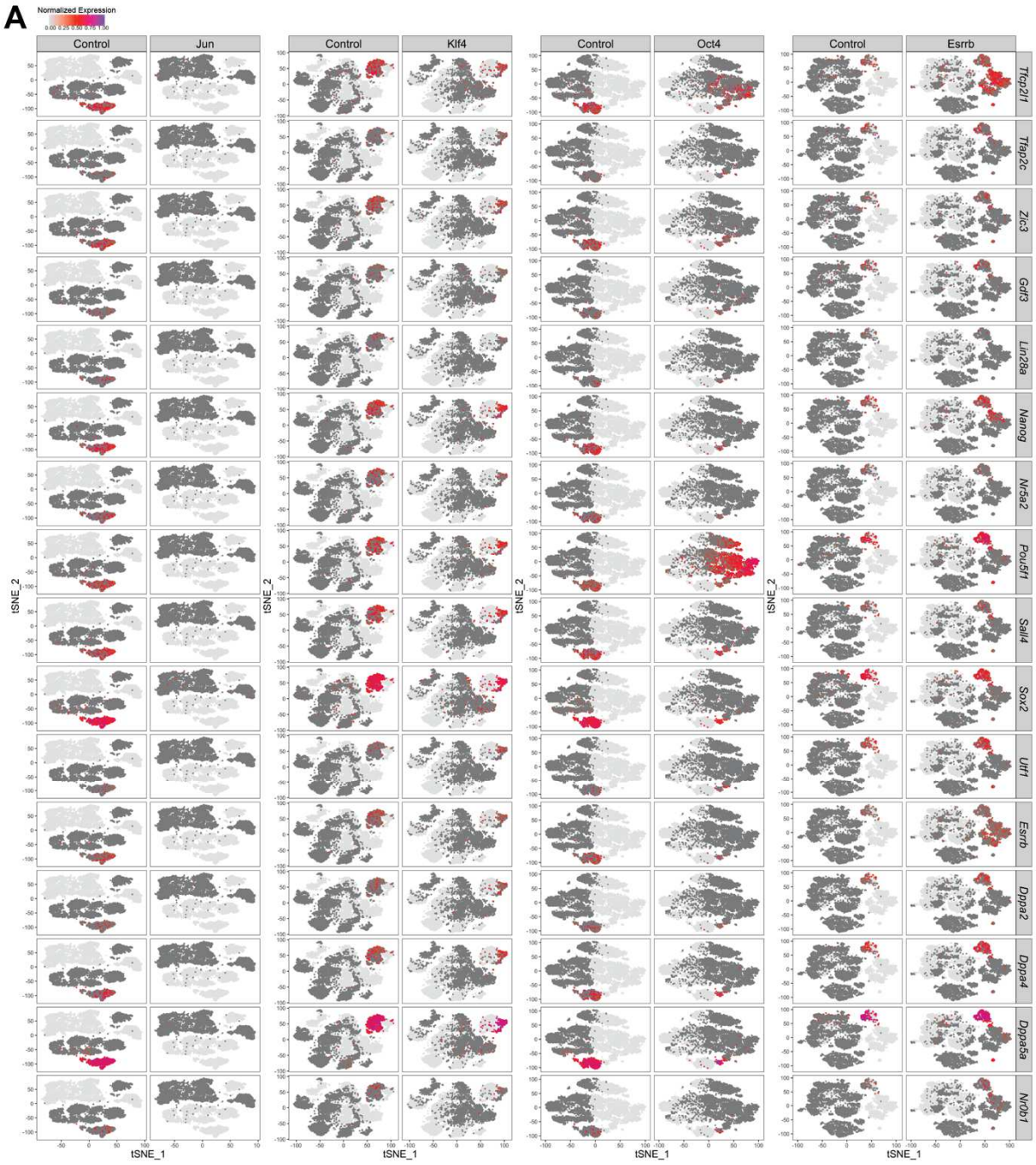


Figure S2-16 – The effect of ectopic TF overexpression on MET signature genes

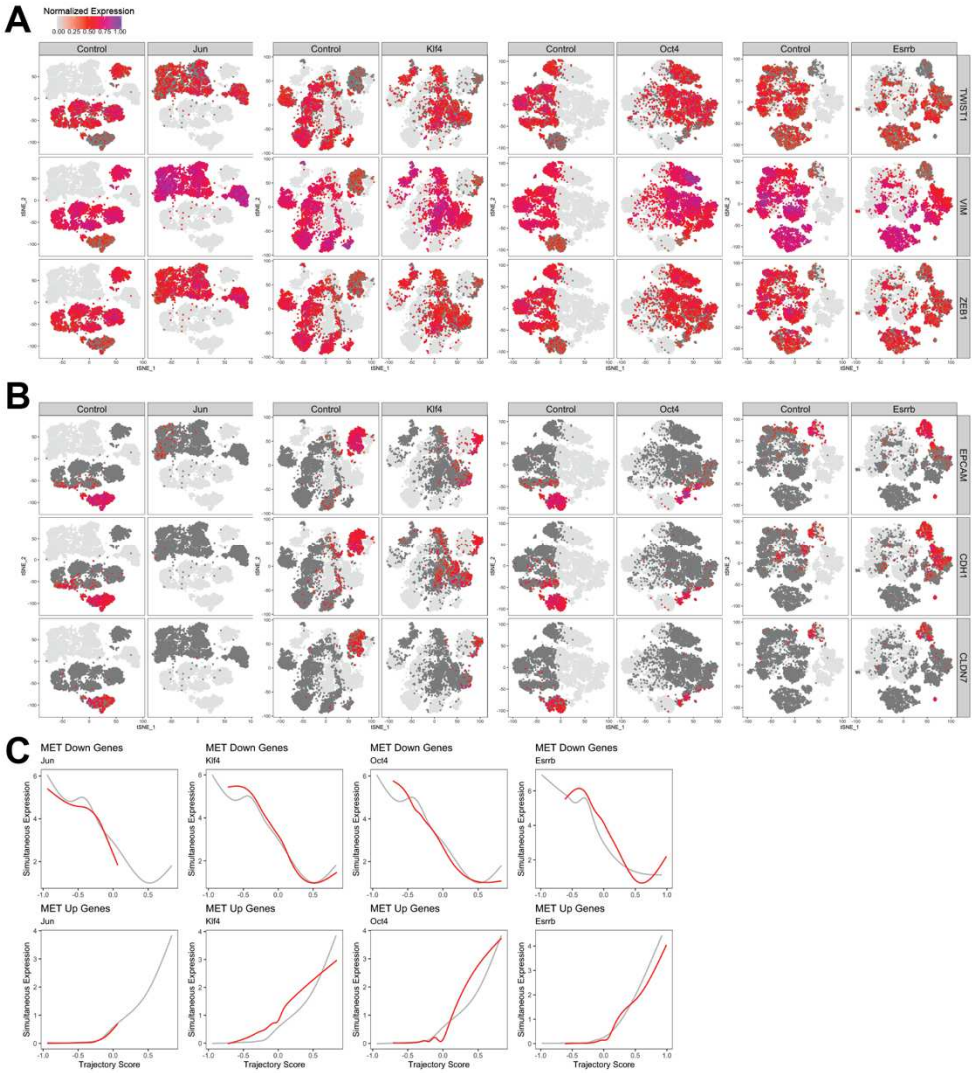


Figure S2-17 – Ectopic programs are enriched for metabolism and various differentiation processes

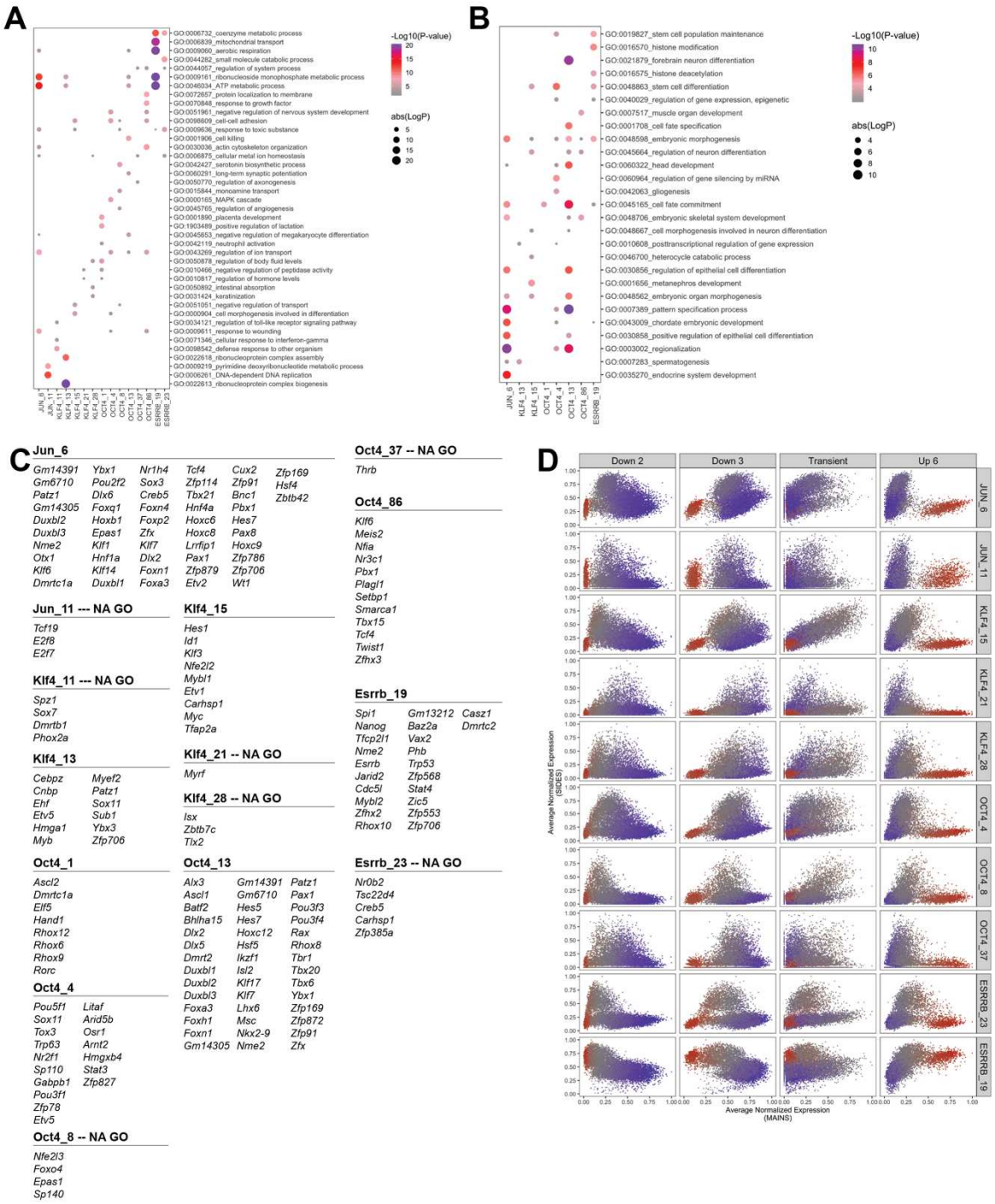


Figure S2-18 – Overexpression of MyoD in addition to OSKM induces skeletal muscle genes and completely blocks iPSC reprogramming

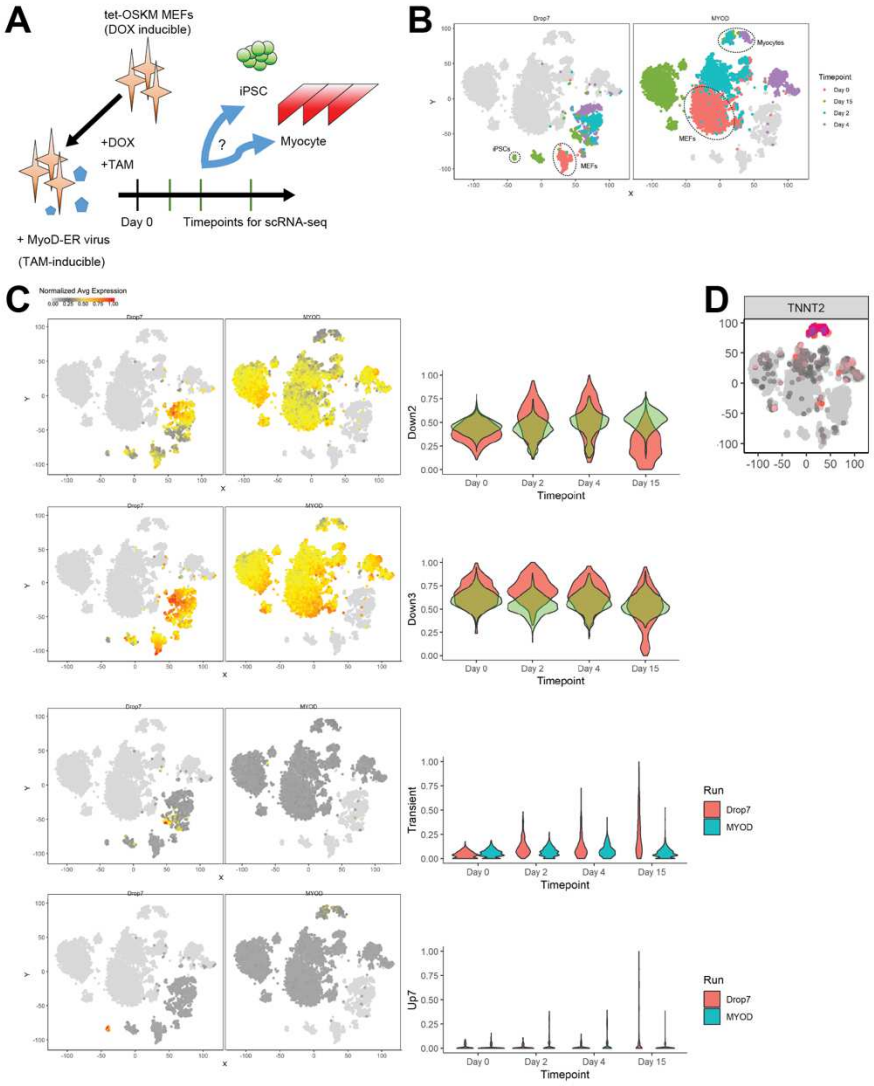


Figure S2-19 – Enabling efficient NSC-to-iPSC reprogramming through the overexpression of Esrrb

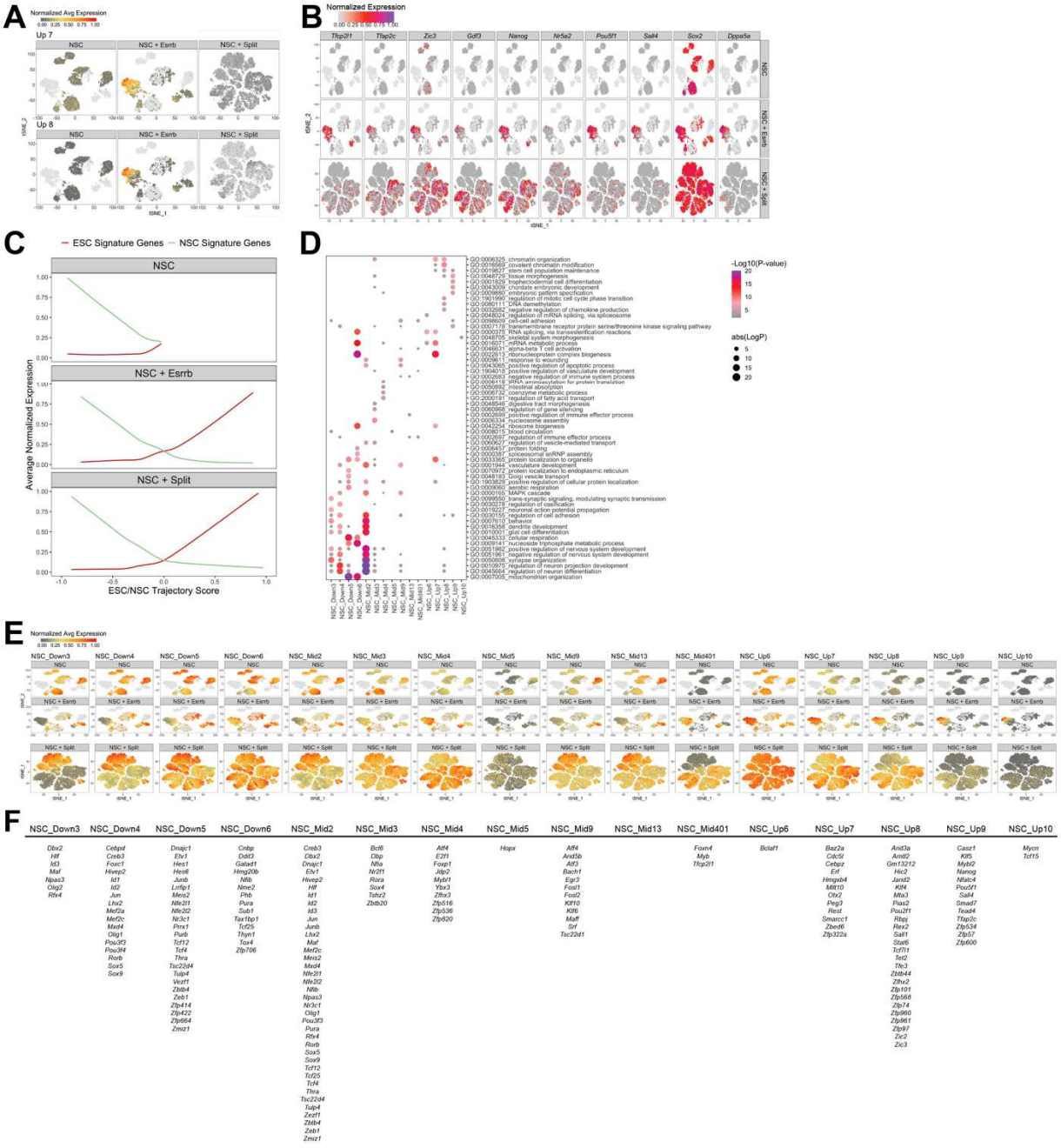


Figure S2-20 – Enabling efficient Keratinocyte-to-iPSC reprogramming though Keratinocyte stimulating substrate fibronectin

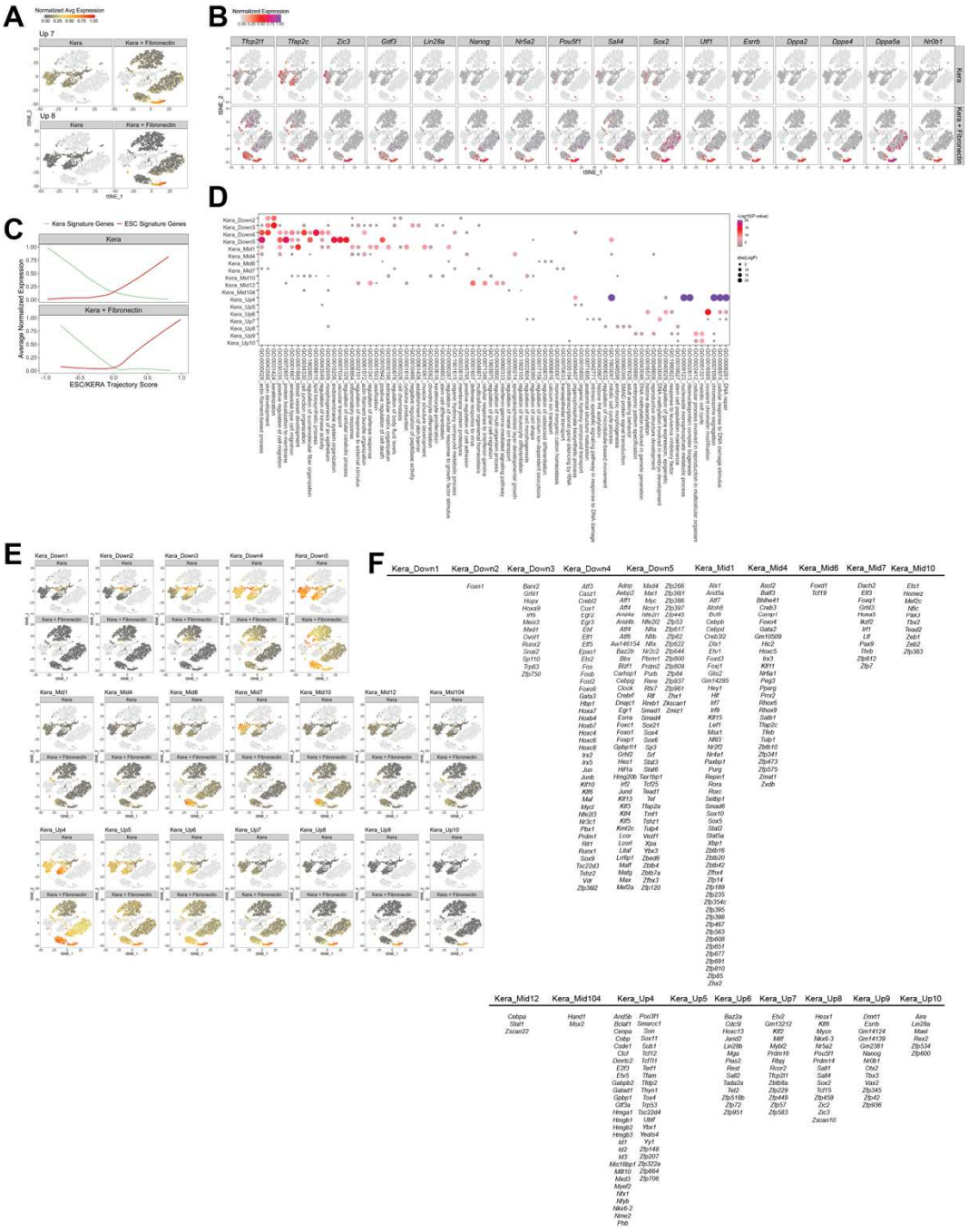


Figure S2-21 – Gene ontology analysis of shared sets of genes among 3 reprogramming systems

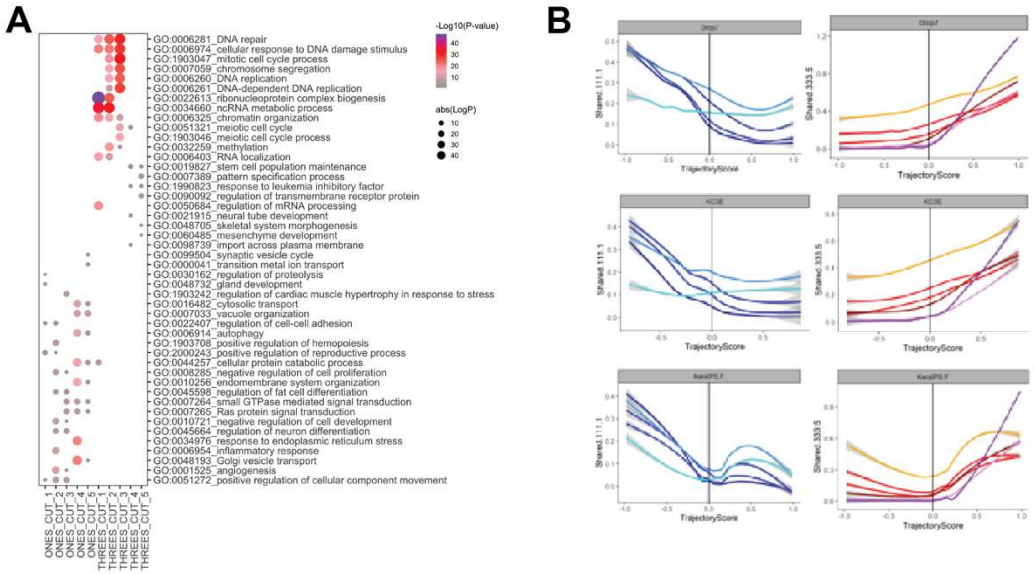


Figure S2-22 – Cell fate transition paradigm is conserved in other TF-induced direct reprogramming systems

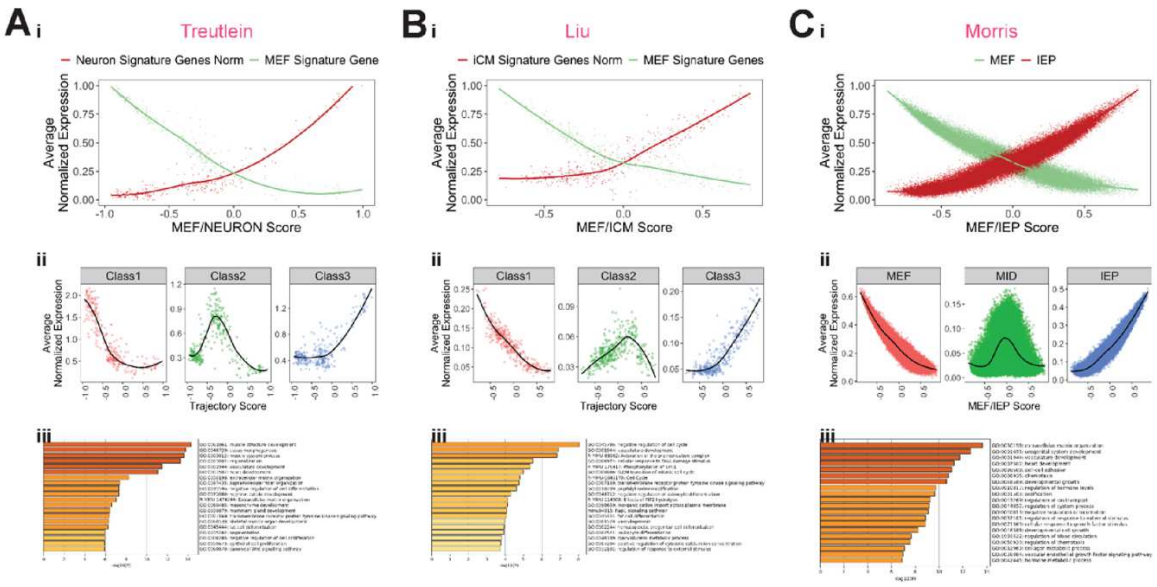


Table 2-1 – Single cell sequencing statistics

Experiment	Time point	Cell #	Average Gene #	Average UMIs	Median Gene #	Median UMIs
Time Course 1 (OSKM)	Day 0	282	1533	3903	1280	2758
	Day 2	514	2263	6718	1786	4125
	Day 3	628	2054	5149	1779	3675
	Day 4	339	1701	4029	1441	2752
	Day 9	1098	1880	4080	1720	3288
	Day 12	750	1737	3547	1582	2918
	Day 15	277	1717	4421	1383	2948
	Day 24	490	1890	4225	1706	3187
	Day 34	397	2622	6982	2409	5322
Time Course 2 (OSKM)	Day 0	1297	2401	7266	2215	5148
	Day 3	996	2139	4964	1956	3850
	Day 6	1379	2257	6024	2058	4464
	Day 9	980	2010	5015	1898	4080
	Day 12	918	2266	5914	2142	4692
	Day 15	692	1729	3970	1550	2902
Time Course 3 (OSKM)	Day 0	947	2990	10711	2764	7815
	Day 3	974	2829	8932	2558	6290
	Day 6	931	2644	7208	2294	4977
	Day 7	1038	3051	9113	2773	6738
	Day 15	1429	2498	6328	2248	4899
Time Course 3E (OSKM+E)	Day 0	1068	2782	9017	2559	6724
	Day 3	870	3064	9902	2806	7254
	Day 6	800	2809	7955	2506	5662
	Day 7	877	2776	7461	2560	5751
	Day 15	943	2589	6966	2370	5327
Culture Conditions (48hrs)	wt MEFs + FBS	535	3036	8546	2771	6307
	wt MEFs KSR + Dox	504	3149	9073	2832	6670
	Tet-O OSKM MEFs + FBS	684	2761	6978	2522	5452
	Tet-O OSKM MEFs KSR + Dox	670	2781	6647	2586	5496
	Tet-O OSKM MEFs FBS + Dox	564	2737	6497	2414	4926
Cdh1 Sort	Unsorted Day 6	1043	1983	4874	1835	3924
	Cdh1+ Day 6	1098	2464	6954	2252	5291
Drop 6, 8, 10 Merge	Day 0	2943	2723	8859	2657	6982
	Day 3	3041	2074	5509	1779	3659
	Day 6	3867	1954	4942	1665	3427
	Day 7	1038	3051	9113	2773	6738
	Day 9	914	2060	5171	1938	4223
	Day 12	854	2330	6116	2201	4980
	Day 13	2057	1642	3585	1508	3018
	Day 15	2062	2282	5664	2076	4367
Drop 10 Ctrl	Day 0	808	2737	8459	2730	7564
	Day 3	1135	1370	3008	1244	2558
	Day 6	1650	1342	2883	1214	2452
	Day 7	1038	3051	9113	2773	6738
	Day 13	2057	1642	3585	1508	3018
	Day 15	1429	2498	6328	2248	4899
	Day 17	262	2556	8936	2468	6881
Drop 10 Jun	Day 0	1055	1539	3022	1413	2521
	Day 3	299	1113	1829	1061	1711
	Day 6	1366	1381	2665	1568	2075
	Day 13	2655	1727	3365	1201	2826
Drop 10 Klf4	Day 0	638	1979	5848	1558	3468
	Day 3	956	2448	6669	2318	5409
	Day 6	1429	1638	3629	1473	2797
	Day 17	1229	1829	4267	1656	3279
Drop 10 Oct4	Day 0	1483	1729	3781	1610	3223
	Day 3	1487	`813	4122	1677	3422
	Day 6	2072	1599	3182	1442	2620
	Day 13	1746	1768	3630	1570	2864
Kera Ctrl	Day 0	93	1405	3678	1054	2303
	Day 3	1070	1080	2743	900	1736
	Day 10	9	997	2151	979	2102
	Day 17	833	1298	2709	1137	2104
Kera + Fib	Day 0	93	1405	3678	1054	2303
	Day 3	1957	412	914	286	527
	Day 13	875	1682	3657	1442	2597
	Day 20	152	1483	3779	1170	2364
NSC Ctrl	Day 0	629	1694	3163	1569	2751
	Day 2	869	1253	2382	1151	2041
	Day 4	639	1470	2720	1344	2379
	Day 6	840	1227	2399	1118	2055
	Day 8	170	861	1615	808	1456
NSC + ESRRB	Day 1	391	1296	2426	1214	2156
	Day 4	605	1559	2878	1445	2509
	Day 6	323	1556	2879	1422	2458
	Day 8	939	1209	2480	1063	1944
Xist	Day 8	1453	1276	2245	1155	1893
	Day 29	1168	1292	2415	1155	1964

Table 2-2 – Differentially expressed genes that define MEF- and ESC-signature genes based on population RNA-seq

GeneID	Mean Expression	Log2FC	SE	Wald Statistic	Wald Test P-Value	BH Adjusted P-Value
Lin28a	493.72	-16.24	0.87	-18.66	1.00E-77	4.01E-75
Zfp42	438.16	-15.94	0.84	-19.03	1.06E-80	4.94E-78
Tdgf1	1625.07	-14.65	0.99	-14.83	1.01E-49	9.36E-48
Sall4	342.08	-14.56	0.84	-17.31	4.07E-67	9.89E-65
Tdh	611.70	-14.54	0.85	-17.17	4.80E-66	1.11E-63
Dppa4	208.03	-13.91	0.87	-15.98	1.64E-57	2.34E-55
Fgf4	520.41	-13.88	0.82	-16.97	1.39E-64	2.99E-62
L1td1	1911.79	-13.77	0.79	-17.37	1.37E-67	3.45E-65
Triml2	100.57	-13.61	0.91	-15.00	7.63E-51	7.51E-49
Esrrb	10228.58	-13.58	0.42	-32.03	3.80E-225	3.20E-221
Triml1	64.54	-13.36	0.97	-13.82	1.98E-43	1.30E-41
Dppa5a	1247.25	-13.25	1.04	-12.78	2.04E-37	9.37E-36
Nr0b1	251.38	-13.19	0.97	-13.64	2.42E-42	1.46E-40
Mageb16	48.31	-12.70	1.02	-12.44	1.68E-35	6.94E-34
Zscan10	209.17	-12.62	0.67	-18.84	3.80E-79	1.64E-76
Fam169a	320.20	-12.51	0.61	-20.61	2.21E-94	1.86E-91
Cecr2	177.87	-12.45	0.85	-14.66	1.11E-48	9.66E-47
Tcl1	51.18	-12.38	1.01	-12.21	2.72E-34	1.05E-32
Gpa33	122.38	-12.31	0.72	-17.01	6.38E-65	1.39E-62
Ano9	53.66	-11.97	0.82	-14.58	3.92E-48	3.26E-46
Grb7	68.62	-11.88	0.94	-12.63	1.50E-36	6.48E-35
Grhl2	41.70	-11.77	1.01	-11.70	1.33E-31	4.07E-30
Pou5f1	22813.48	-11.75	0.34	-34.61	1.84E-262	3.10E-258
Esrp1	93.35	-11.73	0.97	-12.08	1.40E-33	5.09E-32
Tcf15	38.43	-11.67	0.95	-12.33	6.63E-35	2.67E-33
Trh	235.86	-11.67	0.89	-13.04	6.98E-39	3.49E-37
Trap1a	118.02	-11.66	0.91	-12.78	2.12E-37	9.67E-36
Tex19.1	91.01	-11.54	0.98	-11.80	4.06E-32	1.32E-30
Dnmt3l	365.04	-11.32	0.50	-22.84	2.03E-115	3.11E-112
Cldn6	61.10	-11.23	0.80	-13.97	2.45E-44	1.69E-42
Gm13242	117.25	-11.19	0.86	-12.97	1.81E-38	8.87E-37
AU018091	355.23	-11.11	0.64	-17.33	2.65E-67	6.57E-65
Rbmxl2	67.07	-10.94	0.95	-11.52	1.09E-30	3.14E-29
Tcfap2c	172.78	-10.92	0.77	-14.11	3.32E-45	2.36E-43
LOC100303645	45.82	-10.86	0.98	-11.03	2.77E-28	6.75E-27
Nefh	53.21	-10.81	0.74	-14.61	2.46E-48	2.08E-46
Ap1m2	48.39	-10.69	0.88	-12.13	7.29E-34	2.71E-32
Epcam	282.34	-10.66	0.63	-16.82	1.66E-63	3.41E-61
Nodal	60.57	-10.62	1.05	-10.16	3.02E-24	5.39E-23
Tdrd12	74.06	-10.57	0.72	-14.73	4.45E-49	3.98E-47
Gm7325	89.17	-10.55	0.76	-13.81	2.09E-43	1.36E-41
Lefty2	90.63	-10.52	0.85	-12.43	1.90E-35	7.76E-34
2410007B07Rik	66.83	-10.44	1.00	-10.45	1.46E-25	2.91E-24
Dppa2	151.83	-10.39	0.69	-15.14	9.13E-52	9.60E-50
Foxh1	44.86	-10.04	0.80	-12.63	1.48E-36	6.44E-35
Slc28a1	114.48	-9.89	0.85	-11.67	1.82E-31	5.51E-30
Zfp819	38.35	-9.86	0.83	-11.95	6.70E-33	2.32E-31
Mcf2	41.10	-9.78	0.83	-11.72	1.05E-31	3.27E-30
Cdh1	961.04	-9.74	0.60	-16.16	1.03E-58	1.59E-56
Dsg2	482.99	-9.71	0.48	-20.08	1.03E-89	6.90E-87
Camkv	43.98	-9.64	0.81	-11.95	6.53E-33	2.27E-31
Gldc	327.40	-9.63	0.50	-19.31	4.60E-83	2.50E-80
Elf3	96.66	-9.57	0.93	-10.34	4.82E-25	9.19E-24
Tex14	76.71	-9.42	0.56	-16.75	5.30E-63	1.04E-60
Morc1	107.22	-9.33	0.61	-15.39	1.90E-53	2.19E-51
Lrrc34	63.53	-9.27	0.73	-12.78	2.24E-37	1.02E-35
Nphs1	131.56	-9.17	0.88	-10.43	1.75E-25	3.43E-24
Gm949	37.81	-9.12	0.76	-12.05	1.88E-33	6.74E-32
Gm13247	70.40	-9.11	0.63	-14.50	1.13E-47	9.22E-46
Bex1	322.81	-9.09	0.63	-14.43	3.56E-47	2.86E-45
Sox2	2181.29	-9.05	0.36	-25.15	1.51E-139	5.08E-136
Olig2	37.66	-9.05	1.00	-9.07	1.14E-19	1.39E-18
Hsd17b14	95.62	-9.05	0.55	-16.37	3.38E-60	5.68E-58
Slc39a4	70.74	-8.99	0.60	-14.93	1.97E-50	1.89E-48
Glt1d1	77.72	-8.73	0.68	-12.90	4.37E-38	2.08E-36
Gm6792	57.43	-8.69	0.92	-9.45	3.40E-21	4.78E-20
Pecam1	140.19	-8.67	0.77	-11.24	2.52E-29	6.61E-28
Upp1	585.26	-8.66	0.76	-11.42	3.15E-30	8.84E-29
Dppa3	47.53	-8.66	1.12	-7.76	8.20E-15	6.43E-14
Nr5a2	78.70	-8.64	1.02	-8.44	3.28E-17	3.17E-16
Syt9	119.84	-8.61	0.62	-13.88	8.12E-44	5.47E-42
Esx1	51.55	-8.59	1.03	-8.35	6.57E-17	6.17E-16
Zic3	381.23	-8.56	0.54	-15.77	5.32E-56	6.73E-54
Fbxo15	668.09	-8.53	0.68	-12.61	1.96E-36	8.43E-35
Krt17	39.31	-8.48	0.90	-9.45	3.33E-21	4.69E-20
Tex21	44.15	-8.43	0.97	-8.68	4.12E-18	4.31E-17
4930461G14Rik	100.86	-8.36	0.76	-11.01	3.47E-28	8.39E-27
Cldn4	227.16	-8.35	0.98	-8.53	1.47E-17	1.46E-16
Ocln	65.84	-8.34	0.70	-11.93	8.70E-33	2.99E-31
Phlda2	46.19	-8.33	0.67	-12.51	6.94E-36	2.90E-34
Zbtb32	150.11	-8.32	0.51	-16.45	8.20E-61	1.44E-58
Insm1	57.41	-8.32	0.88	-9.44	3.73E-21	5.22E-20
Sh3gl2	52.40	-8.30	0.57	-14.66	1.24E-48	1.07E-46
Hsf2bp	96.95	-8.26	0.44	-18.98	2.57E-80	1.17E-77
Ildr1	76.56	-8.24	0.91	-9.06	1.36E-19	1.65E-18
Sox1	178.36	-8.24	0.83	-9.88	5.21E-23	8.52E-22
Gm10664	297.48	-8.21	0.69	-11.92	9.10E-33	3.11E-31
Mtap7d2	39.18	-8.17	0.55	-14.95	1.51E-50	1.46E-48
Gng3	43.61	-8.17	0.90	-9.10	8.86E-20	1.09E-18
Aqp3	400.57	-8.11	0.59	-13.85	1.19E-43	7.98E-42
4930500J02Rik	66.90	-8.09	0.87	-9.25	2.26E-20	2.91E-19
Nefl	344.02	-8.08	0.69	-11.72	1.01E-31	3.17E-30
Trim71	280.49	-8.07	0.72	-11.26	2.10E-29	5.53E-28
Gdf3	117.53	-8.07	0.54	-15.04	4.19E-51	4.24E-49
Fgfbp1	1177.21	-8.07	0.75	-10.78	4.25E-27	9.42E-26
Lin28b	79.77	-8.03	0.90	-8.91	5.20E-19	5.94E-18
Wfdc2	118.43	-8.02	0.81	-9.85	6.72E-23	1.09E-21
Sox15	39.40	-8.01	0.62	-12.82	1.29E-37	6.00E-36

Mreg	188.55	-7.98	0.45	-17.88	1.83E-71	5.23E-69
Tcfcp2l1	927.22	-7.98	0.63	-12.75	2.98E-37	1.34E-35
Nanog	37.06	-7.98	0.70	-11.33	9.06E-30	2.45E-28
Slc35f2	470.42	-7.94	0.36	-22.28	5.44E-110	7.04E-107
Trim6	271.33	-7.92	0.53	-14.80	1.46E-49	1.35E-47
Grik3	86.29	-7.91	0.76	-10.36	3.62E-25	6.95E-24
Kit	163.67	-7.89	0.57	-13.77	3.97E-43	2.54E-41
Gm13051	153.20	-7.88	0.73	-10.77	4.52E-27	9.96E-26
Bex4	125.13	-7.88	0.82	-9.61	7.05E-22	1.05E-20
Spnb3	47.65	-7.88	0.93	-8.44	3.17E-17	3.07E-16
Ina	57.66	-7.85	0.79	-9.90	4.36E-23	7.18E-22
Ppp2r2c	175.50	-7.80	0.67	-11.70	1.28E-31	3.96E-30
Gcnt3	570.51	-7.78	0.80	-9.69	3.39E-22	5.13E-21
Fut9	202.23	-7.71	0.69	-11.21	3.49E-29	9.05E-28
1600029D21Rik	72.45	-7.66	0.76	-10.14	3.70E-24	6.57E-23
Pipox	299.60	-7.61	0.67	-11.42	3.36E-30	9.37E-29
Alpk3	41.40	-7.60	0.91	-8.37	5.97E-17	5.62E-16
Padi4	42.21	-7.58	0.74	-10.27	9.83E-25	1.83E-23
Usp44	54.40	-7.52	0.82	-9.16	5.39E-20	6.72E-19
Cgn	116.88	-7.49	0.58	-12.90	4.73E-38	2.24E-36
Mycn	277.26	-7.46	0.52	-14.39	5.71E-47	4.49E-45
Calcoco2	36.69	-7.42	0.90	-8.23	1.85E-16	1.67E-15
1700019D03Rik	89.98	-7.41	0.72	-10.23	1.48E-24	2.71E-23
Pou3f1	44.00	-7.34	0.52	-14.25	4.60E-46	3.50E-44
Ybx2	128.64	-7.33	0.55	-13.29	2.67E-40	1.45E-38
Dlgap3	100.40	-7.32	0.61	-12.06	1.69E-33	6.08E-32
Dnahc8	62.98	-7.27	0.66	-10.99	4.45E-28	1.06E-26
1190003J15Rik	353.87	-7.25	0.78	-9.26	2.05E-20	2.66E-19
Podxl	1560.83	-7.23	0.63	-11.39	4.85E-30	1.34E-28
Rpp25	188.04	-7.22	0.41	-17.77	1.17E-70	3.18E-68
Tgm3	643.17	-7.20	0.64	-11.32	1.05E-29	2.84E-28
2310057J16Rik	137.38	-7.17	0.61	-11.77	5.38E-32	1.73E-30
Alpl	542.64	-7.15	0.59	-12.19	3.72E-34	1.42E-32
1-Sep	183.74	-7.15	0.37	-19.23	1.90E-82	9.71E-80
Rimklb	83.82	-7.13	0.69	-10.32	5.75E-25	1.09E-23
Rnf17	141.43	-7.12	0.39	-18.05	7.60E-73	2.37E-70
Bspry	50.66	-7.12	0.63	-11.36	6.73E-30	1.85E-28
Ush1c	212.70	-7.10	0.87	-8.14	4.05E-16	3.56E-15
Dtx1	52.36	-7.08	0.48	-14.62	2.11E-48	1.80E-46
Bex2	58.76	-7.08	0.77	-9.15	5.68E-20	7.08E-19
Casz1	112.03	-7.07	0.72	-9.77	1.47E-22	2.31E-21
Rasgrp2	65.68	-7.05	0.54	-13.04	7.40E-39	3.68E-37
Hap1	249.73	-7.05	0.55	-12.78	2.17E-37	9.91E-36
Cyp4f39	53.38	-7.02	0.93	-7.57	3.79E-14	2.80E-13
Hook1	192.82	-7.01	0.39	-17.88	1.77E-71	5.14E-69
Zfp936	38.48	-6.98	1.14	-6.13	8.75E-10	4.01E-09
Smtnl2	91.93	-6.98	0.58	-12.10	9.94E-34	3.65E-32
Vax2os2	52.80	-6.97	0.85	-8.22	2.01E-16	1.81E-15
Ckmt1	122.54	-6.97	0.77	-9.07	1.18E-19	1.44E-18
Slain1	41.55	-6.88	0.63	-10.91	1.01E-27	2.36E-26
Grhl3	57.50	-6.87	0.92	-7.44	9.86E-14	6.97E-13
Mapt	166.23	-6.82	0.52	-13.20	9.03E-40	4.79E-38
Slc2a3	1888.54	-6.81	0.38	-18.02	1.25E-72	3.84E-70
Clstn3	45.61	-6.77	0.72	-9.46	3.12E-21	4.40E-20
Ccdc88c	480.18	-6.75	0.49	-13.70	9.94E-43	6.19E-41
Ooep	81.68	-6.73	0.72	-9.35	9.06E-21	1.22E-19
Palm3	132.28	-6.71	0.55	-12.27	1.30E-34	5.15E-33
Sigirr	390.07	-6.68	0.70	-9.54	1.44E-21	2.09E-20
Zfp534	108.65	-6.68	0.92	-7.25	4.15E-13	2.73E-12
Olig1	84.46	-6.66	0.72	-9.23	2.73E-20	3.49E-19
Cyp2b23	326.41	-6.63	0.79	-8.40	4.35E-17	4.16E-16
Atp2c2	121.92	-6.62	1.06	-6.21	5.15E-10	2.43E-09
Robo4	38.94	-6.55	1.06	-6.19	5.94E-10	2.79E-09
Ttn	54.04	-6.53	0.89	-7.35	1.92E-13	1.31E-12
Ryr1	78.37	-6.51	0.50	-12.90	4.57E-38	2.17E-36
Mtap7	407.91	-6.51	0.29	-22.17	7.46E-109	8.96E-106
Folr1	79.36	-6.51	0.61	-10.61	2.73E-26	5.73E-25
Snhg11	53.79	-6.49	0.67	-9.75	1.90E-22	2.94E-21
Krt8	376.27	-6.49	0.88	-7.41	1.28E-13	8.86E-13
Slc38a5	3865.91	-6.48	1.01	-6.39	1.69E-10	8.41E-10
Dpysl5	44.28	-6.43	0.61	-10.47	1.14E-25	2.27E-24
St14	622.16	-6.42	0.48	-13.24	5.32E-40	2.84E-38
Cacna2d2	725.34	-6.39	0.67	-9.58	9.88E-22	1.45E-20
Gstp2	114.73	-6.36	0.69	-9.19	3.96E-20	5.01E-19
Tgm1	217.17	-6.32	0.56	-11.24	2.52E-29	6.61E-28
Esyt3	196.72	-6.32	0.79	-8.04	8.75E-16	7.46E-15
Hif3a	94.47	-6.28	0.66	-9.50	2.17E-21	3.09E-20
Tet1	1183.09	-6.27	0.52	-11.95	6.27E-33	2.19E-31
B4galnt3	55.30	-6.26	0.55	-11.32	1.05E-29	2.84E-28
Lad1	112.56	-6.24	0.83	-7.53	5.21E-14	3.79E-13
Dnmt3b	410.85	-6.22	0.26	-24.27	4.13E-130	9.94E-127
Mylpf	160.43	-6.22	0.47	-13.37	9.04E-41	5.02E-39
Piwil2	79.45	-6.21	0.54	-11.41	3.70E-30	1.02E-28
Nccrp1	1842.72	-6.18	0.92	-6.71	1.99E-11	1.11E-10
Fcho1	268.01	-6.16	0.49	-12.56	3.36E-36	1.43E-34
Zfp459	41.29	-6.11	0.66	-9.28	1.66E-20	2.19E-19
Klk1	3133.80	-6.11	0.94	-6.48	9.19E-11	4.73E-10
Spint1	222.60	-6.11	0.77	-7.96	1.72E-15	1.43E-14
Cpn1	100.09	-6.08	1.01	-6.00	1.94E-09	8.53E-09
Gli1	105.08	-6.08	0.55	-11.08	1.50E-28	3.75E-27
Nlrp1a	272.77	-6.06	0.58	-10.40	2.43E-25	4.72E-24
Lsr	109.38	-6.04	0.64	-9.40	5.42E-21	7.47E-20
Tjp3	72.74	-6.04	0.72	-8.40	4.63E-17	4.42E-16
Mapk4	54.54	-6.01	0.51	-11.90	1.16E-32	3.92E-31
Cbx7	287.20	-6.00	0.55	-10.85	2.09E-27	4.71E-26
Lrrc2	152.03	-5.99	0.71	-8.39	4.83E-17	4.60E-16
Lrp2	317.53	-5.95	0.64	-9.29	1.61E-20	2.12E-19
Ccnb1ip1	39.84	-5.93	0.68	-8.72	2.69E-18	2.87E-17
Syt7	64.43	-5.93	0.38	-15.45	7.18E-54	8.63E-52
Zfp473	82.76	-5.91	0.39	-15.18	4.60E-52	4.90E-50
Chchd10	1058.04	-5.90	0.63	-9.41	4.76E-21	6.61E-20

Zfp296	182.71	-5.87	0.55	-10.66	1.51E-26	3.25E-25
Pkp3	133.91	-5.87	0.53	-11.01	3.43E-28	8.30E-27
Cubn	288.23	-5.86	0.67	-8.77	1.80E-18	1.96E-17
Mybl2	2078.26	-5.81	0.22	-26.52	6.27E-155	3.51E-151
Itgb7	875.82	-5.80	0.54	-10.71	9.57E-27	2.08E-25
Hal	49.13	-5.78	0.55	-10.54	5.80E-26	1.18E-24
Tmprss5	107.70	-5.78	0.90	-6.41	1.44E-10	7.23E-10
Cyp2s1	699.41	-5.72	0.62	-9.21	3.17E-20	4.04E-19
Jag2	449.13	-5.69	0.45	-12.76	2.65E-37	1.20E-35
Lefty1	132.89	-5.69	0.47	-12.18	3.78E-34	1.44E-32
Sall3	41.92	-5.68	0.64	-8.95	3.68E-19	4.27E-18
Celsr1	481.46	-5.64	0.39	-14.44	2.90E-47	2.33E-45
Rec8	40.90	-5.64	0.86	-6.57	4.91E-11	2.62E-10
Slc13a5	39.59	-5.63	0.67	-8.43	3.36E-17	3.24E-16
Spnb1	218.58	-5.63	0.66	-8.48	2.21E-17	2.16E-16
Car2	323.69	-5.61	0.60	-9.37	7.42E-21	1.01E-19
Myh14	117.50	-5.61	0.69	-8.14	4.11E-16	3.61E-15
Liph	118.56	-5.61	0.55	-10.20	2.02E-24	3.64E-23
Pllp	146.23	-5.60	0.68	-8.20	2.39E-16	2.15E-15
Plekhh1	304.84	-5.57	0.50	-11.22	3.10E-29	8.07E-28
B4galnt4	647.92	-5.56	0.46	-11.99	4.20E-33	1.47E-31
Rhpn2	79.58	-5.54	0.56	-9.83	8.52E-23	1.37E-21
Rab11fip4	64.09	-5.54	0.54	-10.32	5.47E-25	1.04E-23
Nup210	564.36	-5.52	0.65	-8.53	1.44E-17	1.44E-16
Ifitm1	1435.37	-5.49	0.55	-9.90	4.10E-23	6.78E-22
B3gnt7	205.48	-5.48	0.57	-9.58	9.96E-22	1.46E-20
Spnb4	109.50	-5.45	0.45	-12.24	1.84E-34	7.13E-33
Spna1	93.93	-5.45	0.53	-10.36	3.87E-25	7.40E-24
Hpcal4	39.16	-5.43	0.96	-5.64	1.72E-08	6.80E-08
Grtp1	240.13	-5.41	0.46	-11.74	7.74E-32	2.44E-30
Bcam	431.06	-5.41	0.39	-13.75	5.15E-43	3.26E-41
Kcnk5	120.80	-5.41	0.36	-15.03	4.41E-51	4.44E-49
Slc37a1	180.03	-5.41	0.53	-10.19	2.20E-24	3.96E-23
Amt	153.69	-5.39	0.41	-13.16	1.50E-39	7.83E-38
Ptch2	49.95	-5.39	0.66	-8.20	2.48E-16	2.22E-15
Pnldc1	37.20	-5.35	0.56	-9.48	2.44E-21	3.47E-20
Tesc	74.99	-5.35	0.60	-8.98	2.73E-19	3.20E-18
Zfp750	44.23	-5.32	0.61	-8.72	2.71E-18	2.89E-17
Dpp4	54.65	-5.31	0.71	-7.44	9.84E-14	6.95E-13
Frem2	288.20	-5.30	0.55	-9.71	2.86E-22	4.37E-21
Atp1a3	157.66	-5.26	0.82	-6.43	1.28E-10	6.48E-10
Epb4.9	121.32	-5.25	0.44	-11.85	2.20E-32	7.27E-31
Foxd3	63.82	-5.24	0.75	-6.96	3.38E-12	2.03E-11
Gm5860	44.82	-5.22	0.78	-6.69	2.29E-11	1.27E-10
Dmkn	131.57	-5.21	0.69	-7.58	3.41E-14	2.54E-13
Exoc3l	95.57	-5.18	0.59	-8.84	9.79E-19	1.09E-17
Rps6ka6	196.71	-5.17	0.48	-10.77	4.98E-27	1.09E-25
Cobl	1709.64	-5.16	0.38	-13.43	4.07E-41	2.32E-39
Gbx2	64.09	-5.14	0.88	-5.88	4.20E-09	1.79E-08
Dusp9	267.21	-5.13	0.61	-8.45	2.81E-17	2.73E-16
Fabp3	190.21	-5.12	0.38	-13.52	1.20E-41	6.99E-40
E130012A19Rik	444.81	-5.10	0.42	-12.28	1.09E-34	4.36E-33
C130074G19Rik	53.49	-5.06	0.60	-8.47	2.51E-17	2.44E-16
Zic5	80.69	-5.06	0.69	-7.32	2.47E-13	1.66E-12
N4bp3	257.54	-5.04	0.36	-13.82	1.83E-43	1.21E-41
Celsr3	48.05	-5.03	0.50	-10.08	6.45E-24	1.13E-22
Gata4	40.11	-5.03	0.81	-6.24	4.33E-10	2.06E-09
Prr15l	39.57	-5.03	1.18	-4.25	2.16E-05	5.84E-05
Krtdap	81.29	-5.02	1.01	-4.98	6.31E-07	2.09E-06
Enpp3	1882.63	-5.00	0.41	-12.15	5.50E-34	2.07E-32
Hoxa11as	454.75	5.01	0.54	9.26	1.98E-20	2.58E-19
Msrb3	4245.78	5.01	0.28	18.17	9.07E-74	3.05E-71
Pcdhb5	48.64	5.01	0.77	6.50	8.21E-11	4.26E-10
Hoxd13	1474.80	5.01	0.54	9.26	1.96E-20	2.55E-19
C430049B03Rik	89.98	5.02	0.49	10.23	1.41E-24	2.59E-23
Nckap5	91.99	5.03	0.75	6.75	1.49E-11	8.38E-11
Hoxd8	196.54	5.03	0.33	15.45	7.67E-54	9.15E-52
Col16a1	4685.18	5.04	0.25	20.55	7.50E-94	6.01E-91
Ptprd	1029.32	5.04	0.59	8.51	1.75E-17	1.73E-16
Rgs4	1177.96	5.04	0.44	11.36	6.70E-30	1.84E-28
Mid2	850.48	5.05	0.40	12.75	3.04E-37	1.36E-35
Cxcl12	8003.82	5.05	0.54	9.44	3.79E-21	5.31E-20
Adam12	3466.06	5.06	0.57	8.93	4.38E-19	5.03E-18
Slc2a10	368.31	5.06	0.39	12.93	3.20E-38	1.54E-36
Ptprc	166.58	5.07	0.84	6.01	1.90E-09	8.37E-09
Abca9	69.86	5.07	0.82	6.17	6.90E-10	3.21E-09
Ctso	381.79	5.07	0.36	14.11	3.20E-45	2.30E-43
Ano5	50.63	5.08	0.83	6.10	1.06E-09	4.82E-09
Fndc1	4271.59	5.08	0.53	9.52	1.66E-21	2.38E-20
Dpep1	398.06	5.08	0.59	8.69	3.65E-18	3.85E-17
Olfml2b	1712.23	5.09	0.35	14.34	1.23E-46	9.57E-45
Plscr2	191.57	5.09	0.61	8.37	5.90E-17	5.57E-16
Trp63	73.65	5.10	0.90	5.63	1.80E-08	7.10E-08
Fam20a	1142.70	5.10	0.51	10.05	9.52E-24	1.64E-22
Prkg1	600.55	5.10	0.52	9.75	1.76E-22	2.75E-21
Twist2	481.90	5.11	0.47	10.78	4.51E-27	9.94E-26
Csf1r	848.79	5.11	0.74	6.93	4.35E-12	2.59E-11
Akr1c14	390.30	5.11	0.80	6.42	1.39E-10	6.96E-10
Rnf150	1503.86	5.11	0.46	11.19	4.71E-29	1.21E-27
Zfp521	1707.38	5.11	0.50	10.32	5.99E-25	1.13E-23
Fam101b	1998.90	5.11	0.43	11.78	4.85E-32	1.57E-30
Hgf	159.17	5.13	0.74	6.98	2.97E-12	1.79E-11
Foxc1	878.69	5.14	0.44	11.77	5.63E-32	1.80E-30
Pcdh11x	63.28	5.14	0.75	6.86	7.00E-12	4.06E-11
Plek	391.11	5.14	0.74	6.94	3.93E-12	2.34E-11
Pcdhb20	94.23	5.15	0.46	11.14	8.38E-29	2.12E-27
Meox1	142.67	5.15	0.78	6.59	4.27E-11	2.29E-10
Col4a5	3739.62	5.16	0.36	14.25	4.43E-46	3.39E-44
Sla	44.92	5.16	0.91	5.68	1.36E-08	5.44E-08
Il16	61.68	5.17	0.80	6.47	1.01E-10	5.18E-10

Nfix	2564.54	5.17	0.58	8.87	7.00E-19	7.93E-18
Ddr2	6861.93	5.17	0.32	16.39	2.09E-60	3.63E-58
Matn4	356.17	5.18	0.81	6.40	1.51E-10	7.58E-10
Igsf10	3224.11	5.18	0.62	8.29	1.10E-16	1.01E-15
Itm2a	5095.47	5.18	0.63	8.29	1.15E-16	1.06E-15
Dnm3os	2181.03	5.18	0.47	11.10	1.32E-28	3.32E-27
Shox2	146.64	5.18	0.55	9.40	5.72E-21	7.85E-20
Prr16	136.02	5.18	0.70	7.45	9.54E-14	6.75E-13
Col6a1	34014.76	5.19	0.49	10.61	2.61E-26	5.49E-25
Kcne4	83.80	5.19	0.74	7.01	2.38E-12	1.45E-11
Hand2	200.53	5.20	0.40	13.07	4.68E-39	2.36E-37
Stab1	748.05	5.20	0.88	5.93	3.04E-09	1.31E-08
Hoxa13	134.35	5.20	0.66	7.89	3.04E-15	2.48E-14
Pcdhb14	61.26	5.22	0.58	8.94	3.83E-19	4.43E-18
Slc9a9	209.49	5.24	0.66	7.99	1.35E-15	1.13E-14
Sfrp2	4689.54	5.24	1.05	5.00	5.60E-07	1.87E-06
Tox	176.96	5.25	0.67	7.86	3.95E-15	3.19E-14
Gdf10	254.32	5.26	0.98	5.35	9.01E-08	3.30E-07
Galnt13	229.43	5.26	1.03	5.10	3.35E-07	1.15E-06
Gm15663	38.81	5.26	0.64	8.21	2.12E-16	1.91E-15
Angpt1	772.83	5.27	0.69	7.58	3.37E-14	2.51E-13
Robo2	325.61	5.27	0.74	7.13	9.76E-13	6.17E-12
Slc8a1	912.22	5.27	0.48	10.89	1.24E-27	2.85E-26
Hs3st3a1	131.07	5.29	0.43	12.39	3.14E-35	1.28E-33
Adra1d	619.77	5.29	0.36	14.91	2.84E-50	2.70E-48
Gjb2	695.60	5.30	0.62	8.52	1.53E-17	1.52E-16
Spock3	103.66	5.30	0.83	6.40	1.53E-10	7.66E-10
Snai2	870.13	5.31	0.45	11.73	8.99E-32	2.82E-30
Cmya5	134.71	5.32	0.79	6.73	1.76E-11	9.81E-11
Mafb	474.84	5.32	0.62	8.63	6.09E-18	6.29E-17
Wnt16	343.67	5.32	0.62	8.58	9.73E-18	9.89E-17
Dcn	4019.44	5.32	0.52	10.18	2.52E-24	4.53E-23
Tspan11	285.22	5.34	0.76	7.06	1.64E-12	1.02E-11
Fbn1	40815.00	5.36	0.45	11.82	3.06E-32	1.00E-30
Hoxd12	348.47	5.37	0.76	7.11	1.18E-12	7.40E-12
A930038C07Rik	332.67	5.37	0.71	7.52	5.31E-14	3.86E-13
Clec2d	423.56	5.37	0.55	9.80	1.08E-22	1.71E-21
Gpr64	978.53	5.38	0.67	8.00	1.23E-15	1.04E-14
Scn7a	109.21	5.38	0.81	6.62	3.65E-11	1.98E-10
6330406I15Rik	2575.58	5.40	0.34	15.89	7.19E-57	9.69E-55
Nkx3-2	47.82	5.41	0.81	6.71	1.92E-11	1.07E-10
Fam198a	79.51	5.42	0.97	5.59	2.25E-08	8.82E-08
Col6a2	24286.75	5.45	0.52	10.52	7.02E-26	1.42E-24
Fat4	3610.19	5.46	0.74	7.41	1.30E-13	9.00E-13
Kif26b	1583.58	5.46	0.63	8.68	4.11E-18	4.30E-17
Cdh11	10018.54	5.47	0.40	13.76	4.27E-43	2.73E-41
Mylk	1029.10	5.47	0.54	10.13	4.26E-24	7.49E-23
Lair1	37.71	5.48	0.87	6.28	3.28E-10	1.58E-09
Il21r	43.52	5.49	0.90	6.07	1.27E-09	5.70E-09
Rbms3	1354.09	5.50	0.38	14.40	4.81E-47	3.80E-45
Adamts1	8031.58	5.50	0.59	9.32	1.19E-20	1.59E-19
Wisp1	6508.52	5.50	0.46	12.01	3.20E-33	1.13E-31
Htra3	763.20	5.53	0.67	8.24	1.77E-16	1.61E-15
Mef2c	371.99	5.53	0.63	8.81	1.28E-18	1.42E-17
Igf2	23821.61	5.53	0.57	9.69	3.20E-22	4.86E-21
Bmp3	45.57	5.54	0.98	5.66	1.51E-08	6.03E-08
Adamtsl3	1563.37	5.56	0.61	9.06	1.29E-19	1.56E-18
Dhrs9	202.75	5.56	0.86	6.48	9.30E-11	4.78E-10
Tgfbr2	5467.67	5.56	0.48	11.54	8.59E-31	2.50E-29
Col28a1	343.18	5.57	0.64	8.75	2.09E-18	2.25E-17
Dclk1	2001.81	5.57	0.62	8.95	3.49E-19	4.06E-18
Rspo3	909.35	5.58	0.68	8.20	2.34E-16	2.10E-15
C1qtnf6	2167.18	5.58	0.31	17.85	2.74E-71	7.70E-69
Col5a1	88100.99	5.58	0.39	14.15	1.93E-45	1.40E-43
Pcdhb19	106.65	5.59	0.52	10.82	2.83E-27	6.33E-26
Csgalnact1	1655.92	5.59	0.38	14.84	8.52E-50	7.96E-48
Fxyd1	221.61	5.60	0.90	6.22	5.11E-10	2.42E-09
Irx1	693.99	5.61	0.34	16.31	8.10E-60	1.31E-57
Dkk2	1004.07	5.62	0.67	8.34	7.72E-17	7.18E-16
Il1r1	1095.29	5.63	0.54	10.43	1.78E-25	3.49E-24
Igf1	1788.99	5.63	0.80	7.07	1.59E-12	9.83E-12
Lsp1	2011.06	5.65	0.34	16.64	3.61E-62	6.68E-60
Gm885	38.38	5.66	0.90	6.28	3.37E-10	1.62E-09
Clec11a	991.74	5.66	0.70	8.11	5.12E-16	4.45E-15
Chrdl1	166.06	5.69	0.77	7.38	1.64E-13	1.13E-12
Ldb2	191.76	5.69	0.61	9.40	5.31E-21	7.32E-20
H19	27870.60	5.70	0.57	10.00	1.55E-23	2.64E-22
Slc24a3	330.52	5.71	0.71	8.01	1.12E-15	9.49E-15
Hapln1	389.69	5.72	0.93	6.16	7.21E-10	3.35E-09
Pappa2	103.78	5.73	0.65	8.76	1.89E-18	2.05E-17
Gdf6	220.29	5.74	0.79	7.30	2.97E-13	1.99E-12
Col12a1	66704.18	5.74	0.55	10.42	2.04E-25	3.97E-24
Lrrn3	145.35	5.74	0.83	6.94	3.99E-12	2.38E-11
Gpm6b	1448.42	5.74	0.67	8.57	1.05E-17	1.06E-16
Sox5	137.38	5.75	0.68	8.43	3.58E-17	3.45E-16
Gm12824	736.29	5.75	0.82	7.00	2.62E-12	1.59E-11
Mrc1	1741.61	5.75	0.93	6.17	6.74E-10	3.14E-09
Cxcl5	2547.20	5.76	0.82	6.99	2.74E-12	1.66E-11
Sec16b	399.83	5.76	0.55	10.42	2.10E-25	4.08E-24
Tbx15	2992.59	5.77	0.38	15.19	3.95E-52	4.24E-50
Penk	7277.17	5.77	1.09	5.31	1.12E-07	4.07E-07
1500015O10Rik	1100.98	5.77	0.71	8.11	5.13E-16	4.46E-15
Erg	234.41	5.77	0.78	7.40	1.39E-13	9.64E-13
Pcdhb16	140.25	5.79	0.61	9.50	2.16E-21	3.08E-20
Cd300ld	41.40	5.81	0.95	6.11	1.02E-09	4.63E-09
Pcdh10	176.68	5.82	0.72	8.06	7.78E-16	6.66E-15
Thbs2	27576.84	5.82	0.49	11.79	4.50E-32	1.45E-30
Ptgfr	245.96	5.85	0.67	8.69	3.57E-18	3.77E-17
Vipr2	38.79	5.85	0.97	6.04	1.52E-09	6.77E-09
Dmrt2	44.74	5.87	0.76	7.75	9.13E-15	7.14E-14

Arsi	717.30	5.88	0.72	8.21	2.27E-16	2.04E-15
Slc27a6	46.10	5.89	0.82	7.16	8.12E-13	5.17E-12
Tlr7	67.95	5.93	0.86	6.93	4.24E-12	2.52E-11
Wisp2	9559.04	5.94	0.59	10.14	3.83E-24	6.77E-23
Hoxa10	729.25	5.94	0.37	16.24	2.76E-59	4.30E-57
Podn	1810.86	5.95	0.64	9.25	2.19E-20	2.83E-19
Tmem119	2341.21	5.95	0.29	20.23	5.40E-91	4.13E-88
Col5a2	90968.52	5.96	0.34	17.51	1.17E-68	3.02E-66
Pgm5	374.54	5.96	0.67	8.90	5.35E-19	6.10E-18
Svep1	9399.53	5.96	0.76	7.80	6.00E-15	4.75E-14
Dpt	902.70	5.97	0.74	8.02	1.09E-15	9.21E-15
Hmcn1	1741.18	6.02	0.88	6.83	8.58E-12	4.93E-11
Eya4	642.03	6.02	0.56	10.77	4.60E-27	1.01E-25
Cpa6	142.14	6.03	0.50	12.06	1.75E-33	6.29E-32
Gprin3	98.16	6.04	0.84	7.15	8.65E-13	5.50E-12
Ptx3	4912.69	6.06	0.77	7.83	4.81E-15	3.85E-14
Fibin	774.77	6.06	0.54	11.23	2.99E-29	7.81E-28
Ebf3	450.67	6.06	0.48	12.52	5.73E-36	2.41E-34
Mfap4	2565.34	6.07	0.65	9.35	8.63E-21	1.17E-19
Fam180a	172.85	6.10	0.75	8.14	3.90E-16	3.44E-15
Pcdh18	1600.31	6.10	0.40	15.25	1.58E-52	1.72E-50
Pik3cg	80.17	6.13	0.96	6.41	1.45E-10	7.26E-10
Osr1	795.25	6.16	0.67	9.13	6.82E-20	8.40E-19
Chodl	117.55	6.23	0.85	7.30	2.84E-13	1.90E-12
Clec14a	287.37	6.24	0.80	7.77	7.76E-15	6.09E-14
C1qtnf3	1993.10	6.25	0.48	12.90	4.32E-38	2.06E-36
BC055004	107.80	6.27	0.96	6.55	5.70E-11	3.01E-10
Agtr2	842.16	6.27	1.02	6.13	8.69E-10	3.99E-09
Foxp2	281.25	6.27	0.69	9.05	1.40E-19	1.68E-18
Col1a1	203933.49	6.28	0.55	11.48	1.62E-30	4.61E-29
Capn6	2241.47	6.30	0.60	10.50	8.27E-26	1.66E-24
Tnn	1615.94	6.31	0.59	10.66	1.62E-26	3.48E-25
F13a1	289.46	6.31	0.88	7.18	7.05E-13	4.52E-12
Maf	690.29	6.32	0.67	9.43	3.98E-21	5.56E-20
Srpx	1027.35	6.33	0.46	13.82	2.00E-43	1.30E-41
Lum	3100.31	6.34	0.87	7.28	3.39E-13	2.25E-12
Col1a2	196788.78	6.36	0.35	18.15	1.25E-73	4.05E-71
Col11a1	16441.58	6.37	0.42	15.21	2.85E-52	3.08E-50
Tnmd	294.57	6.38	0.53	12.09	1.20E-33	4.40E-32
Ptprq	219.86	6.47	0.90	7.20	5.93E-13	3.83E-12
Gucy1a3	160.48	6.48	0.57	11.47	1.93E-30	5.48E-29
Abi3bp	5115.94	6.49	0.98	6.61	3.96E-11	2.14E-10
Col14a1	706.85	6.49	0.57	11.42	3.37E-30	9.37E-29
Fbln5	14206.22	6.53	0.25	25.66	3.18E-145	1.34E-141
Ccl12	88.35	6.58	0.98	6.74	1.55E-11	8.73E-11
Gxylt2	423.40	6.64	0.66	10.04	1.01E-23	1.74E-22
C1qtnf7	111.86	6.65	0.78	8.48	2.35E-17	2.29E-16
Adamts16	66.86	6.68	0.82	8.19	2.72E-16	2.42E-15
Fmod	5087.00	6.69	0.62	10.85	1.98E-27	4.47E-26
Zcchc5	1204.93	6.70	0.72	9.31	1.34E-20	1.78E-19
Lgr5	985.90	6.70	0.90	7.48	7.29E-14	5.22E-13
Gas1	4408.80	6.71	0.79	8.48	2.27E-17	2.22E-16
Ptn	4269.23	6.72	0.51	13.09	3.99E-39	2.02E-37
Cxcl15	551.42	6.74	0.98	6.90	5.38E-12	3.17E-11
Cbr2	811.93	6.81	0.69	9.86	6.51E-23	1.06E-21
Epha3	647.56	6.84	0.63	10.82	2.64E-27	5.93E-26
Igfbp5	5458.87	6.90	0.68	10.20	1.96E-24	3.55E-23
Gpr88	57.17	6.90	0.94	7.35	1.99E-13	1.35E-12
Cx3cr1	198.65	7.07	0.78	9.01	1.98E-19	2.35E-18
Col8a2	1783.00	7.09	0.84	8.41	4.13E-17	3.95E-16
Col3a1	129904.26	7.09	0.55	12.97	1.85E-38	9.04E-37
Palmd	603.68	7.17	0.67	10.70	1.01E-26	2.18E-25
Postn	48109.06	7.18	0.42	17.05	3.28E-65	7.27E-63
Tmem26	169.59	7.31	0.91	8.01	1.15E-15	9.71E-15
Fcrls	697.23	7.32	0.87	8.45	2.83E-17	2.75E-16
Fbn2	9045.40	7.37	0.59	12.55	3.91E-36	1.66E-34
Aspn	4866.84	7.38	0.55	13.48	2.14E-41	1.24E-39
Ogn	9428.26	7.38	0.67	11.04	2.37E-28	5.84E-27
Lrrc17	2845.52	7.79	0.38	20.76	1.07E-95	1.06E-92
Nov	3150.28	7.91	0.65	12.14	6.65E-34	2.49E-32
Eln	18771.38	7.93	0.74	10.71	8.82E-27	1.92E-25

Method Details

Cell lines and culture conditions

Primary MEFs harboring a heterozygous R26-M2rtTA allele and dox-inducible polycistronic transgenic cassette coding for OSKM in the Col1A locus (tetO-OSKM)[16,17] were harvested from day 14.5 embryos of timed mouse pregnancies. The Chancellor's Animal Research Committee at University of California Los Angeles approved our animal breeding and research protocols for this purpose. For the reprogramming experiment involving the CDH1-positive population sort, MEFs were harvested from R26-M2rtTA mice which additionally carried a HMRPS-Bcl2 transgene[32]. For Esrrb overexpression, a lentiviral construct encoding the tet-inducible Essrb cDNA, obtained from [13], was transfected alongside viral packaging vectors (pMDLg, pRSV-REV, pCMV-VSVG) into 293T cells using the CalPhos mammalian transfection kit (Clontech 062013) as per manufacturer's instructions. Lentiviral production was performed for 48 hours, and the harvested supernatant used to infect tetO-OSKM/M2rtTA MEFs twice, once every 24 hours. For a control reprogramming time course for the Esrrb+ reprogramming experiment, tetO-OSKM/M2rtTA MEFs were infected with a pMX retrovirus encoding the fluorophore Tomato (time course 3). MEFs were grown in ESC media containing knockout DMEM, 15% fetal bovine serum (FBS), recombinant leukemia inhibitory factor, b-mercaptoethanol, 1x penicillin/streptomycin, L-glutamine, and non-essential amino acids. Reprogramming cultures were split by manual disruption in trypsin and were plated onto irradiated MEFs (feeders). To initiate reprogramming and Esrrb expression, respectively, cells were cultured in ESC medium containing 2mg/ml doxycycline, in media containing knockout serum replacement instead of FBS. For sorting of CDH1-positive cells, a day 6 reprogramming culture was collected by trypsin digestion followed by incubation with antibodies against CDH1 (Abcam ab11512) then incubated with anti-rat IGG Alexa 488 secondary antibody (Abcam ab150157), and sorting was performed as previously described 13 on a FACS Aria III sorter.

Immunostaining

Cells were grown on coverslips pretreated with 0.3% gelatin (Sigma G2500). After fixation with 4% paraformaldehyde, cells were washed with 1xPBS-0.05% Tween, and permeabilized with 1xPBS-0.5% Triton-X. Primary antibody incubation was carried out at 4C overnight, secondary antibody incubation at RT for 30min, each in blocking buffer (5% donkey serum, 1xPBS, 0.2% Tween, and 0.2% fish skin gelatin). Antibodies used: anti-Crym (Abcam ab54669 1/200) and anti-CDH1 (Abcam ab11512, 1/200).

Single cell RNA-sequencing

Reprogramming cultures from one well of a 6-well plate were harvested upon trypsin treatment, passed through a 40uM filter, and resuspended in 0.01% BSA in 1xPBS at approximately 150 cells/ul. For time course 2, human cells (UCLA9 hESC) were mixed 1:1 into the cell mixture for the purpose of confirming lack of cross-species cell mixing (data not shown). Cells were co-flowed with barcodes beads (Chemgenes) in a microfluidics device (PDMS Drop-seq device, Flowjem), and isolated for reverse transcription as described 25. Libraries were constructed with KAPA polyermase and Nextera XT preparation kit as previously described 25.

Processing, read alignment and digital gene expression (DGE) matrix construction

Raw sequencing data were filtered by read quality, adapter- and polyA-trimmed, and reads satisfying a length threshold of 30 nucleotides were aligned to the mouse (mm9) genome using Bowtie2 (v2.2.9 with the '--very-sensitive' mode). For all time course experiments, we filtered human and mixed species barcodes by a cutoff of <20% of mapped human reads allowable. Aligned reads were tagged to gene exons using Bedtools Intersect (v2.26.0). DGE matrices were then generated by counting gene transcripts for all cells within each time point from all Drop-seq experiments using custom Python scripts. To correct for any bead synthesis errors/read errors leading to false barcodes, reads with the same corresponding cell barcode were aggregated

together, and unique molecular identifiers (UMIs) and cell barcodes were merged within 1 Hamming and 2 Levenshtein distances, respectively 52. We excluded cells with <500 expressed (>0 transcripts) genes and <1250 transcripts from all downstream analyses. For separation of transgenic and endogenous Pou5f1 and Sox2 reads, we took only reads which mapped to the respective unique 3' UTR sequences (to the endogenous loci or the transgenic UTR) to determine transcript number.

Dimensionality reduction of single cell RNA-seq data

DGE matrices were normalized by the total number of transcripts per cell in log space by dividing raw counts by the total number of transcripts per cell, then multiplied by 10,000. Cells were projected onto a 2D embedding using t-Distributed Stochastic Neighbor Embedding (tSNE, perplexity set to 30), Uniform Manifold Approximation and Projection (UMAP, neighbors set to 30), SPRING (standalone version 1.5, neighbors set to 5) or Palantir (standalone version 0.2.1, default parameters) with cell loadings associated with 30 principal components utilizing all expressed (>0 transcripts) genes as input (R packages 'irlba', 'Rtsne', 'umap').

Trajectory score assignment

MEF, ESC, NSC, and Keratinocyte bulk RNA-seq data were used to calculate signature genes. Signature genes were defined as those most differentially (log2FoldChange>5 and FDR<0.01) expressed genes between the starting cell type and the ending cell type. The average gene expression of these differential genes was computed for each cell, linearly transformed between 0 and 1, and used as signature gene scores for each cell. The difference between these normalized scores, for each cell, was termed the trajectory score (ending signature gene score-start signature gene score). For ESC- and MEF-signature scores, time course 1 had the maximum values across all experiments therefore, the trajectory score calculations for all other experiments used these maxima. Trajectory score ordering was compared to the pseudo-temporal ordering of

cells based on Monocle2 (R package ‘monocle’). Monocle2 analyses were performed using all expressed genes (>0 transcripts across all cells, just as the tSNE embedding was preformed) and default settings.

Determination of gene expression changes during reprogramming

To build gene expression networks, we define a gene’s weight within a given cell as a function of its expression and the cell’s progression along the trajectory score. A weight is computed for each cell and each gene, then averaged across cells to create a meta-weight. This meta-weight is used as a metric to bin genes into ‘rough’ networks. Each of these ‘rough’ networks are pruned by computing independent Pearson gene-to-gene correlation matrix using the expressed genes of a network and identify genes with correlation values greater than 0.20. We linked these genes with edges to its next most correlated gene and recursively expand the network until the correlation threshold is not met. We observe that gene networks typically stop expanding after 7 rounds of edge linking. Next, we K-means cluster our cells into 95 clusters reasoning roughly 50 cells per cluster, then the average expression of all gene members in a given network was computed for each cluster by taking the average of all normalized expression value for all genes and cells in the network. Because the number of gene networks are large and many gene networks exhibit similar expression, we recursively merge gene networks on the basis of expression correlation by computing a Pearson correlation matrix on the average expression values and link clusters with correlation values greater than 0.70. For plotting on tSNE graphs, average normalized expression of all genes in the network was calculated per cell.

Stratification of genes according to CpG content

Based on CpG content, we classified mouse promoters (-500bp +50 bp around the TSS of mm9 annotated genes) into three groups, high CpG promoters (HCP), intermediate CpG promoters (ICP) and low CpG promoters (LCP) categories. CpG-poor correspond to LCP, weak CpG islands

to ICP and strong CpG islands to HCP. Classification was obtained by calculating the CpG observed vs expected ratios (R) within the aforementioned windows according to published criteria; LCPs R<=0.48, ICPs, 0.48<R<0.75, HCPs, R>=0.75.

The single cell RNA-seq data are uploaded to Gene Expression Omnibus (GEO) and will be made publicly available upon publication.

References

1. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
2. Cherry, A. B. C. & Daley, G. Q. Reprogrammed Cells for Disease Modeling and Regenerative Medicine. *Annu. Rev. Med*. **64**, 277–290 (2013).
3. Passier, R., Orlova, V. & Mummery, C. Complex Tissue and Disease Modeling using hiPSCs. *Cell Stem* Cell **18**, 309–321 (2016).
4. Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J. & Plath, K. Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442–459.e20 (2017).
5. Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., van Oudenaarden, A. & Jaenisch, R. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150**, 1209–1222 (2012).
6. Lujan, E., Zunder, E. R., Ng, Y. H., Goronzy, I. N., Nolan, G. P. & Wernig, M. Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* **521**, 352–356 (2015).
7. O'Malley, J., Skylaki, S., Iwabuchi, K. A., Chantzoura, E., Ruetz, T., Johnsson, A., Tomlinson, S. R., Linnarsson, S. & Kaji, K. High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature* **499**, 88–91 (2013).
8. 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. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell* **151**, 1617–1632 (2012).
9. Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G. P. A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. *Cell Stem Cell* **16**, 323–337 (2015).
10. Kim, D. H., Marinov, G. K., Pepke, S., Singer, Z. S., He, P., Williams, B., Schroth, G. P., Elowitz, M. B. & Wold, B. J. Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell* **16**, 88–101 (2015).
11. Pasque, V., Tchieu, J., Karnik, R., Uyeda, M., Sadhu Dimashkie, A., Case, D., Papp, B., Bonora, G., Patel, S., Ho, R., Schmidt, R., McKee, R., Sado, T., Tada, T., Meissner, A. & Plath, K. X Chromosome Reactivation Dynamics Reveal Stages of Reprogramming to Pluripotency. *Cell* **159**, 1681–1697 (2014).
12. Guo, L., Lin, L., Wang, X., Gao, M., Cao, S., Mai, Y., Wu, F., Kuang, J., Liu, H., Yang, J., Chu, S., Song, H., Li, D., Liu, Y., Wu, K., Liu, J., Wang, J., Pan, G., Hutchins, A. P., … Chen, J. (2019). Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. *Molecular Cell*, **73**(4), 815–829.e7.
13. 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),
14. Tran, K. A., Pietrzak, S. J., Zaidan, N. Z., Siahpirani, A. F., McCalla, S. G., Zhou, A. S., … Sridharan, R. (2019). Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. *Cell Reports*, **27**(6), 1726–1741.e5.
15. Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., Hecht, J., Filion, G. J., Beato, M., Marti-

Renom, M. A., & Graf, T. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics*, **50**(2), 238–249.
16. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58**, 610–620 (2015).
17. Brambrink, T., Foreman, R., Welstead, G. G., Lengner, C. J., Wernig, M., Suh, H. & Jaenisch, R. Sequential Expression of Pluripotency Markers during Direct Reprogramming of Mouse Somatic Cells. *Cell Stem Cell* **2**, 151–159 (2008).
18. 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., Kim, J.-I., Mar, J. C., Shakiba, N., Ritchie, W., Rasko, J. E. J., Grimmond, S. M., Zandstra, P. W., Wells, C. A., Preiss, T., Seo, J.-S., Heck, A. J. R., Rogers, I. M. & Nagy, A. Genome-wide characterization of the routes to pluripotency. *Nature* **516**, 198–206 (2014).
19. Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H., Beyer, T. A., Datti, A., Woltjen, K., Nagy, A. & Wrana, J. L. Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. *Cell Stem Cell* **7**, 64–77 (2010).
20. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. *Cell Stem Cell* **2**, 230–240 (2008).
21. Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., Plath, K. & Hochedlinger, K. Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution. *Cell Stem Cell* **1**, 55–70 (2007).
22. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. & Trapnell, C. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
23. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol*. **32**, 381–386 (2014).
24. 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., Gallagher, P. G., Krause, D. S., Fan, R. & Lu, J. Nonstochastic Reprogramming from a Privileged Somatic Cell State. *Cell* **156**, 649–662 (2014).
25. Hong, H., Takahashi, K., Ichisaka, T., Aoi, T., Kanagawa, O., Nakagawa, M., Okita, K. & Yamanaka, S. Suppression of induced pluripotent stem cell generation by the p53–p21 pathway. *Nature* **460**, 1132–1135 (2009).
26. Li, H., Collado, M., Villasante, A., Strati, K., Ortega, S., Cañamero, M., Blasco, M. A. & Serrano, M. The Ink4/Arf locus is a barrier for iPS cell reprogramming. *Nature* **460**, 1136–1139 (2009).
27. Ruiz, S., Panopoulos, A. D., Herrerías, A., Bissig, K.-D., Lutz, M., Berggren, W. T., Verma, I. M. & Izpisua Belmonte, J. C. A High Proliferation Rate Is Required for Cell Reprogramming and Maintenance of Human Embryonic Stem Cell Identity. *Curr. Biol*. **21**, 45–52 (2011).
28. Smith, Z. D., Nachman, I., Regev, A. & Meissner, A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nat. Biotechnol.* **28**, 521–526 (2010).
29. 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., Zhang, B., Zeng, L., Lai, L., Esteban, M. A. & Pei, D. A Mesenchymal-to-Epithelial Transition Initiates and

Is Required for the Nuclear Reprogramming of Mouse Fibroblasts. *Cell Stem Cell* **7**, 51–63 (2010).
30. 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. Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. *Nat. Cell Biol*. **15**, 872–882 (2013).
31. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. *Nat. Methods* **7**, 53–55 (2010).
32. Lagasse, E. & Weissman, I. L. bcl-2 inhibits apoptosis of neutrophils but not their engulfment by macrophages. *J. Exp. Med.* **179**, 1047–1052 (1994).

Chapter 3. Genome–wide deconvolution of single cell type chromatin features from single cell gene expression data

Introduction

Understanding the mechanisms of how genes are regulated requires the quantification of the transcriptome and knowledge of how the chromatin is packed and modified. Population-based sequencing assays such as RNA-seq, ATAC-seq and ChIP-seq allow for the profiling of bulk tissues but are insensitive to cellular heterogeneity, such that they return an ensembled profile that smooth over rare or interesting subpopulations, thereby limiting their resolution when considering regulatory diversity underlying heterogeneous cell populations. Although *in vivo* cell types can be flow-sorted and studied, this is labor intensive and requires *a priori* knowledge of sorting markers. Beyond this, there are also practical limitations to using these population-based assays such as the lack of biological material when working with precious samples or the quality of available antibodies for detecting locus-specific enrichments (Kidder et al., 2011).

Recent advancements in the field have allowed for the ability to profile the transcriptome (Klein et al., 2015; Macosko et al., 2015; Vickovic et al., 2016, Zheng et al., 2017) and epigenome (Buenrostro et al., 2015; Cusanovich et al., 2015; Satpathy et al., 2019) at single cell level and in a high-throughput manner. Currently, the standard for single cell methods include single cell RNA sequencing (scRNA-seq) (Zheng et al., 2017) for profiling the transcriptomic landscape, and single cell ATAC sequencing (scATAC-seq) (Satpathy et al., 2019) for measuring chromatin accessibility at single cell level. This has given researchers the ability to unravel a variety of biological contexts including tumor cell heterogeneity (Khoo et al., 2016; Satpathy et al., 2019) and iPSC reprogramming (Buganim et al., 2012; Polo et al., 2012; Schieginger et al. 2019), while also allowing the ability to create cell atlases to characterize a number of cell types and understand cell-to-cell relationships (Cao et al., 2017; Fincher et al., 2018; Han et al., 2018; Karaiskos et al., 2017; The Tabula Muris Consortium et al., 2017). Nevertheless, the output signal from these technologies is sparse and has a narrow dynamic range which is not enough to

accurately describe the activity of regulatory elements within a single cell. For reference, in a typical scATAC-seq dataset, each cell has 10^3–10^5 sequence reads (Zhou et al., 2019). In contrast, the human genome contains 10^6–10^7 cis-regulatory elements (CREs) (Zhou et al., 2019), many of which are constituitively enriched. It has been shown that there are, on average, 4,100 cell type-specific regulatory elements which are likely to be cell-selective regulatory regions (ranging from 1,700 in NHLF to 6,600 in GM12878) (Chen et al., 2013). Therefore, in either case in a typical cell, most of these CREs receive sparse read support.

Recent advancements in the field have allowed for single cell multi-omics profiling unveiling the ability to jointly profile the transcriptiome and epigenome in the same cell. These include, but are not limited to, parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing (scPaired-seq) (Zhu et al., 2019), single cell nucleosome occupancy and methylome-sequencing (scNOMe-seq) (Pott, 2017), and single cell nucleosome, methylation and transcription sequencing (scNMT-seq) (Clark et al., 2018). However, these multi-omics methods have lower throughput to analyze cells and they do not provide the throughput comparable to performing independent scRNA-seq or scATAC-seq to analyze massive numbers of cells. These complex assays are non-standard and require operational knowledge and high cost resulting in discrete and sparse output which limits the widespread promotion of these technologies.

Profiling chromatin that has been post-translationally ultered with histone modifications or remodeled nucleosomes via regulatory proteins provides a means to explore CREs and their underlying regulation. Histone modifications, such as H3 lysine 27 acetylation (H3K27ac) or H3 lysine 4 monomethylation (H3K4me1), allow for the separation of active and poised enhancers (Creyghton et al., 2010). Profiling these histone marks at the single cell level allows for the characterization of cell type-specific enhancer activity from heterogeneous samples. Drop-ChIP, a method that utilize microfluidics, DNA barcoding and next generation sequencing, has the ability

to produce low coverage maps of chromatin state in single cells (Rotem et al., 2015). However, the resulting single cell data are also sparse, capturing on the order of 1000 promoters or enhancers per cell (Rotem et al., 2015). The aggregation of chromatin information within a population of cells into a 'pseudo-bulk' sample, thereby smoothing over the chromatin landscapes of single cell types, is required to make any meaningful interpretations of the data. Taken together, although single cell genomic technologies are rapidly evolving and significant progress has been made in the field to overcome these limitations, accurately measuring the chromatin landscape for single cells remains a challenge.

Previous studies have established computational frameworks to predict gene expression levels from histone modifications (Karlic et al., 2010; Dong et al., 2012) suggesting a relationship between gene expression levels and chromatin features. Building on this, Zhou et al (Zhou et al., 2017) found that chromatin accessibility measured by DNase I hypersensitivity (DH) in a bulk sample can be predicted using the sample's gene expression profile measured by Affymetrix exon array. The study is limited to using Affymetrix exon array data as predictiors, rather than using RNA-seq, which is considered the gold-standard for transcriptomic analysis and offers the ability to measure the transcriptome in small cell numbers. A follow-up study from the same group showed that various chromatin profiling technologies can be predicted by scRNA-seq (Zhou et al., 2019) using a reference atlas of cell types but prediction accuracy is largely based on the similarity between new samples and cell types within the atlas. When a new RNA-seq sample represents a unique new cell type considerably different from all cell types within the atlas data, then the prediction accuracy drops due to instability of extrapolation (Zhou et al., 2019). Predictions are also limited to only genomic regions with DH enrichments found within the atlas cell types. Another study takes a similar approach to deconvolve bulk samples into subpopulation-specific data through a series of linear convolutions. Their method, DC3 (De-Convolution and Coupled-Clustering) (Zeng et al., 2019) takes a simplified approach and models a cost function as a linear relationship between the enhancer-promoter interaction strength from single cell Hi-C

(Kim et al., 2019), gene expression from scRNA-seq, and enhancter openness from scATAC-seq to deconvolve bulk signal (i.e., bulk sample loop counts) into subpopulation-specific signals. DC3 requires, and relys heavily on, the input from these three sets of data. Furthermore, scHi-C (Kim et al., 2019) is used to estimate the regulatory potential between active regulatory elements and target genes in single cells, but the assay itself has not yet been established in the field (unpublished), and lacks the ability to capture a set of genomic interactions that distinguish between cell types. Additionally, the course scale (500kb bin size) is not fine enough to accurately link regulatory elements with target genes. Taken together, while existing scRNA-seq and ChIP-seq technologies are readily available, there does not exist a framework for accurately deconvolving ChIP-seq profiles at the single cell type level using scRNA-seq alone.

Here we investigate the feasibility of predicting the genome-wide chromatin maps at the single cell type level using a reference atlas of 146 purified cell types with matched population-level RNA-seq and ChIP-seq data types. Using this atlas, we take a three-step modeling approach to accurately deconvolve bulk chromatin signal into subpopulation-level signal: (1) across-cell type modeling to learn the relationship between the transcriptome and epigenome across many cell types, (2) within-cell type modeling to leverage cell type-specific features, and (3) the integration of these techniques into an accurately predicted and unified signal intensity track. In this work, we deconvolve histone modification mark H3K27ac from ChIP-seq of a fetal brain organoid that was characterized into a variety to neuronal cell types by scRNA-seq. Doing so allows for the classification of cell type-specific regulatory elements, such as enhancers, and investigation of cell type-to-cell type variability of different regulatory elements. We validate our deconvolution accuracy by analyzing our predicted tracks for H3K27ac enrichment and relate these regions to cell type-specific motifs and transcription factor expression. We formalize this framework into a novel computational method called DeconR and can be used for the deconvolution of individual cell type ChIP-seq profiles from population-level ChIP-seq and single

cell RNA-seq data from the same sample or tissue (R package; https://github.com/ShanSabri/deconR).

Results.

The DeconR algorithm

We formulate the deconvolution of population-level ChIP-seq data into its underlying cell types by profiling single cell RNA-seq on the same sample or tissue. For each cell type in the single cell data, the goal is to predict genome-wide chromatin signal intensity maps that allow us to define cell type-specific chromatin features. DeconR achieves accurate deconvolution by leveraging a cell type reference atlas that consists of 146 distinct cell types with matched bulk RNA-seq and ChIP-seq data types (Figure 3-1, Supp Figure S3-1).

As a first step, the input scRNA-seq data must be imputed for "dropout" artifacts, or the excess of zero counts due to the low amounts of mRNA sequenced within individual cells. We leverage the tool scImpute (Li and Li, 2018) to accurately identify likely dropout events and perform imputation without introducing new bias to the rest of the data. As a next step, the imputed scRNA-seq data must be clustered to define unique cell types. We utilize Seurat's workflow (Butler et al., 2018) for data normalization, unsupervised graph-based clustering, followed by dimentionality reduction by Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to achieve this. To validate that our scRNA-seq clusters correspond to unique cell types, we measure and visualize the expression of canonical marker genes that are established in literature with the expectation that these markers are well-defined in the data. For single cell types that are not well characterized in literature, we perform differential gene expression analysis through Seurat (Butler et al., 2018) to identify marker genes enriched in the single cell type of interest and link these markers to Gene Ontology categories through Metascape (Zhou et al., 2019).

Once single cell types have been established in the scRNA-seq data, then we aggregrate the gene expression measurements for all cells contained within a given single cell type. The

creation of “pseudo-bulk” samples are needed in order to overcome the sequencing depth bias between scRNA-seq data and bulk RNA-seq data contained within the reference atlas. The use of dropout imputation (described above) also aids in correcting for this effect. The “pseudo-bulk” samples derived from the scRNA-seq data are used as input to DeconR.

DeconR leverages two modeling techniques, within-cell type and across-cell type, to accurately predict the chromatin landscape for populations of single cells derived from scRNA-seq data (Figure 3-1). The across-cell type modeling technique is used to learn the relationship across cell type within the atlas. This approach consists of modeling a regression to predict the signal intensity for a genomic locus given gene expression features across training cell types in the reference atlas. However, modeling each locus independently requires one to deal with a challenging big data regression problem which involves fitting a model for each data point (200bp bin). This equates to training 15,181,508 regression models on the human genome, each with a large number of predictors (18,436 expressed genes). To cope with the high dimensionality and computational intensity, we cluster our genomic loci (200bp bins) across all training cell types into 2000 clusters, hereon referred to as loci groups, then model each loci group independently. Rather than predicting the signal intensity at a given locus, we are now predicting the average signal intensity for a group of genomic loci. Clustering is performed using a big-data optimized version of K-means clustering (see Methods). Loci groups are not bound to spatial restrictions but rather assembled by accessing signal intensity trends across all training cell types. For each loci group, we compute an average signal intensity value at the cell type level. This reduced the number of models needed to train from >15,000,000 to 2,000, and as this loci group parameter is increased, the across-cell type model will approach a bin-by-bin fit. We show that as we increase this loci group parameter, the model accuracy plateaus indicating that there is not much to improve by stratifying into more groups thereby increasing the runtime and complexity (Supp Figure S3-2). For each grouping of genomic loci, we utilize all expressed genes (features) in a K-nearest-neighbor regression framework to predict the average signal intensity value (response).

Using all expressed genes as features is reasonable as many regulatory elements are known to control genes over a long genomic distance and sometimes across many other genes. We have explored popular regression techniques such as Lasso (Tibshirani, 1996) and Elastic-net (Zou and Hastie, 2005) regression but found runtime to be significantly slower with no accuracy gain. Once genomic loci group-level models are trained, group-level predictions can be made and the predicted average signal intensity values are then substituted into each locus within the respective group.

In addition to the across-cell types modeling, we establish a within-cell type modeling technique to capture cell-type specific regions that are typically not bound to a single loci group. The within-cell type modeling techniques holds the general assumption that the regulatory potential of a gene increases as a function of how close the gene is to a genomic regulatory element. Therefore, in this modeling technique we do not account for the entire transcriptome but rather only the gene neighborhood for which the regulatory element resides. To formulate this modeling technique as a regression, we take each genomic loci's signal intensity value (response) and model it as a function of its nearby genes' expression values, respective linear genomic distance to TSSs, and a binary encoded expression indicator to denote if a gene is expressed (features) for the entire genome (see Methods). In order to avoid a large feature space, we consider the five closest genes (15 total features for 5 of the closest genes: 5 gene expression values, 5 genomic distances in bp, 5 binary expression encodings) to each genomic locus (Supp Figure S3-2). We show that accounting for fewer genes will not provide as accurate predictions and including many genes has little accuracy gain with increased model complexity. For each cell type in the reference atlas, we split its genome into training and test sets by stratifying chromosomes into even number and odd number partitions. We train a genome-wide model to predict single intensity on the training set of chromosomes using the features described above, then evaluate the model accuracy on the held-out test set of chromosomes. The output prediction from each cell type-specific model is then averaged together to generate a robust prediction.

The modeling techniques described above are integrated together using Ridge linear regression to produce the final signal track prediction. Here, we include a feature that is predictive of cell type specificity such as the number of cell types within the reference atlas containing a peak at a given locus, in addition to the across- and within-cell type predictions. For a random, held-out set of 20 cell types from our reference atlas, we show this method of integration matches (3/20 cell types) or outperforms (17/20 cell types) a genome-wide static proportional split between both modeling techniques (Supp Figure S3-3). Interestingly, we see the highest accuracy gain for using this integration technique over an across- or within-cell type only model (ACROSS_100_WITHIN_0 or ACROSS_0_WITHIN_100) in the BSS00333 cell type, although overall predictive accurately for this cell type is poor and warrants further investigation. Integration models are trained on a cell type specific basis similar to the within-cell type training process and predictions are averged together. This model integration step may result in a better tradeoff between the prediction bias since the across-cell type modeling leverages information across all training cell types in the reference atlas and the within-cell type modeling captures cell type specific loci more accurately (Supp Figure S3-4B, Supp Figure S3-5).

Lastly, the integrated prediction single cell type tracks are used to de-convolve the bulk/convolved signal track into its underlying cell type fractions. For a given locus, we proportionally scale each predicted signal intensity value across all predicted single cell type tracks to a relative proportion that sums to the signal intensity value of the convolved signal track. The deconvolved tracks can be used downstream analyses, such as for the classification of cell type-specific regulatory elements.

Predicting and validating genome-wide H3K27ac ChIP-seq signal intensity from Within- and Across-cell type modeling techniques

We use matched RNA-seq and H3k27ac ChIP-seq samples generated Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium) and Roadmap Epigenomics (Bernstein et

al., 2010) consortiums that covered a wide variety (n=146) of human cell types to train models to predict the genome-wide chromatin signal intensity from gene expression features (Supp Figure S3-1, Table 3-1). These data for all cell types within the reference atlas are processed uniformly (see Methods) such that genes expression values are size-factor normalized and signal intensity profiles are processed into 200bp non-overlapping bins and normalized to fragments per kilobase of exon model per million reads mapped (FPKM). We show that the relationship between the H3K27ac signal intensity around gene transcriptional start sites (TSSs +/-500bp) is strongly correlated (average Spearman R = 0.66) with gene expression (Supp Figure S3-6) consistent with previous claims that gene expression provides valuable information for predicting chromatin signal intensity (Karlic et al., 2010; Dong et al., 2012, Zhou et al., 2017; Zhou et al., 2019; Zeng et al., 2019). We leverage this relationship to train models to learn the associations between gene expression features and signal intensity at a given locus.

Using a random set of held-out cell types from the reference atlas, we use our within- and across-cell type models to predict H3K27ac signal intensity tracks for a held out set cell types in the reference atlas. With these tracks, we are able to call peaks and annotate genomic regions of H3K27ac enrichment (see Methods). Doing so allows us to compare regulatory elements in the ground truth signal track with those in the predicted signal tracks. To evaluate the performance of our predictions, we annotate the patterns of peak overlap as a binary (0=no peak, 1=peak present) three digit encoding, where the indices correspond to the across-cell type peaks, within-cell type peaks, and true peaks, respectively. For example, an encoding of 101 corresponds to the fraction of peak overlap between the across-cell type prediction and the ground truth track. By analyzing the data in this way, we are able to define 7 distinct patterns (011, 110, 010, 100, 101, 001, 111) of agreement in peak calls between the two modeling techniques and the ground truth. Each modeling technique captures slightly different information but the vast majority of peaks are shared among the across, within and ground truth tracks (Supp Figure S3-4A). In the case of cell type BSS01849, a transverse colon cell type, we see that over half (pattern 111, 55.63%) of all

called peaks match between the within- and across-cell type model predictions and ground truth. Interestingly, we see that 18.97% of peaks in this cell type are unique to the across-cell type prediction and ground truth (pattern 101) and that 1.19% of peaks are unique to the within-cell type prediction and ground truth (pattern 011). Downstream analyses of 011 peaks in the BSS00476 (GM23248, Primary Fibroblast) cell type conclude their enrichment over cell type-specific genes (Supp Figure S3-4B). In this example, the within-cell type model is able to capture peaks over ICAM1, SBSN, DKMN, among other genes that are known to be upregulated in inflammatory diseases and expressed in the differentiated layers of skin (Piela-Smith et al., 1992; Horikoshi et al., 1995; Hasegawa et al., 2012).

It is also worthy to note that predictive accuracy is a function of the strength of the relationship between the transcriptome and epigenome. Of the cell types in Figure S3-4A, BSS00333, a stromal fibroblast cell type, has the largest proportion of peaks (45.27%) that are not found in the respective predictive tracks (pattern 001). This cell type also contains the weakest relationship (Spearman R = 0.2541) between its gene expression and H3K27ac signal intensity distribution around TSSs, and has very poor predictive accuracy (Supp Figure S3-3, Supp Figure S3-4A, Supp Figure S3-6).

H3K27ac deconvolution of a fetal brain organoid using scRNA-seq expression features

We next asked whether one can use the DeconR framework to deconvolve a bulk heterogenous sample into its underlying cell types defined by scRNA-seq on the same sample. As a proof of concept, we preform single cell RNA-seq on a fusion of fetal brain organoids from the H9 cell line harvested at d149, d157 and d161 to demonstrate DeconR's deconvolution performance (Figure 3-2A). The protocol used to generate these organoids are well-established and result in organoids containing known cell types previously characterized (Watanabe et al., 2017; Samarasinghe et al., 2019). After scRNA-seq and downstream filtering, preprocessing and normalization, 7,287 single cell transcriptomes (median UMI: 5,151, median genes: 2,596) with expression

measurements from 24,324 genes were obtained for downstream modeling. Through clustering and differential gene expression analyses, we are able to identity six unique neuronal-related cell types: maturing neurons (Stmn2+, Neurod6+), radial glia (Pax6+/Hopx+), interneurons (Lhx1+/Gabrg1+), intermediate progenitors (Eomes+), nigral neurons (Cebpb+/Vgf+/Hexim1+), and early astrocyte/glial cells (Igfbp7+/Ttr+/Cxcl14+) (Figure 3-2A). For each one of these single cell types, our goal is to predict the genome-wide H3K27ac signal intensity profile using the modeling approach previously described (Figure 3-1). We focus on chromatin modification H3K27ac because profiling this mark at single cell level is difficult and inefficient, and will shed light to the enhancer landscape for the underlying cell types without the need for additional assays. In order to systematically evaluate the performance of the predictions we perform motif and transcription factor binding analysis on the predictive output tracks and relate these data to the defined cell types supported by scRNA-seq.

To overcome the inherent problem of sparsity and to accurately model the single cell data from a reference atlas of population-level cell types, we aggregate cells within each cell type create `pseudo-bulk` samples prior to normalization. To this point, we also apply scImpute (Li and Li, 2018) to impute the gene expression dropout prior to the creation of pseudo-bulk samples. DeconR was then applied to the imputed pseudo-bulk gene expression values to make signal intensity predictions at the single cell type level. The output prediction tracks from the across- and within-cell type modeling techniques are integrated together to form final predictions. The predicted single cell type tracks are then used to deconvolve the bulk chromain track.

To evaluate the performance of the predictive signal intensity tracks we performed peak calling through MACS2 (Zhang et al., 2008; Feng et al., 2012) to identity areas in the genome that have been enriched for H3K27ac histone modifications, likely targeting enhancers and proximal and distal regions of TSSs. Constitutive peaks, or peaks that are enriched across all six single cell type prediction tracks, make up the majority of peak overlaps (n = 17,406) (Figure 3-2B).

However there exists a large set of peaks (n = 4,168) that are unique to the interneuron, maturing neuron, intermediate progenitor and nigral neuron cell types, which agrees with the hierarchal clustering clade of the correlation heatmap between these cell types, and the localized embedding of these single cell types on UMAP, indicating that the similarity in peak overlap is reflected by the similarity in gene expression for which the predictions are based. This suggests that DeconR is not introducing new bias and that the predicted signal intensity tracks are inline with the gene expression relationships between these single cell types. Interestingly, we notice the early astrocyte cell type being isolated from the other cell types on the UMAP embedding of gene expression, while also identified as the outgroup from the pseudo-bulk gene expression correlations, indicating that this cell type is most dissimilar from others in both the gene expression and chromain landscapes. In agreement with this, we show that there exists a large set of peaks (n = 8,098) unique to the early astrocyte cell type suggesting that its unique nature of gene expression agrees with its unique chromatin landscape.

Using these peak calls, we performed regulatory element analysis through HOMER (Heinz et al., 2010) to discovery motif enrichments. The top significant motifs that overlap with corresponding differential genes for each cell type is displayed as a heatmap (Figure 3-2C). Our results show that enriched motifs agree with the expression of their corresponding transcription factors in each cell type. For example, nigral neurons have previously been characterized by their upregulation of Cebpb (Hu, 2011). In addition to Cebpb being most highly expressed in the pseudo-bulk nigral neuron cell type, the enriched peaks within this cell type contain Cebpb binding motifs. To this point, our results also show that the Smad3 motif and gene expression is most highly enriched in the early astrocyte cell type peaks. This data agrees with been previously shown data that Smad3 is a master regulator for early glial and astrocyte cell types (Stipursky and Gomes, 2007; Hamby et al., 2010; Stipursky et al., 2012). In addition, we show that Nanog gene expression and motifs are enriched within the neural progenitor cell type, which is in line with previous studies showing Nanog expression, in synchrony with WNT signaling, regulates

neural patterning, a process during neurogenesis which neural progenitor cells differentiate into neurons with distinct functions (Su et al., 2018). Our results also show an abundance of Sox gene family expression and motifs in the radial glia cell type. It has previously been shown that Sox2 expression levels alone can distinguish radial glia from intermediate neural progenitors (Hutton and Pevny, 2011). It has also been shown that Cux2 is a master regulator for the devleiopment of neural progenitors (Lulianella et al., 2008) and that Cux1/2 are expressed in the developing brain, particularly in subventricular zone (SVZ) cells and their maturing layers (Cubelos et al., 2007). These findings are in line with expression and motifs enriched within the maturing neurons cell type. Collectively, the analyses in this section show that predicting chromatin modification H3K27ac signal intensity tracks using scRNA-seq data is feasible. The prediction accuracy based on motif enrichment and RNA expression is highly consistent with previously published reports.

Discussion.

This study examines the feasibility of deconvolving the chromatin landscape for cell types defined by single cell RNA sequencing by employing machine learning techniques. We formalize this framework into a user-friendly R package called DeconR (https://github.com/ShanSabri/deconR). DeconR is a novel computational method that can be used to unravel gene expression information to study the chromatin landscape at the single cell level. DeconR predictions can be used as a first pass measure to provide insights to hypothesis-driven questions or the design of follow up experiments without the need for an experimental assay, thereby forgoing the limitations of many single cell chromatin assays. Although the deconvolution made in this study is based on chromatin modification H3K27ac, DeconR is a general framework that leverages the relationship between the transcriptome and epigenome and can be applied to other data types for which a reference atlas can be created, such as DNase-seq or ATAC-seq data types with matched RNA-seq.

In our analyses, we show the accurate deconvolution of six distinct cell types within a human fetal brain organoid sample by predicting the chromatin landscape, specifically targeting ChIP-seq of H3K27ac chromatin modification, for these populations and relating regions enriched for H3K27ac chromatin modifications to cell type-specific bound transcription factors. Our results provide correlative insight into the underlying epigenetic heterogeneity without the need for additional experiments. DeconR utilizes the associations across an atlas of cell types to make chromatin prediction, but unlike previous studies, it does not rely solely on this. DeconR's within-cell type modeling technique is able to accurately predict regions that are cell type-specific and ultimately allows for the prediction and characterization of novel cell types that may not be found in a reference atlas. Though as more training data becomes available, one can create a diverse atlas that will cover a variety of input tissues and cell types.

Conventionally, single cell gene expression measurements are collected to explore the transcriptome by characterizing heterogeneous cell types with underlying gene signatures, pathways, and ontologies. We hope that DeconR can be used to add a new component of this

workflow by unleashing insights into the epigenetic landscape by using gene expression. DeconR can be readily applied to a number of gene expression studies and impact how gene expression is data is used.

Future Directions.

In this study we describe a method of linking the epigenome to transcriptomic features in order to learn the relationship between these data. In doing so, we measure distance as a linear function along the genome, though in practice the genome is compacted into three-dimensional confirmation that are able to link promoters to distal enhancers. Next-generation chromosome capture technologies, such as Hi-C (Belton et al., 2012) or scHI-C (Kim et al., 2019), are able to measure the linkage of enhancer elements to promoters in a more accurate manner. Therefore, like previous studies, we believe that the integration of Hi-C data may be useful in tuning the distance feature of the within-cell type model to provide for more accurate linking. Though with this comes additional data dependencies for non-standard assays that may not be readily available.

In this work we focus on the deconvolution of ChIP-seq of the H3K27ac histone modification but we acknowledge that this framework is generalizable and can be applied to other histone modifications or assays, such as ATAC-seq or DNase-seq, to deconvolve bulk samples into their underlying heterogenous sub-samples. This study may be extended in the near future to overcome these limitations.

Figure Legends

Figure 3-1. Schematic diagram of DeconR. Outline of the method for which a bulk chromatin signal intensity track is deconvolved into its underlying cell type fractions based on the subpopulations/clusters defined by scRNA-seq. DeconR utilizes two methods of modeling chromatin signal as a function of gene expression, both leverage an atlas of purified cell types with matched RNA-seq and chromatin data. In the across-cell type model, loci from the training set within the atlas are grouped based on signal intensity similarities and, for each group, a model is trained to predict the average signal intensity value based on a feature space containing all expressed genes. The within-cell type model leverages cell type-specific model to predict signal intensity of a locus given features that characterize that locus, such as the expression of nearby genes and their respective distance to their TSSs. Both of these modeling techniques are locally integrated together to form a final deconvolved track.

Figure 3-2. Deconvolution of fetal brain organoids into 6 distinct neuronal subtypes. (A) tSNE embedding of single cell transcriptomes from a fetal brain organoid identifies six distinct cell types (maturing neurons, radial glia, interneurons, intermediate progenitors, nigral neurons, and early astrocyte/glial cells) though differential gene expression analysis. Highly differential genes, shown as a heatmap, are used to annotate cell types through expert knowledge and literature. For each of these defined cell types, DeconR computes pseudo-bulk aggregates at the cell type level and predicts the H3K27ac ChIP-seq landscape as shown as signal intensity tracks within a genome browser. The output cell type-specific tracks contain constitutive and differential peaks that can be used for downstream analyses such as motif and TF binding enrichment, as well as to gain a deeper understanding of their functional dynamics and relationships. (B) For each predicted track, we call peaks and access their similarity of overlap using an upset plot. The most common set of peaks (n = 17,406) are shared among all cell types. The second most common set (n = 8,098) is unique to the early astrocyte cell type which agrees with the inset heatmap, as it is the outlier

when considering the Spearman correlation coefficient between the transcriptome of the pseudo-bulk aggregates, indicating the most dissimilar of the six defined cell types in (A). (C) For peak enriched in each cell type, motif analysis was preformed and the normalized expression of the TFs corresponding to these motifs are shown as a heatmap. The heatmap is column z-scaled on normalized expression values.

Figure S3-1. Spearman correlation heatmap of reference atlas containing 146 cell types profiling the RNA-seq and ChIP-seq landscapes. Spearman correlation heatmaps of (Top) the gene expression landscape containing all expressed genes and (Bottom) the genome-wide chromatin landscape between all 146 cell types in the reference atlas.

Figure S3-2. Selecting model hyperparameters as a function of accuracy. Box plots illustrating the distribution of RMSE for a random set of 30 cell types from the reference atlas as a function of (Top) across-cell type modeling loci-group stratification and (Bottom) within-cell type modeling of the number of nearest genes to each loci bin. By default, DeconR will set these hyperparameters to 2000 loci groups and five nearest genes for across- and within-cell type modeling techniques, respectively.

Figure S3-3. The integration of across- and within-cell type models improves the overall prediction accuracy. Heatmap showing the Pearson correlation matrix of the agreement between a variety of integration techniques with ground truth tracks (y-axis) for 20 randomly held out cell types from the reference atlas (x-axis). Integration methods include using a static grid (step side of 10%) to proportionally allocated each modeling technique genome-wide. For example, integration technique ACROSS_60_WITHIN_40 corresponds to utilizing 60% of the across-cell type prediction with 40% of the within-cell type prediction genome-wide. The TRUTH ~ ACROSS + WITHIN integration models the true signal intensity as a function of the across and within

predictions, estimating genome-wide coefficients for integration. The TRUTH ~ ACROSS + WITHIN * NUM_CELLTYPES integration technique includes an interaction terms that denotes the number of cell types within the reference atlas that contain a peak at a given locus. We benchmark these grid integration methods with using regression-based approaches and show that, in all 20 cell types, the integration which utilizes the interaction term matches (3/20 cell types) or out-performs (17/20 cell types) the ACROSS_100_WITHIN_0 integration.

Figure S3-4. Majority of predicted peaks overlap with ground truth peaks and across-cell type modeling captures proportionally more peaks in the ground truth than the within-cell type model. (A) Stacked bar graphs showing the proportion of six patterns of peak overlaps between across- and within-cell type modeling techniques with ground truth peak calls for a random set of 20 held out cell types from the reference atlas. (B) Genome browser view of the across-, within-cell type predictions, and ground truth tracks for a primary fibroblast cell type showing example cases of peak pattern 011 over cell type-specific genes.

Figure S3-5 – Within-cell type modeling more accurately predicts cell type-specific regions than across-cell type modeling. (A) Density distribution for the number of cell types within the reference atlas containing a peak. The non-overlapping bimodality implies there are many peaks that are either cell type-specific (found in <10 cell types within the reference atlas) or constitutive (found in > 130 cell types within the reference atlas). (B) Boxplots measuring within- and across-cell type model accuracy, for a randomly held-out set of 20 cell types form the reference atlas, as a function of peak cell type specificity, or the number of cell types in the reference atlas containing a peak. Smoothed regression line overlays illistrate the overall trend of RMSE from cell type specific peaks to constitutive peaks. (C) As in (B) but showing the smoothed regression lines as a function of the cumulative fraction of all peaks. Here, we see the substantial improvement in model

accuracy for within-cell type modeling of cell type-specific peak, and across-cell type modeling for constitutive peaks.

Figure S3-6. Spearman distribution of gene expression with average H3K27ac signal intensity around gene TSSs (+/- 500bp). (Top) Scatter plot of Pearson and Spearman correlation metrics measuring the agreement between normalized gene expression and average signal intensity around corresponding TSS (+/- 500bp) windows for each cell type in the reference atlas. Cell type BSS00333 is labeled as an outlier with both correlation metrics. (Bottom) Density distribution of the Spearman correlation agreement from (Top).

Table 3-1. Metadata of EpiMap cell types with matched H3K27ac ChIP-seq and RNA-seq data in reference atlas. A table containing metadata for the 146 cell types within the reference atlas. Metadata fields include: ID, SampleName, Tissue, Project Source, Sex, Age and tissue/group classifications.

Figures

Figure 3-1 – Schematic diagram of DeconR

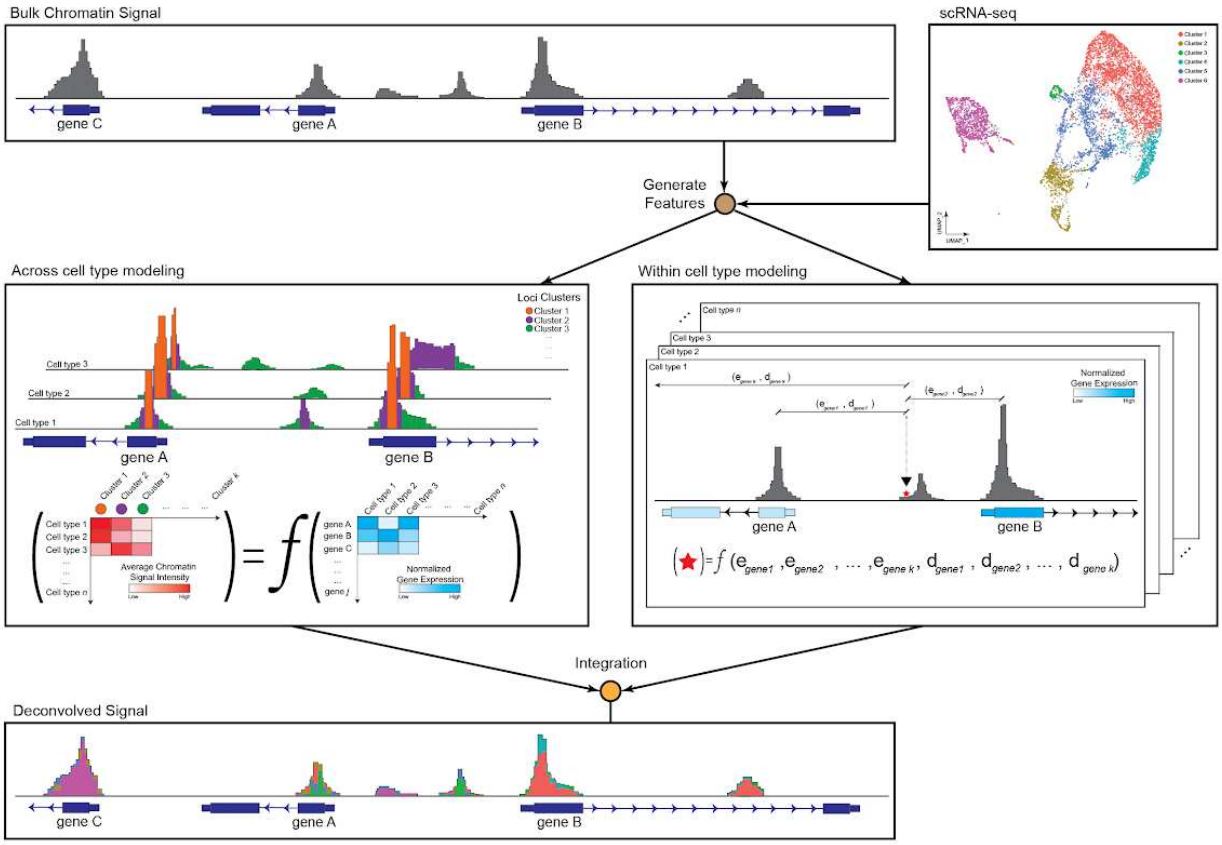


Figure 3-2 – Deconvolution of fetal brain organoids into 6 distinct neuronal subtypes

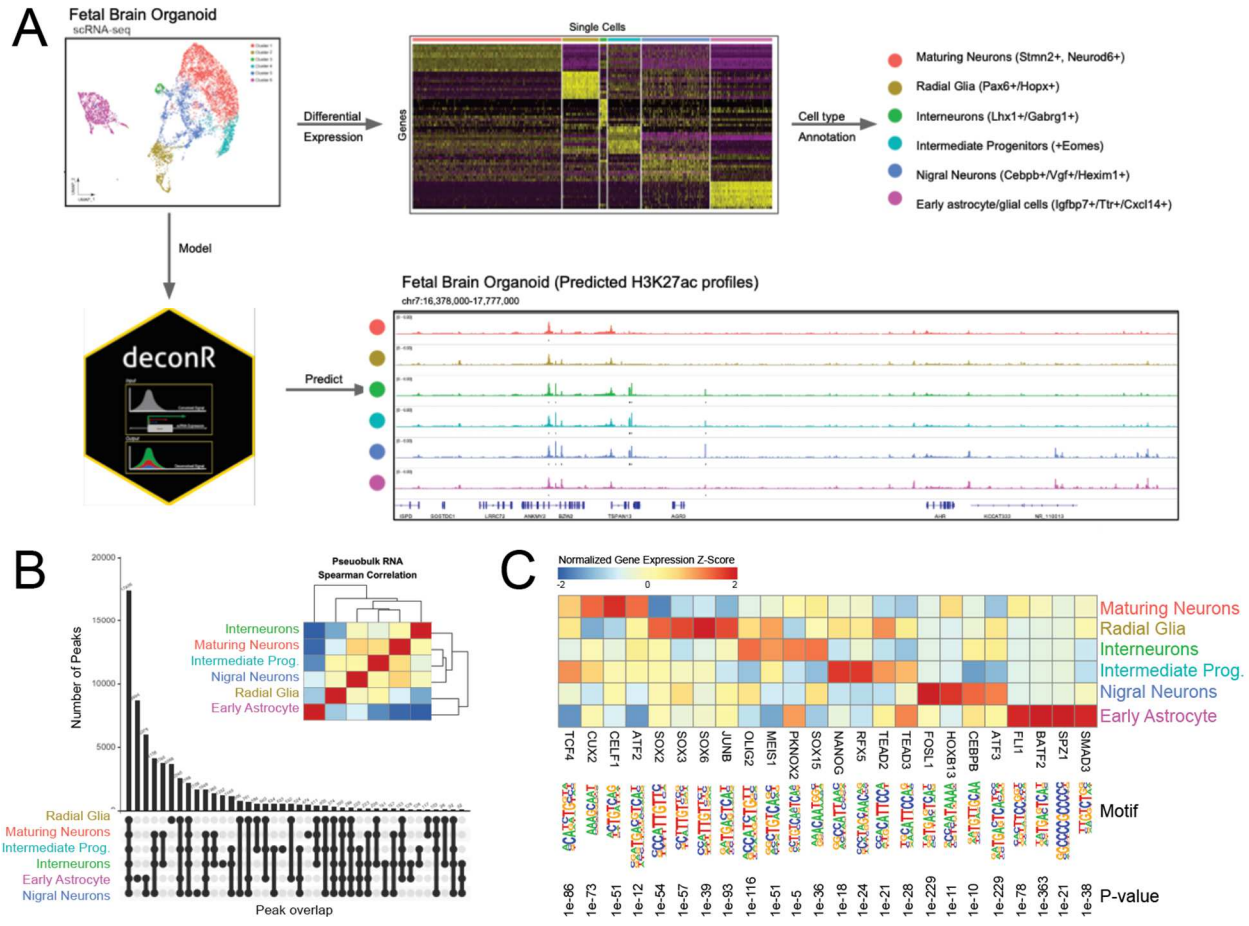


Figure S3-1 – Spearman correlation heatmap of reference atlas containing 146 cell types profiling the RNA-seq and ChIP-seq landscapes

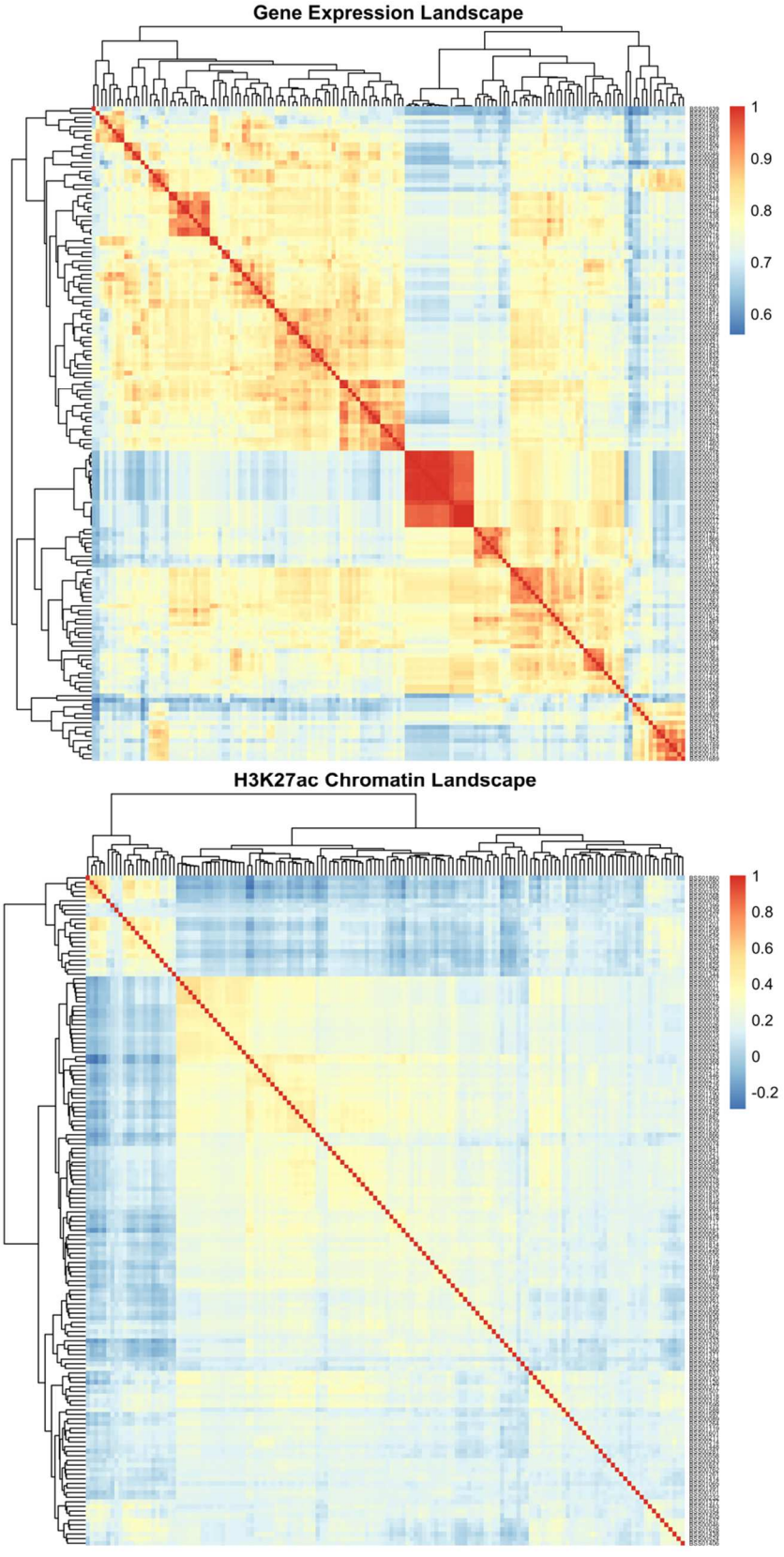


Figure S3-2 – Selecting model hyperparameters as a function of accuracy

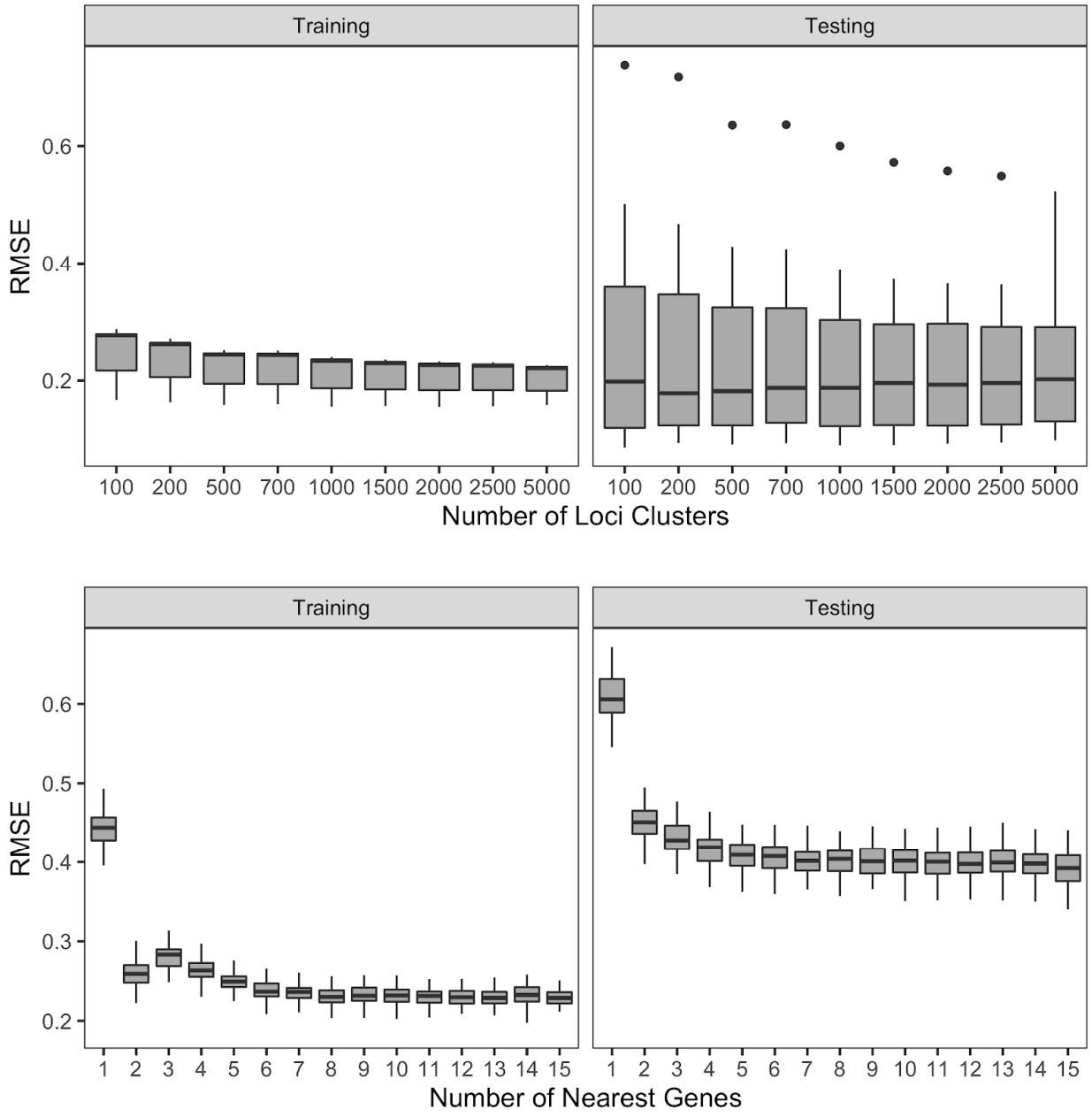


Figure S3-3 – The integration of across- and within-cell type models improves the overall prediction accuracy

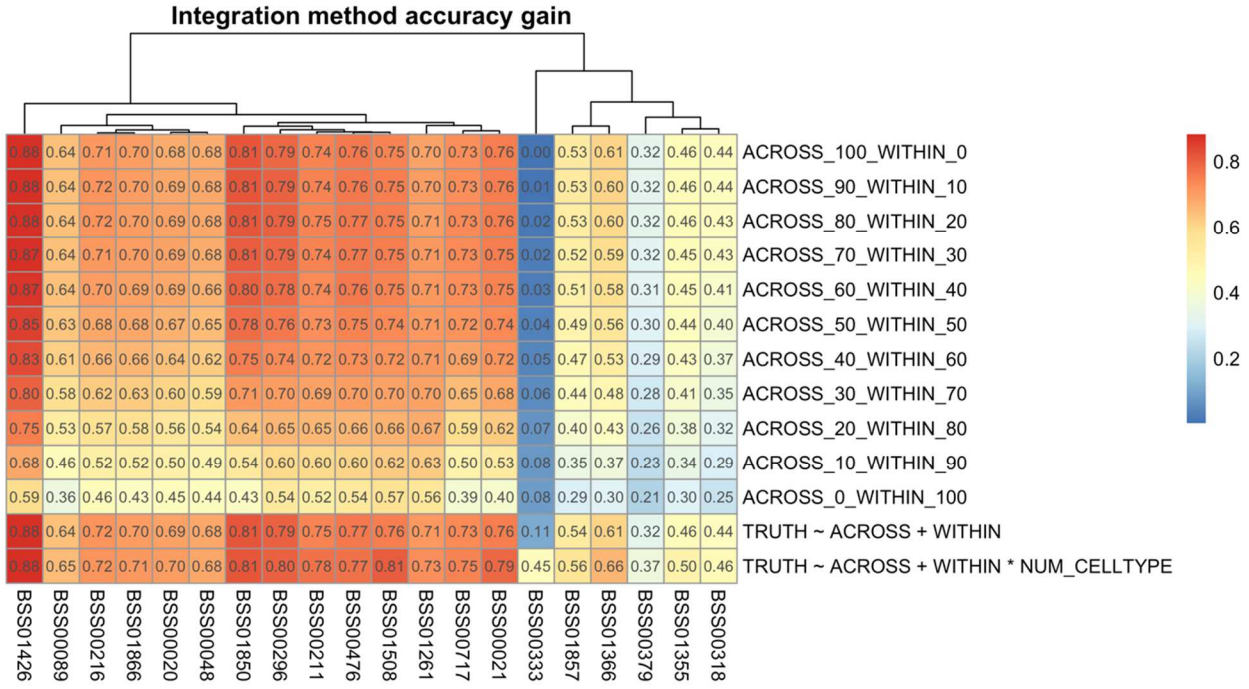


Figure S3-4 – Majority of predicted peaks overlap with ground truth peaks and across-cell type modeling captures proportionally more peaks in the ground truth than the within-cell type model

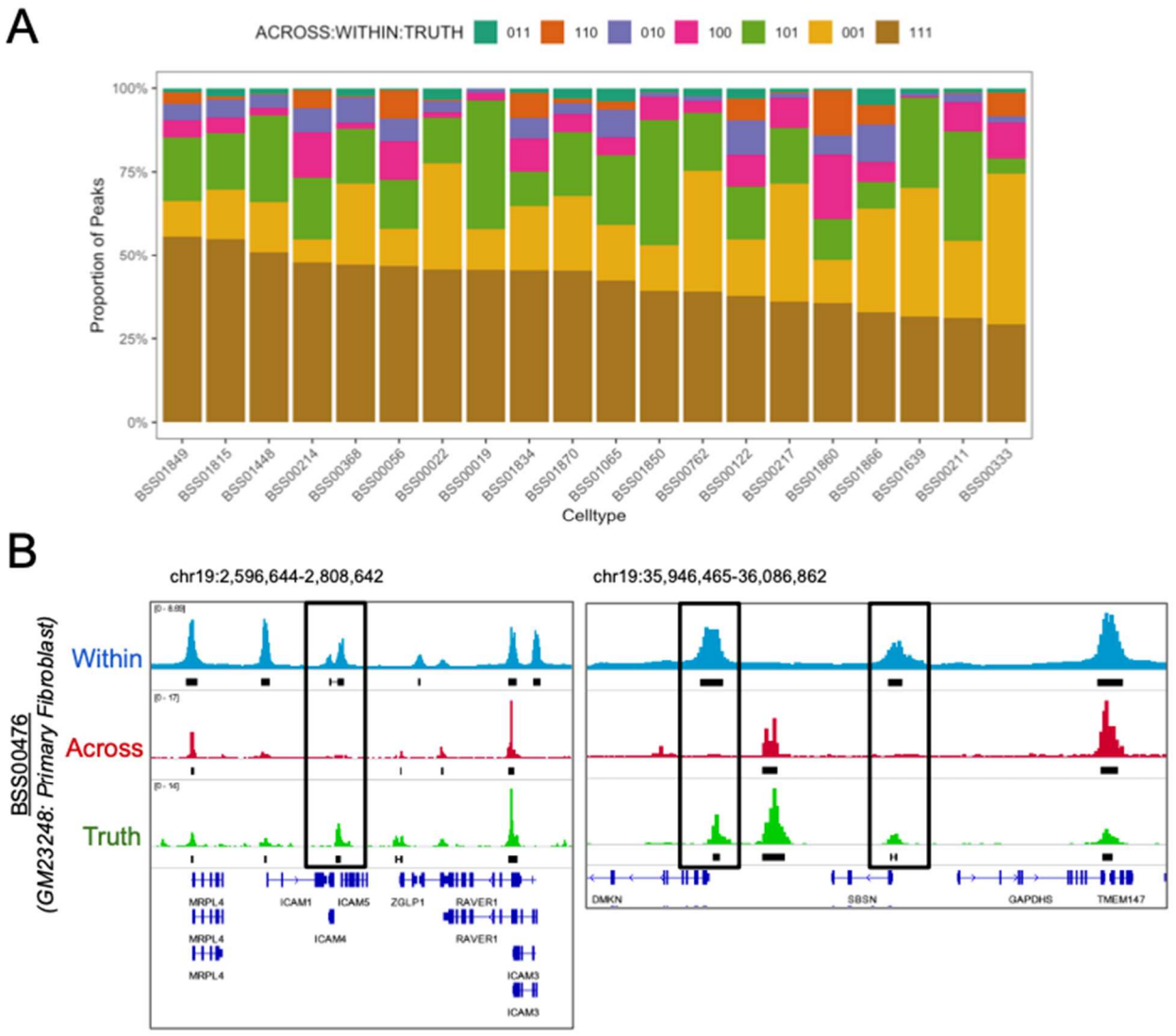


Figure S3-5 – Within-cell type modeling more accurately predicts cell type-specific regions than across-cell type modeling

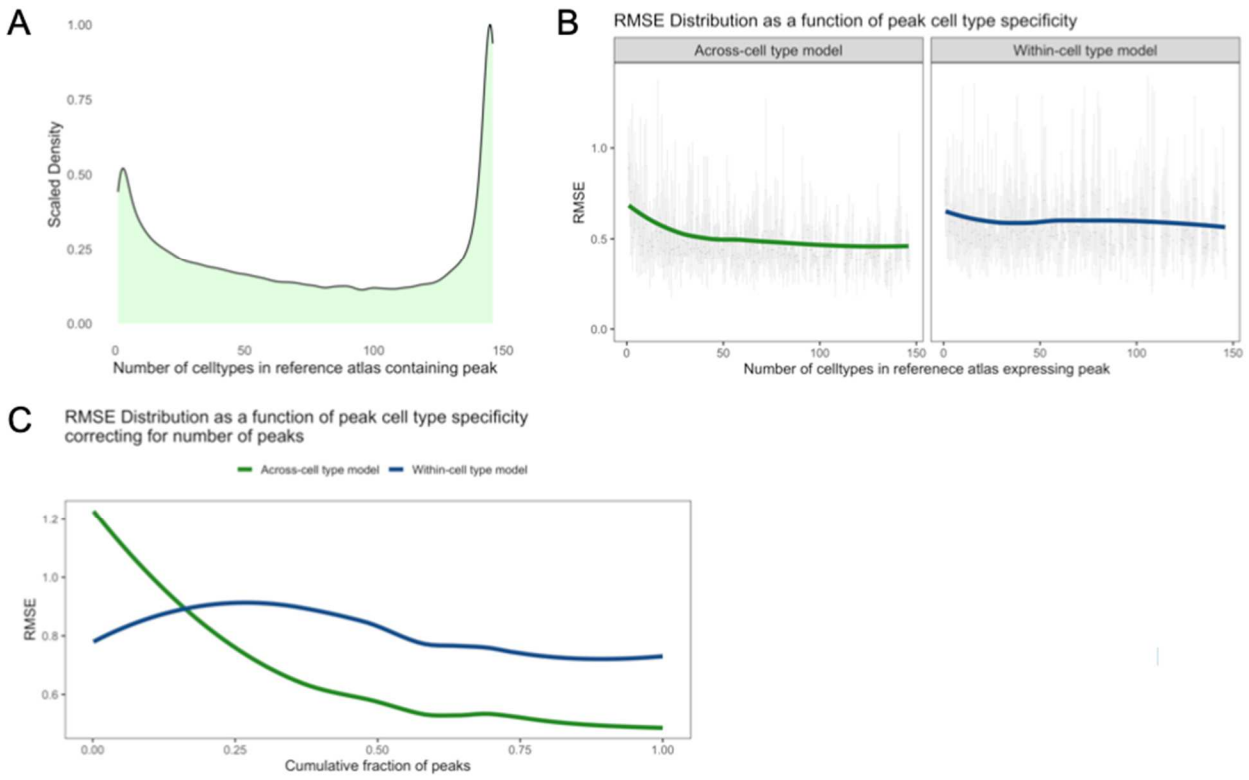


Figure S3-6 – Spearman distribution of gene expression with average H3K27ac signal intensity around gene TSSs (+/- 500bp)

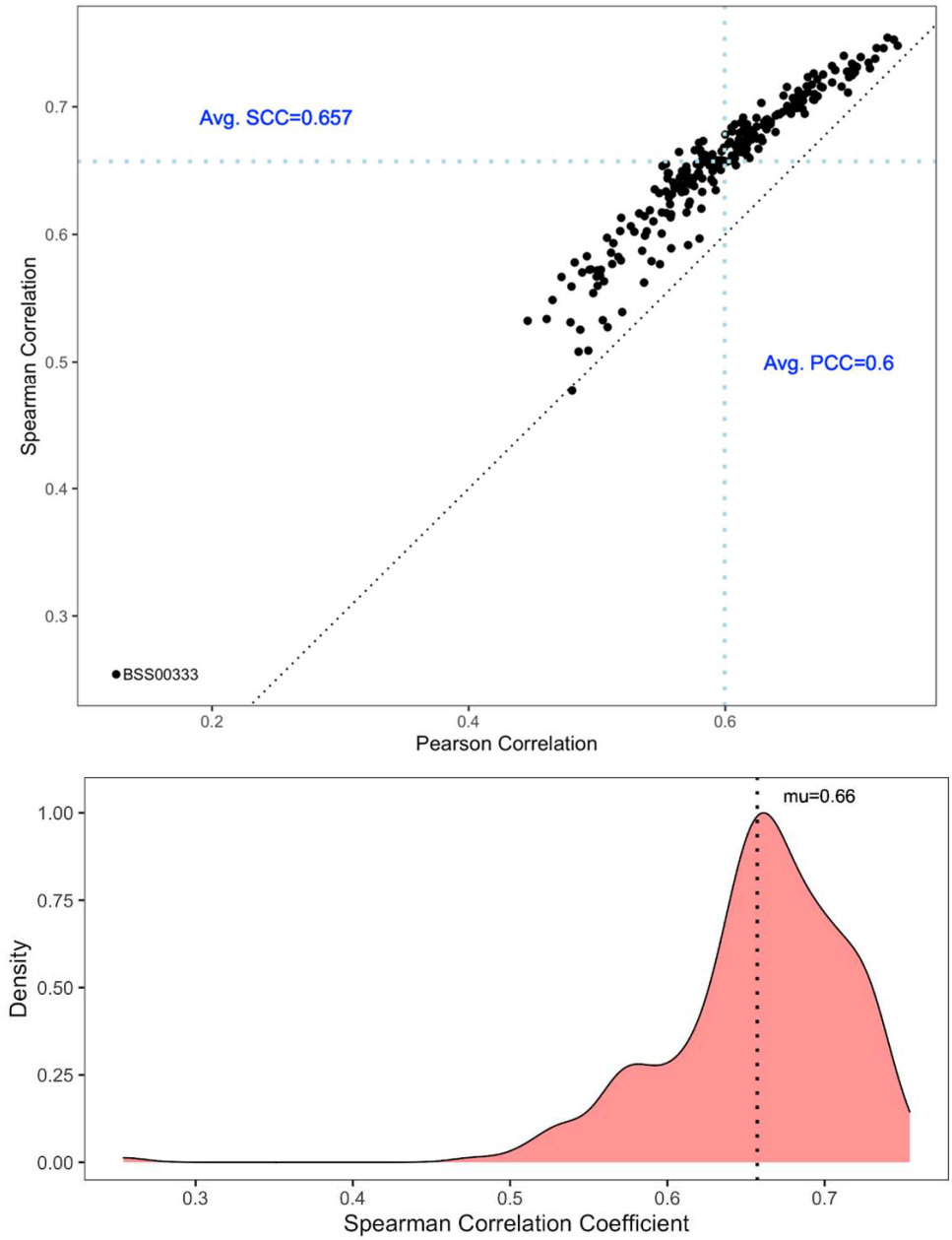


Table 3-1 – Metadata of EpiMap cell types with matched H3K27ac ChIP-seq and RNA-seq data in reference atlas

BSSID	Group	Secondary	Extended.Info	Lifestage	Age	AgeUnits	Sex	Type	Project	SampleName
BSS00043	Adipose	NA	ADIPOSE TISSUE	adult	34	year	male	tissue	Roadmap	adipose_tissue_male_adult_(34_years)
BSS00189	Blood & T-cell	NA	CD4 T CELL	adult	21	year	male	primary cell	Roadmap	CD4-positive,_alpha-beta_T_cell_male_adult_(21_year)
BSS00197	Blood & T-cell	NA	CD8 T CELL	adult	21	year	male	primary cell	Roadmap	CD8-positive,_alpha-beta_T_cell_male_adult_(21_year)
BSS01419	Blood & T-cell	NA	MONONUCLEAR CELL	adult	28	year	female	primary cell	Roadmap	peripheral_blood_mononuclear_cell_female_adult_(28_years)
BSS01424	Blood & T-cell	NA	MONONUCLEAR CELL	adult	39	year	male	primary cell	Roadmap	peripheral_blood_mononuclear_cell_male_adult_(39_years)
BSS01689	Blood & T-cell	NA	T CELL	adult	37	year	male	primary cell	Roadmap	T-cell_male_adult_(37_years)
BSS00089	Brain	NA	ASTROCYTE	unknown	unknown		unknown	primary cell	ENCODE	astrocyte
BSS01126	Brain	NA	HIPPOCAMPUS	adult	81	year	male	tissue	Roadmap	layer_of_hippocampus_male_adult_(81_year)
BSS01562	Cancer	Brain	NEUROBLASTOMA	child	4	year	female	cell line	ENCODE	SK-N-SH
BSS01391	Cancer	HSC & B-cell	B CELL LYMPHOMA	adult	48	year	male	cell line	ENCODE	OCI-LY7
BSS01065	Cancer	HSC & B-cell	B CELL LYMPHOMA	adult	72	year	female	cell line	ENCODE	Karpas-422
BSS00762	Cancer	HSC & B-cell	MYELOGENOUS LEUKEMIA	adult	53	year	female	cell line	ENCODE	K562
BSS00558	Cancer	Liver	HEPATOCELLULAR CARCINOMA	child	15	year	male	cell line	ENCODE	HepG2
BSS00017	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_10_minutes
BSS00019	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_15_minutes
BSS00021	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_20_minutes
BSS00022	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_25_minutes
BSS00027	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_5_minutes
BSS00016	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_10_hours
BSS00020	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_2_hours
BSS00023	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_3_hours
BSS00024	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_30_minutes
BSS00026	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_5_hours
BSS00028	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_6_hours
BSS00029	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_7_hours
BSS00018	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_12_hours
BSS00025	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_4_hours
BSS00030	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	GGR	A549_treated_with_100_nM_dexamethasone_for_8_hours
BSS00007	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	ENCODE	A549
BSS00015	Cancer	Lung	LUNG EPITHELIAL CARCINOMA	adult	58	year	male	cell line	ENCODE	A549_treated_with_100_nM_dexamethasone_for_1_hour
BSS01226	Cancer	Other	MAMMARY GLAND ADENOCARCINOMA	adult	69	year	female	cell line	ENCODE	MCF-7
BSS01405	Cancer	Pancreas	PANCREAS DUCT EPITHELIAL CARCINOMA	unknown	56	year	male	cell line	ENCODE	Panc1
BSS00529	Cancer	Reproductive	CERVIX ADENOCARCINOMA	adult	31	year	female	cell line	ENCODE	HeLa-S3
BSS01414	Cancer	Reproductive	PROSTATE ADENOCARCINOMA	adult	62	year	male	cell line	ENCODE	PC-3
BSS00316	Digestive	NA	ESOPHAGUS	adult	30	year	female	tissue	Roadmap	esophagus_female_adult_(30_years)
BSS00318	Digestive	NA	ESOPHAGUS	adult	34	year	male	tissue	Roadmap	esophagus_male_adult_(34_years)
BSS00321	Digestive	NA	ESOPHAGUS MUSCULARIS MUCOSA	adult	53	year	female	tissue	ENCODE	esophagus_muscularis_mucosa_female_adult_(53_years)
BSS00325	Digestive	NA	ESOPHAGUS SQUAMOUS EPITHELIUM	adult	53	year	female	tissue	ENCODE	esophagus_squamous_epithelium_female_adult_(53_years)
BSS00381	Digestive	NA	GASTROESOPHAGEAL SPHINCTER	adult	53	year	female	tissue	ENCODE	gastroesophageal_sphincter_female_adult_(53_years)
BSS01119	Digestive	NA	LARGE INTESTINE	embryonic	108	day	male	tissue	Roadmap	large_intestine_male_embryo_(108_days)
BSS01426	Digestive	NA	PEYERS PATCH	adult	53	year	female	tissue	ENCODE	Peyer_s_patch_female_adult_(53_years)
BSS01548	Digestive	NA	SIGMOID COLON	child	3	year	male	tissue	Roadmap	sigmoid_colon_male_child_(3_years)
BSS01545	Digestive	NA	SIGMOID COLON	adult	34	year	male	tissue	Roadmap	sigmoid_colon_male_adult_(34_years)
BSS01543	Digestive	NA	SIGMOID COLON	adult	53	year	female	tissue	ENCODE	sigmoid_colon_female_adult_(53_years)
BSS01597	Digestive	NA	SMALL INTESTINE	adult	34	year	male	tissue	Roadmap	small_intestine_male_adult_(34_years)
BSS01599	Digestive	NA	SMALL INTESTINE	child	3	year	male	tissue	Roadmap	small_intestine_male_child_(3_years)
BSS01588	Digestive	NA	SMALL INTESTINE	adult	30	year	female	tissue	Roadmap	small_intestine_female_adult_(30_years)
BSS01601	Digestive	NA	SMALL INTESTINE	embryonic	108	day	male	tissue	Roadmap	small_intestine_male_embryo_(108_days)
BSS01637	Digestive	NA	STOMACH	adult	30	year	female	tissue	Roadmap	stomach_female_adult_(30_years)
BSS01654	Digestive	NA	STOMACH	child	3	year	male	tissue	Roadmap	stomach_male_child_(3_years)
BSS01651	Digestive	NA	STOMACH	adult	34	year	male	tissue	Roadmap	stomach_male_adult_(34_years)
BSS01639	Digestive	NA	STOMACH	adult	53	year	female	tissue	ENCODE	stomach_female_adult_(53_years)
BSS01851	Digestive	NA	TRANSVERSE COLON	adult	54	year	male	tissue	ENCODE	transverse_colon_male_adult_(54_years)
BSS01849	Digestive	NA	TRANSVERSE COLON	adult	53	year	female	tissue	ENCODE	transverse_colon_female_adult_(53_years)
BSS01850	Digestive	NA	TRANSVERSE COLON	adult	37	year	male	tissue	ENCODE	transverse_colon_male_adult_(37_years)
BSS00283	Endocrine	Pancreas	ENDOCRINE PANCREAS	adult	45	year	male	tissue	Roadmap	endocrine_pancreas_male_adult_(45_years)
BSS00281	Endocrine	Pancreas	ENDOCRINE PANCREAS	adult	59	year	unknown	tissue	Roadmap	endocrine_pancreas_adult_(59_years)
BSS01399	Endocrine	Reproductive	OVARY	adult	30	year	female	tissue	Roadmap	ovary_female_adult_(30_years)
BSS00046	Endocrine	NA	ADRENAL GLAND	adult	30	year	female	tissue	Roadmap	adrenal_gland_female_adult_(30_years)
BSS00048	Endocrine	NA	ADRENAL GLAND	adult	53	year	female	tissue	ENCODE	adrenal_gland_female_adult_(53_years)
BSS00054	Endocrine	NA	ADRENAL GLAND	adult	34	year	male	tissue	Roadmap	adrenal_gland_male_adult_(34_years)
BSS00055	Endocrine	NA	ADRENAL GLAND	adult	37	year	male	tissue	ENCODE	adrenal_gland_male_adult_(37_years)
BSS00056	Endocrine	NA	ADRENAL GLAND	adult	54	year	male	tissue	ENCODE	adrenal_gland_male_adult_(54_years)
BSS01834	Endocrine	NA	THYROID GLAND	adult	37	year	male	tissue	ENCODE	thyroid_gland_male_adult_(37_years)
BSS01832	Endocrine	NA	THYROID GLAND	adult	53	year	female	tissue	ENCODE	thyroid_gland_female_adult_(53_years)
BSS01835	Endocrine	NA	THYROID GLAND	adult	54	year	male	tissue	ENCODE	thyroid_gland_male_adult_(54_years)
BSS00296	Endothelial	Placenta & EEM	UMBILICAL VEIN ENDOTHELIAL CELL	newborn	unknown		male	primary cell	ENCODE	endothelial_cell_of_umbilical_vein_male_newborn
BSS00355	Epithelial	NA	FORESKIN KERATINOCYTE	newborn	43500	day	male	primary cell	GGR	foreskin_keratinocyte_male_newborn_(2-4_days)
BSS00361	Epithelial	NA	FORESKIN KERATINOCYTE	newborn	43500	day	male	primary cell	GGR	foreskin_keratinocyte_male_newborn_(2-4_days)_treated_with_1.2_mM_calcium_for_2.5_days
BSS00367	Epithelial	NA	FORESKIN KERATINOCYTE	newborn	43500	day	male	primary cell	GGR	foreskin_keratinocyte_male_newborn_(2-4_days)_treated_with_1.2_mM_calcium_for_5.5_days
BSS00354	Epithelial	NA	FORESKIN KERATINOCYTE	newborn	unknown		male	primary cell	Roadmap	foreskin_keratinocyte_male_newborn
BSS01068	Epithelial	NA	KERATINOCYTE	unknown	unknown		female	primary cell	ENCODE	keratinocyte_female
BSS00112	ES-deriv	Brain	BIPOLAR NEURON DERIV	adult	53	year	male	in vitro differentiated cells	ENCODE	bipolar_neuron_originated_from_GM23338_treated_with_0.5_mg_mL_doxycycline_hyclate_for_4_days
BSS01366	ES-deriv	Brain	NEURAL DERIV	embryonic	unknown		male	in vitro differentiated cells	ENCODE	neural_cell_originated_from_H1-hESC
BSS01370	ES-deriv	Brain	NEURAL PROGENITOR DERIV	embryonic	5	day	female	in vitro differentiated cells	ENCODE	neural_progenitor_cell_originated_from_H9
BSS01371	ES-deriv	Brain	NEURAL PROGENITOR DERIV	embryonic	unknown		male	in vitro differentiated cells	Roadmap	neural_stem_progenitor_cell_originated_from_H1-hESC
BSS00171	ES-deriv	Heart	CARDIAC MUSCLE DERIV	embryonic	unknown		unknown	in vitro differentiated cells	ENCODE	cardiac_muscle_cell_originated_from_RUES2
BSS00556	ES-deriv	Liver	HEPATOCYTE DERIV	embryonic	5	day	female	in vitro differentiated cells	ENCODE	hepatocyte_originated_from_H9
BSS01261	ES-deriv	Mesench	MESENCHYMAL STEM DERIV	embryonic	unknown		male	in vitro differentiated cells	Roadmap	mesenchymal_stem_cell_originated_from_H1-hESC
BSS01857	ES-deriv	Placenta & EEM	TROPHOBLAST DERIV	embryonic	unknown		male	in vitro differentiated cells	Roadmap	trophoblast_cell_originated_from_H1-hESC
BSS01612	ES-deriv	Sm. Muscle	SMOOTH MUSCLE DERIV	embryonic	5	day	female	in vitro differentiated cells	ENCODE	smooth_muscle_cell_originated_from_H9
BSS00287	ES-deriv	NA	ENDODERMAL DERIV	embryonic	unknown		male	in vitro differentiated cells	Roadmap	endodermal_cell_originated_from_HUES64
BSS01263	ES-deriv	NA	MESENDODERM DERIV	embryonic	unknown		male	in vitro differentiated cells	Roadmap	mesendoderm_originated_from_H1-hESC
BSS01264	ES-deriv	NA	MESODERMAL DERIV	embryonic	unknown		male	in vitro differentiated cells	Roadmap	mesodermal_cell_originated_from_HUES64
BSS01866	ESC	NA	ESC	embryonic	unknown		female	cell line	Roadmap	UCSF-4
BSS00717	ESC	NA	ESC	embryonic	unknown		male	cell line	Roadmap	HUES64
BSS00478	ESC	NA	ESC	embryonic	unknown		male	cell line	ENCODE	H1-hESC
BSS00079	Heart	NA	AORTA	adult	30	year	female	tissue	Roadmap	aorta_female_adult_(30_years)
BSS00080	Heart	NA	AORTA	adult	34	year	male	tissue	Roadmap	aorta_male_adult_(34_years)
BSS00088	Heart	NA	ASCENDING AORTA	adult	53	year	female	tissue	ENCODE	ascending_aorta_female_adult_(53_years)
BSS00513	Heart	NA	HEART LEFT VENTRICLE	child	3	year	male	tissue	Roadmap	heart_left_ventricle_male_child_(3_years)
BSS00512	Heart	NA	HEART LEFT VENTRICLE	adult	34	year	male	tissue	Roadmap	heart_left_ventricle_male_adult_(34_years)
BSS00507	Heart	NA	HEART LEFT VENTRICLE	adult	53	year	female	tissue	ENCODE	heart_left_ventricle_female_adult_(53_years)
BSS01508	Heart	NA	HEART RIGHT ATRIUM	adult	34	year	male	tissue	Roadmap	right_cardiac_atrium_male_adult_(34_years)
BSS01507	Heart	NA	HEART RIGHT ATRIUM	adult	53	year	female	tissue	ENCODE	right_atrium_auricular_region_female_adult_(53_years)
BSS00524	Heart	NA	HEART RIGHT VENTRICLE	adult	34	year	male	tissue	Roadmap	heart_right_ventricle_male_adult_(34_years)
BSS00525	Heart	NA	HEART RIGHT VENTRICLE	child	3	year	male	tissue	Roadmap	heart_right_ventricle_male_child_(3_years)
BSS01815	Heart	NA	THORACIC AORTA	adult	54	year	male	tissue	ENCODE	thoracic_aorta_male_adult_(54_years)
BSS01814	Heart	NA	THORACIC AORTA	adult	37	year	male	tissue	ENCODE	thoracic_aorta_male_adult_(37_years)
BSS00101	HSC & B-cell	NA	B CELL	adult	37	year	male	primary cell	Roadmap	B_cell_male_adult_(37_years)
BSS00178	HSC & B-cell	NA	CD14 MONOCYTE	unknown	unknown		female	primary cell	ENCODE	CD14-positive_monocyte_female
BSS00232	HSC & B-cell	NA	CD34 CMP	adult	33	year	female	primary cell	Roadmap	common_myeloid_progenitor,_CD34-positive_female_adult_(33_years)
BSS01355	HSC & B-cell	NA	NK CELL	adult	37	year	male	primary cell	Roadmap	natural_killer_cell_male_adult_(37_years)
BSS01519	Liver	NA	LIVER	adult	53	year	female	tissue	ENCODE	right_lobe_of_liver_female_adult_(53_years)
BSS01201	Lung	NA	LUNG	child	3	year	male	tissue	Roadmap	lung_male_child_(3_years)
BSS01190	Lung	NA	LUNG	adult	30	year	female	tissue	Roadmap	lung_female_adult_(30_years)
BSS01870	Lung	NA	LUNG	adult	53	year	female	tissue	ENCODE	upper_lobe_of_left_lung_female_adult_(53_years)
BSS00439	Lymphoblastoid	NA	LYMPHOBLASTOID CELL LINE	adult	unknown		female	cell line	ENCODE	GM12878
BSS00378	Muscle	NA	GASTROCNEMIUS MEDIALIS	adult	37	year	male	tissue	ENCODE	gastrocnemius_medialis_male_adult_(37_years)
BSS00377	Muscle	NA	GASTROCNEMIUS MEDIALIS	adult	53	year	female	tissue	ENCODE	gastrocnemius_medialis_female_adult_(53_years)
BSS00379	Muscle	NA	GASTROCNEMIUS MEDIALIS	adult	54	year	male	tissue	ENCODE	gastrocnemius_medialis_male_adult_(54_years)
BSS01460	Muscle	NA	PSOAS MUSCLE	adult	30	year	female	tissue	Roadmap	psoas_muscle_female_adult_(30_years)
BSS01462	Muscle	NA	PSOAS MUSCLE	adult	34	year	male	tissue	Roadmap	psoas_muscle_male_adult_(34_years)

BSS01463	Muscle	NA	PSOAS MUSCLE	child	3	year	male	tissue	Roadmap	psoas_muscle_male_child_(3_years)
BSS01344	Myosat	NA	MYOTUBE	adult	22	year	male	in vitro differentiated cells	ENCODE	myotube_originated_from_skeletal_muscle_myoblast
BSS01377	Neurosph	NA	NEUROSPHERE	embryonic	15	week	unknown	primary cell	Roadmap	neurosphere_embryo_(15_weeks)_originated_from_ganglionic_eminence
BSS00146	Other	NA	BREAST EPITHELIUM	adult	53	year	female	tissue	ENCODE	breast_epithelium_female_adult_(53_years)
BSS00368	Other	NA	FORESKIN MELANOCYTE	newborn	unknown		male	primary cell	Roadmap	foreskin_melanocyte_male_newborn
BSS00122	Pancreas	NA	BODY OF PANCREAS	adult	53	year	female	tissue	ENCODE	body_of_pancreas_female_adult_(53_years)
BSS01406	Pancreas	NA	PANCREAS	adult	30	year	female	tissue	Roadmap	pancreas_female_adult_(30_years)
BSS01407	Pancreas	NA	PANCREAS	adult	34	year	male	tissue	Roadmap	pancreas_male_adult_(34_years)
BSS00074	Placenta & EEM	NA	AMNION	embryonic	16	week	male	tissue	Roadmap	amnion_male_embryo_(16_weeks)
BSS00211	Placenta & EEM	NA	CHORION	embryonic	40	week	female	tissue	Roadmap	chorion_female_embryo_(40_weeks)
BSS00212	Placenta & EEM	NA	CHORION	embryonic	16	week	male	tissue	Roadmap	chorion_male_embryo_(16_weeks)
BSS00215	Placenta & EEM	NA	CHORIONIC VILLUS	embryonic	40	week	female	tissue	Roadmap	chorionic_villus_female_embryo_(40_weeks)
BSS00216	Placenta & EEM	NA	CHORIONIC VILLUS	embryonic	16	week	male	tissue	Roadmap	chorionic_villus_male_embryo_(16_weeks)
BSS00217	Placenta & EEM	NA	CHORIONIC VILLUS	embryonic	38	week	male	tissue	Roadmap	chorionic_villus_male_embryo_(38_weeks)
BSS00214	Placenta & EEM	NA	CHORIONIC VILLUS	embryonic	16	week	unknown	tissue	Roadmap	chorionic_villus_embryo_(16_weeks)
BSS01448	Placenta & EEM	NA	PLACENTA	embryonic	38	week	male	tissue	Roadmap	placental_basal_plate_male_embryo_(38_weeks)
BSS01446	Placenta & EEM	NA	PLACENTA	embryonic	40	week	female	tissue	Roadmap	placental_basal_plate_female_embryo_(40_weeks)
BSS01438	Placenta & EEM	NA	PLACENTA	embryonic	113	day	female	tissue	Roadmap	placenta_female_embryo_(113_days)
BSS01860	Placenta & EEM	NA	TROPHOBLAST	embryonic	40	week	female	tissue	Roadmap	trophoblast_female_embryo_(40_weeks)
BSS01841	PNS	NA	TIBIAL NERVE	adult	53	year	female	tissue	ENCODE	tibial_nerve_female_adult_(53_years)
BSS01884	Reproductive	NA	UTERUS	adult	53	year	female	tissue	ENCODE	uterus_female_adult_(53_years)
BSS01887	Reproductive	NA	VAGINA	adult	53	year	female	tissue	ENCODE	vagina_female_adult_(53_years)
BSS01628	Spleen	NA	SPLEEN	adult	30	year	female	tissue	Roadmap	spleen_female_adult_(30_years)
BSS01634	Spleen	NA	SPLEEN	child	3	year	male	tissue	Roadmap	spleen_male_child_(3_years)
BSS01631	Spleen	NA	SPLEEN	adult	34	year	male	tissue	Roadmap	spleen_male_adult_(34_years)
BSS01630	Spleen	NA	SPLEEN	adult	53	year	female	tissue	ENCODE	spleen_female_adult_(53_years)
BSS00062	Stromal	Lung	LUNG FIBROBLAST	embryonic	12	week	male	cell line	ENCODE	AG04450
BSS00720	Stromal	Lung	LUNG FIBROBLAST	embryonic	16	week	female	cell line	ENCODE	IMR-90
BSS00332	Stromal	NA	BREAST FIBROBLAST	adult	17	year	female	primary cell	Roadmap	fibroblast_of_breast_female_adult_(17_years)
BSS00333	Stromal	NA	BREAST FIBROBLAST	adult	26	year	female	primary cell	Roadmap	fibroblast_of_breast_female_adult_(26_years)
BSS00353	Stromal	NA	FORESKIN FIBROBLAST	newborn	unknown		male	primary cell	ENCODE	foreskin_fibroblast_male_newborn
BSS00476	Stromal	NA	SKIN FIBROBLAST	adult	53	year	male	cell line	ENCODE	GM23248
BSS01825	Thymus	NA	THYMUS	child	3	year	male	tissue	Roadmap	thymus_male_child_(3_years)

Method Details

Organoid dissociation.

The papain dissociation reagents were prepared according to manufacturer recommendations (http://www.worthington-biochem.com/PDS/cat.html), with a slight modification. Papain was resuspended in 5mL Hibernate E, to yield a final concentration of 20U/mL, to negate the need for 95% O2/5% CO2 equilibration. DNAse was resuspended in EBSS as recommended and mixed gently to avoid shearing before being added to the papain solution. The final papain/DNAse solution was then incubated at 37C for 10 minutes prior to use to ensure complete solubilization. To dissociate, organoids were washed twice with Hibernate E in a 1.5mL eppendorf before being transferred to a 10cm dish in fresh Hibernate E. Organoids were gently diced using a single edge razor blade into small chunks. These chunks were then transferred to a 15mL tube and pelleted to remove the Hibernate E. Diced organoid chunks were subsequently resuspended in 5mL of papain/DNAse solution at a final concentration of 20U/mL. Organoids were incubated at 37C with constant agitation for 30 minutes. After 30 minutes, the organoids were manually titrated 5 times using a 5mL stripette to break up clumps, then placed at 37C for a further 15 minutes. After 15 minutes, organoids were very gently titrated 10 times with a P1000 tip and placed for a further 15 minutes at 37C. In total, organoids were incubated in papain for 1h to obtain a single cell solution. Resulting cells were then filtered through a 40μM strainer into a fresh 15mL tube and centrifuged at 300g for 10 minutes. The papain/DNAse solution was removed and cells were resuspended in Hibernate E and centrifuged again. This process was repeated once more to completely remove papain and the majority of cell debris. Finally, the cells were resuspended in 1mL of PBS with 0.04% BSA and counted using live/dead stain Trypan blue on the countess II (Thermo Fisher #AMQAX1000). Cell solution was >80% live for subsequent single cell sequencing. The cell concentration was then adjusted to 1000 cells/μL for loading into the 10X scRNA chip and 5000 cells/μL for the scATAC chip.

Single cell sequencing, processing and analysis.

Single cell RNA-seq and single cell ATAC-seq libraries of the human H9 fetal brain organoid were constructed using the Chromium Single-Cell 3' Library Kit (10X Genomics) for enzymatic fragmentation, end-repair, A-tailing, adapter ligation, ligation cleanup, sample index PCR, and PCR cleanup. After library quality control, sequencing libraries were loaded on an Illumina NovaSeq 6000 for sequencing and fastq file generation.

Raw fastq files for scRNA and scATAC were processed with the corresponding Cell Ranger software (10X Genomics) to map reads and filter barcodes. Briefly, the raw sequencing reads were aligned to the transcriptome using STAR (REF), using a hg19 human transcriptome reference from GENCODE. Expression counts for each gene in all samples were collapsed and normalized to unique molecular identifier (UMI) counts, yielding a large digital matrix with cell barcodes as rows and gene identities/peaks as columns.

Digital gene expression matrices were processed with the Seurat toolkit (REF) for downstream quality control followed by dimensionality reduction and the generation of cluster/cell type assignments and differential genes expression analysis. Peak matrices were processed in a similar fashion with the Signac toolkit (REF).

Reference atlas generation.

We used matched RNA-seq and H3K27ac ChIP-seq samples generated the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium) and Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010) consortiums, aggregrated in a unified format by EpiMap (for Epigenome Integration across Multiple Annotation Projects) (REF), that covered a wide variety of human cell types to train models to predict the genome-wide chromatin signal intensity from gene expression features. RNA-seq of EpiMap samples were downloaded from https://personal.broadinstitute.org/cboix/epimap/extended_data/rnaseq_data/ and ChIP-seq data

from https://epigenome.wustl.edu/epimap/data/observed/. An atlas of 146 cell types was generated with matching RNA-seq and H3K27ac ChIP-seq data types (Table 3-1).

The ChIP-seq data was subset to include chromosomes 1-22 and X because samples contained a mixture of male and female cell types. These data were then uniformly processed into 200bp non-overlapping bins measuring reads as normalized fragments per kilobase of exon model per million reads mapped (FPKM). RNA-seq data was downloaded as a Log2 FPKM normalized table but had to be reshaped from long-format (3-column format: cellt ype, gene, fpkm value) to wide-format (gene by cellt ype matrix).

DeconR model training and integration.

DeconR takes two modeling approaches, across- and within-cell type modeling, to accurately deconvolve the chromatin landscape in single cells. The across-cell type model requires the cell types within the reference atlas to be split into training, validation and test sets. As a first step, the genome-wide signal intensity values (15,181,508 200bp bins) of the training set are clustered into 2000 groups, termed loci groups, for each chromosome using an optimized K-means algorithm that leverages the Armadilla C++ library for large data (Struyf et al., 1996). For each group, within each chromosome, we train a K-nearest neighbor (KNN) regression model to predict the average signal intensity value for each cell type, given gene expression measurements for the respective cell type. In order to select the optimal number of nearest neighbors for each loci group, we perform a grid search by training each model with 1-73 nearest neighbors with a step size of 2, then select the number of nearest neighbors which minimizes the RMSE of the validation set. We choose a maximum of 73 neighbors as this corresponds to 50% of the number of cell types within the reference atlas. Allowing for more than 73 neighbors in the regression tends to over-smooth the predictions leading to less accurate estimates. Model performance is evaluated on the test set of cell types using root mean square error (RMSE) as a measure of accuracy.

The within-cell type model takes advantage of cell type-specific relationships between the gene expression features and signal intensity values. A model is built for each cell type in the reference atlas by engineering features to describe positions along the genome. Each locus (200bp bin) within each cell type is associated with its five nearest genes. Features are constructed to denote the expression, linear genomic distance to TSSs, and expression encoding (0 = not expressed, 1 = expressed) for all nearest genes to the locus. We leverage the H2O.ai framework (REF; H2O.ai. H2O AutoML, June 2017. URL http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html. H2O version 3.30.0.1.) for model selection and automatic hyperparameter optimization through the use of the h2o.automl() function (https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html). This function will optimally find a combination of a collection of prediction algorithms including, but not limited to, Gradient Boosting Machines (GBMs), Random Forests (RFs), a fixed grid of Generalized Linear Models (GLMs), and a near-default Deep Neural Net through a process known as stacking. This function can also take advantage of Nvidia GPUs for highly optimized model training. The predictions from each cell type-specific model is averaged to robustly estimate the signal intensity at a given locus.

The predicted output tracks from the across- and within-cell type models are integrated together using Ridge linear regression. Here, we include a feature that denotes the number of cell types within the reference atlas that contain a given peak (e.g num_celltypes). Coefficient are estimated for the across, within, and num_celltypes features on a cell type-specific basis. As with the within-cell type modeling, the predictions from each integration model are averaged together to provide a final prediction. The final predictions are used downstream to preform the deconvolution by using the predicted values at each locus as a proxy for single cell type fraction of the bulk, convolved chromatin signal.

Overcoming dropout by imputation and pseudo-bulk aggregation of scRNA-seq cell types.

To overcome the inherent problem of dropout in single cell assays, we perform imputation with scImpute (Li and Li, 2018) to correct for gene expression values identified likely as dropout without introducing new biases to the data. After imputation, the single cell gene expression data is summed together at the gene-level for each cell type, creating pseudo-bulk cell types. Pseudo-bulk cell types are size factor normalized to correct the total transcript count. Pseudo-bulk cell types are used to generated features and predictions for within- and across-cell type modeling techniques.

Peak calling, motif and peak pattern analysis.

Peaks are called on the integrated tracks using MAC2 (Zhang et al., 2008; Feng et al., 2012) with the command `macs2 bdgpeakcall -i ${file} -o ${file}_peaks.bed`. The resulting peak file is used for motif analysis through HOMER (Heinz et al., 2010) with the command `findMotifsGenome.pl ${file}_peaks.bed hg19${file} -p 32 -size 200 -mask `.

A binarized peak consensus matrix is computed for each modeling techniques and the corresponding ground truth, for each cell type within the reference atlas. All called peaks for all signal intensity tracks are constructed into a binary matrix where 1 and 0 corresponding to the presence and absence of a peak, respectively. By constructing a matrix in this way we can analyze and sort for peaks within a given cell type as a three digit encoding (across, within, truth) or across single cell cell types previously described (Fig 3-2) as a six digit encoding (cluster1, cluster2, … , cluster6 *or* maturing_neurons, … , early_astrocyte_glial).

Data, pre-trained models, and code availability

The DeconR software is available through GitHub at https://github.com/ShanSabri/deconR. Instructions to download the reference atlas with pre-trained models for each of the 146 cell types used in this study, as well as source code, within-/across-/ integrated-cell type prediction tracks

with peaks and motif calls are uploaded to https://github.com/ShanSabri/deconR/blob/master/data.md.

References

1. Kidder, B. L., Hu, G., & Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. *Nature Immunology*, **12**(10**)**, 918–922.
2. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, **161**(5), 1187–1201.
3. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**(5), 1202–1214.
4. Vickovic, S., Ståhl, P. L., Salmén, F., Giatrellis, S., Westholm, J. O., Mollbrink, A., Navarro, J. F., Custodio, J., Bienko, M., Sutton, L.-A., Rosenquist, R., Frisén, J., & Lundeberg, J. (2016). Massive and parallel expression profiling using microarrayed single-cell sequencing. *Nature Communications*, **7**(1).
5. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**(7561), 486–490.
6. Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., … Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**(6237), 910–914.
7. Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**(1).
8. Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., … Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, **37**(8), 925–936.
9. Khoo, B. L., Chaudhuri, P. K., Ramalingam, N., Tan, D. S. W., Lim, C. T., & Warkiani, M. E. (2016). Single-cell profiling approaches to probing tumor heterogeneity. *International Journal of Cancer*, **139**(2), 243–255.
10. Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., van Oudenaarden, A. & Jaenisch, R. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150**, 1209–1222 (2012).
11. 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. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell* **151**, 1617–1632 (2012).
12. 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 (2019).
13. Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., … Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, **357**(6352), 661–667.
14. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science*, **360**(6391), eaaq1736.

15. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, **172**(5), 1091–1107.e17.
16. Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., & Zinzen, R. P. (2017). The Drosophila embryo at single-cell transcriptome resolution. *Science*, **358**(6360), 194–199.
17. The Tabula Muris Consortium et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**(7727), 367–372.
18. Zhou, W., Ji, Z., Fang, W., & Ji, H. (2019). Global prediction of chromatin accessibility using small-cell-number and single-cell RNA-seq. *Nucleic Acids Research*, **47**(19), e121–e121.
19. Chen, C., Zhang, S., & Zhang, X.-S. (2013). Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. *Nucleic Acids Research*, **41**(20), 9230–9242.
20. Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., … Ren, B. (2019). An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural & Molecular Biology*, **26**(11), 1063–1070.
21. Pott, S. (2017) Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife*, **6**, e23203.
22. Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., & Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, **9**(1).
23. Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, **107**(50), 21931–21936.
24. Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, **33**(11), 1165–1172.
25. Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, **107**(7), 2926–2931.
26. Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E., & Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, **13**(9), R53.
27. Zhou, W., Sherwood, B., Ji, Z., Xue, Y., Du, F., Bai, J., Ying, M., & Ji, H. (2017). Genome-wide prediction of DNase I hypersensitivity using gene expression. *Nature Communications*, **8**(1).
28. Zhou, W., Ji, Z., Fang, W., & Ji, H. (2019). Global prediction of chromatin accessibility using small-cell-number and single-cell RNA-seq. *Nucleic Acids Research*, **47**(19), e121–e121.
29. Zeng, W., Chen, X., Duren, Z., Wang, Y., Jiang, R., & Wong, W. H. (2019). DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nature Communications*, **10**(1).
30. Kim, H.-J., Gürkan Yardımcı, G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J., Duan, Z., & Stafford Noble, W. (2019). Capturing cell type-specific chromatin structural patterns by applying topic modeling to single-cell Hi-C data. *Cold Spring Harbor Laboratory*.
31. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, **36**(5), 411–420.

32. McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3**(29), 861.
33. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., … Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, **10**(1).
34. Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, **9**(1).
35. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.
36. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.
37. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.
38. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, **28**(10), 1045–1048.
39. Horikoshi, T., Ezoe, K., Nakagawa, H., Eguchi, H., Hanada, N., & Hamaoka, S. (1995). Up-regulation of ICAM-1 expression on human dermal fibroblasts by IFN β in the presence of TNF-α. *FEBS Letters*, **363**(1–2), 141–144.
40. Piela-Smith TH, Broketa G, Hand A, Korn J H. (1992). Regulation of ICAM-1 expression and function in human dermal fibroblasts by IL-4. *The Journal of Immunology*, **148**(5) 1375-1381
41. Hasegawa, M., Higashi, K., Yokoyama, C., Yamamoto, F., Tachibana, T., Matsushita, T., Hamaguchi, Y., Saito, K., Fujimoto, M., & Takehara, K. (2012). Altered expression of dermokine in skin disorders. *Journal of the European Academy of Dermatology and Venereology*, **27**(7), 867–875.
42. Watanabe, M., Buth, J. E., Vishlaghi, N., de la Torre-Ubieta, L., Taxidis, J., Khakh, B. S., Coppola, G., Pearson, C. A., Yamauchi, K., Gong, D., Dai, X., Damoiseaux, R., Aliyari, R., Liebscher, S., Schenke-Layland, K., Caneda, C., Huang, E. J., Zhang, Y., Cheng, G., … Novitch, B. G. (2017). Self-Organized Cerebral Organoids with Human-Specific Features Predict Effective Drugs to Combat Zika Virus Infection. *Cell Reports*, **21**(2), 517–532.
43. Samarasinghe, R. A., Miranda, O. A., Mitchell, S., Ferando, I., Watanabe, M., Buth, J. E., Kurdian, A., Golshani, P., Plath, K., Lowry, W. E., Parent, J. M., Mody, I., & Novitch, B. G. (2019). Identification of neural oscillations and epileptiform changes in human brain organoids. *Cold Spring Harbor Laboratory.*
44. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, **9**(9), R137.
45. Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, **7**(9), 1728–1740.
46. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, **38**(4), 576–589.
47. Hu, W. (2011). Parkinson's disease is a TH17 dominant autoimmune disorder against accumulated alpha-synuclein. *Nature Precedings*.
48. Hamby, M. E., Hewett, J. A., & Hewett, S. J. (2010). Smad3-dependent signaling underlies the TGF-β1-mediated enhancement in astrocytic iNOS expression. *Glia*, **58**(11), 1282–1291.
49. Stipursky, J., & Gomes, F. C. A. (2007). TGF-β1/SMAD signaling induces astrocyte fate commitmentin vitro: Implications for radial glia development. *Glia*, **55**(10), 1023–1033.

50. Stipursky, J., Francis, D., & Gomes, F. C. A. (2012). Activation of MAPK/PI3K/SMAD Pathways by TGF-β1 Controls Differentiation of Radial Glia into Astrocytes in vitro. *Developmental Neuroscience*, **34**(1), 68–81.
51. Su, Z., Zhang, Y., Liao, B., Zhong, X., Chen, X., Wang, H., Guo, Y., Shan, Y., Wang, L., & Pan, G. (2018). Antagonism between the transcription factors NANOG and OTX2 specifies rostral or caudal cell fate during neural patterning transition. *Journal of Biological Chemistry*, **293**(12), 4445–4455.
52. Hutton, S. R., & Pevny, L. H. (2011). SOX2 expression levels distinguish between neural progenitor populations of the developing dorsal telencephalon. *Developmental Biology*, **352**(1), 40–47.
53. Iulianella, A., Sharma, M., Durnin, M., Vanden Heuvel, G. B., & Trainor, P. A. (2008). Cux2 (Cutl2) integrates neural progenitor development with cell-cycle progression during spinal cord neurogenesis. *Development*, **135**(4), 729–741.
54. Cubelos, B., Sebastián-Serrano, A., Kim, S., Moreno-Ortiz, C., Redondo, J. M., Walsh, C. A., & Nieto, M. (2007). Cux-2 Controls the Proliferation of Neuronal Intermediate Precursors of the Cortical Subventricular Zone. *Cerebral Cortex*, **18**(8), 1758–1770.
55. Belton, J.M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods*, **58**(3), 268–276.
56. Struyf, A., Hubert, M., & Rousseeuw, P. (1996). Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, **1**(4).

Chapter 4. Utilization of Single Cell Technologies to Create Newfound Cellular and Molecular Atlases in Human Developmental Systems

4.1 Developmental Trajectory of Human Skeletal Muscle Progenitor and Stem Cells across Development and from Pluripotent Stem Cells

Introduction.

Skeletal myogenesis starts early during development, which initially gives rise to prenatal skeletal muscle progenitor cells (SMPCs) and later on postnatal satellite cells (SCs) (Applebaum and Kalcheim, 2015; Cerletti et al., 2008; Chal and Pourquie, 2017; Sambasivan and Tajbakhsh, 2007). Both populations are endowed with muscle stem cell properties including, in addition to the expression of the essential myogenic transcription factor (TF) PAX7, the ability to expand and fuse to generate new myofibers in vitro or in vivo (Sacco et al., 2008; Tierney et al., 2016). However, the molecular and functional differences between SMPCs and SCs are only beginning to be unveiled. In vivo, mouse SMPCs contribute to muscle establishment and growth, whereas SCs in mature muscles are typically quiescent and enter the cell cycle in the event of injury (Tierney and Sacco, 2016). In vitro, isolated mouse SMPCs proliferate and maintain PAX7 expression longer than SCs. Moreover, following transplantation after muscle injury, mouse SCs are superior to fetal SMPCs to repopulate the stem cell niche and support long-term regeneration (Tierney et al., 2016). Despite studies on developmental myogenesis in model organisms, our knowledge of muscle ontogeny in human is limited (Schiaffino et al., 2015).

Following developmental cues, we and others have developed directed differentiation protocols using human pluripotent stem cells (hPSCs) to generate myogenic cells including SMPCs or SC-like cells (Borchin et al., 2013; Chal et al., 2015; Hicks et al., 2018; Magli and Perlingeiro, 2017; Shelton et al., 2014; Xi et al., 2017; Xu et al., 2013). These cells may serve as potential sources for personalized cell replacement therapies for degenerative muscle diseases or sarcopenia. However, they have not been fully characterized and compared to in vivo human SMPCs or SCs to facilitate their proper translation to clinical usage.

Here, we performed a comprehensive single cell RNA-sequencing (scRNA-seq) analysis

of myogenesis in human limb tissues across development. We identified skeletal muscle (SkM) cells as well as other supportive cell types present at distinct developmental stages. We also evaluated the myogenic and non-myogenic cell populations from three different directed differentiation strategies from hPSCs. Using the developmental trajectory built from the in vivo SMPCs and SCs, we mapped hPSC-derived progenitor cells to a developmental period corresponding to the embryonic-to-fetal transition (7-12 weeks prenatal) across all protocols. Further analysis identified gene groups differentially regulated across developmental stages and provided potential TF candidates that may regulate stage transitions. In summary, this work provides a critical resource to understand the developmental networks defining human skeletal myogenesis and can be used to aid molecular identification of myogenic cells derived from hPSCs. This work will enable the development of new approaches to mature and support the most regenerative cells from hPSCs for use in cell-based therapies.

Results.

Identification of skeletal myogenic and supportive cell types using scRNA-seq across in vivo human development

To gain a comprehensive view of cell populations present during human SkM ontogeny, we used scRNA-seq to evaluate human limb muscle tissues from embryonic (week 5-8), fetal (week 9-18) as well as postnatal juvenile (year 7-11) and adult (year 34-42) stages (see STAR Methods). To universally identify skeletal myogenic cells from different samples, we developed a computational tool called "Muscle.Score" that examines the average expression of a list of conserved genes representing myogenic cells of distinct developmental and differentiation status (PAX3, PAX7, PITX2, MYF5, MYF6, MYOD1, MYOG, NEB and MYH3). Using "Muscle.Score", we were able to readily identify SkM cells at each developmental stage (Figure 4-1). Within mononucleated cells from whole limbs, SkM cells gradually increased in proportion from early embryonic (week 5-6; ~5%) to the beginning of fetal (week 9; above 20%) stage (Figures 4-1A-D, 4-1I-L and S4-1I). At early fetal stage (week 12-14; ~35%), SkM cells constituted a major cell type of the non-endothelial/hematopoietic lineages in limbs (Figures 4-1E, 4-1M and S4-1I). This proportion decreased during later fetal development (week 17-18; ~15%) and further dropped in postnatal juvenile and adult limb SkM tissues (below 10%) (Figures 4-1F-H, 4-1N-P and S4-1I).

In addition to SkM cells, we also found various non-myogenic populations at distinct developmental timepoints. One highly dynamic population is formed by mesenchymal cell types. Early on (week 5-6), the mesenchyme of the developing limbs was relatively homogeneous, mainly comprised of DUSP6+ multipotent limb mesenchymal progenitors (Limb.Mesen) (Gros and Tabin, 2014; Reinhardt et al., 2019) (Figures 4-1A, 4-1I and S4-1A). As limbs develop (week 6-9), the multipotent progenitors became more lineage restricted and SHOX2+ prechondrogenic (PreChondro) and SOX9+ chondrogenic (Chondro) progenitors became prominent (Akiyama et al., 2005; Barna and Niswander, 2007; Neufeld et al., 2014) (Figures 4-1B-D, 4-1J-L and S4-1B-D). During fetal development (week 12-18), the mesenchymal cells expressed the mesenchymal

stromal cell (MSC) marker NT5E/CD73 (Figures 4-1E, 4-1F, 4-1M, 4-1N, S4-1E and S4-1F). At postnatal stage, the mesenchymal/stromal population was highly enriched for PDGFRA, a marker for fibro-adipogenic progenitors (FAPs) found in adult mouse SkM (Joe et al., 2010; Uezumi et al., 2010) (Figures 4-1G, 4-1H, 4-1O, 4-1P, S4-1G and S4-1H). Other cell types present at various levels across limb development include dermal fibroblasts and progenitors (Dermal; TWIST2+), Schwann cells (CDH19+), smooth muscle cells (SMCs; MYLK+) and tenogenic cells (Teno; TNMD+). Skin cells (KRT19+), endothelial cells (ECs; ESAM+) and the hematopoietic (Hema; SRGN+) lineages including red and white blood cells (RBCs and WBCs; HEMGN+ and AIF1+, respectively) were detected at early stages (week 5-9) (Figures 4-1A-1D, 4-1I-L and S4-1A and S4-1D), and only residuals of these cell types were found at later stages (week 12 and later) as they were either removed during tissue dissection (skin) or depleted during cell sorting (EC and Hema). In summary, using our scRNA-seq pipeline, we were able to identify dynamic cell populations of both myogenic and non-myogenic nature across human limb development.

Skeletal myogenic subpopulations vary throughout human development

At embryonic week 5-6, the myogenic population in the developing hindlimbs was relatively homogeneous and mainly consisted of PAX3+ myogenic progenitors (MPs) (Figure 4-2A). Later at week 6-7, a small subset of differentiating myoblasts-myocytes (MB-MC) were observed that expressed commitment and terminal differentiation markers including MYOD1, MYOG and MYH3 (Figure 4-2B). At the same time, MPs increased PAX7 while decreasing PAX3 expression. The differentiating MB and MC subpopulations became more prominent during week 7-9 (Figures 4-2C and 4-2D), consistent with the rapid expansion of SkM needed to support prenatal growth. During fetal week 12-18, we found a reduction of MBs and MCs (Figures 4-2E and 4-2F), possibly due to incorporation of most differentiated myogenic cells into multi-nucleated myofibers. At postnatal stage, SkM cells were mainly comprised of PAX7+ SCs with little to no differentiating cells detected (Figures 4-2G and 4-2H).

In addition to the myogenic subpopulations reflecting distinct differentiation status, we also found another subpopulation transiently present between weeks 7 and 18. This subset expressed the canonical myogenic markers, albeit at slightly lower levels. Compared to the main myogenic subpopulations (MP, MB and MC), these cells uniquely expressed genes suggesting a more mesenchymal-like nature, such as PDGFRA and OGN, and we termed them SkM mesenchymal subtype (SkM.Mesen) (Figures 4-2C-F).

To better understand the molecular differences among myogenic subpopulations, we focused on fetal week 9 as an example as all four subpopulations were readily detected at this time point. We performed differential gene expression of the subpopulations, followed by gene ontology (GO) analysis as well as Gene Set Enrichment Analysis (GSEA). As expected, MCs were enriched for muscle contraction genes compared to MPs. Moreover, MCs highly expressed genes involved in mitochondria and oxidative phosphorylation (OxPhos) as well as calcium signaling (Figures 4-2I, 4-2L and S4-2A). Proliferating MPs were enriched for genes regulating cell cycle progression, RNA splicing and protein translation (Figures 4-2J and 4-2L). MYC and WNT/β-catenin pathways were also enriched in MPs compared to MCs (Figures 4-2L and S4-2B). Another major category of genes enriched in MPs was the extracellular matrix (ECM), which included several members of the laminin family (Figures 4-2L and S4-2B). Interestingly, compared to the main myogenic subpopulations, SkM.Mesen cells were also highly enriched for ECM genes including collagens and regulators of collagen biosynthesis (Figures 4-2K, 4-2L and S4-2C). To rule out the possibility that the SkM.Mesen subtype was an artifact of misclassification of mesenchymal or skeletogenic cells into the myogenic population by using Seurat (Butler et al., 2018), we employed Monocle (Cao et al., 2019), another commonly used scRNA-seq analysis package to independently confirm this population, and found that the vast majority of SkM.Mesen cells were co-clustered with the main SkM subpopulations (Figure S4-2D). Although SkM.Mesen cells expressed some pro-chondrogenic genes such as COL11A1 and OGN, they barely expressed the core chondrogenic determination genes such as SOX9 and COL2A1 compared to

the Chondro population (Figure S4-2E). Moreover, SkM.Mesen cells in general expressed higher levels of mesenchymal/fibroblastic markers (e.g., PDGFRA, DCN, and COL3A1) than the main myogenic subpopulations, but lower than the mesenchymal cell types (Limb.Mesen or PreChondro) (Figure S4-2E).

To better characterize SkM.Mesen cells, we first performed immunohistochemical (IHC) stainings of PAX7, along with PDGFRA which is enriched in the SkM.Mesen subpopulation (Figures 4-2C-F). We found in human embryonic and fetal limb sections that a subset of PAX7-expressing myogenic cells were co-stained with PDGFRA (Figures 4-3A and 4-3B), corroborating the presence of this myogenic subpopulation in vivo. To further explore myogenic subpopulations, we examined cell surface markers enriched in the SkM lineage over other cell types and identified CDH15 as a potential surface marker to isolate the total SkM population from human embryonic and fetal limbs (Figure 4-3C). Next, we performed flow cytometry analysis of CDH15 and PDGFRA (Figure 4-3D) and sorted cell fractions based on these two markers. In freshly sorted cells, myogenic genes PAX7, MYOD1 and MYOG were upregulated in both CDH15+ fractions compared to the CDH15- ones, which validated the usage of this marker for enriching the total myogenic cells. Interestingly, compared to the CDH15+PDGFRA- (15+P-) cells, the CDH15+PDGFRA+ (15+P+) cells showed lower expression of myogenic genes but higher expression of genes involved in osteogenesis (RUNX2 and COL1A1) as well as mesenchyme and ECM (PDGFRA, OGN and DCN) (Figure 4-3E). When subjected to myogenic and osteogenic differentiation in vitro, respectively, 15+P- cells showed more prevalent formation of MyHC+ myotubes and higher expression of terminal myogenic differentiation genes (Figures 4-3F and 4-3G), while 15+P+ cells displayed increased Alizarin Red S-stained calcium depots and higher expression of osteogenic differentiation markers (Figures 4-3H and 4-3I). By focusing on SkM cells in the developing human hindlimbs, we were able to detect myogenic subpopulations representing not only various commitment status but also unique myogenic/osteogenic bipotential differentiation properties.

Skeletal muscle progenitor and stem cells at distinct stages of human development exhibit different gene expression programs

We next isolated SMPCs (only the MP subpopulations from prenatal samples) and SCs (from postnatal samples) in silico and subjected them to trajectory analysis. These cells formed a developmental trajectory in the diffusion map (DM) space (Haghverdi et al., 2015) consistent with the ages of individual human samples. Unbiased clustering divided the trajectory into 5 major stages (Figures 4-4A, 4-4B and S4-3A). Stage 1 mainly consisted of week 5-6 early embryonic SMPCs, while stage 2 harbored the majority of cells beyond embryonic week 6 to early week 7. Late week 7-8 embryonic SMPCs and those from week 9 in fetal development distributed relatively equally between stages 2 and 3. During fetal development of week 12-18, cells gradually progressed from stage 3 to 4. We observed some degree of overlap among sample ages and computationally calculated "stages", suggesting early prenatal myogenic development is a continuous process. Postnatal SCs from both juvenile and adult muscles comprised stage 5, and they diverged from the prenatal SMPCs on a separate trajectory (Figures 4-4A, 4-4B and S4-3A).

Although SMPCs and SCs share some common molecular markers and functionalities (Sacco et al., 2008; Tierney et al., 2016), our developmental trajectory analysis indicates that they display significant differences at the transcriptomic level. To further investigate this, we examined differentially expressed genes (DEGs) between distinct stages and found multiple biological processes and pathways differentially regulated across development. Postnatal SCs were enriched for P53 pathway components (Figure 4-4C) while expressing virtually no cell cycle promoting genes (Figure 4-4D), consistent with their quiescent state in homeostatic SkM tissues (Flamini et al., 2018). Nevertheless, several growth factor/cytokine signaling genes were enriched in SCs (Figure 4-4E), suggesting SCs use specific pathways to actively maintain their quiescence (Price et al., 2014; Shea et al., 2010; Tierney et al., 2014). Two other major differentially regulated biological processes were ECM and cellular metabolism (Figures 4-4F and 4-4G). Multiple ECM

components showed dynamic expression patterns including collagens and laminins (Figure 4-4F). For example, while COL2A1 was uniquely enriched in early embryonic SMPCs (stage 1), COL5A1 gradually increased up to later fetal period (stage 4) and was virtually undetectable at postnatal stage 5. Interestingly, genes facilitating major cellular metabolic pathways (e.g., glycolysis, TCA cycle and OxPhos) were progressively downregulated from early to late developmental stages, while metabolic inhibitors such as TXNIP and PDK4 were increased (Figure 4-4G). Other dynamically expressed gene sets included mesenchymal-like markers, myogenic cell surface molecules and Notch signaling components (Figures 4-4H-J). Interestingly, genes encoding the major components of the dystrophinglycoprotein complex (DGC) including dystrophin, dystroglycan and sarcoglycans, were increased along prenatal development with the highest expression at fetal week 17-18, and then decreased postnatally (Figure 4-4K).

When examining the canonical TFs involved in myogenesis, we also found distinct expression patterns at each developmental stage (Figure 4-4L). EYA1, SIX1 and PITX2 showed gradually decreased expression as development progresses. PAX3 progressively decreased while PAX7 increased along development. To corroborate our in silico findings, we performed IHC stainings of PAX3 and PAX7 proteins. At week 6, developing human hindlimbs contain only PAX3+ and not PAX7+ SMPCs, and no MyHC+ myofibers could be detected (Figure S4-3B). At week 7, both PAX3 and PAX7 were detected in limbs, with the proximal region containing PAX7 single positive cells while distal region harboring SMPCs transitioning from PAX3 to PAX7 expression. At this stage, thin myofibers were present with single or low number of myonuclei (Figure S4-3C). In later fetal and adult stage muscles examined (quadriceps), myofibers continued to grow in size and SMPCs and SCs were exclusively PAX7+ (Figures S4-3D and S4-3E). These results confirmed the findings of PAX3 and PAX7 transcript changes across development from our scRNA-seq analysis.

To explore the common features distinguishing between postnatal SCs and prenatal SMPCs, we performed differential gene expression analysis comparing stage 5 SCs to each

individual stage SMPCs from stage 1-4. We intersected the upregulated genes in stage 5 SCs from each of the above comparisons and generated a list of 140 genes commonly enriched in SCs compared to SMPCs (Figure 4-3M). GO analysis showed several biological processes and signaling pathways were significantly overrepresented, including metabolic and nutrient regulation, intracellular trafficking, ECM organization and cell adhesion as well as FOXO-mediated cell cycle regulation (Figure 4-3N). Interestingly, FOXO3 has been shown to promote quiescence of adult SCs in mice (Gopinath et al., 2014), suggesting that the FOXO family and related signaling pathways might play an important role in regulating the transition of proliferative prenatal SMPCs to quiescent postnatal SCs.

Although we have identified CDH15 as a cell surface marker capable of isolating SkM cells from embryonic week 7 to fetal week 19 human limbs, this marker was not shown in our scRNA-seq dataset to be enriched in the myogenic population in embryonic week 5-6 limb tissues (Figure S4-4F), and no prospective markers for myogenic cell isolation have been established for this developmental stage. Thus, we performed differential gene expression analysis between myogenic and non-myogenic cells and found some known SkM cell surface markers enriched in myogenic vs. non-myogenic populations, such as MET and CXCR4 (Bareja et al., 2014; Yin et al., 2013). However, other markers were not expressed at this stage (e.g., CD82) (Alexander et al., 2016; Uezumi et al., 2016) or not distinguishing between myogenic and other cells (e.g., NCAM1 and ITGB1) (Figure S4-3G) (Castiglioni et al., 2014; Xu et al., 2015). Next, we examined co-expression of PAX3 and MET proteins in week 5-6 human limbs using IHC (Figure S4-3H). We found nearly overlapping expression patterns of these two proteins at the ventral or dorsal level, but there was a condensed population of PAX3- cells expressing low levels of MET across a small length at the central level. When co-stained with CDH2 (Hayashi and Ozawa, 1995; Yajima et al., 1999), these central cells were found to be PAX3-METlow/+CDH2- (Figure S4-3G; right panel mosaic images). Thus, we used MET and CDH2 to sort cells from human week 5-6 limbs (Figure S4-3I), and found the MET+CDH2+ (M+C+) fractions highly enriched for PAX3 and

LBX1 transcripts compared to the MET- fractions or unsorted cells (Figure S4-3J). When cultured in vitro, only M+C+ cells were supported by the myogenic growth medium and expressed PAX3 proteins, and they could form MyHC+ myotubes after switching to fusion conditions (Figure S4-3K).

Taken together, we mapped SMPCs and SCs from different in vivo stage human samples onto a developmental trajectory, and unequivocally demonstrated the highly dynamic gene expression profiles of these cells across development. We showed striking differences in expression of genes regulating cellular processes, including ECM and metabolism, and confirmed the observed in silico differences of the classical PAX3 and PAX7 myogenic TFs at the protein levels in human tissues. We also identified cell surface markers that enabled prospective isolation of the earliest PAX3+ myogenic population from week 5-6 developing human limbs.

Directed myogenic differentiation of hPSCs generate heterogeneous cell types including both myogenic and non-myogenic cells

Although there are numerous reports describing generation of SkM cells from hPSCs, there is often a large variation in efficiency and consistency in directed differentiation protocols (Kim et al., 2017). We reasoned that by using scRNA-seq, we could identify the different cell types present across representative protocols (See STAR Methods). The balance of myogenic and non-myogenic populations may modulate the effectiveness of each differentiation towards SMPCs or SC-like cells. Using our recently published protocol (termed HX protocol) (Xi et al., 2017), we differentiated hPSCs towards the SkM lineage and profiled all live mononuclear cells in culture from 3-8 weeks of differentiation. To better track PAX7+ cells during differentiation, we used CRISPR-Cas9 directed homologous recombination to construct an endogenous PAX7-driven GFP reporter in hPSC cell lines (Figures S4-4A and S4-4B). These reporter cells were validated to enrich for PAX7 when GFP+ cells were sorted after artificial activation of the PAX7 locus by

the dCas9-VPR system (Figures S4-4C-S4-4E) or from directed myogenic differentiation (Figures S4-4F-I).

At week 3, the earliest differentiation time point examined, we detected very few SkM cells in dissociated and live-sorted cultures by our scRNA-seq approach. When the reporter cells were used to enrich for PAX7-GFP+ cells at this time point, the sorted populations mainly consisted of the neural lineage including neural progenitor cells (NPCs; SOX2+) and differentiated neurons (DCX+), while no skeletal myogenic cells could be detected (Figure 4-5A). Interestingly, the proportion of SkM cells dramatically increased one week later at week 4 in live sorted populations. At the same time, SkM cells increased to close to half of the PAX7-GFP+-sorted populations, which was accompanied by a significant decrease in the proportion of neural cells (Figures 4-5B and S4-5C). During week 5-6 of differentiation, the proportions of SkM cells in live-sorted populations were relatively stable, and they represented the major cell type in PAX7-GFP+-sorted populations (Figures 4-5C, 4-5D and S4-5C). The SkM cell proportions at week 8 of differentiation were slightly decreased in both live- and PAX7-GFP+-sorted populations (Figures 4-5E and S4-5C). Our scRNA-seq approach also confirmed the enrichment of SkM cells by using a combination of surface markers recently published by our group (Hicks et al., 2018) (Figure 4-5C).

In addition to SkM and neural cells, we also found multiple other cell types dynamically present in live-sorted populations during the course of differentiation. At week 3 of differentiation, we saw a large portion of chondrogenic cells (SOX9+/COL2A1+) and SMCs (MYLK+) dominating the cultures (Figure 4-5A), and these populations decreased over time and were absent at 6-8 week time points (Figures 4-5D and 4-5E). Meanwhile, a mesenchymal population expressing high levels of PDGFRA and THY1 but not the chondrogenic markers SOX9 or COL2A1, arose at week 4 and increased in proportion towards later time points of differentiation (Figures 4-5B-E). Another small but persistent cell type seen during the course of directed differentiation (except week 5) was the Schwann cell population (CDH19+) (Figures 4-5A, 4-5B, 4-5D and 4-5E).

Using our scRNA-seq strategy, we also examined the directed myogenic differentiation cultures from two additional protocols widely used by our lab and others published by Chal et al and Shelton et al (here termed JC and MS protocol, respectively) (Chal et al., 2015; Shelton et al., 2014). At week 5 of differentiation by JC protocol, we observed both myogenic and nonmyogenic populations present in cultures (Figure S4-5A). The latter included NPCs, neurons, Schwann cells as well as a mesenchymal population expressing high levels of PDGFRA, THY1 and DCN which is likely composed of subpopulations indicated by varying degrees of expression of additional markers (e.g., ALCAM, LUM and COL11A1). The cellular composition of the differentiation culture at week 6-7 using MS protocol were found to be quite different from that obtained from HX and JC protocols (Figure S4-5B). In addition to SkM cells, we observed a robust population highly expressing genes encoding cytokeratins (e.g., KRT19) or those pertaining to keratinization (e.g., PERP), and therefore is likely involved in epithelium development. There was another major population enriched for genes involved in skeletal development (e.g., COL1A1 and OGN) but lacking strong expression of the canonical commitment markers for the osteogenic, chondrogenic or tenogenic lineages. We also found a small subset of cells enriched for genes participating in cholesterol biosynthesis (CRABP1 and CRABP2) but the accurate identity of this population is yet to be determined.

In conclusion, our scRNA-seq approach identified dynamic cellular compositions, both myogenic and non-myogenic, during the course of hPSC SkM directed differentiation across multiple protocols. This provides a unique resource to not only further explore hPSC-derived myogenic cells, but also other cell types present in the differentiation cultures and their potential influences on in vitro hPSC myogenesis.

Skeletal muscle cells derived in vitro from hPSCs harbor multiple myogenic subpopulations during the course of directed differentiation

Similar to our approach on studying in vivo human myogenesis, we bioinformatically isolated the SkM cells from cultures examined during 4-8 weeks of in vitro hPSC directed differentiation using

HX protocol. Consistent with our in vivo findings, we also found subpopulations representing different myogenic commitment status, i.e., MP, MB and MC cells at all time points of directed differentiation, and the relative distribution of these three subpopulations largely stayed constant regardless of differentiation timing or enrichment strategies (Figures 4-6A-D). Of note, we detected MBs and MCs within the SkM populations even from PAX7-GFP+-sorted fractions. This is likely due to the low expression of PAX7 in early committed MBs (Figures 4-6A-D, middle panels) and the high stability of the GFP proteins (Li et al., 1998) retained in committed cells that have previously expressed PAX7. Interestingly, MPs at 4 weeks of directed differentiation from live-sorted populations could be further subdivided into two subsets, enriching for PAX3 and PAX7, respectively. As expected, MPs at this differentiation time point from PAX7-GFP+-sorted SkM cells were mainly comprised of PAX7+ with only barely detectable PAX3+ progenitors (Figure 4-6A). However, at later time points MPs from either live- or PAX7-GFP+-sorted fractions did not show obvious expression of PAX3 and only expressed PAX7 (Figures 4-6B-D). This is similar to the PAX3 to PAX7 transition that we observed at early in vivo human limb myogenesis (Figures 4-2A2H, 4-4L and S4-3B-E). Reminiscent of the SkM.Mesen subpopulation found during week 7-18 of prenatal development (Figures 4-2C-F), we observed a small but consistent “side” population in all examined directed differentiation time points (we also termed these cells “SkM.Mesen” but in an in vitro context). This subset of cells showed slightly higher expression of myogenic activation and commitment markers MYOD1, MYOG and MYH3 than MPs, but much lower than MBs and MCs, suggesting they are not fully committed terminally differentiated muscle cells. Meanwhile, they showed appreciably lower expression of the stem/progenitor marker PAX7 than MPs, and indeed this subpopulation was only detectable in live-sorted but not PAX7-GFP+-enriched cell fractions (Figures 4-6A-D).

When examining the SkM subpopulations from JC and MS protocols, we found similar MP, MB and MC subsets, though their relative proportions varied across different protocols (Figures S4-5D and S4-5E). Again, we observed the SkM.Mesen subpopulations from both

protocols that share many of the enriched genes and biological processes with similar populations from HX protocol as well as in vivo week 9 fetal samples (Figures 4-6E and S4-5F).

Here, we consistently identified, across multiple hPSC myogenic differentiation protocols, major and rare subpopulations within the SkM cells. This allows us to better understand the dynamics of myogenic lineage development modeled in vitro by hPSCs.

hPSC-SMPCs generated from multiple protocols align to a developmental stage of late embryonic to early fetal transition

To determine the molecular identity of hPSC-derived SMPCs, we mapped the MP subpopulations from all differentiation time points generated from HX protocol along with the in vivo progenitor and stem cells on DM space. The in vivo cells largely retain their developmental trajectory from stage 1 to 5 as previously analyzed (Figure 4-4A), with minor changes possibly due to variations introduced by adding in the in vitro cells. SMPCs derived from hPSCs aligned to the stage 2-3 in vivo SMPCs along the DM1 component and diverged along DM2 which likely results from culture-related effects (Figures 4-7A and 4-7B). To more quantitatively assess the developmental timing of the cells, we developed a more linear method to calculate each cell's developmental score ("Dev.Score"), where we took into account the expression levels of postnatal vs. embryonic enriched genes in individual cells (see STAR Methods). Using this independent method, we again found in vitro hPSC-derived SMPCs aligned to in vivo SMPCs of stage 2 to 3, which corresponds to the embryonic week 7 to fetal week 12 transition period (Figure 4-7C). Furthermore, we included additional SMPCs generated from JC and MS protocols in our analysis pipeline and found that hPSC-SMPCs derived from all protocols mapped to a similar late embryonic to early fetal transition stage of human myogenesis (Figures S4-6A-D).

To further explore the differences underlying the separation of in vivo and in vitro SMPCs, we compared the gene expression profiles of hPSC-derived myogenic progenitors from all three protocols to in vivo progenitors from stage 2 and 3, a developmental period that the hPSC-SMPCs

most closely align to. Hierarchical clustering of these five groups of cells showed major segregation based on source of in vivo or in vitro derivation, and within the in vitro hPSCSMPCs those generated from HX and JC protocols were closer to each other than those from MS protocol (Figure S3-6E). Next, we performed differential gene expression analysis between each of the three hPSC-SMPC populations compared to the stage 2 or 3 populations and found genes that are commonly enriched in either in vivo stage 2 or 3 cells (Figures S4-6F and S4-6G), and vice versa (Figures S4-6H and S4-6I). Subsequently, GO analysis of these genes revealed biological processes and signaling pathways consistently upregulated in SMPCs from in vivo stages compared to all of the three in vitro myogenic protocols. These include both positive (CCND1 and CDK6) and negative (SPRY1 and DUSP1) regulation of cell cycle indicating more orchestrated cell cycle progression, RNA splicing (RPS26 and RBM39), WNT signaling pathways (FRZB and TCF12) and SkM development (MYF5, MSTN and VGLL2) (Figures S4-6F, S4-6G and S4-6J). On the other hand, processes and pathways consistently enriched in in vitro derived cells from all three protocols include muscle contraction (MYL1, CKB and KLHL41), cell motility (NEFL and YBX3), lipid metabolism (FDFT1, NPC2 and TSPO) and ECM (DCN and MGP) (Figures S4-6H, S4-6I and S4-6K). These findings suggest that there are fundamental differences between SMPCs derived in vivo compared to in vitro, although they might represent a similar developmental stage.

To better understand the gene regulatory networks distinguishing the different myogenic stem and progenitor cells arising during in vivo human development and derived from hPSC directed differentiation, we performed gene co-regulation analysis on our scRNA-seq data (see STAR Methods). We found co-regulated gene groups differentially expressed at distinct stages of myogenesis (Figure 4-7D) and performed GO analysis to explore the key biological processes/pathways enriched in these gene networks (Figure 4-7E). For example, gene groups 12, 8 and 21 were upregulated in the in vitro hPSC-SMPCs compared to the in vivo cells, and they were enriched for GO terms such as ECM, muscle contraction and reactive oxygen species.

Cell cycle, translation, energy metabolism as well as morphogenesis and patterning were enriched in gene groups upregulated in early embryonic- as well as hPSCSMPCs, such as groups 4, 1, 9 and 6. For gene groups upregulated in postnatal SCs (groups 10, 11, 2 and 5), enriched biological processes were in general involved in maintaining cellular homeostasis. Group 20 was found to be uniquely expressed at high levels in stage 4 SMPCs (fetal week 17-18) and was enriched for genes participating in neuromuscular junction establishment. Group 79 was expressed at a relatively stable level across prenatal development, but at low levels in hPSC-SMPCs or postnatal SCs, and this group enriched for processes such as limb morphogenesis. Next, we focused on the TFs within each of the gene groups, as they have been shown to be the master regulators in cell fate decisions in multiple systems (Oh and Jang, 2019). We found distinct TF programs that were differentially enriched in embryonic/in vitro, fetal and postnatal stages (Figures 4-7F-H). These TFs included some canonical myogenic factors such as PITX2 and SIX1 that were enriched in SMPCs from early in vivo stages and derived from hPSCs (Figure 4-7F), which is consistent with our previous findings (Figure 4-4L). However, most of these TFs are not classic myogenic genes, which indicates that maturation of myogenic progenitor and stem cells involves processes beyond the regulation of myogenic identity. Furthermore, using RNAscope coupled with IHC, we confirmed the dynamic expression patterns of selected TFs (NFIX, NFIC, KLF9 and CEBPD) in PAX7+ SMPCs/SCs in limb tissues from different embryonic, fetal and adult stages (Figure S4-7). Overall, these analyses provide potential candidate pathways and TFs to manipulate the maturation status of SMPCs in the future.

Discussion.

Myogenesis occurs from early embryonic to postnatal periods and involves myogenic as well as other supportive cell types. Yet myogenic development in human is poorly understood. Although recent work has profiled skeletal muscles using scRNA-seq (Barruet et al., 2020; De Micheli et al., 2020; Dell'Orso et al., 2019; Giordani et al., 2019; Rubenstein et al., 2020; Tabula Muris Consortium et al., 2018), the scope was limited to adult tissues. In this work, we provide a comprehensive roadmap of in vivo human limb myogenesis at the single cell level across development from as early as embryonic week 5 up to adulthood. We also interrogated in vitro hPSC myogenic differentiation from multiple published protocols. Through trajectory analysis, we showed that myogenic progenitor and stem cells from different developmental stages possess distinct gene expression profiles, and hPSC-derived SMPCs align to an in vivo stage of late embryonic to early fetal transition.

One interesting observation is the identification of a resident embryonic and fetal SkM subpopulation that expresses reduced canonical myogenic markers but increased levels of mesenchymal (e.g., PDGFRA, OGN, THY1 and DCN) and skeletal lineage genes (e.g., RUNX2, COL1A1, MGP and TNMD) (Figures 4-2, 4-3 and S4-2). When isolated and cultured in vitro, this SkM.Mesen subpopulation showed weaker myogenic fusion but stronger osteogenic differentiation capacities. These unique cells could represent a transient subset of myogenic cells existing during early myogenic development that have a higher propensity of osteogenic fate adoption. Indeed, it has been shown that human second trimester fetal SkM cells harbor myogenic and osteogenic bipotency when isolated and cultured in vitro (Castiglioni et al., 2014; Tanaka et al., 2012). A similar “side” population of SkM.Mesen was also detected from all three hPSC myogenic differentiation protocols (Figures 4-6 and S4-5). However, whether these in vitro SkM.Mesen cells are the same as those detected in vivo or a small subset drifting away from their myogenic identity due to culture conditions, needs to be further explored. It will also be interesting to fully characterize other cell types during the transition from prenatal to postnatal limb

development. Deciphering and co-opting the roles of supportive cells in vivo could increase our ability to mature and improve the functional potential of SMPCs derived from hPSCs in vitro.

Our scRNA-seq pipeline enabled us to focus on the differences of the progenitor and stem cell subpopulations within the SkM lineage across development, avoiding potential influences such as commitment status from the other myogenic subpopulations. Accordingly, we were able to confidently map the developmental trajectory of SMPCs and SCs across development and identify gene expression differences that distinguishes each of them (Figures 4-4 and S4-3). Striking differences in ECM components have been recently reported in fetal and postnatal mouse myogenic progenitor and stem cells, and they are critical for the differential regenerative capacities of cells from different developmental stages (Tierney et al., 2016). Here, we also found that ECM gene expression is one of the key features that significantly changed across human development (Figure 4-4F), suggesting ECM remodeling as a critical process in response to both the intrinsic cues and extrinsic cell-cell/cell-matrix interactions during the SMPC-to-SC transition in human development.

Metabolism is becoming a key feature of cell fate regulation in model organisms, including somite specification and mouse SC states (Koopman et al., 2014; Oginuma et al., 2017; Pala et al., 2018; Ryall, 2013; Ryall et al., 2015; Yucel et al., 2019), but has not been carefully evaluated throughout embryonic and fetal to adult development. We found that multiple genes participating in central metabolism were expressed at higher levels in early embryonic SMPCs and gradually decreased as cells transition to postnatal SCs. Consistently, negative metabolic regulators such as TXNIP, an inhibitor of glucose uptake and glycolysis and PDK4, which downregulates pyruvate entry into the mitochondrial TCA cycle, were found to be upregulated in postnatal SCs (Figure 4-4G). This gene expression pattern most likely reflects the changing metabolic demands as actively expanding SMPCs during prenatal muscle establishment transition to quiescent SCs in postnatal homeostasis, and suggests that metabolic wiring distinguishes SMPC and SC states.

Although there are multiple protocols reporting generation of SkM cells from hPSCs, the heterogeneity and dynamics of cell types present in culture and within the myogenic populations have not been adequately studied. Using scRNA-seq, we undoubtedly found myogenic as well as significant numbers of non-myogenic populations from all three representative protocols examined (Figures 4-5 and S4-5). Both HX and JC protocols employ a sequential specification through presomitic mesoderm, somite, dermomyotome and SkM, and they yielded similar cell types in the differentiation cultures. Both of these two protocols generated neural cell types including NPCs, neurons and Schwann cells. It is worth noting that the WNT activation and BMP and TGFβ inhibition approach used in these protocols have also been employed in strategies to differentiate hPSCs towards neural crest (NC) cells (Chambers et al., 2009), which are ancestors of multiple cell types including peripheral neurons, Schwann cells, SMCs and craniofacial cartilage and bone, among others (Cheung et al., 2019). Thus, it is conceivable that the neural cell types generated from these protocols might be derived from NC cells that were specified along with somite early on during differentiation. Moreover, in HX protocol, we observed SMCs and chondrogenic cells present at early time points (week 3-4) but with decreased proportions (week 5) and eventually undetectable (week 6-8) towards later time points. These populations could be derivatives from either NC or somite cells (Brent and Tabin, 2002), and the decrease of their presence might reflect the unsuitableness of the myogenic conditions to support them in long-term culture. The origin of the mesenchymal populations starting at week 4 will be interesting to explore further and might be derived from a rare population generated early on during differentiation, or from cells not well-supported in culture that drift away from their original identities. Future in vitro lineage tracing and depletion experiments will be required to delineate the origins of these non-myogenic populations and their influences on the myogenic specification efficiencies of the protocols.

This resource provides the ability for any lab performing hPSC differentiation to map the developmental identity of myogenic progenitor or stem cells. It is very striking that across all three

different protocols, SMPCs derived from hPSCs align comparably to the in vivo embryonic-to-fetal transition stage and are not equivalent to the postnatal juvenile and adult SCs (Figure 4-7 and Figure S4-6). Prolonging the length of directed differentiation (HX protocol; up to 8 weeks) does not seem to drive hPSC-SMPCs beyond this transitioning stage. Of note, even compared to the in vivo SMPCs at embryonic-to-fetal transition, hPSC-SMPCs still show fundamental differences in a wide range of biological processes (Figure S4-6). These suggest that stringent evaluation is required to correctly determine cell identity, molecular property and functional potential of myogenic derivatives across differentiation strategies from hPSCs.

To better understand the regulatory network underlying myogenic development, we performed gene co-regulation analysis and identified developmental stage specific gene group signatures. Focusing on TFs within each group, we provide key TF programs that can serve as potential maturation factors for manipulating progenitor and stem cell states across development (Figure 4-7). We found canonical myogenic specification factors such as SIX1, PITX2 and PAX7. We also found other genes known to regulate SkM, such as ID2 and TCF12 (Zhao and Hoffman, 2004) that were enriched in the embryonic and hPSC-derived SMPCs, SMAD1 (BMP signaling) (Sartori and Sandri, 2015) and PROX1 (Kivela et al., 2016; Petchey et al., 2014) that were increased from early embryonic to late fetal stage and decreased postnatally, and FOXO3 (Sanchez et al., 2014) that was specifically expressed at high levels in postnatal SCs. Interestingly, we found all of the Nuclear Factor I family members (NFIA, NFIB, NFIC and NFIX) expressed at higher levels in late fetal or postnatal stages, suggesting this TF family might play an important role in myogenic maturation. In fact, NFIX has been reported to control the switch from embryonic-to-fetal myogenesis in both mouse and zebrafish (Messina et al., 2010; Pistocchi et al., 2013; Taglietti et al., 2018). Moreover, it is worth noting that the majority of the identified network genes are not typical myogenic TFs. For example, the Kruppel Like Factor family members KLF2, KLF4 and KLF9 were all enriched in postnatal SCs. This family of genes participates in the development and homeostasis of numerous tissues (McConnell and Yang,

2010), and KLF4 is well-known of its ability in induced pluripotency by acting as a pioneer factor that facilitates large scale chromatin remodeling (Schmidt and Plath, 2012; Takahashi and Yamanaka, 2016). Along this line, we also found other chromatin modifiers differentially expressed across development, including ARID5B, NCOA1 and NR3C1. These observations suggest a model where concerted efforts from canonical myogenic TFs as well as epigenetic and chromatin regulators are required to shape the gene regulatory landscapes and drive SMPC-to-SC transition during development. This intricate interplay will also likely be required to instruct hPSCs to gain a SC-like state and maintain their cell fate identity in culture.

In summary, this work serves as a resource for advancing our knowledge of human myogenesis. It also provides a tool for molecular identification of hPSC-derived SMPCs, and targets to guide the generation of the most regenerative cells for translational applications in SkM-based regenerative medicine.

Figure Legends

Figure 4-1. scRNA-seq identifies dynamic cell types across human limb development. See also Figure S4-1. (A-H) Left panels: single cells from human biological replicates grouped by age on tSNE plots and colored by cell type. Right panels: tSNE plots showing color-scaled "Muscle.Score" (purpleto-gray: high-to-low expression). SkM populations red-circled. (I-P) Bar plots of cell type distribution in biological replicates within age groups.

Figure 4-2. Different skeletal myogenic subpopulations are present across human development. See also Figure S4-2. (A-H) Left panels: single cells classified as "SkM" within each age group on tSNE plots and colored by myogenic subtype. Right panels: dot plots of selected subtype markers. (I-K) Selected enriched GO terms from DEGs enriched in MC vs. MP (I), MP vs. MC (J) or SkM.Mesen vs. the main SkM subpopulations (MP, MB and MC) (K). (L) Heatmap of selected markers of different pathways across averaged SkM subpopulations.

Figure 4-3. Prospective isolation and in vitro differentiation potential of the SkM.Mesen subpopulation in human embryonic and fetal limbs. See also Figure S4-2. (A and B) IHC staining of PAX7 and PDGFRA in human limb sections. Images in (B) show enlarged area of the boxed region in (A). Cross (x), PAX7+PDGFRA+; arrow, PAX7-PDGFRA+; arrowhead, PAX7+PDGFRA-. Scalebars represent 50 (A) or 20 (B) μm. Representative images are shown from 4 week 7-17 human embryonic and fetal limbs. (C) tSNE plots of CDH15 (purple-to-gray: high-to-low expression). (D) Flow cytometry analysis of CDH15 and PDGFRA co-expression. Representative FACS plots are shown from 3-4 samples for each age group. (E) Freshly sorted CDH15 (15) and PDGFRA (P) subpopulations were subjected to qRT-PCR for myogenic, osteogenic as well as mesenchymal and ECM gene expression. (F-I) Sorted 15+P- and 15+P+ cells subjected to myotube fusion followed by IF of MyHC (F) and qRT-PCR of myogenic commitment genes (G), or osteogenic conditions followed by Alizarin Red S staining (H) and qRT-

PCR of osteogenic differentiation markers (I). Scalebars in (F) represent 100 μm. Data shown in (E-I) are representative of 2-3 human fetal limbs. Data of qRT-PCR are normalized to RPL13A as mean+SD of technical triplicates.

Figure 4-4. Skeletal myogenic progenitor and stem cells display dynamic gene expression signatures across human development. See also Figure S4-3. (A) DM plot of single cells of in vivo SMPCs and SCs computationally clustered into 5 major stages. (B) Proportions of cells from each biological sample assigned to each computational stage. (C-L) Dot plots of selected markers for each labelled category. Pst, postnatal (including juvenile and adult). (M) Venn diagram of upregulated genes in stage 5 SCs compared to each stage of SMPCs from stage 1-4. (N) Selected enriched pathways from the 140 genes (M) commonly upregulated in SCs compared to each stage of SMPCs.

Figure 4-5. scRNA-seq identifies skeletal myogenic populations as well as other cell types during hPSC differentiation. See also Figures S4-4 and S4-5. (A-E) From left to right. First panels: single cells from hPSC-derived samples using HX protocol grouped by differentiation time on tSNE plots and colored by cell type. Second panels: tSNE plots showing color-scaled "Muscle.Score" (purple-to-gray: high-to-low expression). The tiny SkM population at week 3 is red-circled. Third panels: bar plots of cell type distribution in enriched or unenriched samples at similar differentiation time points. Fourth panels: tSNE plots of selected cell type markers (purple-to-gray: high-to-low expression). Small populations are boxed for easy visualization.

Figure 4-6. scRNA-seq identifies myogenic subpopulations during hPSC myogenic differentiation. See also Figures S4-4 and S4-5. (A-D) Left panels: single cells classified as "SkM" derived using HX protocol at similar time points on tSNE plots and colored by myogenic subtype. Middle panels: dot plots of selected subtype markers. Right panels: bar plots of subtype

distribution in enriched or unenriched samples at similar differentiation time points. (E) DEGs upregulated in SkM.Mesen vs. the main SkM subpopulations (MP, MB and MC) from three hPSC differentiation protocols as well as human fetal week 9 samples were subjected to GO enrichment analysis. Heatmap clustering of the top 20 shared GO groups based on enrichment p values.

Figure 4-7. In vitro hPSC-SMPCs align to an embryonic-to-fetal transition stage of in vivo human myogenesis. See also Figures S4-6 and S4-7. (A) DM plot of single cells of in vivo and in vitro (HX protocol) SMPCs and SCs. (B) DM plots highlighting cells (in red) from individual in vivo or in vitro (HX protocol) stages. (C) Ridge plot of developmental score ("Dev.Score") distribution across in vivo or in vitro (HX protocol) stages. (D) Heatmap of selected co-regulated gene groups (gene number > 50) across averaged in vivo or in vitro (HX protocol) stages. (E) Two selected enriched GO terms from each gene group are plotted and color-coded. (F-H) Dot plots of selected TFs differentially enriched in embryonic/in vitro, fetal and postnatal stages.

Figures

Figure 4-1 – scRNA-seq identifies dynamic cell types across human limb development

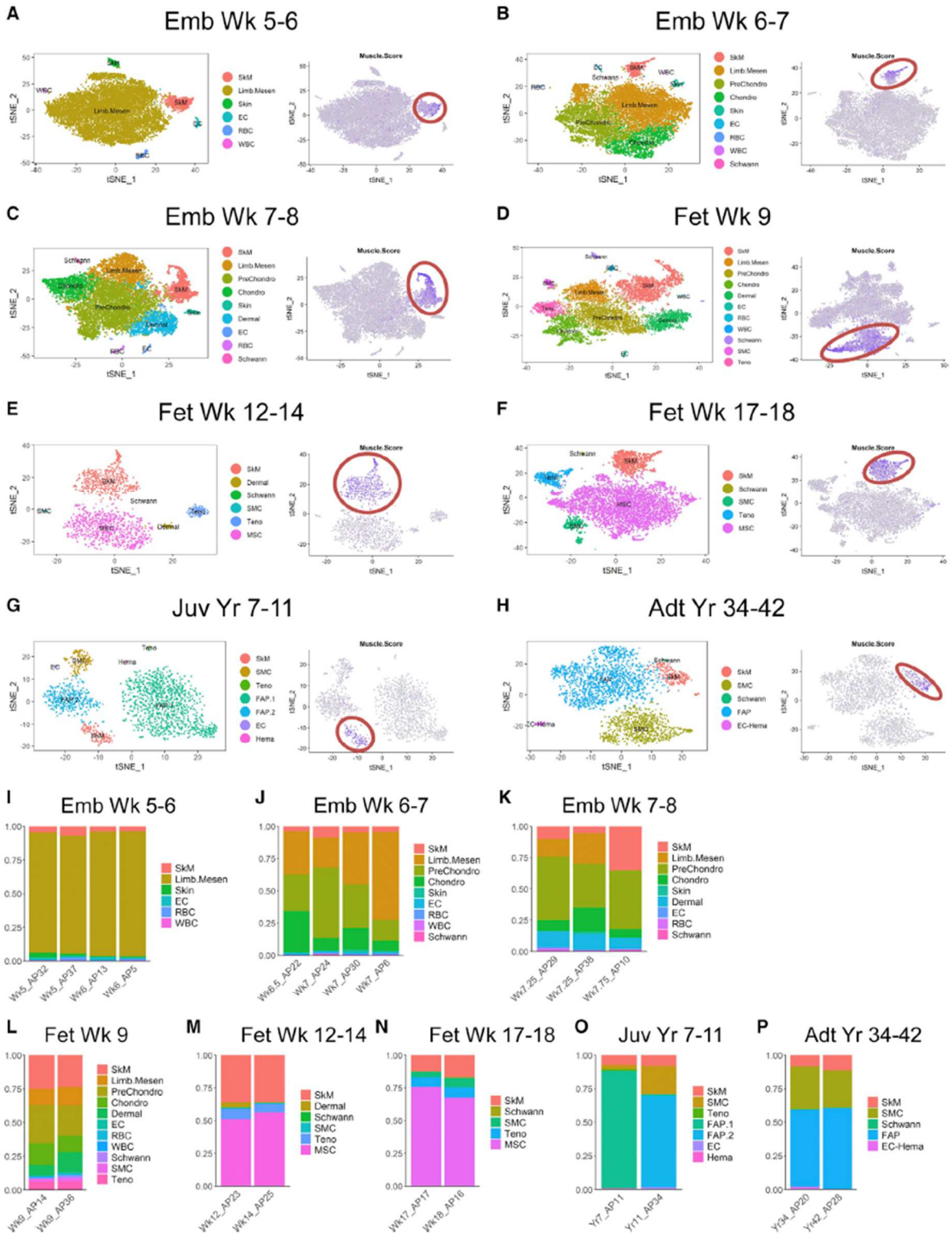


Figure 4-2 – Different skeletal myogenic subpopulations are present across human Development

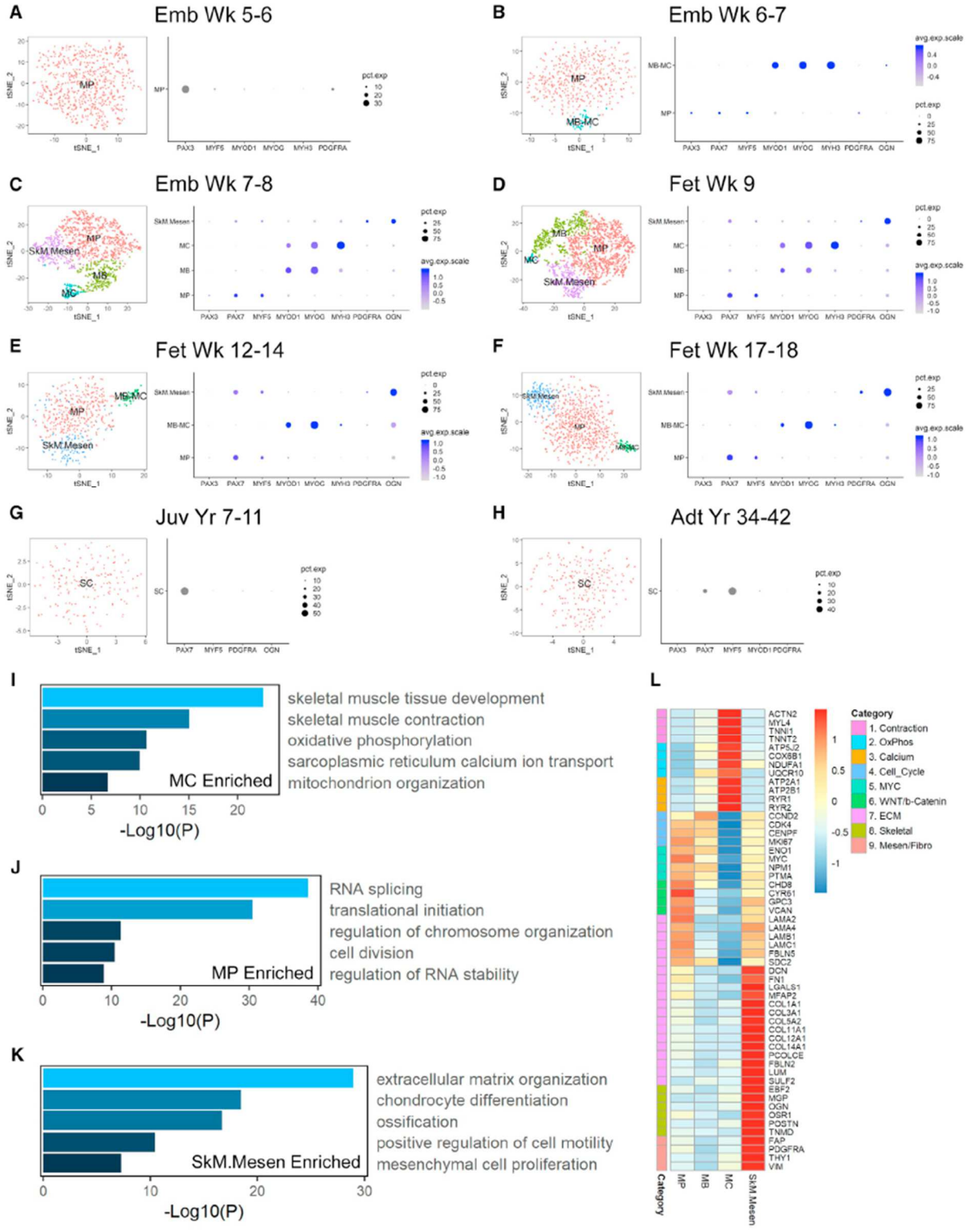


Figure 4-3 – Prospective isolation and in vitro differentiation potential of the SkM.Mesen subpopulation in human embryonic and fetal limbs.

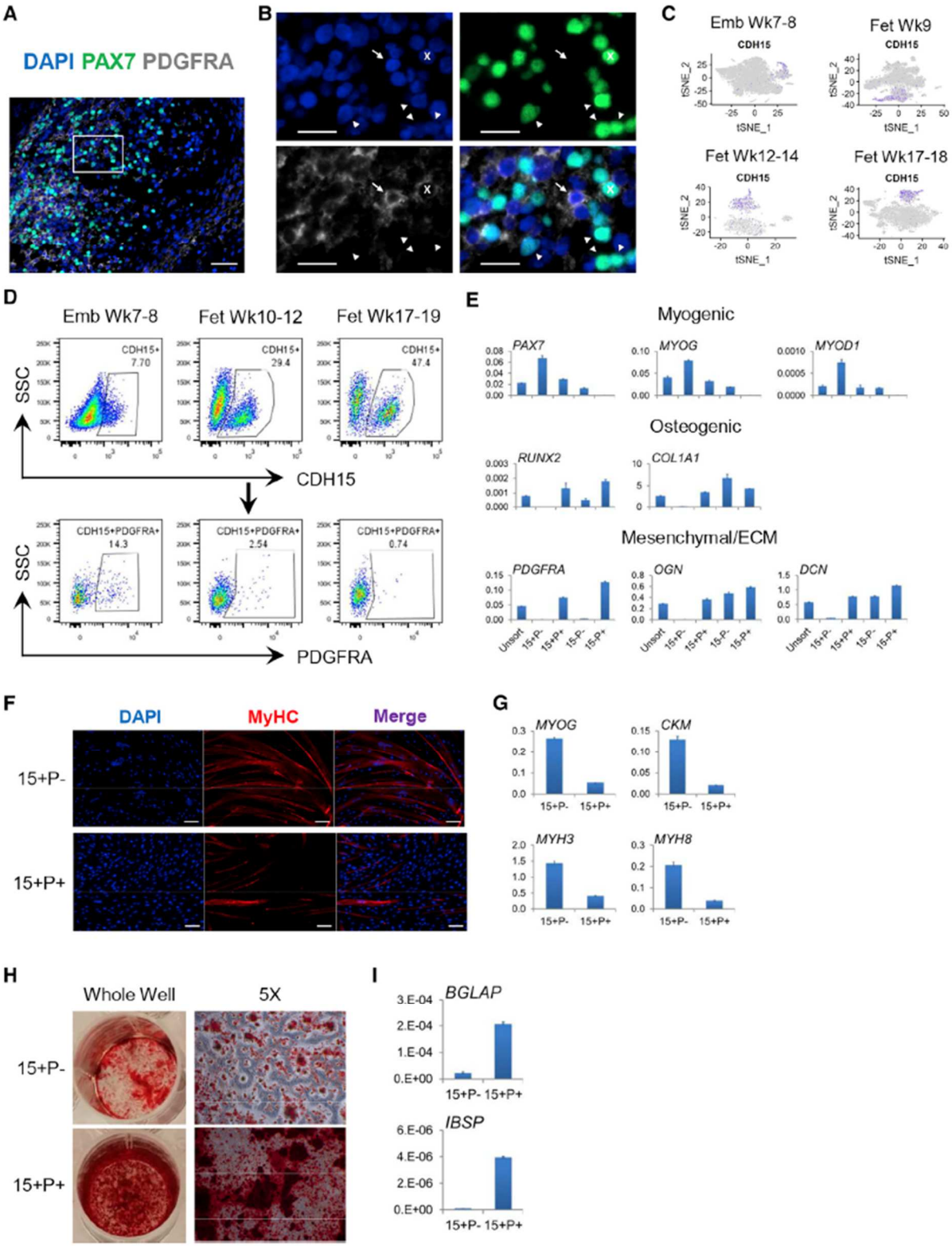


Figure 4-4 – Skeletal myogenic progenitor and stem cells display dynamic gene expression signatures across human development

Figure 4-5 – scRNA-seq identifies skeletal myogenic populations as well as other cell types during hPSC differentiation

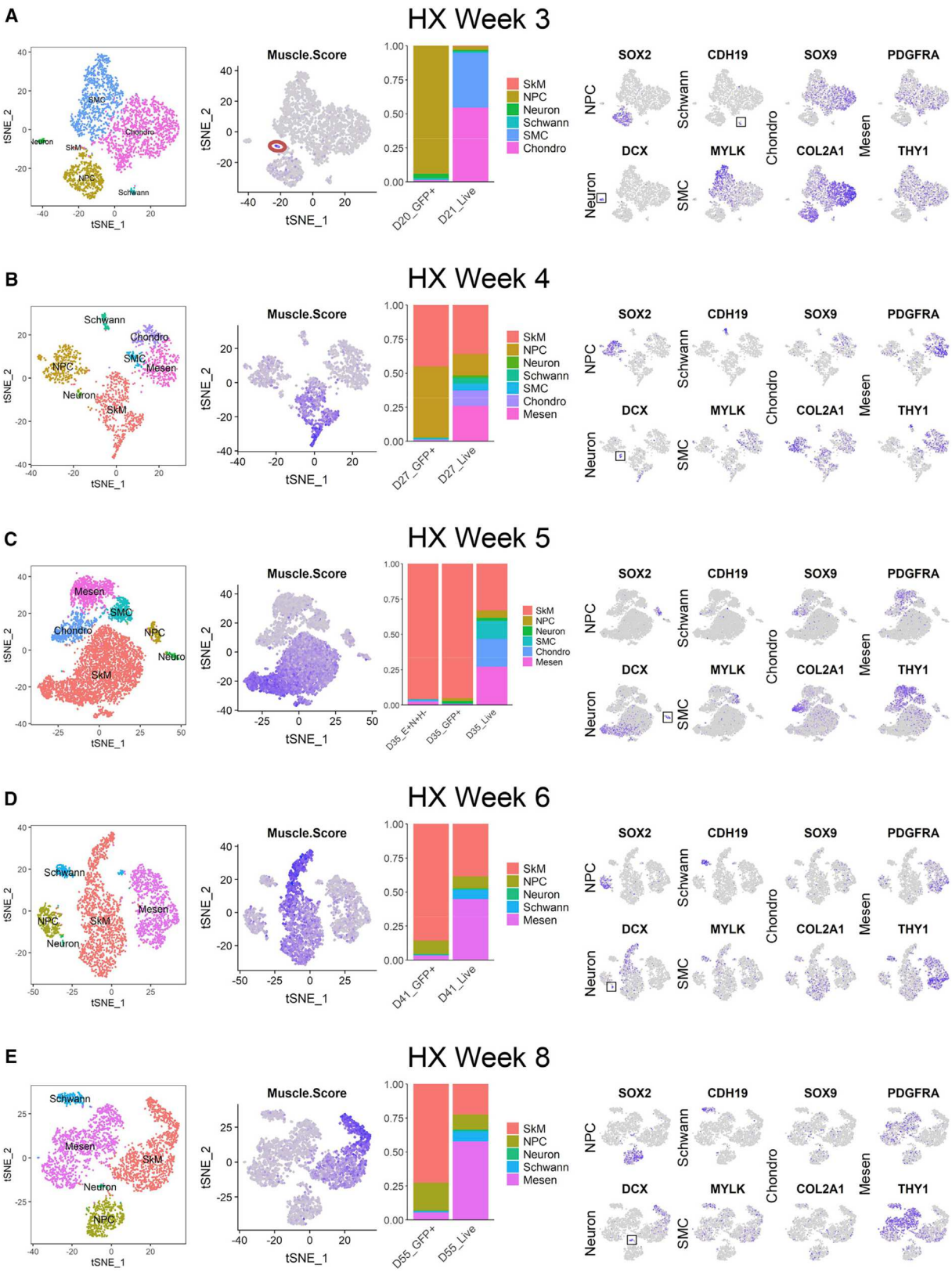


Figure 4-6 – scRNA-seq identifies myogenic subpopulations during hPSC myogenic differentiation

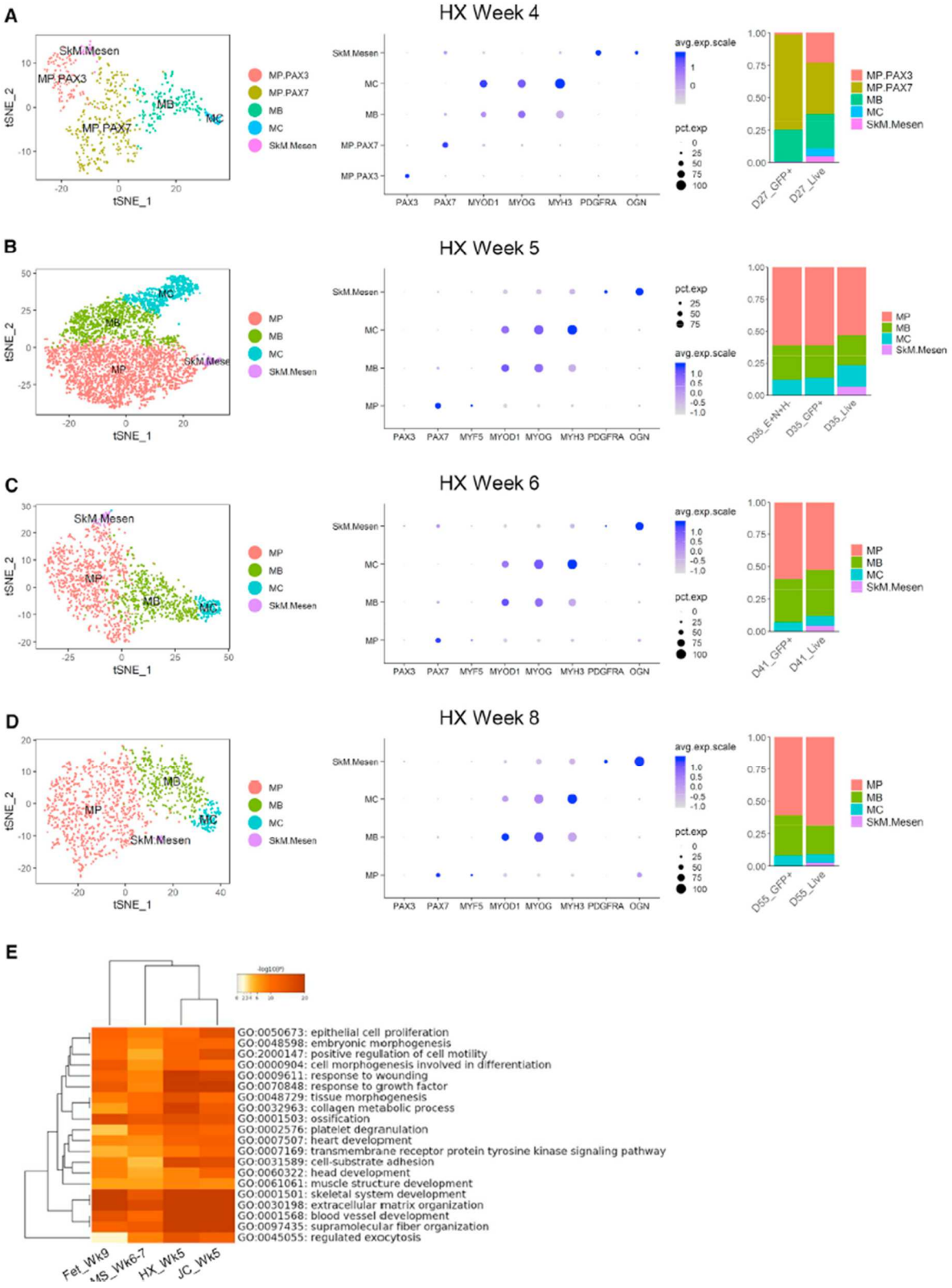


Figure 4-7 – *In vitro* hPSC-SMPCs align to an embryonic-to-fetal transition stage of *in vivo* human myogenesis.

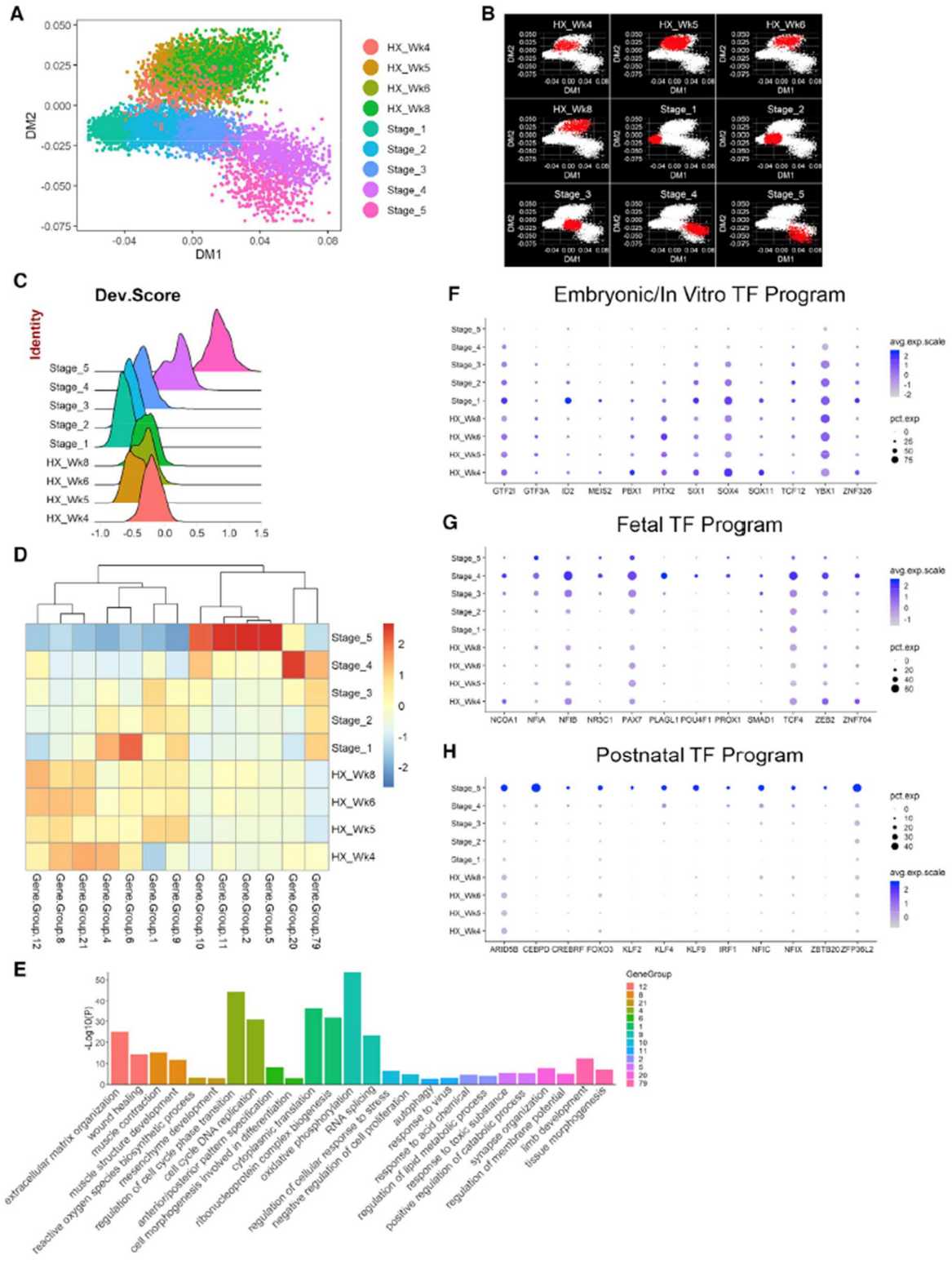


Figure S4-1 – Cell types present in limbs and skeletal muscle tissues at different human developmental stages

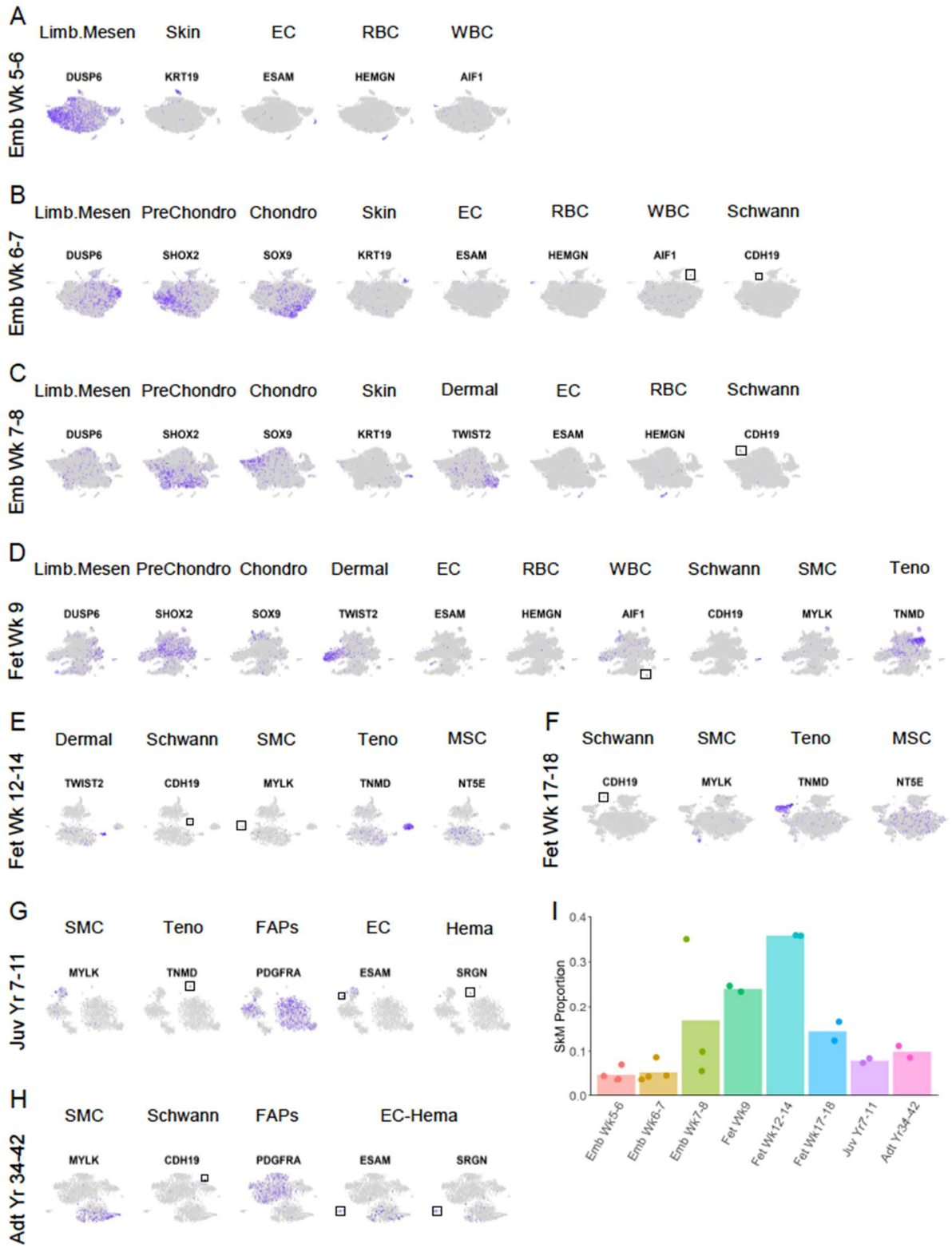


Figure S4-2 – Characterization of skeletal myogenic subpopulations in human fetal limbs

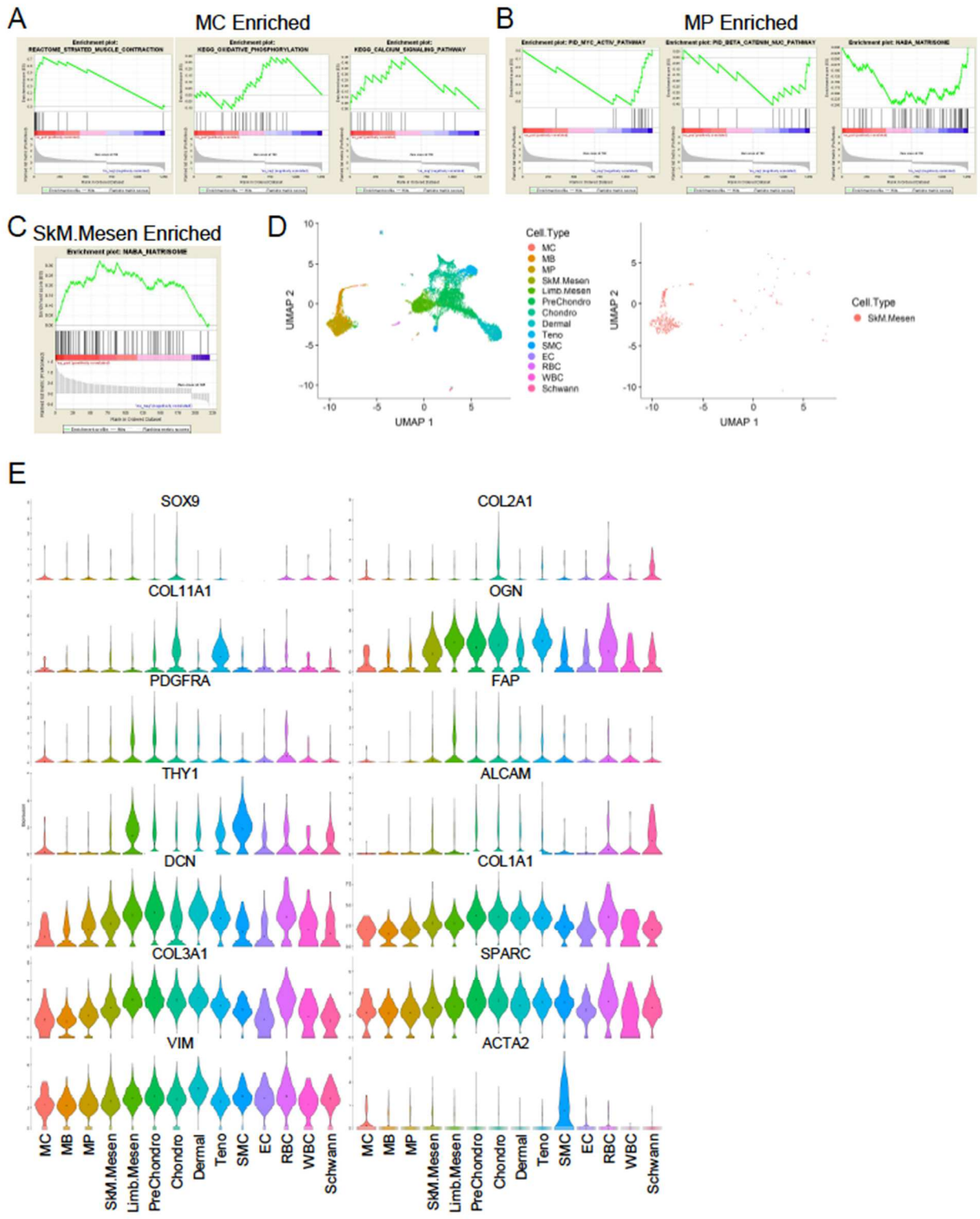


Figure S4-3 – Distinct SMPC and SC populations across human development and isolation of myogenic cells from early human embryonic limbs

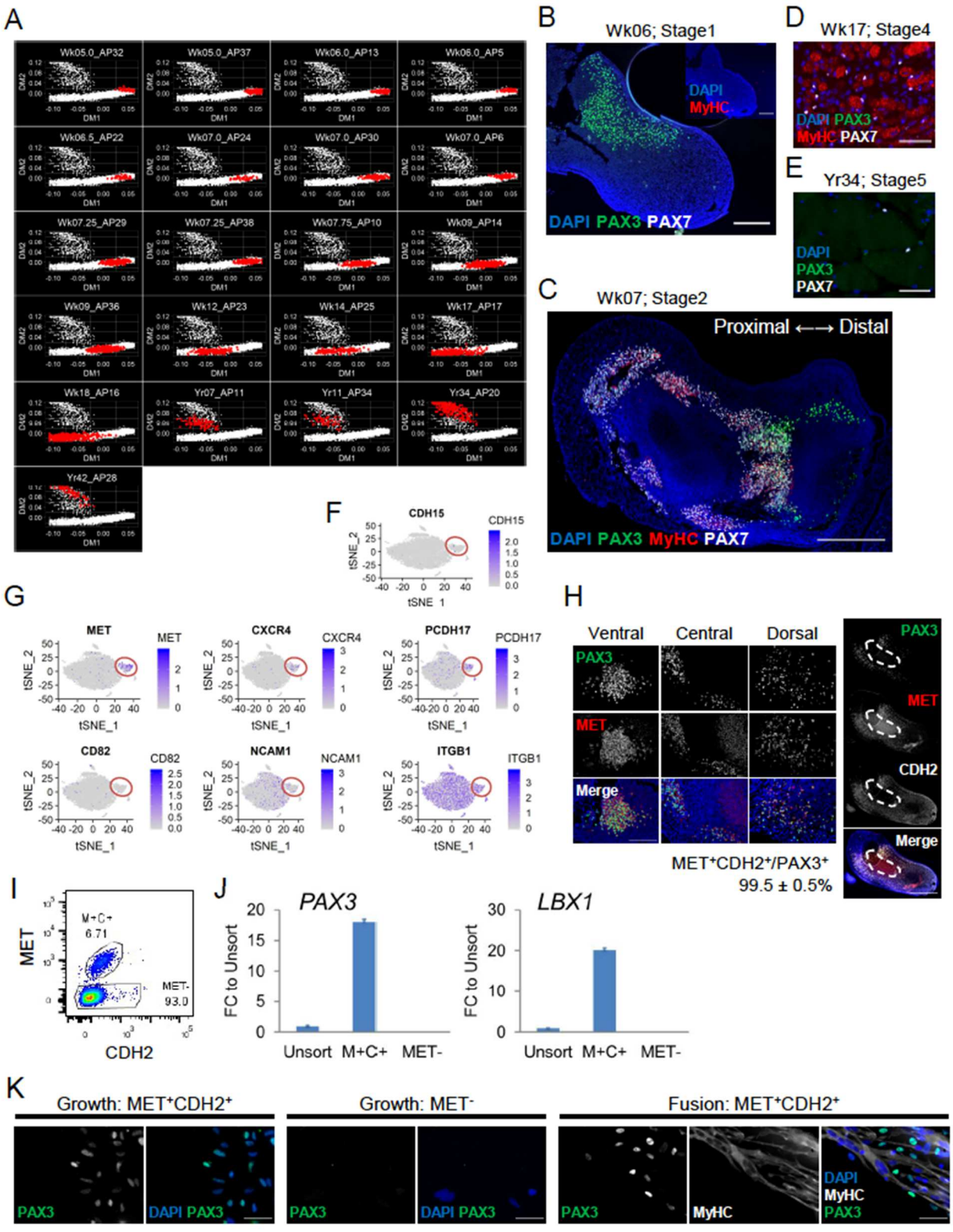


Figure S4-4 – Construction of the PAX7-GFP reporter cell lines

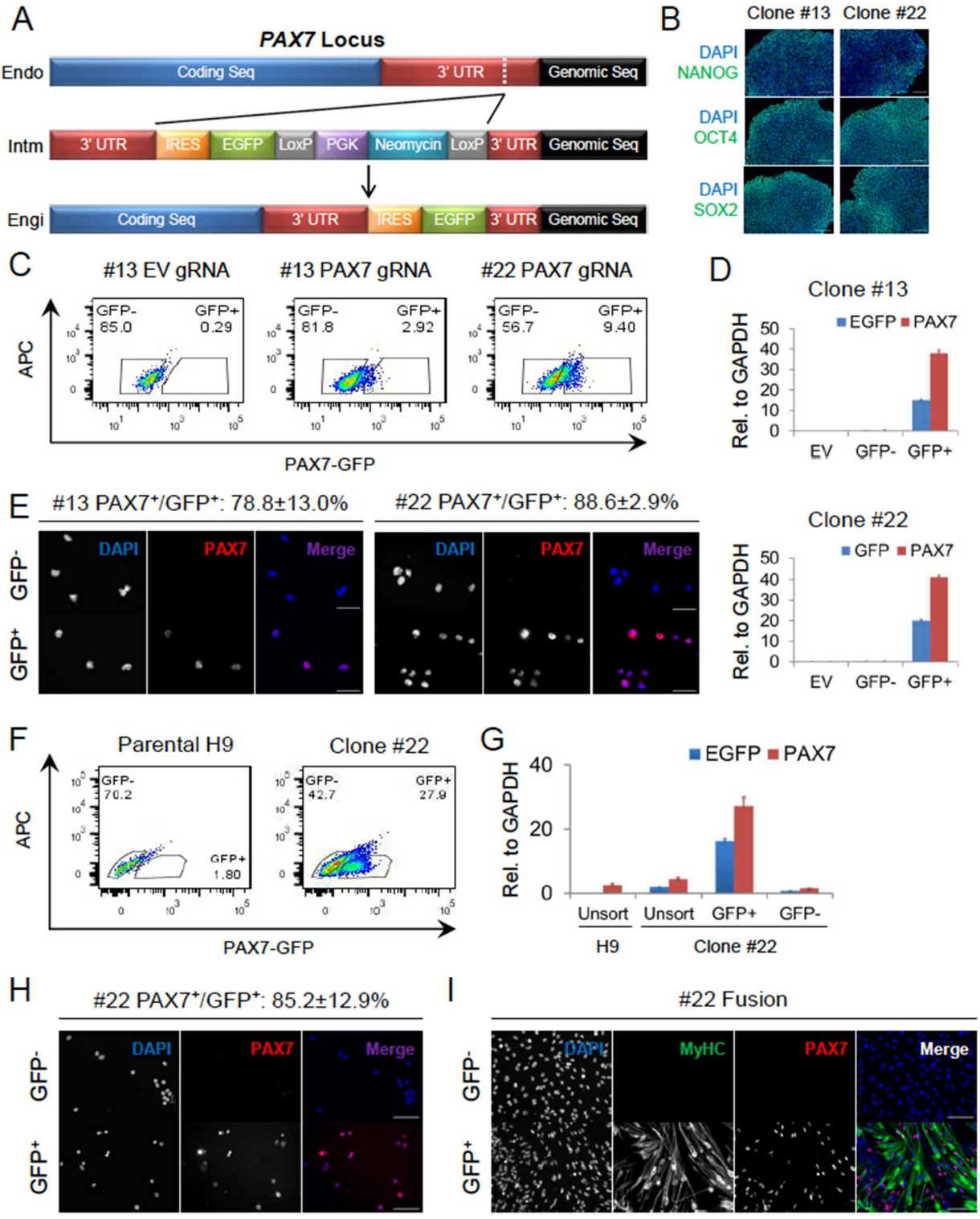


Figure S4-5 – scRNA-seq reveals heterogeneous cell types and skeletal muscle subpopulations from additional hPSC myogenic differentiation protocols

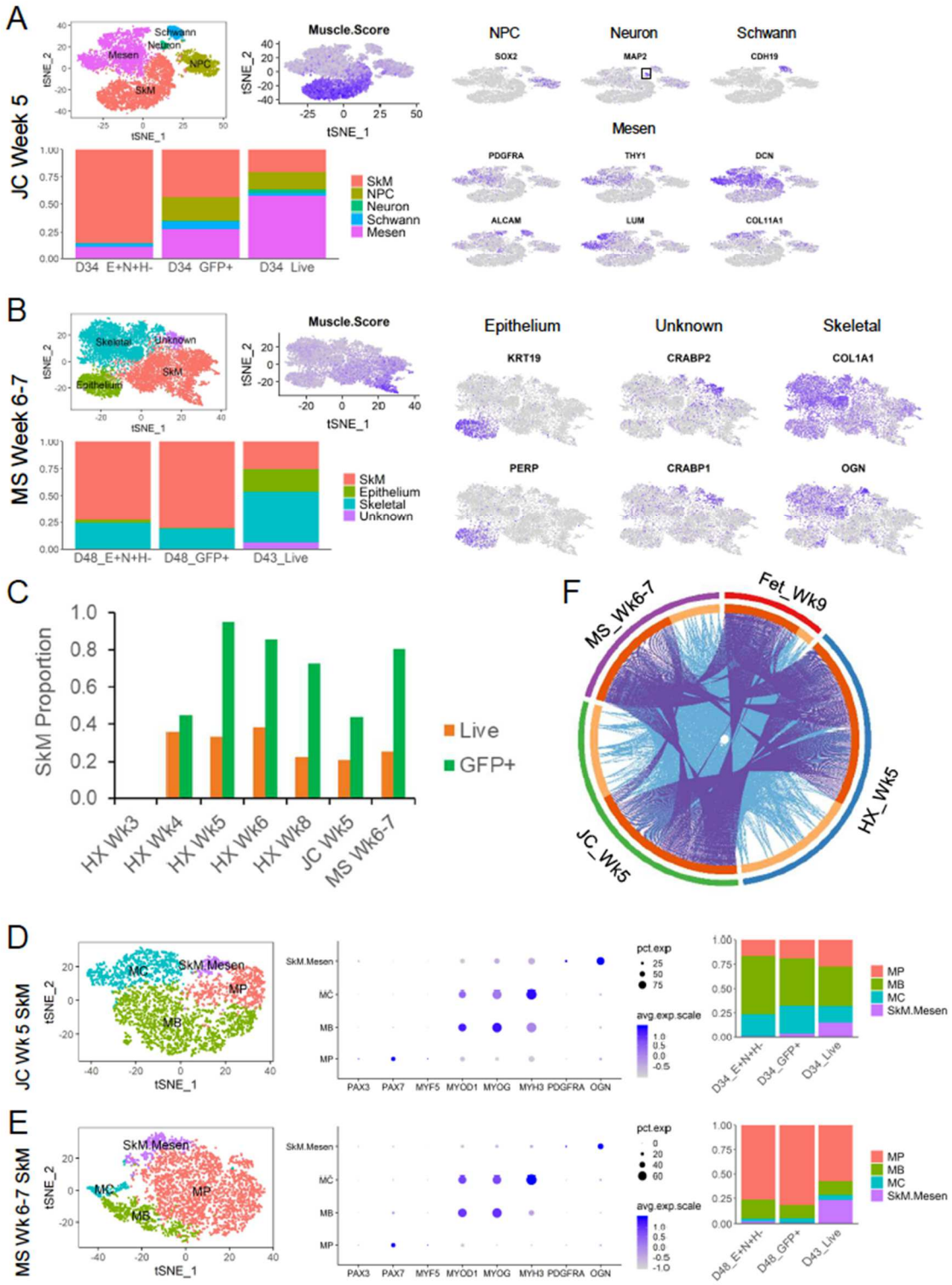


Figure S4-6 – *In vitro* SMPCs derived from multiple hPSC myogenic differentiation protocols are different from *in vivo* human myogenic progenitor cells during embryonic-to-fetal transition

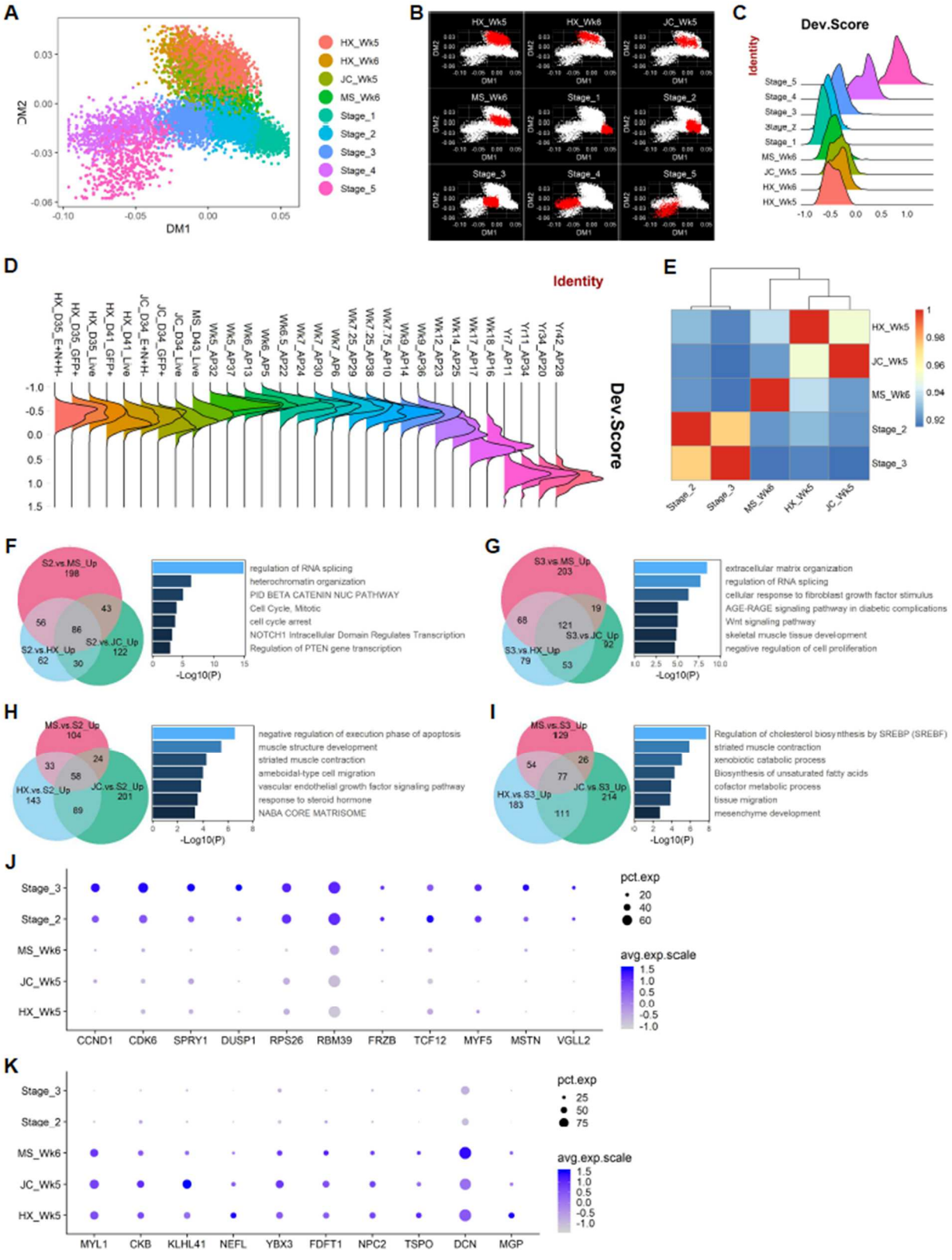
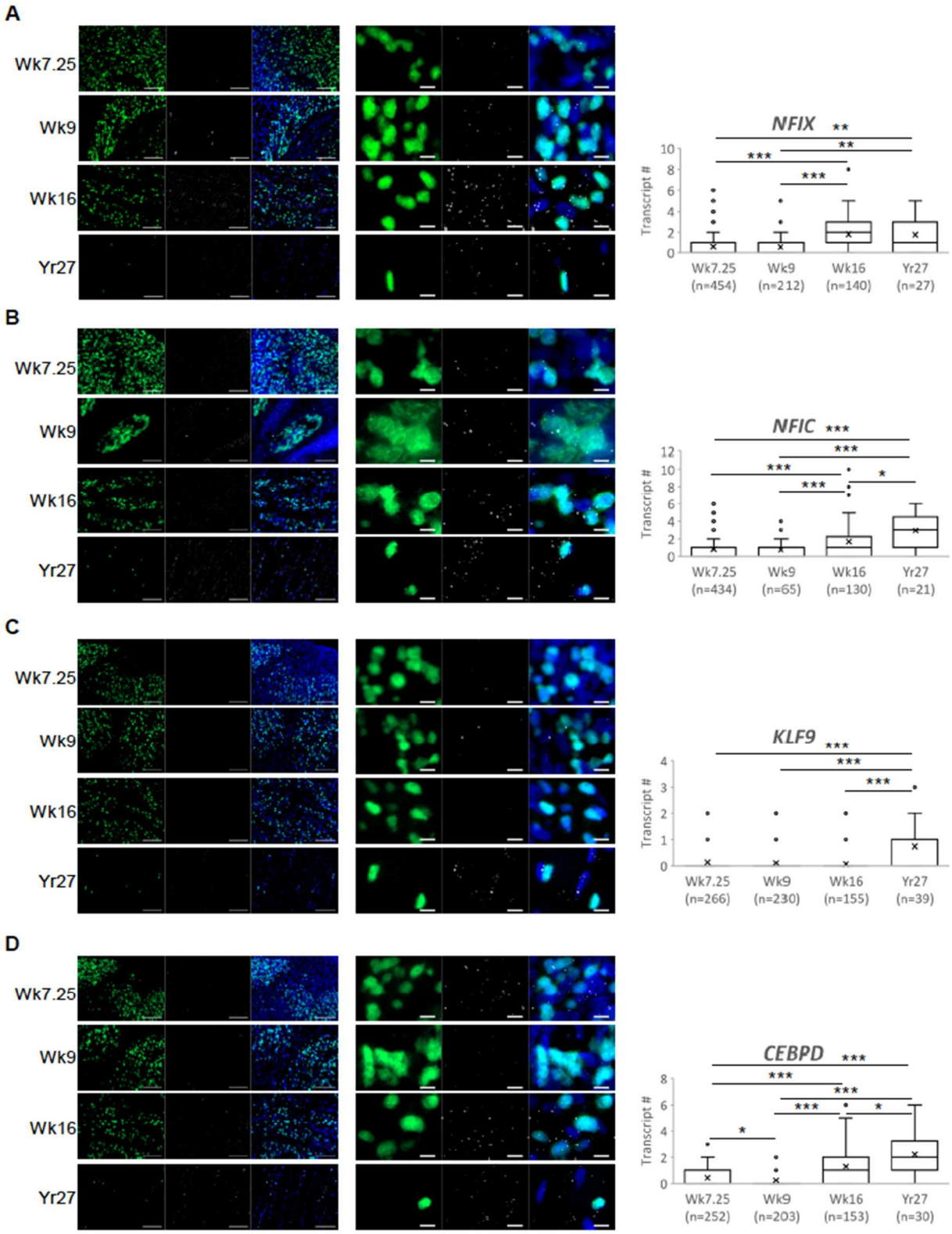


Figure S4-7 – Validation of expression of TFs differentially expressed across human developmental stages



Method Details

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, April D. Pyle (apyle@mednet.ucla.edu).

Materials availability

Plasmids generated in this study will be provided upon request.

Data and code availability

Both raw sequencing reads and processed digital gene expression (DGE) matrices of scRNAseq datasets are deposited at NCBI GEO with accession number GSE147457. Interactive scRNA-seq data exploration can be accessed at skeletal-muscle.cells.ucsc.edu or aprilpylelab.com/datasets. General codes for computational analysis follow the instructions of the respective software and customized modifications will be available upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human tissues

Human tissues of 9 weeks of gestation or younger were obtained from electively aborted embryos and fetuses following informed consent and de-identification in accordance with institutional guidelines, which was approved by the local research ethics committee of the University of Tübingen (#312/2016BO1 and #634/2017BO1). Human tissues of 12-18 weeks of gestation were obtained from the University of California Los Angeles (UCLA) Center for AIDS Research (CFAR) Gene and Cellular Therapy Core using institutional review board (IRB)-approved de-identified and consented electively aborted human fetuses. Skeletal muscles from the 7 years old human juvenile subject were obtained from leftover tissues from surgical procedures approved by the

UCLA institutional IRB, with patient consent and de-identification. Skeletal muscles from the 11 years old human juvenile subject and the two adult human subjects were obtained from donor autopsy provided by the National Disease Research Interchange (NDRI) with de-identification. Use of human tissues was IRB exempt by the UCLA Office of the Human Research Protection Program (IRB #15-000959).

Cell lines

The H9 human embryonic stem cells (WA09; WiCell Research Institute) are registered in the NIH Human Embryonic Stem Cell Registry with the Approval Number: NIH hESC-10-0062 (https://grants.nih.gov/stem_cells/registry/current.htm?id=414). The PAX7-GFP reporter cell lines are derived from the H9 cells.

Cell preparation for single cell RNA-sequencing

Embryonic week 7.25 and younger samples

Whole limbs were washed with wash buffer consisting of DMEM/F12, 10% fetal bovine serum (FBS), 1% Penicillin-Streptomycin (P/S) and 0.1% Amphotericin. Tissues were then mechanically chopped into small pieces at room temperature (RT) in digestion buffer consisting of wash buffer supplemented with 2 mg/ml of Collagenase IV and 1 mg/ml of Dispase II. Chopped tissues were further incubated in digestion buffer on a shaker at 37°C for 10-20 minutes with intermittent trituration. Digestion was stopped by adding surplus amount of Dropseq buffer consisting of phosphate-buffered saline (PBS) supplemented with 0.01% bovine serum albumin (BSA). Digested tissues were filtered twice through 40 μm cell strainers, spun down and resuspended in small volumes of Drop-seq buffer. Cell number was counted and resuspended cells were kept on ice until subjected to the Drop-seq flow procedures.

Embryonic week 7.75 and fetal week 9 samples

Whole limbs excluding feet were washed with wash buffer and then mechanically chopped into small pieces at RT in digestion buffer consisting of wash buffer supplemented with 2 mg/ml of Collagenase II and 1 mg/ml of Dispase II. Chopped tissues were further incubated in digestion buffer on a shaker at 37°C for 20-25 minutes with intermittent trituration. Digestion was stopped by adding surplus amount of Drop-seq buffer. Digested tissues were filtered twice through 40 µm cell strainers, spun down and resuspended in small volumes of Drop-seq buffer. Cell number was counted and resuspended cells were kept on ice until subjected to the Drop-seq flow procedures.

Fetal week 12-18 samples

Skeletal muscles from whole limbs were separated from bones and skin. Muscles were washed with wash buffer and then mechanically chopped into small pieces at RT in digestion buffer consisting of wash buffer supplemented with 2 mg/ml of Collagenase II, 1 mg/ml of Dispase II and 50 µg/ml of DNase I. Chopped tissues were further incubated in digestion buffer on a shaker at 37°C for 20-25 minutes with intermittent trituration. Digestion was stopped by adding surplus amount of fluorescence-activated cell sorting (FACS) buffer consisting of PBS supplemented with 1% FBS and 1% P/S. Digested tissues were filtered through 100 µm cell strainers and spun down. Cell pellets were resuspended in FACS buffer, filtered through 70 µm cell strainers, spun down and resuspended again in small volumes of FACS buffer. Cells were then incubated on ice with antibodies against CD31, CD45 and CD235a. Stained cells were sorted on BD FACSAria sorters to collect the DAPI-/CD31-/CD45-/CD235a- fraction (live and depletion of the endothelial and hematopoietic lineages). Sorted cells were washed with Dropseq buffer, spun down and resuspended in small volumes of Drop-seq buffer. Cell number was counted and resuspended cells were kept on ice until subjected to the Drop-seq flow procedures.

Postnatal juvenile and adult samples

Skeletal muscles from autopsy or surgical procedures were washed with wash buffer and then mechanically chopped into small pieces at RT in primary digestion buffer consisting of wash buffer supplemented with 2 mg/ml of Collagenase II. Chopped tissues were further incubated in primary digestion buffer on a shaker at 37°C for 10-20 minutes with intermittent trituration. Primary digestion was stopped by adding surplus amount of wash buffer and tissues spun down. Next, supernatant was removed and tissues were resuspended in secondary digestion buffer consisting of wash buffer supplemented with 7 mg/ml of Collagenase D, 1.5 mg/ml of Dispase II and 50 µg/ml of DNase I. Tissues were further digested on a shaker at 37°C for 15-20 minutes with intermittent trituration. Secondary digestion was stopped by adding surplus amount of FACS buffer. Digested tissues were filtered through 100 µm cell strainers and spun down. Cell pellets were resuspended in FACS buffer, filtered through 70 µm cell strainers, spun down and resuspended again in small volumes of FACS buffer. Cells were then incubated on ice with antibodies against CD31, CD45 and CD235a. Stained cells were sorted on BD FACSAria sorters to collect the DAPI-/CD31-/CD45-/CD235a- fraction (live and depletion of the endothelial and hematopoietic lineages). Sorted cells were washed with Drop-seq buffer, spun down and resuspended in small volumes of Drop-seq buffer. Cell number was counted and resuspended cells were kept on ice until subjected to the Drop-seq flow procedures.

Human pluripotent stem cell-derived samples

At the end of directed differentiation, cells were dissociated by 2 mg/ml of Collagenase IV for about 5 min, followed by TrypLE Express for another 5-7 minutes. Dissociation was stopped by adding surplus amount of FACS buffer and dissociated cells were filtered sequentially through 100 and 70 µm cell strainers. Cells were spun down and resuspended in small volumes of FACS buffer. For some samples, cells were incubated on ice with antibodies against ERBB3, NGFR and HNK1. Cells were sorted on BD FACSAria sorters to collect the total live (DAPI-), DAPI-/ERBB3+/NGFR+/HNK1- or DAPI-/GFP+ fractions. Sorted cells were washed with Drop-seq

buffer, spun down and resuspended in small volumes of Drop-seq buffer. Cell number was counted and resuspended cells were kept on ice until subjected to the Drop-seq flow procedures.

Cell capture and library construction for single cell RNA-sequencing

Prepared single cell solutions were subjected to single cell capture and droplet formation following instructions in the online Drop-seq protocol v.3.1 (http://mccarrolllab.org/download/905/) and those published in the original Drop-seq paper (Macosko et al., 2015). In brief, cells at 150,000 cells/ml, barcoded beads at 175,000 beads/ml and droplet generation oil were co-flowed at a rate of 4, 4, and 15 ml/hour, respectively, in a PDMS microfluidics chip to generate oil droplets containing beads and lysed cells. Post flow, droplets were breakdown and reverse transcription performed. Complementary DNA was PCR amplified, magnetically cleaned up and subjected to tagmentation and sequencing library construction. Prepared libraries were cleaned up and sequenced via Illumina HiSeq2500, HiSeq4000 or NovaSeq.

Human PSC maintenance

The parental H9 cells and engineered PAX7-GFP reporter cells were maintained on Matrigelcoated tissue culture plates in mTeSR1 medium. Cells were fed with fresh medium every day and passaged with 0.5 mM of EDTA every 4-6 days.

Human PSC skeletal myogenic directed differentiation

HX protocol

Differentiation was performed following procedures published by Xi, et al. (Xi et al., 2017) with minor modifications. Briefly, on day -1 hPSC colonies were dissociated into single cells with TrypLE Express and seeded on Matrigel-coated tissue culture plates at 12,500-25,000 cells/cm2 in mTeSR1 medium containing 10 μM of Y-27632. Differentiation was initiated the next day (day

0) when medium was switched to DMEM/F12 medium containing 1% ITS-G, 0.5% P/S and 3 μM of CHIR99021 (CHIR) for 2 days. On day 2, cells were switched to DMEM/F12 medium containing 1% ITS-G, 0.5% P/S, 200 nM of LDN193189 (LDN) and 10 μM of SB431542 (SB) for another 2 days. On day 4, LDN and SB from the previous medium were replaced with 10 μM of CHIR and 20 ng/ml of FGF2 for 2 days. On day 6, medium was switched to DMEM medium containing 0.5% P/S, 15% KSR (Knockout Serum Replacement), 10 ng/ml of HGF and 2 ng/ml of IGF1 until the end of differentiation. Cells were fed with fresh medium every day until day 6 and every other day thereafter.

JC protocol

Differentiation was performed following procedures published by Chal, et al. (Chal et al., 2015; Chal et al., 2016). Briefly, on day -1 hPSC colonies were dissociated into single cells with TrypLE Express and seeded on Matrigel-coated tissue culture plates at 15,000 cells/cm2 in mTeSR1 medium containing 10 μM of Y-27632. Differentiation was initiated on day 0 by switching to a medium containing DMEM/F12, 1% ITS-G, 1% nonessential amino acids (NEAA) and 0.5% P/S supplemented with 3 μM of CHIR and 0.5 μM of LDN. On day 3, 20 ng/ml of FGF2 was added to the differentiation medium for an additional 3 days. On day 6, medium was changed to a medium containing DMEM/F12, 15% KSR, 1% NEAA, 0.5% P/S and 0.1 mM of 2-mercaptoethanol supplemented with 10 ng/ml of HGF, 2 ng/ml of IGF1, 20 ng/ml of FGF2 and 0.5 μM of LDN for 2 days. On day 8, medium was changed to DMEM/F12 containing 15% KSR, 1% NEAA, 0.5% P/S and 0.1 mM of 2-mercaptoethanol supplemented with 2 ng/ml of IGF1. On day 12 until the end of differentiation, 10 ng/ml of HGF was added to the previous medium. Cells were fed with fresh medium every day until day 12 and every other day thereafter.

MS protocol

Differentiation was performed following procedures published by Shelton, et al. (Shelton et al., 2014) with minor modifications (Hicks et al., 2018). Briefly, on day -1 hPSC colonies were dissociated into single cells with TrypLE Express and seeded on Matrigel-coated tissue culture plates at 37,500 cells/cm2 in mTeSR1 medium containing 10 μM of Y-27632. On the next day (day 0), differentiation was initiated by switching to the E6 medium containing 0.5% P/S supplemented with 10 μM of CHIR for 2 days. On day 2, cells were switched to E6 medium containing 0.5% P/S for 10 days. On day 12, medium was changed to StemPro-34 medium supplemented with 0.5% P/S, 2 mM of L-glutamine, 0.45 mM of 1-thioglycerol, 11 μg/ml of human transferrin and 5 ng/ml of FGF2 for 6 to 8 days. On around day 20, medium was switched to E6 medium containing 0.5% P/S for about 10-15 days with the medium during the last 5-7 days of this period supplemented with 10 ng/ml of IGF1. From around day 30-35, medium was changed to DMEM/F12 containing 1.2% N2 supplement, 1% ITS-G, 0.5% P/S and 10 ng/ml of IGF1 for about 5 days. From then on cells were cultured in the same medium supplemented with 3 μM of SB until the end of differentiation. Cells were fed with fresh medium every day.

PAX7-GFP reporter cell construction

Candidate guide RNAs (gRNAs) targeting the 3' untranslated region (UTR) of PAX7 transcript variant 3, which is conserved across species, were designed using the online tool at crispr.mit.edu. The targeting region was limited to the last 1600 bp of the 3' UTR to exclude the potential human miR206/miR1-1/miR1-2 binding sites predicted by miRbase (http://www.mirbase.org/), as the mouse counterparts of these miRNAs have been shown to regulate Pax7 expression (Chen et al., 2010). Next, each of the candidate gRNAs in both the regular 20 bp form and short 17 bp form (which has been reported to increase specificity by (Fu et al., 2014)) was cloned into a gRNA cloning vector (Addgene, #41824; (Mali et al., 2013)) using the Gibson Assembly Cloning Kit following manufacturer's instructions. The final gRNA used was selected based on the highest cleavage efficiencies in hPSCs when a hCas9 plasmid (Addgene,

#41815; (Mali et al., 2013)) was co-expressed. The PAX7 targeting homology arms were then PCR amplified from the H9 cell genomic DNA based on the gRNA targeting region selected. For homologous recombination (HR) vector, the Oct4-IRES-eGFP-PGK-Neo plasmid (Addgene, #48681; (Yang et al., 2013)) was used and the Oct4 targeting homology arms were replaced by the ones targeting PAX7 using the Gibson Assembly Cloning Kit following manufacturer's instructions. Plasmids encoding gRNA, hCas9 and the HR construct (2 µg each) were nucleofected together into 800,000 H9 cells following the Lonza Amaxa 4D guideline with program CA-137. Four days post nucleofection, neomycin/G418 selection at 50 µg/ml was applied for 5 days and then increased to 100 µg/ml afterwards. Individual resistant clones were expanded and genotyped to confirm correct insertion of the reporter cassette. One of the confirmed clones was incubated with recombinant TAT-Cre protein (a gift from Dr. William Pastor, McGill University) to remove the PGK-neomycin cassette between the LoxP sites. Single cell clones were selected, expanded and confirmed by genotyping and they regained sensitivity to neomycin/G418. Two of the final clones, #13 and #22 were used for downstream functional validation and clone #22 were used for directed differentiation for scRNA-seq experiments. Both clones were confirmed to express the pluripotency markers (NANOG, OCT4 and SOX2) by immunofluorescence staining. They were also examined and showed normal karyotypes.

PAX7-GFP reporter validation

Method of dCas9-VPR

Four gRNAs targeted to the PAX7 promoter region (Murmann et al., 2000) were designed using crispr.mit.edu. Each gRNA was cloned individually into the gRNA cloning vector (Addgene, #41824) similarly to previously described (Mali et al., 2013). In brief, 50 ng AflII-digested empty gRNA plasmid was mixed with 3.8 ng of the forward and reverse oligos and combined using the NEBuilder HiFi DNA Assembly Master Mix according to the manufacturer's instructions. To activate endogenous PAX7 locus, plasmids encoding for all 4 gRNAs along with one for dCas9-

VPR (Addgene, #63798; (Chavez et al., 2015)) were co-transfected using ViaFect according to manufacturer's instructions. To limit nucleofection-related toxicity and increase transfection efficiency, H9 cells were dissociated into single cells with TrypLE Express and seeded on Matrigel-coated tissue culture plates at 25,000 cells/cm2 in mTeSR1 medium containing 10 μM of Y-27632. The next day medium was changed to DMEM/F12 medium containing 1% ITS-G, 0.5% P/S and 3 μM of CHIR for 2 days. One day before transfection, cells were dissociated into single cells and seeded on Matrigel-coated tissue culture plates at 75,000 cells/cm2 in DMEM medium supplemented with 20% FBS, 1% chicken embryo extract and 20 ng/ml of FGF2. One day after, cells were co-transfected in the same medium with 0.5 μg of each plasmids. Cells were grown for 3 more days with medium changing every day to express the vectors and activate the PAX7-GFP reporter cassette. Cells were then harvested and purified by FACS. Cells co-transfected with dCas9-VPR plasmid and the empty gRNA vector were used as controls. The GFP+ and GFP- cell fractions were collected and subjected to downstream analysis.

Method of directed differentiation

PAX7-GFP reporter cells were subjected to directed differentiation by the HX protocol as described above. Cells were harvested and purified by FACS. The H9 parental cells were differentiated alongside the reporter cells and used as controls. The GFP+ and GFP- cell fractions were collected and subjected to downstream analysis.

FACS cell sorting

Single cell solutions were filtered through 40 μm cell strainers and incubated with 1 μg/ml of DAPI as a live/dead cell indicator. When cell surface labelling was needed, cells were first blocked by Human TruStain FcX at RT for 5-10 minutes, followed by fluorophore-conjugated primary antibodies on ice for 20-30 minutes. For antigens requiring 2-step antibody staining, cells were stained on ice for 20-30 minutes with unconjugated primary antibodies followed by fluorophore-

conjugated secondary antibodies on ice for another 20-30 minutes. Stained cells were washed with FACS buffer and processed as described above. Cells were sorted by BD FACSAria sorters with FACSDiva software. Standard gating strategies were applied to exclude the debris, doublets and dead cells. Marker specific gating was set up using fluorescenceminus-one stained controls. The parental H9 cells were used for GFP gating. Sorted cells were collected into buffers containing 10% FBS and kept cold until downstream processing. FACS plots were generated using FlowJo.

Immunofluorescence

Cells were fixed with 4% PFA for 10 minutes, followed by permeabilization with 0.3% Triton X-100 in PBS at RT for 10 minutes at RT. Samples were then blocked with 3% BSA, 10% goat serum and 0.3% Triton X-100 in PBS for 60 minutes at RT. Primary antibodies were applied for overnight at 4°C and fluorophore-conjugated secondary antibodies for 60 minutes at RT. Nuclei were counter stained with DAPI at 1 µg/ml. Images were captured using a Zeiss Axio Observer.Z1 microscope equipped with an AxioCamMR3 camera. Image processing and quantification were performed using Fiji/ImageJ (Schindelin et al., 2012) or Zeiss ZEN 3.1 (blue edition).

Cytospin

Sorted cells were spun down onto Superfrost Plus microscope slides using Shandon Double Cytofunnel in a Shandon Cytocentrifuge. Attached cells were processed for immunofluorescence (IF) staining and imaging as described above. Immunohistochemistry with tyramide signal amplification Human embryos and tissues were fixed with 4% PFA for one day at 4°C, washed and embedded in paraffin. To reduce tissue autofluorescence for samples of fetal week 9 and older, they were subjected to a dehydration-bleaching-rehydration process before embedding as described by The Collection of Immunolabeled Transparent Human Embryos and Fetuses project (https://transparent-human-embryo.com/?page_id=649) (Belle et al., 2014). Tissue blocks were

then sectioned at a 4 μm interval onto Superfrost Plus microscope slides. For immunohistochemistry (IHC) staining, sections were deparaffinized with Xylene and rehydrated through EtOH/water gradient. Antigen retrieval was performed with a pressure cooker using 10 mM of sodium citrate buffer, followed by blocking with 3% BSA, 10% goat serum and 0.1% Tween 20 in PBS for 60 minutes at RT. Primary antibodies were applied for overnight at 4°C and HRP-conjugated secondary antibodies were applied for 45-60 minutes at RT. Tyramide signal amplification (TSA) was performed using the TSA Plus Fluorescence kits per the manufacturer's instructions to amplify the fluorescent signals. Slides were mounted with DAPI nuclei counterstaining and proceeded to image capture and analysis as described above. Images showing whole limbs of early embryonic development were captured in a mosaic mode and stitched together using the Zeiss software.

RNAscope with Immunohistochemistry

Human tissues were processed similar to regular IHC procedures as described above, except that fixation was performed at RT instead of 4°C and the bleaching step was omitted according to manufacturer's recommendations. Sections were hybridized with cataloged or custom designed RNAscope probes and signal developed per manufacturer's instructions using the RNAscope Multiplex Fluorescent Reagent Kit v2, with in-house protease treatment optimization (Protease Plus 15 minutes). Probe-hybridized sections were further subjected to IHC staining of PAX7 with TSA and imaged as described above. Quantification of RNAscope signals and PAX7 cells was performed using Zeiss ZEN 2.6 Pro (blue edition) software. RNAscope negative probes were applied on sections from different individual samples to set the threshold for positive signal counting.

Quantitative real time-PCR

Cells were harvested and RNA extracted using RNeasy Plus Mini or Micro Kit. Complementary DNA was synthesized using iScript Reverse Transcription Supermix and quantitative real time-PCR (qRT-PCR) was performed using SsoAdvanced Universal SYBR Green Supermix with technical triplicates on a Bio-Rad CFX384 Touch Real-Time PCR Detection System or a Thermo Fisher Scientific QuantStudio 6 Pro Real-Time PCR System. All primer pairs were selected from PrimerBank (Spandidos et al., 2010) or designed using Primer-BLAST (Ye et al., 2012) and tested in-house to ensure an amplification efficiency between 90-110%. Primer sequences for BGLAP, CKM, DCN, eGFP, IBSP, MYH8, OGN and RPL13A are listed in Methods S1. Other primer pairs are the same as previously reported (Xi et al., 2017).

In vitro myotube fusion assay

Sorted cells were resuspended in Lonza SkGM2 medium supplemented with 20 ng/ml of FGF2 and plated onto Matrigel-coated culture wells. Cells were cultured for 5-7 days until they reached >70-80% confluency. Then, medium was switched to DMEM/F12 medium containing 1% ITS-G, 0.5% P/S and 1% N2 supplement to induce fusion for 4-6 days. Medium was refreshed every other day during the culture, and cells at the end of fusion were subjected to IF staining and imaging as described above.

In vitro myogenic and osteogenic bipotential differentiation assays

Sorted cells were plated onto Matrigel-coated culture wells and expanded for 4-6 days in expansion medium (DMEM/F12 medium containing 20% FBS, 1% GlutaMAX, 1% NEAA, 1mM sodium pyruvate, 0.5% P/S and 20 ng/ml of FGF2). Cells were then split and cultured for another 2-3 days in expansion medium until they reached >70-80% confluency. For myogenic differentiation, medium was switched to fusion medium for 4-6 days as described above and cells were subjected to IF staining or harvested for qRT-PCR at the end of the fusion period. For osteogenic differentiation, medium was switched to Thermo Fisher Scientific StemPro

Osteogenesis Differentiation medium for 2-3 weeks. At the end of the osteogenic period, cells were subjected to Alizarin Red S staining as previously reported (Xi et al., 2017) or harvested for qRT-PCR analysis. Medium was refreshed every other day during expansion and differentiation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing, read alignment and digital gene expression matrix generation

The raw sequencing reads were processed using the Drop-seq_tools-1.13 pipeline from the McCaroll lab (https://github.com/broadinstitute/Drop-seq/releases/tag/v1.13), following the general guidelines from the Drop-seq Alignment Cookbook v1.2 (https://github.com/broadinstitute/Dropseq/files/2425535/DropseqAlignmentCookbookv1.2Jan2016.pdf) (Macosko et al., 2015). Briefly, reads were indexed and filtered by read quality. Sequencing adapter and polyA sequences were trimmed, and reads were further filtered to retain those of a length of at least 30 nucleotides. Processed reads were aligned to the human reference genome (hg19) using Bowtie2 (v2.2.9 with the '--very-sensitive' mode) (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) (Langmead and Salzberg, 2012). Aligned reads were tagged to gene exons using Bedtools Intersect (v2.26.0) (https://github.com/arq5x/bedtools2) (Quinlan and Hall, 2010). Knee plots of cell-to-read fraction were generated to estimate the number of cell barcodes representing true cells. Digital gene expression matrices (DGEs) were then generated by counting gene transcripts for the number of cell barcodes selected based on the infliction points in the knee plots. To correct for any bead synthesis errors/read errors leading to false barcodes, transcript barcodes (unique molecular identifiers; UMIs) or cell barcodes were merged when they were within 1 Hamming or 2 Levenshtein distances, respectively. Barcodes containing < 2500 reads were excluded from the DGEs.

Computational analysis using Seurat

Data filtration, normalization and scaling Downstream computational analyses of scRNA-seq data were mainly performed using the R package Seurat v2.3.3 (https://github.com/satijalab/seurat/releases/tag/v2.3.3) (Butler et al., 2018) by largely following the standard guidelines from the Satija lab (https://satijalab.org/seurat/). Seurat objects were generated with DGEs constructed as described above. Violin plots of number of expressed genes and unique transcripts (nGene and nUMI, respectively) of each cell were generated and outliers with too high or too low nGene and nUMI were removed to exclude potential cell doublets/aggregates or low quality cells/cell debris, respectively. As sequencing depth and cell type compositions vary across different samples, this filtration step was performed on a sample-to-sample basis. In general, prenatal and hPSC derived samples were filtered with a minimum nGene of 500-1000. We consistently observed lower number of genes expressed from postnatal juvenile and adult samples, although in general they have the lowest unique read fraction levels (suggesting higher sequencing coverage) among all samples. Therefore, we set the nGene threshold of these samples to 250-400. After the cell filtration step, expression counts of each cell were normalized with the default Seurat setting using “NormalizeData”. To mitigate the cell cycle effects on potentially grouping different cell types based on their cell cycle states, we assigned “S.Score” and “G2M.Score” on each cell with the average normalized expression levels of core cell cycle genes using “CellCycleScoring” following the Seurat instructions (Tirosh et al., 2016). To reduce the effects of dissociation-related stress on gene expression analysis, we obtained the core stress genes identified from scRNA-seq studies on both mouse skeletal muscle and acinar (van den Brink et al., 2017), and assigned each cell a “Stress” score using the core stress gene list through the Seurat “AddModuleScore” function. Briefly, this function first assigned each of the genes to be analyzed into different bins based on the genes’ average expression across single cells. It then calculated a residual for each analyzed gene in each cell by subtracting the average expression of the control gene set from the expression level of the gene being analyzed, where the control genes were randomly selected from the bin that the analyzed gene was assigned to.

This process was then reiterated through all the genes in the provided list, and the resulted aggregated expression was assigned as the score of the property the provided gene list represents. After this step, data scaling was performed using “ScaleData”, with “S.Score”, “G2M.Score” and “Stress” passed onto the “vars.to.regress” argument. At the same time, “nMUI” was also included in regression to control for the effects of cell size and/or sequencing depth.

Muscle.Score and Dev.Score

To readily detect skeletal muscle cells at various developmental or differentiation states, we assigned each cell a “Muscle.Score” using the above described “AddModuleScore” function using a list of conserved muscle cell genes (PAX3, PAX7, PITX2, MYF5, MYF6, MYOD1, MYOG, NEB and MYH3). To quantify the developmental status of myogenic progenitor and stem cells, we first used “AddModuleScore” to assign each cell a postnatal score (“Pst.Score”) using genes that were found to be upregulated in stage 5 SCs compared to stage 1 and 2 embryonic SMPCs (Figure 4-4). Similarly, we assigned each cell an embryonic score (“Emb.Score”) using genes upregulated in stage 1 and 2 SMPCs compared to stage 5 SCs. Finally, we calculated each cell’s myogenic developmental score (“Dev.Score”) by subtracting its “Emb.Score” from “Pst.Score”. Thus, a cell with a developmental “age” close to postnatal SCs would have a higher value of “Dev.Score”, and that similar to embryonic SMPCs a lower value.

Dimensional reduction and clustering

First, the most highly variable genes within each Seurat object were calculated and selected using “FindVariableGenes” (1500-2500 genes). Principle components (PCs) were calculated using the selected top variable genes by “RunPCA”, and the PCs were plotted using “PCElbowPlot”. Significant PCs were selected based on the elbow plot and used to further reduce data dimensionality using the T-distributed stochastic neighbor embedding (tSNE) method by “RunTSNE”. Cell clustering was performed by a shared nearest neighbor (SNN) modularity

optimization based clustering algorithm using the Seurat function “FindClusters” with “reduction.type” set to “pca”. Identification of clusters/cell types were aided by known cell type specific markers as well as the distribution of cells on the tSNE space.

Differential gene expression analysis

Differentially expressed genes (DEGs) between one cell cluster versus all remaining cells or between individual clusters were identified by “FindAllMarkers” or “FindMarkers”, respectively. For both functions, “test.use” was set to “negbinom” to fit for the sparse data type generated from scRNA-seq, and “return.thresh” (p values) less than 0.01 (finding cluster markers) or 0.05 (comparing two clusters). We passed the same parameters as we did when scaling the data (“S.Score”, “G2M.Score”, “Stress” and “nUMI”) to the “latent.var” argument to regress out the effects of cell cycle, dissociation-related stress as well as cell size/sequencing depth on the identification of DEGs. In addition, DEGs must also meet the following default criteria in Seurat: 1) average expression difference exceeding 1.28-fold between the comparing group of cells (“logfc.threshold = 0.25”), and 2) detected in a minimum of 10% of cells in either of the comparing populations (“min.pct = 0.1”).

Trajectory analysis

For trajectory analysis, we reduced the dimensionality of the data by diffusion map (DM) (Haghverdi et al., 2015) using the top variable genes of the objects via the Seurat “RunDiffusion” function. For in vivo SMPC and SC only analysis, we further clustered the cells using “FindClusters” with “reduction.type = “dm” (using the first 2 DM dimensions) into distinct developmental stages. For analysis combining in vivo SMPCs and SCs as well as hPSCSMPCs, “RunDiffusion” was performed using the top variable genes from the in vivo only dataset as a reference gene set. The developmental stage labels of the in vivo cells and the sample identities of the hPSC-derived cells were transferred and maintained from the original objects.

Analysis using Monocle3

Analysis in the Monocle3 R package (Cao et al., 2019) was performed according to Trapnell lab guidelines (https://cole-trapnell-lab.github.io/monocle3/monocle3_docs/). Gene expression data and cell metadata including cell type labels were carried over from the Seurat object. Parameters to regress were set similarly to analysis in Seurat by passing “nUMI”, “S.Score” “G2M.Score” and “Stress” to the “residual_model_formula_str” argument in the “preprocess_cds” function. Significant PCs were calculated and selected to further reduce the data dimensionality using uniform manifold approximation and projection (UMAP). Cells were plotted onto the UMAP space for visualization of their distribution and cell type identities.

Gene ontology enrichment analysis

Gene ontology (GO) enrichment was performed using Metascape (http://metascape.org/gp/index.html#/main/step1) (Zhou et al., 2019) against GO terms belonging to “Biological Processes”. Enriched GO terms with similar properties were further assigned to a common group, and the top 20 groups were retrieved. Select representative GO terms (members) from the consolidated groups were plotted against their negative Log10-transformed p values (no more than one member was selected from each group).

Gene set enrichment analysis

The gene set enrichment analysis (GSEA) (http://software.broadinstitute.org/gsea/index.jsp) (Subramanian et al., 2005) was performed with the “GSEAPreranked” mode against the “Canonical Pathways” (c2.cp.v6.1) gene sets database. The “enrichment statistic” was set to “classic” and enriched gene sets containing more than 500 or less than 10 genes were excluded from the final enriched gene sets. The “normalization mode” was set to “meandiv” and permutations were performed 1000 times.

Co-regulated gene network analysis

To build the co-regulated gene network, the dataset containing all stages of in vivo SMPCs and SCs and in vitro hPSC-SMPCs derived using the HX protocol (Figure 4-7A), was used to compute a Pearson gene-to-gene correlation matrix and determine groups/networks including genes with correlation values greater than 0.125. Similar networks were condensed by segregation of cells into ample numbers of small cell clusters (roughly 50 cells per cluster), from which the expression of the primary networks was calculated and compared to each other again via a Pearson network-to-network correlation matrix, followed by merging similar networks with expression correlation of 0.7 or higher to generate the final networks. The expression level of a given gene group/network was calculated by averaging the normalized expression values of all genes in the group in a given cell. We manually inspected the gene groups to exclude those that were driven by an extremely high expression level of a few genes in random rare cells. To retrieve TFs from the gene groups, we intersected our identified genes with those annotated as transcription factors/regulators by the Animal Transcription Factor Database (bioinfo.life.hust.edu.cn/AnimalTFDB/). To mitigate the effects of tissue/cell dissociation-induced stress signatures, we compiled a common stress gene list (411 genes) from published literature. Genes included in this list were chosen based on the following criteria: 1) included in the stress regression gene list as described above, or 2) significantly changed in the same direction (both induced or reduced) in response to dissociation-related stress as reported by van Velthoven et al and Machado et al (Machado et al., 2017; van Velthoven et al., 2017). Mouse genes were converted to their homologs in human and those mice only genes were removed from the final list. These common stress genes were intersected with the gene groups and TF sub-lists to exclude them from the final gene/TF lists for downstream analysis.

For gene group heatmaps, the average expression of selected groups was calculated for each developmental or directed differentiation stage/sample using the Seurat

"AverageExpression" function. Only groups containing 50 or more genes were plotted. For GO analysis, all genes contained in a given group were used as input to Metascape for enrichment analysis against "GO Biological Processes".

Hierarchical clustering

Average gene expression levels of single cells belonging to the same groups were calculated using the Seurat "AverageExpression" function. Spearman correlation coefficients were calculated between the averaged group of cells and visualized using the R package pheatmap. Hierarchical clustering was performed via the same package with default settings.

Gene list intersection and Venn diagram generation

Individual gene lists were supplied as input to the R package eulerr. All possible intersections of input gene lists were calculated and visualized in Venn diagram format with region areas proportional to the number of events in the regions.

References

1. Akiyama, H., Kim, J.E., Nakashima, K., Balmes, G., Iwai, N., Deng, J.M., Zhang, Z., Martin, J.F., Behringer, R.R., Nakamura, T., and de Crombrugghe, B. (2005). Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14665-14670.
2. Alexander, M.S., Rozkalne, A., Colletta, A., Spinazzola, J.M., Johnson, S., Rahimov, F., Meng, H., Lawlor, M.W., Estrella, E., Kunkel, L.M., and Gussoni, E. (2016). CD82 Is a Marker for Prospective Isolation of Human Muscle Satellite Cells and Is Linked to Muscular Dystrophies. *Cell. Stem Cell*. **19**, 800-807.
3. Applebaum, M., and Kalcheim, C. (2015). Mechanisms of myogenic specification and patterning. *Results Probl. Cell Differ.* **56**, 77-98.
4. Bareja, A., Holt, J.A., Luo, G., Chang, C., Lin, J., Hinken, A.C., Freudenberg, J.M., Kraus, W.E., Evans, W.J., and Billin, A.N. (2014). Human and mouse skeletal muscle stem cells: convergent and divergent mechanisms of myogenesis. *PLoS One* **9**, e90398.
5. Barna, M., and Niswander, L. (2007). Visualization of cartilage formation: insight into cellular properties of skeletal progenitors and chondrodysplasia syndromes. *Dev. Cell.* **12**, 931-941.
6. Barruet, E., Garcia, S.M., Striedinger, K., Wu, J., Lee, S., Byrnes, L., Wong, A., Xuefeng, S., Tamaki, S., Brack, A.S., and Pomerantz, J.H. (2020). Functionally heterogeneous human satellite cells identified by single cell RNA sequencing. *Elife* **9**, 10.7554/eLife.51576.
7. Belle, M., Godefroy, D., Dominici, C., Heitz-Marchaland, C., Zelina, P., Hellal, F., Bradke, F., and Chedotal, A. (2014). A simple method for 3D analysis of immunolabeled axonal tracts in a transparent nervous system. *Cell*. Rep. **9**, 1191-1201.
8. Borchin, B., Chen, J., and Barberi, T. (2013). Derivation and FACS-mediated purification of PAX3+/PAX7+ skeletal muscle precursors from human pluripotent stem cells. *Stem Cell. Reports* **1**, 620-631.
9. Brent, A.E., and Tabin, C.J. (2002). Developmental regulation of somite derivatives: muscle, cartilage and tendon. *Curr. Opin. Genet. Dev.* **12**, 548-557.
10. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411-420.
11. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C., and Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502.
12. Castiglioni, A., Hettmer, S., Lynes, M.D., Rao, T.N., Tchessalova, D., Sinha, I., Lee, B.T., Tseng, Y.H., and Wagers, A.J. (2014). Isolation of progenitors that exhibit myogenic/osteogenic bipotency in vitro by fluorescence-activated cell sorting from human fetal muscle. *Stem Cell. Reports* **2**, 92-106.
13. Cerletti, M., Shadrach, J.L., Jurga, S., Sherwood, R., and Wagers, A.J. (2008). Regulation and function of skeletal muscle stem cells. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 317-322.
14. Chal, J., Al Tanoury, Z., Hestin, M., Gobert, B., Aivio, S., Hick, A., Cherrier, T., Nesmith, A.P., Parker, K.K., and Pourquie, O. (2016). Generation of human muscle fibers and satellite-like cells from human pluripotent stem cells in vitro. *Nat. Protoc.* **11**, 1833-1850.
15. Chal, J., Oginuma, M., Al Tanoury, Z., Gobert, B., Sumara, O., Hick, A., Bousson, F., Zidouni, Y., Mursch, C., Moncuquet, P., et al. (2015). Differentiation of pluripotent stem cells to muscle fiber to model Duchenne muscular dystrophy. *Nat. Biotechnol*. **33**, 962-969.
16. Chal, J., and Pourquie, O. (2017). Making muscle: skeletal myogenesis in vivo and in vitro. *Development* **144**, 2104-2122.

17. Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* **27**, 275-280.
18. Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J., et al. (2015). Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* **12**, 326-328.
19. Chen, J.F., Tao, Y., Li, J., Deng, Z., Yan, Z., Xiao, X., and Wang, D.Z. (2010). microRNA-1 and microRNA-206 regulate skeletal muscle satellite cell proliferation and differentiation by repressing Pax7. *J. Cell Biol*. **190**, 867-879.
20. Cheung, M., Tai, A., Lu, P.J., and Cheah, K.S. (2019). Acquisition of multipotent and migratory neural crest cells in vertebrate evolution. *Curr. Opin. Genet. Dev*. **57**, 84-90.
21. De Micheli, A.J., Laurilliard, E.J., Heinke, C.L., Ravichandran, H., Fraczek, P., Soueid-Baumgarten, S., De Vlaminck, I., Elemento, O., and Cosgrove, B.D. (2020). Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration. *Cell*. Rep. **30**, 3583-3595.e5.
22. Dell'Orso, S., Juan, A.H., Ko, K.D., Naz, F., Perovanovic, J., Gutierrez-Cruz, G., Feng, X., and Sartorelli, V. (2019). Single cell analysis of adult mouse skeletal muscle stem cells in homeostatic and regenerative conditions. *Development* **146**, 10.1242/dev.174177.
23. Flamini, V., Ghadiali, R.S., Antczak, P., Rothwell, A., Turnbull, J.E., and Pisconti, A. (2018). The Satellite Cell Niche Regulates the Balance between Myoblast Differentiation and Self-Renewal via p53. *Stem Cell. Reports* **10**, 970-983.
24. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol*. **32**, 279-284.
25. Giordani, L., He, G.J., Negroni, E., Sakai, H., Law, J.Y.C., Siu, M.M., Wan, R., Corneau, A., Tajbakhsh, S., Cheung, T.H., and Le Grand, F. (2019). High-Dimensional Single-Cell Cartography Reveals Novel Skeletal Muscle-Resident Cell Populations. *Mol. Cell* **74**, 609-621.e6.
26. Gopinath, S.D., Webb, A.E., Brunet, A., and Rando, T.A. (2014). FOXO3 promotes quiescence in adult muscle stem cells during the process of self-renewal. *Stem Cell. Reports* **2**, 414-426.
27. Gros, J., and Tabin, C.J. (2014). Vertebrate limb bud formation is initiated by localized epithelialto-mesenchymal transition. *Science* **343**, 1253-1256.
28. Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single cell analysis of differentiation data. *Bioinformatics* **31**, 2989-2998.
29. Hayashi, K., and Ozawa, E. (1995). Myogenic cell migration from somites is induced by tissue contact with medial region of the presumptive limb mesoderm in chick embryos. *Development* **121**, 661-669.
30. Hicks, M.R., Hiserodt, J., Paras, K., Fujiwara, W., Eskin, A., Jan, M., Xi, H., Young, C.S., Evseenko, D., Nelson, S.F., et al. (2018). ERBB3 and NGFR mark a distinct skeletal muscle progenitor cell in human development and hPSCs. *Nat. Cell Biol*. **20**, 46-57.
31. Joe, A.W., Yi, L., Natarajan, A., Le Grand, F., So, L., Wang, J., Rudnicki, M.A., and Rossi, F.M. (2010). Muscle injury activates resident fibro/adipogenic progenitors that facilitate myogenesis. *Nat. Cell Biol*. **12**, 153-163.
32. Kim, J., Magli, A., Chan, S.S.K., Oliveira, V.K.P., Wu, J., Darabi, R., Kyba, M., and Perlingeiro, R.C.R. (2017). Expansion and Purification Are Critical for the Therapeutic Application of Pluripotent Stem Cell-Derived Myogenic Progenitors. *Stem Cell. Reports* **9**, 12-22.
33. Kivela, R., Salmela, I., Nguyen, Y.H., Petrova, T.V., Koistinen, H.A., Wiener, Z., and Alitalo, K. (2016). The transcription factor Prox1 is essential for satellite cell differentiation and muscle fibre-type regulation. *Nat. Commun*. **7**, 13124.
34. Koopman, R., Ly, C.H., and Ryall, J.G. (2014). A metabolic link to skeletal muscle wasting and regeneration. *Front. Physiol.* **5**, 32.

35. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359.
36. Li, X., Zhao, X., Fang, Y., Jiang, X., Duong, T., Fan, C., Huang, C.C., and Kain, S.R. (1998). Generation of destabilized green fluorescent protein as a transcription reporter. *J. Biol. Chem.* **273**, 34970-34975.
37. Machado, L., Esteves de Lima, J., Fabre, O., Proux, C., Legendre, R., Szegedi, A., Varet, H., Ingerslev, L.R., Barres, R., Relaix, F., and Mourikis, P. (2017). In Situ Fixation Redefines Quiescence and Early Activation of Skeletal Muscle Stem Cells. *Cell*. *Rep.* **21**, 1982-1993.
38. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214.
39. Magli, A., and Perlingeiro, R.R.C. (2017). Myogenic progenitor specification from pluripotent stem cells. Semin. *Cell Dev. Biol.* **72**, 87-98.
40. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826.
41. McConnell, B.B., and Yang, V.W. (2010). Mammalian Kruppel-like factors in health and diseases. *Physiol. Rev*. **90**, 1337-1381.
42. Messina, G., Biressi, S., Monteverde, S., Magli, A., Cassano, M., Perani, L., Roncaglia, E., Tagliafico, E., Starnes, L., Campbell, C.E., et al. (2010). Nfix regulates fetal-specific transcription in developing skeletal muscle. *Cell* **140**, 554-566.
43. Murmann, O.V., Niggli, F., and Schafer, B.W. (2000). Cloning and characterization of the human PAX7 promoter. *Biol. Chem*. **381**, 331-335.
44. Neufeld, S.J., Wang, F., and Cobb, J. (2014). Genetic interactions between Shox2 and Hox genes during the regional growth and development of the mouse limb. *Genetics* **198**, 1117-1126.
45. Oginuma, M., Moncuquet, P., Xiong, F., Karoly, E., Chal, J., Guevorkian, K., and Pourquie, O. (2017). A Gradient of Glycolytic Activity Coordinates FGF and Wnt Signaling during Elongation of the Body Axis in Amniote Embryos. *Dev. Cell*. **40**, 342-353.e10.
46. Oh, Y., and Jang, J. (2019). Directed Differentiation of Pluripotent Stem Cells by Transcription Factors. *Mol. Cells* **42**, 200-209.
47. Pala, F., Di Girolamo, D., Mella, S., Yennek, S., Chatre, L., Ricchetti, M., and Tajbakhsh, S. (2018). Distinct metabolic states govern skeletal muscle stem cell fates during prenatal and postnatal myogenesis. *J. Cell. Sci*. **131**, 10.1242/jcs.212977.
48. Petchey, L.K., Risebro, C.A., Vieira, J.M., Roberts, T., Bryson, J.B., Greensmith, L., Lythgoe, M.F., and Riley, P.R. (2014). Loss of Prox1 in striated muscle causes slow to fast skeletal muscle fiber conversion and dilated cardiomyopathy. *Proc. Natl. Acad. Sci. U. S. A*. **111**, 9515-9520.
49. Pistocchi, A., Gaudenzi, G., Foglia, E., Monteverde, S., Moreno-Fortuny, A., Pianca, A., Cossu, G., Cotelli, F., and Messina, G. (2013). Conserved and divergent functions of Nfix in skeletal muscle development during vertebrate evolution. *Development* **140**, 1528-1536.
50. Price, F.D., von Maltzahn, J., Bentzinger, C.F., Dumont, N.A., Yin, H., Chang, N.C., Wilson, D.H., Frenette, J., and Rudnicki, M.A. (2014). Inhibition of JAK-STAT signaling stimulates adult satellite cell function. *Nat. Med*. **20**, 1174-1181.
51. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.
52. Reinhardt, R., Gullotta, F., Nusspaumer, G., Unal, E., Ivanek, R., Zuniga, A., and Zeller, R. (2019). Molecular signatures identify immature mesenchymal progenitors in early mouse limb buds that respond differentially to morphogen signaling. *Development* **146**, 10.1242/dev.173328.

53. Rubenstein, A.B., Smith, G.R., Raue, U., Begue, G., Minchev, K., Ruf-Zamojski, F., Nair, V.D., Wang, X., Zhou, L., Zaslavsky, E., et al. (2020). Single-cell transcriptional profiles in human skeletal muscle. *Sci. Rep*. **10**, 229-019-57110-6.
54. Ryall, J.G. (2013). Metabolic reprogramming as a novel regulator of skeletal muscle development and regeneration. *Febs j.* **280**, 4004-4013.
55. Ryall, J.G., Dell'Orso, S., Derfoul, A., Juan, A., Zare, H., Feng, X., Clermont, D., Koulnis, M., Gutierrez-Cruz, G., Fulco, M., and Sartorelli, V. (2015). The NAD(+)-dependent SIRT1 deacetylase translates a metabolic switch into regulatory epigenetics in skeletal muscle stem cells. *Cell. Stem Cell.* **16**, 171-183.
56. Sacco, A., Doyonnas, R., Kraft, P., Vitorovic, S., and Blau, H.M. (2008). Self-renewal and expansion of single transplanted muscle stem cells. *Nature* **456**, 502-506.
57. Sambasivan, R., and Tajbakhsh, S. (2007). Skeletal muscle stem cell birth and properties. Semin. *Cell Dev. Biol*. **18**, 870-882.
58. Sanchez, A.M., Candau, R.B., and Bernardi, H. (2014). FoxO transcription factors: their roles in the maintenance of skeletal muscle homeostasis. *Cell Mol. Life Sci.* **71**, 1657-1671.
59. Sartori, R., and Sandri, M. (2015). Bone and morphogenetic protein signalling and muscle mass. *Curr. Opin. Clin. Nutr. Metab. Care* **18**, 215-220.
60. Schiaffino, S., Rossi, A.C., Smerdu, V., Leinwand, L.A., and Reggiani, C. (2015). Developmental myosins: expression patterns and functional significance. Skelet Muscle 5, 22-015-0046-6. *eCollection* 2015.
61. 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.
62. Schmidt, R., and Plath, K. (2012). The roles of the reprogramming factors Oct4, Sox2 and Klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation. *Genome Biol*. **13**, 251-2012-13-10-251.
63. Shea, K.L., Xiang, W., LaPorta, V.S., Licht, J.D., Keller, C., Basson, M.A., and Brack, A.S. (2010). Sprouty1 regulates reversible quiescence of a self-renewing adult muscle stem cell pool during regeneration. *Cell. Stem Cell*. **6**, 117-129.
64. Shelton, M., Metz, J., Liu, J., Carpenedo, R.L., Demers, S.P., Stanford, W.L., and Skerjanc, I.S. (2014). Derivation and expansion of PAX7-positive muscle progenitors from human and mouse embryonic stem cells. *Stem Cell. Reports* **3**, 516-529.
65. Spandidos, A., Wang, X., Wang, H., and Seed, B. (2010). PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. *Nucleic Acids Res*. **38**, D792-9.
66. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545-15550.
67. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372.
68. Taglietti, V., Angelini, G., Mura, G., Bonfanti, C., Caruso, E., Monteverde, S., Le Carrou, G., Tajbakhsh, S., Relaix, F., and Messina, G. (2018). RhoA and ERK signalling regulate the expression of the transcription factor Nfix in myogenic cells. *Development* **145**, 10.1242/dev.163956.
69. Takahashi, K., and Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.* **17**, 183-193.

70. Tanaka, K., Matsumoto, E., Higashimaki, Y., Katagiri, T., Sugimoto, T., Seino, S., and Kaji, H. (2012). Role of osteoglycin in the linkage between muscle and bone. *J. Biol. Chem*. **287**, 11616-11628.
71. Tierney, M.T., Aydogdu, T., Sala, D., Malecova, B., Gatto, S., Puri, P.L., Latella, L., and Sacco, A. (2014). STAT3 signaling controls satellite cell expansion and skeletal muscle repair. *Nat. Med.* **20**, 1182-1186.
72. Tierney, M.T., Gromova, A., Sesillo, F.B., Sala, D., Spenle, C., Orend, G., and Sacco, A. (2016). Autonomous Extracellular Matrix Remodeling Controls a Progressive Adaptation in Muscle Stem Cell Regenerative Capacity during Development. *Cell. Rep*. **14**, 1940-1952.
73. Tierney, M.T., and Sacco, A. (2016). Satellite Cell Heterogeneity in Skeletal Muscle Homeostasis. *Trends Cell Biol*. **26**, 434-444.
74. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H.,2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196.
75. Uezumi, A., Fukada, S., Yamamoto, N., Takeda, S., and Tsuchida, K. (2010). Mesenchymal progenitors distinct from satellite cells contribute to ectopic fat cell formation in skeletal muscle. *Nat. Cell Biol*. **12**, 143-152.
76. Uezumi, A., Nakatani, M., Ikemoto-Uezumi, M., Yamamoto, N., Morita, M., Yamaguchi, A., Yamada, H., Kasai, T., Masuda, S., Narita, A., et al. (2016). Cell-Surface Protein Profiling Identifies Distinctive Markers of Progenitor Cells in Human Skeletal Muscle. *Stem Cell. Reports* **7**, 263-278.
77. van den Brink, S.C., Sage, F., Vertesy, A., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935-936.
78. van Velthoven, C.T.J., de Morree, A., Egner, I.M., Brett, J.O., and Rando, T.A. (2017). Transcriptional Profiling of Quiescent Muscle Stem Cells In Vivo. *Cell. Rep*. **21**, 1994-2004.
79. Xi, H., Fujiwara, W., Gonzalez, K., Jan, M., Liebscher, S., Van Handel, B., Schenke-Layland, K., and Pyle, A.D. (2017). In Vivo Human Somitogenesis Guides Somite Development from hPSCs. *Cell. Rep*. **18**, 1573-1585.
80. Xu, C., Tabebordbar, M., Iovino, S., Ciarlo, C., Liu, J., Castiglioni, A., Price, E., Liu, M., Barton, E.R., Kahn, C.R., Wagers, A.J., and Zon, L.I. (2013). A zebrafish embryo culture system defines factors that promote vertebrate myogenesis across species. *Cell* **155**, 909-921.
81. Xu, X., Wilschut, K.J., Kouklis, G., Tian, H., Hesse, R., Garland, C., Sbitany, H., Hansen, S., Seth, R., Knott, P.D., Hoffman, W.Y., and Pomerantz, J.H. (2015). Human Satellite Cell Transplantation and Regeneration from Diverse Skeletal Muscles. *Stem Cell. Reports* **5**, 419-434.
82. Yajima, H., Yoneitamura, S., Watanabe, N., Tamura, K., and Ide, H. (1999). Role of N-cadherin in the sorting-out of mesenchymal cells and in the positional identity along the proximodistal axis of the chick limb bud. *Dev. Dyn.* **216**, 274-284.
83. Yang, H., Wang, H., Shivalila, C.S., Cheng, A.W., Shi, L., and Jaenisch, R. (2013). One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**, 1370-1379.
84. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134-2105-13-134.
85. Yin, H., Price, F., and Rudnicki, M.A. (2013). Satellite cells and the muscle stem cell niche. *Physiol. Rev.* **93**, 23-67.

86. Yucel, N., Wang, Y.X., Mai, T., Porpiglia, E., Lund, P.J., Markov, G., Garcia, B.A., Bendall, S.C., Angelo, M., and Blau, H.M. (2019). Glucose Metabolism Drives Histone Acetylation Landscape Transitions that Dictate Muscle Stem Cell Function. *Cell*. *Rep.* **27**, 3939-3955.e6.
87. Zhao, P., and Hoffman, E.P. (2004). Embryonic myogenesis pathways in muscle regeneration. *Dev. Dyn*. **229**, 380-392.
88. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun*. **10**, 1523-019-09234-6.

4.2 A Molecular Atlas of Proximal Airway Identifies Subsets of Known Airway Cell Types Revealing Details of the Unique Molecular Pathogenesis of Cystic Fibrosis

Introduction.

Cystic fibrosis (CF) is a lethal autosomal recessive disorder that afflicts in excess of 70,000 people globally. People with CF experience multi-organ dysfunction resulting from aberrant electrolyte transport across polarized epithelia due to mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. CF-related lung disease is by far the most significant determinant of morbidity and mortality. In this study we report results from a multi-institute consortium in which single cell transcriptomics were applied to define disease-related changes to the proximal airway of CF donors (n=19) undergoing transplantation for end-stage lung disease compared to the proximal airway of previously healthy lung donors (n=19). We found that all major airway epithelial cell types were conserved between control and CF donors. Disease-dependent differences were observed, including an overabundance of epithelial cells transitioning to specialized ciliated and secretory cell subtypes coupled with an unexpected decrease in cycling basal cells. This study developed a molecular atlas of the proximal airway epithelium that will provide insights for the development of new targeted therapies for CF airway disease.

Results.

Transcriptional analysis of single cells from control and CF airways

There is a great deal of interest in defining human bronchial epithelial (hBE) cell subtypes in Cystic Fibrosis (CF) airways as a means to develop gene therapeutic strategies to effect long-term correction of CFTR function. To address this knowledge gap we sought to produce single cell reference atlases of proximal airway epithelium isolated from lung tissue from donors with no evidence of chronic lung disease (CO, considered control for these experiments; n=19) compared to explant tissue from patients undergoing transplantation for end-stage CF lung disease (CF, n=19). Isolation of single cells from proximal airways was performed at three different institutions (Fig 4-8a), using similar but distinct methodologies (Fig 4-8b & Materials and Methods). After initial quality control and filtering, datasets from the three institutions were integrated for subsequent analyses. Data were visualized through UMAP dimensional reduction. The distribution of cells from each institution on UMAP projections showed homogeneous data integration (Supp Fig S4-8a,b). While all datasets integrated well, expression of some genes, particularly those associated with metabolic state, did show differential expression according to institution (Supp Fig S4-8c-f). Accordingly, data that were reproducibly observed across datasets from each of the three institutions were highlighted in this study.

UMAP projections of data comparing cells from CO versus CF samples revealed a high degree of overlap (Fig 4-8c). Using the cell type gene signatures from Plasschaert et al[1], we were able to identify all major human airway epithelial cell types including basal, secretory and ciliated, in addition to rare cell types including ionocytes, neuroendocrine (NE) and FOXN4+ cell populations (Supp Fig S4-8g,h). We then performed differentially expressed gene (DEG) analysis between clusters to discern cell subtypes with unique molecular characteristics. Among the three major cell types we were able to resolve 3 ciliated, 5 secretory, and 5 basal cell subtypes (Fig 4-8c). These subtypes of each major airway epithelial cell type were found in similar proportions

between CO versus CF samples and institutions (Fig 4-8d, Supp Fig S4-8i). We considered the functions conferred by differential genes to distinguish each cell subtype.

Basal cells were subdivided into five clusters (Basal1-5) (Fig 4-8e). The Basal1 cluster is characterized by high expression of canonical basal cell markers including tumor protein P63 (TP63) and the cytokeratins 5 and 15 (KRT5 and KRT15) (Fig 4-8e)[1]. Cells of the Basal2 cluster show enrichment for transcripts such as DNA Topoisomerase II Alpha (TOP2A) and the Marker of Proliferation Ki-67 (MKI67) and have a transcriptomic signature indicative of proliferating basal cells (Fig 4-8e). The Basal3 cluster is enriched for transcripts of the serpin family, members of which are known to regulate protein folding associated with secretory cell maturation and may represent basal cells transitioning to a secretory phenotype[2]. The Basal4 cluster is characterized by the highest expression of the AP-1 family members JUN and FOS, and Basal5 uniquely expressed b-catenin (CTNNB1).

Secretory cells were partitioned into five specific subsets (Secretory1-5) that share defining gene signatures in CO and CF datasets (Fig 4-8e). The Secretory1 cluster includes cells characterized by high expression of Secretoglobin Family Member 1A1 (SCGB1A1) and various members of the Serpin family. Serpins regulate protein folding associated with maturation of secretory proteins[2] and define cells undergoing maturation into secretory cell type with similarities to club cells of bronchiolar airways[3]. The Secretory2 cluster is composed of cells sharing expression of mucins MUC5B and MUC5AC, anterior gradient 2 (AGR2) and SAM-pointed domain–containing Ets-like factor (SPDEF), suggesting they are goblet cells[4]. Cells in the Secretory3 cluster lack expression of known canonical secretory cell markers and can be distinguished from other secretory cluster subsets by their expression of Dynein Axonemal Heavy Chain proteins (DNAHs), Ankyrin Repeat Domain proteins (ANKRDs), and the mucins MUC16 and MUC4, suggesting that the Secretory3 cluster acts as progenitor cell for ciliated cell differentiation. The Secretory4 cluster is defined by expression of MUC5B and Trefoil Factor family domain peptides (TFF1 and TFF3) and represents a subtype of mucous-like cells that is

distinct from goblet cells5. The Secretory5 cluster contains a serous-like signature5, including expression of Lysozyme (LYZ), Proline-Rich Proteins (PRBs, and PRRs), and Lactoferrin (LTF), and may represent glandular cell types of submucosal glands (SMGs) or their surface airway epithelial counterparts.

Ciliated cells were subdivided in to three clusters (Ciliated1-3) (Fig 4-8e), all sharing expression of the lineage marker / master regulator of ciliogenesis, Forkhead box protein J1 (FOXJ1)[6]. The Ciliated1 cluster contains the highest expression of markers of cilia pre-assembly[7], including Sperm Associated Antigen 1 (SPAG1), Leucin Rich Repeat Containing 6 (LRRC6) and Dynein Axonemal Assembly Factor 1 (DNAAF1), whereas cells within the Ciliated2 cluster show the highest expression of markers of mature ciliated cells including TUBA1A and TUBB4B. The Ciliated3 cluster is characterized by the expression of Serum Amyloid A proteins (SAA1 and SAA2), reflective of a pro-inflammatory state[8], suggesting that this subset of ciliated cells is either responding to or regulating immune responses.

In light of the cellular heterogeneity observed among freshly isolated airway epithelial cells we sought to determine the extent to which this was recapitulated in commonly used culture models, most notably the primary human bronchial epithelial (hBE) cell air liquid interface (ALI) culture system. We performed single cell RNA sequencing on well differentiated ALI cultures[9] generated from hBE cells matched to a subset of CO and CF donors used for analysis of freshly isolated cells. Previously identified cell types[10] observed in fresh isolates (basal, secretory, ciliated, FOXN4+, ionocyte, and NE) were also observed in ALI cultures (Supp Fig S4-8j), for both CO and CF-derived samples (Supp Fig S4-8k). Based on gene expression differences in ALI cultures, we were able to further define subtypes of basal (ALI Basal1-4), secretory (ALI Secretory 1-4), and ciliated cells (ALI Ciliated 1) (Fig 4-8f). ALI Basal1, 2, and 4 showed overlapping marker gene expression with Basal1 (Canonical), Basal3 (Serpin-enriched), and Basal2 (proliferating) cells from freshly isolated tissue, respectively (compare Fig 4-8e and 4-8g). ALI Basal3 identified cells with high KRT14 expression that lacked a counterpart basal cell cluster in the fresh tissue

data sets (Fig 4-8e, g). ALI secretory and ciliated cell clusters lacked markers observed in the respective subtypes of the freshly isolated tissue (Fig 4-8e, g). The comparison of gene expression profiles between cells from ALI cultures and fresh tissue confirm that while the major cell types are present in ALI cultures, significant differences are observed in subtype states (Fig 4-8h, i, j). We conclude that ALI cultures recapitulate major hBE cell types observed among freshly isolated airway epithelial cells, as a regenerative model, but the cultures do not fully recapitulate the heterogeneity of cell subtypes observed in native airways at steady state.

Despite differences across the donors, isolation techniques, and sequencing methods, we found all cell subtypes in fresh airway tissue were recapitulated in each CO and CF patient (Fig 4-8k, Supp Fig S4-8k). Interestingly, we observed an average proportionate depletion in CF samples of 46.8% less cells in the proliferative Basal2 subtype and of 26% in the club cell-like Secretory1 subtype compared to CO, while a 44.6% increase in the proportion of cells in the inflammatory Ciliated3 subtype was observed (Fig 4-8k). These changes were observed in all three institutions, showing that the rigorous subtyping of human bronchial epithelium allows the identification of reproducible differences in cell states between CF and CO airways.

We next used our molecular atlas to examine cystic fibrosis transmembrane regulator (CFTR) expression in different cell types in CO and CF airways. Recent studies have proposed ionocytes as specialized cells with high CFTR expression that may represent primary tractable targets for restoration of CFTR expression in CF[10,11]. We found that CFTR is expressed in a wide selection of cells, with overall higher expression in CF compared to CO (Fig 4-8l). While more than 30% of all ionocytes expressed CFTR in both CO and CF samples, the majority of CFTR-expressing cells were secretory cells, followed by basal cells, with ionocytes constituting a minor fraction (Fig 4-8m, n). Analysis of the proportional expression of CFTR among secretory and basal cells showed that secretory cells and not ionocytes are the major producer of CFTR in both CO and CF tissue (Supp Fig S4-8l). Secretory2 (goblet-like) cells and Basal3 (serpin-expressing) cells were the major contributors to CFTR expression among the identified cell subtypes (Fig 4-8n,

Supp Fig S4-8l). The comparison of CFTR expression between CO and CF samples showed cell type- specific differences, with increases of expression in CF samples in the ionocyte, Secretory1 (Club-like), Secretory2 (Goblet-like), Basal1 (Canonical), and Basal3 (serpinexpressing) cell subtypes (Fig 4-8o). Overall, our analysis, while confirming the specialized role of ionocytes for CFTR expression, establishes that secretory cells are the main cells that express CFTR and that secretory and basal cells together contribute the vast majority to CFTR expression in the proximal airway epithelium. Therefore, both secretory and basal cells should be considered as plausible candidates for therapeutic restoration of CFTR expression in CF in addition to ionocytes.

Secretory cells show a secretory signature with increased antimicrobial activity in CF donors

The secretory cells of the proximal airway epithelium play an important role in host defense by producing serous and mucous secretions that trap and clear microbial organisms. People with CF develop dehydrated mucus and problems with mucus clearance which result in chronic bacterial infections. We explored the gene expression differences of secretory cell subtypes between CO and CF samples to define transcriptional differences that may be associated with these issues.

In order to find differences in each secretory cell subtype, we applied DEG analysis to identify the most specifically expressed genes in a given secretory subtype in CO or CF donors, and selected gene expression changes that were cross validated between all three datasets (Fig 4-9a). In the Secretory1 (Club-like) subtype, CF samples showed downregulation of members of the S100 gene family12, which are important for tissue repair, differentiation and inflammation, suggesting possible repair defects in CF donors. In the Secretory2 (Goblet-like) subtype, we found that immune response genes such as BPIFA1 and BPIFB1[13] were upregulated in CF samples. The Secretory3 (DNAHs enriched) subtype shows CF-specific increased expression of specific dyneins (DNAH5,11,12, DNAAF1), transcripts that are usually associated with cilium assembly[14]. In the Secretory4 (mucous-like) subtype, Angiogenin (ANG) and TFF1, two molecules with a role in antimicrobial defense[15,16], were upregulated in CF compared to CO samples. Secretory5

(serous-like) subtype showed fewer differences between CO and CF compared to other subtypes (Fig 4-9a).

To better understand the alterations to co-regulated sets of genes in CO compared to CF samples, we additionally applied an unbiased gene expression network discovery method that employs correlation between transcript levels to group genes to define co-expression networks that are most prominently expressed in secretory cells. We focused on seven of these networks (Net S1-S7) with statistical significance for differences in CO versus CF samples across each institution's data (Fig 4-9b, Supp Fig S4-9). Secretory networks 1-6 (Net S1-S6) are more highly expressed in CF vs CO secretory cells, in particular S2-S4, whereas S7 was lower in secretory cells from CF samples (Fig 4-9c, Supp Fig S4-9). Gene ontology analysis revealed an antimicrobial signature[17] for S1 and S4, S2 was related to ER stress[18] and S3 linked to metabolic processes (Fig 4-9b). The antimicrobial gene program S1 was most highly expressed in the Secretory4 and Secretory5 subgroups and expression of S4 was high specifically within the Secretory4 cluster (Fig 4-9c,d). This indicated that the Secretory5 (serous-like) and Secretory4 (mucous-like) cells in CF lungs have a highly specialized response in a particular subtype of cells and increased levels of antimicrobial activity in response to disease. Elevated ER-stress seen with S2 was more pronounced among Secretory4 and Secretory2 (goblet-like) cells (Fig 4-9c,d). S3 described a metabolic difference between Secretory2 (goblet-like) and Secretory1 (club-like) cells from CF versus CO samples (Fig 4-9c,d), indicating the surface airway epithelial secretory cells may be more exhausted. S5, marked by developmental ontology and containing the Wnt signaling gene FRZB, and S6, which contained Notch gene HEY1, were also elevated in CF samples (Supp Fig S4-9). Only one secretory network, S7, had a notable upregulation in CO compared to CF samples and marked a small group of cells expressing members of the KLK family, reported to be expressed in epithelial cells of the lung[19], and implicated in regulation of airway inflammatory response (Fig S4-9c,d).

Overall, gene expression and network differences identified between CO and CF secretory cell subtypes demonstrate overactive mucosal secretion, humoral immunity, antimicrobial activity and stress-related organelle maintenance, suggesting an increase in secretory function in the CF airway epithelium.

An expanded ciliated cell gene expression program reveals aberrant ciliogenesis and altered cellular lineages in CF airways

Multi-ciliated cells provide the necessary mechanical force for the directional clearance of contaminants trapped in the mucus layer for optimal airway homeostasis and host defense. Although CF-related defects in electrolyte transport across the epithelial lining and associated dehydration of airway surface liquid impair ciliary function and muco-ciliary transport, we know little about how ciliated cells respond to this perturbation. To comprehensively assess CF-related changes in ciliated cells, we compared their single cell transcriptomes between CO and CF patient samples. Ciliated cell gene expression is driven by a complex multi-ciliated cell specific gene expression network that turns on downstream of cell fate acquisition to generate the hundreds of structural and regulatory components of cilia[20]. In general, the temporal expression of specific ciliary genes reflects their role in sequential stages of ciliogenesis, and many transcripts are first strongly induced, then are eventually downregulated to a maintenance expression level[21]. Differential gene analysis revealed genes that were specific to either CO or CF cells in each of the ciliated subtypes, again reproducible between datasets from all three institutions (Fig 4-10a). The Ciliated1 subtype (cells undergoing ciliogenesis) showed higher expression for ciliogenesis transcripts such as Dynein Axonemal Heavy Chain 5 (DNAH5), Spectrin Repeat Containing Nuclear Envelope Protein 1 and 2 (SYNE1 and SYNE2) in CF compared to CO tissue, suggesting an attempt to boost cilium biogenesis in lungs from CF donors. Cells of the Ciliated2 (mature ciliated) subtype, showed higher expression of Anterior Gradient 3 (AGR3) and LRRC6 in CF samples, genes that play a role in ciliary beat frequency and motility[22,23]. CF cells of the Ciliated3

(cells involved in immune defense) subtype showed higher expression of Major Histocompatibility Complex, Class II, DP Alpha 1 and DR Beta 1 (HLA-DPA1 and HLA-DRB1), genes that play an important role in the immune system in antigen presenting cells.

Through gene expression network discovery, we also defined ten expression networks that are highly expressed in ciliated cells (Fig 4-10b, Supp Fig S4-10). Despite each network having distinct genes, many networks showed enrichment of ontology terms related to ciliogenesis and cilium movement (Net C1-C4, C8; Fig 4-10b, Supp Fig S4-10). Network C3 was associated with respiratory electron transport, C7 related to cellular repair and networks C3, C5, and C6 contained genes with immune functions (Supp Fig S4-10). The smaller networks C9 and C10 had no ontology but also contained immune and ciliary genes (Supp Fig S4-10). Interestingly, the Ciliated3 subtype consistently showed an increase in expression of all of these networks in CF compared to CO cells and often displayed the largest gene expression difference between these samples (Fig 4-10c,d; Supp Fig S4-10). We also found that the microtubule and ciliogenesis-related networks C1-C4 and C8 had higher expression among nonciliated cells in CF compared to CO tissues (Fig 4-10b, c).

Given this specific and unexpected upregulation of various cilium-related genes in non-ciliated cells in CF samples, we wondered if particular cell subtypes were affected. To address this question, we interrogated a manually curated list based on published observations, containing a total of 10 categories and 491 genes, representing different phases of ciliogenesis (Fig 4-10e, Supp Fig S4-10). We then compared the proportion of cells from each hBE cell subtype that expressed a given ciliogenesis signature above a specific cutoff between CO and CF samples. For the Ciliated1 and Ciliated2 subtypes, we found that more cells expressed centriole assembly and centriole-to cilium conversion gene signatures in CF than in CO samples suggesting a defect in ciliogenesis kinetics. All 10 ciliogenesis signatures were expressed by a higher proportion of CF cells in the “immune defense” Ciliated3 cell subtype, suggesting aberrant regulation of ciliogenesis in this subtype of ciliated cells in CF airways.

Among non-ciliated subtypes, Basal4, Basal5 and Secretory3 clusters had higher expression of nearly all categories of ciliogenesis signature genes in CF compared to CO samples, indicating a potential commitment toward the ciliated lineage in these CF non-ciliated cells (Fig 4-10e). Interestingly, FOXN4+ cells, previously reported to represent transitional FOXJ1+ cells undergoing multiciliogenesis[10], were also found to express ciliogenesis signature genes at a higher level in CF compared to CO samples. Taken together, these data suggest that CF airways include an overabundance of cells transitioning to the ciliated cell phenotype compared to CO airways. We speculate that this may result in generation of more structurally and functionally aberrant ciliated cells and more immune defensive Ciliated3 cells in CF tissues. Furthermore, the epithelial lining of CF airways exhibits a more plastic and stressed phenotype consistent with known airway defects resulting from electrolyte and ASL imbalances in the CF airway.

CF basal cells show depletion of metabolic stability and proliferation

Basal cells are considered to be the primary stem cells of the proximal airways that are capable of proliferation, long-term self-renewal, and differentiation to yield specialized luminal cell types[24,25]. Analysis of differentially expressed genes between basal cells of CO and CF samples revealed reproducible subtype-specific differences (Fig 4-10a). The CF Basal2 (Proliferating) cell subtype showed a general reduction of transcripts involved in cell division, whereas the CF Basal3 (Serpin-expressing) subtype showed lower expression of keratinization associated genes[26,27] including Cystatin A (CSTA) and Heat Shock Protein Family B (Small) Member 1 (HSPB1). The CF Basal4 (activated) subtype displayed increased expression of Fos and FosB Proto-Oncogene, AP-1 Transcription Factor Subunit (FOS, and FOSB), whereas the AP-1 complex companion transcription factors Jun and JunB (JUN and JUNB) were unchanged.

Using the gene network analysis approach, we defined 10 gene expression networks that were differentially regulated between CO and CF samples and were prominent in basal cells.

Eight networks (Net B1-B4, B7-B10) were higher in CO samples and two networks (B5 and B6) were higher in CF samples (Fig 4-10b, Supp Fig S4-12). The CF-enhanced B5 and B6 networks are related to surfactant metabolism and immune function, and, interestingly, were expressed in the smallest proportion of basal cells, specifically in the Basal5 subtype for B5 and the Basal4 subtype for B6 (Fig 4-10b, Supp Fig S4-12). Many networks showing down-regulation in CF compared to CO samples demonstrated gene ontologies related to metabolic processes and oxidative stress, cell division, epithelial cornification, immune functions, and response to wounding (Fig 4-10c, Supp Fig S4-12). Networks B1, B2 and B8 were more highly expressed in CO compared to CF samples (Fig 4-10c,d) and may signify patient specific wound healing related to intubation. Several other molecular pathways were also downregulated in the basal cells of CF samples compared to CO, including those related to response to oxidative stress, and ATP synthesis (Net B2, B4, B10, Fig 4-10c,d). Strikingly, networks B3 and B7 revealed widespread downregulation of genes related to cell cycle in CF samples across all basal subtypes but most notably among cells of the Basal2 (proliferating) cluster (Fig 4-10b,c,d), which may be related to the fact that the number of Basal2 cells is lower in CF than CO samples.

Our finding of reduced proliferative capacity among basal cells of CF compared to CO airways has important implications for the ability of endogenous stem/progenitor cells to maintain the specialized epithelial lining of CF airways. To confirm the depletion of dividing basal cells in intact CF mucosa that are inferred from single cell RNA-Seq data, we analyzed immunofluorescent co-staining for a proliferative marker (PCNA) and the basal marker, KRT5, in the same proximal airway samples used for transcriptomic analysis. We found that the PCNA-proliferative index of KRT5-immunoreactive cells in CF proximal airways was significantly reduced compared to comparable airway regions of CO tissue (Fig 4-10e,f, Supp Fig S4-13). Furthermore, analysis of cell-cycle transitional state signatures in transcriptomes of the proliferative Basal2 cell cluster confirmed a general reduction in all phases of the cell cycle among CF samples compared to their CO counterparts (Fig 4g). Taken together, the reduction in proliferation of the basal cells

of the surface hBE cells has important implications for airway repair in CF and cellular gene targeting of long-lived stem/progenitor cells in CF.

Discussion.

Taken together, we report both novel identification of proximal airway epithelial basal, secretory, and ciliated molecular subtypes and insights about the transcriptional differences at the single cell level between airways from CF and control subjects. Histological reports have described basal cell hyperplasia in the CF airways[28,29], but this was not corroborated in our study. We attribute this discrepancy to the increased sampling power and sample access associated with our study, in which we observed similar proportions of basal cells in airways of CO and CF lung, but find a notable decrease in the dividing basal subtype in CF airways. Interestingly, it is the secretory cell subtype associated with mucosal immunity that is most highly upregulated in CF. Also, we report that secretory cells account for the largest fraction of CFTR transcript expression among all cell populations in both control and CF samples. In addition, the ciliated cells were found to have an increased number of transitioning precursors in the lungs of CF donors compared to controls suggesting that they may have more plasticity than their counterparts from CO donors. Lastly, upon examining the hBE cell ALI model system, we found that the diversity of cell subtypes in ALI cultures is different to that found in fresh tissue, presumably due to the effect of a uniform culture microenvironment.

By leveraging the analysis of 38 patient samples across a 3-institution consortium and assessing gene expression patterns that are common between datasets, we have generated molecular atlases of control and CF proximal airway epithelium. This molecular atlas was used to examine CF-lung disease dependent changes in the transcriptional phenotype of lung epithelial cells but can also be utilized as a hypothesis generating tool for other airway conditions. Our data suggest that specific subtypes of the main airway cell types have potential to play a role in CF lung disease, although in vitro and in vivo validation is still needed to assess the functional potential of these cell subtypes in health and disease. These studies provide valuable, novel insights into the molecular pathogenesis of CF lung disease and have potential to impact development of new therapies to ameliorate CF-related airway dysfunction.

Figure legends.

Figure 4-8. Single cell transcriptome atlas of the epithelium lining proximal airways of control donors and donors with end-stage CF lung disease. (a) Locations of cell procurement for single-cell RNA sequencing. (b) Methodology used for cell isolation by each institution. (c) Dimensional reduction of data generated from freshly isolated control and CF airway epithelium, visualized by UMAP, with cells colored by subtypes as shown in key. (d) Distribution of cell subtypes by institution. (e) Scaled expression of the top differentially expressed genes that inform specific cell subtypes, for k-groups of control and CF cells further separated by subtype, visualized by heatmap. (f) Dimensional reduction of data generated from air liquid interface cultures (ALI) derived from samples shown above. Cells are colored by ALI-specific subtypes, shown in key at right. (g) Heatmap of the scaled expression of the same fresh tissue subtype genes from (e), but shown for groups of ALI- control and CF cells split by subtype. (h-j) Comparison of subtype-specific gene expression among fresh tissue subtypes and cultured cells. (k) Distribution of the average proportion of cell subtypes per sample, comparing CO and CF cells (l-o) CFTR expression in subtype groups, key at right. (l) CFTR expression across all subtypes, shown on the UMAP projection and as a boxplot of CO/CF versus expression level (m) Proportion of CFTR expressing cells per each subtype. (n) Proportion of CFTR+ cells per cell subtype. (o) Boxplots showing the distribution of CFTR expression in all subtypes, for CFTR+ cells only, divided by CO and CF status. P values shown at right indicate the significance of distribution differences between CO and CF per subtype, bolded if p value < 0.05.

Figure 4-9. Expansion of secretory function, including mucus secretion and antimicrobial activity, in cystic fibrosis secretory cells. (a) Dot plot indicating the expression level and frequency of differentially expressed genes from each secretory subtype, across all subtypes in CO and CF cells. Genes are expressed higher in either CO or CF, as indicated by label at top. (b) For gene networks preferentially located in secretory cells, shown is a gene ontology heatmap of the top 3

associated terms for each network with the term enrichment –log(p-value) colored as displayed in key. Networks with no associated ontology terms are blank (Net S6/S7). (c) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Secretory or non-Secretory, and CO or CF classification (d) Bar plots showing the average expression of all genes in individual secretory networks per secretory subtype, in CO or CF cells.

Figure 4-10. Cilia related gene expression is vastly expanded outside of the main cilia subgroups in CF. (a) Dot plot indicating the expression level and frequency of differentially expressed genes in each ciliated subtype, for CO or CF cells. (b) For gene networks preferentially expressed in ciliated cells, shown is a gene ontology heatmap of the top 3 associated terms for each network with the term enrichment –log(p-value) colored as displayed in key. (c) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Ciliated or non-Ciliated, and CO or CF classification (d) Bar plots showing the average expression of all genes in individual ciliated networks per ciliated subtype group, in CO or CF cells. (e) For distinct categories of genes related to cilia biogenesis, the expansion of cilia gene expression is shown by a heatmap indicating the proportional percent change in amount of cells in each subtype expressing each category above a threshold, towards CF(+%) versus CO(-%) cells. The percent change number between CF and CO samples is given in each heatmap cell and colored as indicated in key at right.

Figure 4-11. Depletion of metabolic stability, basal epithelial function, and cellular division is widespread in CF lung basal cells. (a) Dot plot indicating the expression level and frequency of differentially expressed genes in each basal subtype, for CO or CF. (b) For gene networks highly expressed in basal cells, shown is a gene ontology heatmap of the top 3 associated terms for

each network with the term enrichment –log(p-value) colored as displayed in key. (c) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Basal or non-Basal, and CO or CF classification (d) Bar plots showing the average expression of all genes in individual basal networks per basal subtype group, in CO or CF cells. (e) Immunostaining for KRT5 (green) and PCNA (red) in sections from CF and CO lung tissue. Nuclei are stained with DAPI. Arrow indicate points of interest, while insets show magnification of the basal cell layer. (f) Quantification of KRT5+ PCNA+ basal cells in CO and CF. (g) Expression distributions of cell cycle genes in CO and CF cells, in the proliferating Basal2 subtype.

Figure S4-8. The distribution of cells from each institution on UMAP projections showed homogeneous data integration. (a) Visualization of the distribution of cells from the three institutions in the integrated embedding, showed by institution and (b) by samples of origin, visualized by UMAP. (c-f) Network distributions with differences between institutions, visualized by UMAP. (g) Major cell types identified using previously described markers, visualized by UMAP. (h) Ionocyte and NE cell clusters analyzed independently of other cell types, visualized by UMAP. (i) CO and CF sample contribution to cell populations and subclusters, visualized by a stacked column chart. (j) Signatures of major cell types in ALI cells, created using previously published ALI gene lists, shown by violin plots. (k) Distribution of major cell type proportions in freshly isolated and ALI datasets. (l) Proportion of CFTR expressing cells, key at right, visualized by a stacked column chart.

Figure S4-9. Expression differences between secretory gene networks S5 and S6. (a) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Secretory or non-Secretory, and CO or CF classification (b) Bar plots

showing the average expression of all genes in individual secretory networks per secretory subtype group, in CO or CF cells.

Figure S4-10. Expression differences between ciliated gene networks C5-C10. (a) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Ciliated or non-Ciliated, and CO or CF classification (b) Bar plots showing the average expression of all genes in individual ciliated networks per ciliated subtype group, in CO or CF cells.

Figure S4-11. Expression distribution change of distinct gene categories stratified by CO and CF subtypes. (a-j) For distinct categories of genes related to cilia biogenesis, the expansion of cilia gene expression is shown by violin plots and UMAP, indicating the changes in CO and CF for each cell subtype.

Figure S4-12. Signature gene expression visualized on UMAP for selected Basal gene networks. (a) For each cell, the average mean expression of the genes in a given network is shown, visualized on a UMAP. Cells are split by Basal or non-Basal, and CO or CF classification (b) Bar plots showing the average expression of all genes in individual basal networks per basal subtype group, in CO or CF cells.

Figure S4-13. PCNA-proliferative index of KRT5-immunoreactive cells in CF proximal airways was significantly reduced compared to comparable airway regions of CO tissue. (a) Representative IF images of airways showing KRT5 (green) and PCNA (cyan), all nuclei are counterstained with DAPI (blue) in the merged image. (b) Representative examples of watershed segmentation for isolated KRT5 and PCNA staining. (c) Representative images indicating counting of KRT5 (green) and PCNA (cyan) expressing cells in the segmented images. Red and

yellow boxes highlight areas that provide 4x zoomed images. (d) Segmentation data assumes a normal distribution. Each data point represents a possible cell and its corresponding area. Red line represents the mean area of the data and black line represents two standard deviations above the mean area. Representative tiles scan regions taken at 20x magnification for non-CF (e) and CF (f) subjects stained for KRT5 (green), PCNA (cyan) and nuclei are counterstained with DAPI (blue). Dimensions of the airways are indicated by the white lines.

Figure S4-14. Utilization of FACS for isolating epithelial cells. Representative FACS for isolation of epithelial cells to use in scRNAseq with 10X Genomics. Cell debris were excluded on the basis of FSC-A versus SSC-A, then doublets were removed using Trigger Pulse Width versus FSC-A (Influx). Dead cells were identified and excluded on the basis of staining with DAPI. Negative gating for CD45, CD31, and CD235a, combined with positive gating for EPCAM (CD326) were used to identify epithelial cells.

Figures

Figure 4-8 – Single cell transcriptome atlas of the epithelium lining proximal airways of control donors and donors with end-stage CF lung disease.

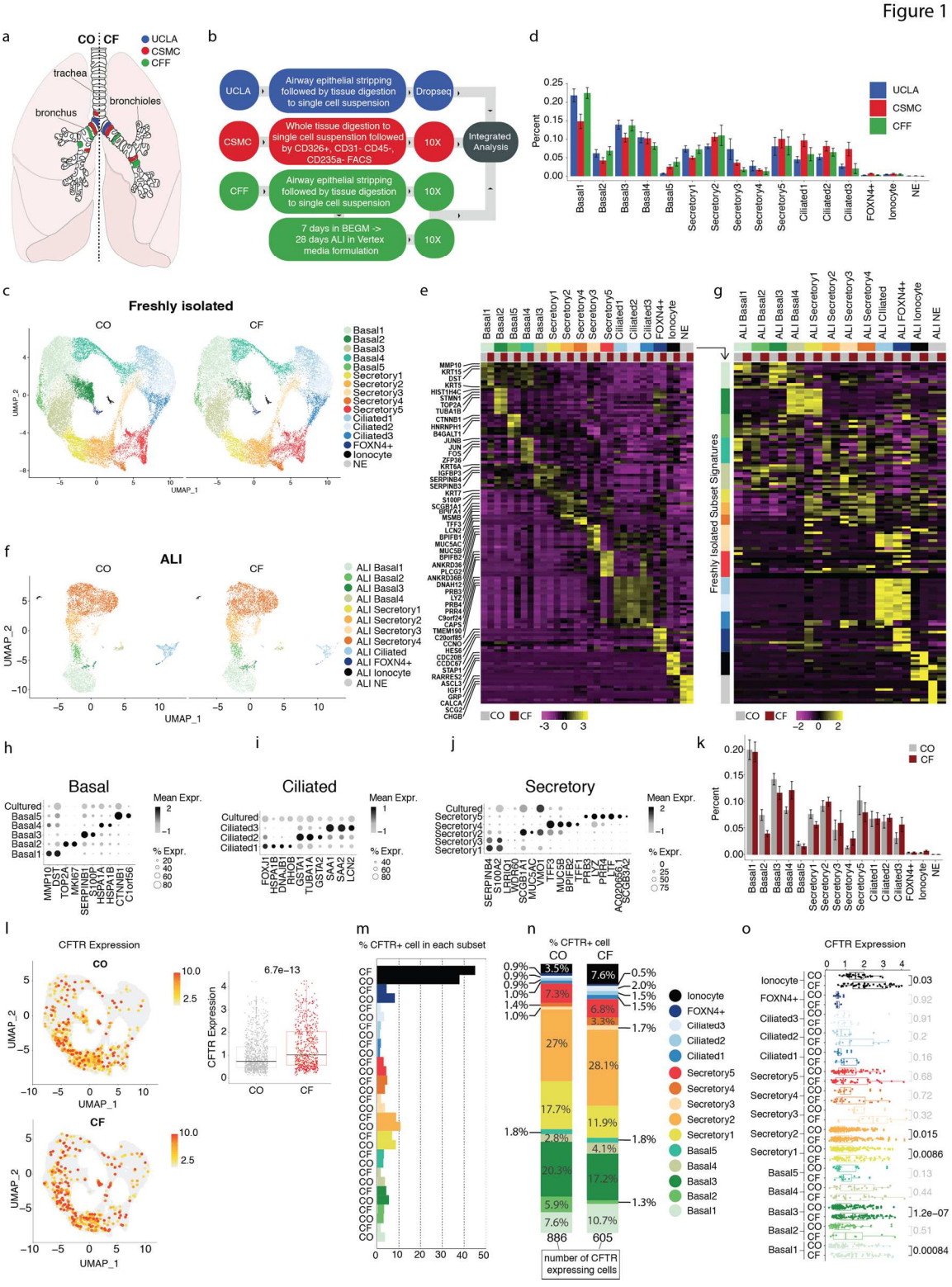


Figure 4-9 – Expansion of secretory function, including mucus secretion and antimicrobial activity, in cystic fibrosis secretory cells

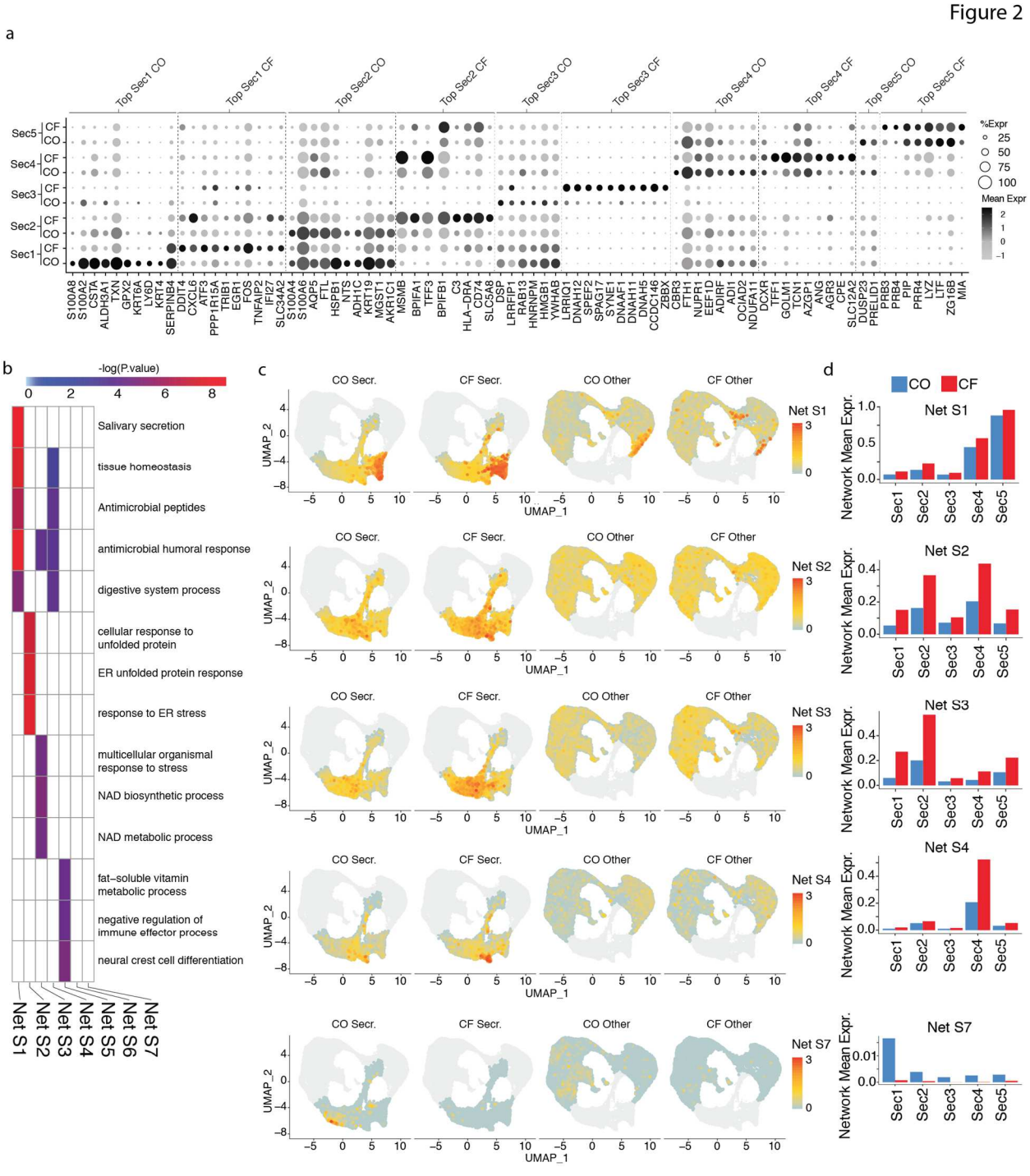


Figure 4-10 – Cilia related gene expression is vastly expanded outside of the main cilia subgroups in CF

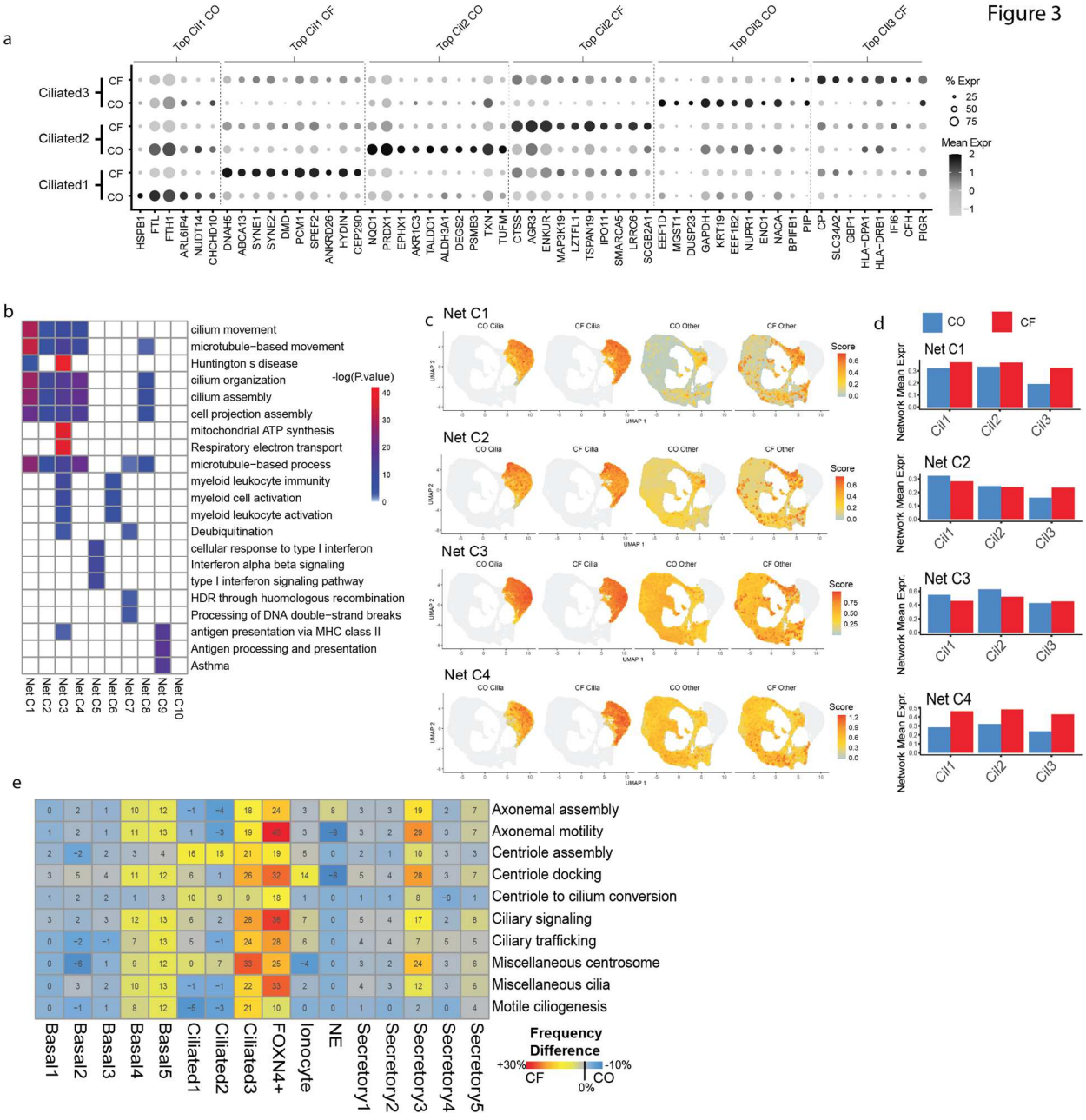


Figure 4-11 – Depletion of metabolic stability, basal epithelial function, and cellular division is widespread in CF lung basal cells

Figure 4

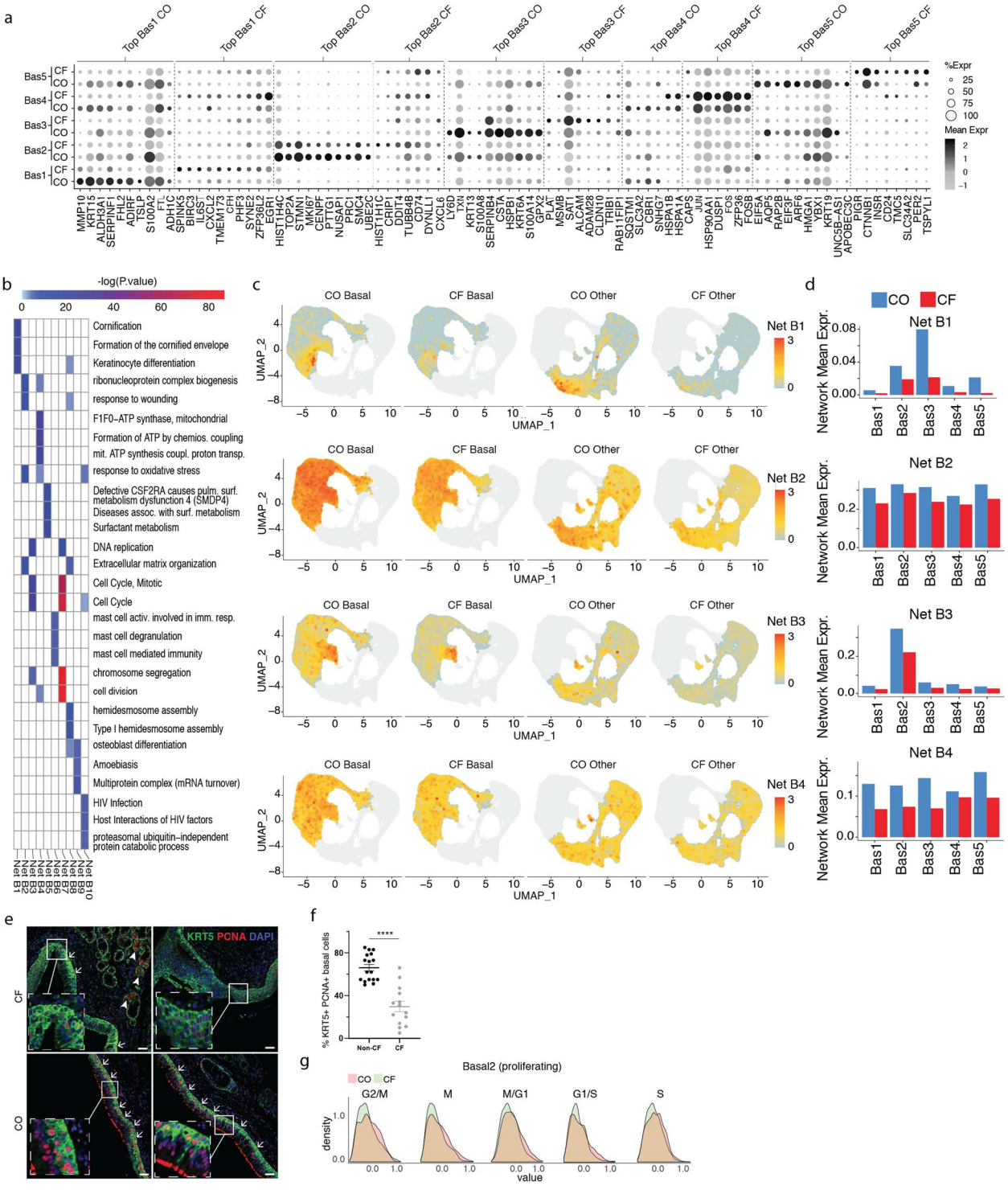


Figure S4-8 – The distribution of cells from each institution on UMAP projections showed homogeneous data integration

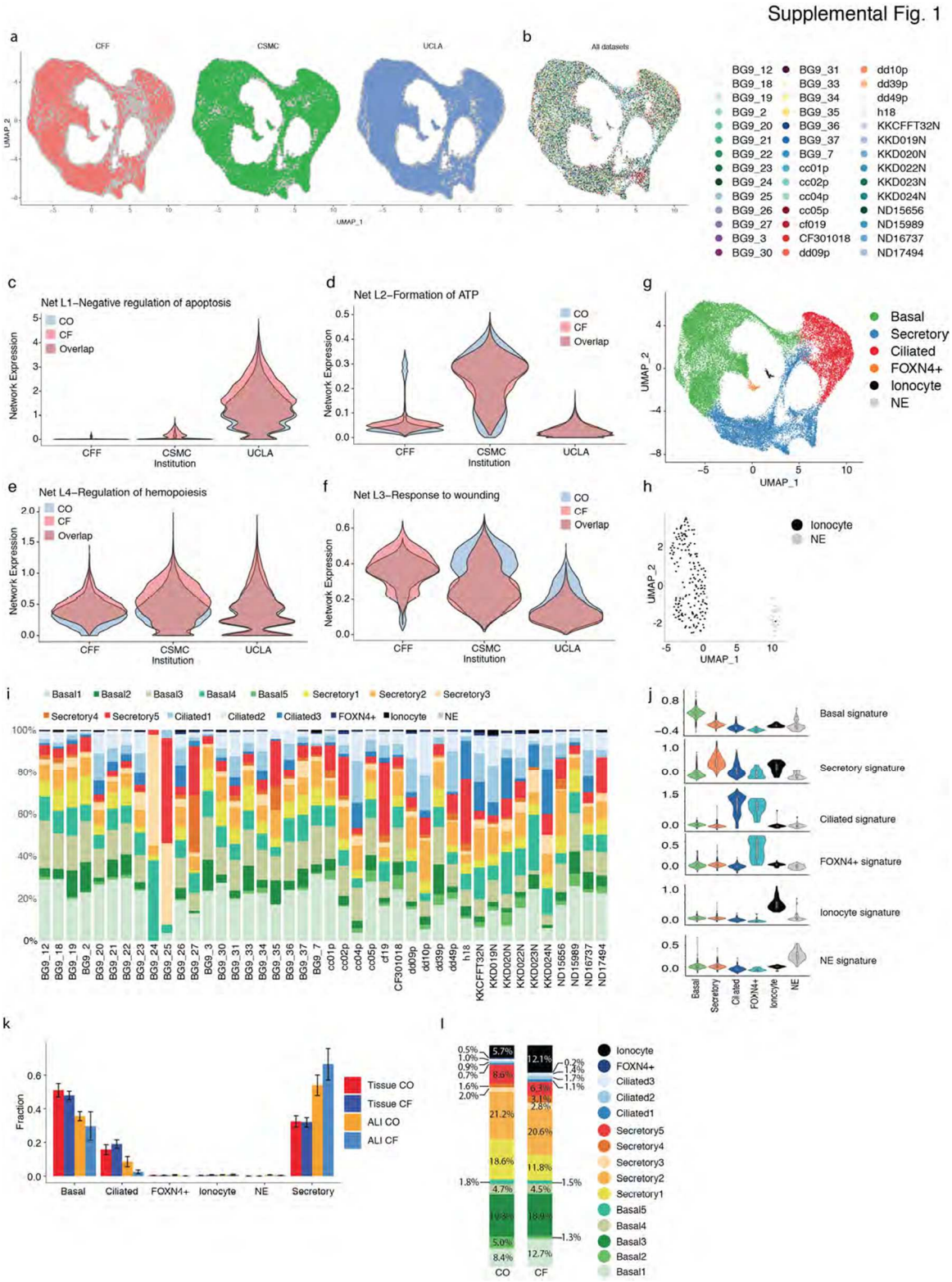


Figure S4-9 – Expression differences between secretory gene networks S5 and S6

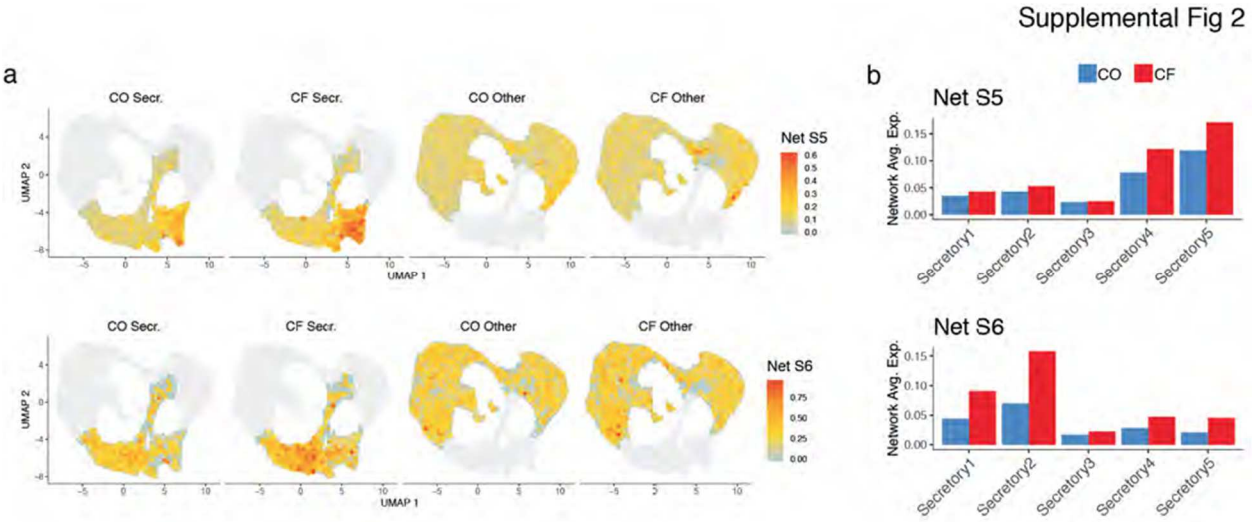


Figure S4-10 – Expression differences between gene ciliated networks C5-C10

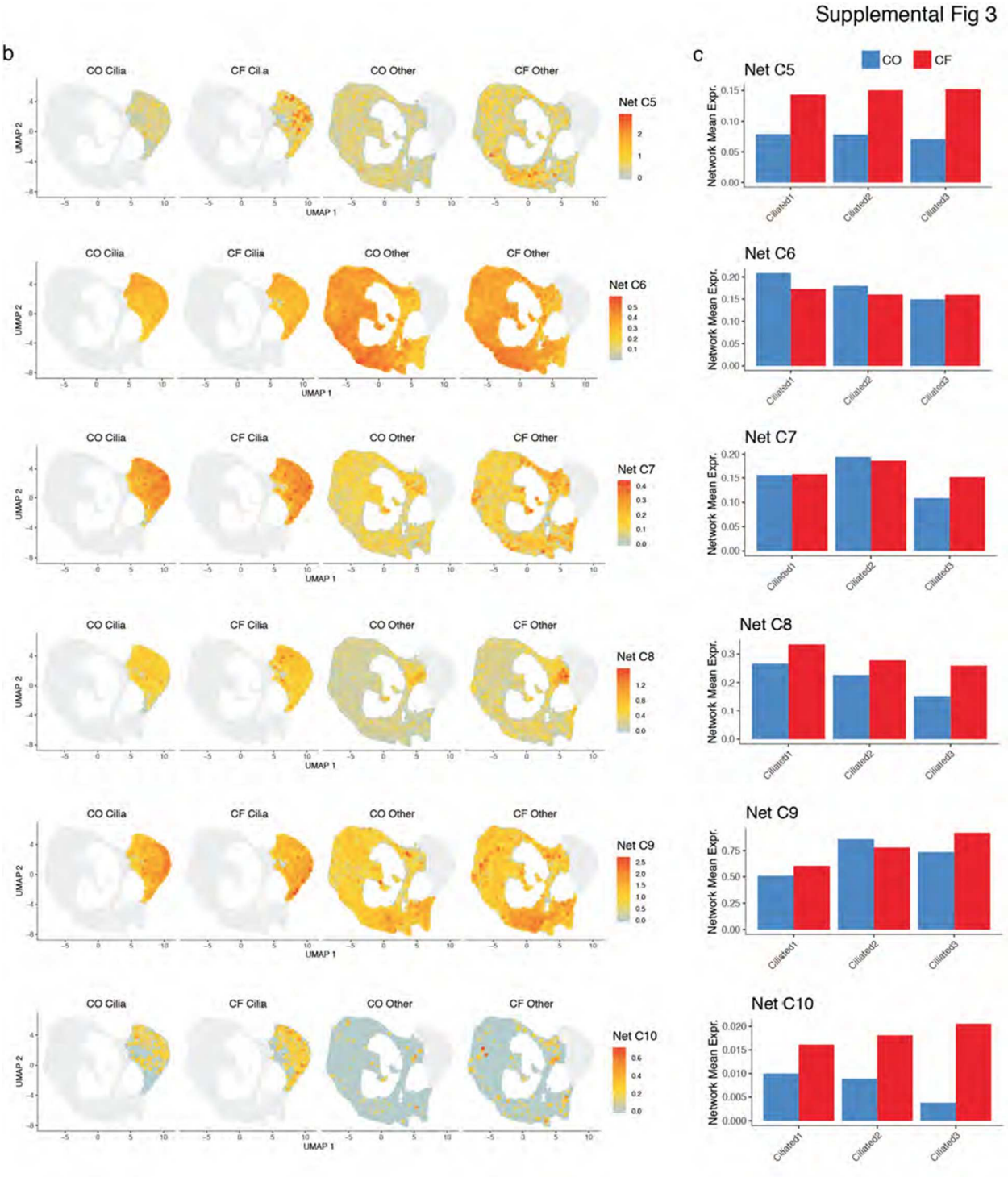


Figure S4-11 – Expression distribution change of distinct gene categories stratified by CO and CF subtypes

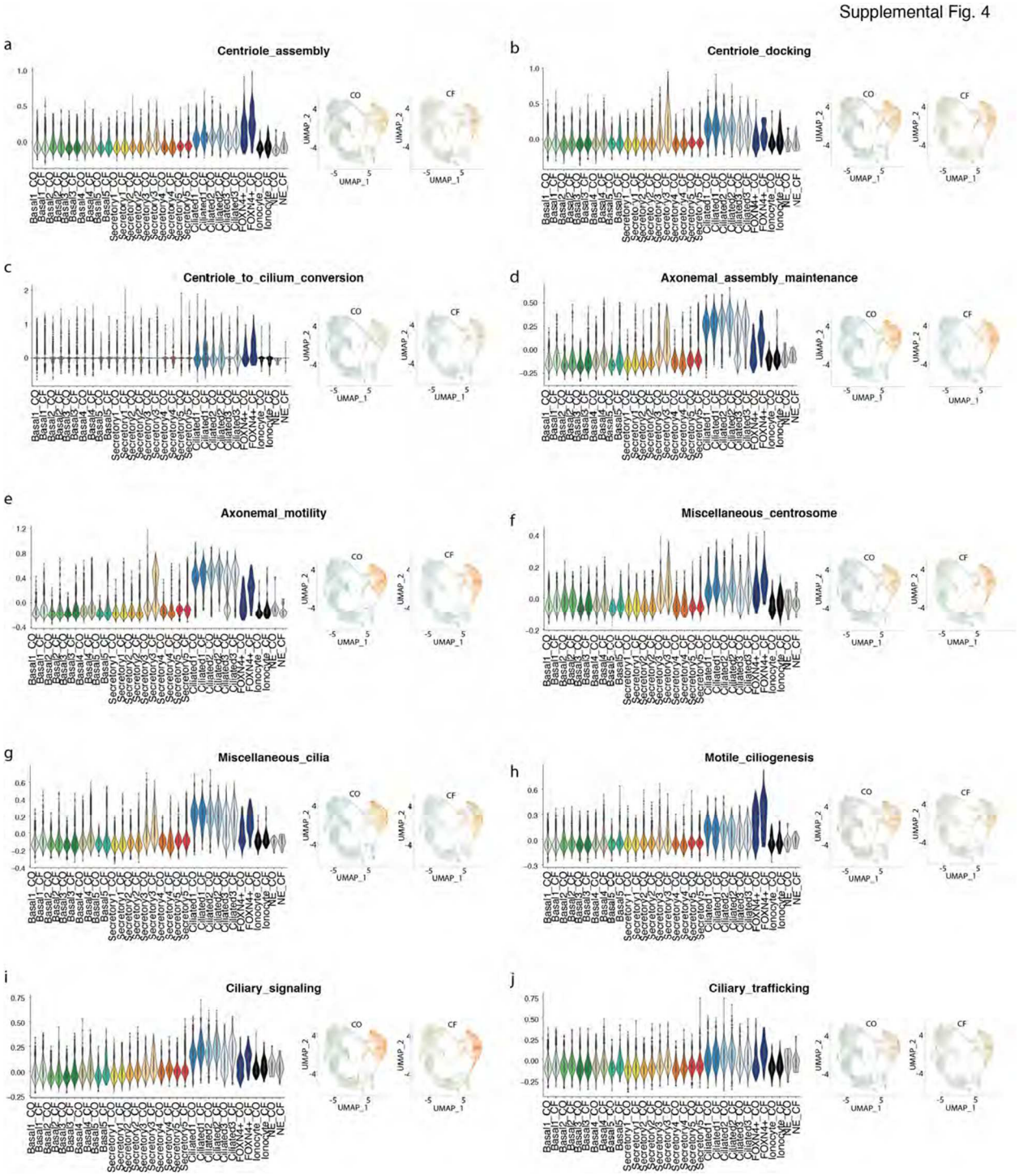


Figure S4-12 – Signature gene expression visualized on UMAP for selected Basal gene networks

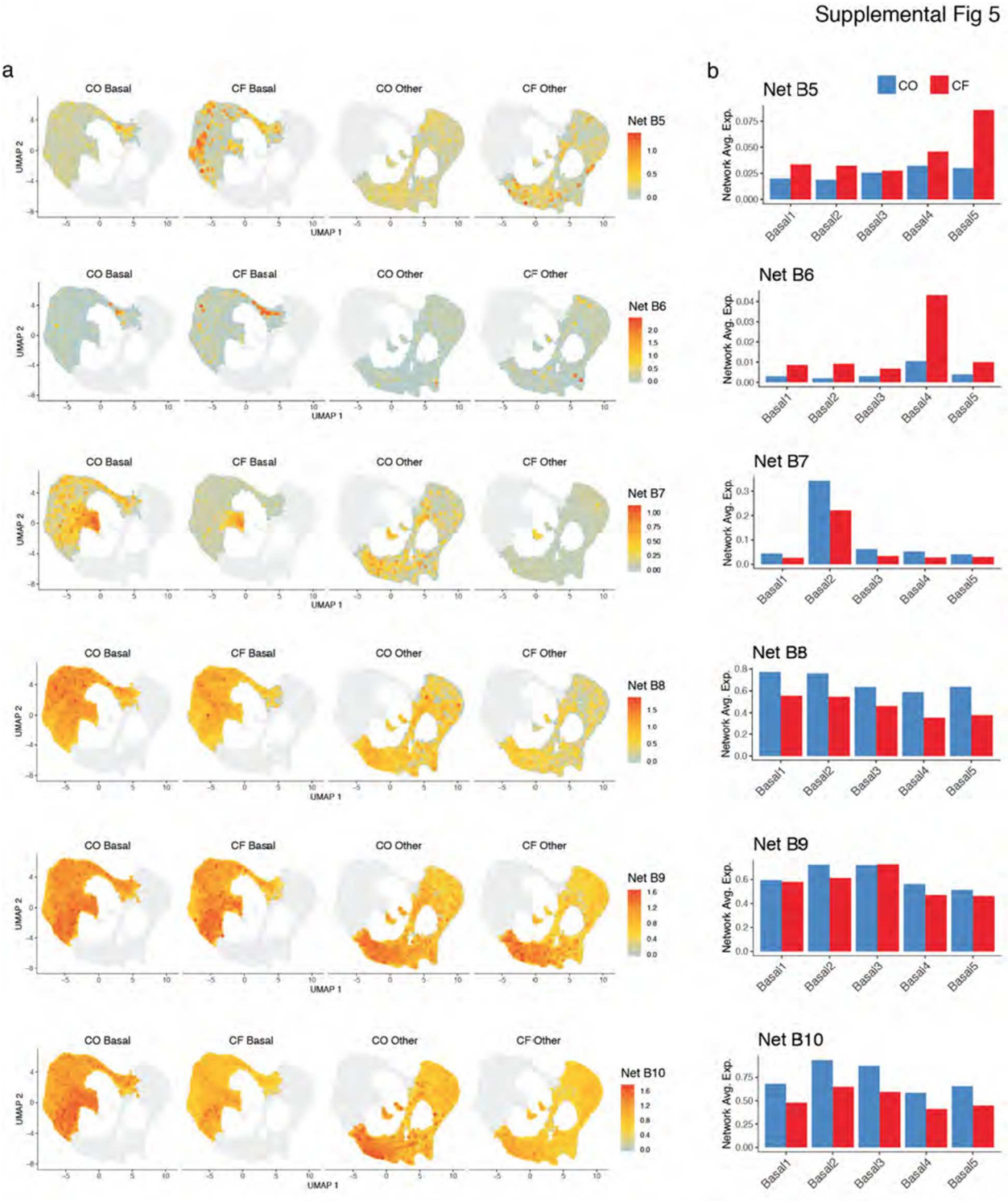


Figure S4-13 – PCNA-proliferative index of KRT5-immunoreactive cells in CF proximal airways was significantly reduced compared to comparable airway regions of CO tissue

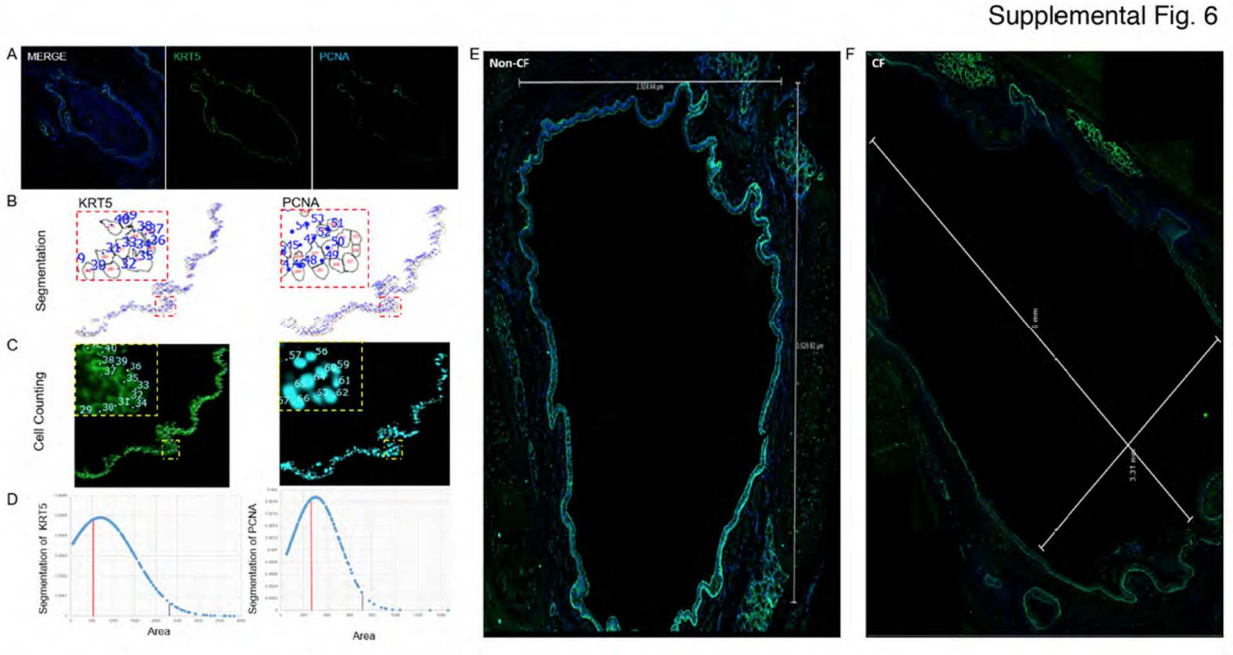
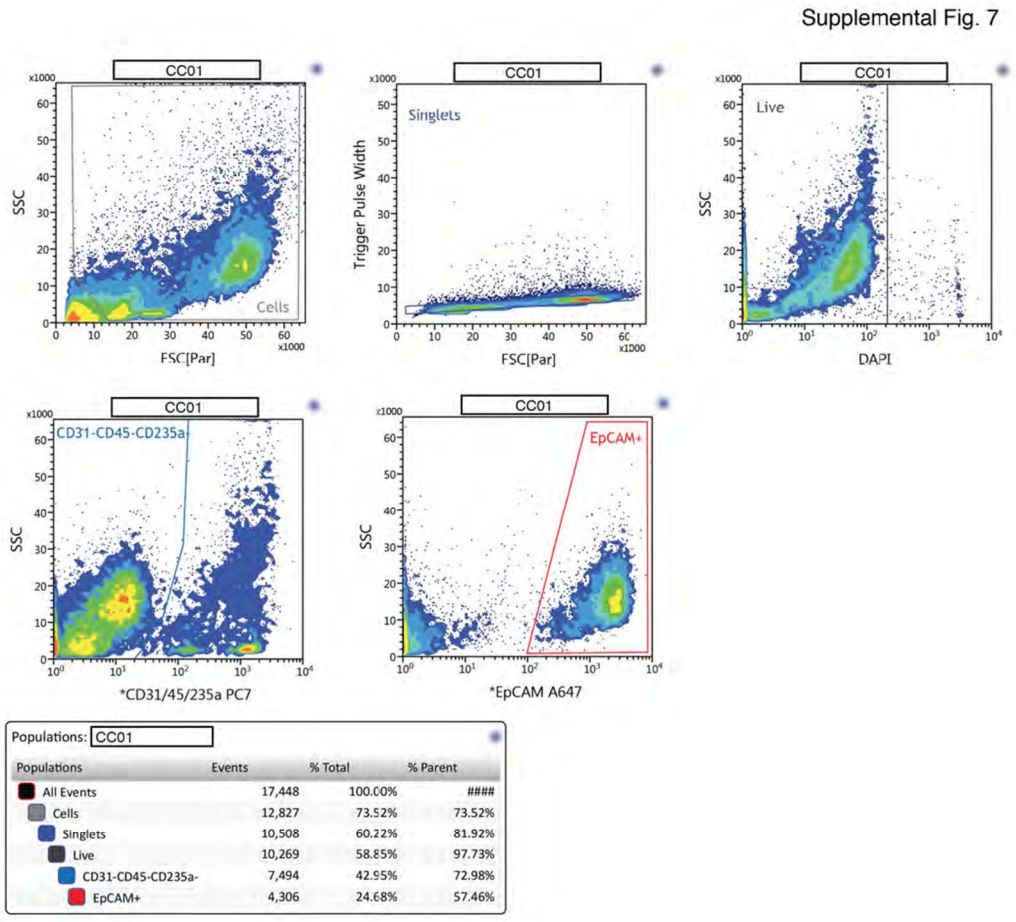


Figure S4-14 – Utilization of FACS for isolating epithelial cells



Method Details

Study population

Human lung tissue was obtained at the Cedars-Sinai Medical Center (CSMC) and at the University of North Carolina at Chapel Hill (UNC) Cystic Fibrosis Center Tissue Procurement and Cell Culture Core. Cystic fibrosis tissue was obtained from donors undergoing transplantation, while human lungs unsuitable for transplantation were obtained from Carolina Donor Services (Durham, NC), the National Disease Research Interchange (Philadelphia, PA), or the International Institute for Advancement of Medicine (Edison, NJ). Human lung tissues were procured under protocols #98-1015 and #03-1396 approved by the CSMC IRB and the UNC Biomedical IRB, respectively. Informed consent was obtained from lung donors or authorized representatives.

Human lung tissue at the Cystic Fibrosis Foundation (CFF) site, was obtained from the National Disease Research Interchange (NDRI), or from University of Texas Southwestern (UTSW) under IRB protocol approved by the CFF and WCG-Copernicus Group wIRB (Study# 1172286). Informed consent was obtained and maintained by NDRI. CF explant tissue at the University of California Los Angeles (UCLA) site was obtained from donors undergoing lung transplantation with end stage CF at UCLA, University of Southern California (USC), University of Iowa or UNC in compliance with each institutions IRB. All samples from UCLA were deidentified under protocol IRB#16-000742. Control de-identified lung specimens were obtained from lung transplant donors obtained from UCLA, USC or University of Iowa.

Data availability

All transcriptome data were deposited in GEO: Accession number pending.

Histology

Proximal airway from control donors and CF explant tissues were fixed in formalin for 24 hours, embedded in paraffin and sectioned at 10 µm thickness. Sections were deparaffinized at 60℃ followed by washes in Xylene (VWR 89370-088) and rehydrated through a gradient of decreasing ethanol concentration (Fisher Scientific BP28184). Heat-induced epitope retrieval was performed using a steamer (Hamilton-Beach 37530) in antigen retrieval solution (Vector Laboratories H-3301). Slides were blocked with 5% normal donkey serum and normal goat serum in IF buffer (1x PBS/1% BSA/0.3% Triton™ X-100) for 1 hour at room temperature, and incubated with primary antibodies, PCNA (Cell Signaling, 13110), and KRT5 (Biolegend, 905901) overnight at 4℃. After washes in 1xTBS sections were incubated with secondary antibody for 1 hour at room temperature followed by incubation in DAPI (ThermoFIsher D1306). Sections were mounted in Fluomount G (SouthernBiotech 0100-01) and imaged at 20x magnification using a Leica DMi8. Tile scans covering the entire section were created using Leica's LAS X software (Leica Microsystems, Germany). Tile scans were cleaned using Photoshop (Adobe Inc., San Jose, CA) by creating a masking layer to select for KRT5 expressing cells and from this KRT5 mask, PCNA expressing cells were isolated (Supp Fig. S4-13). These images were then converted to 8-bit and analyzed on FiJi (Image J with plugins)(Schindelin et al. 2012) by setting appropriate thresholds, creating a binary mask, and performing a watershed segmentation (Supp Fig. S4-13). Segmented images were then measured, and counts obtained using a minimum area of 100 pixels and a maximum area of two standard deviations above the mean area of pixels (Supp Fig. S4-13). The basal cell proliferative index was obtained by dividing the number of isolated PCNA-immunoreactive nuclei by the total number of KRT5-immunoreactive cells. Representative tile scan images are shown in Supp Fig. 6 for CO and CF subjects, respectively. All data were compared using an unpaired student's t-test; results were considered significant when $p<0.05$.

Cell isolation

Tissue at the CSMC site was processed to generate single cell suspensions of isolated epithelial cells as described previously[30], with the following modifications. Tissue was enzymatically digested with Liberase followed by gentle scraping of epithelial cells off the basement membrane. Remaining tissue was then finely minced and washed with rocking in Ham's F12 (Corning) at 4C for 5 minutes, followed by centrifugation at 4C for 5 minutes at 600g. Minced cleaned tissue was then incubated in DMEM/F12 (Thermo Fisher Scientific) containing 1X Liberase (Sigma-Aldrich), incubated at 37C with rocking for 45 minutes. Dissociated cells recovered by scraping or by tissue mincing were then combined and epithelial cells enriched in a two-step process involving 1). Magnetic bead (MicroBeads, Miltenyi Biotec) depletion of erythrocytes, leukocytes and endothelial cells using antibodies to CD235a (MACS, CD235a 130-050-501), CD45 (MACS, CD45 130-045-801, Miltenyi Biotec), CD31 (MACS, CD31 130-091-935, Miltenyi Biotec). FACS enrichment of epithelial cells based upon negative surface staining for CD235a (HI264, 349106), CD45 (2D1,368522), and CD31 (WM59,303124) (Biolegend) and positive staining for CD326 (CO17-1A, 369820) (Biolegend). Stained cells were washed in HBSS containing 2% FBS, resuspended and placed on ice for fluorescence-activated cell sorting (FACS) using a BD Influx cell sorter (Becton Dickinson) (Supp Fig S4-14). Viability was determined by staining cell preparations with either 7AAD (Biolegend), Propidium Iodide (Biolegend) or DAPI (ThermoFisher Scientific), 15 minutes prior to cell sorting.

Tissue at the CFF site was processed as previously described9. Briefly, large airways (8 mm in diameter and larger) were rinsed with PBS and soft tissue and lung parenchyma was dissected away, exposing intrapulmonary airways. Isolated airways were cut into ~2-3 cm segments and along their longitudinal axis to expose the airway lumen. Post dissection, tissue was collected and washed in ice cold PBS supplemented with 65mg diothreitol (DTT) and 1.25 mg of Deoxyribonuclease I (DNase). Tissue was then washed with cold basal BronchiaLife Airway media (LifeLine Cell Technology, catalog # LL-0023), prior to digestion for 6-24hr in 0.25% Protease XIV (Sigma) supplemented with ACT-V [Amphotericin B (Sigma, catalog# A2942),

Antibiotic-Antimycotic (Gibco, catalog#15240-062) Ceftazidime HCL (Sigma, catalog# C3809), Tobramycin (Sigma, catalog# T4014), and Vancomycin (Sigma, catalog# V8138)]. After digestion the luminal side of bronchial tissue was scrapped using a convex scalpel and rinsed to remove airway epithelial cells. Isolated airway epithelial cells were then either: 1) Treated with Accumax (Sigma, catalog# A7089) to yield a single cell suspension and processed for single cell transcriptional analysis, or 2) Plated and grown on collagen coated flasks in BronchiaLife Media + ACT-V until clearance of bacterial / fungal infections. Standard culture techniques followed, using complete BronchiaLife media.

Tissue at the UCLA site was processed as previously described (Hegab et al., 2012a; Hegab et al., 2014; Hegab et al., 2012b; Paul et al., 2014, Aros et al., 2020). Tissue from the bronchi and carina were dissected, cleaned, and incubated in 16U/mL Dispase for 1 hour at room temperature. Tissues were then incubated in 0.5mg/mL DNase for another hour at room temperature. The airway epithelium was then stripped and incubated in Accumax (Sigma, catalog# A7089) for 1 hour with shaking at 37°C, cells were filtered, centrifuged at 800g for 5 mins and the cell pellet was resuspended in media to a single cell suspension before being used immediately for Dropseq. For submucosal gland microdissection, the remaining tissue after airway epithelial stripping was left in Liberase at 4C overnight (diluted fresh 1:40 with PBS from 2.5mg/ml stock) and submucosal glands recovered by microdissection. Isolated submucosal glands were digested in trypsin for 30 mins to yield a single cell suspension. An equal volume of media was added to neutralize the Trypsin and filtered through 40um filter to generate a suspension of single cells. Cells were centrifuged at 800g for 5 mins, the cell pellet was suspended in media and then immediately processed for Dropseq.

Generation of air liquid interface cultures

Human bronchial epithelial cells (hBE) were cultured as previously described9. Briefly, after initial airway expansion in BronchiaLife (LifeLine Cell Technology, catalog # LL-0023) on BioCoat

collagen coated t-75 flasks (Corning, catalog# 356487), cells were lifted by Versene (Gibco, catalog# 15040-066) followed by Accutase (Sigma, catalog# SCR005) incubations, and plated to transwell filter membranes (Corning, catalog# 3470). hBE seeding density of transwell filters was 5.0*105/cm2 in BronchiaLife media for 24hrs, followed by media change to the Vertex ALI9 media formulation. Cultures remained submerged for first 96hrs, prior to removal of apical chamber and initiated the ALI time course. hBE ALI cultures were maintained for 28 days, with 48hrs media changes. On day 28, hBE ALI samples were collected by a thorough PBS wash followed by incubation in AccuMax (Sigma, catalog# A7089) for 1-2hrs followed by microscopic evaluation until a single cell suspension was identified. After a wash with cold PBS, cells were passed through a 40mm filter and counted prior to single cell capture and RNA sequencing.

Single cell library generation and sequencing

Single cells at the CSMC and CFF sites were captured using a 10X Chromium device (10X Genomics) and libraries prepared according to the Single Cell 3' v2 or v3 Reagent Kits User Guide (10X Genomics, https://www.10xgenomics.com/products/single-cell/). Cellular suspensions were loaded on a Chromium Controller instrument (10X Genomics) to generate single-cell Gel Bead-In-EMulsions (GEMs). Reverse transcription (RT) was performed in a Veriti 96-well thermal cycler (ThermoFisher). After RT, GEMs were harvested, and the cDNA underwent size selection with SPRIselect Reagent Kit (Beckman Coulter). Indexed sequencing libraries were constructed using the Chromium Single-Cell 3' Library Kit (10X Genomics) for enzymatic fragmentation, end-repair, A-tailing, adapter ligation, ligation cleanup, sample index PCR, and PCR cleanup. Libraries QC was performed by the Agilent Technologies Bioanalyzer 2100 using the High Sensitivity DNA kit (Agilent Technologies, catalog# 5067-4626) and quantitated using the Universal Library Quantification Kit (Kapa Biosystems, catalog# KK4824. Sequencing libraries were loaded on a NextSeq 500 (Illumina) for the CFF site and a NovaSeq 6000 (Illumina) for the CSMC site.

At UCLA, cells were resuspended in 0.01% BSA in 1xPBS at approximately 150 cells/ul. Cells were coflowed with barcoded beads (Chemgenes) in a Flowjem microfluidics device (PDMS Drop-seq) and isolated for reverse transcription as described according to the Drop-Seq protocol[31]. Libraries were constructed with KAPA polymerase and Nextera XT preparation kit as previously described and paired-end sequenced on a HiSeq 4000 (Illumina).

Data analysis

For the CSMC and CFF sites, Cell Ranger software (10X Genomics) was used for mapping and barcode filtering. Briefly, the raw reads were aligned to the transcriptome using STAR[32], using a hg38 transcriptome reference from GENCODE 25 annotation. Expression counts for each gene in all samples were collapsed and normalized to unique molecular identifier (UMI) counts, yielding a large digital expression matrix with cell barcodes as rows and gene identities as columns.

At UCLA, raw sequencing data were filtered by read quality, adapter- and polyA-trimmed, and reads satisfying a length threshold of 30 nucleotides were aligned to the human genome using Bowtie2. Aligned reads were tagged to gene exons using Bedtools Intersect (v2.26.0). DGE matrices were then generated by counting gene transcripts for all cells within each sample using custom Python scripts. Cell barcodes were merged within 1 Hamming distance.

Data analysis was performed with Seurat 3.0[33] with some variation that will be described. For all data, quality control and filtering were performed to remove cells with low number of expressed genes (threshold n>=200) and elevated expression of apoptotic transcripts (threshold mitochondrial genes < 15%). Only genes detected in at least 3 cells were included. Each dataset was run with SoupX analysis package to remove contaminant ‘ambient’ RNA derived from lysed cells during isolation and capture (Young MD et al., https://doi.org/10.1101/303727). Correction was performed on the basis of genes with a strong bimodal distribution and for which the ‘ambient’ RNA expression was overlapping with a gene signature of a known cell type. The ‘adjustCounts’ function of SoupX was used to generate corrected count matrices. To minimize doublet

contamination for each dataset quantile thresholding was performed to identify high UMI using a fit model generated using the multiplet's rate to recovered cells proportion, as indicated by 10X Genomics (https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-canbe-profiled-). The raw expression matrix was processed with SCTransform wrapper in Seurat. Mitochondrial and ribosomal mapping percentages were regressed to remove them as source of variation. Each dataset was first processed separately with Principal Component Analysis (PCA) using the 5000 most variable genes as input, followed by clustering with Leiden algorithm[34] using the first 30 independent components and a resolution of 0.5 for clustering. Two-dimensional visualization was obtained with Uniform Manifold Approximation and Projection (UMAP)[35]. Identified AT2 (SFTP+), immune (CD45+), and endothelial (PECAM1+) contaminating clusters were removed by subsetting the Seurat object, using the 'subset' function, before proceeding to data integration. After removal of contaminating cells, the raw expression matrix was processed with SCTransform. Log1p logarithmically transformed data were obtained for each dataset and scaled as Pearson residuals. Pearson residual data were then used to integrate datasets following Seurat workflow, using the PrepSCTIntegration function. Integrated datasets were used for downstream analysis. Datasets were processed with PCA using the 5000 most variable genes as input, followed by clustering with Leiden algorithm using the first 30 independent components and a resolution of 3 for fine clustering. Two-dimensional visualization was obtained with UMAP. To identify differentially expressed genes between clusters, Modelbased Analysis of Single-cell Transcriptomics (MAST)[36] was used within Seurat's FindMarkers function. For this analysis the p-value adjustment was performed using Bonferroni correction based on the total number of genes. To identify major cell types in our normal integrated datasets, previously published lung epithelial cell type specific gene lists[10] were used to create cell type-specific gene signatures using a strategy previously described[37]. All analyzed features were binned based on averaged expression and the control features were randomly selected from each bin. Clusters identified with the Leiden algorithm were

assigned to major cell types on the basis of rounds of scoring and refinement. Each refinement was produced using transcripts differentially expressed within the best identified clusters from the previous scoring. Within each major cell type, Leiden clustering and differential gene expression were used to infer sub-clustering. Gene lists used as cell type- and cluster-specific signatures are shown in supplementary table. Violin plots show expression distribution and contain a boxplot showing median, interquartile range, and lower and upper adjacent values.

For all genes, institution-based CO to CF ratios were calculated across all cells with thresholds set at 25% increase/decrease for being classified as reproducibly changing. Gene expression network discovery was implemented by computing a Pearson gene-to-gene correlation matrix and identified genes with correlation values greater than 0.20. Genes were then grouped into subnetworks based on their pairwise connections and then collapsed into larger networks based on correlation of subnetwork gene expression into a final set. Expression threshold differences of networks was determined by applying a cutoff to all cell's average expression of a network, set at 30% of the third max cell's expression level, for CO and CF cells separately to determine percentile of each cell in each subtype cluster, and then subtracting them to report the difference in those percentiles. Gene ontology enrichments were determined using the Metascape tool[38].

References

1. Rock, J.R., Randell, S.H. & Hogan, B.L. Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. *Dis Model Mech* **3**, 545-556 (2010).
2. Pan, S., Iannotti, M.J. & Sifers, R.N. Analysis of serpin secretion, misfolding, and surveillance in the endoplasmic reticulum. *Methods Enzymo*l **499**, 1-16 (2011).
3. Rokicki, W., Rokicki, M., Wojtacha, J. & Dzeljijli, A. The role and importance of club cells (Clara cells) in the pathogenesis of some respiratory diseases. *Kardiochir Torakochirurgia Pol* **13**, 26-30 (2016).
4. Chen, G., et al. SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. *J Clin Invest* **119**, 2914-2924 (2009).
5. Widdicombe, J.H. & Wine, J.J. Airway Gland Structure and Function. *Physiol Rev* **95**, 1241-1319 (2015).
6. Yu, X., Ng, C.P., Habacher, H. & Roy, S. Foxj1 transcription factors are master regulators of the motile ciliogenic program. *Nat Genet* **40**, 1445-1453 (2008).
7. Horani, A., et al. Establishment of the early cilia preassembly protein complex during motile ciliogenesis. *Proc Natl Acad Sci U S A* **115**, E1221-E1228 (2018).
8. Ather, J.L., et al. Serum amyloid A activates the NLRP3 inflammasome and promotes Th17 allergic asthma in mice. *J Immunol* **187**, 64-73 (2011).
9. Neuberger, T., Burton, B., Clark, H. & Van Goor, F. Use of primary cultures of human bronchial epithelial cells isolated from cystic fibrosis patients for the pre-clinical testing of CFTR modulators. *Methods Mol Biol* **741**, 39-54 (2011).
10. Plasschaert, L.W., et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381 (2018).
11. Montoro, D.T., et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319-324 (2018).
12. Xia, C., Braunstein, Z., Toomey, A.C., Zhong, J. & Rao, X. S100 Proteins As an Important Regulator of Macrophage Inflammation. *Front Immunol* **8**, 1908 (2017).
13. Akram, K.M., et al. An innate defense peptide BPIFA1/SPLUNC1 restricts influenza A virus infection. *Mucosal Immunol* **11**, 1008 (2018).
14. Thomas, J., et al. Transcriptional control of genes involved in ciliogenesis: a first step in making cilia. *Biol Cell* **102**, 499-513 (2010).
15. Mihalj, M., et al. Differential Expression of TFF1 and TFF3 in Patients Suffering from Chronic Rhinosinusitis with Nasal Polyposis. *Int J Mol Sci* **20** (2019).
16. Eckmann, L. Defence molecules in intestinal innate immunity against bacterial infections. *Curr Opin Gastroenterol* **21**, 147-151 (2005).
17. Bals, R., Weiner, D.J. & Wilson, J.M. The innate immune system in cystic fibrosis lung disease. *J Clin Invest* **103**, 303-307 (1999).
18. Tang, A.C., et al. Endoplasmic Reticulum Stress and Chemokine Production in Cystic Fibrosis Airway Cells: Regulation by STAT3 Modulation. *J Infect Dis* **215**, 293-302 (2017).
19. Petraki, C.D., Papanastasiou, P.A., Karavana, V.N. & Diamandis, E.P. Cellular distribution of human tissue kallikreins: immunohistochemical localization. *Biol Chem* **387**, 653-663 (2006).
20. Brooks, E.R. & Wallingford, J.B. Multiciliated cells. *Curr Biol* **24**, R973-982 (2014).
21. Hoh, R.A., Stowe, T.R., Turk, E. & Stearns, T. Transcriptional program of ciliated epithelial cells reveals new cilium and centrosome components and links to human disease. *PLoS One* **7**, e52166 (2012).
22. Inaba, Y., et al. Transport of the outer dynein arm complex to cilia requires a cytoplasmic protein Lrrc6.*Genes Cells* **21**, 728-739 (2016).

23. Bonser, L.R., et al. The Endoplasmic Reticulum Resident Protein AGR3. Required for Regulation of Ciliary Beat Frequency in the Airway. *Am J Respir Cell Mol Biol* **53**, 536-543 (2015).
24. Wells, J.M. & Watt, F.M. Diverse mechanisms for endogenous regeneration and repair in mammalian organs. *Nature* **557**, 322-328 (2018).
25. Teixeira, V.H., et al. Stochastic homeostasis in human airway epithelium is achieved by neutral competition of basal cell progenitors. *Elife* **2**, e00966 (2013).
26. Tezuka, T., Takahashi, M. & Katsunuma, N. Cystatin alpha is one of the component proteins of keratohyalin granules. *J Dermatol* **19**, 756-760 (1992).
27. O'Shaughnessy, R.F., et al. AKT-dependent HspB1 (Hsp27) activity in epidermal differentiation. *J Biol Chem* **282**, 17297-17305 (2007).
28. Voynow, J.A., Fischer, B.M., Roberts, B.C. & Proia, A.D. Basal-like cells constitute the proliferating cell population in cystic fibrosis airways. *Am J Respir Crit Care Med* **172**, 1013-1018 (2005).
29. Leigh, M.W., Kylander, J.E., Yankaskas, J.R. & Boucher, R.C. Cell proliferation in bronchial epithelium and submucosal glands of cystic fibrosis patients. *Am J Respir Cell Mol Biol* **12**, 605-612 (1995).
30. Xu, Y., et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* **1**, e90558 (2016).
31. Macosko, E.Z., et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
32. Dobin, A., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
33. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
34. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).
35. Becht, E., et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2018).
36. Finak, G., et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015).
37. Tirosh, I., et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309-313 (2016).
38. Zhou, Y., et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).