Spring 5-15-2022

Unraveling Population Heterogeneity using Single-Cell Analysis

Wenjun Kong
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Samantha A. Morris, Chair
Cristina De Guzman Strong
José E. Figueroa-López
Nancy L. Saccone
Nathan Stitziel

Unraveling Population Heterogeneity using Single-Cell Analysis

by
Wenjun Kong

A dissertation presented to
the Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May, 2022
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

I am extremely thankful to my PI and mentor, Samantha Morris, who gave me the opportunity to join the lab, grow, and learn to become a scientist. During my Ph.D. training, Sam was always generous with her time, supportive of new ideas, and patient with my writing. I cherish the training from Sam since our first brainstorming meeting to the last defense preparation. Without Sam, this work would not have been made possible and I would not be the person I am today.

I would like to express my great appreciation to my committee – Dr. Nancy Saccone, Dr. Cristina De Guzman Strong, Dr. José E. Figueroa-López, and Dr. Nathan Stitziel. Each member of my thesis committee has offered helpful and constructive feedback to shape and refine this work. Thank you especially to Dr. Saccone, the chair of my committee, who has been a great mentor since day one of graduate school and given me scientific guidance and support. I would also like to thank Dr. Eric Reyes and Ms. Margaret Hurdlik, my undergraduate professor and mentor, who has continued to provide career and personal guidance to this date.

I would like to express my deepest gratitude to the Morris Lab, alumni and present members, who have been the source of inspiration and support in this work. The lab is full of amazing people, nurturing collaborations, new ideas, and motivate new discoveries. I am fortunate to be a part of this group. Thanks especially to Yuheng Fu, Kunal Jindal, Xue Yang, Emily Holloway, Kenji Kamimoto, and Guillermo Rivera-Gonzalez for their patience, support, and assistance with both experimental and computational parts of this work; and to Chuner Guo, Sarah Waye, and Brent Biddy for their kind guidance and support since I joined the lab. I am very lucky to have established collaborations with the lab of Dr. Jeffery Magee, Dr. Andrew

Yoo, and Dr. Esteban Mazzoni at NYU, who have expanded my horizon, knowledge, and curiosity in life science.

I want to extend my sincere appreciation to the Computational and Systems Biology program and the Department of Developmental Biology and Genetics for offering a great training environment. Thank you especially to Jeanne Silvestrini, Sara Holmes, Dr. Ting Wang, and Dr. Dantas Gautam for their support and leadership of the program during my time.

I am grateful to have been supported by the Douglas Covey Fellowship from Department of Developmental Biology. This work in was supported by National Institutes of Health (NIH) grants R01-GM126112; Silicon Valley Community Foundation, Chan Zuckerberg Initiative Grants HCA2-A-1708-02799; The Children's Discovery Institute of Washington University and St. Louis Children's Hospital MI-II-2016-544 and DR2019726.

In my life, many friends and family have offered unconditional encouragement and support. In particular, I want to thank Catie, Christy, Chuner, Emily, Hanyue, Huiming, Jiang, Lei, Sadie, Sarah, and Xue for their continuous support and company during this journey. I would like to express my extreme gratitude to my parents, Huanhong and Qingwei, who have given me unconditional love and allowed me to pursue my dreams. To James, thank you for always being there for me while I was frustrated and for your love, kindness, and patience. Lastly, I would like to thank Einsteinie, my dog, and all the fur animals I have encountered during this journey for bringing me much happiness.

Wenjun Kong

Washington University in St. Louis

May 2022

ABSTRACT OF THE DISSERTATION

Unraveling Population Heterogeneity using Single-Cell Analysis

by

Wenjun Kong

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2022

Dr. Samantha A. Morris, Chair

The human body contains approximately 100 trillion cells, encompassing distinct cell types that serve diverse functions. Understanding cell population heterogeneity is vital for uncovering different biological functions and mechanisms. In addition, cells at transition during continual processes, such as development, reprogramming, and disease, are essential for painting the entire blueprint and highlighting critical stages of the progression trajectory. For instance, cell fate engineering holds much promise for generating clinically valuable cell types from mature somatic cells. Nonetheless, current reprogramming protocols are inefficient, and charting the changes in cell identity during such processes can help design strategies to mitigate the off-target and increase efficiency. RNA-sequencing allows us to study transcript abundance and dissect different genetic features. Prior to single-cell level sequencing, bulk-level transcriptomics have demonstrated power at a lower resolution to distinguish populations and identify differential gene markers. The advent of single-cell RNA-sequencing technologies has brought us a new era of exploring the small world inside individual cells via their transcriptome profiles. Single-cell RNA-sequencing takes a snapshot of individual cells, enabling the dissection of population composition and capture of cells at different states in complex biological systems.

Cell type annotation has been a long-standing interest in understanding cell identities from gene profiles. Yet, manual annotations require prior knowledge of cell-type-specific gene signatures and are labor-intensive and time-consuming. Automated annotation approaches are in demand for exponentially growing single-cell datasets.

In response to such demand, many computational approaches have been developed. However, they classify cells in a discrete, categorical manner, limiting their application in continuous biological systems. Focusing on continual processes, we designed a computational tool, 'Capybara,' to measure cell identity as a continuum at a single-cell resolution. This approach enables the classification of discrete cell identities and recognizes cells harboring hybrid identities, supporting a quantitative cell-fate transition metric. After benchmarking against other classifiers and validation with "ground-truth" lineage data, we apply Capybara to a diverse range of cellular programming and reprogramming protocols: The application to direct cardiac reprogramming uncovers a patterning bias and a hybrid state between atrial and ventricular cardiomyocytes; Capybara reveals previously uncharacterized patterning deficiencies in motor neuron programming, instructing a new approach to alleviate the lack of proper patterning; Further, we apply Capybara to our in-house system, direct reprogramming of fibroblast to induced endoderm progenitors, and find a putative in vivo correlate for this engineered cell type that has, to date, remained poorly defined. These findings highlight the utility of Capybara to dissect cell identity and fate transitions in development, reprogramming, and disease. Finally, we further explore the direct cardiac reprogramming system using the comprehensive set of tools developed in the lab. We resolve lineage relationships in this system using CellTagging, find key regulatory transcription factors using CellOracle, and evaluate small molecules' effect on the patterning bias using Capybara.

In summary, I have developed a tool to highlight cell fate transitions and reveal insight into cellular heterogeneity in different continuous biological processes. Further investigation in the transition states by integration with other data modalities and experimental approaches may help pinpoint key checkpoints for successful reprogramming, allowing future interventions to improve the efficiency and fidelity of cell fate engineering.

Chapter 1: Introduction

1.1 Background and Motivation

Complex biological systems are composed of many different cell types. These diverse cellular compositions govern specific functions and mechanisms of these biological systems (Altschuler and Wu, 2010, Paszek *et al.*, 2010). Understanding such heterogeneity is key to uncover hidden mechanisms and allow future advancements in medicine and biology, such as improvement in efficiency of cell regeneration (Takahashi and Yamanaka, 2006, Guo and Morris, 2017), insight into cell fate decisions (Stegle *et al.*, 2015), and target identification in tumors (McGranahan and Swaton, 2017). As disparate cell types transcribe unique combinations of genes, heterogeneity can be analyzed using transcriptome-based approaches (Wit, 2017). The development of transcriptome profiling technologies has enabled the capture and quantification of complete sets of transcripts in cells (Wang *et al.*, 2009).

Bulk RNA-sequencing, at the population level, has enabled insights in comparisons between different tissues, treatments, or cell populations. Yet, concerns, such as loss of cell-to-cell variation and the stochastic nature of the single-cell transcriptome (Elowitz et al., 2002), have emerged with growing demands of decoding cellular heterogeneity at a higher resolution. In the past decade, single-cell RNA-sequencing (scRNA-seq) techniques have been flourishing, enabling the field to explore the world inside of an individual cell, increasing the resolution of sequencing to a new standard (Sandberg, 2014). Starting with techniques that sequence the transcriptome following the manual isolation of one cell (Tang *et al.*, 2009), researchers have developed highly parallel scRNA-seq techniques to profile individual cells in a larger population

simultaneously (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017). This rapidly evolving technique has brought broad opportunities to assess cell composition and dynamics in complex systems at an individual-cell resolution. A demand that naturally arises with the broad adoption of these relatively nascent technologies is to catalog cell identities. Recent endeavors to create cell atlases across a range of organisms have brought interest in systematic cell type classification (Han et al., 2018, Regev et al., 2017; Tabula Muris Consortium et al., 2018). With these atlases primarily focused on defining discrete cell types of tissues under homeostasis, classifying cell identities in dynamic contexts, such as development, disease, and reprogramming, poses a challenge as cell types and states are in continual transition (Cahan et al., 2014).

Inspired by this challenge, this dissertation will focus on dissecting cellular heterogeneity via application and analysis of scRNA-seq data in complex, continuous biological systems, particularly development and reprogramming. Specifically, we will develop computational tools and apply our in-house technologies to understand heterogeneity of *in vitro* systems, and further feed into potential improvements of *in vitro* models.

The following subsections (1.1.1-1.1.4) will include a literature review, focusing on four major topics: 1) single-cell RNA-sequencing technologies, 2) cell type classification algorithms, 3) cellular reprogramming, with a subsection focusing on direct cardiac reprogramming, and 4) transition cells in continuous processes.

1.1.1 From Bulk to Single Cell, One Cell to Cell Atlases

As the genetic material being determined in the mid-20th century, efforts have been made to develop technologies to decipher DNA sequences. Starting with the genome of bacteriophage

(Sanger et al., 1982) to the Human Genome Project (Abdellah et al., 2004; Lander et al., 2001), the fidelity and efficiency of DNA sequencing have been brought to the next scale level. Such sequencing techniques have enabled the genome assembly of multiple organisms, identification of genetic variations to benefit different clinical applications, and the initiation of using sequencers to quantify molecules (Shendure et al., 2017). At the same time as the technology evolves, the definition of "gene" has grown to be more complex over the years – from the unit of heredity to the Central Dogma. The ENCODE project, seeking to interpret the human DNA sequence, revealed that protein-encoding genes occupy 1% of the entire genome while over 80% were responsible for gene activity regulation (Birney et al., 2007). In addition, quantification of RNA molecules has expanded the vision of RNA beyond single protein-coding transcripts to include multiple isoforms from one gene and non-coding RNAs, such as miRNAs, snoRNAs, etc. (Gerstein et al., 2007).

Early years of transcriptomic studies have largely relied on probe-based microarray approaches, where cDNA was hybridized to custom-made oligo probes and expression was measured via scanning of the fluorescent signal. The probes can be designed to hybrid at splicing junctions, allowing detailed mapping of transcript isoforms (Z. Wang et al., 2009). Yet, the hybridization probes rely on prior knowledge regarding the genome sequence, limiting its ability to capture the full range of transcripts. In addition, quantification of hybridization-based fluorescence can be obscured by high-background cross-hybridization (F Ozsolak, 2011; Wang et al., 2009). In 2008, RNA-sequencing (RNA-seq) was introduced via shortgun sequencing of the cDNA using next-generation DNA sequencing (NGS). The introduction of this approach allows massive scale-up in RNA molecules at test and detection of nascent transcripts without

prior knowledge to the genomic regions, offering a comprehensive view of gene expression for various cell types (F Ozsolak, 2011; Shendure et al., 2017; Z. Wang et al., 2009).

The development of RNA-seq warrants new applications in understanding cellular functions and molecular mechanisms via the lens of gene expression, such as cell growth and differentiation and regulation under different stimuli. Further, it introduced potential clinical application, such as biomarker detection or gene fusion for disease diagnosis and treatment improvements (X. Li & Wang, 2021). Studies of RNA-seq at the population level provides the opportunity to study expression patterns across different populations or species and adds information in genetic modification at the transcription layer. Yet, as bulk RNA-seq takes the average expression across the population, cell types in heterogeneous environment can be challenging to distinguish and rare cell populations can be masked (Elowitz et al., 2002; X. Li & Wang, 2021). Such concerns sparked the initiation of the new era to profile individual cells.

Single-cell RNA-sequencing (scRNA-seq) started with the capture of the entire transcriptome of a single mouse blastomere cell manually isolated from 4-cell stage. The transcriptome was reverse-transcribed and double-stranded cDNA was synthesized, amplified, sheared, and ligated with an adaptor to be sequenced on a NGS system. Using this approach, transcripts of more genes and nascent splicing junctions were detected compared to previous microarray attempts, demonstrating promising results in analyzing complex profiles of individual cells (Tang et al., 2009). In the last decade, highly parallel scRNA-seq techniques have been developed to profile individual cells in a large population simultaneously. These techniques can be broadly categorized based on isolation methods employed. Droplet-based microfluidic techniques use individual droplets as reaction chambers, where a cell is encapsulated in an oil droplet together with lysis buffer and a bead. The bead is primed with oligos including PCR

primer, cell barcode, unique molecular identifier, and poly-T capture sequence. As the individual cell lyse, it releases its mRNA within the droplet, hybridizing to the oligos on the bead. The samples are further collected, washed, and processed to produce the final libraries (X. Zhang, Li, et al., 2019). Three representative droplet-based systems include Drop-seq (Macosko et al., 2015), inDrop (Klein et al., 2015), and 10X Genomics Chromium (Zheng et al., 2017). All three strategies share similar designs to incorporate microfluidics but differ in their design of beads and primed oligos leading to different experimental processing of the library. For instance, Drop-seq uses small, inflexible beads, while inDrop and 10X use hydrogel beads, enabling higher pairing between the cells and the beads, allowing higher capture efficiency. With detachable priming oligos, 10X and inDrop enable higher capture efficiency of mRNA molecules (X. Zhang, Li, et al., 2019).

Another major type of scRNA-seq technology uses wells as reaction chambers for individual cells, such as CytoSeq (Fan et al., 2015), Microwell (J. Yuan & Sims, 2016), and Seq-Well (Gierahn et al., 2017). The wells are first loaded with a cell suspension, during which cells settle into individual wells by gravity. The barcoded and primed beads are then loaded by a similar mechanism. Microfluid and semi-permeable membranes have been used in combination with the wells to avoid cross-contamination between wells and loss of transcripts in individual wells (Gierahn et al., 2017). To further improve the throughput, split-pool combinatorial barcoding strategies have been developed, such as SPLiT-seq (Rosenberg et al., 2018) and sci-RNA-seq (Cao et al., 2017). Using the split-pool approach, cells are not required to be encapsulated in their own reaction chamber, but multiple cells are allowed in each well of 96- or 384-well plates. In brief, with each well labelled with a unique barcode, cells are fixed and first randomly split into wells and labelled with first set of barcoded primers. The cells are then

pooled and re-distributed into another well-plate to be labelled with a second set of primers. With repeating rounds of split-pool, cells are uniquely labelled with combinations of barcodes. In SPLiT-seq, four rounds of spilt and pool were performed, producing over 21 million barcode combinations, enabling the profiles of over 100,000 single cells to be captured (Rosenberg et al., 2018). While in sci-RNA-seq, the re-distribution of cells is performed with fluorescence activated cell sorting (FACS) on DAPI, ensuring 10 to 100 cells per well, before further barcode indexing (Cao et al., 2017).

Technology development and advancement lays the foundation of scientific discoveries. The rapidly improving scRNA-seq technologies enable researchers to further address the long-standing interest of understanding biology from the cells. As the basic building blocks of life, cells of different types play unique and pivotal roles in the overall function of each organ to further support essential functions of life. Leveraging scRNA-seq technologies, comprehensive sequencing of cells from different organs at homeostatic conditions in *Mus Musculus* has been flourishing since 2018, empowering the identification of cell types and subtypes and global analysis of same cell types across tissues (X. Han et al., 2018; Schaum et al., 2018). In addition to the adult mouse map, continuous effort has been made to chart transcriptomic changes sampling at multiple time points during gastrulation and organogenesis, mapping out major developmental trajectories (Cao et al., 2019; Pijuan-Sala et al., 2019). Beyond mouse, the Human Cell Atlas (HCA) was first proposed in 2017 (Regev et al., 2017). Since then, collaborative efforts have been made to sequence different cells in various human systems, such as lung (Schupp et al., 2021), heart (Litviňuková et al., 2020), intestine (Elmentaite et al., 2021), pancreas (Tosti et al., 2021), immune system (Suo et al., 2022) etc. In light of the COVID-19 pandemic, the meta-analysis of these single-cell data has shed light in understanding the

pathology, venue and target of infection, and immune responses (Loske et al., 2021; Melms et al., 2021; Stephenson et al., 2021). In addition to adult cells, the HCA was further expanded to the Human Development Cell Atlas (HDCA), empowering the molecular basis for human organogenesis and development (Cao et al., 2020; Haniffa et al., 2021).

Accompanying the advancement in technologies and exploration of many systems using single-cell sequencing (scRNA-seq), the volume of data generated has escalated exponentially. Compared to gene expression profile in bulk, transcriptomes within individual cells posed new challenges in analysis. Due to the low number of starting transcripts, scRNA-seq data is usually noisy and shallow (Cao et al., 2019; X. Han et al., 2018). To allow detection of low transcriptome levels, the mRNA is reversed transcribed, and the cDNA is amplified for them to be detected, leading to potential distortions and bias (Sandberg, 2014). In addition, low-abundance transcripts, such as those encoded by transcription factors and regulatory genes, fail to be detected in scRNA-seq dataset, presenting large proportion of zero values and sparse matrices. These unique features of single-cell data posed a demand in development of statistical and computational frameworks (Andrews et al., 2021).

To tackle potential quantitative bias introduced during cDNA synthesis and amplification, unique molecular identifiers (UMIs) were introduced to become a direct and more accurate measurement of gene expression compared to RNA reads. The usage and incorporation of UMI barcodes in the bead-primed oligos assists the quality of downstream analysis (Andrews et al., 2021; Islam et al., 2014). After the libraries are sequenced, the data was processed to produce the cell-barcode matrices, which is analyzed to further uncover the biological relevance. While multiple frameworks have been developed over the years, this work mainly used the 10x Genomics Cell Ranger pipeline to process the raw sequencing data (Zheng et al., 2017). In

general, data is filtered for correct cell barcodes, aligned to transcriptome using STAR aligner, further filtered to include unique-hit genes and correct UMI sequences, and counted to create the final UMI count matrices (Zheng et al., 2017). The matrices generated are next preprocessed to reduce noise and support further biological interpretation of the data.

Single-cell RNA-seq matrices are analyzed via the general five steps, including quality control (QC), normalization, feature selection, dimensional reduction, and clustering. Quality control steps involves addition selection of the genes as well as cells. Genes and cells are selected based on the thresholds of minimum number of cells expressing the gene and minimum number of genes detect per cell, respectively (Andrews et al., 2021). Further, mitochondria gene percentage is used to remove potential damaged or dying cells.

A critical next step in analyzing RNA-seq data is normalization. RNA-seq data needs to be normalized and scaled against different factors, such as library size and depth of reads, prior to further analysis. As a result of the intrinsic stochasticity of single-cell gene expression (Elowitz et al., 2002), single-cell datasets show high cell-to-cell variability, making normalization a particularly important step in analysis. Bulk transcriptome data is usually normalized to a size factor, determined based on the sequencing depth of each sample. Though a similar approach may apply to single cells, single cell data is more prone to variation due to multiple factors, such as diverse gene expression levels in different cell types, uneven sequencing depths among cells, cell cycle stages, and high sparsity, etc. One approach to normalization is to robustly estimate the size factor by pooling across cells (Lun et al., 2016). An alternative involves the usage of synthetic spike-in RNAs, External RNA Control Consortium (ERCC) (Jiang et al., 2011), or expression of housekeeping genes (Anders & Huber, 2010; Robinson & Oshlack, 2010). Yet, to have known and controlled concentration, synthetic spike-in

RNAs are mostly amenable to well-based single-cell technologies while hard to adapt for droplet-based approaches. In addition to normalizing across the entire dataset, other approaches, such as scTransform (Hafemeister & Satija, 2019), account for the expression of each gene to take in the differences between highly and lowly expressed genes (Andrews et al., 2021).

Single-cell matrices demonstrate the feature of high dimensionality, represented by the inclusion of ~20,000 genes. Yet, a lot of genes are not expressed or captured in an experiment depending on different cell types, protocols, or technologies. Such highly sparse datasets make downstream analysis challenging, such as inaccurate distance metrics and uncomfortable matrix calculations. To alleviate such challenges, feature selection is a key step to select genes of strong biological relevance to reduce noise and simplify further analysis (Andrews et al., 2021). The common approach used in the field is to choose highly variable genes based on cell-to-cell variations. However, focusing solely on the variance of expression profile could overlook the technical noise and dependence between the variance and mean of the data. Thus, methods, such as Seurat, fit a mean-variance relationship in the data to compute expected variances, which is then used as the metric for gene selection (Stuart et al., 2019). With such feature selection, the data is then scaled to the standard normal and processed for the proceeding dimensional reduction step.

With the initial dimensional reduction using feature selection, additional dimensional reduction steps are performed to further remove the challenges posed by high dimensionality. The feature-selected dataset first undergoes principal component analysis (PCA) and components that explain significant fractions of variance are chosen for the next step (Peres-Neto et al., 2005; S. Sun et al., 2019). Further dimensional reduction approaches, such as Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) and t-distributed Stochastic

Neighbor Embedding (t-SNE; van der Maaten & Hinton, 2008), are applied for visualization in two-dimensions. UMAP first constructs a manifold, a particular weighted k-neighbor graph, where the data is approximately uniformly distributed. The graph is then simplified and projected on the lower dimensional space. As the goal of UMAP algorithm is to preserve the structure of the manifold, UMAP performs better at maintaining the connected global structure of the data compared to t-SNE (McInnes et al., 2018), whose focus is to capture the local structure and separate clusters in the global space (van der Maaten & Hinton, 2008). The application of either UMAP or t-SNE can be dependent on the emphasis of the analysis. Yet, as neither preserve the cell-to-cell distances, the projection results should be used with caution in downstream analysis, such as clustering (Andrews et al., 2021).

Clustering is an essential step in analyzing scRNA-seq data to group and distinguish cells based on the similarity in expression profiles. Cells can be separated into different clusters driven by various reasons, such as similar cell types or cell cycle states. Starting with the *k*-means clustering algorithm (SC3; Kiselev et al., 2017), the single-cell field has moved to adopt graph-based clustering algorithms, such as Louvain (Blondel et al., 2008) and Leiden algorithms (Traag et al., 2019). This category of algorithms constructs a graph network among the cells and based on connectivity, identify distinct cell-modules in the graph. Identified cell groups can be projected on the two-dimensional space for visualization and a preliminary interpretation.

Downstream analyses post clustering includes three major categories: 1) differential gene expression analysis, 2) pseudotime analysis or trajectory inferences, and 3) cell type annotation. Differential expression (DE) analysis is an essential analysis to distinguish the genes that separate different cell groups. Yet, traditional methods comparing averaged populational expressions in bulk RNA-seq are challenging as single-cells represent gene expression across

cells as a distribution. From the non-parametric Wilcoxon rank sum test to scRNA-seq specific method (e.g., MAST; Finak et al., 2015), different statistical tests have been proposed and evaluated to identify significant expression differences in distributions between populations. Another approach is to average the gene expression across cell populations such that the averaged profiles can be compared, similarly to bulk RNA-seq data (Baran et al., 2019). As diverse methods emerge, DE analysis enables further biological interpretation of the data, such as identification of potential markers or perturbation targets to specific populations.

During a continuous biological process, it is potentially inappropriate to place cells into distinct buckets of cell types as they follow a continual trajectory. Pseudotime analysis concerns cells collected from such biological processes, such as development and reprogramming. A common approach used in trajectory inference methods includes two steps: 1) construction of a manifold of the data in a lower dimensional space, and 2) identify a graph that best capture the topology of the manifold (Andrews et al., 2021; Saelens et al., 2019). For instance, Monocle reduces the data to lower dimensionality, build a minimum-spanning tree (MST), and orders the cells following the MST (Trapnell et al., 2014). Over the years, trajectory inference methods have been flourishing to produce better representations of continuous biological processes (e.g., PAGA, Wolf et al., 2019; Slingshot, Street et al., 2018; etc.). Evaluation of these methods has offered comprehensive criteria based on the data to choose the appropriate algorithm for optimal inference (Saelens et al., 2019).

A natural question that follows the clustering step is to ask “What are these cells? What are their *in vivo* correlates?” Starting from manual annotation focusing on marker expression, many automated cell-type classification algorithms have emerged over the past years. In the next section, I will review the current existing algorithms and the challenges to be tackled.

1.1.2 Classification Algorithms

Before annotating cell identity, a key question needs to be discussed: what is the definition of cell type or cell identity? It has been a long-lasting interest to break down and decipher the cells that constitute living organisms. Starting from histological approaches, the idea of cell identity has been constantly growing with the advancement of technologies to include multiple facets that define cell type. The identity of a cell has been expanded and proposed in recent years to include its phenotype and function, lineage, and present states. Phenotype and function widely include different aspects of a cell, such as molecular composition, morphology, spatial context *in vivo*, functionality, differentiation potential, and so on. The lineage aspect reveals where the cell starts, its potential destination, and its timestamp in a continuous biological process, such as development. The present cell state describes the response of a cell to its environmental stimuli, such as if it is perturbed or in homeostasis (Morris, 2019; Savulescu et al., 2020). With a complex definition of cell identity, it is worth noting that single-cell RNA-sequencing enables us to decode cell identity from the molecular angle of RNA. Yet, it might not exactly correlate to the interpretation of cellular function (Morris, 2019; Pasquini et al., 2021; Savulescu et al., 2020). Future consideration to incorporate multi-omics data could lead us to a more comprehensive blueprint of a cell's identity.

The exploration of cell identity based on RNA-seq data started in bulk. A few methods, such as over-representation analysis (ORA) and gene set enrichment analysis (GSEA), were developed to look for genotype-phenotype relations (Diaz-Mejia et al., 2019; Pasquini et al., 2021). The unique features, such as high sparsity and noise, of single-cell dataset posed new challenges and demand for annotation algorithms. Current methodologies can be broadly categorized to two types: reference-based or reference-free classification. Reference-based

algorithms normally relies on a reference dataset, such as cell atlas and marker databases. These approaches further break down to four types based on the choice of their essential building blocks, including marker gene, statistical test, tree-like relationship, and machine learning (Pasquini et al., 2021; Y. Zhang et al., 2022). Reference-free classification does not rely on a reference but tries to group cells sharing similar transcriptional patterns. In the following subsections, a selection of reference-based methods will be discussed, followed by a brief discussion of reference-free methods.

1.1.2.1 Reference-Based Methods

Inspired by bulk analysis, initial approaches to annotate single-cell datasets is based on gene markers. As aforementioned, general analysis of single-cell RNA-sequencing (scRNA-seq) data involves preprocessing, dimensional reduction, and clustering. Differentially expressed genes (DEGs) for each cluster can be identified. One approach to categorize cells is to compare the DEGs for each cluster with known cell-type specific gene set in a marker gene dataset. As scRNA-seq datasets grow and become publicly available, the cell type markers used in manual annotation in each dataset have been analyzed and deposited into available databases, such as CellMarker (X. Zhang, Lan, et al., 2019), PanglaoDB (Franzén et al., 2019), and CancerSEA (H. Yuan et al., 2019). These databases are generally constructed from extensive mining of public resources, such as PubMed and Gene Expression Omnibus (GEO). For instance, CellMarker is curated based on extensive search on PubMed, cross-confirmation across literature, and further unification by comparison to public resources, such as UniProt (Bateman et al., 2021), Human Cell Atlas (Regev et al., 2017), etc. Via exhaustive search and careful construction, CellMarker contains 9,148 markers for 389 cell types of 81 tissues in mouse and 13,605 markers for 467 cell types from 158 tissues in human (X. Zhang, Lan, et al., 2019). These databases provide valuable

resources for manual comparison and annotation of single-cell datasets using gene signatures and lay the foundation for further development of automated classification tools. Yet, manual search and comparison for marker genes can be labor-intensive and time consuming, with limited reproducibility (Diaz-Mejia et al., 2019; Pasquini et al., 2021).

Methods, such as scCATCH (Shao et al., 2020) and SCINA (Z. Zhang et al., 2019), that automate the usage of marker genes have been developed to alleviate the limitations of such manual annotation. ScCATCH gathers markers from different databases and unifies them to assemble a tissue specific cell type-marker database, CellMatch. Next, it identifies cluster-specific marker genes of the single-cell dataset and uses evidence-based scoring with CellMatch to score each cluster for each cell type. The clusters are then annotated based on the scores (Shao et al., 2020). SCINA employs a probabilistic framework based on expectation-maximization algorithm. This approach performs clustering on the sample data and assign known cell types to clusters by fitting a bimodal distribution to the expression of known gene signatures in public marker databases. In addition to known cell types, SCINA defines and captures novel cell types for clusters without high expression of any known signatures (Z. Zhang et al., 2019).

Moving away from prior knowledge of marker genes, other approaches have been developed to annotate clusters based on statistical metrics. The intuitive metric at choice is correlation assuming profile similarity between single-cell clusters and the reference. CIPR (Ekiz et al., 2020) and ClustifyR (R. Fu et al., 2020) calculate the centroids of the clusters and the average expression of each cluster as its pseudo-bulk expression. Clusters are then annotated based on correlation between the pseudo-bulk expression and the reference datasets (Ekiz et al., 2020; R. Fu et al., 2020). These approaches improve the automation and enable more accurate classification. Nonetheless, they are bounded by the definition of the clusters, which can be

affected by user-defined resolutions that potentially create within cluster heterogeneity (Pasquini et al., 2021).

Independent of clustering, methods are developed to evaluate cell identity using the transcriptome on a cell-by-cell basis. SingleR performs differential expression analysis in the bulk reference set and identifies the top variable genes. Correlation is then computed between each single-cell profile and each cell type in the bulk reference. Top correlated cell types are selected to iterate through the steps until only one top cell type remains to be the cell type of the cell (Aran et al., 2019). With cell atlases becoming publicly available, scmap leverages the high resolution of single-cell atlases, aiming to map individual cells to cells or clusters in the reference. Scmap first identifies biologically relevant features and using those features, computes similarity metrics, including cosine similarity, Pearson, and Spearman correlation. Scmap-cluster maps individual cells in the sample to centroids of clusters in the reference, while scmap-cell maps sample cells to cells in the reference. For a cell to be assigned a cell type, scmap-cell requires agreement in at least two of the similarity metrics, and such similarities hold for at least three nearest neighbors around the reference cell (Kiselev et al., 2018). Considering the correlation can be unreliable under the circumstance of high sparsity, more complex statistical metrics have been employed. For instance, SciBet selects features based on statistical entropy, where features are more cell-type specific with larger entropy. It then models the expression of each selected feature as well as expression across genes. Each sample cell profile is evaluated based on the likelihood over all models and cell type with maximum likelihood estimation is assigned to the test cell (C. Li et al., 2020).

Tree-based approaches concern the hierarchical relationships within the reference and cell types are assigned via searches traversing through the tree. These approaches share a similar

initial step to construct a hierarchical tree of the reference based on distance metrics, such as correlation or average linkage, across reference cell-type profiles (de Kanter et al., 2019; Lin et al., 2020; Pliner et al., 2019). Post tree construction, sample cells will travel from top to bottom through the tree and receive classifications. CHETAH finds the gene signatures for each cell type at evaluation in the tree, scores the input cell for each cell type, measures the confidence of the scores, and assigns a cell type (de Kanter et al., 2019). ScClassify takes a different approach of ensemble learning, taking advantages from different methods for each step. It takes an ensemble of gene selection approaches and similarity methods to establish a collection of classifiers, capturing different characteristics for a cell type. Using the bundle of classifiers collectively allows the algorithm to give accurate assignments to cells (Lin et al., 2020). Garnett requires user input that contains gene marker information. With the tree built, it employs an elastic net classifier to classify cells into different cell types. In addition to scRNA-seq data, Garnett can also classify single-cell ATAC dataset based on gene activity scores (Pliner et al., 2019).

As machine learning methodologies prosper, it has been adapted and applied in analysis of single-cell datasets with their increasing dimensionality and complexity. Machine learning-based classification tools are developed on a different basis, such as support vector machine (SVM), random forest, transfer learning, and so on (Pasquini et al., 2021). To avoid significant individual gene effect, 16cPred uses PCA-transformed gene expression matrices. It builds a SVM model on the training data and applies the model to the testing data, computing a conditional probability. Benchmarked by a threshold, the probabilities calculated allows the assignment of cell types or unknowns. The incorporation of SVM with radial kernels allows the capture of multi-collinearity and non-linear relationships in high-dimensional data (Alquicira-Hernandez et al., 2019). SingleCellNet, like CellNet (Cahan et al., 2014; Morris, Cahan, Li,

Zhao, San Roman, et al., 2014), applies a random forest algorithm. It identifies features unique to each cell type and finds the most discriminating sets of gene pairs to binarize the data. A multi-class random forest model is further trained on the binarized data for final classification. Uniquely, singleCellNet creates a randomly transformed single-cell profile that differs from any cell types in the reference, enabling the mapping of unknown cell types (Tan & Cahan, 2019). ScID utilizes Fisher's linear discriminant analysis to compute discriminative weights in each cell for each cell-type specific genes identified from the reference. The scores for gene markers are collectively evaluated to assign cells to reference clusters (Boufea et al., 2020). With the goal to achieve accuracy, efficiency, and consistency, CaSTLe implements transfer learning. It performs feature selection based on multiple filtering criteria internal to the tool and trains a XGBoost classification model on the selected features. The trained model is applied to sample scRNA-seq data for cell identity categorization (Lieberman et al., 2018). Finally, ACTINN is established on a three hidden-layer neural network, trained, and tested using Tabula Muris (Schaum et al., 2018). It has been demonstrated to have high accuracy in cell subtypes and is robust against different single-cell profiling techniques (Ma & Pellegrini, 2020).

1.1.2.2 Reference-Free Methods

Compared to reference-based methods, reference-free approaches remove the dependencies on the previously established data and build solely on the test dataset to identify patterns of cells or features that separate different groups.

Concerning the limited sequencing depth per cell in scRNA-seq and large proportion of unmapped reads, scSimClassify is a reference-free and alignment-free annotation tool. Instead of genes, features are defined as *k*-mers in the reads, which are further preprocessed and filtered to include only informative *k*-mers. The selected *k*-mers are then used to generate a *n*-bit fingerprint

and grouped into compressed *k*-mer groups (CKGs) based on similar abundances. Rather than cell-by-gene matrices, scSimClassify creates cell-by-CKG matrices that are further used to identify cell patterns and correlates to gene features (Q. Sun et al., 2021).

ScCoGAPS considers cell identity as the result from collective effects of diverse processes, leading to various cell states that may not be found in publicly available references. Focusing on the decomposition of latent spaces and comparison across multiple datasets, scCoGAPS employs non-negative matrix factorization (NMF) algorithms and transfer learning to allow target cells to align with the same latent space learned in the source and learn differential features represented in identified patterns. Applications to developing retina dataset demonstrates its ability to capture cell patterns during continual process, aligning with temporal progress during development (Stein-O'Brien et al., 2019).

Though not an exhaustive demonstration of all current available classifiers, the above discussion represents the major categories of classification algorithms. The rapid development of tools to annotate cells opens exciting opportunities for downstream analyses and generates hypotheses for experimental validations and discoveries.

As most of the classifiers attempt to categorize cells in a discrete manner, limiting their application in continuous biological processes, such as development, reprogramming, and disease. In the next subsection, we will focus on the review of cellular reprogramming and lay the foundation for understanding heterogeneity in these processes.

1.1.3 Cellular Reprogramming

In 1957, Waddington proposed a landscape model to depict development from stem cells to defined terminal fates. In this landscape, differentiation from stem cells is portrayed as a ball rolling down from the top of a mountain toward different valleys, where the top defines the

pluripotent state, and the valleys describes diverse differentiated cell identities. This model explains embryonic development to be a unidirectional, progressive, and irreversible restriction of cell fate determination (Waddington, 1957). Yet, the breakthrough in pushing cells "back up the hill" to a more potent state or "crossing the valleys" to a different fate has expanded the landscape to be more flexible in directionality and reversibility (Ladewig et al., 2013; Morris, 2019).

The spark of rewiring the cell from a differentiated to a totipotent state was initiated in 1958 by John Gurdon. Prior to this discovery, Robert Briggs and Thomas King have attempted and succeed in nuclear transfer experiments in single-celled organisms (R. Briggs & King, 1952). Following such endeavor, via somatic cell nuclear transfer (SCNT) – transferring the nucleus of one cell into an enucleated oocyte, Gurdon and his colleague demonstrated that the recombinant egg was able to successfully develop into a viable animal. These experiments demonstrated the possibility of reprogramming the somatic epigenome to pluripotency under internal cues, challenging the unidirectional Waddington landscape, laying the foundation for cellular reprogramming (Gurdon et al., 1958; Morris, 2019). In addition to efforts to push differentiated cells toward to pluripotency, it was found that a differentiated cell can be reprogrammed to another differentiated fate with the guidance of ectopic expression of transcription factors. Pioneering the field of cell fate engineering between somatic cell types was the identification of a transcription factor, *Myod1*, in 1987. The gene was identified via a screening of a pool of cDNA probes and overexpression of the transcription factor was shown to fibroblasts to myoblasts (Davis et al., 1987). Since then, large efforts have been made to identify sets of factors that can induce pluripotency and conversion across somatic cell types (H. Wang et al., 2021).

In 2006, Kazutoshi Takahashi and Shinya Yamanaka identified that overexpression of four transcription factors, including *Oct3/4, Sox2, c-Myc*, and *Klf4*, can reprogram differentiated cells to an induced pluripotent state, where cells have the potential to differentiate to multiple cell types. Considering the potential of embryonic stem cells (ESC) in the treatment of diseases, this Nobel-prize winning discovery of induced pluripotent stem cells (iPSCs) shed light in regenerative medicine to circumvent the challenges of ESC isolations (Takahashi & Yamanaka, 2006). Under proper exogenous cues with small molecules and growth factors, somatic cell-derived iPSCs can be guided to desired cell types, such as spinal motor neurons (Dimos et al., 2008) and cardiomyocytes (Lian et al., 2012). Yet, this approach is lengthy with potential tumorigenic risk with intermediate induced pluripotency. In addition, the engineered cells do not faithfully recapitulate the target cell identities (H. Wang et al., 2021).

As the intermediate pluripotent state introduces uncertainty, the previous identification of MyoD inspired a series of discoveries to bypass intermediate states and induce direct conversion between somatic cell types. Compared to going through iPSCs, direct cell fate conversions are faster with higher fidelity and have unique niche of *in vivo* tissue repair (ref). Broadly, these reprogramming protocols can be divided into two categories, including conversions between cell types originated from different germ layers or from the same germ layers (Ladewig et al., 2013).

In the Morris lab, a relevant process to cell fate reprogramming across germ layers involves the conversion form fibroblasts to the endoderm lineage, such as hepatocytes (Sekiya & Suzuki, 2011). Starting from mouse embryonic fibroblasts (MEFs), this protocol overexpresses two transcription factors, *Hnf4* and *Foxa1/2/3*. The resulting cell type was first reported as an ‘induced hepatocyte’ with the capacity to engraft damaged liver (Sekiya & Suzuki, 2011). Nonetheless, it was later found that these cells have broader potential to engraft damaged

intestine, leading to their new name, ‘induced endoderm progenitors’ (Guo et al., 2019; Morris, Cahan, Li, Zhao, San Roman, et al., 2014). Using lineage tracing and single-cell sequencing, recent work from the Morris lab has illuminated the bifurcation of the resulting population into successfully reprogrammed cells and cells that fail to achieve the target identity (Biddy et al., 2018).

Another illustrative procedure demonstrates the engineering of fibroblasts to the ectoderm lineage, such as neurons (Ladewig et al., 2013). With screening of transcription factors (TFs) in the neuronal lineage, a combination of three factors, including *Ascl1, Brn2*, and *Myt1l*, was identified to reprogram MEFs to induced neurons (Pang et al., 2011; Vierbuchen et al., 2010), The induced neurons adopt neuronal morphology with expression of neural markers and firing of action potential. Further, modulation with additional neural subtype-specific factors can further specify different subtypes of neurons, such as motor neurons (Aydin & Mazzoni, 2019; Son et al., 2011). In addition to TF-mediated reprogramming, other regulatory factors, such as micro-RNAs, have been demonstrated to induce reprogramming. For instance, the combination of *miR-9/9** and *miR-124* can introduce the switch from fibroblast to neuron identity (Yoo et al., 2011). Though mechanistically distinct between different regulatory factors, the diverse approaches further support the plasticity of cells and the ability to cross the “long-distance” barrier created by different developmental history (H. Wang et al., 2021).

Compared to cross germ layer-reprogramming, within germ layer-reprogramming concerns the trans-differentiation between cell types that are developmentally related (Ladewig et al., 2013). A relevant example involves such lineage switching is within the hematopoiesis lineage, from B cells to macrophages. Starting from committed B-cells in the lymphoid lineage, overexpression of transcription factor, *Cebp*, can induce macrophage phenotypes within 72 hours

of induction (H. Xie et al., 2004). Like neural reprogramming, addition of other factors, such as *Gata2*, can redirect lymphoid cells to other hematopoietic lineages (Iwasaki et al., 2006). However, it was found that the converted cell identity was not sustained, with marker expression of the starting cell type regained over time (Morris, Cahan, Li, Zhao, San Roman, et al., 2014).

Direct cardiac reprogramming exemplifies another protocol that engineers within the same germ layer. As it serves a major system in the second half of this thesis, I will discuss this cell fate conversion process in detail in the next subsection.

1.1.3.1 Direct Cardiac Reprogramming

Heart disease is a leading cause of death worldwide. One of the key pathologies contributing to cardiovascular disease is the loss of functional cardiomyocytes, whose post-mitotic nature limits regenerative potential (Ieda et al., 2010; Qian et al., 2012; Song et al., 2012). Induced pluripotent stem cell (iPSC) reprogramming from somatic cells opened broad opportunities in direct differentiation of mature cardiomyocytes from a stem-cell like state (Lian et al., 2012; Takahashi & Yamanaka, 2006). Yet, the inefficient conversion to iPSCs along with limited differentiation to cardiomyocytes produces a heterogeneous population, hindering the clinical utility of these cells (Ieda et al., 2010). Direct reprogramming between somatic cell types is a promising alternative for faster and more efficient production of the target cell types (H. Wang et al., 2021). Composing 50% of the heart, cardiac fibroblasts serve important structural and signaling functions in normal heart. Upon injury, cardiac fibroblasts are activated to proliferate and respond to form the scar tissue. This endogenous source of cells holds much promise for reprogramming and repair to generate functional cardiomyocytes *in vivo* (Ieda et al., 2010; Qian et al., 2013; Song et al., 2012; H. Wang et al., 2021).

In 2010, without known master regulators, a similar screening approach was employed to find appropriate transcription factors (TFs) for cardiac conversion. In brief, cardiac fibroblasts and cardiomyocytes were subject to the microarray assay, and 14 higher expressing factors in cardiomyocytes were identified as potential factors. Via iterative removal of one factor from the combination, the combination of three key transcription factors, *Gata4*, *Mef2c*, and *Tbx5* (GMT), was found to be sufficient to generate cardiomyocytes with cardiac-specific expression, sarcomere structure, and electrophysiology. Transplantation of TF-transfected fibroblasts demonstrates conversion and functional engraftment *in vivo* (Ieda et al., 2010). Nonetheless, the overall conversion efficiency remains low, ranging from 1.5% to 3% (Qian et al., 2013; Song et al., 2012). An optimal set of TFs, including GMT and *Hand2* (GHMT), was thus explored and established to reprogram with an increased efficiency between 7% and 9% (Song et al., 2012).

Taking this protocol beyond *in vitro*, *in vivo* reprogramming induced by direct delivery of TFs into diseased heart is shown to be effective to improve cardiac function in the damaged myocardium post myocardial infarction (MI). In brief, the TFs are locally delivered to the infarcted region via retrovirus, and the region is allowed to reprogram and recover. Evaluations at 12 weeks post MI show significant improvement in cardiac functionality and decrease in scar size with the TF-delivery compared to control dsRed-delivery. Notably, in contrast to *in vitro* induced cardiomyocytes, *in vivo* reprogrammed cells are reported to achieve more mature phenotypes, such as contractility, and more closely mimic resident cardiomyocytes in the heart. It is suggested that cues from the microenvironment *in vivo* could have facilitate the process to maturity, inspiring additional approaches to improve the protocol (Qian et al., 2012).

To further enhance reprogramming outcomes, chemical enhancement and additional factors have been tested to facilitate the reprogramming process. For instance, in GHMT-guided

reprogramming, addition of Akt1, protein kinase B, enhances cardiomyocyte reprogramming by 2-fold increase in efficiency and enhanced cardiac-specific transcription, structural protein, and morphology (H. Zhou et al., 2015). Additionally, alternative modulations with GMT, such as a set of two small molecules that inhibit Wnt and TGF pathways, are found to improve the efficiency by over three-fold compared to GMT only (Mohamed et al., 2017). Beyond mice, direct cardiac reprogramming from human skin fibroblast has been explored considering its potential to improve heart diseases. With further optimization, *miR-133* is identified as a key regulator with MGT to induce efficient conversion in human cells (Y. Zhou et al., 2019).

Though we have seen continuous advancement in reprogramming protocols the underlying molecular mechanisms driving this process remain unclear. Recent studies have elucidated important features of this cell fate conversion (de Soysa et al., 2019; Stone et al., 2019; Y. Zhou et al., 2019). Single-cell transcriptomics has revealed that cells during early conversion transit through a bifurcation, with one route leading to reprogramming while the other becomes refractory to cardiac conversion (Y. Zhou et al., 2019). In support of this, in joint epigenetic and single-cell analysis, transition cells are found to determine their terminal fate in the first 24-48 hours and take one of the two paths, either to cardiac fate or to non-cardiac identity resembling fibroblast or vascular developmental cell states (Stone et al., 2019). Analysis of chromatin landscape changes illustrates enhancers activated at early transition resemble mostly neonatal cardiac development, gradually moving toward postnatal and adult stages. This demonstrates that early reprogrammed cells adopt an immature identity, mimicking embryonic development, before specification and activation of mature cardiac enhancers toward an adult fate (Hashimoto et al., 2019).

The development of technologies has expedited the discoveries to understand the underlying mechanism of cardiac reprogramming, bringing us closer to future clinical applications. Current studies rely on clustering to identify cellular subpopulations before, during, and after reprogramming (Y. Xie et al., 2022). As aforementioned, the limitations of clustering-based approach could miss the identification of rare cells or cells in progress of changing. In addition, present trajectories established are inferred from gene expression. Based on current studies, we ask the following questions. Could we systematically identify and annotate these cells, including cells in transition? Using lineage tracing technologies, could we build a "ground truth" map of lineage, connecting cell ancestry via heritable cell labeling? In addition, could we identify putative TFs that assist reprogramming via gene regulatory network construction?

In a nutshell, reprogramming strategies have grown exponentially during the past two decades, laying the foundation for future regenerative medicine. Yet, the low efficiency and fidelity of the resulting cell types hinders the application in clinical settings. Understanding such heterogeneity and dissection of cell identity and state at different stages can facilitate potential methods for improvement. Laying on a continuum, cells in transition could offer invaluable insights into key stages along the processes. In the next section, we will briefly focus on transition cells during continuous biological processes and further develop the questions of interest in this thesis.

1.1.4 Cells in Transition

The Waddington epigenetic landscape established the idea of continuous development as well as offering a conceptual framework for mathematical modeling. Flourishing single-cell sequencing technologies have brought the resolution to the next level and enabled closer exploration of cells during fate decision-making processes, such as differentiation,

reprogramming, and disease. Due to the asynchronous nature of captured cells, the data contains cells in different states, including differentiated fates, primed progenitor or stem states, and transitions (Brackston et al., 2018; MacLean et al., 2018; Moris et al., 2016). Known fates and states have been defined based on diverse phenotypes, such as marker expression, morphology, and protein level, while transition states take a broader definition as any potential intermediates between different defined states (MacLean et al., 2018). For instance, using hematopoiesis hierarchical lineage as a model, these transition cells could be bipotent progenitors (e.g., granulocyte-monocyte progenitors), states from monocytes to macrophages, or cross-lineage stages between monocytes and neutrophils (Olsson et al., 2016). This broad concept paints transition states as a continuous blueprint with less stability but more fluidity, posing new challenges on inference and modeling (S. Jin et al., 2018).

One approach of addressing such states in continuous processes from single-cell datasets is based on pseudotemporal ordering algorithms, as previously discussed in section 1.1.1. These algorithms align the asynchronous cells based on gradual transcriptional changes (Kester & van Oudenaarden, 2018). Identification of terminal identities on the trajectories connects the cells in between terminals as transition cells. The other category of approaches is established based on the concept of the Waddington hierarchical structures (S. Jin et al., 2018). For instance, entropy is adopted to measure stemness and potency of the current cell state. StemID is a method developed to identify stem cell state based on an entropy metric. RaceID2 is first applied to annotate cell types on the single-cell clusters. Lineage tree is then inferred with prior knowledge regarding topology of the system of interest, where cells are further positioned on the links between pairs of clusters. Cell states with high connectivity and transcriptome entropy are predicted to have a stem cell identity. Whereas low entropy states mark the putative

differentiated states of the system (Grün et al., 2016). Nevertheless, clustering-based methods tend to separate cells into distinct clusters, overlooking potentially rare transiting cell populations (P. Zhou et al., 2021).

To directly dissect transition dynamics, other approaches, incorporating interdisciplinary models from physics, have been proposed to model transition probabilities and infer lineage trees. ScEpath is an approached based on an energy landscape of single-cell processes. Instead of an emphasis on variations between different genes, this approach computes the energy estimation based on the connectivity and interdependence among genes within the gene regulatory network. Principle component analysis and clustering are further performed on the estimated energy to place the cells on a landscape. Transition probabilities for each state are computed based on the energy within each state and the distances in the reduced dimensional space between pairs of clusters. With the inferred probabilities, lineages are constructed following a probability-directed graph, further enabling downstream analysis, such as pseudotime calculation and identification of driver genes in cell fate-decision making (MacLean et al., 2018).

Recently, considering the dynamics within the transition cells, MuTrans took on the idea dynamic modeling to characterize transition cells. Taking the metaphor of Waddington landscape on development, MuTrans considers that, in the hilly landscape, valleys are differentiated fates while peaks are bipotent or multi-potent states. Cells at the peak could have bi-stability or multi-stability, yet those at the valleys have most stability. Any position in between the peak and valley is considered as transitions. To identify a path of movement, MuTrans implements random walks in three scales: 1) to compare cell to cell, 2) to compare cluster to cluster, and 3) to compare cell to cluster. Iterative modeling in the three scales leads to ultimate identification of a potential path of the cell from one fate to the other. Application of MuTrans to dynamic systems, such as iPSC

differentiation, has revealed key cell fate-decision dynamics with their key related driver genes (P. Zhou et al., 2021).

At the same time as development of inference of transitional cell states, another important question posed is "what are the roles of these intermediate cell states?" Five putative roles of these cell states have been proposed. First, these cell states serve as a controlled switch in a bidirectional relationship between two fates. For instance, in the condition of epithelial-to-mesenchymal transition (EMT), the intermediate state is a key transition to push the cells from E to M (MacLean et al., 2018). Second, cells at this cell state harbors multiple characteristics from distinct cell types, namely a hybrid phenotype. This hybrid phenomena have been previously identified in different systems (Farrell et al., 2018; Hong et al., 2012, 2015; Olsson et al., 2016). Hematopoiesis is one of the examples where bistable cells are found to express both monocytic and neutrophilic lineage-specific markers (Olsson et al., 2016). The third role proposed is to maintain the balance of cell populations. Responding environmental cues or changes, cells at intermediate states can fluctuate to attenuate the fluctuations within the different cell populations, such as E and M cells during EMT. The next function in consideration is that these transition states maintain a higher potential to readily expand to their differentiated cell types. An example demonstrating this is EMT in cancer, where the identified bipotent intermediates are considered to have higher stemness and ability to generate both E and M cells. Last proposed function is to serve as a checkpoint during the continual processes, such as checkpoint before full differentiation to terminal fates (MacLean et al., 2018; Sha et al., 2019). These proposed functions highlight the potential pivotal roles of transition states and postulate interesting hypothesis in modulating the dynamics in continuous biological systems.

Investigation in transition cells during differentiation and disease progression have unraveled genes that guide the decision-making process. Similarly, we would like to consider these cell states during transcription factor (TF) -mediated reprogramming. As the current reprogramming protocols are inefficient with low fidelity, we can ask key questions from the scope of transition cells. What are the key transition states between the starting population and target cell types? How can we introduce additional modulation to the core TFs to push the cells toward the proper transitions toward the target?

1.2 Key Questions in the Field and Goals

This dissertation will focus on the following challenges in the field: 1) classification in a continuous biological process, 2) identification of intermediate cells during programming and reprogramming, and 3) systematic characterization of cellular heterogeneity. With these challenges in mind, I formulate two specific project goals:

1. I will develop computational tools to better understand heterogeneity in complex, continuous biological systems.
2. I will apply our in-house technologies and tools in another reprogramming system, direct reprogramming of cardiomyocytes, to study the heterogeneity, lineage, and underline mechanism.

To approach these goals, statistical methods, technologies, and biological systems will be utilized. Major technologies, methods, and systems will include single-cell RNA-sequencing of various systems as discussed in Subsection 1.1.1, classification algorithms as reviewed in Subsection 1.1.2, and direct cardiac reprogramming system as discussed in Subsection 1.1.3.

Chapter 2: Dissecting Heterogeneity via Continuous Measurement of Cell Identity

Adapted from:

Capybara: A computation tool to measure cell identity and fate transitions

Kong W., Fu Y.C., Holloway E.M., Garipler G., Yang X., Mazzoni E.O., Morris S.A. (2022). Capybara: A computational tool to measure cell identity and fate transitions. *Cell Stem Cell*. (Accepted)

Single-cell analysis of clonal dynamics in direct lineage reprogramming: a combinatorial indexing method for lineage tracing

Biddy B.A., Kong W., Kamimoto K., Guo C., Waye S.E., Sun T., Morris S.A. (2018). Single-cell analysis of clonal dynamics in direct lineage reprogramming: a combinatorial indexing method for lineage tracing. *Nature*.

Single-Cell Analysis Reveals Regional Reprogramming during Adaptation to Massive small Bowel Resection in Mice

Seiler K.M.**, Waye S.E.**, Kong W., Kamimoto K., Bajinting A., Goo W.H., Onufer E.J., Courtney C., Guo J., Warner B.W.*, Morris S.A.*, (2019, **co-first, *co-corresponding) Single-Cell Analysis Reveals Regional Reprogramming during Adaptation to Massive small Bowel Resection in Mice. *Cellular and Molecular Gastroenterology and Hepatology.*

Single-Cell Analysis of Neonatal HSC Ontogeny Reveals Gradual and Uncoordinated Transcriptional Reprogramming that Begins before Birth

Li Y.**, Kong W.**, Yang W., Patel R.M., Casey E.B., Okeyo-Owuor T., White J.M., Porter S.N., Morris S.A.*, Magee J.A.*, (2020, **co-first, *co-corresponding) Single-cell Analysis of Neonatal HSC Ontogeny Reveals Gradual and Uncoordinated Transcriptional Reprogramming that Begins before Birth. *Cell Stem Cell.*

Deconstruction Stepwise Fate Conversion of Human Fibroblasts to Neurons by MicroRNAs

Cates K.*, McCoy M.J.*, Kwon J.*, Liu Y.*, Abernathy D.G., Zhang B. Liu S., Gontarz P., Kim W.K., Chen S., Kong W., Ho J.N., Burbach K.F., Gabel H.W., Morris S.A., Yoo A.S., (2020, *co-first) Deconstruction Stepwise Fate Conversion of Human Fibroblasts to Neurons by MicroRNAs.

Cell Stem Cell.

2.1 Abstract

Transitions in cell identity are fundamental to development, reprogramming, and disease. Single-cell technologies empower the dissection of tissue composition on a cell-by-cell basis in complex biological systems. Yet highly sparse single-cell RNA-seq data poses challenges for cell-type annotation algorithms based on bulk RNA-seq data. Furthermore, current methods mostly require prior biological knowledge regarding gene signatures and classify cells in a discrete, categorical manner. Here, we present a bioinformatic tool, Capybara, to measure cell identity on a continuum, at a single-cell resolution. This approach enables the mapping of gradual changes in cell identity throughout development and reprogramming. Furthermore, Capybara not only supports the classification of discrete cell entities but also identifies cells in transition, a 'hybrid' cell state, representing critical connections during cell identity acquisition. We benchmark the performance of Capybara against other existing classifiers and demonstrate its efficacy to accurately annotate cells. Further, we validate Capybara-predicted cell-fate transitions using "ground-truth" lineage information within a well-characterized differentiation hierarchy, hematopoiesis. Our application of Capybara to *in vitro* programming strategies reveals previously uncharacterized patterning deficiencies, instructing a new approach to alleviate the lack of appropriate patterning in motor neuron programming. Further, application to the direct reprogramming of fibroblast to induced endoderm progenitors identifies a putative *in vivo* correlate for this engineered cell type that has, to date, remained poorly defined. These findings prioritize interventions to increase the efficiency and fidelity of these cell engineering strategies, showcasing the utility of Capybara to dissect cell identity and fate transitions in development, reprogramming, and disease.

2.2 Introduction

Cells are the fundamental building blocks of complex biological systems, with each defined cell type governing specific functions and mechanisms (Altschuler & Wu, 2010; Paszek et al., 2010). Uncovering cell identities is essential to establish a comprehensive atlas, standardize cell biology, and allow future advancements in medicine and biology. As genes encode distinct transcripts in disparate cell types, whole-cell transcriptome profiles serve as a characterization metric for cell identities (de Wit, 2017). While population RNA-sequencing was applied to classify a bulk group of cells into categories, concerns, such as loss of cell-to-cell variation and lack of resolution, have emerged with growing demands of decoding cellular heterogeneity at a higher resolution (Alquicira-Hernandez et al., 2019; Elowitz et al., 2002). High-throughput single-cell RNA-sequencing (scRNA-seq) technologies have revolutionized the cellular resolution of sequencing to a new standard (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017). This rapidly evolving technique opened broad opportunities to portray cell dynamics in complex processes and decode cellular composition on a cell-by-cell basis. Nevertheless, scRNA-seq is susceptible to noise and dropout, especially for low-abundance transcripts, such as cell-type-specific and regulatory genes, resulting in highly sparse data. This poses challenges for cell-type classification algorithms previously developed for bulk transcriptome profiles. In recently established atlases (X. Han et al., 2018; Schaum et al., 2018), cell identities are manually annotated via unsupervised clustering and known cell-type-specific marker analysis within each cluster. Manual annotation of scRNA-seq data is time-consuming and labor-intensive as the cell number, especially if prior knowledge of cell type-specific markers is limited.

In addition, compared to the homeostatic systems characterized in these atlases, accurate classification in dynamic systems represents additional challenges as cell states are in a continuum. For instance, previous approaches built based on bulk transcriptome revealed incomplete conversion of cell identity during reprogramming, where the resulting cell population retains the features of the starting cell type, demonstrating continual transition during the process (Cahan et al., 2014; Morris et al., 2014).

To tackle these challenges, various computational and analytical algorithms have emerged to support automated annotation of cell identity from single-cell datasets (Abdelaal et al., 2019). For example, *Garnett* utilizes both scRNA-seq and scATAC-seq data to classify single-cell data in three steps: input of gene markers, identification of representative cells, and classification of other cells (Pliner et al., 2019). In contrast, *ScPred* builds a prediction model based on a training scRNA-seq dataset and estimates the probability of each cell belonging to a cell type, based on which a binary classification is provided (Alquicira-Hernandez et al., 2019). However, as *Garnett* is presented as a supervised method, it is limited by the availability of prior biological knowledge. In addition, discrete and categorical identification of cell type is beneficial in describing cells with deterministic or defined fates. Yet, it can be challenged upon evaluation of cells in continuous biological processes, such as early development and reprogramming, where quantitative and continuous measures of cell identities would be valuable (Tan & Cahan, 2019). *SingleCellNet (SCN)* is an approach that quantitatively assesses identity via comparison to reference single-cell datasets using random forests and top-pair transformation. In this approach, gene selection was first performed to select genes that are preferentially and specifically expressed in each cell type in the training set. Using the selected genes, the training set is binarized to train the random forest classifier, which is further used for unknown single-cell cell-

type classification (Tan & Cahan, 2019). While a thorough selection of gene sets that distinguish cell types, it could be challenging to identify a training set that properly contains all cell types to map the testing single-cell data without prior information. Moreover, as these approaches emphasize discrete and categorical cell types, leading to a potential omission of cells at transition or with mixed identities (Tan & Cahan, 2019).

While cataloging discrete cell types is beneficial in describing cells with deterministic or defined fates, it is limited for the evaluation of cells in continuous biological processes, such as early development and reprogramming. Here, we present Capybara, an unsupervised method to quantitatively assess cell identity as a continuous property against publicly available single-cell datasets. Considering cell identities as snapshots of the continuum of biological processes, we compute quantitative scores to measure the likelihood of each cell belonging to a defined cell class via quadratic programming, a method previously used to evaluate cell identity in reprogramming processes (Biddy et al., 2018; Cates et al., 2021; Treutlein et al., 2016). Based on these continuous identities, we further establish a comprehensive classifier that determines the discrete cell class of each cell. Unlike existing methods, Capybara uses continuous identity scores to allow multiple identities to be assigned to an individual cell, enabling identification of hybrid cell types. Building on this unique feature, we develop a 'transition metric' to quantify cell fate transition dynamics.

The efficacy and robustness of Capybara are evaluated against a range of existing cell type classifiers, demonstrating its accuracy to annotate discrete cell identities (Abdelaal et al., 2019). We validate hybrid cells via experimental lineage tracing data of hematopoiesis, in addition to RNA FISH and immunostaining of one of the hybrids during cardiac reprogramming. We also showcase distinct applications in differentiation, programming, and reprogramming

processes to diagnose and instruct shortcomings. In direct programming of spinal motor neurons from embryonic stem cells (ESCs) and reprogramming of cardiomyocytes, Capybara reveals off-target cell types arising from potential deficiency in patterning. Addition of signaling factors enhances the generation of motor neurons by over four-fold. Finally, beyond relatively well-characterized cell types, analysis of direct reprogramming from fibroblast to induced endoderm progenitors (iEPs) identifies a potential *in vivo* correlate to this relatively mysterious reprogrammed cell type. We further validate this cell type experimentally. Take it together, we highlight the utility of Capybara to interrogate cell fate transitions in dynamic biological systems, emphasizing strategies to enhance efficiency and fidelity of cell fate engineering. Capybara code and documentation are available via https://github.com/morris-lab/Capybara.

2.3 Results

2.3.1 Application of quadratic programming

Quadratic Programming (QP) is a method that has been used to classify single cells into cell types during various processes, such as reprogramming. It models the transcriptome of a single cell as a linear combination of bulk gene expression signatures of possible cell types. Under this model, each single cell receives a set of continuous scores for defined cell classes, enabling quantitative measurements of cell identity (**Methods**) (Treutlein et al., 2016).

Previously, we have applied this approach to chart cell identity changes over the time course of reprogramming from mouse embryonic fibroblasts (MEFs) to induced endoderm progenitors (iEPs). This reprogramming protocol was first reported to generate induced hepatocytes, where the cells were evaluated transcriptionally using microarray and functionally with engraftment in damaged liver (Sekiya & Suzuki, 2011). Leveraging the microarray that

measures the start and terminal cells, we performed a time course experiment to dissect the lineage and gradual shift of cell identity during this process. With identified differentially

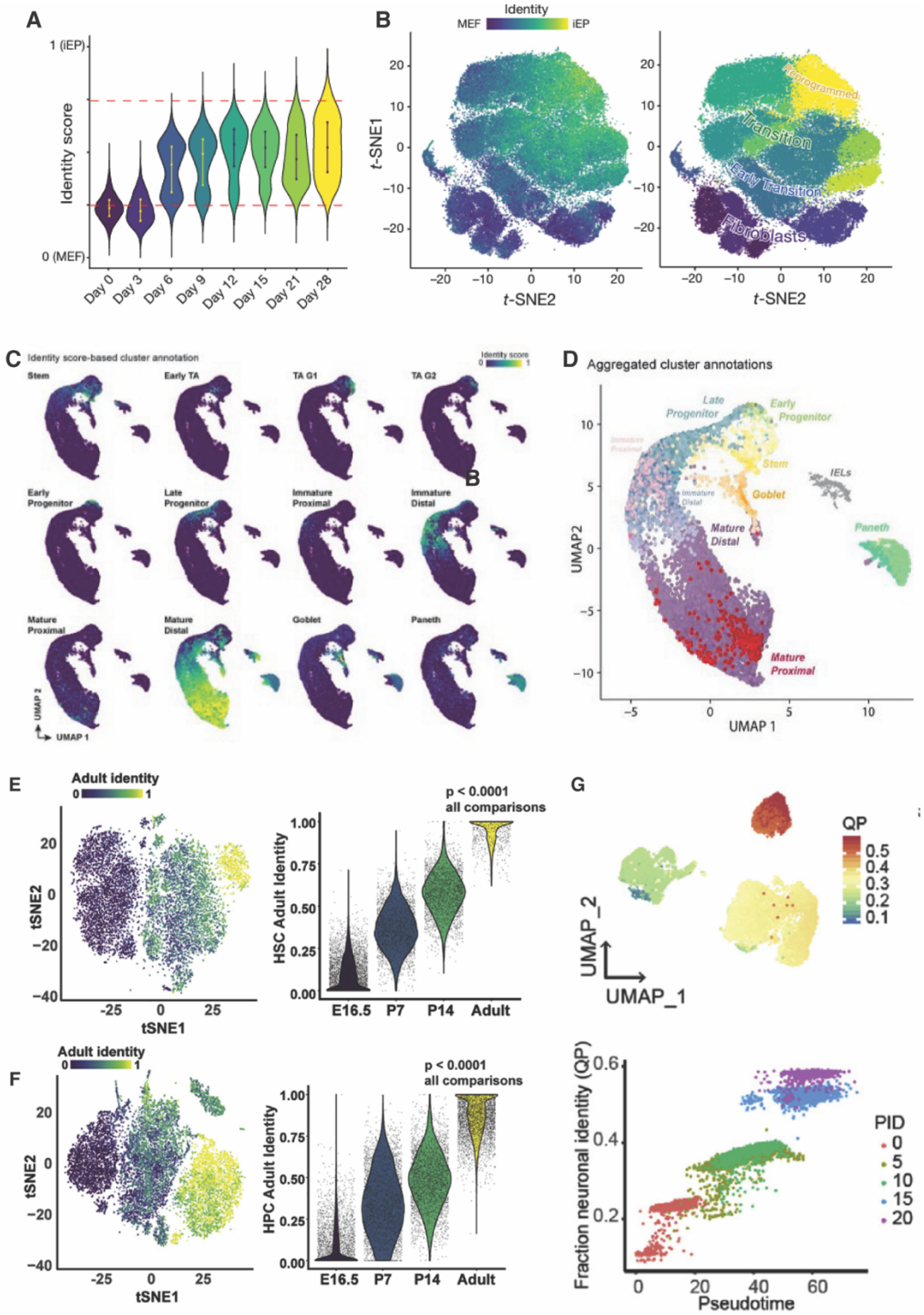

Figure Error! No text of specified style in document..**1**:

expressed genes, QP assigns an identity score between 0 and 1 to each cell. Comparing this quantitative measure of different time points side-by side, we break the data into fibroblast, early transition, transition, and reprogrammed populations (**Figure 2.1 A, B**; (Biddy et al., 2018)**)**.

Although this approach with bulk data grants us a general idea of the cell identities in a single-cell dataset, its resolution is limited by the gene signature selection and low cellular resolution of the bulk dataset with masked sub-populations. Therefore, we have adopted this approach to incorporate single-cell RNA-seq datasets with the whole gene set as a reference to measure cell identity dynamics in the epithelium of small intestine upon small bowel resection (SBR). The small intestine is a powerhouse, where different types of nutrients are absorbed from different regions of the bowel. Upon diseases or injury, such as necrotizing enterocolitis or gastroschisis, a large portion of the bowel is removed with the remaining bowel anastomosed together. This procedure is referred as small bowel resection (SBR). The significant loss of the bowel leads to major loss of absorptive epithelium, resulting in decreased nutrient absorption, potentially resulting in short bowel syndrome (SBS). One potential venue for treatment of SBS resides on the adaptation of the remaining intestine regions to absorb a broader range of nutrients, compensating for the loss of the bowel. To unravel the mechanisms of this adaptation, single-cell RNA sequencing is adopted to understand the epithelium changes post SBR surgery in mice. Using QP with a single-cell dataset surveying the epithelium of small intestine, we categorized the individual cells to detailed intestinal epithelial populations, permitting a higher resolution in cell-type classification (**Figure 2.1 C, D**; (Seiler et al., 2019)).

Further, we applied this approach to investigate the developmental transitions and reprogramming protocols. First, we evaluated fetal-to-adult transitions of hematopoietic stem cells (HSCs) and lineage-committed hematopoietic progenitor cells (HPCs). HSCs and HPCs

harbor different phenotypes, such as reduced self-renewal ability and different lineage bias, between fetal and adult stage. Yet, the mechanistic changes during this process were poorly characterized. Using QP, we revealed that the transitions of HSCs and HPCs from fetal to adult are gradual instead of bimodal as turning on a switch (Y. Li et al., 2020). Compared to this continuous scheme, when applied to neural reprogramming using microRNAs, QP reveals a stepwise cell-fate determination toward neuron identity (**Figure 2.1E-G**; (Cates et al., 2021)).

2.3.2 Capybara overview, benchmarking, and validation

In Capybara, we further develop QP to enable systematic reference construction based on bulk and annotated single-cell atlas datasets as well as identity calculation using reconstructed references. We enable quantitative evaluation of cell identities based on single-cell datasets, benefiting cell state evaluation in continuous biological processes. This further supports an unsupervised classification algorithm in four steps, as follows (**Figure 2.2; Methods**).

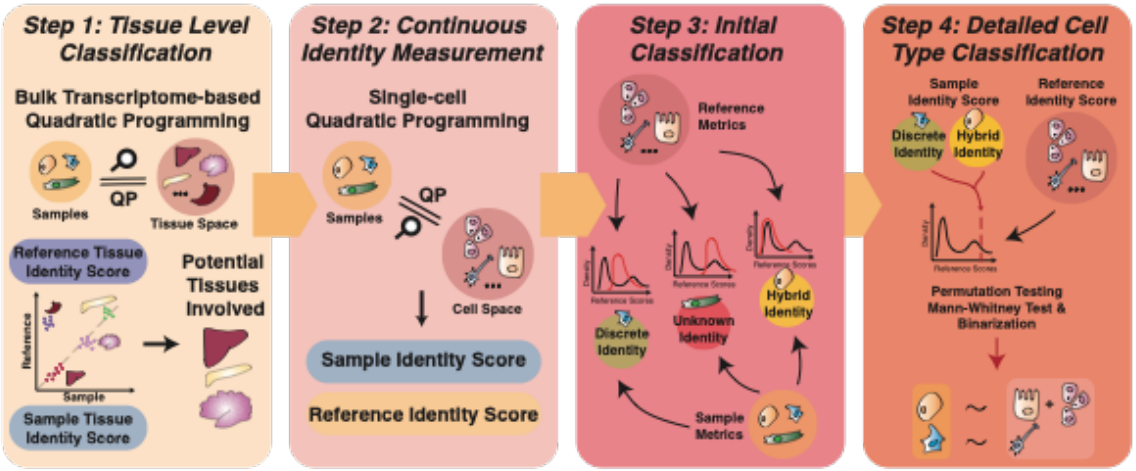

Figure Error! No text of specified style in document.**.2**: Overview of Capybara Workflow.

First, tissue-level classification is performed prior to assessing cell identity at single-cell resolution. Identification of relevant tissues is performed based on bulk transcriptomes from ARCHS4, an exhaustive resource platform comprising the majority of published RNA-seq and ChIP-seq datasets (Lachmann et al., 2018). This step limits the reference cell types to be included in downstream analysis, reducing excessive noise and dependencies caused by correlation across tissues. Then, using the tissue of choice, a reference at higher resolution is constructed from subsetting a larger atlas, such as the Mouse Cell Atlas (MCA; (X. Han et al.,

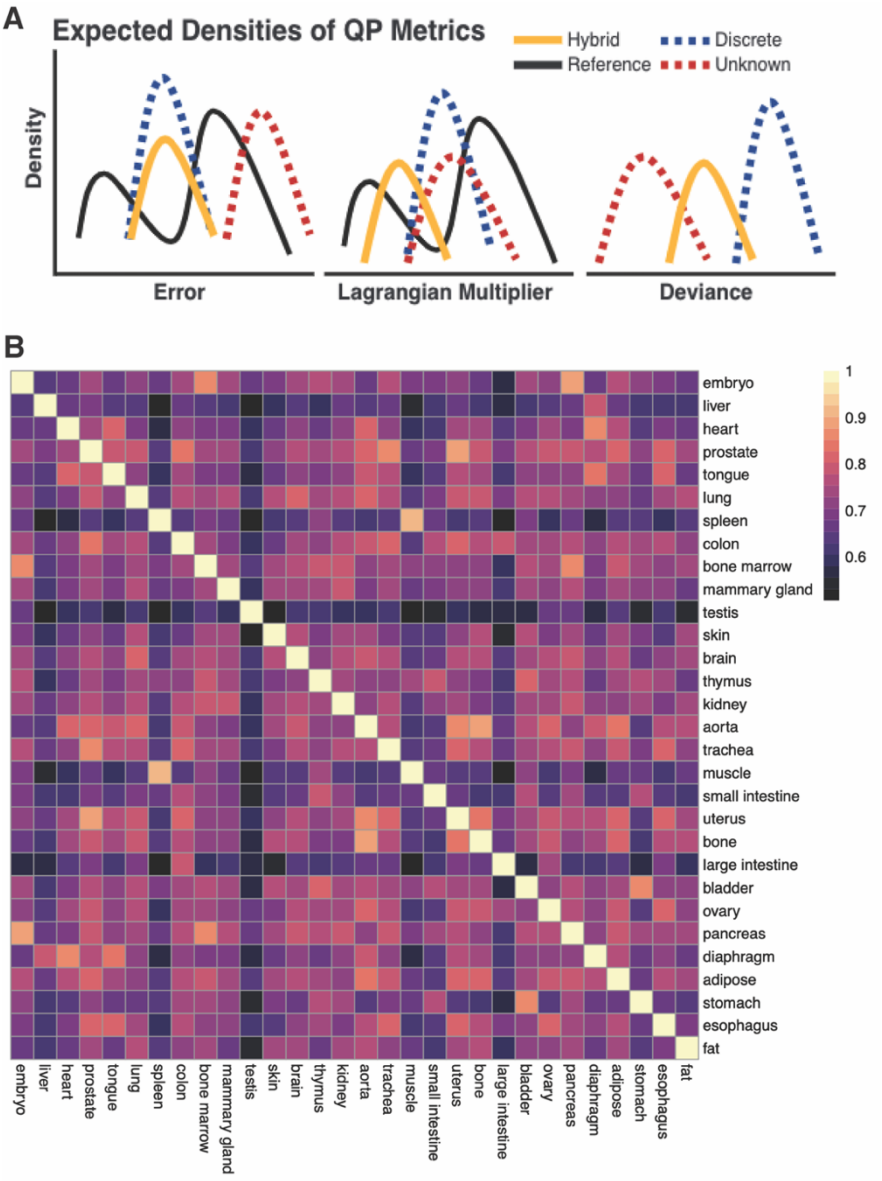

Figure Error! No text of specified style in document.**.3**: QP Metric Demonstration

2018)). To alleviate the effect of gene expression dropout, we sample cells from each defined cell type to create a 'pseudo-bulk' reference. Application of QP using this constructed reference provides continuous measurements of cell identity as a linear combination of all cell types within the reference. In the third step, we categorize the cells into three broad buckets: 1) discrete, 2) hybrid, and 3) unknown, leveraging QP quality metrics (**Figure 2.3A; Methods**). Lastly, a statistical framework is applied to assign discrete identities to cells. In addition, during this step, we characterize cells that resemble multiple cell types, representing potential hybrid cells. Annotations of hybrid cells distinguishes Capybara from other classifiers and allows us to explore the dynamics as well as maintenance of cell identity in complex systems.

To assess the efficacy and robustness of this approach, we first evaluate the sufficiency of this established bulk reference. Accuracy of tissue classification is pivotal as it helps tease apart excessive information and noise from other cells. We evaluate the validity of the tissue reference transcriptome based on the identity scores of annotated single-cell atlases. Though variations are observed within each cohort of cells, we identify a unique pattern of identity scores for each matching tissue, suggesting that this bulk reference is sufficient to imply the correct tissues at relevance (**Figure 2.3B**).

Next, we assess the classification functionality via an automatic evaluation algorithm used in a recent study that benchmarked 20 single-cell classification approaches against various public available datasets (Abdelaal et al., 2019). In brief, 10-fold cross validation is performed using various datasets. We assess predictions from the methods using the multiclass area under the receiver operating characteristic (AUROC), and classification runtime. Based on 5 human pancreatic datasets and the Allen Mouse Brain Atlas, the performance of Capybara suggests a similar and nearly perfect AUROC performance (average = 0.95; rank 5 out of 11) with

reasonable runtime (**Figure 2.4A**). In this automatic benchmarking method, cross-validation provides a relatively large training set compared to the test set. A key feature of Capybara is its flexibility on the size of training set. We identified that a minimum number of 90 cells sampled from each cell type are required to perform classification. Using the minimum number of cells, we evaluate our performance using a recent mouse cell atlas, *Tabula Muris* (Schaum et al., 2018). Using AUROC scores, we benchmark our method against two top-performing classification approaches, scmap (Kiselev et al., 2018) and SingleCellNet (Tan & Cahan, 2019). Here, 90 cells are sampled for each cell type with the remaining cells assessed quantitatively and classified against pseudo-bulk generated. In this manner, we demonstrate the comparable performance of Capybara with almost perfect AUROC scores (**Figure 2.4B**).

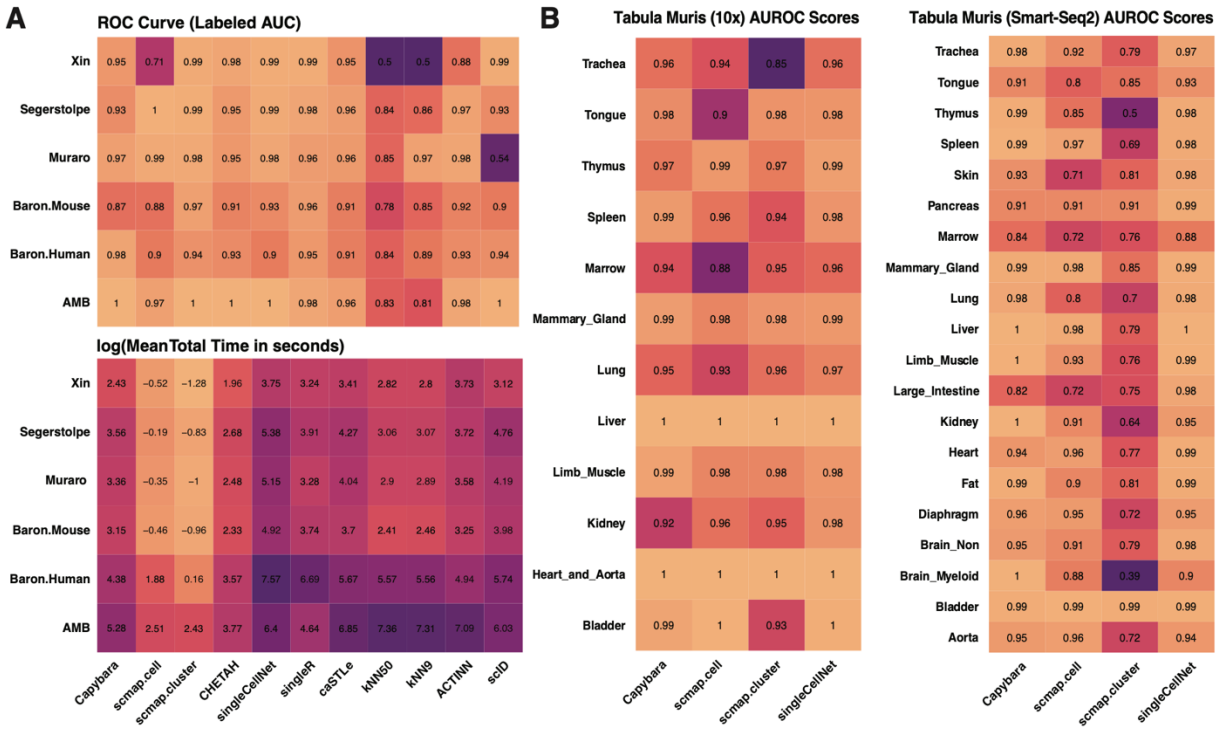

Figure Error! No text of specified style in document.**.4**: Benchmarking of Capybara using Established Pipeline and in-house Cross-Validation.

Moreover, we evaluate Capybara's ability to classify between datasets from different sequencing platforms. Here, we showcase using two of the human pancreatic islet datasets,

Baron *et al* (inDrop, (Baron et al., 2016)) and Muraro *et al* (CEL-seq2, test, (Muraro et al., 2016)). Cross-platform validation presents a diagonal pattern across matching cell types in the two datasets, suggesting agreement between annotation and classification. Nevertheless, four major off-target cell-type classifications are observed. Investigating these cells, we separate them into two categories. One category includes cell types that are annotated in the test set but not in reference. The other includes cell types that are misclassified. In the former, we explore the markers used in the reference for labeling to check if the cells annotated differently in the test set can map the other type in the reference. From Capybara, mesenchymal cells in the test are classified as activated stellate cells. Mesenchymal cells are broadly defined, potentially including stellate cells in the pancreas. Expression of *PDGFRA* shows significant enrichment in the classified activated stellate cells compared to other groups. Similarly, PP cells are determined as gamma cells, whose marker genes, including *PPY, SERTM1* and *CARTPT*, presents upregulation in the classified cells. We further determine cells with unknown cell types to be ductal cell-like. In the latter category, it is interesting to note that Capybara identifies heterogeneity in the annotated acinar cells. Using three acinar cell marker genes, including *REG3A, PRSS1* and *CPA1*, we cannot isolate expression differences between labeled acinar cells and ductal cells.

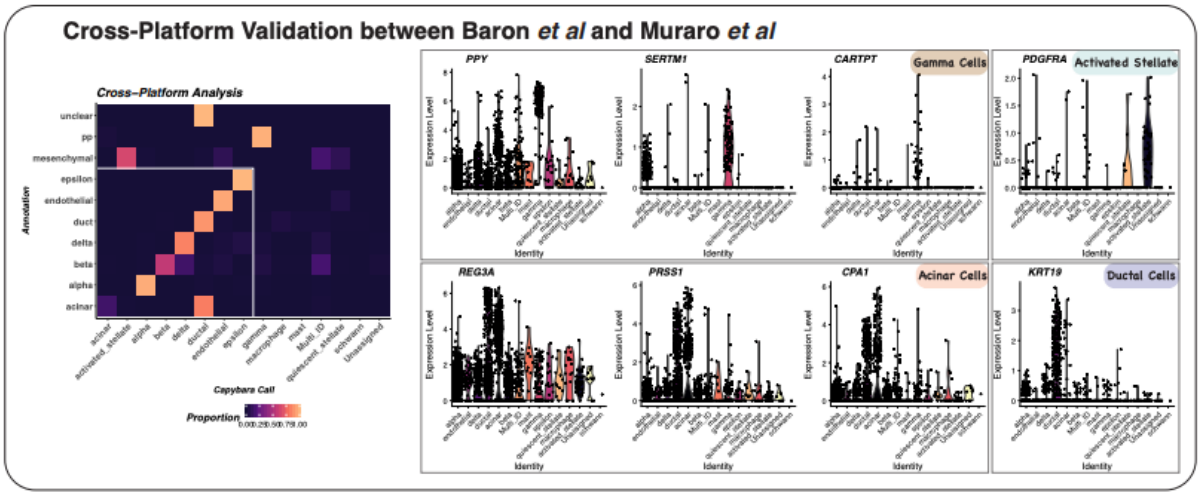

Figure Error! No text of specified style in document.**.5**: Validation of Capybara for Datasets generated

While *KRT19*, a ductal cell marker, reveals a significant upregulation in the labeled ductal cells, indicating that *KRT19* might play a key role in classification of ductal cells (**Figure 2.5**). Additionally, it is worth noting that the original studies use slightly different marker selection for cell type annotation, potentially leading to mismatch between datasets. Overall, we believe this cross-platform analysis demonstrates promising application of Capybara across different sequencing techniques.

As a next step performance validation, we simulate a single-cell dataset comprising different differentiation paths to assess if Capybara can: 1) Capture cells with discrete identities; 2) Identify cells that do not correlate with any cell types in the reference; 3) Characterize hybrid cells that are in transition between discrete identities. We use Splatter, a simulation framework based on gamma-Poisson distribution (**Methods**; (Zappia et al., 2017)), to simulate distinct differentiation paths from a progenitor state (P1), bifurcating toward two discrete states (E1 and P2). P2 progenitor cells bifurcate further toward end states #2 and #3 (E2 and E3, respectively; **Figure 2.6A, B**). We include E1, P2, and E2 in the reference with the remaining cells from, such

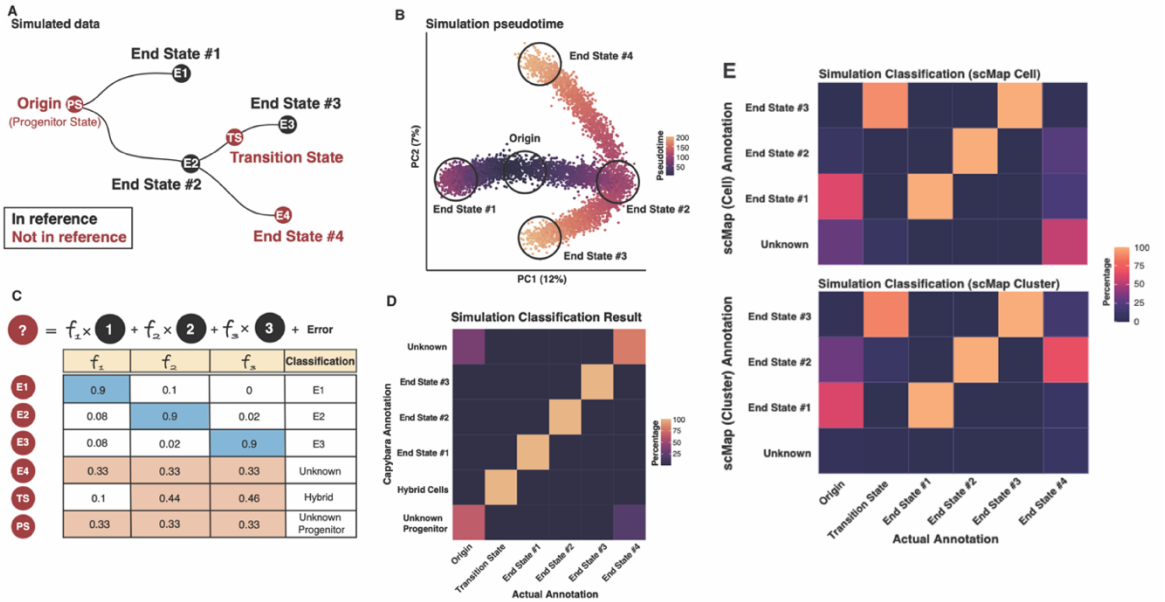

Figure Error! No text of specified style in document.**.6**: Simulation Study for Proof of Concept.

as P1 and E3, to be tested. Capybara accurately classifies cells in the three different identity categories, distinguishing between known discrete identities and cells in transition between them (AUROC = 1; **Figure 2.6A-D**). Further, Capybara can distinguish unknown cell types with 100% accuracy and potentially isolate unknown progenitor states from unknown terminal states, using QP quality metrics with 81% and 65% accuracy, respectively (**Figure 2.6C-D; Methods**). We benchmark our hybrid cell classification against scMap to illustrate how existing cell type classifiers cannot resolve mixed identity cells (**Figure 2.6E**).

Together, our benchmarking and simulation demonstrate the efficacy of our method for cell-type classification of discrete and hybrid cell identities. We next showcase move to evaluate Capybara in a well-characterized continuous differentiation process: hematopoiesis.

2.3.3 Capybara accurately captures cell identity and fate transitions in hematopoiesis

Cells in many biological processes, such as development and reprogramming, reside on a continuous landscape, gradually transitioning from one state to another. Understanding these transition states is essential to uncover the comprehensive overview of dynamic biological systems. Cell fate determination during developmental processes is an important question to be addressed to reveal mechanisms and lineages of development (MacLean et al., 2018). In this process, a single cell differentiates into populations of cells with distinct identities and committed functions. Apart from the cells with dedicated fate, some cells are in progress, where the routes are chosen while the ultimate identities are not attained. These cells, we consider, embrace an intermediate state or in transit, enabling them to be characterized with multiple identities. Variations in the transcriptome profiles of these cells allow us to analyze transitions via single-cell RNA-sequencing data, snapshots of individual cells at different time points of the

continuum. Hematopoiesis provides a well-characterized example of continuous cell state transition and serves as a valuable model to understand cell fate specification (Orkin & Zon, 2008). Even before birth, hematopoiesis has already started transitioning from a fetal clock to an adult clock. As aforementioned, HSCs have distinct self-renewal rate, lineage bias, and transcriptome profiles at fetal stages compared to those in adults. Leveraging single-cell profiling and continuous identity measurements, we identified gradual and synchronous transition from fetal to adult stages. Beyond transcriptome, further analysis in epigenome remodeling uncovers gradual and discordant reprogramming of neonatal landscape in addition to the continuous accessible fetal chromatin regions after adulthood. Yet, it was found that conversion to adult states is initiated before birth and is potentially driven by interferon signaling pathway (Y. Li et al., 2020). In the body of an adult, blood cells are renewed at a rate of seconds through hematopoiesis. In this process, self-renewing hematopoietic stem cells (HSCs) differentiate gradually into their mature cell fate in specific lineages, including erythroid, myeloid and lymphoid lineage. Development of fluorescence-assisted cell sorting (FACS) and identification of key surface markers in distinct lineages enable the establishment of a hierarchical tree of hematopoiesis with a root of HSC growing the leaves of mature blood cells through specific progenitor stages. This lineage paradigm serves as a valuable model in understanding blood formation and, further, the general mechanism of lineage specification in developmental biology (Orkin & Zon, 2008; Paul et al., 2015).

Addition of single-cell RNA-sequencing in this system elevates the cellular resolution, enabling a more comprehensive illustration of the process and demonstrates high heterogeneity in hematopoiesis. Here, we demonstrate an application of Capybara on a massively parallel single-cell RNA-seq (MARS-seq) dataset, containing 2,730-cell snapshot of myeloid progenitor

differentiation (Paul et al., 2015). We first use partition-based graph abstraction (PAGA; (Wolf et al., 2019)) and identify a total of 24 clusters, which we annotate according to Paul *et al*. Initial

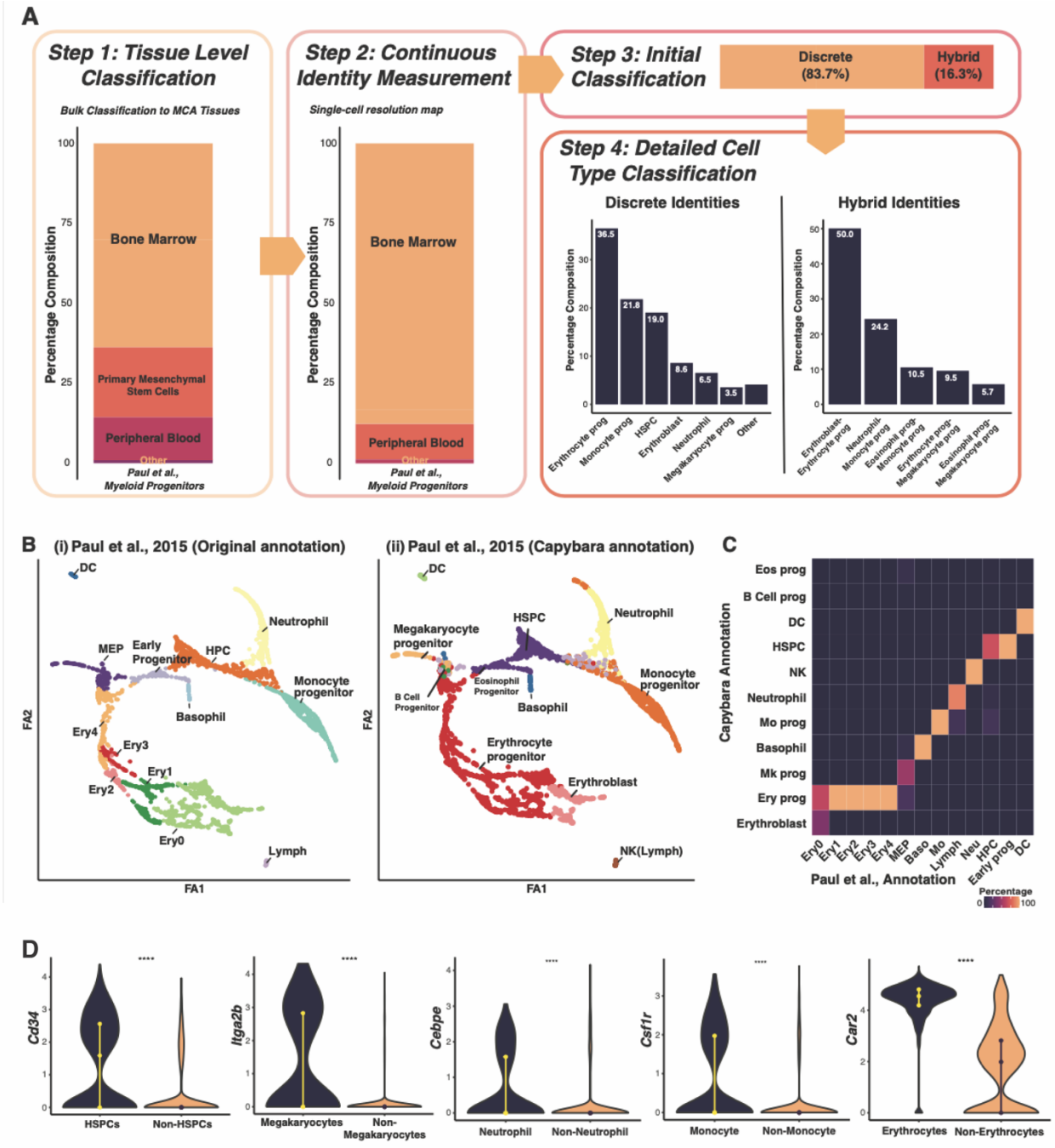

Figure Error! No text of specified style in document.**.7**: Application of Capybara to Classify Hematopoietic Cell Identity.

tissue-level classification using Capybara corresponds mainly to single-cell bone marrow. Using the Mouse Cell Atlas (MCA; (X. Han et al., 2018)), we generate the high-resolution reference, containing two major tissue types: bone marrow, bone marrow (c-Kit), primary mesenchymal stem cells, and peripheral blood (**Figure 2.7A: Step 1**). Continuous measurement of cell identity further returns populations from two major tissues: bone marrow and peripheral blood (**Figure 2.7A: Step 2**). Discrete cell-type classification suggests 0% unknown population and 16.3% (n = 445) received high identities (**Figure 2.7A: Step 3**). Doublet analysis in the single-cell dataset using DoubleFinder (McGinnis et al., 2019) and DoubleDecon (DePasquale et al., 2019) suggests that 7-9% of the hybrid cell population as cell doublets, relative to 4.3-16.9% of the discrete population, ruling out doubles as the source of hybrid signals.

Overall, Capybara classifies the expected myeloid progenitor populations, including erythrocytes, megakaryocytes, hematopoietic stem and progenitor cells (HSPCs), monocytes, and neutrophils (**Figure 2.7A: Step 4**). 13 major populations, identified via PAGA and annotated according to Paul *et al.,* agree with Capybara classification (Weighted Cohen's Kappa = 0.95; **Figure 2.7B-C**). Further, each classified population shows enrichment of established cell-type marker expression (*Cd34, Itga2b, Cebpe, Csf1r,* and *Car2*; P < 2.2E-16, Wilcoxon rank-sum test, **Figure 2.7D**).

Next, we assess the classification using modified diffusion pseudotime estimated by PAGA (Wolf et al., 2019). We first evaluate the identified discrete populations, where Capybara-classified HSPCs coincide with early pseudotime, as expected for this relatively undifferentiated cell population (**Figure 2.8A-B**). Then, we focus on the identified five major hybrid populations (>= 0.5% of the overall population): erythroblast–erythrocyte progenitors, megakaryocyte progenitor–erythrocyte progenitors, monocyte progenitor–neutrophils, megakaryocyte

progenitor–eosinophil progenitors, and monocyte progenitor–eosinophil progenitors (**Figure**

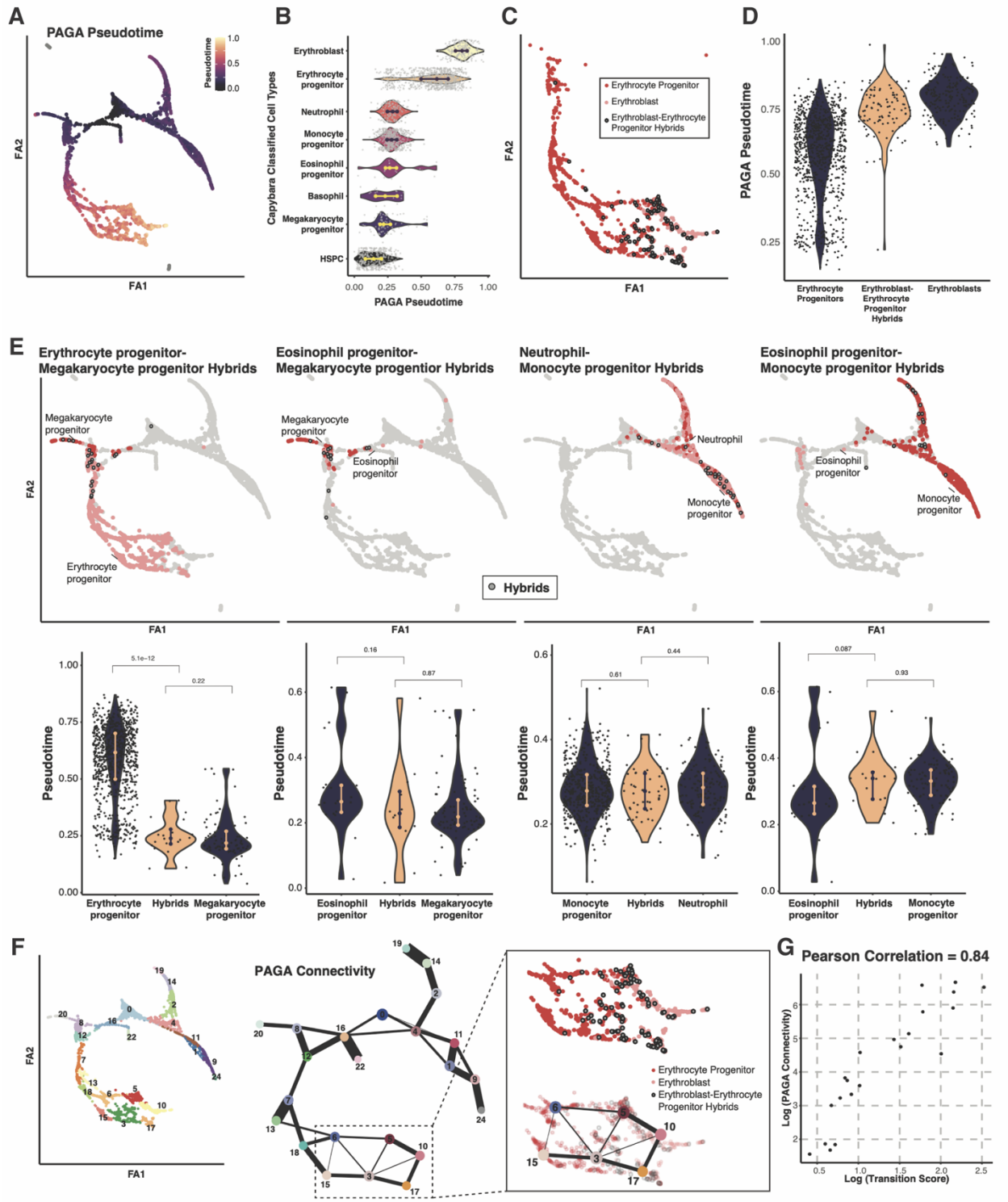

Figure Error! No text of specified style in document.**.8**: Evaluation of Hematopoietic Hybrid Cells against Pseudotime.

2.7A; Step 4). The largest hybrid population contains cells sharing identities between erythrocyte progenitors and more differentiated erythroblast, representing a putative transition state. Leveraging PAGA pseudotime, we evaluate these hybrid cells considering hybrids would likely occupy intermediate pseudotime between defined identities. Comparisons of pseudotime between hybrids and their discrete counterparts demonstrates that the hybrids are located mid-pseudotime. For instance, the hybrids, erythroblast–erythrocyte progenitors, are in between the discrete and erythroblast states (**Figure 2.8C-E**). Further, clusters enriched for hybrid cells are connected based on PAGA connectivity (**Figure 2.8F**). Taken together, this application demonstrates the ability of Capybara to accurately identify major hematopoietic cell populations, in addition to hybrid populations.

2.3.4 Lineage tracing reveals the multi-lineage potential of hybrid-classified cells

To examine hybrid cells, we leverage a single-cell data that simultaneously captures lineage and transcriptome of hematopoiesis (Weinreb et al., 2020). In this previous study, Lin^- $Sca1^+$ Kit^+ (LSK) cells were isolated and labeled with random lentiviral-delivered barcodes. The cells were allowed to differentiate *in vitro* and collected for single-cell RNA sequencing at Days 2, 4, and 6, yielding 72,946 single-cell transcriptomes. The barcoding strategy allows early cell state to be linked to later hematopoietic fate and provides a ground truth dataset to assess the biological relevance of our hybrid cell assignments (**Figure 2.9A**). We first construct a high-resolution reference based on the major Day 6 differentiated myeloid populations, excluding undifferentiated cells due to potential heterogeneity (**Figure 2.10A-B**). Capybara identifies of seven major hybrid cell types, including three largest populations: monocyte-neutrophil, basophil-mast, and basophil-eosinophil hybrids, who contain clones spanning early and late time

points (**Figure 2.9B**). We assessed these hybrid cells via a close look at their cell-type composition of clonal relatives across all time points. This analysis reveals that these clones are significantly enriched for the discrete cell types that constitute each hybrid identity (*: $P < 0.05$; randomized test; **Figure 2.9C**).

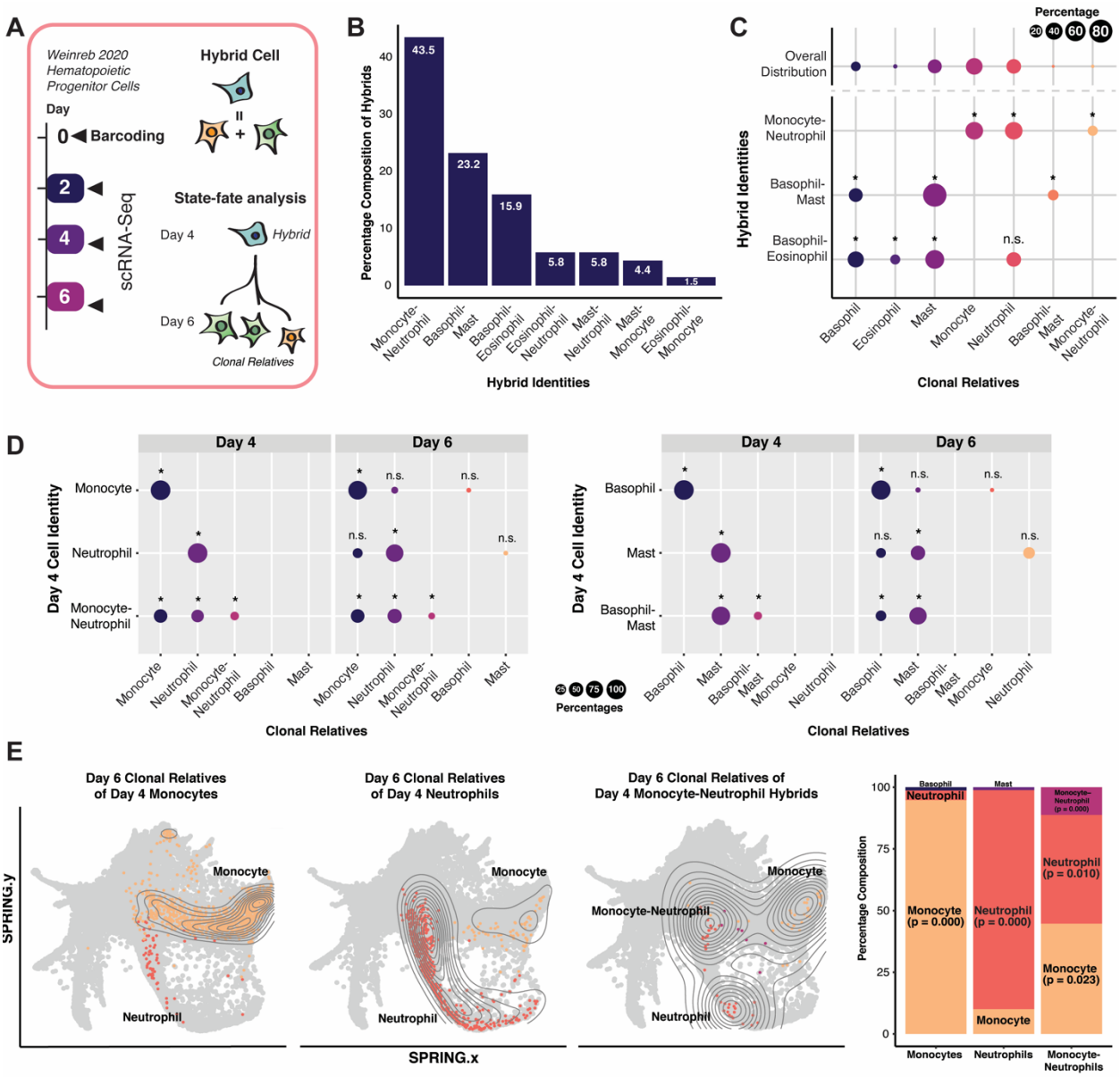

Figure Error! No text of specified style in document.**.9**: Evaluation of Hybrid Cells using Ground-Truth Lineage Tracing.

Particularly, we showcase monocyte-neutrophil and basophil-mast hybrid identities, two major hybrid populations in Day 4 and Day 6. Using Capybara classification, we identified clones on Day 4 that strictly contains discrete cell identities (i.e., monocyte, neutrophil, basophils, or mast cells only) and found that their Day 6 siblings are significantly lineage restricted ($P < 0.05$, randomization test; **Figure 2.9D-E**). Next, we investigate into Day 4 clones containing hybrid cells. To assess the accuracy of hybrid classification, we evaluate if Day 4 clones with hybrid cells would be related to or give rise to significant discrete-identity populations, reflecting the mixed identities. In monocyte-neutrophil containing Day 4 clones, we

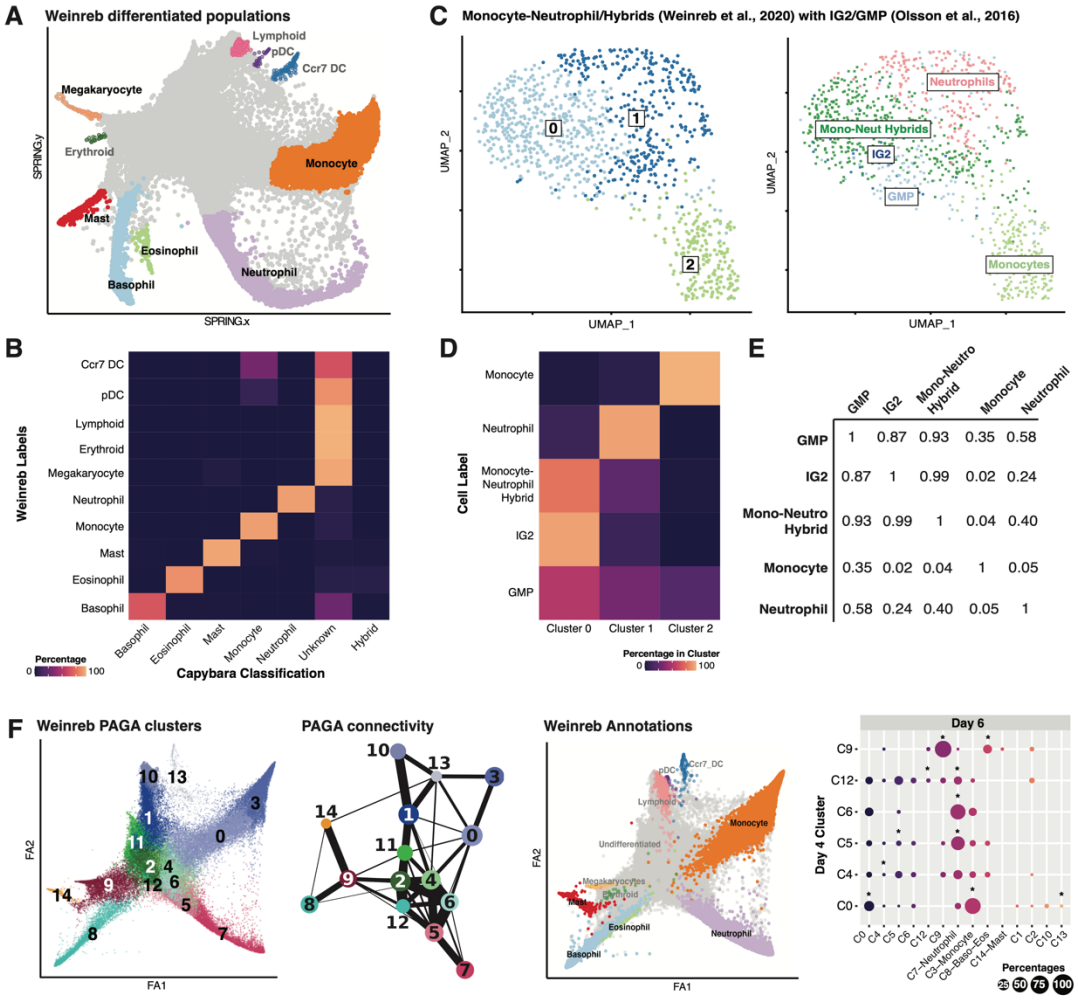

Figure Error! No text of specified style in document.**.10:** Discrete Identities in Ground-Truth Lineage Tracing Dataset and Comparison to Previously Identified bistable states.

identified a significant population of both monocyte and neutrophil on Day 6 **(Figure 2.9D, left; Figure 2.9E)**. Further, via integration of single-cell datasets, we found that these identified monocyte-neutrophil hybrids are similar to previously identified bistable intermediate cell states that can yield both monocytes and neutrophils (**Figure 2.10C-E;** (Olsson et al., 2016)). Similarly in Day 4 clones with basophil-mast hybrid cells, there is a significance mast cell population but not basophil on Day 4, yet both discrete basophil and mast cells are identified as significant populations on Day6 **(Figure 2.9D, right)**. Finally, we compare these results of hybrid cells to trajectory inference approach using PAGA (Wolf et al., 2019). From PAGA, we obtain the connectivity measure between different clusters of cells identified using Louvain algorithm, revealing connected and disconnected regions. Applying similar statistical test as described as above on the Weinreb dataset fails to uncover the hybrid cell states we identified (**Figure 2.10F**). Altogether, using 'ground truth' lineage tracing data, we validate the hybrid cells identified through Capybara to be biologically relevant.

2.3.5 A metric to quantify cell fate transition dynamics

From our application of Capybara to hematopoiesis, the identification of cells harboring hybrid identities represent intermediate transition states (MacLean et al., 2018). Furthermore, the identities and states that these cells are connected to help define the cell transition milestones within a differentiation hierarchy. We leverage this capacity of Capybara to identify hybrid cells to develop a 'transition metric.' Briefly, for each discrete cell identity, we define in a population, we measure the strength and frequency of its connections to hybrid identity states (**Methods; Figure 2.11A**). This generates a metric we define as a 'transition score,' where a high score represents a high information state where identities converge, identifying a putative cell fate transition. This aspect of Capybara is distinct from approaches such as StemID (Grün et al.,

2016) and CytoTrace (Gulati et al., 2020) in that our transition score does not aim to quantify potential. Instead, this score aims to identify those cell fate transitions that are central to a developmental/reprogramming process and is helpful to identify the stages at which a biological system is in flux.

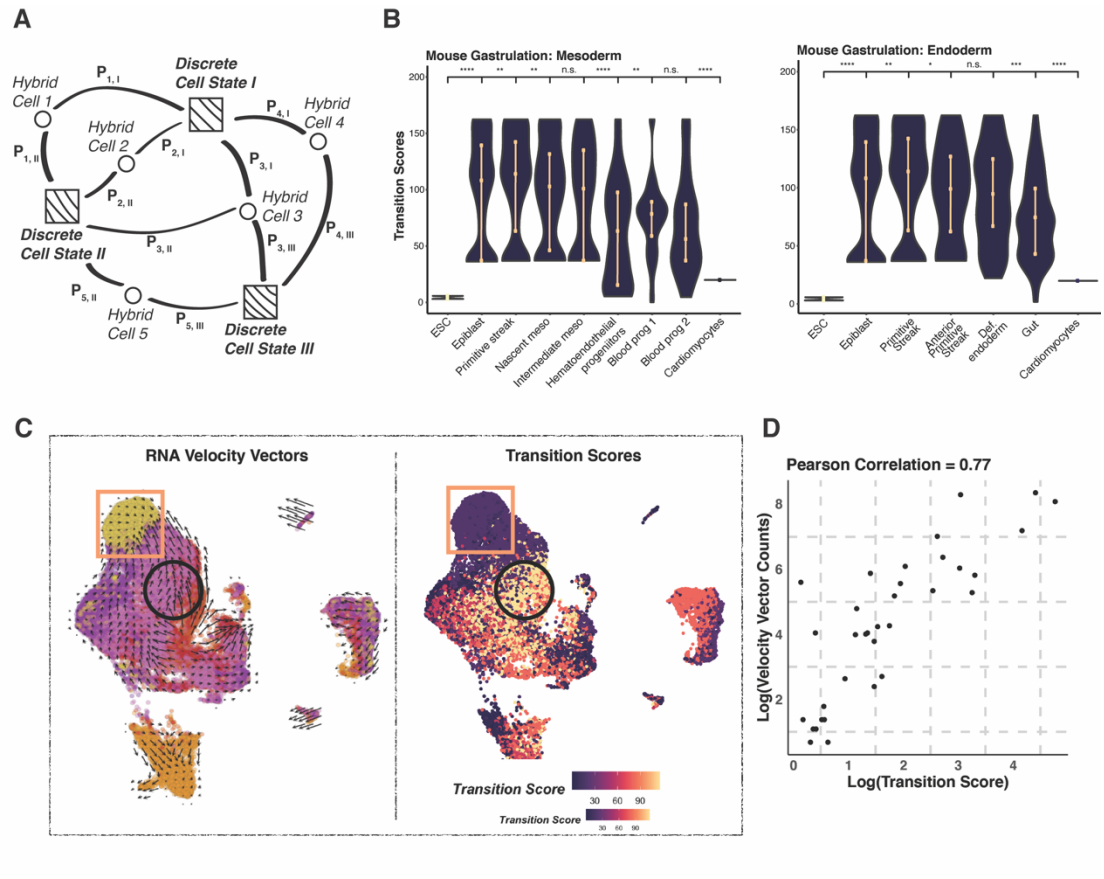

Figure Error! No text of specified style in document.**.11**: Transition Score and Its Validation.

We first validate the concept of transition scores via comparison to cell-to-cell connectivity based on PAGA in hematopoiesis (Paul et al., 2015). From PAGA analysis, we computed the total connections of cells from each cell type to other cell types. Using Pearson's correlation, we observed a strong correlation (r = 0.84) between the total connectivity and transition score (**Figure 2.8G**). Further, we apply RNA Velocity (la Manno et al., 2018) to identify actively transitioning cell states in cardiomyocyte reprogramming (Stone et al., 2019).

Comparing the number of velocity vectors within each defined cell state and transition scores, we observe strong correlation between the two (**Figure 2.11C-D**). We apply the transition score metric to published datasets charting the early development to terminal differentiation (Klein et al., 2015; Pijuan-Sala et al., 2019; Stone et al., 2019). As development progresses and cells specialize, transition scores gradually decrease (**Figure 2.11B**).

Via our validation in hematopoiesis and other systems, we highlight the ability of Capybara to, in an unbiased manner, accurately distinguish discrete cell identity and cell fate transitions. We next apply Capybara to characterize less defined, non-physiological systems, such as reprogramming, to identify key fate specification events and propose potential strategies to improve the protocols.

2.3.6 Characterizing off-target and hybrid cell identity in cardiac lineage reprogramming

The identification of transcription factors to convert somatic cells back to a pluripotent state open new potentials in regenerative medicine to generate therapeutic cells *in vitro* (Takahashi & Yamanaka, 2006). Despite the premise of differentiation from stem cells, limited differentiation to target cell types produces a heterogeneous population, hindering the clinical utility of these cells. Lineage reprogramming with ectopic expression of transcription factors, such as direct cardiac conversion, bypassing progenitor or stem cell states, holds much promise for the generation of clinically valuable cell types from mature somatic cells (H. Wang et al., 2021). A key question in direct reprogramming concerns the overall efficiency, i.e., percentages of successful conversion, and fidelity, i.e., similarity to their *in vivo* counterparts (Guo & Morris, 2017; Morris et al., 2014). Here, we take direct cardiac reprogramming as an example to demonstrate an application of Capybara to provide some insights.

The direct conversion of fibroblast to cardiomyocyte-like cells is introduced via overexpression of three transcription factors: Gata4, Mef2c, and Tbx5 (GMT) (Ieda et al., 2010; Qian et al., 2013; Song et al., 2012). We select a recently published 30,729-cell transcriptome dataset that profiled cardiac reprogramming on day -1, 1, 2, 3, 7, and 14, where transcription factors were added on Day -1 followed by TGFβ inhibitor and Wnt inhibitor treatments on Day 0 and Day 1, respectively (**Figure 2.12A**). On day 14, cells were sorted based on cardiac reporter gene, α-Myosin Heavy Chain (Gulick et al., 1991), before single-cell processing (Stone et al., 2019).

Via Capybara, initial tissue classification, followed by refinement using the MCA, highlights two major tissues, including neonatal skin and neonatal heart (**Figure 2.12B**). Within neonatal skin, we identified two major populations, including macrophages and muscle cells, both of which are mesodermal and resident in the heart (de Soysa et al., 2019). With this reference, 65.1% of cells are assigned with a discrete identity with 19.7% assigned hybrid identities (**Figure 2.12C**). Based on our classification, we identify gradually decreasing trends in non-cardiomyocyte populations across time points with increasing trends in cardiomyocytes population. Cell type composition analysis shows enrichment of atrial and left ventricular cardiomyocytes on Day 7 (39.4%) and Day 14 (83.7%) (**Figure 2.12C**). Interestingly, we find a bias of cardiomyocyte population generated to be more atrial like instead of a balanced distribution between atrial (76%) and ventricular (7.7%) (**Figure 2.12C-D**). We further validate this via examination of expression of some regional specific markers, suggesting a potentially chamber-specific reprogramming (**Figure 2.12G**).

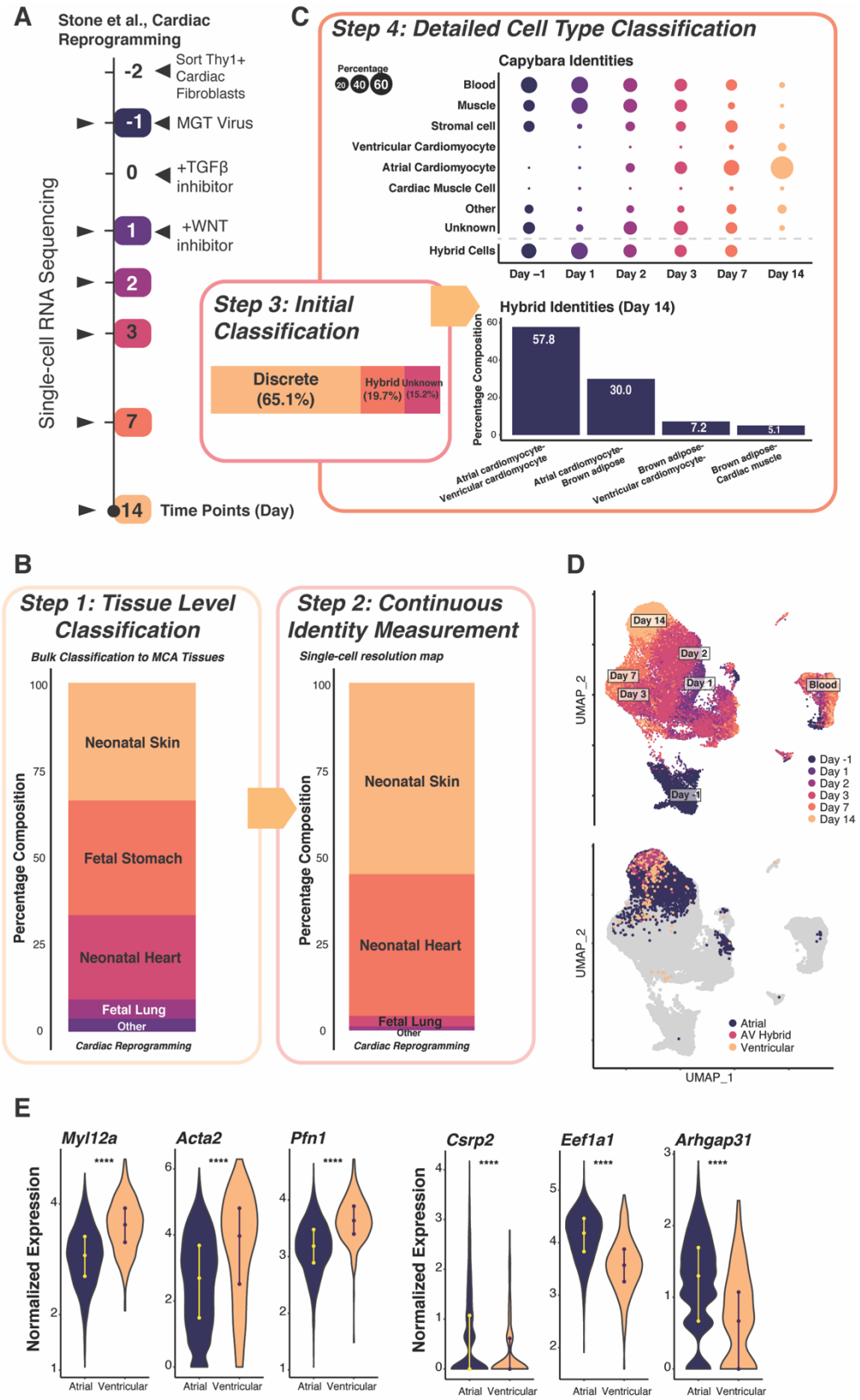

Figure Error! No text of specified style in document.**.12**: Capybara Analysis of Direct

Transition score analysis of this process reveals a significant increase at Day 1 and Day 2 with a progressive decrease on day 3 and further on day 7 to day 14 ($P < 0.0001$, Wilcoxon rank-sum Test; **Figure 2.13A**), implying an active cell fate transition before deterministic commitment. Identification of such a trend reinforces previous findings, where transition cells were found to determine their terminal fate in the first 48hr before they bifurcate into two routes (Stone et al., 2019). Next, we take a close look at hybrid cells in the day 14 sorted population to

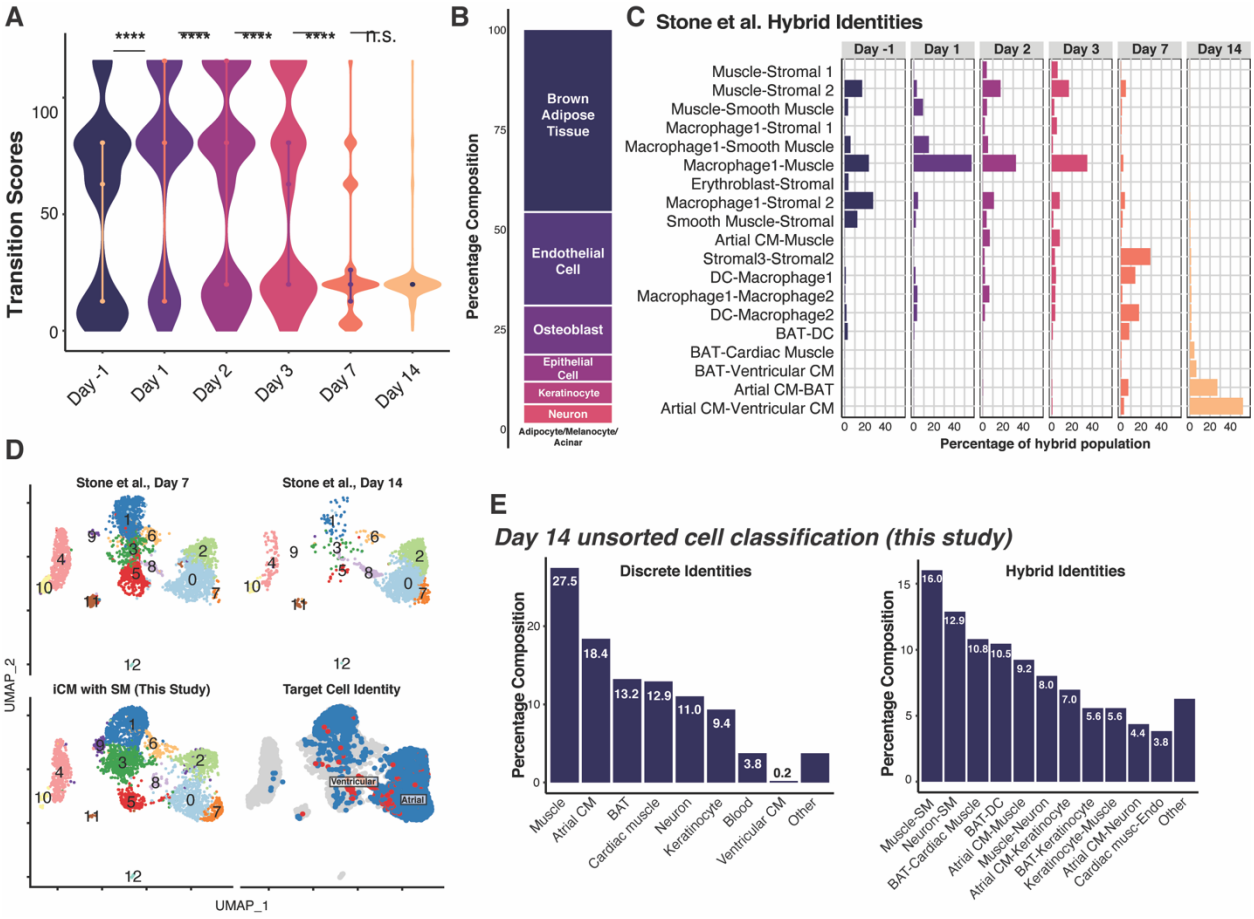

Figure Error! No text of specified style in document.**.13**: Transition Scores and Hybrid Identities of Direct Cardiac Reprogramming.

identify potential intermediate states after commitment to defined fates. Atrial-ventricular (AV) cardiomyocyte occupies a dominant proportion of the hybrid cells (55.9%) with a second major population between brown adipose tissue and cardiac fate (**Figure 2.12C; 2.13B-C**). We

consider the cells with brown adipose tissue fate could be derived from cardiac resident adipogenic progenitor that have been previously reported to repair the heart upon injury (H. J. Chen et al., 2021; K. Wu et al., 2010; Yamada et al., 2006).

To construct a more comprehensive blueprint of the off-target cell types, we performed cardiac reprogramming according to the Stone protocol, without cardiac reporter sorting at day 14. We obtain a total of 5,107 cells across two independent biological replicates. The queality of the data is first assessed via integration with the Stone et al. data, where we observe a cosine similarity of 0.71-0.82, noting the successful recapitulation of the protocol (**Figure 2.13D**). Capybara reveals a similar off-target cell profile to the Stone dataset and an enrichment of atrial cardiomyocytes (**Figure 2.13E**). AV hybrid cells are also present in this dataset, though at a much lower frequency (<1%). We consider that this low occurrence could be a result of not sorting the cells.

To further validate the dominant AV hybrids, we performed RNA fluorescent *in situ* hybridization (FISH) at day 14 of our reprogramming scheme to evaluate canonical markers, *Myh6* (atrial myosin heavy chain) and *Myh7* (ventricular myosin heavy chain). In addition to discrete cells expressing one marker or the other, hybrid cells co-expressing both markers are found (**Figure 2.14B**). Using our own scRNA-seq data, we further select another canonical atrial marker, *Myl4*, along with ventricular enriched expression markers, *Actc1* and *Tnnc1*, to perform RNA FISH, where other AV hybrids are identified (**Figure 2.14A, C**). Interestingly, these hybrid cells are binucleated or have irregular nuclear morphology. Beyond expression, we performed immunostaining (IF) for canonical markers MYL7 (atrial) and MYL2 (ventricular), supporting AV hybrids at the protein level (**Figure 2.14E)**. Moreover, the proportion of hybrid cells identified through IF, scRNA-seq, and RNA FISH are consistent with each other (**Figure 2.14F**).

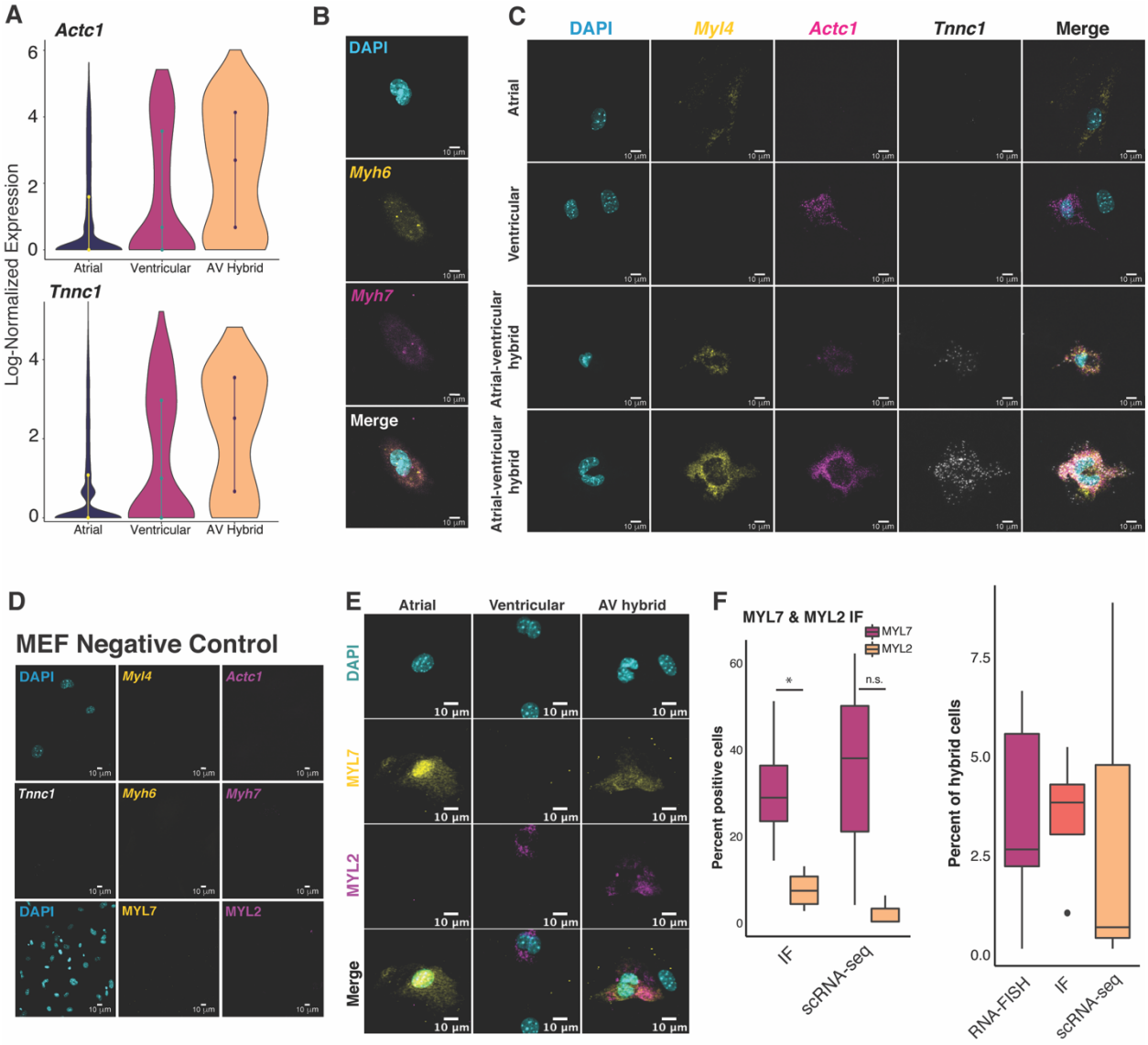

Figure Error! No text of specified style in document.**.14**: Experimental Validation of Hybrid Cells using RNA FISH and immunostaining.

Application of Capybara to direct cardiac reprogramming unravels critical chamber-specification dynamics and off-target cell types, suggesting potential modulations with additional TFs or signaling molecules are necessary to refine reprogramming outcomes.

2.3.7 Capybara reveals a dorsal-ventral patterning deficiency in motor neuron reprogramming

In vitro culture of embryonic stem cells (ESC) and the discovery of induced pluripotent stem cells (iPSC) provide opportunities for disease modeling and hold much promise for generation of clinically valuable cell types to replace damaged tissues (Takahashi & Yamanaka, 2006). To derive specific cell types of interest from stem cells, two major approaches – standard differentiation or direct programming – are employed. Standard differentiation involves sequential induction of developmental signals, mimicking embryonic intermediate states during development, while direct programming introduces exogenous expression of transcription factors following different routes (Briggs et al., 2017). Both protocols have been reported to recapitulate the target cell type efficiently and faithfully. Yet, the relationship between cell types produced *in vitro* procedures and their counterparts *in vivo* development remains in the fog. Recent application of single-cell RNA-sequencing on motor neuron differentiation from mouse ESC (mESC) unveils that both approaches achieve the same cell state resembling MNs but through different paths.

With the experiments performed side-by-side, this dataset provides valuable insights and a venue for direct comparison of the protocols (**Figure 2.15A**; (Briggs et al., 2017)). Specifically in MN generation, direct programming (DP) induces MN fate via overexpression of three transcription factors, Ngn2, Isl1, and Lhx3 (NIL), bypassing developmental progenitor stages (Mazzoni et al., 2013; Velasco et al., 2017). On the other hand, direct differentiation (DD) utilizes sequential induction of Fibroblast Growth Factors (FGFs), Retinoic Acid (RA), and Sonic Hedgehog (SHH) (Wichterle et al., 2002; C. Y. Wu et al., 2012). Here, we primarily focus

on the TF-mediated DP protocol and apply Capybara to compare the cell fate determination to mouse spinal cord development.

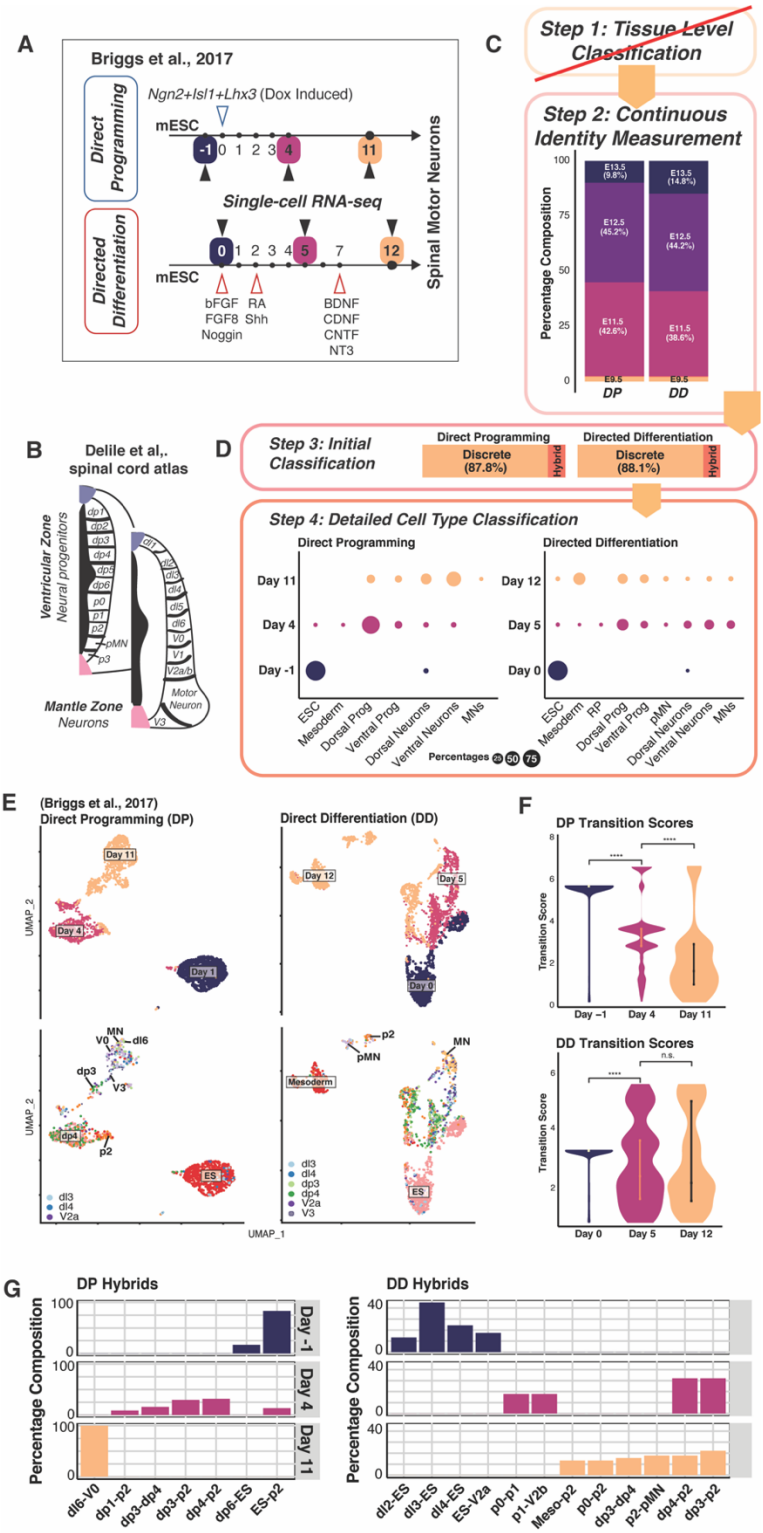

Figure Error! No text of specified style in document.**.15**: Capybara Analysis of Spinal Motor Neuron Differentiation and Programming. (Briggs et al., 2017)

With prior knowledge regarding the system, we start with the construction of high-resolution reference, using a recent single-cell atlas of mouse embryonic spinal cord development (**Figure 2.15B**; (Delile et al., 2019)). This reference contains a total of 118 cell types and states across five different time points from E9.5 to E13.5, including non-neuronal cell types around the spinal cord, such as blood, mesoderm, and neural crest. In combination with the embryonic stem cell profile from the MCA, the aggregated dataset is ideal for our analysis of MN generation, allowing initial tissue selection to be bypassed (**Figure 2.15C**).

With Capybara, 87.8% of cells (n = 4,136/4,704 cells) are classified with discrete identities while 12.2% are classified as hybrids with no cells unclassified (**Figure 2.15D**). We observe a gradual emergence of neuronal identities from the dominant ESC population at day -1, with 63.8% of day 11 cells classified as neurons. Nonetheless, only 3% of this population are classified as motor neurons with majority of them classifying as neurons across the dorsal and ventral regions of the spinal cord (**Figure 2.15D-E**). On the other hand, in direct differentiation, MN generation peaks at 13.4% at the early-stage (day 5), decreasing to 3.4% by the end of the protocol (day 12), concomitant with an increase in off-target mesoderm identity (1% to 32.9%) (**Figure 2.15D-E**). Transition scores significantly decrease as TF-mediated programming progresses ($P < 2.2E-16$; Wilcoxon Test; **Figure 2.15F**) with hybrid cell generation peaking at day 4 (**Figure 2.15G**). Within these hybrid states, we observe very few states that map to known developmental progression, particularly in DP compared to DD, suggesting the cells with multiple identities arise potentially from non-physiological cell fate specification. Altogether, the observed distribution of neurons across the spinal cord raises the possibility of potential deficiency in dorsal-ventral patterning in the *in vitro* systems, particularly direct programming, instructing addition of potential patterning signals could enhance MN production.

2.3.8 Retinoic Acid treatment alleviates off-target identities to enhance MN generation

Regionalization of the spinal cord integrates complex spatial and temporal patterning events (Delile et al., 2019), involving different signaling molecules, such as RA and SHH (Lara-Ramírez et al., 2013; Ribes et al., 2009). Therefore, we hypothesize that these signals could potentially fine-tune dorsal-ventral patterning to increase MN production *in vitro*. We directly programmed ESCs using the original protocol (Mazzoni et al., 2013) with or without 1 μM all-trans RA and/or 0.5 μM smoothened agonist (SAG - a hedgehog pathway activator) (**Figure 2.16C**). Single-cell RNA-sequencing was performed four days following embryoid body (EB) formation and doxycycline-induced reprogramming, capturing 17,163 cells from two independent biological replicates (cosine similarity = 0.988; **Figure 2.16A**). Using Seurat V4, we further integrated our TF-only sample with previous data (DP Days 4 and 11) from Briggs et al., 2017. Compositional analysis in between the two datasets shows that our sample recapitulates the published data (cosine similarity = 0.912). Classification using the same reference as described above reveals 11.6 ± 2.9% of cells as MNs in our dataset, representing at least a three-fold increase than the Briggs protocol. We speculate that this could be an effect due to the initial EB formation. In addition, Capybara annotates 13.7 ± 1.7% dorsal and 10.3 ± 1.2% ventral neurons, mirroring similar results with the Briggs protocol. Apart from cells with discrete identities, 35.4% of cells are classified as hybrids with a major representation by ESC-MN (22.7 ± 0.5%) (**Figure 2.16B**).

Next, we evaluate if the addition of RA and/or SAG is capable of increasing the MN population by reducing off-target cell generation. Breaking down the classifications to each treatment, we observed that RA treatment significantly increases MN generation over four-fold,

from 7.5 ± 1.6% to 33.4 ± 4.9% (P < 2.2E-16, randomization test), and significantly depletes the off-target dorsal population, mainly dl3, from 13.7 ± 2.0% to 5.8 ± 0.9% (P < 2.2E-16,

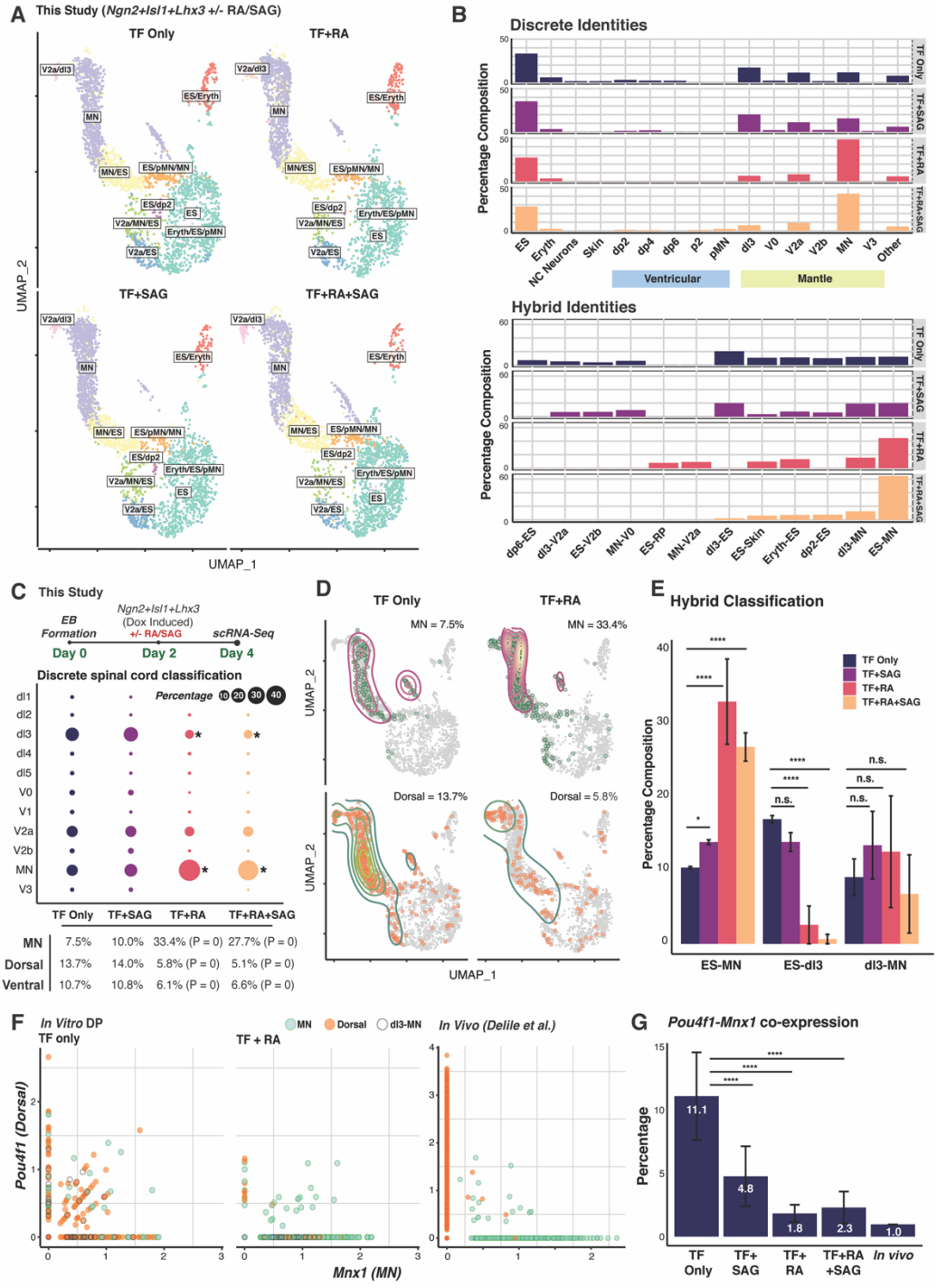

Figure Error! No text of specified style in document.**.16**: Experimental Validation with Modulation of Retinoic Acid and Sonic Hedgehog in MN Programming.

randomization test) (**Figure 2.16B-D**). In addition to the dorsal population, the major off-target ventral (mainly V2a) population is significantly depleted, from 10.7 ± 1.2% to 6.1 ± 0.5% ($P < 2.2E-16$, randomization test). Addition of only SAG slightly enriches for MN generation and provides no additional yields when added together with RA. A detailed exploration on the hybrid identities uncovers a significant enrichment upon the addition of RA ($P < 2.2E-16$, randomization test), whereas a significant depletion of the ESC-dl3 population ($P < 2.2E-16$, randomization test; **Figure 2.16E**). Furthermore, using scRNA-seq data, we assess the co-expression of MN marker, *Mnx1*, and dorsal neuron marker, *Pou4f1*, under different treatment conditions, and *in vivo* (Delile et al., 2019). When treated with RA, the percentage of the co-expressing population decreases by over 6-fold, more in line with the 1% co-expressing cells observed *in vivo*. Treatment with SAG reduced this population by more than half to 4.8% but provides no additional reductions in combination with RA (**Figure 2.16F-G**).

This exploration manifests the ability of Capybara to inspect aberrant dorsal-ventral patterning in MN programming and instruct the addition of signaling, such as RA, to enhance the efficiency and fidelity of MN generation *in vitro*.

2.3.9 An *in vivo* correlate for fibroblast to induced endoderm progenitor reprogramming

To further explore the application of Capybara, we turned our focus to interrogate a relatively uncharacterized direct conversion protocol, from mouse embryonic fibroblasts (MEFs) to induced endoderm progenitors (iEPs). It was first reported that, via the overexpression of two transcription factors, *Hnf4α* and *Foxa1*, MEFs can be reprogrammed to yield hepatocyte-like cells (Sekiya & Suzuki, 2011). Yet, more systematic characterization with CellNet with bulk data

as well as functional studies revealed the capability of these cells to engraft damaged colon and small intestine (Guo et al., 2019; Morris et al., 2014). Based on this potential, these cells have been suggested to resemble an endoderm progenitor-like state. Nevertheless, the *in vivo* correlate to these reprogrammed cells remains unclear.

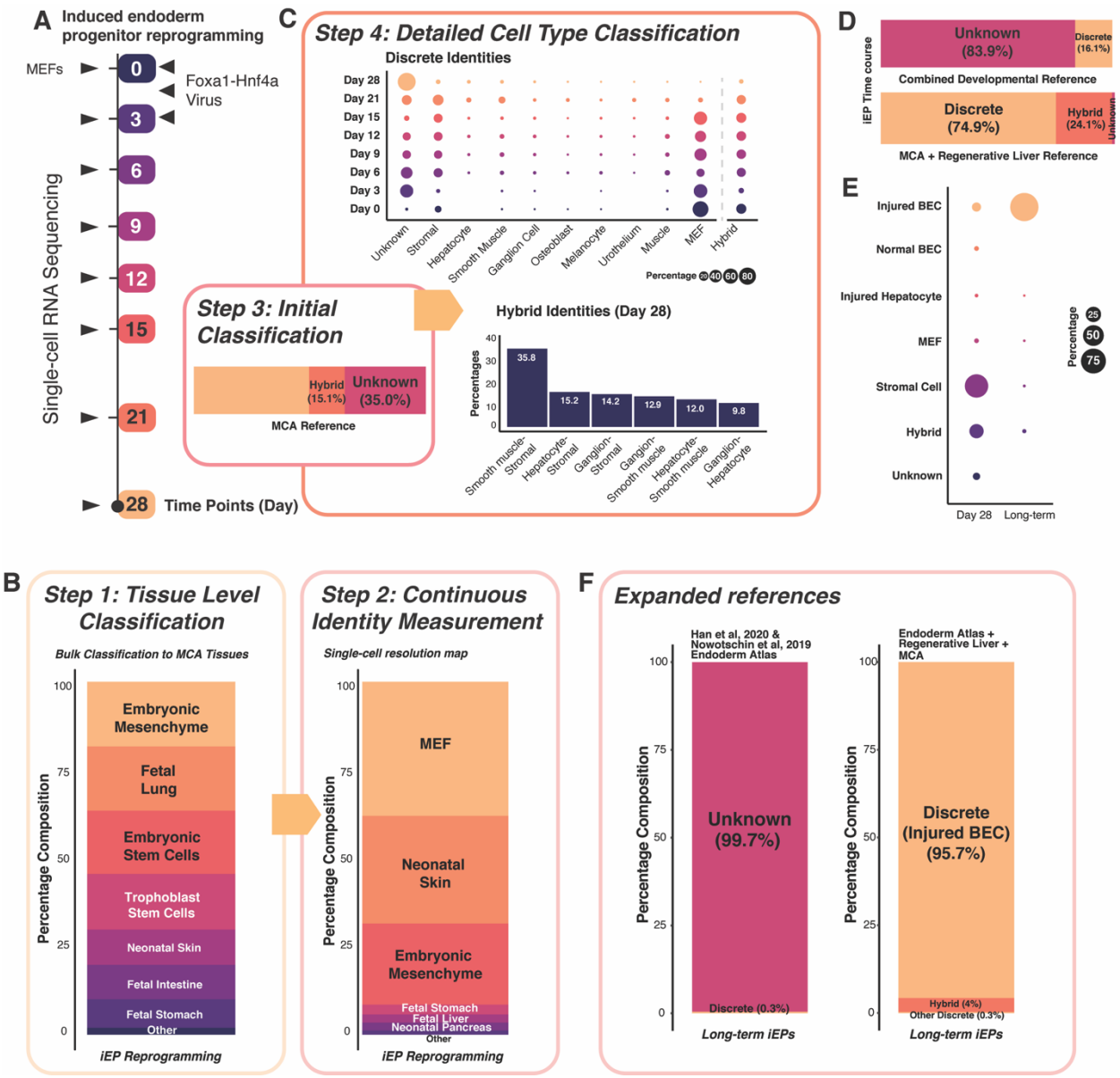

Figure Error! No text of specified style in document.**.17**: Capybara Analysis of fibroblast to induced Endoderm Progenitor (iEP) Reprogramming.

To characterize these cells, we apply Capybara to our previous 85,010 single-cell RNA-sequencing dataset spanning eight time points during the 4-week reprogramming process (**Figure 2.17A**; (Biddy et al., 2018)). Initial tissue selection with the high-resolution continuous measurement suggests major tissues, including embryonic mesenchyme and several endodermal populations (**Figure 2.17B**). Across the time course, epithelial cells gradually emerge to 5.9% at day 28, with few classified hepatocytes, mirroring previous findings using bulk transcriptome information (**Figure 2.17C-D**; (Morris et al., 2014)). Compared to the previously discussed systems, a substantial percentage of cells (35%) remain undetermined, indicating a potential missing cell type in the reference.

As iEPs have progenitor characteristics, we hypothesize that they represent a putative developmental progenitor. Therefore, we constructed an embryonic atlas containing endoderm and foregut tissues across a time course between E3.5 to E9.5 (L. Han et al., 2020; Nowotschin et al., 2019). However, iEPs remain largely unclassified using this reference (**Figure 2.17D-E**). An alternative hypothesis is that iEPs may represent a regenerative cell type considering their capacity to repair liver and colon (Guo et al., 2019; Morris et al., 2014; Sekiya & Suzuki, 2011). Further, it is found that the Hippo signaling effector, Yap1, plays a pivotal role in iEP generation (Kamimoto et al., 2020), resembling injured liver regeneration (Pepe-Mooney et al., 2019). Thus, we built a high-resolution reference based on both homeostatic and regenerative liver epithelial single-cell atlas, containing two main regenerative cell types: hepatocytes and biliary epithelial cells (BECs) (Pepe-Mooney et al., 2019). With this new reference, we classify day28 reprogrammed iEPs, and long-term cultured iEPs (LT-iEPs). Capybara classifies 8.3 ± 4.7% of day 28 reprogrammed iEPs (n = 20,532 cells) and 91.8 ± 7.49 % of LT-iEPs (n = 6,170 cells,

two independent biological replicates) as post-injury BECs (**Figure 2.17E-F**). These results lead us to perform experimental validation for this putative identity.

2.3.10 iEPs possesses characteristics of biliary epithelial cells

Under homeostasis, BECs are quiescent and arrange to form tubular, single-epithelial-layered bile ducts in the liver. When the liver is injured, BECs enter active proliferation and play a key role in regeneration (Kamimoto et al., 2016). Biliary epithelial cells isolated from the injured liver can be cultured *ex vivo*, passaged, and maintained long-term (Okabe et al., 2009). With support from the extracellular matrix, these cells form tubular or ductal structures in 3D-gel culture, mimicking *in vivo* BEC morphology (L. Jin et al., 2013; Lewis et al., 2018). Thus, to interrogate the BEC potential of iEPs, we cultured LT-iEPs, representing highest of injured BEC identity, in a 3D-gel sandwich culture that promotes tubule formation *in vitro* (Ogawa et al., 2015). We observed branching tubular structures after three days of culture and become more evident by day 5. Via immunostaining, we observe significant upregulation of established BEC markers, cytokeratin 19 (CK19), and epithelial cell adhesion molecule (EpCAM) by day 5 (**Figure 2.18A-C**). Moreover, side-by-side, we found that 2D-cultured LT-iEPs express *Ck19* with reduced *Epcam*, recapitulating the reported behavior of injured BECs after expansion *in vitro* for over 30 days (Okabe et al., 2009).

To further characterize 3D-cultured LT-iEPs, we performed single-cell RNA-sequencing on day 5 gel-cultured branching iEPs and obtained a total of 14,047 cells from two independent biological replicates. Application of Capybara on this dataset using the high-resolution reference established in the previous section demonstrates the significant emergence of a normal BEC population in 3D-cultured iEPs (14.3 ± 1.7%; $P < 2.2E-16$, randomization test). On the contrary, classification of cells under 2D culture demonstrates a lack of normal BEC population but an

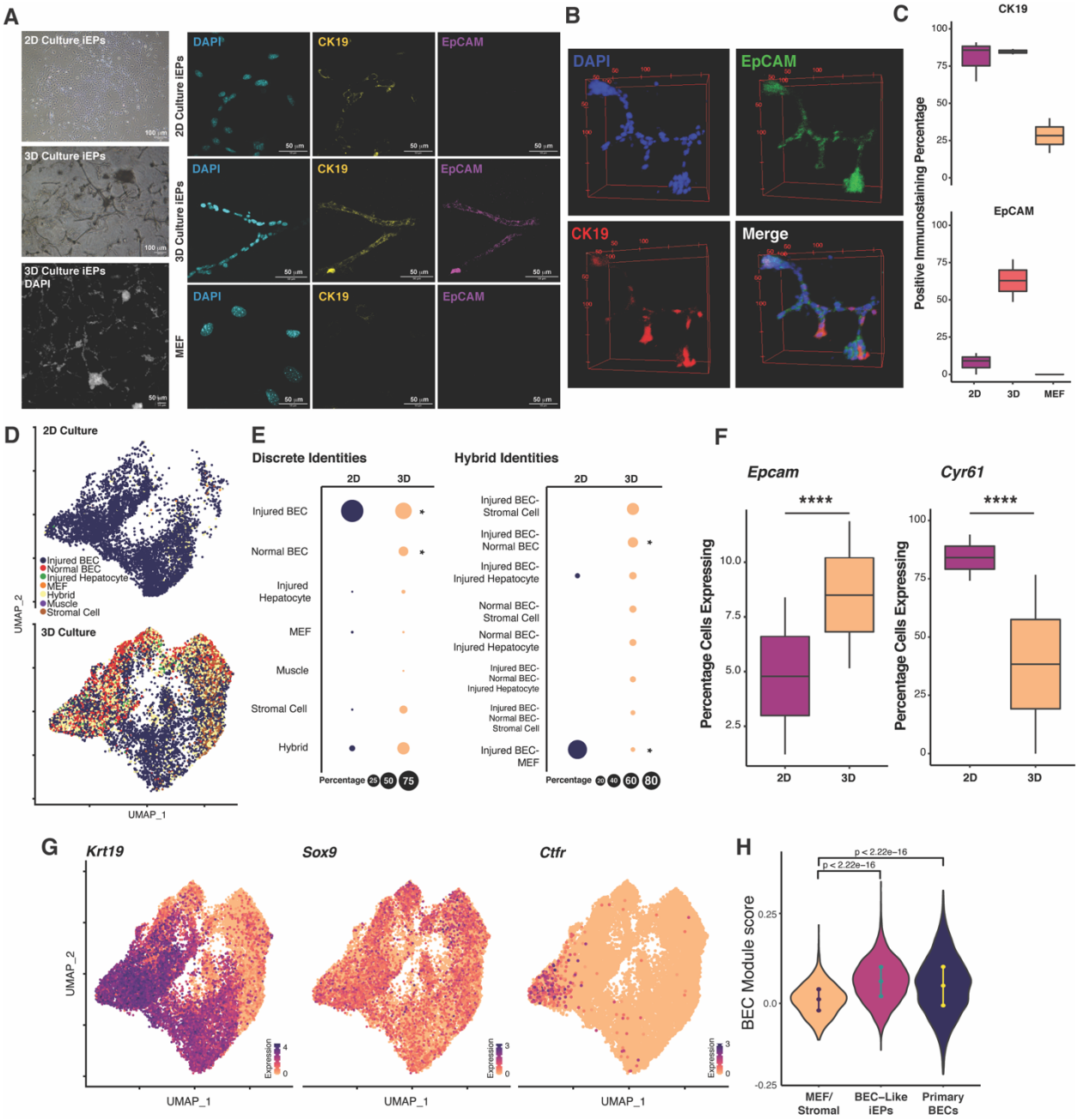

Figure Error! No text of specified style in document.**.18**: Experimental Validation of iEPs resembling injured Biliary Epithelial Cells.

enrichment for post-injury BECs (P < 2.2E-16, randomization test; **Figure 2.18D-E**).

Additionally, under 3D culture conditions, we observe a unique hybrid state between injured BEC and normal BEC hybrid (**Figure 2.18E, right**). As normal BECs emerge in 3D-culture, there is a significant increase of percentage of cells expressing *Epcam* (P < 2.2E-16,

randomization test, **Figure 2.18E**), mirroring the results with the above immunostaining. In addition, there is a significant reduction in the percentage of cells expressing *Cyr61,* a marker of injured BECs ($P < 2.2E-16$; randomization test, **Figure 2.18F**) (Pepe-Mooney et al., 2019). Moreover, these BEC-like cells express other specific BEC markers, such as *Sox9* (**Figure 2.18G**). We further identified a set of specific BEC markers from literature (Verhulst et al., 2019) and performed module score analysis, revealing higher BEC resemblance of the classified BEC-like cells (**Figure 2.18H**).

These results support the resemblance of this relatively uncharacterized cell type, iEPs, to *in vivo* biliary epithelial cells. Together, it highlights the application of Capybara to a relatively uncharacterized reprogramming protocol, incorporating this analysis with our previous lineage tracing studies to offer mechanistic insights into reprogramming, and to identify potential *in vivo* correlates of iEPs, warranting further study.

2.4 Discussion

Here, we have presented Capybara*,* an unsupervised method to quantitatively assess cell identity and fate transitions. A unique feature of Capybara is its ability to measure cell identity on a continuum, allowing the establishment of the statistical framework to classify hybrid cell states. Lineage tracing of myeloid differentiation demonstrates the multi-lineage potential of cells classified as monocyte-neutrophil and basophil-mast hybrids. In addition, we found the monocyte-neutrophil hybrids to be transcriptionally similar to a previously reported rare bistable hybrid that can give rise to both monocyte and granulocyte fates (Olsson et al., 2016). Further, we speculate that basophil-mast hybrids may represent a previously described rare basophil-mast progenitor cell (BMCP). Previously, BMCPs are found to have hybrid transcriptional profile and

have preferred differentiation toward the mast cell and basophil lineages (Dahlin et al., 2018). Moreover, we validate the existence of atrial-ventricular cardiomyocyte hybrids in cardiac reprogramming via RNA FISH and immunostaining, supporting the capacity of Capybara to identify cells sharing features from multiple identities.

Occurring at low frequency with a transient nature, hybrid states have been relatively poorly identified and characterized. High-throughput scRNA-seq brought new opportunities to capture and evaluate hybrid states (MacLean et al., 2018). However, few approaches currently exist to characterize hybrid identities. Though trajectory inference algorithms provide insights on the continuous trajectory, it may overlook the transition states when the terminal fates are undetermined or immature. Capybara represents a unique method in various cell differentiation or reprogramming paradigms to uncover hybrid states that may represent novel progenitor cell types or transitions between discrete identities. Hybrid cell states have been proposed to fulfill vital roles in biological processes as reviewed in Chapter 1 (MacLean et al., 2018). Using Capybara, we report wide-ranging hybrid states in the reprogramming paradigms we have analyzed here. Our analysis in hematopoiesis reveals hybrids that represent rational cell state transitions or reported bistable intermediates. Whereas in reprogramming paradigms, hybrids states are rarely overlapping with known developmental progression, leading us to a hypothesis that the trajectory of TF-mediated conversion involves non-physiological cell states to the target cell identity. Alternatively, the diversity of hybrid states could be from heterogeneity in the starting cell population, which often contains diverse cell types. Characterizing hybrids in this context might provide insight into the origins of successfully reprogramming cells.

The benefits of unsupervised cell-type classification go beyond the characterization of transition states, as we demonstrate via our analysis of several diverse cell fate engineering

strategies. For example, we observed regional patterning dynamics in the TF-mediated generation of cardiomyocytes and motor neurons. In the case of cardiomyocyte reprogramming, atrial cardiomyocytes dominate ventricular cardiomyocytes. An atrial-ventricular hybrid suggests that modulation in the protocol could potentially shift this balance. For instance, signaling factors play pivotal roles in chamber-specification in early cardiac development. Previous protocols where TGFβ signaling is inhibited and Wnt activated yield mainly ventricular cardiomyocytes (H. Wang et al., 2014), whereas protocols inhibiting TGFβ and Wnt generate mostly atrial-like cardiomyocytes, as we demonstrate here. Approaches to fine-tune this balance will be beneficial to increasing production of atrial or ventricular cardiomyocytes for further disease and treatment modeling. For motor neuron programming, our unbiased identification of a range of dorsal-ventral spinal neuron identities instructed the addition of signaling factors (RA and SAG) to alleviate this patterning deficiency. The modulation with such factors yields over 4-fold more motor neurons and increased specification of the resulting population. Finally, Capybara identified injured BECs as a potential *in vivo* correlate for iEPs, a poorly characterized product of reprogramming. Together, these observations highlight the broad application of Capybara to generate quantitative observations, suggest modulation on reprogramming strategies, and reveal the engineered cell identity and potential.

2.5 Limitations of Capybara

It is crucial to note that the performance of Capybara relies on the selection of appropriate reference datasets. We have designed the workflow with this limitation in mind, where initial tissue-level classification identifies the most appropriate tissue-specific single-cell reference to use. If an inappropriate reference is used, Capybara will classify cell identity as 'unknown,' as we demonstrate in our analysis of iEPs, which subsequently led us to a more suitable reference. A strength of Capybara to note here is that references can be constructed from a minimum of 30 cells, increasing the likelihood that rare cell types can be captured from selected references. As more diverse single-cell datasets become publicly available, we anticipate that this will support a much broader classification of cell identities. In addition, it is worth noting that though Capybara can distinguish unknown progenitors from unknown terminal fates, such classification is not to 100% accuracy due to the challenges in deconvolution of distributions. With future distribution modeling in-depth, Capybara will support more accurate classification of these two classes. Moreover, with expansion of complexity in the dataset, Capybara shows a longer runtime due to the exponential increase of rounds of permutation as increase of cell types and cell numbers. As parallelization being implemented in the pipeline, we expect that this will be alleviated to have shorter runtime with increasing computing power.

2.6 Materials and Methods

2.6.1 Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse Monoclonal Anti-MYL7 Antibody (B-10)	Santa Cruz Biotechnology	RRID:AB_10848272
Rabbit Monoclonal Anti-MYL2 Antibody	Abcam	RRID:AB_10563535
Rat Monoclonal Anti-Mouse CD326 (EpCAM) Antibody	BD Biosciences	RRID:AB_394370
Rabbit Monoclonal Anti-Cytokeratin 19 (CK19) Antibody	Abcam	RRID:AB_2281020
CD90.2 (Thy1.2) Monoclonal Antibody, FITC	Invitrogen	RRID:AB_273503
Alexa Fluor 546 Goat Anti-rabbit IgG	Invitrogen	RRID:AB_2534093
Alexa Fluor 488 Goat Anti-mouse IgG	Invitrogen	RRID:AB_2534088
Alexa Fluor 647 Goat Anti-rat IgG	Invitrogen	RRID:AB_141778
Bacterial and Virus Strains		
pMx-MGT	Wang et al., 2015	RRID:Addgene_111810
pGCDNSam-Hnf4-t2a-Foxa1	Morris et al., 2014	
pCL-Eco	Novus Biologicals	RRID:Addgene_12371
Stellar Competent Cells	Takara Bio	Cat #: 636763
Chemicals, Peptides, and Recombinant Proteins		
Fetal bovine serum (FBS)	Gibco	Cat #: 10082147

Fibroblast Medium-2	ScienCell Research Laboratories	Cat #: 2331
Matrigel (GFR Membrane Matrix)	Corning	Cat #: CB-40230
-mercaptoethanol	Life Technologies	Cat #: 21985023
X-tremeGENE9 Transfection Reagent	Sigma Aldrich	Cat #: 6365779001
XAV939	Cayman	Item #: 13031
SB431542	Cayman	Item #: 13596
CHIR99021	BioVision	Cat #: 1677
PD0325901	Sigma	Cat #: PZ0162
Leukemia Inhibitory Factor	Millipore	Cat #: LIF2050
Retinoic Acid (RA)		
Smoothened Agonist (SAG)	Millipore	Cat #: 566660
Epidermal Growth Factor	Sigma Aldrich	Cat #: E5160
Hepatocyte Growth Factor	Sigma Aldrich	Cat #: H9661
Doxycycline (Dox)	Sigma Aldrich	Cat #: D9891
L-Ascorbic Acid	Sigma Aldrich	Cat #: A8960
Insulin-Transferrin-Selenium-Ethanolamine (ITS-X)	Gibco	Cat #: 51500056
Gentle Cell Dissociation Reagent	STEMCELL Technologies	Cat #: 100-0485
Critical Commercial Assays		
RNAscope Multiplex Fluorescent v2 kit	Advanced Cell Diagnostics	Cat #: 323100
EasySep Mouse FITC Positive Selection Kit II	STEMCELL Technologies	Cat #: 17668
Ampure XP SPRI Beads	Beckman	B23318
Chromium Single Cell 3’ Library and Gel Bead Kit v2	10x Genomics	PN-120237

Chromium Single Cell 3' Chip kit v2	10x Genomics	PN-120236
Chromium i7 Multiplex Kit	10x Genomics	PN-120262
Chromium Next GEM Chip G Single Cell Kit	10x Genomics	PN-1000127
Library Construction Kit	10x Genomics	PN-1000196
Chromium Next GEM Single Cell 3' GEM Kit v3.1	10x Genomics	PN-1000130
Chromium Next GEM Single Cell 3' Gel Bead Kit v3.1	10x Genomics	PN-1000129
Dual Index Kit TT Set A	10x Genomics	PN-1000215
Deposited Data		
scRNA-seq	This paper	GEO: GSE145251
Hematopoiesis Development	Paul et al., 2015	GEO: GSE72859
Spinal Motor Neuron Differentiation and Programming	Briggs et al., 2017	GEO: GSE97391
Cardiac Reprogramming	Stone et al., 2019	GEO: GSE131328
MEF to iEP Reprogramming Time course	Biddy et al., 2018	GEO: GSE99915
Normal and Post Injury Hepatocytes and BECs	Pepe-Mooney et al., 2019	GEO: GSE125688
Mouse Gastrulation Atlas	Pijuan-Sala et al., 2019	GEO: GSE87038
Mouse Cell Atlas	Han et al., 2018	https://figshare.com/articles/MCA_DGE_Data/5435866
Tabula Muris	Tabula Muris Consortium et al., 2018	https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733
Developing Mouse Spinal Cord Atlas	Delile et al., 2019	E-MTAB-7320

Experimental Models: Cell Lines		
Mouse Cardiac Fibroblast (CD1, P0)	ScienCell Research Laboratories	Cat #: M6300
293T-17 Cells	ATCC	RRID:CVCL_1926
Primary Mouse Embryonic Fibroblast (C57BL/6, E13.5)		
NIL-V5 inducible ESC line	Mazzoni et al., 2013	
Experimental Models: Organisms/Strains		
Mouse: C57BL/6	The Jackson laboratory	RRID:IMSR_JAX:000664
Software and Algorithms		
ImageJ	Schneider et al., 2012	https://imagej.nih.gov/ij/
Seurat V4	Satija et al., 2015; Butler et al., 2018; Stuart et al., 2019	https://satijalab.org/seurat/articles/get_started.html
Quadprog	Turlach and Weingessel, 2007	https://cran.r-project.org/web/packages/quadprog/index.html
Cell Ranger v5.0.1	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest
Velocyto	La Manno et al., 2018	http://velocyto.org/
MASS	Venables and Ripley, 2002	https://cran.r-project.org/web/packages/MASS/MASS.pdf

mixdist	Macdonald and Du, 2018	https://cran.r-project.org/web/packages/mixdist/mixdist.pdf
Splatter	Zappia et al., 2017	https://github.com/Oshlack/splatter
OpenImageR	Mouselimis, 2021	https://cran.r-project.org/web/packages/OpenImageR/OpenImageR.pdf
PAGA	Wolf et al., 2019	https://github.com/theislab/paga
SCANPY	Wolf et al., 2018	https://scanpy.readthedocs.io/en/stable/
R-4.0.1	R Core Team, 2021	https://www.r-project.org/
RStudio	RStudio Team, 2020	https://www.rstudio.com/
Capybara	This Paper	https://github.com/morris-lab/Capybara
Other		
RNAscope probe Mm-Myh7	Advanced Cell Diagnostics	Cat #: 454741
RNAscope probe Mm-Myh6-C3	Advanced Cell Diagnostics	Cat #: 506251-C3
RNAscope probe Mm-Myl4-C2	Advanced Cell Diagnostics	Cat #: 443801-C2
RNAscope probe Mm-Actc1	Advanced Cell Diagnostics	Cat #: 510361
RNAscope probe Mm-Tnnc1-C3	Advanced Cell Diagnostics	Cat #: 511011-C3

2.6.2 Methods

Capybara Pipeline Overview: The Capybara pipeline comprises four major steps: 1) tissue-level classification, 2) high-resolution custom reference generation and continuous identity measurement, 3) initial classification into discrete, hybrid, or unknown identities, and 4) discrete cell type classification and hybrid identity scoring. Capybara code and documentation are available at: https://github.com/morris-lab/Capybara, along with detailed function descriptions and tutorials.

***Basis of Capybara*: Quadratic Programming (Setup).** Previous studies have measured continuous changes in cell identity using Quadratic Programming (QP) (Biddy et al., 2018; Treutlein et al., 2016), where The R package QuadProg was used for the calculation of QP scores. In brief, the underlying assumption is that each single-cell transcriptome profile can exist as a combination of fractional identities from all possible cell types, described as a linear combination of gene expression profiles from different cell types. This assumption allows us to model cell identity as a multivariate linear regression problem. For ease of biological interpretation, constraints are placed on the coefficients: they are bound between 0 and 1, and the sum of all coefficients does not exceed 1. These constraints limit the use of least squares estimators in this scenario, while QP is an optimization approach that minimizes a quadratic function under the given linear inequalities or equalities.

Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}$ denote the transcriptomic profile of genes g_1, g_2, …, g_n for a query cell, and $X_{g,t}$ denotes the reference dataset of the same set of genes by cell types t_1, t_2, …, t_m. The goal is

then to calculate the identity score vector f_t, such that the random error ϵ is minimized, as described below.

$$\min_f (Y - Xf)^T (Y - Xf) \text{ subject to } \sum_{i=1}^{m} f_i \leq 1, 0 \leq f_t \leq 1 \text{ for } t = 1, 2, \ldots, m$$

In addition to the fractional identity score matrix and the error term, each cell receives a Lagrangian multiplier, gauging how much the solution is pushed toward the constraints. Applying QP offers a quantitative evaluation of cell identity for each cell.

***Basis of Capybara*: Quadratic Programming (Data processing)** Before QP, using raw count matrices, we first perform log-normalization on both the reference and sample dataset. Let $M_{g,c}$ be the matrix with each row representing a gene and each column denoting a cell or a cell type. Let m denote the number of columns, and n denote the number of rows. Then, for each column of the matrix, $M_{*,c}$,

$$Normalized\ M_{*,c} = \frac{M_{*,c}}{\sum M_{*,c}} \times \frac{\sum_{g,c} M_{g,c}}{m}$$

The normalized matrix is then log-transformed with a base of 2 and pseudo-count of 1. The reference dataset undergoes further scaling to ensure that gene expression levels between datasets are comparable. We calculate the scaling factor as the ratio between $\frac{\sum_{g,c} M_{g,c}}{m}$ of the reference and sample. Further, we filter the gene list of both matrices to include only those genes shared between the reference and sample.

***Step 1*: Tissue-level classification.** The performance of Capybara hinges on the selection of an appropriate single-cell reference to classify cell identity. Before assessing cell identity at single-cell resolution, we perform a tissue-level classification designed to restrict the number of reference cell types included in downstream analysis, reducing excessive noise and dependencies caused by correlation across tissues in the final single-cell reference. This tissue-level classification is performed using bulk transcriptomics from ARCHS4, an exhaustive resource platform comprising the majority of published RNA-seq datasets (Lachmann et al., 2018). To achieve a relatively comprehensive and clean evaluation, we take a two-step approach: 1) construct a clean bulk RNA-seq reference and 2) correlation-based tissue classification.

(1) Bulk Reference Construction

ARCHS4, a platform that contains most published RNA-seq and ChIP-seq datasets (Lachmann et al., 2018), was mined for bulk RNA-seq data. ARCHS4 obtained raw datasets from the Gene Expression Omnibus, then realigned and processed through a uniform pipeline. Using this data bank, we first filtered the available datasets to retain only poly-A and total RNA-seq data from C57BL/6 mice. We then calculated Pearson's correlations on every sample pair from the same tissue. The top 90 samples with the highest Pearson's correlation scores for each of 30 tissues comprised the final bulk reference. For tissues with less than 90 samples, we took the entire sample set and randomly sampled with replacement to include 90 total samples. For the selected 90 samples for each tissue, we calculated the average reads per kilobase per million (RPKM) to build the final tissue-level transcriptome profile, containing a total of 30 tissues. We evaluated the quality of this bulk reference by calculating the identity scores of cells from manually annotated single-cell atlases (MCA; (Han et al., 2018) and Tabula Muris; (Tabula Muris Consortium et al., 2018)) based on this reconstructed reference. We randomly selected 90 cells

from each tissue of MCA or Tabula Muris and performed QP using the bulk reference, where we observe high scores when mapping the same tissue between single-cell and bulk datasets.

(2) Tissue-Level Classification

A potential concern of using QP to classify single cells directly is the correlation between similar cell types from different tissues. In this scenario, it could be challenging to tease classification results apart if high similarity to the correct cell type drives the high identity score. Thus, we first perform classification at the tissue level to restrict the number of reference cell types in the downstream analysis, reducing excessive noise and dependencies caused by correlation across tissues in the final single-cell reference. In general, the three primary inputs of this step include the single-cell reference (e.g., MCA), the sample single-cell dataset, and the constructed bulk reference. Using the tissue reference, we calculate QP scores for the single-cell reference as well as the sample, where we obtain two identity matrices. We then compute the Pearson's correlations of QP scores between each cell from the single-cell reference and each cell from the sample. We use a threshold at the 90^{th} percentile of the correlation matrix to binarize the correlation matrix, where a cell-cell pair with a correlation that is greater than the threshold is marked as 1; otherwise, 0. With the binarized matrix, we count the number of cells in each tissue of the reference mapping to the sample. If there is a significant percentage of reference cells of a tissue (over 70%) mapped, we record the tissue label. We then calculate the frequency of each tissue label in the sample. Tissues with a frequency of at least 0.5% sample cells were selected to proceed for further analysis at single-cell resolution. Here, it is worth noting that this tissue-level classification removes most irrelevant tissues but still provides a broad range of tissue types, at which point further downstream analysis removes non-relevant cell types (see 'Cardiomyocyte Reprogramming Analysis,' below). Additionally, having prior information regarding the tissues

involved can be beneficial to narrow down the tissue selection step, as exemplified by our analysis of hematopoiesis and spinal cord below.

***Step 2*: Generation of high-resolution custom references and continuous identity measurement.** Having identified the potential tissues present in a sample from the tissue-level classification, we next assemble a custom single-cell reference dataset containing the relevant cell types to classify the sample cells accurately. An example of such a reference dataset is the Mouse Cell Atlas (MCA; (Han et al., 2018)), which contains both fetal and adult mouse tissues. For each tissue, it offers a detailed breakdown of its cell types, including the same cell type with different marker genes, offering a high-resolution map of cell-type composition. This reference is assembled based on manual annotation of the specific cell types in the tissue involved. A unique feature of scRNA-seq is dropout - the failure to capture and detect known expressed genes and other technical variation (Lun et al., 2016). Due to the highly sparse nature of scRNA-seq data, an individual cell transcriptome may not provide a complete representation of a cell type. To alleviate the effect of these technical variations, we construct pseudo-bulk references for each cell type of each tissue. We sample 90 cells from each cell type for each tissue. For cell types with more than 90 cells, we calculate Pearson's correlations between each cell pair. Based on the correlation matrices, we select the most correlated 45 cells to ensure homogeneity and the least correlated 45 cells to ensure the capture of transcriptional diversity. Cell types that have fewer than 90 cells, but more than 30 cells are sampled with replacement to achieve a total of 90 cells. Summation of the counts of the selected 90 cells is used to construct the final high-resolution reference, assuming homogeneity in the annotated population of the original single-cell

reference. Application of QP using this 'high-resolution' reference generates a continuous measurement of cell identity as a linear combination of all cell types within the reference.

***Step 3*: Initial discrete, hybrid and unknown classification.** As aforementioned, the application of QP generates the continuous scores, from which we calculate a deviance metric of the scores from the expected score. In addition, QP provides two additional metrics: the error and Lagrangian multiplier, together with the continuous scores. Using these three metrics together, we evaluate the likelihood of a cell to have discrete, hybrid, or unknown identities, compared to the scoring metric of reference cells. This step can be evaluated in two parts: 1) deviance, 2) error, and Lagrangian multiplier.

(1) Deviance

The deviance is calculated via comparison between the identity scores to the expected scores ($\frac{1}{\text{number of cell types}}$), assuming a cell is equally similar to every cell type in the reference. We consider that cells with unique identities will have major deviations, while those with unknown identities will have minor deviations from the expectation. Let $f_{i,j}$ denote the score of a cell i on cell type j. The deviance is then calculated as follows:

$$\sum_{j=1}^{\text{number of cell types}} abs(f_{i,j} - \frac{1}{\text{number of cell types}})$$

Assuming the reference cells are accurately annotated with discrete identities, we first calculate the total deviance of each reference cell using the identity score matrix of the reference data. We further model the total deviance from the reference cells as a normal distribution, serving as the

reference distribution of discrete identity cells. Restricting the hybrid cells to have a maximum of two identities, we establish an ideal distribution for the hybrid cells by shifting the density of discrete identities by 2x standard deviation to the left. Lastly, the unknowns are expected to have an even lower deviation than the hybrid cells. We then calculate the total deviance of each sample cell in the same manner. With the established distributions, we obtain probability scores from the evaluation of each distribution by computing $\boxed{P(X \geq x)}$. Cells with $\boxed{P(discrete) \geq 0.01 \,\&\, P(hybrid) \geq 0.95}$ are considered as discrete. Cells with $\boxed{P(discrete) \leq 0.05 \,\&\, P(hybrid) \geq 0.01 \,\&\, P(unknown) \geq 0.95}$ are considered hybrids. Cells with $\boxed{P(hybrid) \leq 0.01 \,\&\, P(unknown) \geq 0}$ are considered unknowns.

(2) Error & Lagrangian Multiplier

The selection of cells to build the high-resolution custom reference includes both highly correlated and uncorrelated cells in the population of the corresponding cell type. Such a selection scheme provides a multimodal distribution for the error and Lagrangian multiplier metric, serving as background distributions for the extreme cases of matching and unmatching cells. Based on the multimodal density, we build an ideal distribution for the test samples, where the mean is the weighted mean of the mixed normal distribution, and the standard deviation is the weighted standard deviation of the mixed distribution. We consider unknown cells will establish higher error (on the right tail). In contrast, hybrid cells will have comparable levels of error but a lower Lagrangian multiplier (on the left tail). In addition, unknowns can potentially be distinguished into unknown progenitors vs. unknown end states by considering the combination of the two distributions. As unknown end states take both higher error and Lagrangian multiplier, unknown progenitors are considered to have a relatively high error, but even lower Lagrangian multiplier compared to the hybrids. Yet, due to the challenges in

deconvolving overlapping distributions, we could partially distinguish the two unknown cell types leveraging the combination of the two metrics.

***Step 4*: Discrete cell type classification and hybrid identity scoring.** While continuous identity scores are informative, discrete cell-type assignment offers a more practical assessment of cell-type composition for a biological system. One approach to call discrete cell types is to apply a threshold to the calculated continuous scores. However, threshold selection and quality of the custom high-resolution reference can bias cell type calling via this approach. To overcome this limitation, we apply QP to score cells in the single-cell reference against the bulk reference. This strategy accounts for reference quality, enabling background matrices to be generated, charting the distributions of possible identity scores for each cell type. We then take a two-step approach to give discrete and hybrid cell type classification: 1) empirical p-value calculation via randomized testing, and 2) Mann-Whitney-based binarization and classification.

(1) Empirical P-Value Calculation via Randomized Testing

With the constructed single-cell reference, we apply QP to both the sample and reference single-cell datasets to generate continuous measurements of cell identity. Let M_R denote the identity score matrix of the reference data with a total of m cell types and $90 \cdot m$ cells, where $f_{R,i,j}$ denotes the score of reference cell i on cell type j. Let M_S denote the identity score matrix of the sample data with a total of m cell types and n cells, where $f_{S,i,j}$ denotes the score of sample cell i on cell type j. We then carry out the following steps to calculate the empirical p-values. (1) For each cell type in M_R, we randomly sampled 1000 times and constructed a

background density of the identity scores, $D_R = [f_{resample,1}, \ldots, f_{resample,1000}]$. (2) For each score in the identity matrices, we calculate the empirical p-value as follows:

$$p_{R,i,j} = \frac{\sum_{h=1}^{1000} \mathbb{I}\left(f_{resample,h} > f_{R,i,j}\right)}{1000}, \; p_{S,i,j} = \frac{\sum_{h=1}^{1000} \mathbb{I}\left(f_{resample,h} > f_{S,i,j}\right)}{1000},$$

where $\mathbb{I}(*) = 1$ if $(*)$ is true; otherwise, $\mathbb{I}(*) = 0$. (3) Next, we repeat steps (1) and (2) for a total of 50 rounds, recording the empirical p-values matrix for each cell of both the reference and the sample. The result of this step includes two lists of p-value matrices: one for the reference and the other for the sample. For each cell, each column of the p-value matrix denotes a cell type, while each row describes each round of 50.

(2) Binarization and Classification

From randomized testing, we construct two lists of empirical p-value matrices: one for all sample cells, P_S, and the other for all reference cells, P_R. Using the list for all reference cells together with its annotation data, we computed a benchmark empirical p-value for each cell type. Specifically, the annotation data contains cell barcodes with their associated annotated cell types. For each cell c and its annotated cell type t^0, we identified the corresponding list of empirical p-values, $P^{(c)}_{R,*,t^0}$. As a result, we construct a possible range of p-values for each cell type, t, from which we generate the benchmark values. For each cell type t, we eliminate the outlier p-values and select the maximum p-value of the remaining cells as the final benchmark score, $B_t = [B_{t1}, \ldots, B_{tm}]$. Outlier p-values are identified based on the definition of outliers in the boxplot (outside of 1.5x the interquartile range above the third quantile or below the first quantile).

Next, we move forward to evaluate the sample list with the initial classification results. If the cell is initially considered as an unknown, it is skipped for this statistical framework

evaluation. The length of the sample list, n, is the number of sample cells. The n^{th} empirical p-value matrix $P_{S,k,t}^{(n)}$ in the list defines empirical p-value for the n^{th} sample cell belonging to reference cell type t under the k^{th} resampling background, where $1 \leq k \leq 50$. We rank all empirical p-values inside the matrix, from the lowest to the greatest, and break any tie by averaging. The rank-sum for each column *t* of $P_{S,k,t}^{(n)}$ is then calculated, and the cell type with the lowest rank-sum, *t**, is determined to be the putative identity for cell *c*. We then compare $\mathrm{mean}(P_{S,*,t^*}^{(c)})$ to B_{t^*} to assign an identity for cell c. To assign cells harboring hybrid identities, recapitulating those identities, we perform a pairwise Mann–Whitney U test between the *t** column and other columns of $P_{R,k,t}^{(n)}$. For any cell type *t'* with rank-sum that is not significantly greater than the rank-sum of *t** (significant level=0.05), we consider *t'* to be one of multiple identities of query *c* along with *t**. Applying this process to each cell, we generated a binary matrix with 1 = putative identities. Further, we generate a classification table with labeled cell types for each cell barcode.

Transition Scoring. Cells with multiple identities (hybrid cells) label critical transition states in different trajectories. Building on this concept, we also measure the strength and frequency of connection to the discrete cell state, which provides a metric that we define as a 'transition score.' The calculation of transition scores only involves cells with hybrid identities. In general, using QP, each cell receives fractional identity scores for different cell types in the reference. Interpreting QP as probabilities of the cell transitioning to each discrete cell identity, we use QP scores as a measure of transition probability.

For a cell marked with multiple identities, we consider a transition between the cell to its terminal cell state as events with the transition probability measured by QP scores $P_{i,j}$, where *i*

denotes the cell and *j* denotes the cell state. Therefore, based on information theory, the information of such transition event can be measured as $I(transition) = -\log(P_{i,j})$. We further consider how much information the terminal cell state has received, which can be defined as:

$$I(received) = P_{i,j} \times I(transition) = -P_{i,j} \times \log(P_{i,j}).$$

Thus, the total amount of information received for cell state *j* from *n* connected cells can be computed as:

$$I(received) = \sum_{i=1}^{n} -P_{i,j} \times \log(P_{i,j}).$$

The measurement appears to be similar to Shannon's entropy, while we note that with each cell independently in transition, probabilities from all events do not necessarily add up to 1, distinguishing it from a measure of entropy. Here, to demonstrate this metric, consider an example as demonstrated in Figure 3F, where Cells 1 to 5 harbor multiple identities connecting Cell State I to III. In this example scenario, the transition score for Cell State II can be calculated as:

$$\begin{aligned} &I(Cell\ State\ II) \\ &\quad = -P_{1,II} \times \log(P_{1,II}) - P_{2,II} \\ &\quad \times \log(P_{2,II}) - P_{3,II} \times \log(P_{3,II}) - P_{5,II} \times \log(P_{5,II}). \end{aligned}$$

Using such measurement, we incorporate the frequency as well as the likelihood of connection such that high information labels a discrete cell state associated with an abundance of dynamic cell transitions.

Benchmarking Capybara. To assess the efficacy and robustness of Capybara to classify cell identity, here we validate each step and demonstrate its basic functionality. In the first step of the Capybara pipeline, tissue-level classification, accuracy is pivotal as it helps reduce noise from other cell types that are not present in the sample. We evaluate the validity of the tissue reference transcriptome based on the identity scores of annotated single-cell atlases (Han et al., 2018; Tabula Muris Consortium et al., 2018). We randomly selected 90 cells from each tissue of MCA and Tabula Muris using the bulk reference, where we observed higher scores mapping of the same tissue between single-cell and bulk.

Next, we assess the classification functionality of Capybara. In this step, we use a benchmarking algorithm that was recently developed to compare a range of single-cell classification approaches using an array of publicly available datasets (Abdelaal et al., 2019). Briefly, we perform 10-fold cross-validation using various datasets. Here, the predictions from the methods are assessed based on the area under the receiver operating characteristics (AUROC) using the multiclass.roc function in R. Based on five human pancreatic datasets and Allen Mouse Brain Atlas, the performance of Capybara indicates similar accuracy (rank 5) and median F1 score (rank 4.2) with reasonable runtime when benchmarked against ten other classifiers (**Figure 2.4**). In this automatic benchmarking method, 5-fold cross-validation provides a relatively large training set (80%) compared to the test set (20%). A key feature of Capybara is its flexible requirement in terms of training set size. We find that a minimum number of 90 cells sampled from each cell type is required to perform accurate classification. For cell types with fewer than 90 cells, we require a minimum of 30 cells, from which a 90-cell sample will be drawn with replacement from the pool. Using this minimum number of cells, we evaluate our performance using the *Tabula Muris* mouse cell atlas (Tabula Muris Consortium et al., 2018).

Using AUROC scores and accuracy, we benchmark our method against two other classification approaches, scmap (Kiselev et al., 2018) and SingleCellNet (Tan & Cahan, 2019). As a result, we demonstrate the comparable performance of Capybara with excellent performance (AUROC > 0.8).

Generation of simulated data. We use Splatter, an R-based simulation framework based on Gamma-Poisson distribution, to simulate a single-cell dataset comprising distinct differentiation paths (Zappia et al., 2017). We design the cell population to originate from a progenitor state (P1) bifurcating toward two discrete states (E1: End State #1; P2: Progenitor State #2). P2 progenitor cells bifurcate further toward end states #2 and #3 (E2 and E3, respectively; **Figure 1B, C**). Using this simulated dataset, we assess if Capybara can 1) Capture cells with unique identities; 2) Identify cells that do not correlate with any cell types in the reference; 3) Characterize transition cells with multiple identities. E1, P2, and E2 cell populations were defined as within 5% variability of the max pseudotime at each terminal. We construct a reference using 90 of the most correlated and diverse cells from E1, P2, and E2 cell populations. Cells in E1, P2, and E2 that did not contribute to the reference are used to test the efficacy of accurate classification. The remaining cell populations are not included in the reference to test how Capybara classifies cells with no correlates in the reference.

Capybara Analysis with Previously Published scRNA-seq data

(1) Paul et al. (2015) Mouse Hematopoiesis Analysis

We obtained the raw hematopoiesis count data from GSE72859 (Paul et al., 2015c). The data was processed and clustered using SCANPY (Wolf et al., 2018) and PAGA (Wolf et al., 2019). From processing, we included 3,451 genes in the dataset of 2,730 cells. We first perform tissue-

level classification with the bulk reference established using ARCHS4, as described in the previous sections. From this, we identified three major relevant tissues: primary mesenchymal stem cells (bone marrow mesenchyme), bone marrow, and bone marrow (c-Kit). Further breakdown of these three major tissues using the MCA (X. Han et al., 2018) resulted in 49 different cell types. We constructed the high-resolution reference using these 49 cell types. 90 cells were selected from each cell type as described above and saved as the reference single-cell dataset. Followed by preprocessing, we applied QP on the reference and sample single-cell dataset, based on which we further categorized them to discrete, hybrid and unknown, calculated empirical p-values, performed binarization and classification. We projected cells with single identities onto the cluster embedding from PAGA. Cells with hybrid identities were isolated, and we extracted the pseudotime for themselves and their terminal cell identities. We re-assessed these multiple identity cells using their scores. If one of the identities scored near zero (score < 10E-3), we considered such identity as inaccurate and discarded it. In this process, we re-evaluated transitioning cells, retaining only those cells with relatively higher shared identity scores. For a hybrid identity to be considered usable, it needs to be represented by more than 0.5% of the sample population. Using this filtering, we alleviate potential transitions due to noise but maintain the more putative transitions. Wilcoxon test was used to compare if the pseudotime density differs comparing hybrids with their discrete identity parts.

(2) Weinreb et al. (2020) Mouse Hematopoiesis Lineage-Tracing Analysis

We obtained the normalized InDrop single-cell data, annotation, and SPRING embedding for mouse hematopoiesis lineage-tracing dataset from https://github.com/AllonKleinLab/paper-data. In this analysis, we mainly focused on the Lin^{-} Sca^{+} $cKit^{+}$ (LSK) population, containing a total of 72,946 cells. We constructed the high-resolution reference using 90 cells in each of the major

day 6 differentiated cell types, including basophils, eosinophils, mast cells, monocytes, and neutrophils. Considering the myeloid differentiation culture condition, we selected these five populations as they are the continuous expanding populations from day 2 to day 6. Following preprocessing, we generated the continuous identity score measurements for the remaining LSK cells using QP, followed by initial classification, binarization, and classification. Leveraging the lineage information in the dataset, we then identified clones that contained hybrid cells on day 4 to evaluate their siblings on day 4 and progeny on day 6. To compare with the hybrid-containing clones, we also identified day 4 clones that are strictly represented by the discrete compartments of the hybrids. For instance, while assessing monocyte-neutrophil hybrids, we compared the siblings and progeny of day 4 clones strictly represented by monocytes or neutrophils. Enrichment of populations was tested via randomization testing. Briefly, for the clones representing the hybrid and its siblings, we randomly sampled the same number of cells as the clones from the entire population and calculated the proportion of the cell type represented in the sample. We iterated this process 10,000 times to establish a distribution. The likelihood of proportions presented in the hybrid family was evaluated based on this density, providing empirical p-values.

The raw count of the data was obtained by taking the reciprocal of the smallest non-zero gene expression of each cell, following https://github.com/AllonKleinLab/paper-data/issues/7. The data was then processed and clustered using SCANPY (Wolf et al., 2018) and PAGA (Wolf et al., 2019), following the tutorials. Connectivity among clusters was obtained from PAGA analysis. Assuming clusters connected to two differentiated fate clusters as potential hybrid clusters, we evaluated if these cells were lineage restricted. From the connected cluster, we

identified the day 4 cells and assessed their siblings and progeny. The enrichment of each population is tested via randomization testing as described above.

(3) Pijuan-Sala et al. (2019) Mouse Gastrulation Transition Score Analysis

We obtained 10x scRNA-seq UMI count data and annotation of mouse gastrulation from GSE87038 (Pijuan-Sala et al., 2019), containing a total of 139,331 cells. The dataset was processed using Seurat (Butler et al., 2018; Satija et al., 2015). We performed classification using all 23 tissues, composed of 361 cell types, in the adult MCA as a reference directly (X. Han et al., 2018). We constructed the high-resolution reference using these annotated cells. Following preprocessing, we generated the continuous identity score measurements for these cells using QP, followed by initial classification, binarization, and classification. We then performed Capybara transition scoring analysis for each sample, analyzing transition score distributions of each annotated cell type from Pijuan-Sala et al.

(4) Stone et al. (2019) Cardiomyocyte Reprogramming Analysis

We obtained the 10x single-cell RNA-sequencing count data from GSE131328 (Stone et al., 2019), containing a total of 30,729 cells. This dataset was processed using Seurat (Butler et al., 2018; Satija et al., 2015) and clustered using UMAP. We used raw data from the filtered cells and genes as input into the Capybara pipeline. We next performed tissue-level classification using ARCHS4 (Lachmann et al., 2018), as described in previous sections, revealing four major tissues, including neonatal skin, neonatal heart, fetal stomach, and fetal lung. Further breakdown of these tissues using MCA (X. Han et al., 2018) contains 57 cell types. We constructed the high-resolution reference using these annotated cells. Following preprocessing, we generated the continuous identity score measures of these cells using QP, based on which we further performed initial classification, binarization, and classification. We calculated the percentage of each

identified cell type in the population. Additionally, we computed the transition scores for the cell states involved in transitions. We performed transition score comparisons using a one-sided Wilcoxon test. We identified region-specific markers from MuscleDB (http://muscledb.org/mouse/mRNA/).

(5) Briggs et al. (2017) *In Vitro* Spinal Cord Motor Neuron Derivation Analysis

We obtained 10x scRNA-seq UMI count data and annotation of developing mouse spinal cord from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7320/ (Delile et al., 2019), including a total of 38,976 cells. We removed unannotated cells and built a high-resolution reference for each cell type at each developmental stage (E9.5 to E13.5), resulting in a total of 118 cell types (19 types for E9.5, 26 for E10.5, 26 for E11.5, 25 for E12.5, 22 for E13.5). We also included embryonic stem cells from the MCA (X. Han et al., 2018). We obtained the InDrop single-cell dataset for *in vitro* spinal cord motor neuron derivation from GSE97391 (Briggs et al., 2017). The *in vitro* datasets were processed and clustered using Seurat (Butler et al., 2018; Satija et al., 2015). From processing, we included 7,860 genes and 7,799 genes in the dataset of 1,984 cells and 2,720 cells in direct programming (DP) and direct differentiation (DD), respectively. We analyzed ESCs from each protocol separately. Following preprocessing, we applied QP using the high-resolution reference on four datasets, including two ESC populations, DP and DD. Based on the identity score matrices, we categorized them into discrete, hybrid, and unknown, calculated the empirical p-value matrices, performed binarization and classification. Cells with discrete identities were separated to calculate the composition in the ventricular zone and mantle zone. The ventricular zone also included the neural crest neurons and mesoderm lineage. Hybrid cells were filtered and refined, as described in the above hematopoiesis section. With the QP scores attached to each identity in the mixed set, we calculated the transition scores for the cell

states involved, as described in the transition scoring section. We compared transition scores between different timepoint via a one-sided Wilcoxon test.

(6) Biddy et al. (2018) MEF to Induced Endoderm Progenitor Analysis

We processed scRNA-seq data of induced endoderm progenitor (iEP) reprogramming, as previously described (Biddy et al., 2018). In brief, Scater was used to normalize (McCarthy et al., 2017) the data across time points, and Seurat (Butler et al., 2018; Satija et al., 2015) was used to integrate biological replicates, perform clustering, and visualize cells using *t*-SNE. We performed tissue-level classification using ARCHS4 (Lachmann et al., 2018), as described in previous sections, highlighting the involvement of 9 potential tissues, containing a total of 73 cell types. Following the construction of a high-resolution reference, we performed preprocessing on the reference and the sample, on which we then applied QP to generate the identity score matrices. Further, we categorized them into discrete, hybrid, and unknown, calculated the p-value matrices, and performed binarization and classification. We calculated the percent composition of each cell type. Cells with hybrid identities were filtered as described in the above hematopoiesis section, represented by more than 0.5% cells of the population.

We obtained scRNA-seq data of biliary epithelial cells (BECs) and hepatocytes, before and after injury, from GSE125688 (Pepe-Mooney et al., 2019). We built a custom high-resolution reference by incorporating additional tissues from the MCA: fetal liver, MEFs, and embryonic mesenchyme. The long-term iEPs were cultured for 12 months before collection and processing. We had previously used these cells to engraft mouse colon (Guo et al., 2019; Morris et al., 2014). The long-term iEP dataset was processed, filtered, and clustered using Seurat, resulting in 2,008 cells. We then constructed the high-resolution reference panel with 20 cell types and performed preprocessing on the reference and single-cell sample. Application of QP

using the processed reference and long-term iEP and iEP reprogramming datasets provides us the continuous metric of identity scores, from which we carried out initial classification, binarization, and classification. Gene expression was compared between groups via Wilcoxon test.

10x alignment, digital gene expression matrix generation.

The Cell Ranger v5.0.1 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to align reads, process, and filter data generated using 10x Chromium single-cell gene expression platform. Following this step, the default Cell Ranger pipeline was implemented to generate the filtered output data for downstream analysis.

scRNA-seq Data Processing

To process and analyze scRNA-seq data, we used the R package, Seurat V4, following the tutorial (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html). Briefly, each sample was pre-processed based on RNA counts and mitochondria percentages and then normalized. The highly variable genes were then identified, followed by scaling and dimensional reduction via PCA. With the selected number of components, graph-based clustering and UMAP plotting were further performed.

(1) Cardiac Reprogramming

When comparing our data with Stone et al., 2019, the data were integrated using canonical correlation analysis and mutual nearest neighbor with the Seurat V4 pipeline (Butler et al., 2018; Stuart et al., 2019). The similarity between the dataset was evaluated based on cosine similarity between the cluster representation in the two datasets.

(2) In Vitro Motor Neuron Programming and iEP Reprogramming

To evaluate reproducibility, datasets for each treatment were integrated across the two biological replicates following the same process described above. The integrated Seurat objects were further integrated to evaluate the effect of different treatment groups.

scRNA-seq Data Capybara Analysis

With the tissues identified from the corresponding publicly available dataset, we started from step 2 of the Capybara pipeline for the single-cell RNA-sequencing data we generated for this study. Using the raw counts, we performed preprocessing on the reference and the sample, on which we then applied QP to generate the identity score matrices. Further, we categorized them into discrete, hybrid, and unknown, calculated the p-value matrices, and performed binarization and classification. We calculated the percent composition of each cell type. Cells with hybrid identities were filtered as described in the above hematopoiesis section, represented by more than 0.5% cells of the population.

2.6.3 Experimental Methods

Reprogramming Virus Production

The retrovirus for cardiac and induced endoderm progenitor (iEP) reprogramming was freshly prepared. 293T cells (RRID:CVCL_1926) were maintained and passaged in fibroblast media (10% FBS, 1x penicillin-streptomycin, 1x -Mercaptoethanol, in DMEM). 293T cells were seeded at a density of 3 million per 10-cm plate the day before transfection. The following day, the cells were transfected with pMX-MGT (RRID:Addgene_111810) or pGCDN-Sam-Hnf4-t2a-Foxa1 with 5g of pCL-Eco (RRID:Addgene_12371), using X-tremeGENE 9 DNA transfection reagent (Sigma, 6365779001) according to the manufacturer's instructions. Media was replaced with

fresh fibroblast media the following day. Retrovirus was harvested the next day by taking the supernatant from the transfected plate and filtered through a 45- m syringe filter. 500x protamine sulfate was added to the viral media prior to transduction of the mouse cardiac fibroblasts (cardiac) or mouse embryonic fibroblasts (iEP).

Cardiomyocyte Reprogramming

Direct cardiac reprogramming was performed using primary cardiac fibroblasts derived from a postnatal day 2 CD1 Mouse (ScienCell, Catalog #M6300) following previously published protocols (Ieda et al., 2010; Qian et al., 2013; Stone et al., 2019). Briefly, cardiac fibroblasts (MCFs) were cultured overnight on gelatin-coated plates in Fibroblast Medium-2 (ScienCell, Cat. #2331). MCFs were passaged 1-2 times, cultured for ~5 days for expansion, and prepared for selection of $Thy1^+$ (RRID:AB_273503) cells by MACS. After sorting, MCFs were plated at a density around 100k~200k per 6-cm dish pre-treated overnight with gelatin (day -1). $Thy1^+$ MCFs were infected with freshly harvested pMX-MGT retrovirus (L. Wang et al., 2015) (day 0). The viral media was replaced with fresh cardiomyocyte media (10% M199, 10% FBS, 1% NEAA, 1% sodium pyruvate, 1x penicillin-streptomycin, 1x Glutamax, in DMEM) containing 2.6M SB431542 (Cayman Chemical, Catalog #13031) or DMSO as a vehicle control (day 1). 5M XAV939 (Cayman Chemical, Catalog #13596) or DMSO was added to the plate without media change (day +1). The media was replaced with fresh cardiomyocyte media two days after the last addition of small molecule (day +3). Media was renewed every 2~3 days. The cells were collected, filtered through a 70m strainer, resuspended in 1% BSA in PBS, and counted on Day 14 for scRNA-seq (see below).

Immunostaining for day 14 Reprogrammed Cells in Cardiac Reprogramming

Mouse cardiac fibroblasts were generated as described above. On day 13 of the reprogramming process, the cells were transferred to 4-Chamber Culture Slides (Falcon). On the next day, the cells were rinsed with 1x DPBS and fixed in 4% paraformaldehyde for 20 minutes at room temperature. The samples were then washed with 1x DPBS three times, permeabilized, and blocked with blocking buffer (0.2% TritonX-100 and 3% FBS in DPBS) for an hour. The primary antibodies, MYL2 (RRID:AB_10563535) and MYL7 (RRID:AB_10848272), were diluted 1:250 (MYL2 and MYL7) in the blocking buffer. The blocking buffer was then removed from the sample, and the primary antibodies were added. The samples were incubated with the primary antibody at 4°C overnight (12hr). The samples were then washed for 5 minutes three times. The secondary antibodies, Alexa Fluor 546 Goat Anti-rabbit IgG (RRID:AB_2534093) and Alexa Fluor 488 Goat Anti-mouse (RRID:AB_2534088), were diluted 1:1000 in the blocking buffer. The secondary antibodies were added and incubated at 4°C overnight (12hr). The samples were washed again for 5 minutes, three times. 100 l of 300 nM DAPI (Invitrogen) was added to each slide chamber and incubated at room temperature for 1 minute. The samples were washed for 5 minutes three times. In the last wash, we aspirate all the DPBS and remove the chamber from the slides. A coverslip was then applied with ProLong Gold Antifade Mountant (Invitrogen). The slides were imaged using an Olympus FV1200 Confocal Microscope with 10x, 20x, and 40x water objectives. The number of positive cells was counted in each channel using ImageJ with “Analyze Particles” function. The total number of cells was determined based on the DAPI counts.

RNA Fluorescent in Situ Hybridization

On day 13 of mouse cardiac reprogramming (above), the cells were transferred to 4-Chamber Culture Slides (Falcon). The next day, cells were rinsed with 1x DPBS and fixed with 10% Neutral Buffered Formalin for 30 minutes at room temperature. RNAscope Multiplex Fluorescent v2 kit (Advanced Cell Diagnostics) was used to perform RNA-FISH to probe *Myh6* and *Myh7* mRNA, following the protocol for cultured adherent cell samples. Briefly, the slides were treated with hydrogen peroxide and RNAscope protease III (Advanced Cell Diagnostics). Then, the slides were incubated to hybridize with the specified probes using the RNAscope HybEZ II Oven (Advanced Cell Diagnostics). Probes were then amplified, and the HRP signal was developed using the TSA Fluorescein Plus Evaluation Kit. Finally, DAPI (Advanced Cell Diagnostics) staining was applied to the slide, and a coverslip was then applied with ProLong Gold Antifade Mountant (Invitrogen). The slides were imaged using an Olympus FV1200 Confocal Microscope with 40x and 60x water objectives. Images were then analyzed using computational quantification: RNA-FISH images were first processed through ImageJ to ensure the same maximum intensity across images. Through ImageJ, individual cells were segmented into smaller regions. All three channels of each selection were stored for further processing with a custom R script to quantify intensity at single-cell resolution. Individual cells were read in as individual matrices, where averaged green and red intensity were calculated and compared.

Motor Neuron Programming from mouse ESC

The NIL (Ngn2-Isl1-Lhx3)-V5 inducible ESC line was previously described (Mazzoni et al., 2013b). All the inducible ESC lines were grown in 2-inhibitors medium (Advanced DMEM/F12:Neurobasal (1:1) Medium (Gibco), supplemented with 2.5% ESC-grade fetal

bovine serum (vol/vol, Corning), N2 (Gibco), B27 (Gibco), 2mM L-glutamine (Gibco), 0.1 mM β-mercaptoethanol (Gibco), 1000 U/ml leukemia inhibitory factor (Millipore), 3μM CHIR (BioVision) and 1 μM PD0325901 (Sigma). To obtain Embryoid bodies (EBs) ESC were trypsinized (Gibco) and 3 x 10^5 cells were plated in each 100 mm dish in AK medium (Advanced DMEM/F12:Neurobasal (1:1) Medium, 10% Knockout SR (vol/vol) (Gibco), Pen/Strep (Gibco), 2mM L-glutamine and 0.1 mM 2-mercaptoethanol) (day -2). After 48 hr, EBs were passed 1:2, and the inducible cassette was induced by adding 3 μg/ml of Doxycycline (Sigma) and/or 1 μM all-trans retinoic acid and/or 0.5 μM smoothened agonist (SAG) (Millipore, 566660). Differentiating EBs were washed three times with PBS, dissociated with Trypsin, and pipetted into single-cell suspensions. After 48 hr, cells were preserved in methanol (Alles et al., 2017) before processing for single-cell profiling (below).

Long-term iEP culture

Mouse Embryonic Fibroblasts were derived from the C57BL/6J strain (RRID:IMSR_JAX:000664). All animal procedures were based on animal care guidelines approved by the Institutional Animal Care and Use Committee. Mouse embryonic fibroblasts were reprogrammed to iHeps/iEPs, as in Sekiya and Suzuki (2011). Briefly, fibroblasts were prepared from E13.5 embryos and serially transduced with polyethylene glycol concentrated Hnf4□-t2a-Foxa1, followed by culture on gelatin for two weeks in hepato-medium (DMEM:F-12, supplemented with 10% FBS, 1 mg/ml insulin (Sigma-Aldrich), dexamethasone (Sigma-Aldrich), 10 mM nicotinamide (Sigma-Aldrich), 2 mM L-glutamine, 50 mM -mercaptoethanol (Life Technologies), and penicillin/streptomycin, containing 20 ng/ml hepatocyte growth factor (Sigma-Aldrich), and 20 ng/ml epidermal growth factor (Sigma-Aldrich)), after which the emerging iEPs were cultured on collagen and passaged twice per week for three months.

Matrigel Sandwich Culture of Long-term iEPs

We adapted the culturing methods in Ogawa et al., 2015 and Okabe et al., 2009. Briefly, 70% Matrigel in DMEM was added as a bottom layer to the plate, 96-well glass-bottom plate (20 l) or glass-bottom 35mm -Dish (100 l; iBidi) or 6-well plate (100 l). The bottom layer was allowed to solidify at 37°C for 30 minutes. Long-term iEPs were dissociated using 0.05% Trypsin-EDTA (diluted from 0.25%; Gibco, Cat #: 25200056). The cells were resuspended in pre-chilled OVM-medium (William's E medium, supplemented with 10% FBS, dexamethasone, 10 mM nicotinamide, 2 mM L-glutamine, 0.2 mM ascorbic acid, 20 mM HEPES, 1% penicillin/streptomycin, 1% sodium pyruvate, 0.15% of 7.5% sodium bicarbonate, 14 mM glucose, containing 1x ITS-X (Gibco), 20 ng/ml hepatocyte growth factor (Sigma-Aldrich), and 20 ng/ml epidermal growth factor (Sigma-Aldrich)). The top layer was prepared with 40% Matrigel, with1.2mg/ml Collagen Type I (Gibco, stock of 3mg/ml), mixed with 20k cells for each well of 96-well plate, or 80k for each well of a 6-well plate. After 30 minutes, the top layer was added to the plate and allowed to solidify and set in the incubator for 45 minutes. After the top layer solidified, pre-warmed OVM medium was added. The medium was changed every other day. After five days of gel culture, the cells were imaged and processed for single-cell RNA-sequencing.

iEP Preparation from Matrigel Culture for Single-Cell Profiling

Cells in 3D gel-culture were dissociated using a combination of Type I Collagenase (Gibco; 100 l of 500 mg/ml Type I Collagenase in 1 ml of OVM) and 1ml of Gentle cell dissociation reagent (STEMCELL Technologies). Briefly, the medium was carefully pipetted off, and 1 ml of enzyme mix was added to each well of the 6-well plate. The plate was incubated at 37°C for 10 minutes.

The partially dissociated gel was collected into a 15-ml Falcon tube, further mixed on a rocker for 15 minutes at room temperature. The cells were then pelleted at 300xg for 5 minutes and washed with 1ml HBSS. The solution was passed through a 27-gauge needle using a 3 ml syringe. The cells were counted, centrifuged, resuspended in 0.04% BSA in 1x DPBS, and passed through a 70 m cell strainer before loading onto the 10x Chromium Single Cell Chip.

Immunofluorescence Staining of Branching iEPs

3D-cultured iEPs were cultured in 96-well glass-bottom plate or glass-bottom 35mm -Dish for imaging. The cells ready for immunostaining was washed with 1x DPBS three times and fixed overnight in 4% paraformaldehyde at 4°C. The fixed sample was then washed twice with 1x DPBS for 15 minutes, permeabilized, and blocked with blocking buffer (0.2% TritonX-100 and 3% FBS in DPBS) for 10 minutes at room temperature. The primary antibodies, EpCAM (RRID:AB_394370) and CK19 (RRID:AB_2281020), were diluted at 1:100 (EpCAM) and 1:200 (CK19) in the blocking buffer. The samples were incubated with the primary antibodies at 4°C overnight (12hr). The sample was then washed for 15 minutes three times. Secondary antibodies Alexa Fluor 546 Goat Anti-rabbit (RRID:AB_2534093), and Alexa Fluor 647 Goat Anti-rat IgG (RRID:AB_141778), were diluted 1:500 (Alexa Fluor 546 and Alexa Fluor 647) in the blocking buffer. 50 l of 300 nM DAPI (Invitrogen) was added with the secondary antibodies and incubated at 4°C overnight (12hr). The samples were washed again for 15 minutes three times. Cells in 96-wells were imaged as 3D z-stack images using Zeiss LSM 880 Confocal with Airyscan with 40x air objective. Samples in the 35-mm dish were transferred to a slide, covered with a coverslip with ProLong Gold Antifade Mountant (Invitrogen), and imaged using an Olympus FV1200 Confocal Microscope with 40x water objective. Representative images were chosen.

Single-cell profiling

For single-cell library preparation on the 10x Genomics platform, we used: the Chromium Single Cell 3′ Library & Gel Bead Kit v2 (PN-120237), Chromium Single Cell 3′ Chip kit v2 (PN-120236), and Chromium i7 Multiplex Kit (PN-120262), according to the manufacturer's instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide. Prior to cell capture, methanol-fixed cells were placed on ice, then spun at 3000rpm for 5 minutes at 4°C, followed by resuspension and rehydration in PBS, according to Alles et al., 2017. 17,000 cells were loaded per lane of the chip, aiming to capture 10,000 single-cell transcriptomes. The resulting cDNA libraries were quantified on an Agilent Tapestation and sequenced on an Illumina HiSeq 2500. For analysis of cardiomyocyte reprogramming, The Chromium Single Cell 3' (v2) Reagent Kits (PN-120237, PN-120236, PN-120262) were used to prepare single-cell RNA-seq libraries, according to manufacturer's guidelines. Libraries were pooled and sequenced on an Illumina NextSeq 550. For motor neuron programming, prior to loading the 10x chip, methanol-fixed cells were counted, spun, resuspended in 1% BSA in PBS, and counted again, according to 10x Genomics methanol fixation protocol. The Chromium Single Cell 3' (v2) Reagent Kits (PN-120237, PN-120236, PN-120262) were used to prepare single-cell RNA-seq libraries, according to manufacturer's guidelines. Libraries were pooled and sequenced on an Illumina NextSeq 550.

2.7 Detailed Figure Legends

Figure 2.1: Previous Applications of Quadratic Programming (QP). (A) Continuous identity score measuring at different time points during MEF to iEP reprogramming using previously available microarray data (Sekiya & Suzuki, 2011). (B) Using the identity scores, single-cell data was labelled from fibroblast state to the reprogrammed states. (C) Continuous identity score measuring of epithelial cells from the small intestine using a reference constructed from a single-cell survey of the epithelium of mouse intestine (Haber et al., 2017). (D) Using the identity scores, single-cell data was labelled of different cell populations in the intestine. (E-F) Gradual identity measures of hematopoietic stem/progenitor cell development from fetal to adult state. (G) Continuous identity score measuring at different time point during fibroblast to neuron reprogramming with microRNAs.

Figure 2.2: Overview of the Capybara Workflow. Four major steps of the Capybara workflow: 1) tissue-level classification; 2) continuous identity measurement; 3) initial classification; 4) discrete cell type classification. In brief, we first perform tissue-level classification to restrict the number of reference cell types in the downstream analysis. We further identify the correlated tissues in a single-cell atlas, such as the Mouse Cell Atlas (MCA; (Han et al., 2018)). Using only the highly correlated tissues, we build a high-resolution reference. Application of quadratic programming (QP) provides a continuous measure of cell identity as a linear combination of all cell types within the reference. With the quality metrics from QP, we perform an initial classification to categorize the sample cells into discrete, hybrid, or unknown identities. Lastly, cells with discrete or hybrid identities are mapped to their corresponding cell types using a statistical framework.

Figure 2.3: QP Metric Demonstration and Tissue-Level Reference Validation. (A) Distributions of quadratic programming metrics, including error, Lagrangian multiplier, and deviance. Black line represents the distribution of these metrics in the reference. Other lines represent the ideal distributions for discrete, hybrid, and unknown, which are modeled based on the reference metrics. Assuming each cell in the reference has a discrete identity, the ideal distribution of deviance for discrete cells is modeled based on the reference identity scores. The ideal distribution of deviance for hybrid is constructed as 2 standard deviations shifted to the left from the one for discrete, and the deviance distribution for unknown is 2-standard-deviation shifted to the left from the one for hybrid. (B) Correlation heatmap of bulk expression between each pair of the 30 identified tissues from ARCHS4. The bulk expression is the logged reads per kilobase per million (RPKM). The RPKM was calculated as the average of the 90 samples selected for each tissue (See Methods). The color of each square denotes the correlation between tissues, where a lighter color indicates a higher correlation.

Figure 2.4: Benchmarking of Capybara using Established Pipeline and in-house Cross-Validation. (A) Evaluation using five human pancreatic datasets and the Allen Mouse Brain atlas against ten other classifiers, using an established benchmarking pipeline (Abdelaal et al., 2019). The performance of the classification was evaluated by the area under the receiver operating characteristic (AUROC) and the mean total time in seconds. The color of each square labels the assessment of each aspect. (B) Cross-validation using Tabula Muris processed with 10x droplet-based or Smart-seq2 technologies. In this evaluation, 90 cells were sampled from each cell type of each tissue to construct the high-resolution reference. Sample cells used for reference construction were used as training sets for scMap and SingleCellNet. The remaining cells were used as test samples. The performance was evaluated by AUROC scores.

Figure 2.5: Validation of Capybara for Datasets generated from Different Single-Cell Platforms. Cross-platform validation between Baron et al., 2016 (InDrop) and Muraro et al., 2016 (CEL-Seq2). Left: Proportions of cells mapping between Capybara classification and the original annotation. A lighter color indicates a higher percentage agreement. Right: Log-normalized expression levels of marker genes across classified cell types, calculated using Seurat. The marker genes were identified from Baron et al., 2016. Gamma cells: PPY, SERTM1, CARTPT; Activated stellate cells: PDGFRA; Acinar cells: REG3A, PRSS1, CPA1; Ductal cells: KRT19.

Figure 2.6: Simulation Study for Proof of Concept. (A) Design of the simulation study with Splatter. Differentiation of single cells is simulated from the progenitor state to two end states. Further, from end-state #2, cell differentiation into two other end states is simulated. Red Node: Test cells; Black Node: Cells to build the reference. (B) Pseudotime presentation of the simulated single-cell dataset with different states marked in circles. (C) Expected classification outcomes using the simulated single-cell dataset. (D) Heatmap comparing the result from Capybara to the ground truth as designed with color showing percentage matched. The transition state is mapped with a hybrid classification; end states are mapped with corresponding discrete identities; unknown cell types are correctly identified. (E) Classification of the simulated dataset (described in Figure 1) using scMap with cell or cluster mapping. The color of each square labels the percentage of the actual annotation mapped to each scMap annotation.

Figure 2.7: Application of Capybara to Classify Hematopoietic Cell Identity. We first applied Capybara analysis to a well-characterized cell differentiation paradigm: hematopoiesis. (A) Stepwise cell-type classification of an existing hematopoiesis dataset (Paul et al., 2015),

consisting of 2,730 myeloid progenitors. 1) Tissue-level classification identifies three major tissues, including bone marrow, primary mesenchymal stem cells, and peripheral blood; 2) Using the higher resolution MCA reference, the main cell types correspond to two primary tissues: bone marrow and peripheral blood; 3) Initial classification places cells into discrete, hybrid, and unknown identity categories; 4) Cells are mapped to discrete and hybrid cell types. 'Prog': Progenitor; 'MPP/HSPC': Multipotent Progenitors/Hematopoietic Stem and Progenitor Cell. 'Other': includes basophils, eosinophil progenitors, B cell progenitors, macrophages, dendritic cells, and NK cells. (B) PAGA embedding of the myeloid progenitor dataset. 'FA': Force Atlas. (I) Manual annotation of clusters, based on Paul et al., 2015 and Wolf et al., 2019. 'DC': Dendritic Cell; 'MEP': Megakaryocyte and Erythroid Progenitor; 'Ery': Erythroid; 'Lymph': Lymphoid; 'GMP': Granulocyte and Monocyte Progenitor. (II) Projection of the major Capybara-classified populations. (C) Heatmap comparing Capybara classifications to the manual annotations. Color denotes the percentage matched. (D) Comparison of log-normalized expression between classified cell types and other cells. The following genes are used as key markers for different populations: MPPs: Cd34; Megakaryocytes: Itga2b; Neutrophils: Cebpe; Monocytes: Csf1r; Erythrocytes: Car2. A Wilcoxon test was used for significance testing (****: P <=0.0001).

Figure 2.8: Evaluation of Hematopoietic Hybrid Cells against Pseudotime. (A) Diffusion pseudotime analysis projected onto the PAGA embedding. (B) Pseudotime for each cell type. MPP/HSPC classified cells, representing early cell states in the hematopoietic lineage, are associated with the lowest pseudotime values. (C) Projection of the major hybrid population, 'erythrocyte progenitor–erythroblasts,' along with discrete erythrocyte progenitors and erythroblasts onto the erythroid lineage. (D) Comparison of pseudotime between the hybrid and

discrete identities shown in (C). (E) Top: Projection of the classified hybrid cell populations on the PAGA-guided clustering, along with their corresponding discrete identities. Bottom: Comparison of pseudotime between hybrids and their discrete identity counterparts. A Wilcoxon test was used for significance testing. (F) PAGA-guided clustering of the Paul et al. dataset, consisting of a total of 2,730 myeloid progenitors enriched from mouse bone marrow, revealing a total of 24 clusters. PAGA connectivity links the clusters containing hybrid cells, as expected, but does not pinpoint the hybrid state. (G) Correlation between log (PAGA connectivity scores) and log (transition scores).

Figure 2.9: Evaluation of hybrid cells using ground-truth lineage tracing. (A) Schematic of the Weinreb 2020 hematopoietic lineage-tracing dataset. Hematopoietic progenitor cells were isolated, barcoded at Day 0, and first collected for scRNA-seq at Day 2. The cells were then cultured under myeloid differentiation conditions and collected at Days 4 and 6 for scRNA-seq. The barcode captured from the single-cell transcriptome allows the capture of clonally related cells. Thus, we can test whether hybrid cells possess significant multi-lineage potential using this dataset. (B) Major hybrid populations identified by Capybara. (C) Cell-type composition of cells clonally related to major hybrid cell types. Upper row: Cell-type distribution of the overall population. Lower rows: Average cell-type breakdown for all clonal relatives of each major hybrid cell population. For example, clonal relatives of Monocyte-neutrophil hybrids are significantly enriched for discrete Neutrophil, Monocyte, and hybrid Monocyte-neutrophil populations (*: $P <= 0.05$, n.s.: $P > 0.05$, randomization test; 24 +/- 4 cells per clone, 10 clones, 243 cells). (D) State-fate analysis: Detailed breakdown of populations across timepoints. We identified clones composed of discrete or hybrid identities at Day 4 and assessed the cell-type composition of their differentiated clonal relatives at Day 6. The top two rows represent the day

6 clonal relatives derived from day 4 lineage-restricted clones. The bottom row represents day 6 clonal relatives derived from day 4 clones containing hybrid cells. Significant population enrichment is assessed using a randomization test (*: P <= 0.05). (E) SPRING projection of cells related to monocyte- or neutrophil-restricted clones and monocyte-neutrophil hybrid clones.

Figure 2.10: Discrete Identities in Ground-Truth Lineage Tracing Dataset and Comparison to Previous Identified bistable states. A) Manual annotation of the major differentiated cell populations identified from the Weinreb et al., 2020 hematopoiesis lineage tracing dataset, projected onto the SPRING embedding. (B) Comparison between Capybara classification and the manual annotation. We selected the major differentiated myeloid cell types, including basophils, eosinophils, mast cells, monocytes, and neutrophils, as a reference. Cell types not included in the reference are correctly identified as 'Unknown.' 95.1% of cells with unknown identities are labeled as undifferentiated in the original Weinreb et al. annotation. The color of each square denotes the percentage agreement between the manual and Capybara-based annotations. (C) Integrative analysis of monocyte-neutrophil hybrids, monocytes, neutrophils (identified by Capybara), and IG2 and GMP cells (identified in Olsson et al., 2016). IG2 cells have the potential to differentiate into monocytes or granulocytes. To evaluate monocyte-neutrophil hybrids, we used Seurat V4 to integrate monocytes, neutrophils, and monocyte-neutrophil hybrids with the IG2 and GMP population, showing the overlap between monocyte-neutrophil hybrids and the IG2 population. (D) The color of each square denotes the percentage of each population in each cluster. (E) Cosine similarity of percentage cluster representations across all populations, showing the highest similarity between monocyte-neutrophil hybrids and the IG2 population. (F) Left: PAGA-guided clustering of the Weinreb dataset, revealing a total of 15 clusters, along with calculated connectivity between these clusters. Right: Projection of

Weinreb annotations onto the PAGA-guided clustering and lineage analysis. We identified cluster 3 as discrete monocytes, cluster 7 as discrete neutrophils, cluster 8 as discrete basophils, and cluster 14 as discrete mast cells. The most connected clusters were then identified based on the connectivity plot to test if they are lineage-restricted using clonal data. A randomization test was used for testing the significant enrichment of populations on day 6 (*: P <=0.05)

Figure 2.11: Transition Scores and Its Validation. (A) Schematic of the principles underlying Capybara's transition metric. Squares represent discrete cell identities. Circles represent cells with hybrid identities. $P_{i,j}$ represents the probability of cell i transitioning to cell type j. Then, we calculate the transition score of each cell type as the accumulated information received from each cell connection. (B) Demonstration of transition scores using the mouse gastrulation atlas (Pijuan-Sala et al., 2019), embryonic stem cell (ESC) dataset (Briggs et al., 2017), and cardiomyocyte dataset (Stone et al., 2019). Transition scores decrease as development progresses during mouse gastrulation. ESCs under maintenance conditions demonstrate low transition scores as they are not actively differentiating. On the other hand, cardiomyocytes show low transition scores as they are terminally differentiated. (****: P <= 0.0001, ***: P <= 0.001, **: P <= 0.01, *: P <= 0.05, Wilcoxon test). (C) Comparison between RNA velocity and Capybara transition scores in cardiac reprogramming (Stone et al., 2019). RNA velocity vectors projected onto the UMAP embedding (Square: area containing a small number of moving vectors; Circle: area containing a large number of moving vectors). Middle: Transition scores projected onto the UMAP embedding (Square: area corresponding to the RNA velocity plot showing low transition scores; Circle: area corresponding to the RNA velocity plot showing high transition scores). (D) correlation between log (velocity vector counts) and log (transition scores).

Figure 2.12: Capybara Analysis of Direct Cardiac Reprogramming. (A) Stone et al., 2019 experimental design: Cardiac fibroblasts were harvested from neonatal mice. Three transcription factors, Mef2c, Gata4, and Tbx5 (MGT), were overexpressed via retroviral transduction (Day -1), followed by the addition of TGFβ inhibitor (Day 0) and Wnt inhibitor (Day 1). Reprogrammed cells were collected on days -1, 1, 2, 3, 7, and 14 for scRNA-seq. (B) Initial tissue-level classification of this dataset reveals four major tissues, which the high-resolution atlas restricts to two major tissues with the higher-resolution reference. (C) Discrete, hybrid and unknown cell composition. Top: Capybara classified cell type composition over the time course. The dot size is proportional to the discrete population size. Bottom: Hybrid cell identities of cells after 14 days of reprogramming. (D) UMAP plot of the cardiac reprogramming dataset. Top: Collection time points projected onto the UMAP embedding; Bottom: Projection of the two major target populations: atrial and ventricular cardiomyocytes. (E) Normalized gene expression of cardiac markers labeling atrial vs. ventricular cardiomyocytes (****: P <=0.0001, Wilcoxon test).

Figure 2.13: Transition Scores and Hybrid Identities of Direct Cardiac Reprogramming. (A) Transition scores across the cardiac reprogramming process (****: P <= 0.0001, ***: P <= 0.001, **: P <= 0.01, *: P <= 0.05, Wilcoxon test). (B) Breakdown of cell types listed as “Other” in Figure 4. (C) Detailed hybrid cell-type breakdown for each time point of the cardiac reprogramming time course. (D) Integration of 10x dataset generated in this study with days 7 and 14 from the Stone et al., 2019 dataset, Cosine similarity = 0.804 between the independent studies. Bottom right: Projection of the two major cardiomyocyte populations, atrial and ventricular cardiomyocytes, onto the integrated UMAP. (E) Detailed cell type and hybrid classification of our in-house reprogramming dataset (n = 5,107 cells, two independent biological

replicates). Note that in our experiment, the day 14 cells were not sorted for the cardiac reporter gene used in Stone et al., 2019. Left: Discrete cell-type composition. Right: Hybrid cell-type composition.

Figure 2.14: Experimental Validation of Hybrid Cells using RNA FISH and immunostaining. (A) Normalized gene expression of Actc1 and Tnnc1 in the classified cells in the scRNA-seq data. (B) RNA FISH of cells expressing only Myh7 or Myh6. (C) RNA FISH for Myl4 (atrial) and Actc1, Tnnc1 (ventricular) showing discrete and hybrid cells. Scale bars = 10mm. (D) Negative staining controls for RNA FISH and immunostaining. (E) Immunofluorescence for MYL7 (atrial) and MYL2 (ventricular) proteins showing cells expressing a single protein (discrete) or co-expressing both proteins (hybrids). (F) Left: Quantification of discrete cells identified from immunostaining and scRNA-seq. MYL7-expressing cells are enriched relative to MYL2-expressing cells, confirming the atrial/ventricular bias observed from scRNA-seq data. Right: Hybrid cell percentages measured by RNA FISH, immunofluorescence, and scRNA-seq.

Figure 2.15: Capybara Analysis of Spinal Motor Neuron Differentiation and Programming (Briggs et al 2017). (A) Differentiation vs. direct programming of motor neurons (MNs) from ESCs (Briggs et al., 2017). (B) Spinal cord domains and regions included in the reference atlas. (C) Capybara classification steps: with prior knowledge that these protocols aim to generate spinal motor neurons, we selected a single-cell spinal cord development atlas (Delile et al., 2019) as the high-resolution reference, omitting the general tissue selection step. We identified the major embryonic development stages corresponding to each protocol. (D) Cell type composition over the differentiation and programming time courses. Dot size is proportional to the discrete

population size. (E) UMAP plot of this dataset, divided by protocol (Direct Programming: DP; Directed Differentiation: DD). Top: Projection of time points onto the UMAP embedding. Bottom: Projection of major discrete cell types, as identified via Capybara analysis, onto the UMAP embedding. (F) Transition scores for each protocol across experimental time points (****: P <=0.0001, n.s. = not significant, Wilcoxon test). (G) Major hybrid populations in the direct programming and differentiation protocols. For each time point, we show the percentage of each hybrid type.

Figure 2.16: Experimental Validation with Modulation of Retinoic Acid and Sonic Hedgehog in MN Reprogramming. (A) UMAP plot of the integrated datasets generated in this study, including four samples with different treatment groups. The major Capybara classification is labelled for each cluster. (B) All discrete and hybrid cell type compositions for each treatment group. (C) Top: Experimental design in this study. After 48 hr of embryoid body (EB) formation, we induced the original reprogramming cocktail (Ngn2, Isl1, Lhx3: NIL) with retinoic acid (RA) and/or smoothened agonist (SAG). Day 4 cells were collected for scRNA-seq (Cells profiled: TF only: 2,926; TF + Shh: 3,340; TF + RA: 2,828; TF + RA +Shh: 8,042; two independent biological replicates per condition). Bottom: Differentiated spinal cord neuron composition and percentage breakdown of dorsal-ventral populations for each treatment group (*: P <= 0.05, ****: P <= 0.0001, randomization test). (D) UMAP plot of MN and dorsal populations comparing TF-only to TF + RA groups. (E) Major hybrid populations across treatment groups (****: P <= 0.0001, *: P <= 0.05; Two sample Chi-squared test). (F) Expression of the dorsal marker, Pou4f1, and motor neuron marker, Mnx1, comparing this study to the in vivo study (Delile et al., 2019). (G) Quantification of co-expressing cells in across treatment groups and in vivo (****: P <= 0.0001; Two sample Chi-squared test)

Figure 2.17: Capybara Analysis of fibroblast to induced Endoderm Progenitor (iEP) Reprogramming. (A) MEF to iEP reprogramming (Biddy et al., 2018). (B) Tissue-level classification of this dataset reveals seven major tissues. With the high-resolution reference, the relevant tissues narrow down to three major tissues. (C) Top: Discrete cell type composition over the time course. Dot size is proportional to the discrete population size. Bottom: Hybrid cell identity proportions of cells after 28 days of reprogramming. (D) Cell composition with a developmental atlas (Han et al., 2020; Nowotschin et al., 2019) or a combined regenerative liver atlas (Han et al., 2018; Pepe-Mooney et al., 2019). (E) Cell type composition of day 28 and long-term cultured iEPs (n = 20,532 and 2,008 cells). (F) Cell-type classification using an expanded reference with embryonic populations. We selected a foregut organogenesis atlas (Han et al., 2020) and a gut tube development atlas (Nowotschin et al., 2018). We combined the two references and performed Capybara analysis. 99.9% of cells were classified as 'unknown.' We then combined this endoderm development atlas with the regenerative liver atlas, where the iEPs were primarily classified as injured BECs. (C) Hybrid populations in day 28 and long-term iEPs.

Figure 2.18: Experimental Validation of iEPs resembling injured Biliary Epithelial Cells. (A) Imaging of 2D and 3D-cultured iEPs. Left: Bright-field images and DAPI field of composite z-stack images. Right: Immunofluorescence images of DAPI, CK19, and EpCAM staining. (B) 3D-rendering of a microscopy z-stack for 3D-cultured iEPs stained for DAPI, EpCAM, and CK19, demonstrating branching. (C) Quantification of the percentage of positively stained cells, with MEFs as a negative control (n = two independent biological replicates, two technical replicates each). (D) UMAP plot of our integrated 2D and 3D single-cell datasets with classified cell types labeled (Two independent biological replicates: n = 9,348 and 4,699 cells). (E) Discrete and hybrid cell type composition of iEPs in 2D and 3D cultures. (*: P <= 0.05,

randomization test). (F) Percentage of 2D- and 3D-cultured iEPs expressing Epcam, Cyr61 ($P <= 0.0001$; Two sample Chi-squared test). (G) Expression of other BEC markers on the UMAP plot, including Krt19, Sox9, and Cftr. (H) Module scores comparing identified injured BECs, MEF/Stromal cells (LT-iEP Dataset), and primary BECs (Pepe-Mooney et al., 2019). BEC markers used for module scores are identified from Verhulst et al., 2019.

2.8 Acknowledgement

We thank members of the Morris laboratory for their helpful discussions. We thank Barbara Treutlein for sharing the original QP code and Dennis Oakley of the Washington University Center for Cellular Imaging (WUCCI) supported by Washington University School of Medicine, The Children's Discovery Institute (CDI-CORE-2015-505 and CDI-CORE-2019-813) and the Foundation for Barnes-Jewish Hospital (3770 and 4642). We appreciate the helpful discussion with Lee Grimes and Nathan Salomonis on monocyte-neutrophil hybrid identity. Thank you also to Colin. This work was funded by National Institute of General Medical Sciences R01 GM126112, and Silicon Valley Community Foundation, Chan Zuckerberg Initiative Grant HCA2-A-1708-02799, both to S.A.M.; S.A.M. is supported by an Allen Distinguished Investigator Award (through the Paul G. Allen Frontiers Group), a Vallee Scholar Award, a Sloan Research Fellowship, and a New York Stem Cell Foundation Robertson Investigator Award; W.K. is supported by a Douglas Covey Fellowship; E.M.H. is supported by NIH/NHLBI T32 HL007317-44.

2.9 Author Contribution

W.K. and S.A.M. conceived the research. W.K. led computational work, assisted by Y.C.F. and X.Y., and supervised by S.A.M. W.K. led experimental work, assisted by E.M.H., X.Y., and supervised by S.A.M. G.G. and E.O.M generated and interpreted the motor neuron programming datasets. All authors participated in data interpretation and manuscript writing.

Chapter 3: Constructing a Comprehensive Lineage Map of Direct Cardiac Reprogramming

Adapted from:

Single-cell analysis of clonal dynamics in direct lineage reprogramming: a combinatorial indexing method for lineage tracing

Biddy B.A., Kong W., Kamimoto K., Guo C., Waye S.E., Sun T., Morris S.A. (2018). Single-cell analysis of clonal dynamics in direct lineage reprogramming: a combinatorial indexing method for lineage tracing. *Nature*.

CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics

Guo C., Kong W., Kamimoto K., Rivera-Gonzalez G.C., Yang X., Kirita Y. Morris S.A., (2019) CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biology*.

CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution

Kong W., Biddy B.A., Kamimoto K., Amrute J.M., Butka E.G., Morris S.A., (2020) CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nature Protocols*.

3.1 Abstract

The Waddington landscape depicts development as a unidirectional descent from a pluripotent state, down valleys toward defined cell identities. Lineage reprogramming, such as direct cardiac conversion, crosses the barrier between valleys, bypassing progenitor or stem cell states. These cell fate engineering strategies hold much promise for generation of clinically valuable cell types from mature somatic cells. However, current reprogramming protocols are inefficient, marked by low conversion rates and failure to fully recapitulate the properties of target cells. Insights into the mechanisms of reprogramming and trajectories leading to undesired, off-target identities will improve the efficiency and fidelity of this process.

Cardiac reprogramming from non-myocyte donor cells to cardiomyocytes via exogenous expression of transcription factors (TFs), provides a promising approach to repair damaged heart tissue. Despite recent advances in cardiac reprogramming, its inefficiency and low fidelity hinders the clinical use of the resulting engineered cells. Application of single-cell technologies together with epigenetic analyses has provided valuable insight into the reprogramming process. However, properties such as reprogramming trajectory – i.e., the transition of cells toward target or off-target identities, are often inferred from gene expression alone thus key information on the progression of cell fate conversion remains unknown. This is an important area of study, as lineage can provide answers to key questions in cardiac reprogramming, such as: 1) what molecular changes accompany transition states on the trajectory? 2) Are conversion trajectories stochastically or deterministically determined? Here, we apply methodologies in the Morris lab to resolve lineage relationships using CellTagging to find key regulatory transcription factors with CellOracle and evaluate the effect of small molecules on patterning biases using Capybara.

3.2 Introduction

Heart disease is a leading cause of death worldwide. One of the key pathologies contributing to cardiovascular disease is the loss of functional cardiomyocytes, whose post-mitotic nature limits regenerative potential (Ieda et al., 2010; Qian et al., 2012; Song et al., 2012). Induced pluripotent stem cell (iPSC) reprogramming from somatic cells opened broad opportunities in direct differentiation of mature cardiomyocytes from a stem-cell like state (Takahashi & Yamanaka, 2006). Yet, the inefficient conversion to iPSCs along with limited differentiation to cardiomyocytes produces a heterogeneous population, hindering the clinical utility of these cells (Ieda et al., 2010). Recent advances in direct lineage reprogramming from somatic cells enable the generation of cardiomyocyte-like cells from fibroblasts with overexpression of three key transcription factors - Gata4, Mef2c and Tbx5 (GMT) (Ieda et al., 2010; Qian et al., 2013; Song et al., 2012; Srivastava & Ieda, 2012). As a major building block of the heart making up of 50% of composition, cardiac fibroblasts have promise as an endogenous source of cells for reprogramming in vivo (Ieda et al., 2010; Qian et al., 2012). Though there have been continuous improvements in cardiac reprogramming protocols, the underlying molecular mechanisms driving this process remain unclear. Recent studies have elucidated important features of this cell fate conversion (de Soysa et al., 2019; Liu et al., 2017; Stone et al., 2019; Y. Zhou et al., 2019). Single-cell transcriptomics has revealed that cells during early conversion transit through a bifurcation, with one route leading to reprogramming while the other becomes refractory to cardiac conversion (Y. Zhou et al., 2019). In support of this, in joint epigenetic and single-cell analysis, transition cells are found to determine their terminal fate in the first 24-48hr and take one of two paths, either to a cardiac destination or to a non-cardiac

identity resembling fibroblast or vascular developmental cell states (Stone et al., 2019). Analysis of chromatin landscape changes illustrates enhancers activated at early transition resemble mostly neonatal cardiac development, gradually moving toward postnatal and adult stages. This demonstrates that early reprogrammed cells adopt an immature identity, mimicking embryonic development, before specification and activation of mature cardiac enhancers toward an adult fate (Hashimoto et al., 2019).

Most current studies rely on enrichment of reprogrammed cells using an α-Myosin Heavy Chain (α-MHC) reporter gene (Gulick et al., 1991), building a restricted lineage rather than a comprehensive map, where off-target identities and interactions between non-myocyte and reprogrammed cells may reside. Moreover, current trajectories leading to reprogrammed and off-target identities are inferred from gene expression. Therefore, the molecular mechanisms underlying these trajectories are also inferred. A "ground truth" map of lineage, connecting cell ancestry via heritable cell labeling, has not been constructed for cardiac reprogramming, representing a critical gap in knowledge for direct cardiac conversion. Unraveling this comprehensive lineage map will help us gauge in vivo delivery of transcription factors or small molecules as putative drugs and to understand potential obstacles to generating functional cardiac tissue. As aforementioned, a key driver of cardiovascular disease is the loss or defect in functional cardiomyocytes, which could potentially be rescued via *in vivo* regeneration.

To construct the "ground-truth" lineage map of this reprogramming process, we leverage CellTagging (Biddy et al., 2018; Kong et al., 2020), joint with single-cell sequencing across different time points to establish trajectories. Further, we apply nascent computational tools, including CellOracle (Kamimoto et al., 2020) and Capybara, developed in the Morris lab to identify potential transcription factors that could potentially promote or block successful

reprogramming outcomes. Lastly, via joint profiling of lineage, transcriptome, and chromatin accessibility, we attempt to characterize and compare the origins of successfully converted cells and learn the diverse outcomes from clonally related siblings during direct cardiac conversion.

3.3 Results

3.3.1 CellTagging: Simultaneous Capture of Lineage and Transcriptomics

The rapid advancement of single-cell technologies opens broad opportunities in uncovering heterogeneity and mechanisms at a single-cell resolution. Yet, during a single-cell experiment, lineage relationships between cells are destroyed, hindering the mechanistic understanding of continuous biological processes, such as reprogramming. To reveal such relationships, we developed CellTagging, a combinatorial cell barcoding approach, enabling

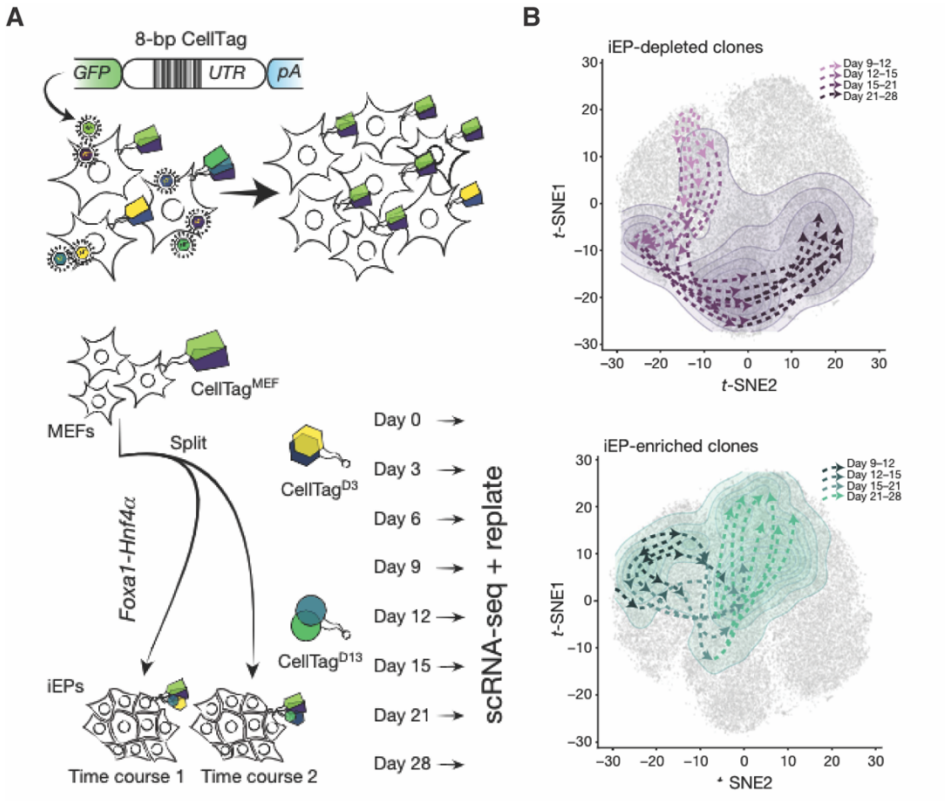

Figure Error! No text of specified style in document..**19:** Overview of the CellTagging System and Previous Discovery in MEF to iEP Reprogramming.

parallel capture of clonal history and transcriptional changes with single-cell RNA sequencing.

In brief, CellTagging labels the cells with an 8-nt heritable lentiviral-delivered barcode, where sequencing cell barcoding during biological processes allows us to construct a multilevel lineage tree **(Figure 3.1A**). Each delivery uses a distinct CellTag library that differs from each other based on a short motif right before the random CellTag region. This design allows us to demultiplex the cells after sequencing. We applied CellTagging to the direct reprogramming of mouse embryonic fibroblast to induced endoderm progenitors (iEPs). As this epithelial-like cell population has been shown to have both intestinal and hepatic potentials (Morris et al., 2014; Sekiya & Suzuki, 2011), it serves as a prototypical example of direct reprogramming, representing low efficiency and fidelity (Cahan et al., 2014; Guo & Morris, 2017; Morris et al., 2014).

In this experiment, initial CellTag library $CellTag^{MEF}$ was transduced into MEFs. After expansion for two days post transduction, the MEFs were split into two replicates for independent reprogramming via overexpression of transcription factors, *Hnf4* and *Foxa1* according to previously reported protocols (Morris et al., 2014; Sekiya & Suzuki, 2011). At the end of 60-hour of transcription factor delivery, $CellTag^{D3}$, a second barcode library, was delivered with a third round of CellTag delivery, $CellTag^{D13}$, on day 13. During the time course of this reprogramming, a portion of cells was collected and fixed in methanol every three days in early reprogramming days (before d15) and every 7 days during the later days (day 21 and day 28), while the remaining cells were replated for continued expansion and reprogramming. Single-cell RNA-seq was employed on all collected and fixed cells at the end of the experiments, following the 10x Genomics protocol.

In total, we recovered over 100k cells during this direct reprogramming protocol with CellTag barcoding information and transcriptomic profiles. The CellTag information was extracted, binarized, filtered, and processed for final computation of clonal or lineage relationships. Using randomized testing with this lineage information, we identified bifurcation trajectories with two terminals – one to successful conversion, expressing *Apoa1* and *Cdh1,* and the other to a "dead-end", where fibroblast gene expression is retained **(Figure 3.1B).** In addition, when investigating early CellTag labels, we found that clonally related cells from the same fibroblast ancestor populations, tend to follow the same trajectory of reprogramming, revealing early initiation of deterministic commitment to these trajectories.

Construction of distinct trajectories led us to explore potential regulators that can enhance reprogramming, where we found that expression of a putative methyltransferase, *Mettl7a1*, was identified to associate with successful reprogramming. Previously, it has been reported that METTL3, which catalyzes N^6-methyladenosine (m^6A) modification of mRNA, plays a key role in stem-cell differentiation and reprogramming to pluripotency (Batista et al., 2014; T. Chen et al., 2015). We performed MEF to iEP reprogramming experiment with the addition of *Mettl7a1* into the cocktail and subjected the cells on day 14 to colony formation assays and single-cell RNA-sequencing. Colony formation assays demonstrate a two-fold increase in E-cadherin-positive colonies by day 14. Analysis of single-cell RNA-sequencing revealed higher expression of *Apoa1* and higher similarity to the successfully reprogrammed cells identified in the previous time course experiment (Biddy et al., 2018).

To enable more streamlined discoveries using CellTagging, we developed a lineage reconstruction pipeline, CellTagR (https://github.com/morris-lab), to analyze CellTag data (Biddy et al., 2018; Kong et al., 2020a). In brief, CellTagR supports two different parts to

analyze CellTag. First, post generation of CellTag library, its complexity needs be accessed via sequencing to further produce an allow list for downstream design of the experiment and analysis of single-cell RNA-sequencing data. CellTags were extracted from FASTQ files, counted, and sorted to provide an allow list (**Figure 3.2A)**. The allow list provides instruction for the design of the experiment to ensure the complexity of the single-cell CellTag library. Then, after generation of the single-cell RNA sequencing data, CellTags were extracted from filtered alignment files and further subject to UMI quantification and count matrix generation. The data was then filtered, binarized, corrected for errors, filtered with the allow list and metrics, and lastly used for clone calling and lineage construction (**Figure 3.2B)**.

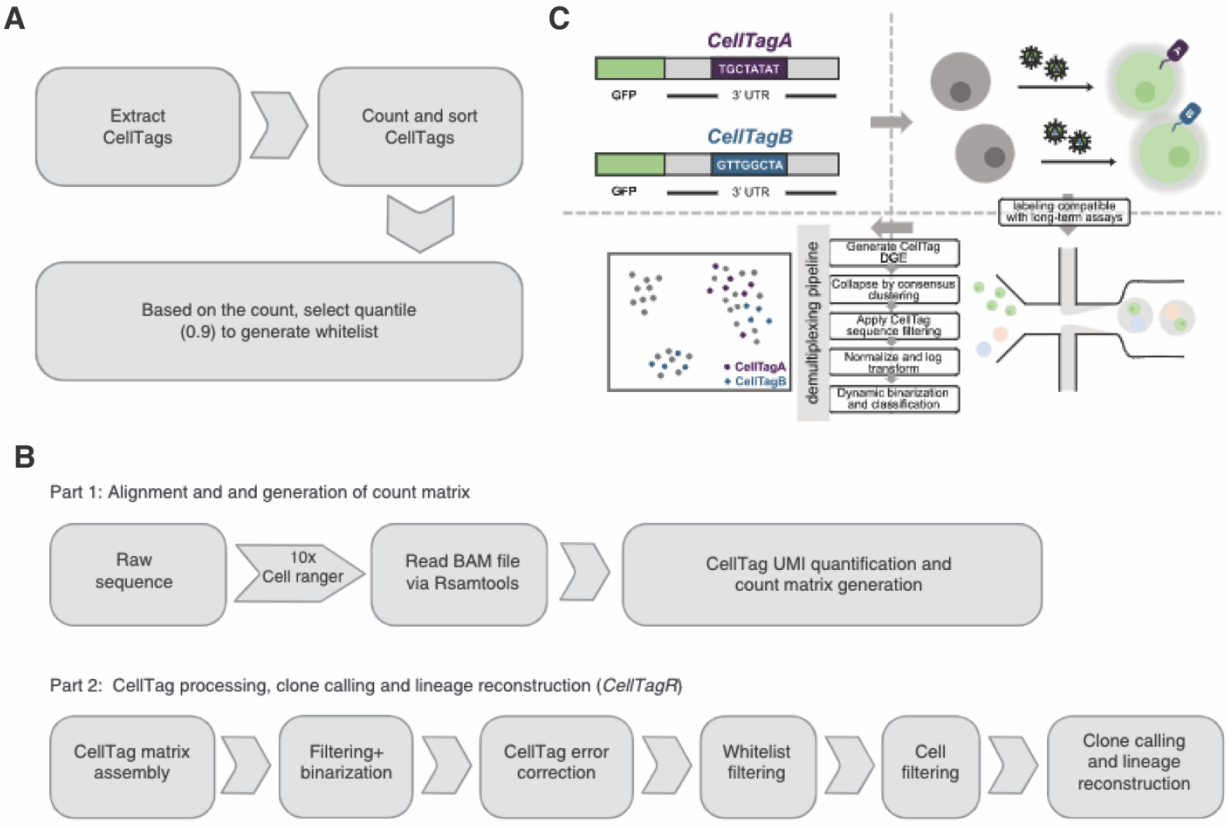

Figure Error! No text of specified style in document.**.20:** Overview of CellTagR pipeline and CellTag Indexing Strategy.

In an effort to alleviate batch effects introduced by differing sample preparation, CellTagging has been adapted for multiplexing of samples for high-throughput single-cell RNA-sequencing. In brief, different CellTag barcodes can be delivered to distinct populations, enabling the separation via barcodes in the downstream analysis post-sequencing. In this manner, samples can be pooled for preparation and single-cell RNA-sequencing. CellTag information was extracted similarly as described above. With such information, we employed similar classification strategy as Capybara in the downstream analysis to classify cells into different groups based on their CellTag profile (**Figure 3.2C**). Using genetically distinct populations, CellTag Index-based multiplexing was shown to be accurate and efficient. In addition, as CellTags are heritable and lentivirus-delivered, CellTag Indexing demonstrates efficacy for long-term live cell multiplexing, supporting cell tracking in a competitive transplant assay *in vivo* (Guo et al., 2019).

Together, CellTagging has enabled us to build a quantitative map of direct iEP reprogramming and empowered us to identify key regulators to enhance reprogramming efficiency and fidelity. Next, we sought to deploy this system in the direct cardiac reprogramming system to build a comprehensive lineage map during this process.

3.3.2 Successful recapitulation of direct lineage reprogramming with CellTag delivery

To deploy CellTagging in cardiomyocyte reprogramming, we started to establish the reprogramming protocol in the lab, following previous established protocols (Qian et al., 2013). Cardiac fibroblasts were harvested from neonatal mouse hearts using an explant protocol (**Methods**). The cells attached and expanded for 7 days prior to the reprogramming experiment. We first sought to test via immunostaining to ensure successful recapitulation of this

reprogramming protocol in the lab. Transcription factors (TFs), Mef2c, Gata4, and Tbx5 (MGT), were retrovirally delivered into the cells, followed by the reprogramming period of 28 days with media change every two to three days. On day 28, we fixed the cells and performed immunostaining labeling cardiac troponin, a regulatory protein in cardiac muscle cells (Sharma et al., 2004). Utilizing confocal microscopy, we identified positive cells only in samples infected with the TFs (**Figure 3.3**), supporting successful conversion of cells from fibroblast to cardiomyocytes.

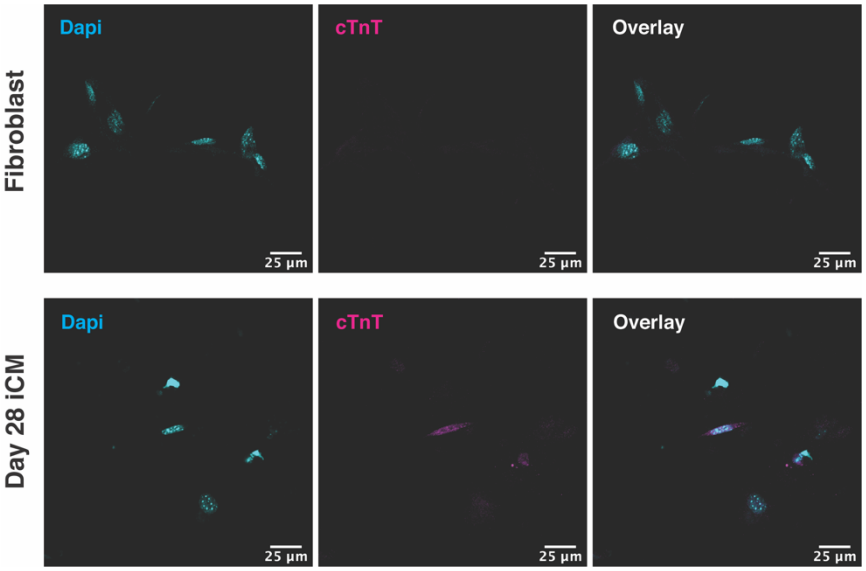

Figure Error! No text of specified style in document..**21:** Immunostaining of Cardiac Troponin Protein.

Next, we applied a similar CellTagging experimental scheme as aforementioned to the conversion of cardiomyocytes. Three days prior to introduction of TFs, the cells were stained and replated from the explant culture, followed by infection of CellTag V1 library on the next day (day -2). After expansion for 48-hr post transduction, the reprogramming cocktail was introduced according to previously reported protocols (Qian et al., 2013). CellTag V2, the second barcode library, was delivered on day 10 with a third round of CellTag delivery, CellTag V3, on day 21. During the process, the cells were collected, and methanol fixed at day 0, 13 and 27, to perform a pilot single-cell experiment using the 10x Genomics platform (**Figure 3.5A**) (Zheng et al.,

2017), where we obtained a total of 3,418 cells (Day0: 1,517; Day 13: 618; Day 27: 1,283) after quality metric filtering. Preliminary clustering analysis using Seurat reveals 14 clusters in total (**Figure 3.4A**). We identified fibroblast markers, *Col1a2* and *Tbx20*, as well as cardiomyocyte

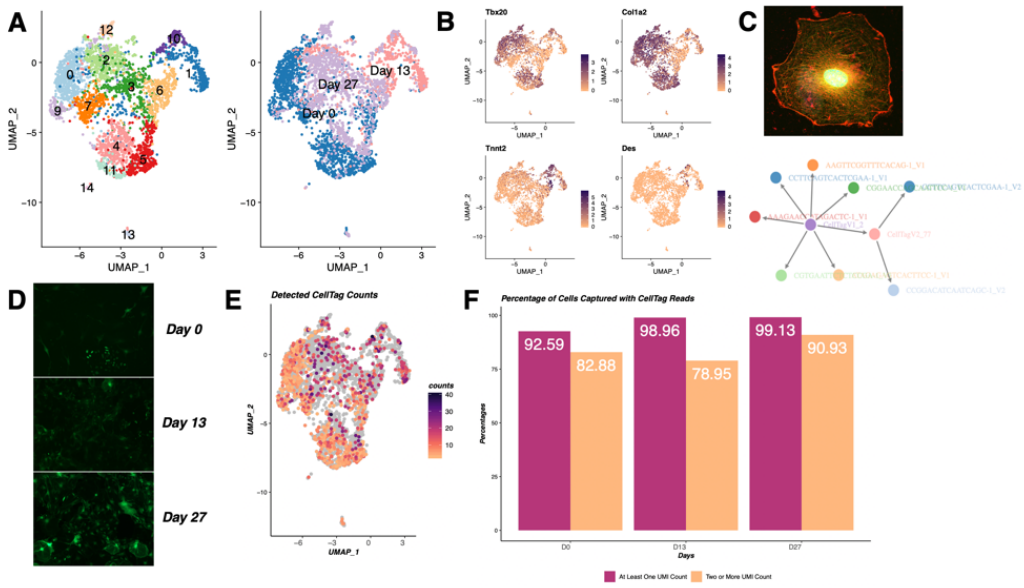

Figure Error! No text of specified style in document.**.22:** Preliminary single cell RNA-sequencing analysis.

markers, *Tnnt2* and *Des*, in distinct clusters (**Figure 3.4B**). Interestingly, we observe most of the cells expressing cardiac gene signatures occur on day 13 with decreasing percentages on day 27, reflecting a possible refractory path as previously reported in Zhou *et al*. Additional immunofluorescence using sarcomeric α-actinin antibody of these cells shows striation and sarcomere structure (**Figure 3.4C**, Top). Moreover, we evaluate CellTag expression by identification of GFP+ cells in culture (**Figure 3.4D**). Along with imaging, we performed initial lineage analysis, demonstrating our ability to capture CellTag expression in this reprogramming process with average of over 95% of cells captured to have at least one UMI count and over 78% with 2 or more counts (**Figure 3.4E, F**). This further enabled successful clone and lineage detection in the system (**Figure 3.4C, Bottom**).

3.3.3 CellTag lineage analysis mainly reveals trajectory toward off-target cell types

Next, we sought to further explore the lineage of this reprogramming process leveraging this preliminary dataset. As this dataset captures a time course of continuous reprogramming, we turned to use partition-based graph abstraction (PAGA; (Wolf et al., 2019)), supporting a better-connected embedding. Similar to Seurat outcomes, PAGA identifies a total of 14 clusters with cells from distinct time points distributed across clusters (**Figure 3.5B**). To evaluate lineage outcomes, we first annotate clusters that represents successfully reprogrammed cardiomyocytes and other off-target cell types. As discussed in previous subsection, we identified distinct clusters expressing fibroblast signatures or cardiac signatures (**Figure 3.5C**). Moreover, using the reference established for Stone et al dataset in Chapter 1, we applied Capybara on this new dataset to distinguish clusters. Continuous scoring reveals enriched atrial cardiomyocyte identity scores in cluster 10 and enriched cardiac muscle cell scores in cluster 1, 2, and 12 (**Figure 3.5E**). Further cell-type classification identifies 20.4% hybrid cells with the remaining being classified as discrete populations. We further assigned cell-type labels to the clusters based on the dominant cell-type represented, leading to the identification of atrial cardiomyocytes in cluster 10 and cardiac muscle cells in cluster 1. Intriguingly, we also recognized a population labelled as brown adipose tissue, mirroring previous findings in Chapter 1 (**Figure 3.5F**).

With the identified target population, we performed randomized testing to assess major clones (more than 10 cells) that were significantly enriched for or depleted of induced cardiomyocytes (**Figure 3.6A**). Interestingly, the test uncovers zero enriched clones with seven depleted clones (p value < 0.05), in which less than 5% of cells coincide with the reprogrammed populations. The major depleted source population largely occupies cluster 0 and 3, representing

mainly cells from day 0, which infers a potential directionality from cluster 0 and 3 toward cluster 4 and 5. Indeed, leveraging diffusion pseudotime analysis, we found a low pseudotime profile in cluster 0 and 3 and relatively higher pseudotime in cluster 4 and 5, supporting more "differentiated" states (**Figure 3.6B-D)**.

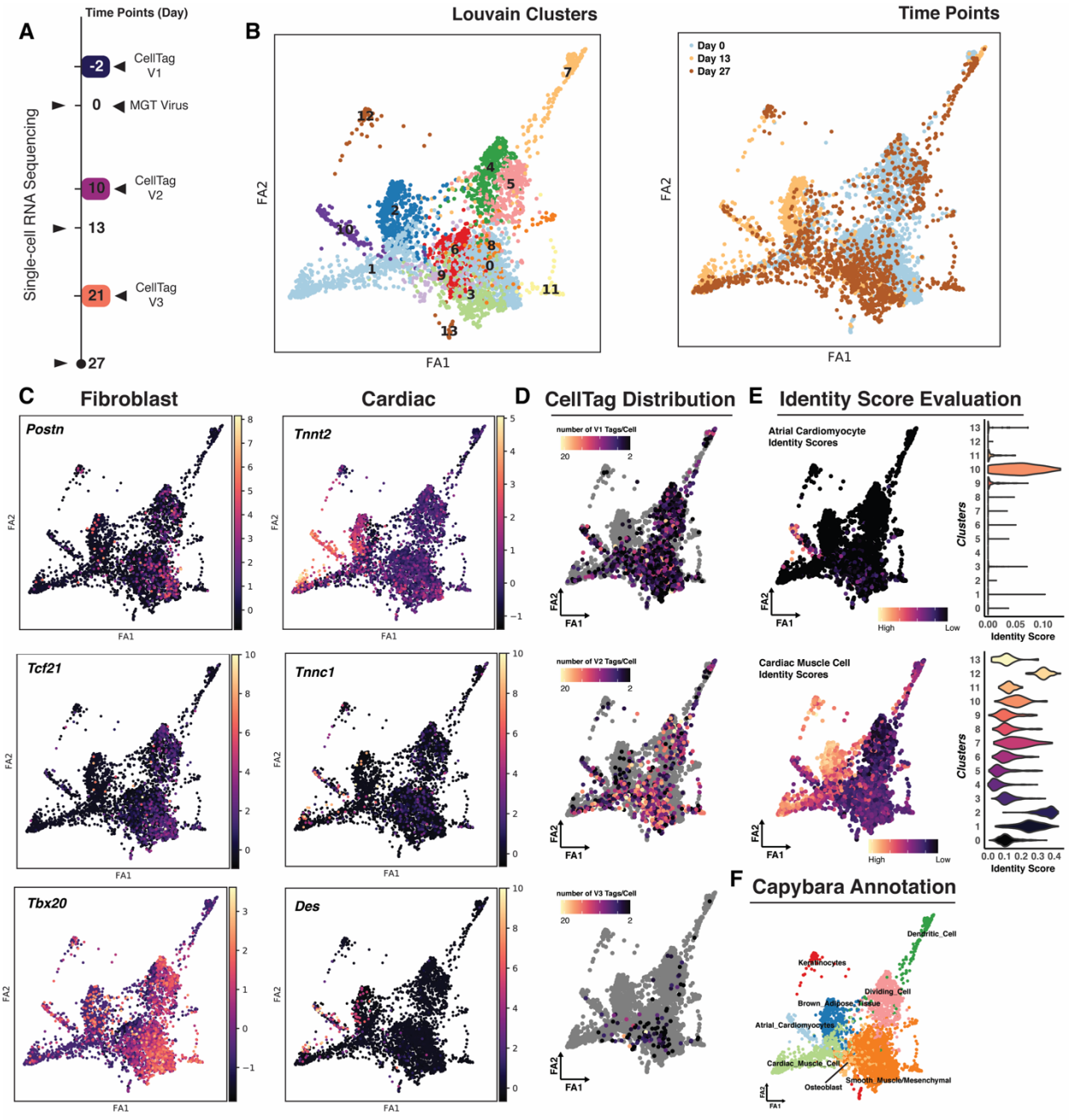

Figure Error! No text of specified style in document..**23:** Further Analysis of the Preliminary Data using SCANPY and PAGA.

Furthermore, we interrogated potential reasons why there is no identified enriched clones. Though large proportion of the cells received qualitied CellTags, we found that the cells associated with a clone were biased toward the off-target lineage, largely overlapping with the categorized iCM depleted clones (**Figure 3.6B-C**). Further, compared to day 0 and day 27, we

A

Clone ID	Enrichment P Value	Depletion P Value
52	0.676	0.093
55	0.46	0.376
56	0.335	0.442
63	0.702	0.163
83	1	0
84	1	0
86	0.868	0
87	0.531	0.224
90	0.919	0
91	0.883	0.045
92	0.998	0
100	0.885	0

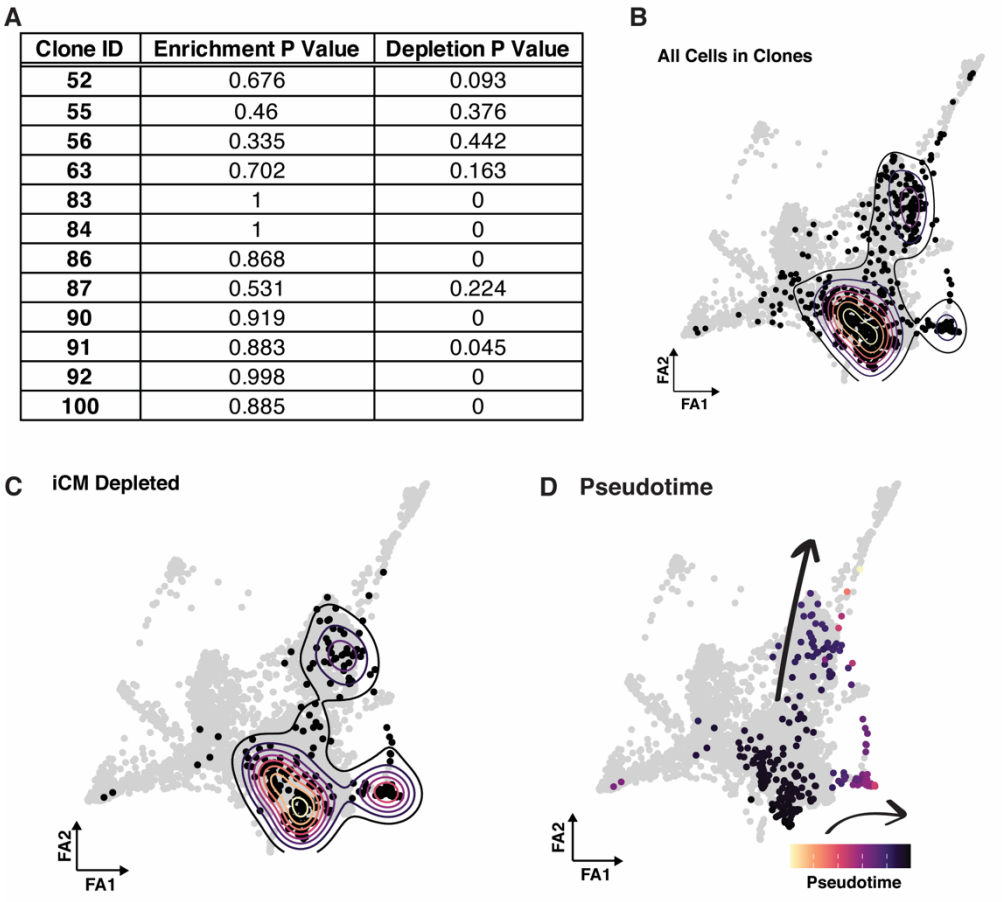

Figure Error! No text of specified style in document.**.24:** Identification of Enriched vs. Depleted clones.

captured at least 50% fewer cells on day 13, where major reprogrammed cardiomyocytes reside (contains 69.1% of atrial cardiomyocyte population). We speculate that this could be due to the post-mitotic nature of cardiomyocytes, where once reprogrammed, the cells would halt proliferation. As previously reported, the number of reprogrammed cardiomyocytes achieve its

peak by day 14 (Qian et al., 2013), explaining the potential capture of all reprogrammed cells at day 13 without their clonally related cells left in culture. This concern potentially can be addressed via restriction of the initial fibroblast population size and careful selection of collection time points, coupled with sequencing more cells to optimize clone detection. Alternatively, this could be alleviated via changing the reprogramming source cell type to an immortalized MEF-T cell line (Vaseghi et al., 2016), where cells do not exit cell cycle as much as primary fibroblasts once reprogrammed. Despite the concern of immaturity of the reprogrammed cells, this cell line could serve as a good model to dissect lineages and mechanisms in this reprogramming process.

In a nutshell, CellTagging enabled us to distinguish the trajectory to the off-target cell types. Though this experiment did not bring detailed information regarding the lineage, it supports the compatibility of CellTagging with the system and instructs high-level information regarding experimental design and time point of CellTag delivery. Future adaptation and adjustment of the experimental design could support better capture of the clonal dynamics and lineage relationships.

3.3.4 Probing cell-type dynamics with potential modulation leverage Capybara

Additional small molecules, such as TGFβ- and Wnt-inhibitors (Y. Fu et al., 2015; Mohamed et al., 2017; Zhao et al., 2015), have been demonstrated to improve the efficiency of reprogramming, relative to conversion with Gata4, Mef2c, and Tbx5 (MGT) alone. In Chapter 1, we utilized Capybara to probe the cell types during reprogramming with MGT and small molecules (SM), where we observe an atrial (76%) vs. ventricular (7.7%) bias in resulting cell

types. As TGFβ and Wnt signaling play key roles in chamber specification in early development of the heart, it has been previously shown that TGFβ inhibition and Wnt agonism yields mostly ventricular cardiomyocytes (H. Wang et al., 2014). We hypothesize that the TGFβ and Wnt inhibitors (SB431542, and XAV939, respectively) applied in the Stone et al., protocol could be potentially responsible for the generation of mostly atrial-like cardiomyocytes that we observed.

To test this hypothesis, we collected cells for scRNA-seq, without sorting, yielding a total of 11,215 cells from two biological replicates (+SM: 5,107; -SM: 6,108) (**Figure 3.7A**).

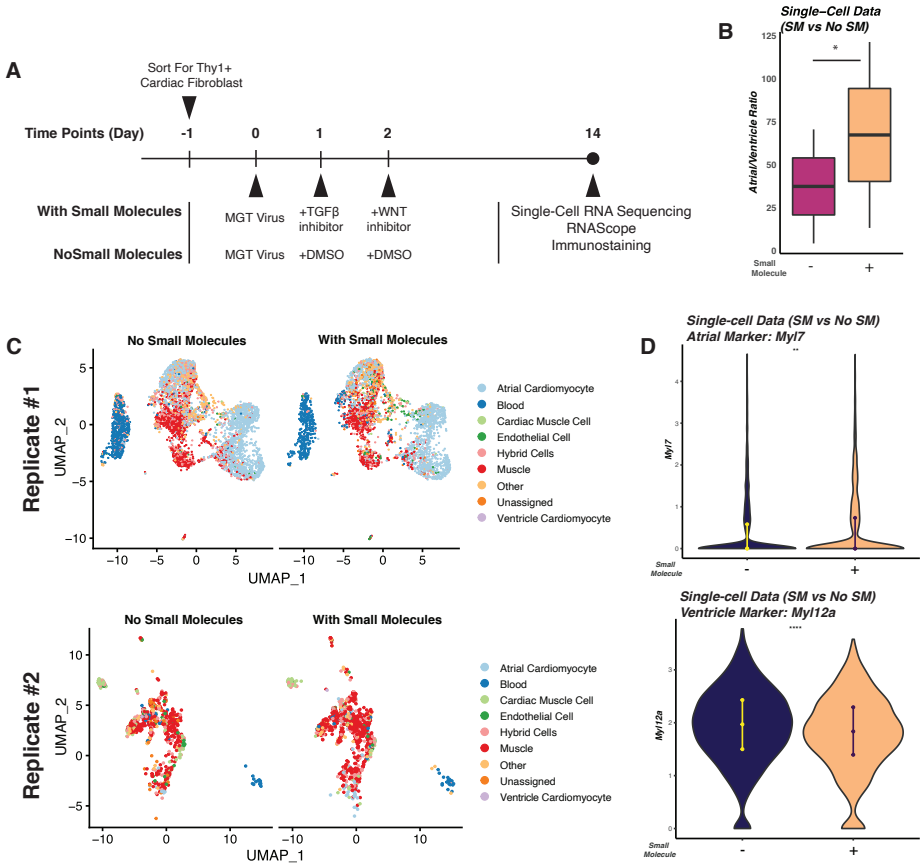

Figure Error! No text of specified style in document.**.25:** Capybara Analysis comparing Cardiac Reprogramming with or without small molecules.

Integration with days 7 and 14 of the Stone et al., data demonstrates successful recapitulation of the protocol (cosine similarity between Day 7 from Stone et al. and +SM samples = 0.71

(Replicate #1); 0.89 (Replicate #2)). Capybara analysis reveals that the addition of SM to the GMT cocktail produces relatively higher percentages of cardiomyocytes (+SM: 25.3 ± 1.06%, -SM: 24.0 ± 0.7%; P = 0.057, randomized test). Breaking down the atrial and ventricular cardiomyocytes reveals that the addition of SM produces a higher ratio of atrial to ventricular cardiomyocytes (P = 0.0287, randomization test; **Figure 3.7B-C**). Confirming this observation, we observed a significant increase in atrial genes, such as *Myl7* (P=0.021, Wilcoxon rank sum test), and decrease in ventricular genes, such as *Myl12a* (P=0.00061, Wilcoxon rank sum test), in the presence of SM (**Figure 3.7D**). Together, this offers support to the hypothesis that small molecules in reprogramming play a role in altering region specification.

In addition to small molecules, other factors, such as transcription factors (Akt1 and Hand2) (Hashimoto et al., 2019; H. Zhou et al., 2015) and microRNA (miR-133) (Jayawardena et al., 2012), have been demonstrated to improve the efficiency and fidelity of reprogramming. Next, we sought to look for potential transcriptional regulators that plays vital parts during this reprogramming process.

3.3.5 CellOracle reveals two transcription factors as key regulators during direct cardiac reprogramming

During reprogramming, transcriptional regulators play key roles in "wiring" cells to their destined cell fate. Identification of such factors can largely facilitate the derivation or differentiation toward their terminal fate. Traditionally, transcription factors were identified via differential gene expression analysis and screened via experimental approach (Ieda et al., 2010; Song et al., 2012; Takahashi & Yamanaka, 2006). Here, we turned to *in silico* approach, using CellOracle, to identify key transcriptional regulators and simulate potential outcomes from knockout or overexpression.

CellOracle reconstructs gene regulatory networks (GRNs) from scRNA-seq profiles. Integrated with chromatin accessibility information, this approach utilizes regularized linear regression to build statistical models for predicting gene expression, the strength of gene-gene interactions, inferring regulatory factors to maintain or reprogram cell identity. Beyond GRN prediction, it provides functionality to simulate expression profile shift once the expression of the transcript factor is altered (Kamimoto et al., 2020).

We first examined the time course reprogramming dataset generated in the preliminary experiment. With the data preprocessed, we constructed GRN models for each of the 14 clusters identified, representing cardiomyocyte fates (cluster 1 and 10) as well as other off-target cell types with the starting population labeled in cluster 0 and 3 (smooth muscle/mesenchymal). Ranking of the degree centrality of the top 30 TFs in cluster 1 and cluster 10 suggests a potential factor, *Klf5* (**Figure 3.8B**). KLF5 has previously been reported as significant in cardiovascular remodeling upon injury. During embryonic development, it is expressed abundantly, yet downregulated as development progress. Nonetheless, upon injury, *Klf5* expression is reinduced to activate downstream wound repair response and remodeling (Nagai et al., 2005). Together with pseudotime (**Figure 3.8A**) and CellOracle gradient vectors, we observe the inferred normal trajectory from cluster 0 and 3 toward different other branches (**Figure 3.8C**). We performed perturbation simulation of *Klf5* to evaluate the potential effect of it on this reprogramming system. The perturbation results reveal that upon knockout, the transitions to cardiac fate are blocked while transitions toward fibroblast fate is promoted (**Figure 3.8E**). On the contrary, overexpression simulation shows promotion toward the cardiomyocyte fate with blocked fibroblast transitions (**Figure 3.8F**), suggesting overexpression of *Klf5* in joint with the MGT cocktail could potentially improve this direct lineage reprogramming.

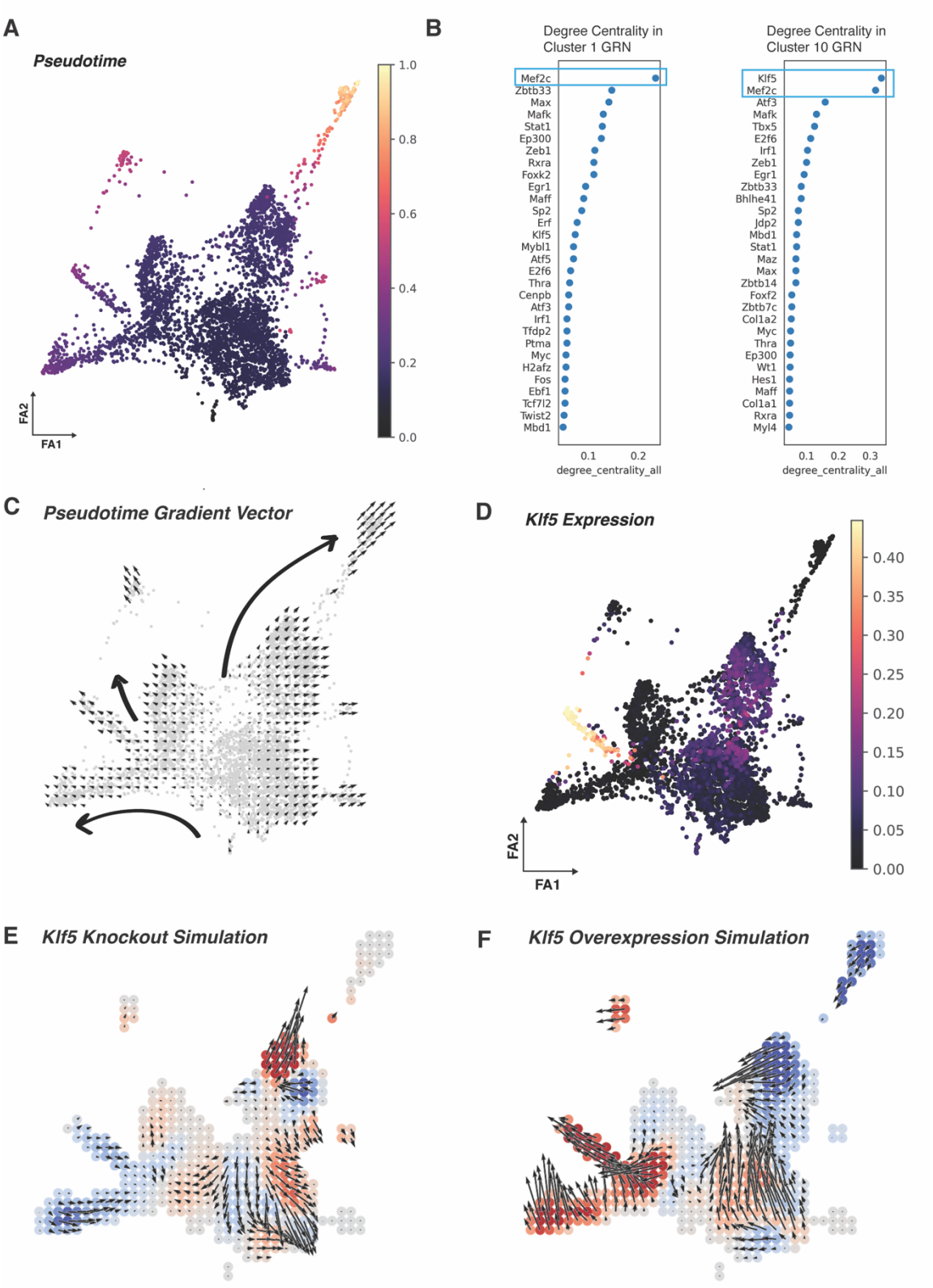

Figure Error! No text of specified style in document..**26:** CellOracle Analysis of the preliminary time course single-cell dataset.

Next, we sought to assess the reprogramming dataset that profiles the reprogramming process initiated with a modified cocktail (MGT + SM). We reanalyzed the data using PAGA (Wolf et al., 2019) to obtain a better-connected embedding, yielding a total of 13 clusters (**Figure 3.9**). Based on Capybara annotation, we labeled the clusters with the most dominant fate

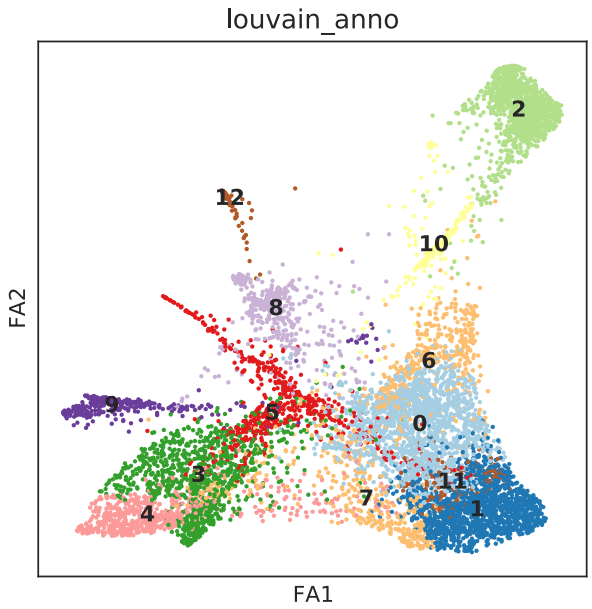

Figure Error! No text of specified style in document.**.27:** SCANPY and PAGA analysis of the dataset with small molecule treatment.

within each cluster and identified cluster 3, 4 and 10 as reprogrammed clusters. Using similar approaches as described above, ranking of the degree centrality of the top 30 TFs in cluster 3, 4, and 10 suggests potential factors, including *Klf5* and *Atf3* (**Figure 3.10B**). Previously, it has been reported that Atf3 is a key regulator for ventricular remodeling to protect the heart from hypertensive stress. The major source of Atf3 response resides in the cardiac fibroblasts, which is the source of reprogramming in this protocol (Y. Li et al., 2017). Similar to *Klf5,* pseudotime (**Figure 3.10A**) and CellOracle gradient vector inferred that the normal trajectory starts from cluster 1 toward different other branches (**Figure 3.10C**). Further perturbation simulation of *Atf3* reveals that upon knockout, the transitions to all cardiac fates are blocked while transitions toward the fibroblastic fate is promoted (**Figure 3.10E**). Interestingly, overexpression simulation shows promotion toward one subpopulation of cardiomyocyte fates (cluster 4) yet with blocked

fibroblast and other cardiac fates (cluster 3) transitions (**Figure 3.10F**), suggesting *Atf3* as a potentially more specific driver factor in determination of fates.

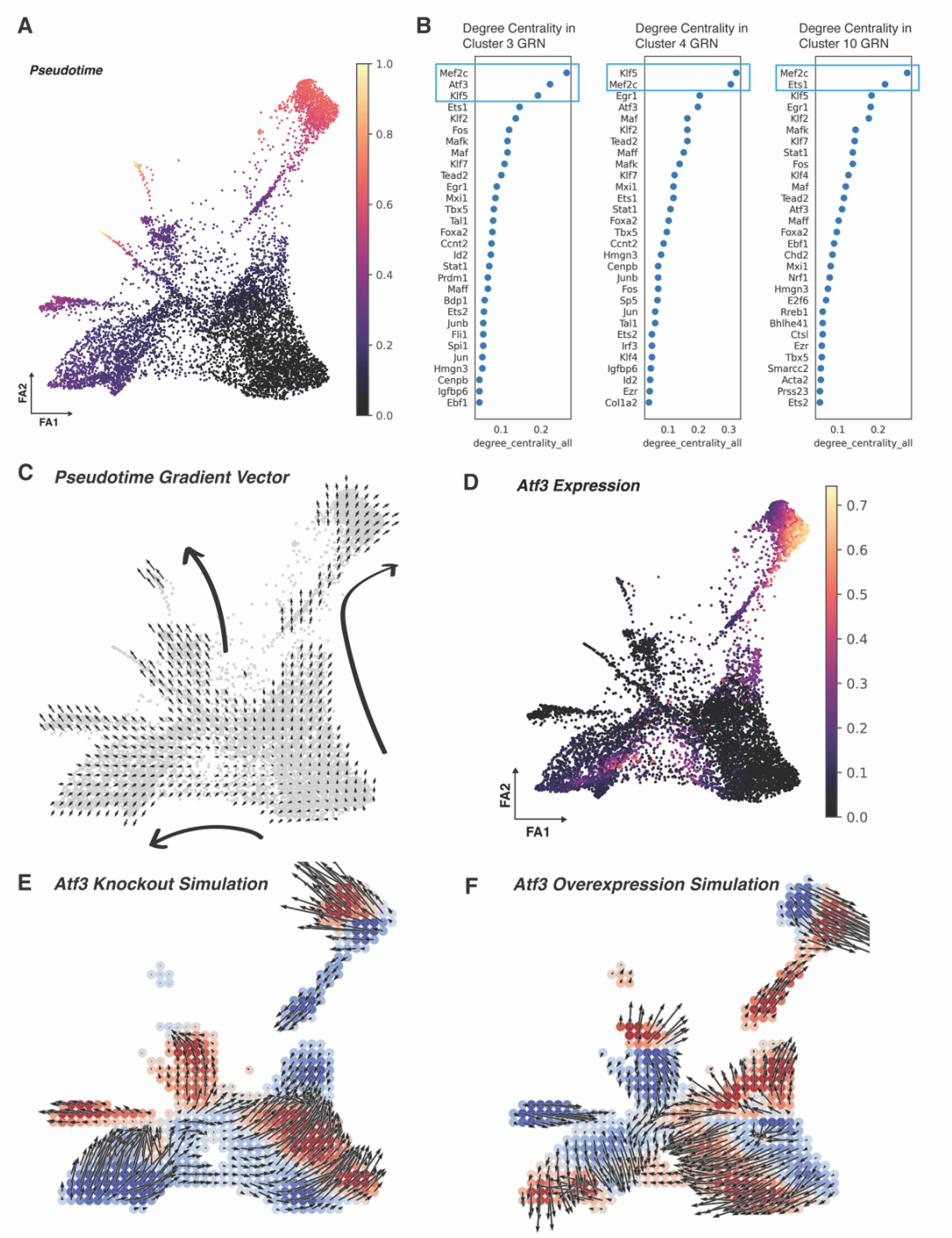

Figure Error! No text of specified style in document.**.28:** CellOracle Analysis of the dataset with small molecules.

Overall, leveraging GRN analysis, I have pinpointed putative transcriptional regulators, overexpression of which could promote the efficiency and fidelity of reprogramming. Intriguingly, both *Klf5* and *Atf3* have been found to serve important roles in cardiovascular remodeling when exposed to injury or stress, suggesting potential similarity between *in vitro* reprogramming and *in vivo* cardiac repair process.

3.3.6 Lineage tracing with immortalized MEF-T cell line uncovers two overlapping putative trajectories in transcriptional profiles

We route back to the big-picture scope of this project to construct a comprehensive lineage map of cardiac reprogramming. To alleviate the previously identified concerns, we moved on to use the inducible MEF-T cell line. In brief, this cell line is constructed from primary cells harvested from transgenic mice, which harbors the reporter under the α-MHC promoter as aforementioned (Gulick et al., 1991). The three transcription factors, MGT, are integrated into the genome in a polycistronic manner and regulated under a Tetracycline inducible promoter. Once doxycycline is introduced, MGT TFs are overexpressed, initiating reprogramming. As the cells initiate cardiac program, they start to express the GFP reporter protein, serving as a benchmark for successful reprogramming. SV40 large T antigen was retrovirally delivered into the primary cells, followed by selection, to establish the immortalized cell line (Vaseghi et al., 2016). As these cells are more prolific than primary MEFs and the reprogramming outcomes are less mature (Stone et al., 2019; Vaseghi et al., 2016), we consider it could help alleviate the concerns as cells become non-proliferative. In addition, doxycycline induces a peak reprogramming within three days post-initiation (Vaseghi et al., 2016), enabling us to probe the initiation and reprogramming in a short time frame.

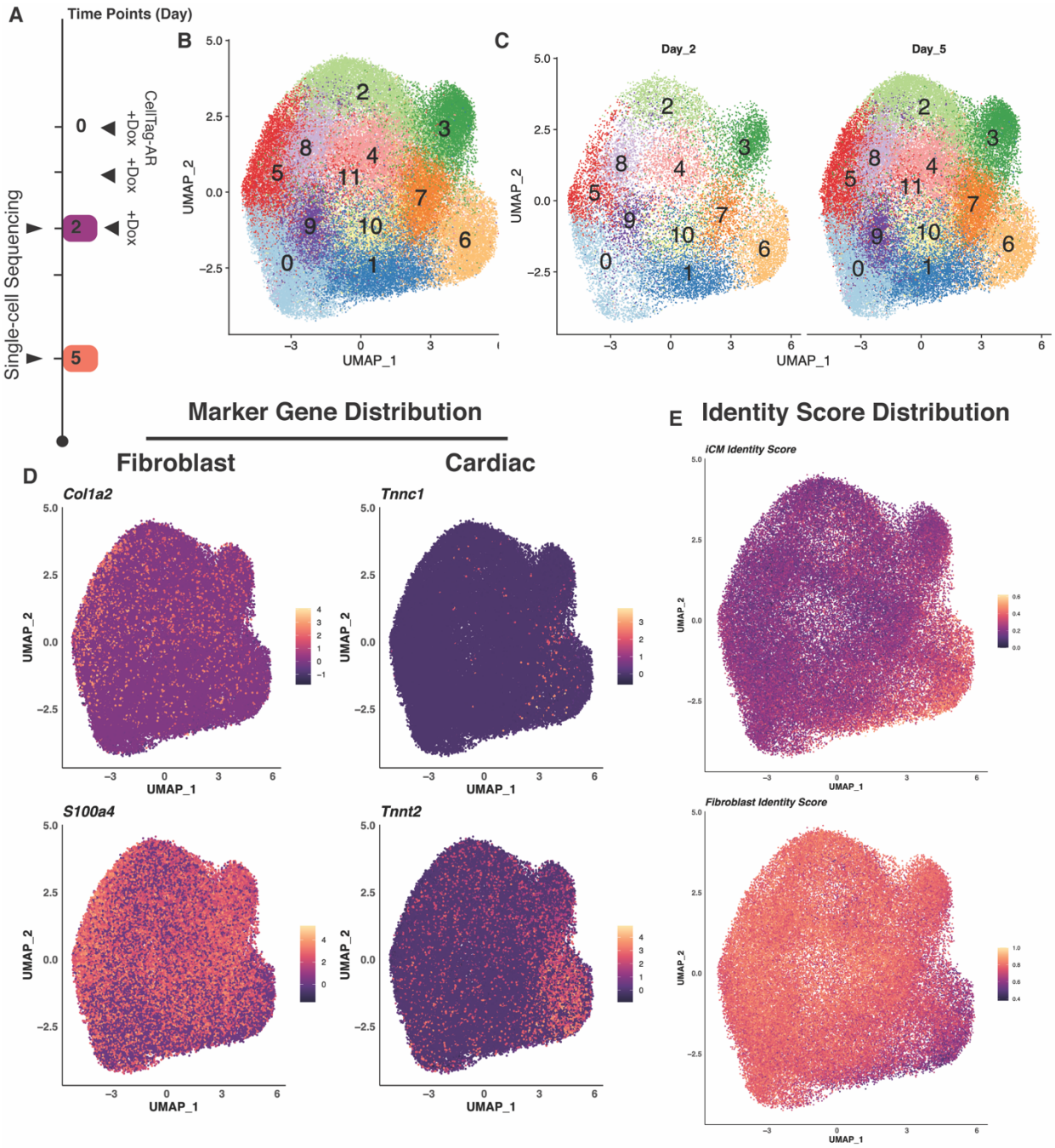

Figure Error! No text of specified style in document.**.29:** State-Fate Experiment (RNA Profile).

Inspired by the state-fate experiment performed in hematopoiesis (Weinreb et al., 2020*)*, we designed a similar experiment in the direct cardiac reprogramming, profiling the state

samples at day 2 and fate samples at day 5 post induction of doxycycline. Accompanying this design is the advancement in the CellTagging system (CellTag-AR), where the 8-nt random barcode is replaced with an 18-nt barcode, enabling higher library complexity and lower MOIs. In addition, CellTagging has been adapted to capture lineage in concurrent with single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) (Kunal Jindal, unpublished, manuscript in preparation). Here, ~10,000 cells were plated (0hr) and transduced with CellTag-AR library 12 hr later with doxycycline induction. The virus was removed 12 hours later, with replacement of fresh media containing doxycycline. Doxycycline was maintained and replaced every day in culture for ~3 days to initiate reprogramming, where the GFP reporter signal was observed by day 2 (data not shown). We collected and counted cells on day 2, two-thirds of which were prepared for single-cell sequencing with the remaining replated for continual expansion. The collected cells were further split for two assays, scRNA-seq and scATAC-seq. After an additional three-day expansion, we collected all the cells at day 5 and subjected them to similar protocols as day 2. Having the cells sharing the same CellTag-AR infection enables us to join RNA and ATAC data together by lineage (**Figure 3.11A**).

Post-sequencing, the RNA data was subject to CellRanger analysis to align, filter, and generate the count matrices. Through scRNA-seq, we obtained a total of 72,834 cells (Day 2: 18,702; Day 5: 54,132) post QC metrics through Seurat analysis. Application of integration with Seurat V4 reveals 12 clusters with even distribution of cells from day 2 and day 5 (**Figure 3.11B-C)**. These clusters dynamically express fibroblastic and cardiac marker genes (**Figure 3.11D**). To have a better idea of identity shifts, we constructed a reference based on the sorted day 14 (iCM) and initial day -1 populations (fibroblast) in Stone et al. Referencing against this dataset, we assigned identity scores to these cells via QP and found the iCM scores to be

enriched in cluster 6 while fibroblast scores to be evenly distributed across all other clusters (**Figure 3.11E)**. Based on the expression level of *Col1a2*, we speculate cluster 5 is the source of fibroblasts. Notably, the iCM scores peak around 0.6 compared to day 14 sorted population in Stone et al, suggesting immaturity in these reprogrammed cells.

Next, we turned our focus to CellTag clonal information. A total of 47,671 RNA and ATAC cells were captured as in clones with at least 2 cells in a clone. From RNA clonal information only, 31,111 RNA cells (Day 2: 7,823; Day 5: 23,288) were identified as in clones with overall even distributions across the clusters (**Figure 3.12A**). Here, following the interest in

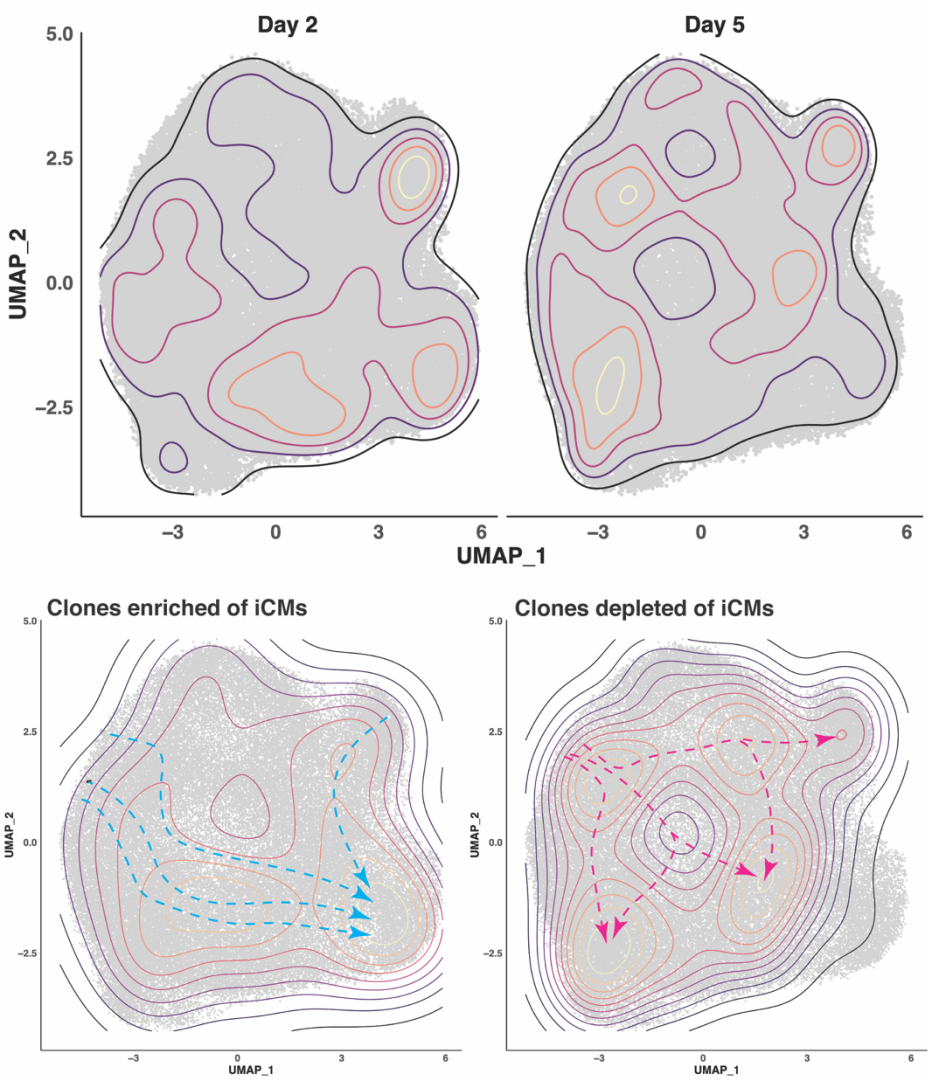

Figure Error! No text of specified style in document..**30:** Identification of Enriched or Depleted Clones based on Transcriptional Lineage.

looking at different time points, we mainly investigate the state-fate clones in this RNA dataset, defined as clones containing cells from day 2 and day 5. Within a total of 12,460 clones, we found 3,477 state-fate (SF) clones. With the identified target population (cluster 6), we performed randomized testing to assess major SF clones (more than 10 cells) that were significantly enriched for or depleted of induced cardiomyocytes, uncovering nine enriched clones with 25 depleted clones (p value < 0.1). 14~16% of the cells in the enriched clones are transcriptionally similar to the identified successfully reprogrammed cluster (cluster 6). The major depleted source population represents 0% of cells in the clone to become reprogrammed, largely locating in cluster 0 and 10 (**Figure 3.12B**). The dominant depleted clones reflect our pilot experiment with serial infection of CellTags. It is worth noting that there is a significant overlap of distribution of the cells identified in the reprogrammed enriched or depleted clones, suggesting similarity in expression profiles of these cells, which further support the immaturity of these reprogrammed cells

3.3.7 Lineage tracing with immortalized MEF-T cell line uncovers two distinctive putative trajectories in chromatin landscapes

Next, we sought to chart the dynamics on the chromatin landscape during this reprogramming protocol. In addition to transcriptional remodeling, direct lineage reprogramming involves significant shifts in chromatin landscape. For instance, in TF-mediated lineage reprogramming, pioneer transcription factors, who can open heterochromatin, alter the chromatin landscape with facilitation from additional co-factors. The chromatin remodeling then drives the transcriptional shift to the expression profile of the target cell types (Guo & Morris, 2017). In cardiac conversion, it has been found that the transcriptional remodeling involves a rapid gain of the cardiac program, yet with gradual loss of the fibroblast profile (Liu et al., 2017; Stone et al.,

2019; Y. Zhou et al., 2019). In effort to chart chromatin remodeling in cardiac reprogramming, a recent study leveraging scATAC-seq reveals key regulatory factors, including Fos and Tcf21, during early stages of the process, providing invaluable insights in epigenomic remodeling (H.

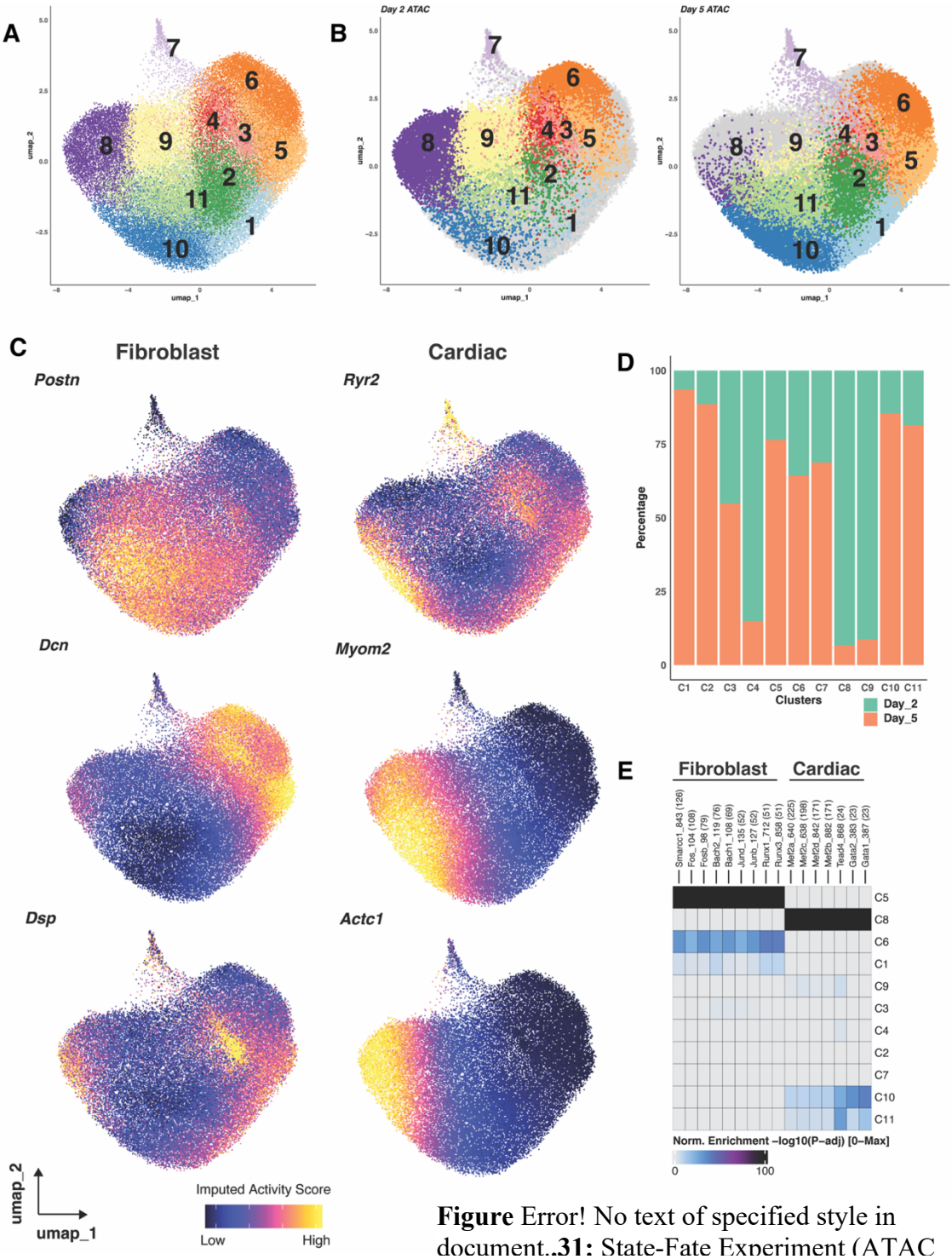

Figure Error! No text of specified style in document..**31:** State-Fate Experiment (ATAC

Wang et al., 2022).

After we obtained the sequencing data, the ATAC data was subject to CellRanger-atac analysis to align, filter, and generate the fragment counts. Through scATAC-seq, we obtained a total of 52,365 cells (Day 2: 20,476; Day 5: 31,889) post QC metrics through ArchR analysis (Granja et al., 2021). Clustering via ArchR reveals 11 clusters with distribution of cells from day 2 mainly occupies the top and those from day 5 lays on the bottom (**Figure 3.13A-B)**. We further compute imputed gene activity scores using ArchR, where we observe these clusters distinctively have high activity scores for fibroblastic and cardiac marker genes (**Figure 3.13C**). To better pinpoint the identities of each cluster, we further perform differential peak analysis, followed by motif enrichment analysis, where we found significant enrichment of previously reported fibroblastic motif, such as Fos and Bach2, in cluster 5 and 6. On the other hand, we found enrichment of cardiac motifs, such as Mef2c and Tead4, in cluster 8 and 10. To further confirm the cluster identities, we performed GREAT analysis, a tool that predict functions of cis-regulatory regions, to assign biological meanings to these regions (McLean et al., 2010), where we found cluster 5 and 6 to be significantly enriched for vascular and immune development. In addition to being fibroblastic, cluster 5 and 6 mirrors previous findings of vascular off-target during this reprogramming ((Stone et al., 2019) (**Figure 3.14A)**. Annotation of cluster 8 and 10 reveals significant enrichment of muscle and cardiac cell development as well as muscle fiber organization, supporting these two clusters as reprogrammed clusters (**Figure 3.14B**). Further comparison of these differential peaks to the binding profile of Mef2c, Gata4, and Tbx5 highlights cluster 8-11 as enriched for these binding sites (**Figure 3.14C)**.

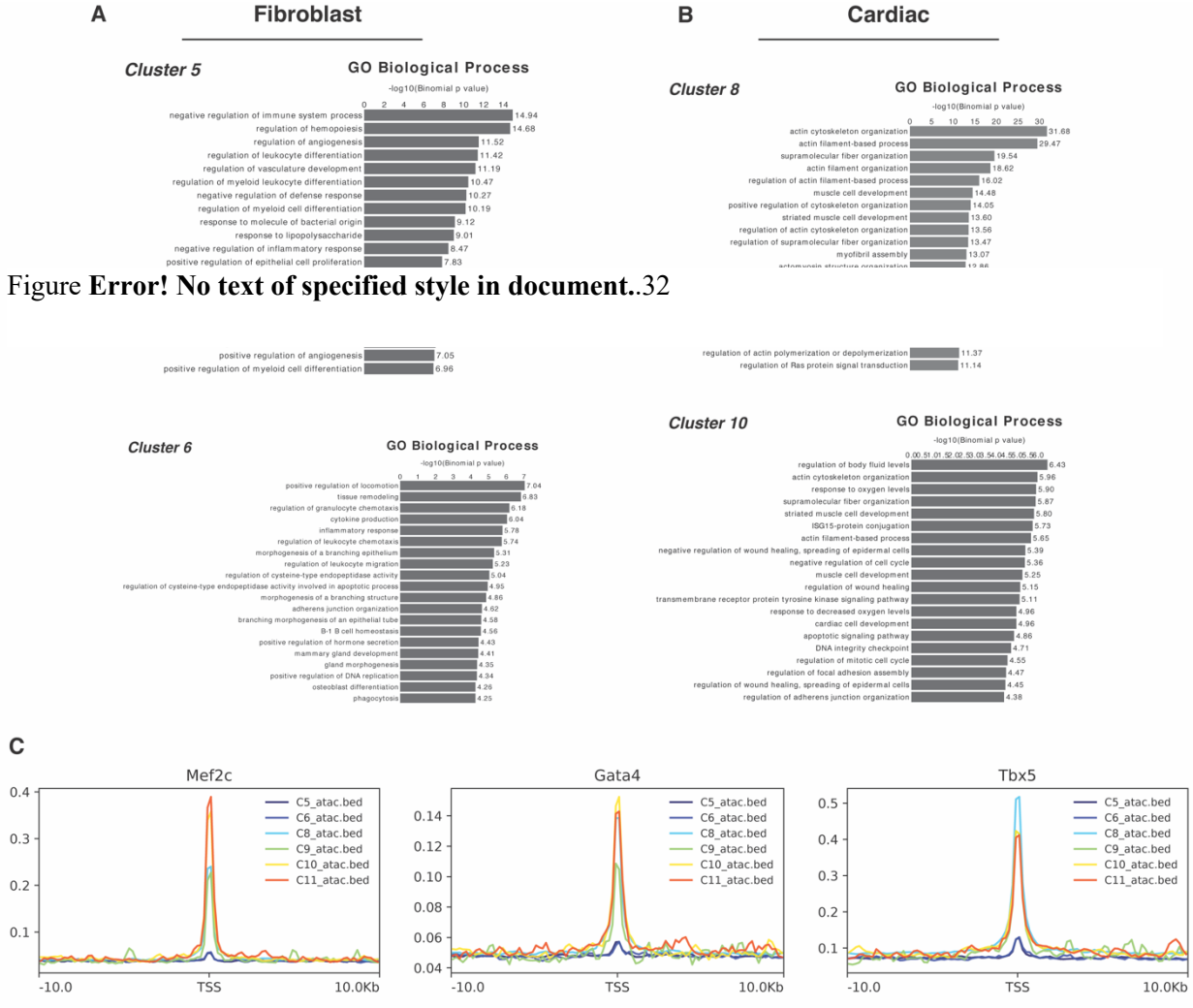

Figure **Error! No text of specified style in document.**.32

Figure Error! No text of specified style in document.**.14:** GREAT analysis and Comparison to Chip-seq Peaks.

Considering cluster 10 is composed mainly of day 5 fate cells, we consider this cluster as the cluster harboring successfully reprogrammed cells. We next performed similar randomized test using the ATAC clonal and lineage information. From ATAC clonal information only, 16,560 ATAC cells (Day 2: 5,157; Day 5: 11,403) were identified as in clones with overall even distributions across the clusters (**Figure 3.15A).** Again, we mainly investigate the state-fate clones in this dataset, where we found 1,305 state-fate (SF) clones out of 9,653 total clones. With the identified target population (cluster 10), we performed randomized testing to assess major SF clones (more than 10 cells) that were significantly enriched for or depleted of induced

cardiomyocytes, uncovering eight enriched clones with eight depleted clones (p value < 0.1).

30~50% of the cells in the enriched clones coincide with the identified successfully

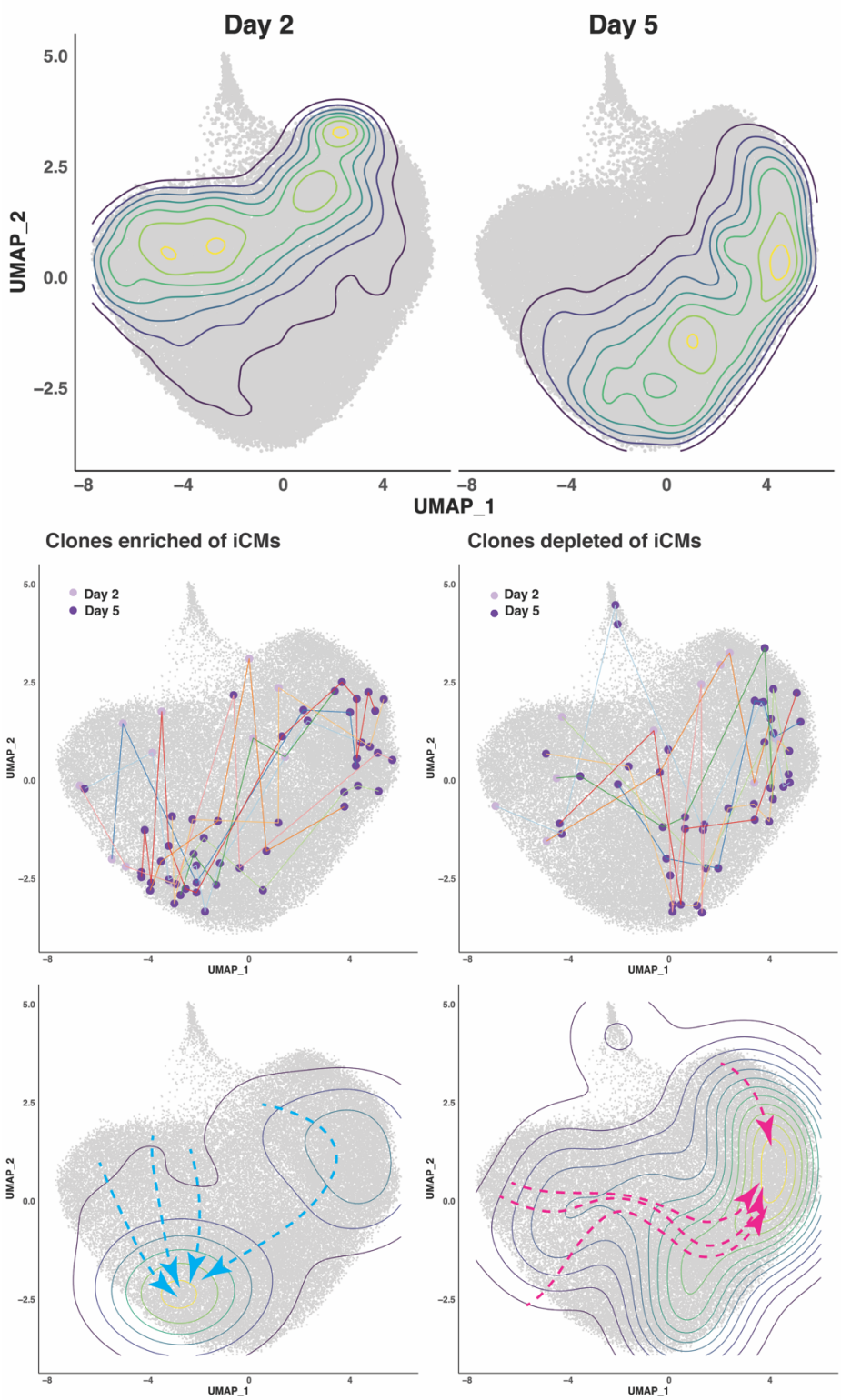

Figure Error! No text of specified style in document.**.33:** Identification of Enriched or Depleted Clones based on Chromatin Landscape.

reprogrammed cluster (cluster 10). The major depleted source population represents 0% of cells in the clone to become reprogrammed, largely locating in cluster 5 and 6 (**Figure 3.15B-C**). It is worth noting that, compared the RNA lineages identified, there is a more distinct distribution of the cells identified in the reprogrammed enriched or depleted clones, suggesting distinctions in their epigenetic landscape to define the two fates. Interestingly, in the depleted clones, we observe clones that remained within the region of cluster 5 and 6 across the two time points as well as clones that traversed cluster 8 to cluster 5/6. As mentioned above, cluster 8 demonstrates strong cardiac profile. This observation mirrors the previous identified refractory trajectory in this reprogramming protocol (Stone et al., 2019; Y. Zhou et al., 2019).

Taking RNA and ATAC together, we demonstrate the rich information presented in combination of lineage, transcriptome, and chromatin landscape. We speculate that the observed more distinct trajectories in the epigenome suggests that the chromatin remodeling has already occurred while the more overlapping trajectories in transcriptome profile reveals the delay in transcriptional remodeling.

3.4 Discussion

In this Chapter, we report the application of our in-house tools on a new direct lineage reprogramming protocol to generate cardiomyocytes. In the pilot experiment, with serial delivery of CellTags, we showcase the compatibility of the CellTag lineage tracing system with this new reprogramming system. In addition, we learn the behavior of these cells, which instruct our future experimental design to take these factors into account.

Brief investigation of atrial-ventricular regionalization with or without small molecules suggests potential patterning events introduced via these modulations. This result requires further

validation as our small molecule treatment group did not show as significant fold increase as previously described in producing cardiomyocytes (Mohamed et al., 2017). We speculate a potential reason that the construct we used for MGT infection is previously optimized and ordered in a polycistronic manner. This construct has previously been shown to significantly increase production of reprogrammed cardiomyocytes, compared to separate transduction of the M, G, T factors (L. Wang et al., 2015), resulting in our observed higher cardiomyocyte rates in the control groups.

In order to identify key transcriptional regulators in this system, we employed CellOracle analysis to construct gene regulatory networks and identified two potential key factors, Klf5 and Atf3. Both factors have been previously reported to play vital roles in cardiovascular remodeling in response to injury and hypertensive stress (Y. Li et al., 2017; Nagai et al., 2005). Perturbation simulation with CellOracle demonstrates that knockout of either one of the factors will block transitions to successful reprogrammed cells. Yet, overexpression of either one will promote transition to more cardiomyocyte-like cell states. In addition, we speculate that overexpression of Atf3 could lead us to a specific cardiac cell type, blocking possibilities to other cardiac fates. This analysis opens new hypothesis for testing to promote the efficiency and fidelity of reprogramming.

Refocusing on the blueprint of this project to identify the comprehensive lineage map, we perform state-fate experiment to capture lineage information together with transcriptome profiles or chromatin landscapes. Leveraging nascent CellTag-AR development, we identified two putative trajectories in both assays – one toward successfully reprogrammed and the other to fibroblastic states. This mirrors our previous discoveries in MEF to iEP reprogramming, where we observed a successfully reprogrammed vs a ‘dead-end’ trajectory. In addition, it reflects

previous reported refractory lineage inferred via RNA velocity (Stone et al., 2019; Y. Zhou et al., 2019). Comparing RNA and ATAC trajectories, we observe a more distinct separation between the two directions from the ATAC clonal information than the more overlapping trajectories in RNA. We speculate that this demonstrates the upstream chromatin remodeling has occurred to turn off fibroblastic program while the downstream transcriptional remodeling is gradually losing their fibroblast expression profile. Indeed, it has been previously reported that the cells acquire a cardiac expression profile rapidly with a gradual loss of their starting transcriptional profiles (Liu et al., 2017; Stone et al., 2019; Y. Zhou et al., 2019).

Through this preliminary analysis, we present this rich dataset profiling both transcriptional and chromatin landscape and demonstrate our ability to track cells in this system. This opens broad opportunities to probe changes occurring in the transcriptome and chromatin level, further providing mechanistic insight in reprogramming to improve cell fate engineering strategies.

3.5 Materials and Methods

3.5.1 Cell Culture

Immortalized mouse embryonic fibroblast cell line was a kind gift from the laboratory of Dr. Li Qian at University of North Carolina, Chapel Hill. The cells were maintained and passaged according to its protocol (Vaseghi et al., 2016) in Dulbecco's Modified Eagle Medium (Gibco) supplemented with 10% Fetal Bovine Serum (Gibco), 1% penicillin/streptomycin (Gibco), and 55 M 2-mercaptoethanol (Gibco).

3.5.2 Mice

Mice used for cardiac fibroblast explant experiment were of C57BL/6J background (The Jackson Laboratory #000664). All animal care procedures were approved by Washington University Institutional Animal Care and Use Committee.

3.5.3 Cardiac Fibroblast Explant Culture

The cardiac fibroblasts were derived via an explant culture, following previous protocol (Qian et al., 2013). In brief, explant media was prepared with Iscove's Modified Dulbecco's Medium (IMDM) supplemented with 20% Fetal Bovine Serum (Gibco), 1% penicillin/streptomycin (Gibco), and 55 M 2-mercaptoethanol (Gibco). hearts were harvested from P0-P3 mice, rinsed with 1x DPBS (Gibco) for 3 times, and minced to small pieces around 0.5 mm^3 in size. The tissue pieces were plated on a 0.1% gelatin-coated 6-well plate with minimum explant media (~0.5 to 1 ml to barely cover the bottom). The plate was incubated at 37°C, 5% CO_2 in an incubator for 2 hours to allow the tissue pieces to attach on the bottom of the plate. Post the 2-hour incubation, 2-ml fresh and pre-warmed explant media was added to each well. The plate was then incubated at 37°C, 5% CO_2 in an incubator for 3 days without disturbance. After the 3 days, medium was changed every three to four days or till medium becomes yellow. Explanted fibroblast was harvested with trypsin, followed by passing through a 70- m cell strainer, on day 7 post initiation of the explant culture. The resulting cells were counted and replated on 0.1% gelatin coated plates at a density of 500k/10-cm dish to be used for reprogramming the next day.

3.5.4 Reprogramming Virus Production

The retrovirus for cardiac reprogramming was freshly prepared. 293T cells (RRID:CVCL_1926) were maintained and passaged in fibroblast media (10% FBS, 1x penicillin-streptomycin, 1x -Mercaptoethanol, in DMEM). 293T cells were seeded at a density of 3 million per 10-cm plate the day before transfection. The following day, the cells were transfected with pMX-MGT (RRID:Addgene_111810) with 5g of pCL-Eco (RRID:Addgene_12371), using X-tremeGENE 9 DNA transfection reagent (Sigma, 6365779001) according to the manufacturer's instructions. Media was replaced with fresh fibroblast media the following day. Retrovirus was harvested the next day by taking the supernatant from the transfected plate and filtered through a 45- m syringe filter. 500x protamine sulfate was added to the viral media prior to transduction of the mouse cardiac fibroblast.

3.5.5 Cardiomyocyte Reprogramming with Primary Cells

Direct cardiac reprogramming was performed using primary cardiac fibroblasts derived from the cardiac fibroblast explant culture. Briefly, cardiac fibroblasts (MCFs) were harvested after seven days of the explant culture and cultured overnight on gelatin-coated plates in MEF media (DMEM supplemented with 10% FBS, 1% penicillin/streptomycin, and 55 M 2-mercaptoethanol). MCFs were plated at a density around 500k per 10-cm dish pre-treated with 0.1% gelatin solution for 20~30 minutes. MCFs were infected with freshly harvested pMX-MGT retrovirus (L. Wang et al., 2015) (day 0). The viral media was replaced with fresh cardiomyocyte media (10% M199, 10% FBS, 1% NEAA, 1% sodium pyruvate, 1x penicillin-streptomycin, 1x Glutamax, in DMEM). In the small molecule comparison experiment, the cardiomyocyte media was supplemented with 2.6M SB431542 (Cayman Chemical, Catalog #13031) or DMSO as a vehicle control (day 1). 5M XAV939 (Cayman Chemical, Catalog #13596) or DMSO was added

to the plate without media change (day +1) for the small molecule group. The media was replaced with fresh cardiomyocyte media two days after the last addition of small molecule (day +3). Media was renewed every 2~3 days. The cells were collected, filtered through a 70-m strainer, resuspended in 1% BSA in PBS, and counted on day 0, day 13, and day 27 for scRNA-seq (see below).

3.5.6 Cardiomyocyte Reprogramming with inducible Fibroblast Cell Line

Direct cardiac reprogramming was performed using the immortalized inducible fibroblast cell line (MEF-T iMGT) (Vaseghi et al., 2016). Briefly, MEF-T iMGT was maintained on gelatin-coated plates in MEF media (DMEM supplemented with 10% FBS, 1% penicillin/streptomycin, and 55 M 2-mercaptoethanol). 12-hour prior to reprogramming initiation, MEF-T iMGT cells were replated at a density around 10k per well on a 48-well plate with 0.1% gelatin solution for 20~30 minutes. The cells were allowed to attach for ~12hrs. The media was aspirated and replaced with cardiomyocyte media with CellTag-AR virus and 1 g/ml doxycycline. The viral media was replaced with fresh cardiomyocyte media (10% M199, 10% FBS, 1% NEAA, 1% sodium pyruvate, 1x penicillin-streptomycin, 1x Glutamax, in DMEM), containing 1 g/ml doxycycline. Doxycycline was maintained in culture for three days and replaced with new treatment every day. The cells were collected, filtered through a 70-m strainer, resuspended in 1% BSA in PBS, and counted on day 2 and day 5 for scRNA-seq.

3.5.7 Immunostaining for day 28 Reprogrammed Cells in Cardiac Reprogramming

Mouse cardiac fibroblasts were generated as described above. On day 27 of the reprogramming process, the cells were transferred to 4-Chamber Culture Slides (Falcon). On the next day, the cells were rinsed with 1x DPBS and fixed in 4% paraformaldehyde for 20 minutes

at room temperature. The samples were then washed with 1x DPBS three times, permeabilized, and blocked with blocking buffer (0.1% TritonX-100 and 5% BSA in DPBS) for an hour. The primary antibodies, Cardiac Troponin T (RRID:AB_11000742) and -Actinin (Sigma-Aldrich, Product No. A7811), were diluted 1:500 (cTnT) and 1:800 (ACTN2) in the 1% BSA in DPBS. The blocking buffer was then removed from the sample, and 250 l of diluted primary antibodies was added to two different wells. The samples were incubated with the primary antibody at 4°C overnight (12~16 hrs). The samples were then washed for 5 minutes three times. The secondary antibodies, Alexa Fluor 568 Goat Anti-mouse IgG (RRID:AB_2534072), were diluted 1:1000 in the blocking buffer. The secondary antibodies were added and incubated at room temperature for 1 hour in the dark on a rocker. The samples were washed again for 5 minutes, three times. 100 l of 300 nM DAPI (Invitrogen) was added to each slide chamber and incubated at room temperature for 1 minute. The samples were washed for 5 minutes once. The final wash with DPBS was aspirated and the chamber was removed from the slides. A coverslip was then applied with ProLong Gold Antifade Mountant (Invitrogen). The slides were imaged using an Olympus FV1200 Confocal Microscope with 60x water objectives.

3.5.8 CellTag Lentivirus Production

293T cells (RRID:CVCL_1926) were maintained and passaged in fibroblast media (10% FBS, 1x penicillin-streptomycin, 1x -Mercaptoethanol, in DMEM). 293T cells were seeded at a density of 3 million per 10-cm plate the day before transfection. CellTag virus was produced by transfecting 293T cells with lentiviral pSMAL vector and packaging plasmids pCMV-dR8.2 dvpr (Addgene plasmid 8455) and pCMV-VSV-G (Addgene plasmid 8454) using X-tremeGENE 9 (Sigma, 6365779001). Media was replaced with fresh fibroblast media the following day.

Virus was collected 48 and 72 hours after transfection by filtering the supernatant through a 0.45-m syringe filter.

3.5.9 CellTag Transduction

The CellTag virus containing supernatant as collected above was used immediately or kept at 4°C for up to a week. Prior to transduction, 500x protamine sulfate was added to the viral media. Media was aspirated from the cells to be transduced. The viral media was immediately added to the cells and left on the plate for a 24-hour transduction period. This transduction was performed on day -2 with V1 library, day 10 with V2 library, and day 21 with V3 library for the primary fibroblast reprogramming or day 0 with CellTag-AR library for inducible MEFs.

3.5.10 Single-Cell RNA Profiling

For single-cell library preparation on the 10x Genomics platform, we used: the Chromium Single Cell 3′ Library & Gel Bead Kit v2 (PN-120237), Chromium Single Cell 3′ Chip kit v2 (PN-120236), and Chromium i7 Multiplex Kit (PN-120262), according to the manufacturer's instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide. Prior to cell capture, methanol-fixed cells were placed on ice, then spun at 3000rpm for 5 minutes at 4°C, followed by resuspension and rehydration in PBS, according to Alles et al., 2017. 17,000 cells were loaded per lane of the chip, aiming to capture 10,000 single-cell transcriptomes. The resulting cDNA libraries were quantified on an Agilent Tapestation and sequenced on an Illumina HiSeq 2500. For analysis of cardiomyocyte reprogramming, The Chromium Single Cell 3' (v2) Reagent Kits (PN-120237, PN-120236, PN-120262) were used to prepare single-cell RNA-seq libraries, according to manufacturer's guidelines. Libraries were pooled and sequenced on an Illumina NextSeq 550. For motor neuron programming, prior to loading the 10x chip, methanol-fixed

cells were counted, spun, resuspended in 1% BSA in PBS, and counted again, according to 10x Genomics methanol fixation protocol. The Chromium Single Cell 3' (v2) Reagent Kits (PN-120237, PN-120236, PN-120262) were used to prepare single-cell RNA-seq libraries, according to manufacturer's guidelines. Libraries were pooled and sequenced on an Illumina NextSeq 550.

3.5.11 Single-Cell ATAC Profiling

For single-cell ATAC library preparation on the 10x Genomics platform, we used: the Chromium Next GEM Single Cell ATAC Library & Gel Bead Kit v1.1 (PN-1000175), Chromium Next GEM Chip H Single Cell Kit v1.1 (PN-1000161), and Single Index Kit N Set A (PN-1000212), according to the manufacturer's instructions in the Chromium Next GEM Single Cell ATAC Reagents Kits v1.1 User Guide. Briefly, the cells were harvested, counted, and subjected to nuclei isolation and light fixation with 4% paraformaldehyde for 4 minutes at room temperature, followed by tagmentation of accessible genome in the nuclei. An *in-situ* reverse transcription step was performed on the nuclei prior loading the chip for better capture of the CellTag-AR barcodes. After the reverse transcription, the nuclei were resuspended and spun at 600xg for 10 min at 4°C. Supernatant was partly removed with measured amount left for loading on the 10x scATAC Chip. 20,000 cells were processed to generate nuclei for loading per lane of the chip, aiming to capture 10,000 single-nuclei accessible genome. The GEM generated from the chip was further prepared to generate the final libraries, according to the manufacturer's instructions. The resulting libraries were quantified on an Agilent Tapestation and sequenced on an Illumina NextSeq 550.

3.5.12 Single-Cell RNA Sequencing Analysis

The Cell Ranger v6.0 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to align reads, process, and filter data generated using 10x Chromium single-cell gene expression platform. Following this step, the default Cell Ranger pipeline was implemented to generate the filtered output data for downstream analysis. To process and analyze scRNA-seq data, we used the R package, Seurat V4, following the tutorial (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html). Briefly, each sample was pre-processed based on RNA counts and mitochondria percentages and then normalized. The highly variable genes were then identified, followed by scaling and dimensional reduction via PCA. With the selected number of components, graph-based clustering and UMAP plotting were further performed.

3.5.13 CellOracle Gene Regulatory Network Analysis

To analyze the scRNA-seq data through CellOracle, the data was processed through SCANPY standard pipeline (https://scanpy.readthedocs.io/en/stable/) (Wolf et al., 2018). The single-cell data was processed through similar preprocessing, filtering, and clustering as described above. The data was then subject to trajectory inference using PAGA (Wolf et al., 2019). The Anndata result from this processing was passed into CellOracle. Gene regulatory network models were constructed on a cluster-by-cluster basis and enabled network analysis of each individual cluster to identify top rank transcription factors. Leveraging the *in-silico* perturbation function with GRNs in CellOracle, the data was the altered to have a gene knocked out or overexpressed. Knockout set the expression level to zero where overexpression set the expression level to the max expression level of that gene in the system. The perturbation was then visualized with gradient vectors and perturbation scores.

3.5.14 CellTag Data Analysis

For CellTag V1, V2, and V3, CellTag information was analyzed as aforementioned (Biddy et al., 2018; Kong et al., 2020). In brief, the CellTags were extracted from aligned bam files, followed by UMI counting to generate the count matrix. With the count matrix, the data was further binarized, passed through an allow list, and QC metric filtered. Finally, Jaccard analysis was performed on the filtered matrix and further graph-based method was used to identify clonally related cells. For CellTag-AR, similar approaches were adapted for extraction, filtering of the data, and clone calling. The clone calling process using CellTag-AR leverage an additional step to break down sparse clones using graph cliques.

3.5.15 Randomized Test for Trajectory Discovery

To identify clones enriched or depleted of iCM generation, we applied similar strategies as previously described (Biddy et al., 2018). Briefly, randomized testing to evaluate whether each clone (of at least 10 cells) has a similar percentage of fully reprogrammed cells relative to a randomly selected population of the same size. The percentage of reprogrammed cells is defined as the proportion of cells within each group found in the reprogrammed cluster, as defined by Seurat together with Capybara and marker gene scores. Let N represent the number of cells in each clone and M represent the remaining cell population size. We pool the two groups of cells (size=N+M) and resample N random cells, without replacement, from the pooled cells (N+M)/N times such that every possible separation with ending groups of size N and M can be sampled and captured. During this process, the percentage is calculated based on the N randomly sampled cells. With the percentage calculated, P values can be evaluated based on the proportion of randomly sampled cells with a percentage greater than or equal to the null percentage.

3.6 Detailed Figure Legends

Figure 3.1: Overview of CellTagging System and Previous Discovery in MEF to iEP Reprogramming. (A) Top: Schematic of the CellTagging barcode cell label, including a GFP with an 8-nt random barcode integrated into its 3'-UTR. Bottom: Outline of CellTagging and trace method for reconstruction of lineage during reprogramming. (B) Application of the lineage tracing approach to MEF to iEP reprogramming uncovers two distinct trajectory – one toward successfully reprogrammed outcome (Bottom) and one toward an off-target cell type (Top).

Figure 3.2: Overview of CellTagR Pipeline and CellTag Indexing Strategy. (A) Pipeline to assess library complexity after sequencing of the CellTag barcode library, prior to application to biological systems. (B) Pipeline for CellTag data extraction, filtering, and clone calling post single-cell RNA sequencing. (C) Schematic of CellTag Indexing Strategy, where unique barcodes are delivered to different populations. Post single-cell sequencing, the barcodes were extracted, normalized, filtered, and provide a classification of the samples toward their labelled population categories.

Figure 3.3: Immunostaining of Cardiac Troponin Protein. Top Row: Negative control group staining fibroblast. Bottom Row: Staining of the protein in cells 28 days after cardiomyocyte reprogramming was initiated.

Figure 3.4: Preliminary single-cell RNA-sequencing analysis. (A) Left: Clustering of the data using Seurat, presenting a total of 14 clusters. Right: the distribution of cells from different time points on the UMAP embedding. (B) Expression of key markers. Cardiac fibroblast: Tbx20 and Col1a2. Cardiomyocyte: Tnnt2 and Des. (C) Top: Immunofluorescent experiment staining for α-

actinin. Red: α-actinin. Bottom: primary lineage analysis using single-cell data, in which each center node labels for a clone for each version of CellTag libraries and each branch node denote a cell in the clone. D) CellTag expression gauged by GFP expression. E) Overall distribution of CellTag counts captured after filtering. F) Percentages of cells captured with 1+ or 2+ CellTags.

Figure 3.5: Further Analysis of the Preliminary Data using SCANPY and PAGA. (A) Schematic of the experiment. (B) Left: PAGA embedding of the time course reprogramming dataset. 'FA': Force Atlas. Right: Distribution of cells from different time points on the PAGA embedding. (C) Expression of additional key markers. Cardiac fibroblast: *Postn, Tcf21,* and *Tbx20*. Cardiomyocyte: *Tnnt2*, *Tnnc1,* and *Des*. (D) Overall distribution of CellTag counts captured after filtering. (E) Cardiomyocyte and fibroblast identity score distribution on the FA plot. (F) Clusters labelled with dominant cell-type classified by Capybara.

Figure 3.6: Identification of Enriched vs Depleted Clones. (A) P-values of major clones tested (>10 cells) using randomized test. (B) Distribution of cells that have clonally related siblings. (C) Overlay of the iCM depleted clones on the FA plot. It demonstrates the depleted clone initiates at the source fibroblast population and move to the off-target branch. (D) Projection of pseudotime of these cells to guide the directionality of cell identity shift.

Figure 3.7: Capybara Analysis comparing Cardiac Reprogramming with or without small molecules. (A) Schematic for the design of the experiment. (B) Comparison of the atrial to ventricular cardiomyocyte ratio between the small molecule and no small molecule treatment. An increase of this ratio was observed with the addition of small molecules (P=0.0287, randomized test). (C) Overlay of Capybara classification on the integrated space of our dataset with Stone et al dataset. (D) Log-normalized expression profile of atrial and ventricular markers

between the two treatment groups. Differential enrichment or depletion of the expression is tested by Wilcoxon rank sum test (*: P <= 0.05; **: P <= 0.01; ***: P <= 0.001; ****: P <= 0.0001)

Figure 3.8: CellOracle Analysis of the Preliminary Time Course Single-Cell Dataset. (A) Diffusion pseudotime calculated via PAGA. (B) Degree centrality for the top 30 TFs in cluster 1 and 10, suggesting a putative factor *Klf5*. (C) Inferred gradient vector under normal development from CellOracle, indicating potential directionality of the cell movement in this dataset. (D) Log-normalized *Klf5* expression projected on the FA plot. (E) CellOracle knockout simulation of *Klf5*, showing transitions toward cardiac fate is blocked. (F) CellOracle overexpression simulation of *Klf5*, showing promoted transition to cardiomyocytes. Blue region: negative perturbation score, transitions toward this region are blocked. Red region: positive perturbation score, transitions toward this direction are promoted.

Figure 3.9: SCANPY and PAGA analysis of the dataset with small molecule treatment. PAGA embedding of the time course reprogramming dataset, representing a total of 13 clusters. 'FA': Force Atlas.

Figure 3.10: CellOracle Analysis of the Dataset with small molecules. (A) Diffusion pseudotime calculated via PAGA. (B) Degree centrality for the top 30 TFs in cluster 3, 4 and 10, suggesting a putative factor *Atf3*. (C) Inferred gradient vector under normal development from CellOracle, indicating potential directionality of the cell movement in this dataset. (D) Log-normalized *Atf3*expression projected on the FA plot. (E) CellOracle knockout simulation of *Atf3*, showing transitions toward cardiac fate is blocked. (F) CellOracle overexpression simulation of *Atf3*, showing promoted transition to cardiomyocytes. Blue region: negative perturbation score,

transitions toward this region are blocked. Red region: positive perturbation score, transitions toward this direction are promoted.

Figure 3.11: State-Fate Experiment (RNA Profile). (A) Schematic of the experimental design. (B) The samples were processed and integrated via Seurat V4, representing a total of 12 clusters. (C) Distribution of cells from different time points on the UMAP projection, showing even distribution across the clusters. (D) Expression of key markers. Fibroblast: *Col1a2* and *S100a4*. Cardiomyocyte: *Tnnt2* and *Tnnc1*. (E) Cardiomyocyte and fibroblast identity score distribution on the UMAP plot, referencing the Stone et al. dataset.

Figure 3.12: Identification of Enriched or Depleted Clones based on Transcriptional Lineage. (A) Distribution of cells that have clonally related siblings, split by time points, showing even clonal relationship across the clusters and days (B) Overlay of the iCM enriched clones on the UMAP plot. (D) Overlay of the iCM depleted clones on the UMAP plot.

Figure 3.13: State-Fate Experiment (ATAC Profile). (A) The samples were processed and aggregated via ArchR, representing a total of 12 clusters. (B) Distribution of cells from different time points on the UMAP projection, showing major occupation of day 2 cells on the top and day 5 cells on the bottom. (C) Imputed gene activity scores of key markers as previously reported (H. Wang et al., 2022). Fibroblast: *Postn, Dcn* and *Dsp*. Cardiomyocyte: *Ryr2, Myom2,* and *Actc1*. (D) Percentage of day 2 and day 5 cells in each cluster. (E) Motif enrichment identified in the differentially accessible peaks for each cluster.

Figure 3.14: GREAT analysis and Comparison to Chip-seq Peaks. (A) Predicted annotation for the genomic regions differentially accessible for Cluster 5 and 6. GREAT analysis predict functional annotation for genomic regions. We note most annotations to be related to immune

response and vascular development. (B) Predicted annotation for the genomic regions differentially accessible for Cluster 8 and 10. We note most annotations to be related to muscle and cardiac development as well as muscle structure organization. (C) Comparison of differential peaks to Chip-seq data of Mef2c, Gata4, and Tbx5, suggesting cells in Cluster 8-11 have accessible peaks for these three reprogramming TFs.

Figure 3.15: Identification of Enriched or Depleted Clones based on Chromatin Landscape. (A) Distribution of cells that have clonally related siblings, split by time points, showing even clonal relationship across the clusters (B) Top: projection of the cells in enriched (left) or depleted (right) clones on the UMAP. Light purple: Day 2 cells, Dark purple: Day 5 cells. Each colored path marks a different clone. Bottom: Overlay of the iCM enriched (left) or depleted (right) clones on the UMAP plot. The arrows are the inferred the direction of movement based on the paths marked by the clones, sourcing from day 2 cells to day 5 cells.

3.7 Acknowledgement

We thank members of the Morris laboratory for their helpful discussions. We thank Dennis Oakley of the Washington University Center for Cellular Imaging (WUCCI) supported by Washington University School of Medicine, The Children's Discovery Institute (CDI-CORE-2015-505 and CDI-CORE-2019-813) and the Foundation for Barnes-Jewish Hospital (3770 and 4642). This work was funded by National Institute of General Medical Sciences R01 GM126112, and Silicon Valley Community Foundation, Chan Zuckerberg Initiative Grant HCA2-A-1708-02799, both to S.A.M.; S.A.M. is supported by an Allen Distinguished Investigator Award (through the Paul G. Allen Frontiers Group), a Vallee Scholar Award, a Sloan Research Fellowship, and a New York Stem Cell Foundation Robertson Investigator Award; W.K. is supported by a Douglas Covey Fellowship. Special thanks go to Guillermo Rivera-Gonzales for assistance in animal protocols, to Kunal Jindal for assistance of the CellTag-AR experiment and analysis, and to Xue Yang for assistance in analysis.

3.8 Author Contribution

Wenjun Kong: study concept and design, data acquisition, analysis, interpretation, figure preparation; Samantha A. Morris: study concept and design, obtained funding, overall study supervision.

Chapter 4: Closing Remarks and Future Directions

In this dissertation, we aim to unravel population heterogeneity in a complex and continuous biological system via development and application of computational and experimental tools, with a focus on reprogramming. The findings of this work, both preliminary and published, advance the field by providing a valuable tool with a unique focus on the hybrid cells and a rich dataset that captures lineage together with transcriptome profiles as well as chromatin landscapes. The discoveries with the application of nascent tools in the Morris lab opens opportunities to answer broad range of questions for continued advancement of the field and research. In this final chapter, we briefly pinpoint interesting questions for future pursuit from this work.

Can we distinguish different types of hybrid cells? In Chapter 2, we present our work, Capybara, in effort to capture transition cells during continual biological processes. It enabled us to observe rare intermediates in different dynamic systems. Yet, as transition cells have been proposed to play diverse roles (MacLean et al., 2018), an interesting problem to pursue is further separation and characterization of these transition cells. For instance, the transition population can mark a bistable state toward two terminal fates or a transient state from one progenitor state to a terminal fate. A potential approach to isolate such population is to consider the stability of the identified hybrid state. A bistable state would demonstrate higher stability compared to a transient state. In addition, as we carefully note that hybrid cells identified in Capybara do not imply directionality, incorporation of trajectory inference or velocity metric could facilitate the identification of directionality of flow in these identified intermediate states.

In Chapter 3, we showcase the joint application of nascent tools in the Morris lab to reveal cell type dynamics and mechanistic insights in direct cardiac reprogramming. Particularly, we reveal two putative trajectories of this reprogramming – one toward successfully reprogrammed state vs. the other toward the fibroblastic fate. The preliminary analysis describes the overarching picture of the lineage map. Taking a deeper dive, we can ask the following questions to further identify changes that define fates.

From fate to state, where are the reprogrammed cells originating from? It has been previously reported that the dynamic cell types in the heterogeneous starting population have different reprogramming efficiency (Biddy et al., 2018; Mahmoudi et al., 2019). Leveraging this state-fate dataset, we can trace from the fate cells to their originating states. Further comparison of their gene expression profile and chromatin landscape can lead us to identify conserved expression or accessible chromatin regions that could be key to a restricted fate. Identification of such can lead to discovery of key regulatory genes in specific fate decision.

From state to fate, do the cells diverge? Previously, we have identified that cell fate decision is stochastic yet becomes deterministic early during reprogramming, leading to restricted lineages (Biddy et al., 2018). Here, we can follow the state cells to their fate cells to assess if we observe similar lineage restriction as we previously observed. If the cells in a clone diverge, additional comparisons between the bifurcation cluster and fate clusters can highlight key transition transcriptional state and chromatin landscape. In this manner, we can further assess if the chromatin has already been remodeled prior to transcriptional changes. Such analysis can further refine our lineage maps for this reprogramming protocol and pinpoint bifurcation dynamics in the system.

Are successfully reprogrammed clones identified with transcription profile also considered successful in chromatin landscape? As described in Chapter 3, we observe more distinctive trajectories from the chromatin landscape point of view compared to expression profiles, reflecting the slow transcriptional remodeling. In addition, in the ATAC lineages, we identified a potential trajectory to depleted clones from an early reprogrammed cluster. Hence, it would be intriguing to explore if their RNA siblings have also lost the cardiac profile with a regain of fibroblastic fate. Considering the delay between upstream chromatin shift to downstream transcriptome modification, this can help identify key accessible genomic regions that are responsible for this refractory lineage.

With the basic lineage map, how is the map altered upon modulation with other factors, such as signaling molecules and other regulatory factors? Previously, it has been shown that other factors can enhance efficiency and fidelity of reprogramming (Mohamed et al., 2017; Hashimoto et al., 2019; Zhou et al., 2015; Jayawardena et al., 2012). Additional functional studies reveal the different roles of the factors in alteration of the chromatin landscape and transcriptional space (Hashimoto et al., 2019). From a lineage perspective, we could ask the question if the lineage is shifted, and the bifurcation observed in this study disappears. State-fate experiments with this modulation can be further employed to identify lineage changes and find key drivers of these changes. In addition, alternation in the lineage map can help pinpoint the factors could achieve both high efficiency and fidelity during reprogramming.

Can we improve the efficiency and fidelity of cardiac reprogramming more? This is a long-standing question in the field of direct cell fate engineering (Guo & Morris, 2017). Utilizing CellOracle (Kamimoto et al., 2020), we highlight two putative transcription factors that could potentially increase the generation of cardiomyocytes *in vitro*. These factors can be tested and

evaluated *in vitro* by performing a reprogramming experiment with the traditional cocktail in joint with the new factors. The resulting cells can be evaluated via qRT-PCR or single-cell sequencing for full profile and immunostaining for structural representation of cardiomyocytes.

In a nutshell, the tool and data presented in this dissertation provide important findings to the field of cell fate engineering. In addition, it laid the foundation for exploration in understanding the mechanisms of *in vitro* regeneration strategies, further improving them closer to future application in medicine.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1795-z

Abdellah, Z., Ahmadi, A., Ahmed, S., Aimable, M., Ainscough, R., Almeida, J., Almond, C., Ambler, A., Ambrose, K., Ambrose, K., Andrew, R., Andrews, D., Andrews, N., Andrews, D., Apweiler, E., Arbery, H., Archer, B., Ash, G., Ashcroft, K., … Kamholz, S. (2004). Finishing the euchromatic sequence of the human genome. *Nature 2004 431:7011*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Alles, J., Karaiskos, N., Praktiknjo, S. D., Grosswendt, S., Wahle, P., Ruffault, P.-L., Ayoub, S., Schreyer, L., Boltengagen, A., Birchmeier, C., Zinzen, R., Kocks, C., & Rajewsky, N. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biology*, *15*(1), 44. https://doi.org/10.1186/s12915-017-0383-5

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., & Powell, J. E. (2019). ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1862-5

Altschuler, S. J., & Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell*, *141*(4), 559–563. https://doi.org/10.1016/j.cell.2010.04.033

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Andrews, T. S., Kiselev, V. Y., McCarthy, D., & Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. In *Nature Protocols* (Vol. 16, Issue 1). Nature Research. https://doi.org/10.1038/s41596-020-00409-w

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163–172. https://doi.org/10.1038/s41590-018-0276-y

Aydin, B., & Mazzoni, E. O. (2019). *Cell Reprogramming: The Many Roads to Success*. https://doi.org/10.1146/annurev-cellbio-100818

Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., & Tanay, A. (2019). MetaCell: Analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1812-2

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, *3*(4), 346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., Bouley, D. M., Lujan, E., Haddad, B., Daneshvar, K., Carter, A. C., Flynn, R. A., Zhou, C., Lim, K. S., Dedon, P., Wernig, M., Mullen, A. C., Xing, Y., Giallourakis, C. C., & Chang, H. Y. (2014). M6A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell*, *15*(6), 707–719. https://doi.org/10.1016/j.stem.2014.09.019

Biddy, B. A., Kong, W., Kamimoto, K., Guo, C., Waye, S. E., Sun, T., & Morris, S. A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. *Nature*, *564*(7735), 219–224. https://doi.org/10.1038/s41586-018-0744-4

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799. https://doi.org/10.1038/NATURE05874

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). https://doi.org/10.1088/1742-5468/2008/10/P10008

Boufea, K., Seth, S., & Batada, N. N. (2020). scID Uses Discriminant Analysis to Identify Transcriptionally Equivalent Cell Types across Single-Cell RNA-Seq Data with Batch Effect. *IScience*, *23*(3). https://doi.org/10.1016/j.isci.2020.100914

Brackston, R. D., Lakatos, E., & Stumpf, M. P. H. (2018). Transition state characteristics during cell differentiation. *PLoS Computational Biology*, *14*(9). https://doi.org/10.1371/journal.pcbi.1006405

Briggs, J. A., Li, V. C., Lee, S., Woolf, C. J., Klein, A., & Kirschner, M. W. (2017). Mouse embryonic stem cells can differentiate via multiple paths to the same state. *ELife*, *6*. https://doi.org/10.7554/eLife.26945

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. https://doi.org/10.1038/nbt.4096

Cahan, P., Li, H., Morris, S. A., Lummertz Da Rocha, E., Daley, G. Q., & Collins, J. J. (2014). CellNet: Network biology applied to stem cell engineering. *Cell*, *158*(4), 903–915. https://doi.org/10.1016/j.cell.2014.07.020

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., Zager, M. A., Aldinger, K. A., Blecher-Gonen, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D., Steemers, F. J., Glass, I. A., Trapnell, C., & Shendure, J. (2020). A human cell atlas of fetal gene expression. *Science*, *370*(6518). https://doi.org/10.1126/science.aba7721

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., & Shendure, J. (2017). Comprehensive single cell transcriptional profiling of a multicellular organism. *Science (New York, N.Y.)*, *357*(6352), 661. https://doi.org/10.1126/SCIENCE.AAM8940

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., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature 2019 566:7745*, *566*(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Cates, K., McCoy, M. J., Kwon, J. S., Liu, Y., Abernathy, D. G., Zhang, B., Liu, S., Gontarz, P., Kim, W. K., Chen, S., Kong, W., Ho, J. N., Burbach, K. F., Gabel, H. W., Morris, S. A., & Yoo, A. S. (2021). Deconstructing Stepwise Fate Conversion of Human Fibroblasts to Neurons by MicroRNAs. *Cell Stem Cell*, *28*(1), 127-140.e9. https://doi.org/10.1016/j.stem.2020.08.015

Chen, H. J., Meng, T., Gao, P. J., & Ruan, C. C. (2021). The Role of Brown Adipose Tissue Dysfunction in the Development of Cardiovascular Disease. In *Frontiers in Endocrinology* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fendo.2021.652246

Chen, T., Hao, Y. J., Zhang, Y., Li, M. M., Wang, M., Han, W., Wu, Y., Lv, Y., Hao, J., Wang, L., Li, A., Yang, Y., Jin, K. X., Zhao, X., Li, Y., Ping, X. L., Lai, W. Y., Wu, L. G., Jiang, G., … Zhou, Q. (2015). M6A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell*, *16*(3), 289–301. https://doi.org/10.1016/j.stem.2015.01.016

Dahlin, J. S., Hamey, F. K., Pijuan-Sala, B., Shepherd, M., Lau, W. W. Y., Nestorowa, S., Weinreb, C., Wolock, S., Hannah, R., Diamanti, E., Kent, D. G., Ottgens, B. G. ¨, & Wilson, N. K. (2018). A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood*. http://ashpublications.org/blood/article-pdf/131/21/e1/1468628/blood821413.pdf

Davis, R. L., Weintraub, H., & Lassar, A. B. (1987). Expression of a Single Transfected cDNA Converts Fibmblasts to Myoblasts. In *Cell* (Vol. 51).

de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., & Holstege, F. C. P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, *47*(16), e95. https://doi.org/10.1093/nar/gkz543

de Soysa, T. Y., Ranade, S. S., Okawa, S., Ravichandran, S., Huang, Y., Salunga, H. T., Schricker, A., del Sol, A., Gifford, C. A., & Srivastava, D. (2019). Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. *Nature*, *572*(7767), 120–124. https://doi.org/10.1038/s41586-019-1414-x

de Wit, E. (2017). Capturing heterogeneity: single-cell structures of the 3D genome. *Nature Structural & Molecular Biology*, *24*(5), 437–438. https://doi.org/10.1038/nsmb.3404

Delile, J., Rayon, T., Melchionda, M., Edwards, A., Briscoe, J., & Sagner, A. (2019). Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development (Cambridge, England)*, *146*(12). https://doi.org/10.1242/dev.173807

DePasquale, E. A. K., Schnell, D. J., van Camp, P. J., Valiente-Alandí, Í., Blaxall, B. C., Grimes, H. L., Singh, H., & Salomonis, N. (2019). DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Reports*, *29*(6), 1718-1727.e8. https://doi.org/10.1016/j.celrep.2019.09.082

Diaz-Mejia, J. J., Meng, E. C., Pico, A. R., MacParland, S. A., Ketela, T., Pugh, T. J., Bader, G. D., & Morris, J. H. (2019). Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Research*, *8*, 296. https://doi.org/10.12688/f1000research.18490.1

Dimos, J. T., Rodolfa, K. T., Niakan, K. K., Weisenthal, L. M., Mitsumoto, H., Chung, W., Croft, G. F., Saphier, G., Leibel, R., Goland, R., Wichterle, H., Henderson, C. E., & Eggan, K. (2008). Induced Pluripotent Stem Cells Generated from Patients with ALS Can Be Differentiated into Motor Neurons. *Science*. https://www.science.org

Ekiz, H. A., Conley, C. J., Stephens, W. Z., & O'Connell, R. M. (2020). CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments. *BMC Bioinformatics*, *21*(1), 191. https://doi.org/10.1186/s12859-020-3538-2

Elmentaite, R., Kumasaka, N., Roberts, K., Fleming, A., Dann, E., King, H. W., Kleshchevnikov, V., Dabrowska, M., Pritchard, S., Bolt, L., Vieira, S. F., Mamanova, L., Huang, N., Perrone, F., Goh Kai'En, I., Lisgo, S. N., Katan, M., Leonard, S., Oliver, T. R. W., … Teichmann, S. A. (2021). Cells of the human intestinal tract mapped across space and time. *Nature 2021 597:7875*, *597*(7875), 250–255. https://doi.org/10.1038/s41586-021-03852-1

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, *297*(5584), 1183–1186. https://doi.org/10.1126/SCIENCE.1070919/SUPPL_FILE/ELOWITZSOM.PDF

F Ozsolak, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet*, *12*, 87–98.

Fan, H. C., Fu, G. K., & Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science*, *347*(6222). https://doi.org/10.1126/SCIENCE.1258367/SUPPL_FILE/FAN-SM.PDF

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392). https://doi.org/10.1126/science.aar3131

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., & Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0844-5

Franzén, O., Gan, L. M., & Björkegren, J. L. M. (2019). PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, *2019*(1). https://doi.org/10.1093/database/baz046

Fu, R., Gillen, A. E., Sheridan, R. M., Tian, C., Daya, M., Hao, Y., Hesselberth, J. R., & Riemondy, K. A. (2020). clustifyr: an R package for automated single-cell RNA sequencing

cluster classification. *F1000Research*, *9*, 223. https://doi.org/10.12688/f1000research.22969.1

Fu, Y., Huang, C., Xu, X., Gu, H., Ye, Y., Jiang, C., Qiu, Z., & Xie, X. (2015). Direct reprogramming of mouse fibroblasts into cardiomyocytes with chemical cocktails. *Cell Research*, *25*(9), 1013–1024. https://doi.org/10.1038/cr.2015.99

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., & Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, *17*(6), 669–681. https://doi.org/10.1101/GR.6339607

Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Christopher Love, J., & Shalek, A. K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods 2017 14:4*, *14*(4), 395–398. https://doi.org/10.1038/nmeth.4179

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, *53*(3), 403–411. https://doi.org/10.1038/s41588-021-00790-6

Grün, D., Muraro, M. J., Boisset, J. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., de Koning, E. J. P., & van Oudenaarden, A. (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, *19*(2), 266–277. https://doi.org/10.1016/j.stem.2016.05.010

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., & Newman, A. M. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. *Science (New York, N.Y.)*, *367*(6476), 405–411. https://doi.org/10.1126/science.aax0249

Gulick, J., Subramaniam, A., Neumann, J., & Robbins, J. (1991). Isolation and characterization of the mouse cardiac myosin heavy chain genes. *Journal of Biological Chemistry*, *266*(14), 9180–9185. https://doi.org/10.1016/s0021-9258(18)31568-0

Guo, C., Kong, W., Kamimoto, K., Rivera-Gonzalez, G. C., Yang, X., Kirita, Y., & Morris, S. A. (2019). CellTag Indexing: Genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1699-y

Guo, C., & Morris, S. A. (2017). Engineering cell identity: establishing new gene regulatory and chromatin landscapes. *Current Opinion in Genetics and Development*, *46*, 50–57. https://doi.org/10.1016/j.gde.2017.06.011

Gurdon, J. B., Elsdale, T. R., & Fischberg, M. (1958). Sexually Mature Individuals of Xenopus Laevis from the Transplantaton of Single Somatic Nuclei. *Nature*, *182*, 64–65.

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1874-1

Han, L., Chaturvedi, P., Kishimoto, K., Koike, H., Nasr, T., Iwasawa, K., Giesbrecht, K., Witcher, P. C., Eicher, A., Haines, L., Lee, Y., Shannon, J. M., Morimoto, M., Wells, J. M., Takebe, T., & Zorn, A. M. (2020). Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis. *Nature Communications 2020 11:1*, *11*(1), 1–16. https://doi.org/10.1038/s41467-020-17968-x

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, *172*(5), 1091-1107.e17. https://doi.org/10.1016/J.CELL.2018.02.001

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B. J., Bader, G. D., Barker, R. A., Camara, P. G., Camp, J. G., Chédotal, A., Copp, A., Etchevers, H. C., Giacobini, P., Göttgens, B., Guo, G., Hupalowska, A., James, K. R., Kirby, E., Kriegstein, A., Lundeberg, J., … Zilbauer, M. (2021). A roadmap for the Human Developmental Cell Atlas. In *Nature* (Vol. 597, Issue 7875, pp. 196–205). Nature Research. https://doi.org/10.1038/s41586-021-03620-1

Hashimoto, H., Wang, Z., Garry, G. A., Malladi, V. S., Botten, G. A., Ye, W., Zhou, H., Osterwalder, M., Dickel, D. E., Visel, A., Liu, N., Bassel-Duby, R., & Olson, E. N. (2019). Cardiac Reprogramming Factors Synergistically Activate Genome-wide Cardiogenic Stage-Specific Enhancers. *Cell Stem Cell*, *25*(1), 69-86.e5. https://doi.org/10.1016/j.stem.2019.03.022

Hong, T., Watanabe, K., Ta, C. H., Villarreal-Ponce, A., Nie, Q., & Dai, X. (2015). An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. *PLoS Computational Biology*, *11*(11). https://doi.org/10.1371/journal.pcbi.1004569

Hong, T., Xing, J., Li, L., & Tyson, J. J. (2012). A simple theoretical framework for understanding heterogeneous differentiation of CD4 + T cells. *BMC Systems Biology*, *6*. https://doi.org/10.1186/1752-0509-6-66

Ieda, M., Fu, J. D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B. G., & Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, *142*(3), 375–386. https://doi.org/10.1016/j.cell.2010.07.002

Islam, S., Zeisel, A., Joost, S., la Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. https://doi.org/10.1038/nmeth.2772

Iwasaki, H., Mizuno, S. I., Arinobu, Y., Ozawa, H., Mori, Y., Shigematsu, H., Takatsu, K., Tenen, D. G., & Akashi, K. (2006). The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes and Development*, *20*(21), 3010–3021. https://doi.org/10.1101/gad.1493506

Jayawardena, T. M., Egemnazarov, B., Finch, E. A., Zhang, L., Payne, J. A., Pandya, K., Zhang, Z., Rosenberg, P., Mirotsou, M., & Dzau, V. J. (2012). *Cellular Biology MicroRNA-Mediated In Vitro and In Vivo Direct Reprogramming of Cardiac Fibroblasts to Cardiomyocytes*. https://doi.org/10.1161/CIRCRESAHA.112

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., & Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551. https://doi.org/10.1101/gr.121095.111

Jin, L., Ji, S., & Sun, A. (2013). Efficient generation of biliary epithelial cells from rabbit intrahepatic bile duct by Y-27632 and Matrigel. *In Vitro Cellular and Developmental Biology - Animal*, *49*(6), 433–439. https://doi.org/10.1007/s11626-013-9627-z

Jin, S., Maclean, A. L., Peng, T., & Nie, Q. (2018). ScEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics*, *34*(12), 2077–2086. https://doi.org/10.1093/bioinformatics/bty058

Kamimoto, K., Hoffmann, C. M., & Morris, S. A. (2020). CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *BioRxiv*. https://doi.org/10.1101/2020.02.17.947416

Kamimoto, K., Kaneko, K., Yuet, C., Kok, Y., Okada, H., Miyajima, A., & Itoh, T. (2016). *Heterogeneity and stochastic growth regulation of biliary epithelial cells dictate dynamic epithelial tissue remodeling*. https://doi.org/10.7554/eLife.15034.001

Kester, L., & van Oudenaarden, A. (2018). Single-Cell Transcriptomics Meets Lineage Tracing. In *Cell Stem Cell* (Vol. 23, Issue 2, pp. 166–179). Cell Press. https://doi.org/10.1016/j.stem.2018.04.014

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., & Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, *14*(5), 483–486. https://doi.org/10.1038/nmeth.4236

Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). Scmap: Projection of single-cell RNA-seq data across data sets. *Nature Methods*, *15*(5), 359–362. https://doi.org/10.1038/nmeth.4644

Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, *15*(5), 359–362. https://doi.org/10.1038/nmeth.4644

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. https://doi.org/10.1016/j.cell.2015.04.044

Kong, W., Biddy, B. A., Kamimoto, K., Amrute, J. M., Butka, E. G., & Morris, S. A. (2020). CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nature Protocols*, *15*(3), 750–772. https://doi.org/10.1038/s41596-019-0247-2

la Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. v. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494–498. https://doi.org/10.1038/s41586-018-0414-6

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., & Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, *9*(1), 1366. https://doi.org/10.1038/s41467-018-03751-6

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., & Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, *9*(1), 1366. https://doi.org/10.1038/s41467-018-03751-6

Ladewig, J., Koch, P., & Brüstle, O. (2013). Leveling Waddington: The emergence of direct programming and the loss of cell fate hierarchies. In *Nature Reviews Molecular Cell Biology* (Vol. 14, Issue 4, pp. 225–236). https://doi.org/10.1038/nrm3543

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland,

J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature 2001 409:6822*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lara-Ramírez, R., Zieger, E., & Schubert, M. (2013). Retinoic acid signaling in spinal cord development. In *International Journal of Biochemistry and Cell Biology* (Vol. 45, Issue 7, pp. 1302–1313). https://doi.org/10.1016/j.biocel.2013.04.002

Lewis, P. L., Su, J., Yan, M., Meng, F., Glaser, S. S., Alpini, G. D., Green, R. M., Sosa-Pineda, B., & Shah, R. N. (2018). Complex bile duct network formation within liver decellularized extracellular matrix hydrogels. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-30433-6

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X., & Zhang, Z. (2020). SciBet as a portable and fast single cell type identifier. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15523-2

Li, X., & Wang, C. Y. (2021). From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science 2021 13:1*, *13*(1), 1–6. https://doi.org/10.1038/s41368-021-00146-0

Li, Y., Kong, W., Yang, W., Patel, R. M., Casey, E. B., Okeyo-Owuor, T., White, J. M., Porter, S. N., Morris, S. A., & Magee, J. A. (2020). Single-Cell Analysis of Neonatal HSC Ontogeny Reveals Gradual and Uncoordinated Transcriptional Reprogramming that Begins before Birth. *Cell Stem Cell*, *27*(5), 732-747.e7. https://doi.org/10.1016/j.stem.2020.08.001

Li, Y., Li, Z., Zhang, C., Li, P., Wu, Y., Wang, C., Lau, W. B., Ma, X. L., & Du, J. (2017). Cardiac fibroblast-specific activating transcription factor 3 protects against heart failure by suppressing MAP2K3-p38 signaling. *Circulation*, *135*(21), 2041–2057. https://doi.org/10.1161/CIRCULATIONAHA.116.024599

Lian, X., Hsiao, C., Wilson, G., Zhu, K., Hazeltine, L. B., Azarin, S. M., Raval, K. K., Zhang, J., Kamp, T. J., & Palecek, S. P. (2012). Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(27). https://doi.org/10.1073/pnas.1200250109

Lieberman, Y., Rokach, L., & Shay, T. (2018). CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE*, *13*(10). https://doi.org/10.1371/journal.pone.0205499

Lin, Y., Cao, Y., Kim, H. J., Salim, A., Speed, T. P., Lin, D. M., Yang, P., & Yang, J. Y. H. (2020). scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular Systems Biology*, *16*(6). https://doi.org/10.15252/msb.20199389

Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C. L., Lindberg, E. L., Kanda, M., Polanski, K., Heinig, M., Lee, M., Nadelmann, E. R., Roberts, K., Tuck, L., Fasouli, E. S., DeLaughter, D. M., McDonough, B., Wakimoto, H., Gorham, J. M., Samari, S., … Teichmann, S. A. (2020). Cells of the adult human heart. *Nature*, *588*(7838), 466–472. https://doi.org/10.1038/s41586-020-2797-4

Liu, Z., Wang, L., Welch, J. D., Ma, H., Zhou, Y., Vaseghi, H. R., Yu, S., Wall, J. B., Alimohamadi, S., Zheng, M., Yin, C., Shen, W., Prins, J. F., Liu, J., & Qian, L. (2017). Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature*, *551*(7678), 100–104. https://doi.org/10.1038/nature24454

Loske, J., Röhmel, J., Lukassen, S., Stricker, S., Magalhães, V. G., Liebig, J., Chua, R. L., Thürmann, L., Messingschlager, M., Seegebarth, A., Timmermann, B., Klages, S., Ralser, M., Sawitzki, B., Sander, L. E., Corman, V. M., Conrad, C., Laudi, S., Binder, M., … Lehmann, I. (2021). Pre-activated antiviral innate immunity in the upper airways controls early SARS-CoV-2 infection in children. *Nature Biotechnology 2021*, 1–6. https://doi.org/10.1038/s41587-021-01037-9

Lun, A. T. L., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0947-7

Ma, F., & Pellegrini, M. (2020). ACTINN: Automated identification of cell types in single cell RNA sequencing. *Bioinformatics*, *36*(2), 533–538. https://doi.org/10.1093/bioinformatics/btz592

MacLean, A. L., Hong, T., & Nie, Q. (2018). Exploring intermediate cell states through the lens of single cells. In *Current Opinion in Systems Biology* (Vol. 9, pp. 32–41). Elsevier Ltd. https://doi.org/10.1016/j.coisb.2018.02.009

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Mahmoudi, S., Mancini, E., Xu, L., Moore, A., Jahanbani, F., Hebestreit, K., Srinivasan, R., Li, X., Devarajan, K., Prélot, L., Ang, C. E., Shibuya, Y., Benayoun, B. A., Chang, A. L. S., Wernig, M., Wysocka, J., Longaker, M. T., Snyder, M. P., & Brunet, A. (2019). Heterogeneity in old fibroblasts is linked to variability in reprogramming and wound healing. *Nature*, *574*(7779), 553–558. https://doi.org/10.1038/s41586-019-1658-5

Mazzoni, E. O., Mahony, S., Closser, M., Morrison, C. A., Nedelec, S., Williams, D. J., An, D., Gifford, D. K., & Wichterle, H. (2013a). Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. *Nature Neuroscience*, *16*(9), 1219–1227. https://doi.org/10.1038/nn.3467

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., & Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, *347*(8), btw777. https://doi.org/10.1093/bioinformatics/btw777

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*(4), 329-337.e4. https://doi.org/10.1016/j.cels.2019.03.003

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Melms, J. C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A. F., Amin, A. D., Schapiro, D., Frangieh, C. J., Luoma, A. M., Filliol, A., Fang, Y., Ravichandran, H., Clausi, M. G., Alba, G. A., Rogava, M., Chen, S. W., … Izar, B. (2021). A molecular single-cell lung atlas of lethal COVID-19. *Nature 2021 595:7865*, *595*(7865), 114–119. https://doi.org/10.1038/s41586-021-03569-1

Mohamed, T. M. A., Stone, N. R., Berry, E. C., Radzinsky, E., Huang, Y., Pratt, K., Ang, Y. S., Yu, P., Wang, H., Tang, S., Magnitsky, S., Ding, S., Ivey, K. N., & Srivastava, D. (2017). Chemical enhancement of in vitro and in vivo direct cardiac reprogramming. *Circulation*, *135*(10), 978–995. https://doi.org/10.1161/CIRCULATIONAHA.116.024692

Moris, N., Pina, C., & Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. In *Nature Reviews Genetics* (Vol. 17, Issue 11, pp. 693–703). Nature Publishing Group. https://doi.org/10.1038/nrg.2016.98

Morris, S. A. (2019). The evolving concept of cell identity in the single cell era. *Development (Cambridge)*, *146*(12). https://doi.org/10.1242/dev.169748

Morris, S. A., Cahan, P., Li, H., Zhao, A. M., San Roman, A. K., Shivdasani, R. A., Collins, J. J., & Daley, G. Q. (2014). Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet. *Cell*, *158*(4), 889–902. https://doi.org/10.1016/j.cell.2014.07.021

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J. P., & van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, *3*(4), 385-394.e3. https://doi.org/10.1016/j.cels.2016.09.002

Nagai, R., Suzuki, T., Aizawa, K., Shindo, T., & Manabe, I. (2005). Significance of the transcription factor KLF5 in cardiovascular remodeling. *Journal of Thrombosis and Haemostasis*, 1569–1576.

Nowotschin, S., Setty, M., Kuo, Y. Y., Liu, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., Boutet, S. C., Church, D. M., Hoodless, P. A., Hadjantonakis, A. K., & Pe'er, D. (2019). The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*, *569*(7756), 361–367. https://doi.org/10.1038/s41586-019-1127-1

Ogawa, M., Ogawa, S., Bear, C. E., Ahmadi, S., Chin, S., Li, B., Grompe, M., Keller, G., Kamath, B. M., & Ghanekar, A. (2015). Directed differentiation of cholangiocytes from human pluripotent stem cells. *Nature Biotechnology*, *33*(8), 853–861. https://doi.org/10.1038/nbt.3294

Okabe, M., Tsukahara, Y., Tanaka, M., Suzuki, K., Saito, S., Kamiya, Y., Tsujimura, T., Makamura, K., & Miyajima, A. (2009). Potential hepatic stem cells reside in EpCAM+ cells of normal and injured mouse liver. *Development*, *136*(11), 1951–1960. https://doi.org/10.1242/dev.031369

Olsson, A., Venkatasubramanian, M., Chaudhri, V. K., Aronow, B. J., Salomonis, N., Singh, H., & Grimes, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, *537*(7622), 698–702. https://doi.org/10.1038/NATURE19348

Orkin, S. H., & Zon, L. I. (2008). Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. In *Cell* (Vol. 132, Issue 4, pp. 631–644). https://doi.org/10.1016/j.cell.2008.01.025

Pang, Z. P., Yang, N., Vierbuchen, T., Ostermeier, A., Fuentes, D. R., Yang, T. Q., Citri, A., Sebastiano, V., Marro, S., Südhof, T. C., & Wernig, M. (2011). Induction of human neuronal cells by defined transcription factors. In *Nature* (Vol. 476, Issue 7359, pp. 220–223). https://doi.org/10.1038/nature10202

Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. In *Computational and Structural Biotechnology Journal* (Vol. 19, pp. 961–969). Elsevier B.V. https://doi.org/10.1016/j.csbj.2021.01.015

Paszek, P., Ryan, S., Ashall, L., Sillitoe, K., Harper, C. v, Spiller, D. G., Rand, D. A., & White, M. R. H. (2010). Population robustness arising from cellular heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(25), 11644–11649. https://doi.org/10.1073/pnas.0913798107

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., … Amit, I. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, *163*(7), 1663–1677. https://doi.org/10.1016/j.cell.2015.11.013

Pepe-Mooney, B. J., Dill, M. T., Alemany, A., Ordovas-Montanes, J., Matsushita, Y., Rao, A., Sen, A., Miyazaki, M., Anakk, S., Dawson, P. A., Ono, N., Shalek, A. K., van Oudenaarden, A., & Camargo, F. D. (2019). Single-Cell Analysis of the Liver Epithelium Reveals Dynamic Heterogeneity and an Essential Role for YAP in Homeostasis and Regeneration. *Cell Stem Cell*, *25*(1), 23-38.e8. https://doi.org/10.1016/j.stem.2019.04.004

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, *49*(4), 974–997. https://doi.org/10.1016/j.csda.2004.06.015

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C., & Göttgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, *566*(7745), 490. https://doi.org/10.1038/S41586-019-0933-9

Pliner, H. A., Shendure, J., & Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nature Methods*, *16*(10), 983–986. https://doi.org/10.1038/s41592-019-0535-3

Qian, L., Berry, E. C., Fu, J. D., Ieda, M., & Srivastava, D. (2013). Reprogramming of mouse fibroblasts into cardiomyocyte-like cells in vitro. *Nature Protocols*, *8*(6), 1204–1215. https://doi.org/10.1038/nprot.2013.067

Qian, L., Huang, Y., Spencer, C. I., Foley, A., Vedantham, V., Liu, L., Conway, S. J., Fu, J. D., & Srivastava, D. (2012). In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature*, *485*(7400), 593–598. https://doi.org/10.1038/nature11044

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Yosef, N. (2017). The human cell atlas. *ELife*, *6*. https://doi.org/10.7554/ELIFE.27041

Ribes, V., le Roux, I., Rhinn, M., Schuhbaur, B., & Dollé, P. (2009). Early mouse caudal development relies on crosstalk between retinoic acid, Shh and Fgf signalling pathways. *Development*, *136*(4), 665–676. https://doi.org/10.1242/dev.016204

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., & Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182. https://doi.org/10.1126/SCIENCE.AAM8999/SUPPL_FILE/PAPV2.PDF

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, *11*(1), 22–24. https://doi.org/10.1038/nmeth.2764

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., & Petersen, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, *162*(4), 729–773. https://doi.org/10.1016/0022-2836(82)90546-0

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. https://doi.org/10.1038/nbt.3192

Savulescu, A. F., Jacobs, C., Negishi, Y., Davignon, L., & Mhlanga, M. M. (2020). Pinpointing Cell Identity in Time and Space. *Frontiers in Molecular Biosciences*, *7*. https://doi.org/10.3389/fmolb.2020.00209

Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M. B., Chen, S., Green, F., Jones, R. C., Maynard, A., Penland, L., Pisco, A. O., Sit, R. v., Stanley, G. M., Webber, J. T., … Weissman, I. L.

(2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature 2018 562:7727*, *562*(7727), 367–372. https://doi.org/10.1038/s41586-018-0590-4

Schupp, J. C., Adams, T. S., Cosme, C., Raredon, M. S. B., Yuan, Y., Omote, N., Poli, S., Chioccioli, M., Rose, K. A., Manning, E. P., Sauler, M., Deiuliis, G., Ahangari, F., Neumark, N., Habermann, A. C., Gutierrez, A. J., Bui, L. T., Lafyatis, R., Pierce, R. W., … Kaminski, N. (2021). Integrated Single-Cell Atlas of Endothelial Cells of the Human Lung. *Circulation*, 286–302. https://doi.org/10.1161/CIRCULATIONAHA.120.052318

Seiler, K. M., Waye, S. E., Kong, W., Kamimoto, K., Bajinting, A., Goo, W. H., Onufer, E. J., Courtney, C., Guo, J., Warner, B. W., & Morris, S. A. (2019). Single-Cell Analysis Reveals Regional Reprogramming During Adaptation to Massive Small Bowel Resection in Mice. *CMGH*, *8*(3), 407–426. https://doi.org/10.1016/j.jcmgh.2019.06.001

Sekiya, S., & Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature*, *475*(7356), 390–395. https://doi.org/10.1038/nature10263

Sha, Y., Haensel, D., Gutierrez, G., Du, H., Dai, X., & Nie, Q. (2019). Intermediate cell states in epithelial-to-mesenchymal transition. In *Physical Biology* (Vol. 16, Issue 2). Institute of Physics Publishing. https://doi.org/10.1088/1478-3975/aaf928

Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., & Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *IScience*, *23*(3). https://doi.org/10.1016/j.isci.2020.100882

Sharma, S., Jackson, P. G., & Makan, J. (2004). Cardiac troponins. In *Journal of Clinical Pathology* (Vol. 57, Issue 10, pp. 1025–1026). https://doi.org/10.1136/jcp.2003.015420

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature 2017 550:7676*, *550*(7676), 345–353. https://doi.org/10.1038/nature24286

Son, E. Y., Ichida, J. K., Wainger, B. J., Toma, J. S., Rafuse, V. F., Woolf, C. J., & Eggan, K. (2011). Conversion of mouse and human fibroblasts into functional spinal motor neurons. *Cell Stem Cell*, *9*(3), 205–218. https://doi.org/10.1016/j.stem.2011.07.014

Song, K., Nam, Y. J., Luo, X., Qi, X., Tan, W., Huang, G. N., Acharya, A., Smith, C. L., Tallquist, M. D., Neilson, E. G., Hill, J. A., Bassel-Duby, R., & Olson, E. N. (2012). Heart repair by reprogramming non-myocytes with cardiac transcription factors. *Nature*, *485*(7400), 599–604. https://doi.org/10.1038/nature11139

Srivastava, D., & Ieda, M. (2012). Critical factors for cardiac reprogramming. In *Circulation research* (Vol. 111, Issue 1, pp. 5–8). https://doi.org/10.1161/CIRCRESAHA.112.271452

Stein-O'Brien, G. L., Clark, B. S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., Liu, S., Qian, J., Colantuoni, C., Blackshaw, S., Goff, L. A., & Fertig, E. J. (2019). Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Systems*, *8*(5), 395-411.e8. https://doi.org/10.1016/j.cels.2019.04.004

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J. S., Wilson, N. K., Mende, N., Jardine, L., Gardner, L. C. S., … Haniffa, M. (2021). Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine 2021 27:5*, *27*(5), 904–916. https://doi.org/10.1038/s41591-021-01329-2

Stone, N. R., Gifford, C. A., Thomas, R., Pratt, K. J. B., Samse-Knapp, K., Mohamed, T. M. A., Radzinsky, E. M., Schricker, A., Ye, L., Yu, P., van Bemmel, J. G., Ivey, K. N., Pollard, K. S., & Srivastava, D. (2019). Context-Specific Transcription Factor Functions Regulate Epigenomic and Transcriptional Dynamics during Cardiac Reprogramming. *Cell Stem Cell*, *25*(1), 87-102.e9. https://doi.org/10.1016/J.STEM.2019.06.012

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., & Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-4772-0

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Sun, Q., Peng, Y., & Liu, J. (2021). A reference-free approach for cell type classification with scRNA-seq. *IScience*, *24*(8). https://doi.org/10.1016/j.isci.2021.102855

Sun, S., Zhu, J., Ma, Y., & Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1898-6

Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.-E., Botting, R. A., Stephenson, E., Engelbert, J., Tuong, Z. K., Polanski, K., Yayon, N., Xu, C., Suchanek, O., Elmentaite, R., Conde, C. D., He, P., Pritchard, S., Miah, M., … Teichmann, S. A. (2022). Mapping the developing human immune system across organs. *BioRxiv*, 2022.01.17.476665. https://doi.org/10.1101/2022.01.17.476665

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, & Principal investigators.

(2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727), 367–372. https://doi.org/10.1038/s41586-018-0590-4

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

Tan, Y., & Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems*, *9*(2), 207-213.e2. https://doi.org/10.1016/j.cels.2019.06.004

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods 2009 6:5*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tosti, L., Hang, Y., Debnath, O., Tiesmeyer, S., Trefzer, T., Steiger, K., Ten, F. W., Lukassen, S., Ballke, S., Kühl, A. A., Spieckermann, S., Bottino, R., Ishaque, N., Weichert, W., Kim, S. K., Eils, R., & Conrad, C. (2021). Single-Nucleus and In Situ RNA–Sequencing Reveal Cell Topographies in the Human Pancreas. *Gastroenterology*, *160*(4), 1330-1344.e11. https://doi.org/10.1053/J.GASTRO.2020.11.010

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-41695-z

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

Treutlein, B., Lee, Q. Y., Camp, J. G., Mall, M., Koh, W., Shariati, S. A. M., Sim, S., Neff, N. F., Skotheim, J. M., Wernig, M., & Quake, S. R. (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature*, *534*(7607), 391–395. https://doi.org/10.1038/nature18323

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. In *Journal of Machine Learning Research* (Vol. 9).

Vaseghi, H. R., Yin, C., Zhou, Y., Wang, L., Liu, J., & Qian, L. (2016). Generation of an inducible fibroblast cell line for studying direct cardiac reprogramming. *Genesis*, *54*(7), 398–406. https://doi.org/10.1002/dvg.22947

Velasco, S., Ibrahim, M. M., Kakumanu, A., Garipler, G., Aydin, B., Al-Sayegh, M. A., Hirsekorn, A., Abdul-Rahman, F., Satija, R., Ohler, U., Mahony, S., & Mazzoni, E. O. (2017). A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. *Cell Stem Cell*, *20*(2), 205-217.e8. https://doi.org/10.1016/j.stem.2016.11.006

Verhulst, S., Roskams, T., Sancho-Bru, P., & van Grunsven, L. A. (2019). Meta-Analysis of Human and Mouse Biliary Epithelial Cell Gene Profiles. *Cells*, *8*(10). https://doi.org/10.3390/cells8101117

Vierbuchen, T., Ostermeier, A., Pang, Z. P., Kokubu, Y., Südhof, T. C., & Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, *463*(7284), 1035–1041. https://doi.org/10.1038/nature08797

Waddington, C. H. (1957). The Strategy of genes. *The Strategy of Genes*, 26–168.

Wang, H., Cao, N., Spencer, C. I., Nie, B., Ma, T., Xu, T., Zhang, Y., Wang, X., Srivastava, D., & Ding, S. (2014). Small molecules enable cardiac reprogramming of mouse fibroblasts with a single factor, oct4. *Cell Reports*, *6*(5), 951–960. https://doi.org/10.1016/j.celrep.2014.01.038

Wang, H., Yang, Y., Liu, J., & Qian, L. (2021). Direct cell reprogramming: approaches, mechanisms and progress. In *Nature Reviews Molecular Cell Biology* (Vol. 22, Issue 6, pp. 410–424). Nature Research. https://doi.org/10.1038/s41580-021-00335-z

Wang, H., Yang, Y., Qian, Y., Liu, J., & Qian, L. (2022). Delineating chromatin accessibility re-patterning at single cell level during early stage of direct cardiac reprogramming. *Journal of Molecular and Cellular Cardiology*, *162*, 62–71. https://doi.org/10.1016/j.yjmcc.2021.09.002

Wang, L., Liu, Z., Yin, C., Asfour, H., Chen, O., Li, Y., Bursac, N., Liu, J., & Qian, L. (2015). Stoichiometry of Gata4, Mef2c, and Tbx5 influences the efficiency and quality of induced cardiac myocyte reprogramming. *Circulation Research*, *116*(2), 237–244. https://doi.org/10.1161/CIRCRESAHA.116.305547

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, *367*(6479). https://doi.org/10.1126/SCIENCE.AAW3381

Wichterle, H., Lieberam, I., Porter, J. A., & Jessell, T. M. (2002). Directed Differentiation of Embryonic Stem Cells into Motor Neurons teristic of specific classes of neurons, for example, mid-brain dopaminergic neurons (Kawasaki et al., 2000; Lee et al., 2000). Despite these advances, the extent to which. In *Cell* (Vol. 110). http://www.cell.com/cgi/content/full/110/3/385/DC1

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, *20*(1), 1–9. https://doi.org/10.1186/s13059-019-1663-x

Wu, C. Y., Whye, D., Mason, R. W., & Wang, W. (2012). Efficient differentiation of mouse embryonic stem cells into motor neurons. *Journal of Visualized Experiments*, *64*. https://doi.org/10.3791/3813

Wu, K., Liu, Z., Wang, H., Zhang, Y., Zhou, J., Lin, Q., Wang, Y., Duan, C., & Wang, C. (2010). Efficient isolation of cardiac stem cells from brown adipose. *Journal of Biomedicine and Biotechnology*, *2010*. https://doi.org/10.1155/2010/104296

Xie, H., Ye, M., Feng, R., & Graf, T. (2004). Stepwise Reprogramming of B Cells into Macrophages. *Cell*, *117*, 663–676.

Xie, Y., Liu, J., & Qian, L. (2022). Direct cardiac reprogramming comes of age: Recent advance and remaining challenges. In *Seminars in Cell and Developmental Biology* (Vol. 122, pp. 37–43). Elsevier Ltd. https://doi.org/10.1016/j.semcdb.2021.07.010

Yamada, Y., Wang, X. di, Yokoyama, S. I., Fukuda, N., & Takakura, N. (2006). Cardiac progenitor cells in brown adipose tissue repaired damaged myocardium. *Biochemical and Biophysical Research Communications*, *342*(2), 662–670. https://doi.org/10.1016/j.bbrc.2006.01.181

Yoo, A. S., Sun, A. X., Li, L., Shcheglovitov, A., Portmann, T., Li, Y., Lee-Messer, C., Dolmetsch, R. E., Tsien, R. W., & Crabtree, G. R. (2011). MicroRNA-mediated conversion of human fibroblasts to neurons. In *Nature* (Vol. 476, Issue 7359, pp. 228–231). https://doi.org/10.1038/nature10323

Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., Xu, L., Luo, T., Yan, H., Long, Z., Shi, A., Zhao, T., Xiao, Y., & Li, X. (2019). CancerSEA: A cancer single-cell state atlas. *Nucleic Acids Research*, *47*(D1), D900–D908. https://doi.org/10.1093/nar/gky939

Yuan, J., & Sims, P. A. (2016). An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Scientific Reports 2016 6:1*, *6*(1), 1–10. https://doi.org/10.1038/srep33883

Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, *18*(1). https://doi.org/10.1186/s13059-017-1305-0

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., Ping, Y., Li, F., Shi, A., Bai, J., Zhao, T., Li, X., & Xiao, Y. (2019). CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, *47*(D1), D721–D728. https://doi.org/10.1093/nar/gky900

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., & Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, *73*(1), 130-142.e5. https://doi.org/10.1016/J.MOLCEL.2018.10.020/ATTACHMENT/E9A876B2-7D35-4E6E-9C06-7B1692F587AF/MMC1.PDF

Zhang, Y., Zhang, F., Wang, Z., Wu, S., & Tian, W. (2022). scMAGIC: accurately annotating single cells using two rounds of reference-based classification. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkab1275

Zhang, Z., Luo, D., Zhong, X., Choi, J. H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E. W., Modrusan, Z., Seshagiri, S., Kapur, P., Hon, G. C., Brugarolas, J., & Wang, T. (2019). Scina: Semi-supervised analysis of single cells in silico. *Genes*, *10*(7). https://doi.org/10.3390/genes10070531

Zhao, Y., Londono, P., Cao, Y., Sharpe, E. J., Proenza, C., O'Rourke, R., Jones, K. L., Jeong, M. Y., Walker, L. A., Buttrick, P. M., McKinsey, T. A., & Song, K. (2015). High-efficiency reprogramming of fibroblasts into cardiomyocytes requires suppression of pro-fibrotic signalling. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9243

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*. https://doi.org/10.1038/ncomms14049

Zhou, H., Dickson, M. E., Kim, M. S., Bassel-Duby, R., & Olson, E. N. (2015). Akt1/protein kinase B enhances transcriptional reprogramming of fibroblasts to functional cardiomyocytes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(38), 11864–11869. https://doi.org/10.1073/pnas.1516237112

Zhou, P., Wang, S., Li, T., & Nie, Q. (2021). Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-25548-w

Zhou, Y., Liu, Z., Welch, J. D., Gao, X., Wang, L., Garbutt, T., Keepers, B., Ma, H., Prins, J. F., Shen, W., Liu, J., & Qian, L. (2019). Single-Cell Transcriptomic Analyses of Cell Fate Transitions during Human Cardiac Reprogramming. *Cell Stem Cell*, *25*(1), 149-164.e9. https://doi.org/10.1016/j.stem.2019.05.020

Curriculum Vitae

Wenjun Kong
kong.wenjun@wustl.edu

Personal Statement

My future goal is to become a scientist in the field of regenerative medicine. I have an extensive background in computer science and research experience in statistics. I have been in training to design and perform experiment in the laboratory in the past couple years at WashU. My current research focus is to develop computational methodologies to explore more potentials from large datasets produced in cell fate engineering. My long-term career goal is to focus on integration of computation, mathematics and biological processes using cutting edge technologies in the field of regenerative medicine and seek translational potential to help the advancement in medicine to provide better patient care.

Education and Training

Aug. 2016 – Apr. 2022	Ph.D. Candidate, Computational and Systems of Biology Department of Genetics, Department of Developmental Biology Washington University in St. Louis, St. Louis, MO Laboratory of Professor Samantha A. Morris, PhD
Aug. 2012 – May 2016	Bachelor of Science, Department of Computer Science Minor in Statistics Rose-Hulman Institute of Technology, Terre Haute, IN Graduated *magna cum laude* honor

Research Experience

Sept. 2017 – Present	Laboratory of Professor Samantha Morris, PhD. Unraveling Population Heterogeneity using Single-Cell Analysis
Feb. 2017 – May 2017	Laboratory of Professor Andrew Yoo, PhD. Multifactor investigation in direct lineage reprogramming from fibroblast to neurons
Aug. 2015 – May 2016	Advisor - Professor Eric M. Reyes *Rose-Hulman Institute of Technology, Terre Haute, IN, USA* Variable Screening in Nonlinear Models via Complete Least Squares and Distance Correlation

May 2014 – Oct. 2014	Advisor – Professor Richard Anthony and Professor David Goulet *Rose-Hulman's Team in the International Genetically Engineered Machine (iGEM) competition* Synthetic Unity: creating symbiotic relationship between yeast and *E. Coli* and building educational modules (Victor the Vector) for teaching synthetic biology.
Nov. 2013 – Oct. 2014	Advisor – Professor Galen C. Duree *Rose-Hulman Institute of Technology, Terre Haute, IN, USA* Characterization of Gaussian Beam

Industry Experience

May 2015 – Aug. 2015	Software Development Engineer Internship *Amazon Corporate LLC, Herndon, VA*

Awards and Honors

Sept. 2016	1st Place: CAUSE Undergraduate Research Project Competition (USRESP), Methodological Research Subcategory
Apr. 2016	1st Place: Predictive Analytics Competition (Predict likelihood of an establishment failing a food inspection, sponsored by Allstate), Undergraduate Mathematics Conference, Rose-Hulman Institute of Technology
Dec. 2020	Douglas Covey Graduate Student Fellowship

Services

Teaching

Jan. 2018 – May 2018	Teaching Assistant in Genomics, Washington University in St. Louis, St. Louis, MO
Sept. 2014 – May 2016	Learning Center Tutor, Rose-Hulman Institute of Technology, IN
Nov. 2012 – May 2016	Homework Hotline Tutor, Rose-Hulman Institute of Technology, IN

Services

Jun. 2018 – Present	Volunteer, SSM Health St. Mary's hospital – St. Louis
Dec. 2020 – Present	Volunteer, American Red Cross – St. Louis
Aug. 2014	Orientation Leader, Rose-Hulman Institute of Technology, IN
Oct. 2013	Volunteer – STEM Buster Tutor, Honey Creek Middle School, IN

Publications

Peer Reviewed Journal Articles

Biddy B.A., **Kong W.**, Kamimoto K., Guo C., Waye S.E., Sun T., Morris S.A., (2018) Single-cell analysis of clonal dynamics in direct lineage reprogramming: a combinatorial indexing method for lineage tracing. *Nature*.

Guo C., **Kong W.**, Kamimoto K., Rivera-Gonzalez G.C., Yang X., Kirita Y. Morris S.A., (2019) CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biology*.

Seiler K.M.**, Waye S.E.**, **Kong W.**, Kamimoto K., Bajinting A., Goo W.H., Onufer E.J., Courtney C., Guo J., Warner B.W.*, Morris S.A.*, (2019, **co-first, *co-corresponding) Single-Cell Analysis Reveals Regional Reprogramming during Adaptation to Massive small Bowel Resection in Mice. *Cellular and Molecular Gastroenterology and Hepatology*.

Kong W., Biddy B.A., Kamimoto K., Amrute J.M., Butka E.G., Morris S.A., (2020) CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nature Protocols*.

Li Y.**, **Kong W.****, Yang W., Patel R.M., Casey E.B., Okeyo-Owuor T., White J.M., Porter S.N., Morris S.A.*, Magee J.A.*, (2020, **co-first, *co-corresponding) Single-cell Analysis of Neonatal HSC Ontogeny Reveals Gradual and Uncoordinated Transcriptional Reprogramming that Begins before Birth. *Cell Stem Cell*.

Cates K.*, McCoy M.J.*, Kwon J.*, Liu Y.*, Abernathy D.G., Zhang B. Liu S., Gontarz P., Kim W.K., Chen S., **Kong W.**, Ho J.N., Burbach K.F., Gabel H.W., Morris S.A., Yoo A.S., (2020, *co-first) Deconstruction Stepwise Fate Conversion of Human Fibroblasts to Neurons by MicroRNAs. *Cell Stem Cell*.

Kong W., Fu Y.C., Holloway E.M., Garipler G., Yang X., Mazzoni E.O., Morris S.A. (2022), Capybara: A computational tool to measure cell identity and fate transitions. *Cell Stem Cell*. (Accept)

Other Journal Articles (Not peer reviewed)

Kong W., Morris S.A. (2018), Evaluation of Wu et al.: Comprehending Global and Local Structure of Single-Cell Datasets, *Cell Systems*.

Presentation

Talks

Apr. 11th, 2015	*Analysis of Power Output of a Laser Using a Bayesian Approach* Chicago Area SIAM Student Conference 2015, Illinois Institute of Technology, Chicago, IL
Apr. & Oct., 2016	*Variable Screening via Distance Correlation and Complete Least Squares* • Undergraduate Mathematics Conference 2016, Rose-Hulman Institute of Technology, Terre Haute, IN (Apr. 24th, 2016) • Electronic Undergraduate Statistics Research Conference 2016, USRESP (CAUSE), Online (Oct. 21st, 2016)
Apr. 22nd, 2021	*Capybara: A computational tool to measure cell identity and fate transition (Presented in tandem with Dr. Samantha Morris)* Stowers Research Conferences: Developmental Cell Biology virtual meeting, Online
Oct. 29th, 2021	*Capybara: A computational tool to measure cell identity and fate transition* HSG/MGG/CSB/IMSD Friday Talks, Washington University in St. louis, St. Louis, MO
Dec. 14th, 2021	*Capybara: A computational tool to measure cell identity and fate transition* Inaugural Tanaka Fellowship Symposium, Washington Univeristy in St. Louis, St. Louis, MO

Posters

Aug. 12th, 2019	*scClassifier: A Bioinformatic Tool to Assess Identity at Single-Cell Resolution* Next-Generation Genomics Conference 2019, New York University, New York City, NY
Mar. 18th, 2021	*Capybara: A computational tool to measure cell identity and fate transition (Live poster session)* Single Cell Biology | EK26, Keystone Symposia, Online