UNIVERSITY OF CALIFORNIA SAN DIEGO

Studying human development by single-cell profiling of primary human tissues and genetic perturbations of novel developmental models

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

in

Bioengineering

by

Yan Wu

Committee in charge:

Professor Kun Zhang, Chair
Professor Prashant Mali, Co-chair
Professor Trey Ideker
Professor Pablo Tamayo
Professor Sheng Zhong

2020

The Dissertation of Yan Wu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-chair

_____
Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

bp: Base Pair

cDNA: Complementary DNA

DNA: Deoxyribonucleic Acid

FACS: Fluorescence Activated Cell Sorting

HAR: Human Accelerated Region

mRNA: Messenger RNA

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

QC: Quality Control

RNA: Ribonucleic Acid

SWNE: Similarity Weighted Nonnegative Embedding

t-SNE: t-Stochastic Neighbor Embedding

UMAP: Uniform Manifold Approximation and Projection

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

SWNE_Factor_Annotatation.xlsx. Table of genes used to give each factor a biological interpretation

SWNE_Runtime_Analysis.xlsx. Runtime comparison of SWNE, t-SNE, and UMAP

Teratoma_Cell_Type_Summary.xlsx. Metadata for each teratoma cell type

Teratoma_Cluster_Identification.xlsx. Genes used to classify each teratoma cluster.

Teratoma_Metrics.xlsx. QC Metrics for each processed teratoma

Teratoma_Primer_Table.xlsx. Table of all primers used for Aim 2.

ACKNOWLEDGEMENTS

I would like to acknowledge my advisers and committee co-chairs, Dr. Kun Zhang and Dr. Prashant Mali for all of their mentorship and support during my PhD. I'm grateful that under their guidance, I had the freedom to explore, make mistakes, and learn from those mistakes. Specifically, Daniella McDonald was the co-first author on the teratoma project from the experimental side with what can only be called a heroic effort and has been extremely helpful and generous with her advice and support. Dr. Blue Lake, Dr. Song Chen, Dr. Brandon Sos, and Dr. Elizabeth Duong ran the scTHS-seq experiments and Blue Lake helped with the analysis and interpretation of much of the data. Other labmates and collaborators across the Zhang and Mali that helped with analysis, experiments, or advice were Dr. Dinh Diep, Dr. Andrew Richards, Dr. Noi Plongthongkum, Sarah Urata, Udit Parekh, Kyle Ford, Dhruva Katrekar, and Dr. Dongxin Zhao. Dongxin was a collaborator on a CRISPR knockout screening project that while unfortunately did not the produce the results we had hoped for, didlaid the groundwork for all of the future work that used CRISPR knockout screens with a scRNA-seq readout work. I'd like to acknowledge my committee members, especially Dr. Pablo Tamayo for all of his advice and guidance in adapting Onco-GPS into SWNE. I also want to acknowledge the rest of my Zhang and Mali labmates, all of whom contributed advice or words of encouragement. Science is a fundamentally collaborative field, and I couldn't have done any of this without my peers and mentors.

I want to acknowledge my friends and classmates for supporting me through this process. I don't think I could've survived five and a half years of grad school without them. When I moved to San Diego, I didn't know anyone. Now, I feel like San Diego is home, and I'm grateful to have met such a great group of people. I'd like to especially thank my partner, Elaine, for supporting me through the thesis writing process. You bring so much joy into my life, and I wish we'd met sooner.

Finally, I'd like to thank my family, specifically my parents, for all of their sacrifice, support, and guidance throughout the years. They endured hardships I could never even imagine, to bring me to this country and give me the opportunities that I've had. To Mom and Dad: every day I feel like the luckiest son in the world: I'm not sure any words could express the gratitude I feel for what you've given me.

Chapter 1, in full, is a reprint of the material as it appears in Cell Systems (Wu, Yan; Tamayo, Pablo; Zhang, Kun. 2018.). The dissertation author was the primary author of this paper.

Chapter 2, in full, is a reprint of unpublished material being prepared for submission (McDonald, Daniella*; Wu, Yan*; Dailamy, Amir; Tat, Justin; Parekh, Udit; Zhao, Dongxin; Hu, Michael; Tipps, Ann; Zhang, Kun; Mali, Prashant. 2020) The dissertation author was a primary author of this paper. Chapter 2, in full, is also a reprint of the material as it appears in Cell Systems (Parekh, Udit*; Wu, Yan*; Zhao, Dongxin; Worlikar, Atharv; Shah, Neha; Zhang, Kun; Mali, Prashant. 2018)

*These authors contributed equally

Chapter 3, in full, is a reprint of unpublished material being prepared for submission (Wu, Yan; Sos, Brandon; Chen, Song; Lake, Blue; Duong, Elizabeth; Dong, Weixiu; Limaye, Sid; Mali, Prashant; Zhang, Kun). The dissertation author was the primary author of this papers.

VITA

Bachelor of Science in Engineering, Princeton University, 2014

Doctor of Philosophy, University of California San Diego, 2020

PUBLICATIONS

Bertalan, Tom; Wu, Yan; Laing, Carlos; Kevrekidis, IG. “Coarse-Grained Descriptions of Dynamics for Networks with Both Intrinsic and Structural Heterogeneities” Frontiers in Computational Neuroscience 11(43), 2017. https://doi.org/10.3389/fncom.2017.00043

Wu, Yan; Tamayo, Pablo; Zhang, Kun. “Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding”. Cell Systems, vol. 7, 2018. https://doi.org/10.1016/j.cels.2018.10.015

Parekh, Udit*; Wu, Yan*; Zhao, Dongxin; Worlikar, Atharv; Shah, Neha; Zhang, Kun; Mali, Prashant. “Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout”. Cell Systems, vol. 7. 2018.
*These authors contributed equally

FIELDS OF STUDY

Major Field: Bioengineering

Studies in Genomics, Molecular Biology, Computational Biology, Bioinformatics, and Developmental Biology

Supervised by Professors Kun Zhang and Prashant Mali

ABSTRACT OF THE DISSERTATION

Studying human development by single-cell profiling of primary human tissues and genetic perturbations of novel developmental models

by

Yan Wu

Doctor of Philosophy in Bioengineering

University of California San Diego, 2020

Professor Kun Zhang, Chair
Professor Prashant Mali, Co-Chair

Understanding human development is critical to understanding human evolution, treating developmental disorders, and creating regenerative therapeutics. Single-cell methods have enabled the high-resolution profiling of developmental trajectories and novel organoid models have facilitated a greater understanding of difficult to access developmental periods, especially through the use of perturbation experiments. Here, we used a novel multi-lineage developmental model along with profiling of the developing human prefrontal cortex to better understand human development and evolution. First, we developed a novel visualization method, Similarity Weighted Nonnegative Embedding (SWNE), which both preserves the structure of single-cell datasets and enables key marker genes and relevant

genesets to be embedded alongside the cells. We then leveraged a novel multi-lineage developmental model, the teratoma, to study the role of key developmental genes. We conducted a pooled CRISPR knockout screen of those regulators in the teratoma with a single cell RNA-seq readout, enabling us to better understand the function of these genes across all major human lineages. Finally, we used both single-cell RNA-seq and single-cell chromatin accessibility profiling to study the role of human accelerated regions (HARs), genomic regions thought to influence human-specific evolution, in human corticogenesis. The chromatin accessibility enabled us to assess the activity of HARs in specific developmental cell types and link those HARs to genes using co-accessibility of the HARs and gene promoters, while the RNA-seq enabled us to validate the expression of those HAR-linked genes.

INTRODUCTION

Deciphering the processes that underlie how humans develop from a single-cell embryo to a multi-trillion cell person has long been a goal of developmental biologists. A greater understanding of the molecular biology of human development can help clinicians better treat or manage developmental diseases, such as Autism Spectrum Disorder, as well as offer the potential to harness the regenerative power of development for treating disease and injury, such as in repairing cardiac tissue after a heart attack (DiCicco-Bloom et al. 2006; Xin, Olson, and Bassel-Duby 2013). Additionally, understanding human specific development can shed light on the evolutionary origins of humans as a species(Arthur 2002; Levchenko et al. 2018).

Genomic assays that can shed light on the gene expression, chromatin accessibility, or transcription factor binding sites of developing tissues have shed light on the dynamic developmental processes (Y. Y. Zhu et al. 2001; de la Torre-Ubieta et al. 2018; Jolma et al. 2013). However, these methods have traditionally relied on pooling cells to obtain a bulk signal, potentially obscuring heterogeneity among individual cells. Many developmental processes exist along a continuum between progenitor and mature cell populations, requiring us to measure things like gene expression at a single-cell level (Marioni and Arendt 2017; Griffiths, Scialdone, and Marioni 2018). Thus, advances in our ability to profile gene expression and chromatin accessibility at a single-cell level have unlocked our understanding of processes such as hematopoiesis and neurogenesis (Macosko et al. 2015; Rosenberg et al. 2018; Cao et al. 2019; Buenrostro et al. 2015; B. B. Lake et al. 2017; Paul et al. 2015; Dulken et al. 2017).

These single-cell datasets are extremely high-dimensional, meaning they measure up to hundreds of thousands of variables for each cell. In single-cell RNA-seq each variable corresponds to a gene, and oftentimes tens of thousands of genes will be relevant for a given

dataset (Macosko et al. 2015). In single-cell chromatin accessibility, each variable corresponds to a genomic region, which oftentimes results in the measurement of accessibility at hundreds of thousands of genomic regions (Buenrostro et al. 2015). One approach to understanding these datasets is to visualize the relationships between cells in 2 or 3 dimensional scatterplot (Becht et al. 2018; Kobak and Berens 2019). Ideally these visualizations would maintain both the structure of the data while facilitating interpretability. While much research has been done on generating visualizations that can faithfully distinguish different cell types while preserving global structures, such as distances between cell types and trajectories, very little work has been done on ensuring that these visualizations are easily interpretable (Becht et al. 2018; Kobak and Berens 2019; Moon et al. 2019). We developed a method, Similarity Weighted Nonnegative Embedding (SWNE), which can preserve global structure at least as well as existing methods while enabling key cell type marker genes, as well as genesets to be visualized alongside the cells (Wu, Tamayo, and Zhang 2018). These marker genes and genesets serve as landmarks to guide the viewer through the single-cell dataset, allowing for greater interpretability.

Another key aspect of understanding human development is the use of model systems. Since human development itself is difficult to study in an ethical manner, model systems such as animals or cell culture models are key research tools that can be analyzed and perturbed. Animal models, however, oftentimes fail to recapitulate human specific features of development, especially brain development (Hodge et al. 2019). 2D culture models are limited in their ability to generate mature cell types due to the lack of 3D cellular context (Camp et al. 2017). And while 3D organoid models are able to more faithfully recreate human developmental cell types, they are limited to a single lineage (such as kidney or brain), while development is fundamentally a multi-lineage process (Combes et al. 2019; Velasco et al. 2019). We harnessed the multi-lineage power of the teratoma in order to

generate a model of human development that can capture all major germ layers and developmental lineages, while also creating a 3D vascularized environment where cells can mature. We demonstrated this multi-lineage functionality using a CRISPR knockout screen of key developmental regulators in teratomas, assessing the effect of the knockouts using single-cell RNA-seq.

Finally, understanding human specific development can help with understanding human evolution. One approach to dissecting the molecular biology behind human evolution is to look for Human-specific Accelerated Regions (HARs) via comparative genomics(Pääbo 2014; Levchenko et al. 2018). These are regions of the genome that are thought to be of functional importance, while also showing signs of increased mutation in the human specific branch of the evolutionary tree (Pääbo 2014; Levchenko et al. 2018). While there have been efforts to study the biological role of these HARs using organoid systems, enhancer reporter assays, and chromatin capture analysis, the cell-type specific role of these HARs in human brain development has yet to be assessed (Kanton et al. 2019; Doan et al. 2016; Capra et al. 2013; Won et al. 2019; Ryu et al. 2018). We generated a single-cell chromatin accessibility and gene expression atlas of human cortical development from week 16 and week 18 prefrontal cortex. The single-cell map of chromatin accessibility enabled us to directly assess the activity of HARs in each developmental cell type, as well as identify genes that appear to be regulated by these HARs, enabling us to dissect the cell-type specific role of these HARs during development.

CHAPTER 1. Developing a single-cell visualization method that preserves data structure and facilitates data interpretation

1.1. Introduction

Single cell gene expression profiling has enabled the quantitative analysis of many different cell types and states, including human brain cell types (Lake et al. 2016; Lake et al. 2017) and cancer cell states (Tirosh et al. 2016; Puram et al. 2017), while also enabling the reconstruction of cell state trajectories during reprogramming and development (Trapnell et al. 2014; Qiu et al. 2017; Setty et al. 2016). Recent advances in droplet based single cell RNA-seq technology (Macosko et al. 2015; Lake et al. 2017) as well as combinatorial indexing techniques (Cao et al. 2017a; Rosenberg et al. 2017) have improved throughput to the point where tens of thousands or even millions of single cells can be sequenced in a single experiment, creating an influx of single cell gene expression datasets. In response to this influx of data, computational methods have been developed for latent factor identification (Buettner et al. 2017), clustering (B. Wang et al. 2017), cell trajectory reconstruction (Qiu et al. 2017; Setty et al. 2016), and differential expression (Kharchenko, Silberstein, and Scadden 2014). However, visualization of these high dimensional datasets is critical to their interpretation, and existing visualization methods often distort properties of the data, while lacking in biological context.

A common visualization method is t-Stochastic Neighbor Embedding (t-SNE), a non-linear visualization method that tries to minimize the Kullback-Leibler (KL) divergence between the probability distribution defined in the high dimensional space and the distribution in the low dimensional space (Maaten and Hinton 2008; van der Maaten 2014). This property enables t-SNE to find local patterns in the data that other methods, such as Principal Component Analysis (PCA) (Abdi and Williams 2010) and Multidimensional Scaling (MDS)

(Kruskal 1964), cannot (Maaten and Hinton 2008). However, t-SNE often fails to accurately capture global structure in the data, such as distances between clusters, making interpreting higher order features of t-SNE plots difficult. While a recent method, UMAP, addresses the issue of capturing global structure in discrete datasets, it seems to still distort single cell gene expression trajectories (McInnes and Healy 2018).

Additionally, visualizations such as t-SNE and UMAP lack biological context, such as which genes are expressed in which cell types, requiring additional plots or tables for interpretation. Dual-tSNE creatively addressed this issue by plotting genes and samples in parallel tSNE plots, which enabled users to link gene expression in one plot to specific samples in the partner plot, and vice versa (Huisman et al. 2017). Genetically Weighted Connectivity Analysis linked gene sets to the physical connectome using spatial transcriptomics (Ganglberger et al. 2018), and Onco-GPS enabled users to embed biologically interpretable factors alongside samples (J. W. Kim et al. 2017). However, to our knowledge, there are still no methods that allow for features and samples to be embedded onto the same plot.

Here, we developed a method for visualizing high dimensional single cell gene expression datasets, Similarity Weighted Nonnegative Embedding (SWNE), which captures both local and global structure in the data, while enabling the genes and biological factors that separate the cell types and trajectories to be embedded directly onto the visualization. SWNE adapts the Onco-GPS NMF embedding framework (J. W. Kim et al. 2017) to decompose the gene expression matrix into latent factors, embeds both factors and cells in two dimensions, and smooths both the cell and factor embeddings by using a similarity matrix to ensure that cells which are close in the high dimensional space are also close in the visualization. In this way, SWNE maintains fidelity when visualizing the global and local structure of the data for both developmental trajectories and discrete cell types.

1.2. Results

1.2.1 SWNE overview and methodology

SWNE combines Nonnegative Matrix Factorization (NMF) and Shared Nearest Neighbors (SNN) networks to generate a two dimensional visualization of both genes and cells. First, SWNE uses NMF (Lee and Seung 1999; Franc, Hlaváč, and Navara 2005) to create a parts based factor decomposition of the data (**Figure 1a**). The number of factors (k) is chosen by selecting the highest *k* that results in a decrease in reconstruction error above the decrease in reconstruction error for a randomized matrix (Frigyesi and Höglund 2008) **(Methods)**. With NMF, the gene expression matrix (A) is decomposed into: (1) a *genes by factors* matrix (W), and (2) a *factors by cells* matrix (H) (**Figure 1a**). SWNE then uses the similarity matrix, specifically an SNN network (Houle et al. 2010a), to smooth the H matrix, resulting in a new matrix H_{smoo} . SWNE calculates the pairwise distances between the rows of the H_{smoo} matrix, and uses Sammon mapping (Sammon 1969) to project the distance matrix onto two dimensions (**Figure 1a**). Next, SWNE embeds cells relative to the factors using the cell scores in the unsmoothed H matrix, and embeds genes relative to the factors using the gene loadings W matrix. Finally, SWNE uses the SNN network to smooth the cell coordinates so that cells which are close in the high dimensional space are close in the visualization (**Figure 1a**).

1.2.2 SWNE faithfully captures local and global structure in simulated datasets

To benchmark SWNE against t-SNE, UMAP, and other visualization methods, we used the Splatter single-cell RNA-seq simulation method (Zappia, Phipson, and Oshlack 2017) to generate two synthetic datasets. We generated a 2700 cell dataset with five discrete groups, where Groups 2 – 4 were relatively close and Groups 1 & 5 were further apart

(**Figure 1b**). We also generated a simulated branching trajectory dataset with 2730 cells and four different paths, where Path 1 branches into Paths 2 & 3, and Path 4 continues from Path 3 (**Figure 1b**).

For the discrete simulation, the t-SNE plot qualitatively distorts the cluster distances, making Groups 1 & 5 closer than they should be to Groups 2 – 4 **(Figure 1c**). The SWNE and UMAP plots both accurately show that Groups 1 & 5 are far from each other and Groups 2 – 4, while still separating Groups 2 – 4 **(Figure 1c)**. PCA, LLE, and MDS do a better job of accurately visualizing cluster distances, but have trouble visually separating Groups 2 – 4 **(Figure S1a)**. For the branching trajectory simulation, the t-SNE and UMAP plots incorrectly expand the background variance of the paths, while the SWNE plot does a better job of capturing the important axes of variance, resulting in more clearly defined paths **(Figure 1d)**. PCA, LLE, and MDS again do a better job of capturing the trajectory-like structure of the data, but still expand the background variance more than SWNE **(Figure S1b)**.

To quantitatively benchmark the visualizations, we developed metrics to quantify how well each embedding captures both the global and local structure of the original dataset. For the discrete simulation, we calculated the pairwise distances between the group centroids in the original gene expression space, and then correlated those distances with the pairwise distances in the 2D embedding space to evaluate the embeddings' ability to capture global structure **(Methods)**. To evaluate local structure, we calculated the average Silhouette score (Rousseeuw 1987), a measure of how well the groups are separated, for each embedding **(Methods)**. For maintaining global structure, SWNE outperforms t-SNE, performs similarly to UMAP, and performs about as well as PCA, MDS, and Diffusion Maps **(Figure 1e)**. SWNE also outperforms every other method, including t-SNE and UMAP, in cluster separation **(Figure 1e)**.

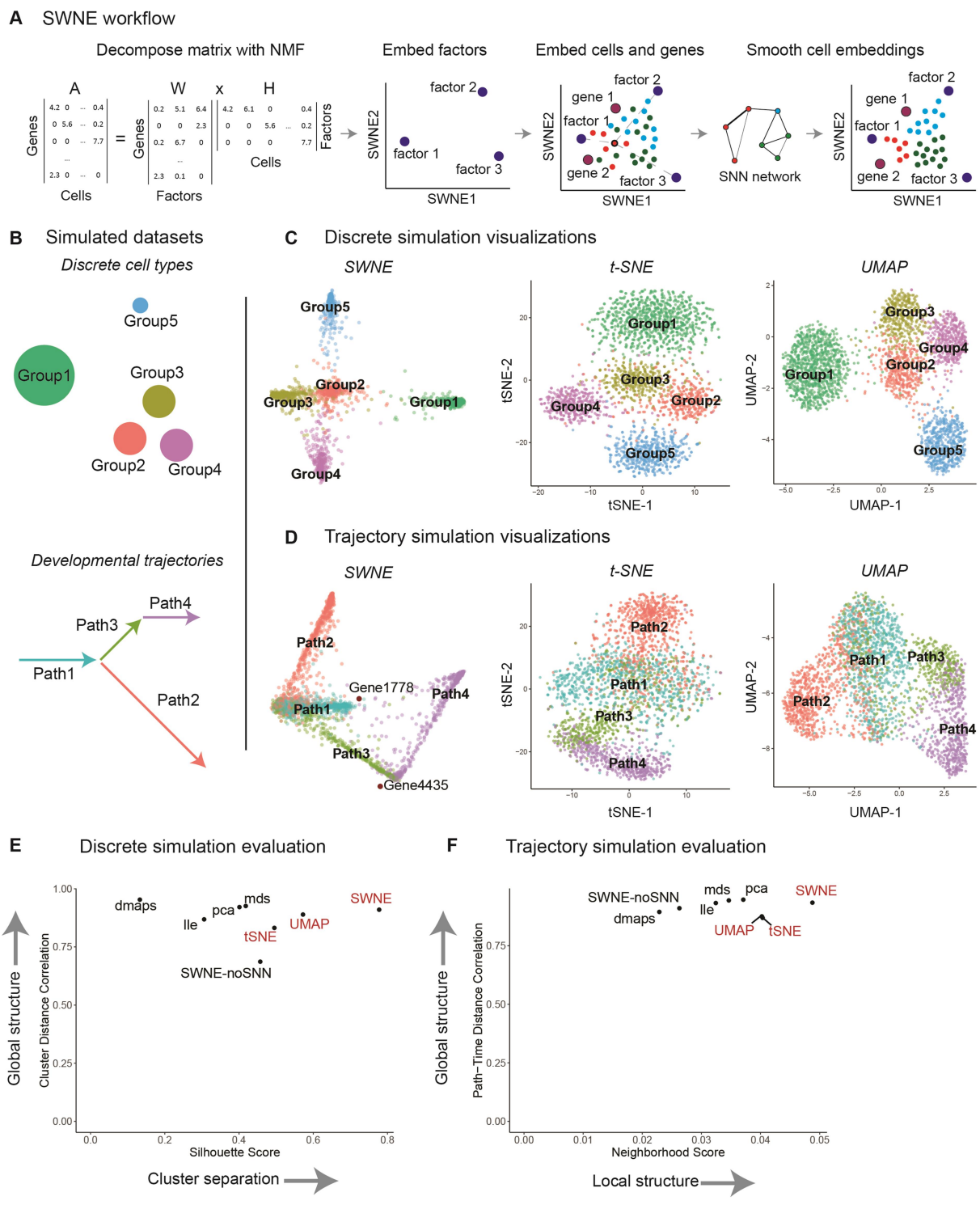
For the trajectory simulation, since we know the simulated pseudotime for each cell, we divide each path into groups of cells that are temporally close **(Methods)**. We then evaluate global structure by calculating pairwise distances between each path-time-group in the original gene expression space and the 2D embedding space, and then correlating those distances **(Methods)**. We can evaluate local structure by constructing a ground truth neighbor network by connecting cells from adjacent pseudotimes, and then computing the Jaccard similarity between each cell's ground truth neighborhood matches and its 2D embedding neighborhood **(Methods)**. SWNE outperforms t-SNE and UMAP in capturing global structure, and performs about as well as PCA, MDS, and LLE **(Figure 1f)**. For capturing neighborhood structure, SWNE again outperforms every other embedding, including t-SNE and UMAP **(Figure 1f)**. Finally, both the qualitative and quantitative benchmarks show that SNN smoothing of the cell and factor embeddings is critical to SWNE's performance, especially for capturing local structure in the data **(Figure 1e2b, 1f, Figure S1a, S1b)**.

We assessed how changing the number of factors affects both the quantitative of qualitative performance of SWNE on the trajectory and discrete simulated datasets. Visually, using too few factors results in sub-optimal cluster separation, while using too many factors results in only a minor decrease in visualization quality **(Figure S1c, S1d)**. The quantitative performance of SWNE is fairly robust across the number of factors used, although again there is more of a penalty for using too few factors than too many **(Figure S1e, S1f)**.

Additionally, we assessed SWNE's runtime alongside UMAP and t-SNE on simulated datasets. It seems like SWNE scales linearly with the number of samples, and visualizes 50,000 cells using the top 3,000 over-dispersed genes in about 8 minutes **(Supplementary Files)**. In comparison, t-SNE and UMAP visualize the same dataset, using the top 40

principal components as input, in about 8 minutes and 2 minutes respectively **(Supplementary Files)**.

Figure 1: SWNE overview and ability to capture local and global structure in simulated datasets. **(a)** The gene expression matrix (A) is decomposed into a gene loadings matrix (W) and a factor matrix (H) using NMF, selecting the number of factors by taking the highest number of factors that still results in a reduction in reconstruction error above noise(Frigyesi and Höglund 2008) (Methods). The factor matrix (H) is smoothed using the SNN network, and factors (rows of H) are embedded in 2 dimensions via Sammon mapping of their pairwise distances. Cells are embedded relative to the factors using the cell scores matrix (H), and selected genes are embedded relative to the factors using the gene loadings matrix (W). Finally, the cell embeddings are refined using the SNN network. **(b)** Simulating a discrete dataset with five clusters, and a branching trajectory dataset with four paths. **(c)** SWNE, t-SNE, and UMAP plots of the simulated discrete dataset (see Figure S1e for additional plots). **(d)** SWNE, t-SNE, and UMAP plots of the simulated trajectory dataset (see **Figure S1f** for additional plots). **(e)** Quantitative evaluation of SWNE and existing visualization methods on the discrete simulation. Global structure is evaluated by correlating pairwise cluster distances in the embedding with distances in the gene expression space. Cluster separation is evaluated with the Silhouette score. **(f)** Quantitative evaluation of SWNE and existing visualization methods on the trajectory simulation. Global structure is evaluated by dividing each path up into time steps, and correlating pairwise path-time-step distances in the embedding with distances in the gene expression space. Local structure is evaluated by taking the Jaccard similarity of the nearest neighbors in the embeddings with the true nearest neighbors.


A SWNE workflow
Decompose matrix with NMF
Embed factors
Embed cells and genes
Smooth cell embeddings
Genes
Cells
Factors
SWNE1
SWNE2
factor 1
factor 2
factor 3
gene 1
gene 2
SNN network
B Simulated datasets
Discrete cell types
Group1
Group2
Group3
Group4
Group5
Developmental trajectories
Path1
Path2
Path3
Path4
C Discrete simulation visualizations
SWNE
t-SNE
UMAP
tSNE-1
tSNE-2
UMAP-1
UMAP-2
D Trajectory simulation visualizations
Gene1778
Gene4435
E Discrete simulation evaluation
Global structure
Cluster Distance Correlation
Silhouette Score
Cluster separation
dmaps
lle
pca
mds
SWNE
UMAP
tSNE
SWNE-noSNN
F Trajectory simulation evaluation
Path−Time Distance Correlation
Neighborhood Score
Local structure

1.2.3 Illuminating the branching structure of hematopoiesis

We then applied SWNE to analyze the single cell gene expression profiles of hematopoietic cells at various stages of differentiation(Paul et al. 2015). Briefly, single cells were sorted from bone marrow and their mRNA was sequenced with single cell RNA-seq(Paul et al. 2015) (**Figure 2a**). The differentiation trajectories of these cells were reconstructed using Monocle2(Qiu et al. 2017), a method built to identify branching trajectories and order cells according to their differentiation status, or "pseudotime" (**Figure 2a**). The branched differentiation trajectories are shown in the tree in **Figure 2a**, starting from the monocyte and erythrocyte progenitors (MP/EP) and either moving to the erythrocyte (Ery) branch on the right, or the various monocyte cell types on the left(Qiu et al. 2017). We selected the number of factors for SWNE using our error reduction above noise selection method (**Figure S2a, S2b**, **Methods**).

We benchmarked SWNE performance on the hematopoiesis dataset using the same metrics we applied to the simulated trajectory dataset. To evaluate global structure, we divided the cell type clusters into groups that are temporally close according to their Monocle2 pseudotime, and then correlated pairwise distances between each cluster-pseudotime-group in the original gene expression space with distances in the 2D embedding space **(Methods)**. We evaluated local structure by computing the Jaccard similarity between each cell's neighborhood in the gene expression space and its neighborhood in the embedding space **(Methods)**. SWNE outperforms t-SNE and UMAP, as well as other embedding methods, when it comes to maintaining global structure in the dataset **(Figure 2b)**. SWNE performs about as well as UMAP in capturing neighborhood structure, and is slightly out-performed by t-SNE **(Figure 2b)**.

Qualitatively, the SWNE plot does a better job of capturing the two dominant branches: erythrocyte and the monocyte, and shows that those two branches are the primary

axes of variation in this dataset (**Figure 2c**). While the t-SNE plot captures the correct orientation of the cell types, it disproportionately expands the more differentiated cell types, obfuscating the branch-like structure of the data (**Figure 2c**). The UMAP plot also disproportionately expands the mature cell types, while placing the monocyte and erythrocyte branches too far apart. Qualitatively, SWNE, t-SNE and UMAP seem to all visually separate the cell types well. However, none of the methods accurately orient the different monocyte cell types in the monocyte branch, most likely because the variance is dominated by the erythrocyte – monocyte split, and the extent of differentiation.

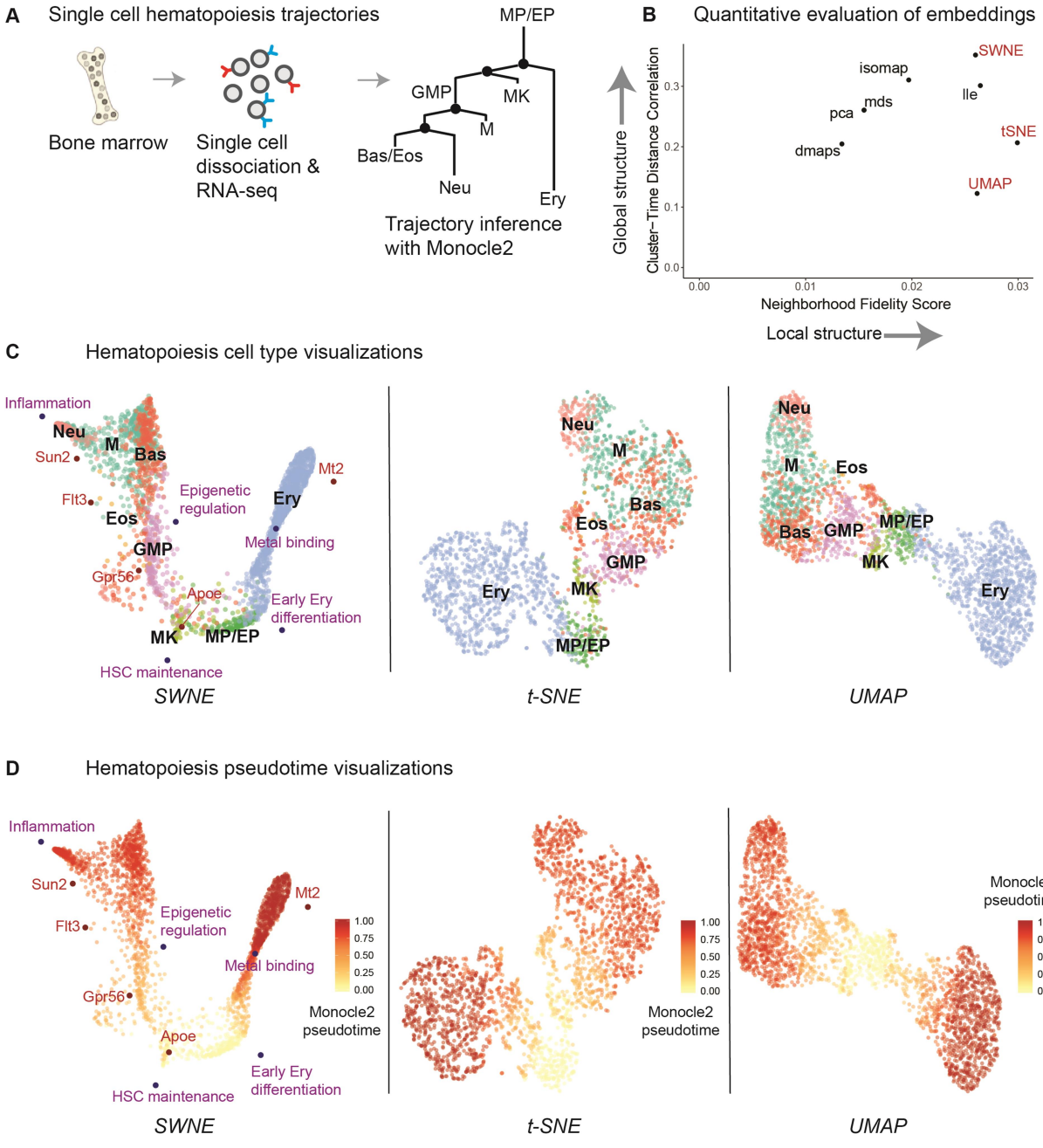
We also used Monocle2 to calculate differentiation pseudotime for the dataset, which is a metric that orders cells by how far along the differentiation trajectory they are (Qiu et al. 2017). We then overlaid the pseudotime score on the SWNE, t-SNE, and UMAP plots (**Figure 2d, 2e**). In the SWNE plot, there is a clear gradient of cells at different stages of differentiation along the two main branches (**Figure 2d**). The gradient in the t-SNE and UMAP plots is not as visible, most likely because t-SNE and UMAP obscure the branching structure by expanding the more differentiated cell types (**Figure 2e**).

Additionally, we compared the SWNE visualization with the two types of trajectory plots generated by Monocle2, which uses reversed graph embedding (RGE) to learn the underlying graph that best represents the data (Qiu et al. 2017). The Monocle2 plot of two RGE components is able to resolve the main erythrocyte and monocyte branches, but cannot visually separate the monocyte cell types **(Figure S2c)**. With ten RGE components, Monocle2's tree-based visualization can resolve the different monocyte branches **(Figure S2d)**. Nevertheless, SWNE is able to both capture the two main branches of the data while still visually separating the monocyte cell types **(Figure 3c)**. Additionally, the Monocle2 visualizations assume the data is continuous, and are specific to the Monocle2 analysis

framework, while SWNE makes no such assumptions and is meant to be used for both discrete cell types/states and continuous cellular trajectories (Qiu et al. 2017).

Furthermore, SWNE provides an intuitive framework to show how specific genes and biological factors contribute to the visual separation of cell types or trajectories by embedding factors and genes onto the visualization. We used the gene loadings matrix (W) to identify the top genes associated with each factor, as well as the top marker genes for each cell type, defined using Seurat (Butler et al. 2018; Satija, Butler, and Hoffman 2018) (**Methods, Supplementary Files**). We chose five factors and five genes that we found biologically relevant (**Figure 4a, 4c, Supplementary Files**). The genes are: *Apoe, Flt3, Mt2, Sun2, and Gpr56*. The factors are: Inflammation, Epigenetic regulation, Metal binding, HSC maintenance, and Early erythrocyte differentiation, and factor names were determined from the top genes associated with each factor (**Supplementary Files**). These factors and genes enable the association of biological processes and genes with the cell types and trajectories shown in the data visualization. For example, erythrocytes (Ery) are associated with metal binding and express *Mt2*, a key metal binding protein, while neutrophils (Neu) are associated with inflammation (**Figure 2c**). Additionally, the embedded factors and genes allow for interpretation of the overall differentiation process (**Figure 2d**). Undifferentiated progenitors (MP/EP) express *Apoe*, granulocyte-monocyte progenitors (GMP) express *Flt3*, while more differentiated neutrophils (Neu) express *Sun2* (**Figure 2d**).

Figure 2: Illuminating the branching structure of hematopoiesis. (a) Paul et al sorted single hematopoietic cells from bone marrow, sequenced them with single cell RNA-seq (Mars-Seq), and identified the relevant cell types. The hematopoiesis trajectories were reconstructed using Monocle2, and the cells were ordered according to their Monocle2 differentiation pseudotime. **(b)** Quantitative evaluation of SWNE and other embeddings on the hematopoiesis dataset. Global structure is evaluated by dividing cell type clusters into groups of cells with similar pseudotime, and correlating pairwise cluster-pseudotime-group distances in the embedding with distances in the gene expression space. Local structure is evaluated by taking the Jaccard similarity of the nearest neighbors in the embeddings with the nearest neighbors in the gene expression space. **(c)** SWNE plot of the hematopoiesis dataset, with selected genes and biological factors displayed (see **Figure 4a, 4c, Supplementary Files** for gene and factor annotations), alongside the t-SNE and UMAP plots. **(d)** SWNE, t-SNE, and UMAP plots of the hematopoiesis dataset, with developmental pseudotime calculated using Monocle2 overlaid onto the plot.


A
Single cell hematopoiesis trajectories
Bone marrow
Single cell dissociation & RNA-seq
MP/EP
GMP
MK
M
Bas/Eos
Neu
Ery
Trajectory inference with Monocle2
B
Quantitative evaluation of embeddings
SWNE
isomap
lle
mds
pca
tSNE
dmaps
UMAP
Global structure
Cluster−Time Distance Correlation
0.0
0.1
0.2
0.3
0.00
0.01
0.02
0.03
Neighborhood Fidelity Score
Local structure
C
Hematopoiesis cell type visualizations
Inflammation
Neu
M
Bas
Sun2
Mt2
Flt3
Epigenetic regulation
Ery
Eos
Metal binding
GMP
Gpr56
Apoe
Early Ery differentiation
MK
MP/EP
HSC maintenance
SWNE
t-SNE
UMAP
D
Hematopoiesis pseudotime visualizations
Monocle2 pseudotime
1.00
0.75
0.50
0.25
0.00
SWNE
t-SNE
UMAP

1.2.4 Creating an interpretable map of the human visual cortex and cerebellum

We also applied SWNE to a single nucleus RNA-seq (snDrop-Seq) human brain dataset (Lake et al. 2017) from the visual cortex (13,232 cells) and the cerebellum (9,921 cells) (**Figure 3a**). Briefly, single nuclei were dissociated from the visual cortex and cerebellum of a single donor and sequenced using snDrop-Seq (**Figure 3a**) (Lake et al. 2017). Again, the number of factors for SWNE was selected using the error reduction above noise method (**Figure S2e, S2f**).

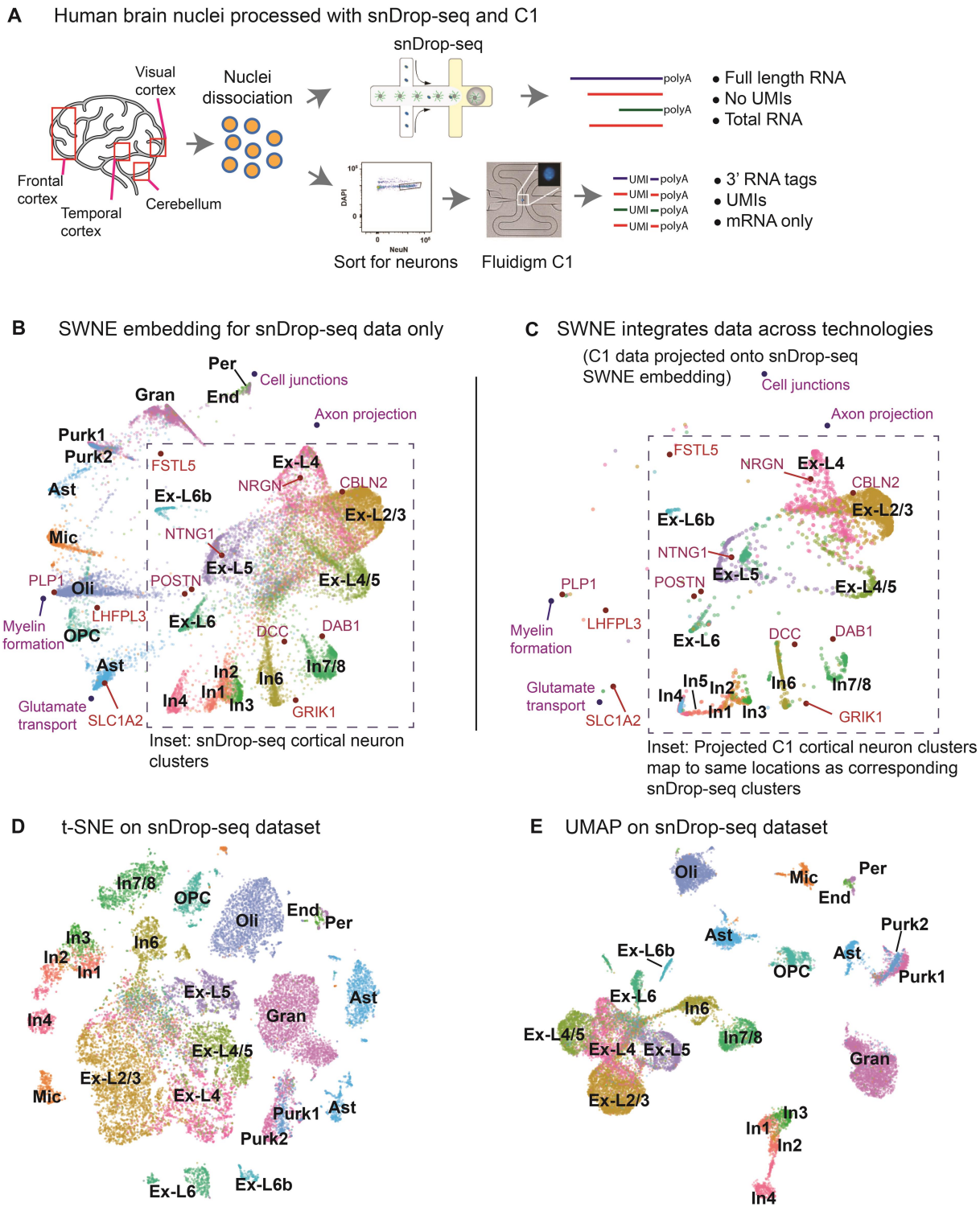
As with the hematopoiesis dataset, SWNE is able to visually separate cell types while providing an intuitive framework to visualize the contributions of specific genes and factors to that visual separation **(Figure 3b)**. We selected four factors (Myelin formation, Cell Junctions, Glutamate transport and Axon projection) and ten genes (*PLP1, GRIK1, SLC1A2, LHFPL3, CBLN2, NRGN, FSTL5, POSTN, DCC, DAB1, NTNG1*) to embed onto the SWNE plot using cell type markers and gene loadings (**Figure 4b, 4d, Supplementary Files**), adding biological context to the spatial placement of the cell types (**Figure 3b**). *CBLN2*, a gene known to be expressed in excitatory neuron types (Seigneur and Sudhof 2017), is visually close to Layer 2/3 excitatory neurons (Ex_L2/3) and *GRIK1*, a key glutamate receptor (Sander 1997), is close to inhibitory neurons (**Figure 3b, Figure 4d**). Additionally, the Myelin formation biological factor is near Oligodendrocytes (Oli), consistent with their function in creating the myelin sheath (Bunge 1968) (**Figure 3b**). The Cell junction biological factor is very close to Pericytes (Per) and Endothelial cells (End), reinforcing their functions as the linings of blood vessels, while the Axon projection factor is close to the excitatory neuron clusters, reflecting their role in transmitting action potentials (**Figure 3b**).

We also demonstrate that SWNE is able to project data across technologies by projecting a 3000-cell cortical neuron dataset, generated from a different individual, using Smart-seq+ on a Fluidigm C1 microfluidic system onto the snDrop-Seq SWNE embedding

(Figure 3a) (B. Lake et al. 2016). The Smart-seq+ protocol generates full length, total RNA without UMIs while the snDrop-Seq system generates 3' mRNA tags with UMIs **(Figure 3a)** (Lake et al. 2016; Lake et al. 2017). Despite the major differences in technologies, the cortical neuron cell types in the C1 data project onto the same locations where the corresponding cell types in the snDrop-Seq data were embedded **(Figure 3c)**. Plotting the C1 and snDrop-Seq data together shows that the technology specific batch effects are minimal **(Figure S2g)**. Thus, SWNE's ability to project new data onto existing embeddings can be used to integrate datasets across technologies and individuals.

t-SNE (**Figure 3d**) and UMAP **(Figure 3e)** are also able to visually separate the various brain cell types. Again, t-SNE seems to distort distances between cell types. For example, the Inhibitory neuron 7/8 (In7/8) cluster is equidistant from both the In6 cluster and Oligodendrocyte Progenitors (OPCs) **(Figure 3d)**. Based off of their biological functions, In7/8 and In6 should be close and both clusters should be far from OPCs. Both SWNE and UMAP are able to more accurately visualize cluster distances **(Figure 3b, 3e)**. UMAP in particular seems to generate the qualitatively cleanest visual separation between cell type clusters, while also maintaining the global structure of the data **(Figure 3e)**.

Figure 3: Creating an interpretable map of the human visual cortex and cerebellum. (a) Single nuclei were dissociated from the human cortex and cerebellum, and sequenced using both single nucleus Drop-Seq (snDrop-Seq) and the Fluidigm C1 platform (Lake et al. 2016; Lake et al. 2017). snDrop-Seq uses unique molecular indexes (UMIs), and only captures the 3' end of mRNA transcripts. The C1 method does not use UMIs and captures full length total RNA. **(b)** SWNE plot of cells from the visual cortex and cerebellum generated using snDrop-Seq, with selected genes and factors displayed (see **Figure 4b, 4d, Supplementary Files** for gene and factor annotations). **(c)** C1 data projected onto the snDrop-Seq SWNE embedding. The grey inset outlines the region where cortical neurons are embedded. **(d)** t-SNE plot of cells from the visual cortex and cerebellum generated using snDrop-Seq. **(e)** UMAP plot of cells from the visual cortex and cerebellum generated using snDrop-Seq.


A
Human brain nuclei processed with snDrop-seq and C1
Visual cortex
Frontal cortex
Temporal cortex
Cerebellum
Nuclei dissociation
snDrop-seq
polyA
polyA
• Full length RNA
• No UMIs
• Total RNA
DAPI
NeuN
Sort for neurons
Fluidigm C1
UMI polyA
UMI polyA
UMI polyA
UMI polyA
• 3' RNA tags
• UMIs
• mRNA only
B
SWNE embedding for snDrop-seq data only
Per
Cell junctions
Gran
End
Axon projection
Purk1
Purk2
FSTL5
Ex-L4
NRGN
CBLN2
Ast
Ex-L6b
Ex-L2/3
NTNG1
Mic
Ex-L5
PLP1
Oli
POSTN
Ex-L4/5
LHFPL3
Myelin formation
Ex-L6
DCC
DAB1
OPC
Ast
In2
In7/8
In6
In1
In4
In3
Glutamate transport
SLC1A2
GRIK1
Inset: snDrop-seq cortical neuron clusters
C
SWNE integrates data across technologies
(C1 data projected onto snDrop-seq SWNE embedding)
Cell junctions
Axon projection
FSTL5
NRGN
Ex-L4
CBLN2
Ex-L2/3
Ex-L6b
NTNG1
Ex-L5
POSTN
Ex-L4/5
PLP1
Myelin formation
LHFPL3
Ex-L6
DCC
DAB1
In5
In2
In6
In7/8
In4
In1
In3
Glutamate transport
SLC1A2
GRIK1
Inset: Projected C1 cortical neuron clusters map to same locations as corresponding snDrop-seq clusters
D
t-SNE on snDrop-seq dataset
In7/8
OPC
Oli
End
Per
In3
In6
In2
In1
Ex-L5
Ast
In4
Gran
Ex-L4/5
Ex-L2/3
Ex-L4
Mic
Purk1
Ast
Purk2
Ex-L6
Ex-L6b
E
UMAP on snDrop-seq dataset
Oli
Mic
Per
End
Ast
Purk2
Ex-L6b
Ast
OPC
Purk1
Ex-L6
In6
Ex-L4/5
In7/8
Ex-L4
Ex-L5
Gran
Ex-L2/3
In3
In1
In2
In4

1.2.5 Validating and assessing gene embeddings

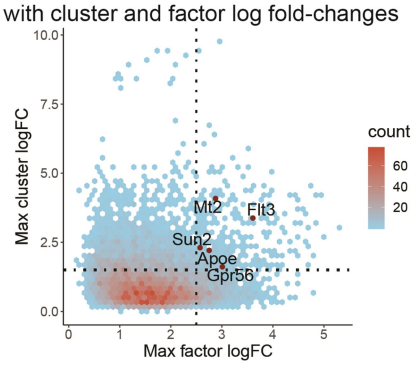
To check if the embedded genes in the hematopoiesis and human brain datasets are indeed informative, we plotted the top cluster log fold-change vs top factor loading log fold-change for each gene **(Figure 4a-b)**. Genes with high cluster specific expression are more likely to be biologically relevant, and genes that have high factor loading specificity are more likely to be visually informative. The genes we chose to embed for both datasets fell above the cluster and factor log fold-change cutoffs **(Figure 4a-b)**. Additionally, we generated cell type expression heatmaps for the embedded genes to show in which cell type(s) each embedded gene is expressed **(Figure 4c-d)**.

We also evaluated where differentially expressed (DE) genes and non-differentially expressed (non-DE) genes would be embedded. To start we looked at examples of where DE and non-DE genes would embed. We picked the DE genes and non-DE genes by ranking genes in each dataset by the average of the cluster log fold-change and the factor log fold-change, and picking genes from the top and bottom of the list. For the hematopoiesis dataset, we chose *Apoe* as the DE gene, specific to monocotyte and erythrocyte progenitors (MP/EP), and *Snap29* as the non-DE gene, overlaying their respective expression levels onto the SWNE plot **(Figure 4e)**. *Apoe* is visually close to MP/EP cell type, while *Snap29* seems to be equidistant from all cells **(Figure 4e)**. For the human brain dataset, we chose *PLP1*, an oligodendrocyte (Oli) marker, as the DE gene, and *CADM2* as the non-DE gene. Again, *PLP1* embeds close to the cluster that expresses it, while *CADM2* embeds near the middle of the plot **(Figure 4f)**. For a more systematic evaluation of gene embedding locations, we generated heatmaps of gene embedding locations. For the hematopoiesis dataset, DE genes tend to embed near the edges of the plot, while non-DE genes mostly embed towards the center **(Figure 4g)**. For the human brain

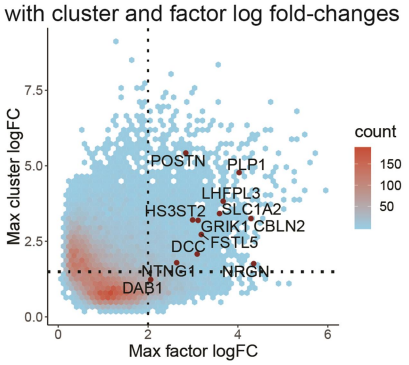
dataset, the DE genes are slightly more spread out but the non-DE genes still mostly embed near the center **(Figure 4h)**.

Figure 4: Identifying and validating gene embeddings. (a – b) Top cluster expression log fold-changes vs top factor loading log-fold changes for genes in the hematopoiesis **(a) (Figure 2)** and human brain **(b) (Figure 3)** datasets, with genes chosen for embedding labeled. Genes with both high cell type and factor log fold-changes are high quality candidates for embedding (top right quandrant). **(c – d)** Cell type specific gene expression for embedded genes in the hematopoiesis **(c) (Figure 2)** and human brain **(d) (Figure 3)** datasets. **(e)** An example of a differentially expressed gene (Apoe) and a non-differentially expressed gene (Snap29) embedded onto the hematopoiesis SWNE plot with corresponding expression overlaid **(Figure 2)**. **(f)** An example of a differentially expressed gene (PLP1) and a non-differentially expressed gene (CADM2) embedded onto the human brain SWNE plot with corresponding expression overlaid **(Figure 3)**. **(g)** Heatmap showing the locations of embedded differentially expressed and non-differentially expressed genes on the hematopoiesis SWNE embedding. **(h)** Heatmap showing the locations of embedded differentially expressed and non-differentially expressed genes on the human brain SWNE embedding.

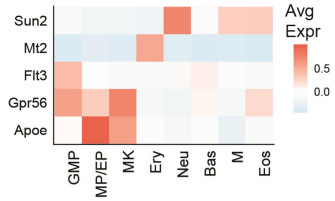
A Checking hematopoiesis gene embeddings with cluster and factor log fold-changes



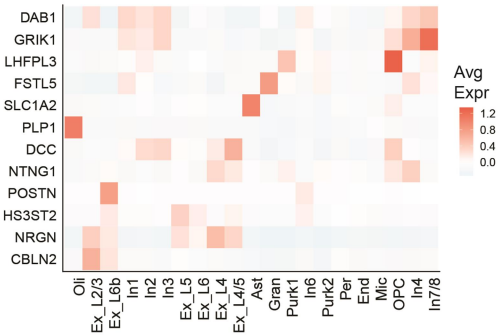
B Checking human brain cell type gene embeddings with cluster and factor log fold-changes



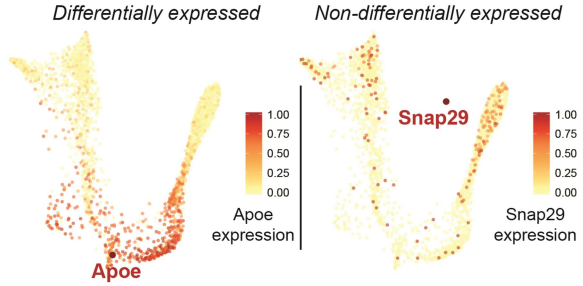
C Hematopoiesis cell type expression of embedded genes



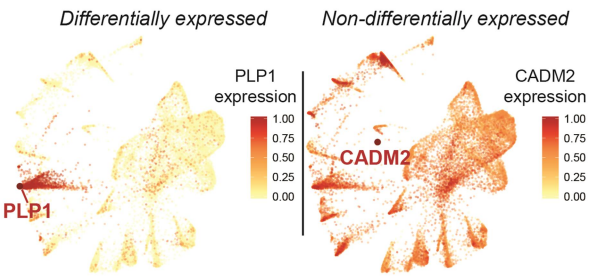
D Human brain cell type expression of embedded genes



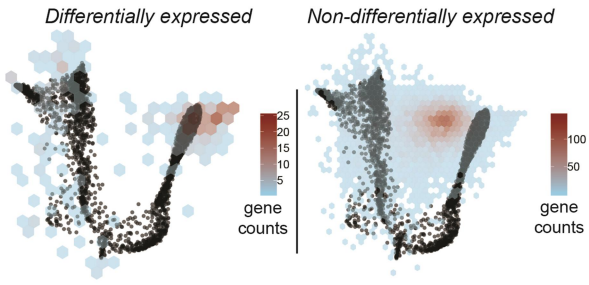
E Hematopoiesis gene embedding examples



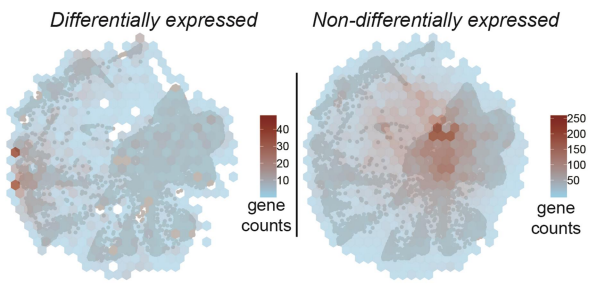
F Human brain gene embedding examples



G Hematopoiesis gene embedding heatmaps



H Human brain gene embedding heatmaps



1.3 Discussion

1.3.1 SWNE improves visualization fidelity for both continuous and discrete datasets

Interpretation and analysis of high dimensional single cell gene expression datasets often involves summarizing the expression patterns of tens of thousands of genes in two dimensions, creating a map that shows viewers properties of the data such as the number of cell states or trajectories, and how distinct cell states are from each other. However, while t-SNE, the most popular visualization method, can visualize subtle local patterns of expression that other methods cannot, it often distorts global properties of the dataset such as cluster distances and sizes **(Figure 1e, 1f)**. This is especially apparent in t-SNE visualizations of developmental datasets, as t-SNE tends to exaggerate the size of cell types instead of visualizing the axes of differentiation **(Figure 2c, 2d)**. While UMAP, a more recent visualization method, addresses these issues for discrete datasets **(Figure 3e)**, it also has limitations when visualizing continuous time-series data with developmental trajectories, and actually performs worse than t-SNE in capturing the trajectories in some cases **(Figure 2b, 2c, 2d)**.

Here, we integrated NMF with a Nearest Neighbors smoothing method to create SWNE, a visualization method that preserves global and local properties of the data for both continuous and discrete datasets. A key factor in SWNE's performance is the Shared Nearest Neighbors (SNN) network weighting. Without SNN weighting, the quantitative and qualitative performance of SWNE drops off **(Figure 1e, 1f, S1a, S1b)**. We believe SNN weighting reduces the effect of biological or technical noise, collapsing the data onto the biologically relevant components of heterogeneity. Surprisingly, this ability to minimize noise enables SWNE to capture local structure in the data better than t-SNE, and in some cases, UMAP **(Figure 1e, 1f)**. This ability to capture local structure enables SWNE to be effective at illuminating the branch-like structure in developmental trajectory datasets **(Figure 1f, 2c, 2d)**.

1.3.2 SWNE adds biological context to visualizations and projects data across technologies

Additionally, t-SNE, UMAP, and other existing methods only display cells, forcing important biological context, such as cell type marker genes, to be shown in separate plots. One of SWNE's key advantages is that the nonnegative factor embedding framework allows for embedding of genes and cells on the same visualization. The factors act as a skeleton for the data, as both cells and genes are embedded relative to these factors. The closer a group of cells is to a gene or a factor on the visualization, the more of that gene or factor the cells express **(Figure 4e, 4g)**. If one thinks of visualizations as maps, these embedded genes and factors act as landmarks, adding key biological waypoints to features of the visualization. Embedding genes and factors also streamlines the presentation of the data, eliminating the need for separate plots of marker genes or gene sets.

Batch effects in single cell RNA-seq are a well-known issue, and multiple methods have recently been developed for dataset integration (Butler et al. 2018; Haghverdi et al. 2018). SWNE's framework enables new data to be projected onto an existing SWNE embedding, which we demonstrated by projecting data generated using the Fluidigm C1 microfluidic system onto an embedding generated from snDrop-Seq **(Figure 3c)**. Despite the differences between the Fluidigm C1 and snDrop-Seq technologies, the C1 cortical neuron cell clusters map closely to the corresponding snDrop-Seq cell clusters in the embedding. Thus, SWNE's ability to project data onto existing embeddings can be used to analyze datasets across technologies or individual patient samples.

1.3.3 SWNE limitations and future work

SWNE's runtime is currently dominated by the NMF decomposition, so future work could focus on improving NMF speed, or substituting NMF with a faster matrix decomposition

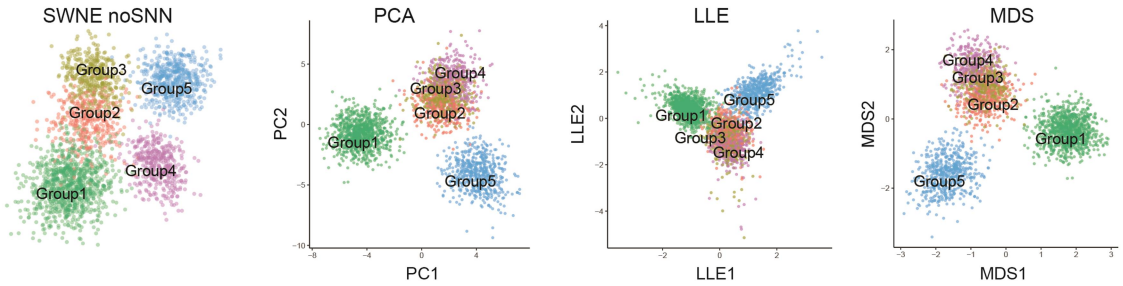
method such as f-scLVM or Pagoda/Pagoda2 (Buettner et al. 2017; Fan et al. 2016). More recent methods that enable the decomposition of heterogeneity such as Nonnegative Independent Factor Analysis (NIFA) and deep learning methods such as Single-cell Variational Inference (scVI) could also improve the interpretability and performance of the matrix decomposition (Mao et al. 2020; Lopez et al. 2018). Additionally, SNN weighting occurs sequentially after embedding the cells, factors, and genes. This causes the genes and factors to sometimes be further from cell clusters than they should be, although they are still generally closest to the most relevant cell cluster. Future work could involve developing a more elegant method that allows factor embeddings to shift relative to the cell embeddings. Finally, hyperbolic embeddings have been shown to be a very useful visualization for continuous trajectories, as they simultaneously enable a continuous modeling of lineage trees (Klimovskaia et al. 2019).

Overall, we developed a projection and visualization method, SWNE, which captures both the local and global structure of the data for continuous and discrete datasets, and enables relevant biological factors and genes to be embedded directly onto the visualization. Capturing global structure enables SWNE to address issues of distortion that occurs with t-SNE and in some cases, UMAP, creating a more accurate map of the data. Capturing local structure with the SNN network smoothing enables SWNE to accurately visualize the key axes of variation. This enables SWNE to illuminate differentiation trajectories that are not apparent in other visualizations, such as t-SNE or UMAP. Finally, embedding key marker genes and relevant biological factors adds important biological context to the SWNE visualization. As single cell gene expression datasets increase in size and scope, we believe that SWNE's ability to create an accurate, context-rich map of the datasets will enable more complete and meaningful biological interpretation.

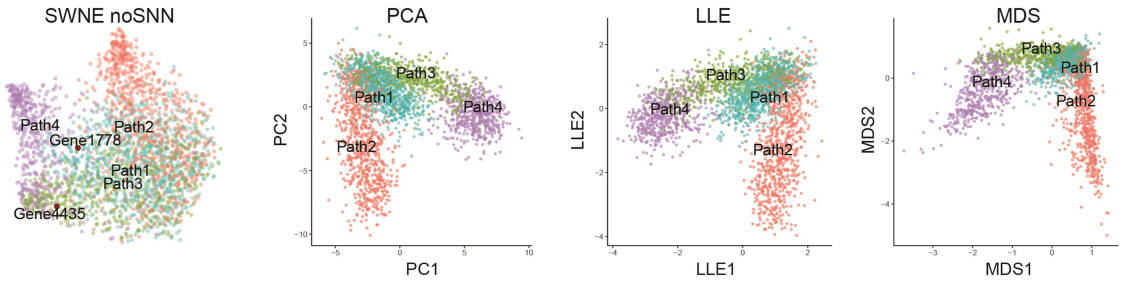
1.4 Supplemental Figures and Supplemental Table Captions

Supplementary Figure 1: SWNE model selection stability and additional visualizations of simulated datasets. Related to Figure 1. (a) Additional visualizations for the discrete simulation: SWNE without SNN weighting, PCA, locally linear embedding (LLE), multidimensional scaling (MDS). **(b)** Additional visualizations for the trajectory simulation: SWNE without SNN weighting, PCA, locally linear embedding (LLE), multidimensional scaling (MDS). **(c)** SWNE visualizations of the discrete simulation across a range of k. **(d)** SWNE visualizations of the trajectory simulation across a range of k. **(e)** Quantitative evaluation of SWNE performance across a range of k for the discrete simulation. **(f)** Quantitative evaluation of SWNE performance across a range of k for the trajectory simulation.

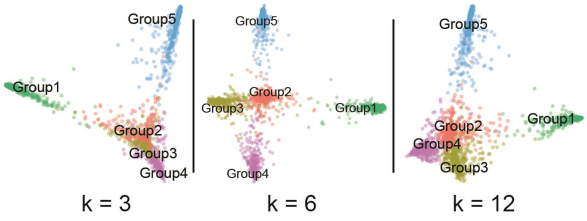
A Additional visualizations of the discrete simulation



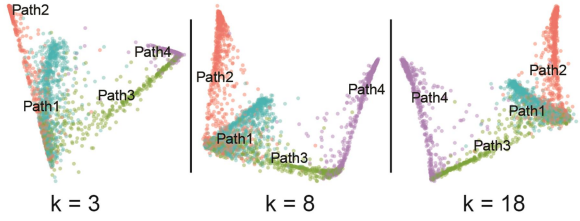
B Additional visualizations of the trajectory simulation



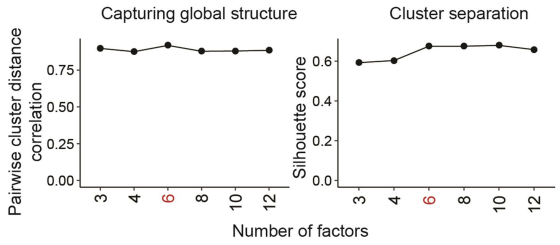
C Effect of factor selection on discrete simulation



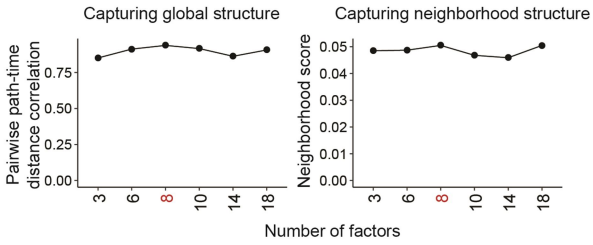
D Effect of factor selection on trajectory simulation



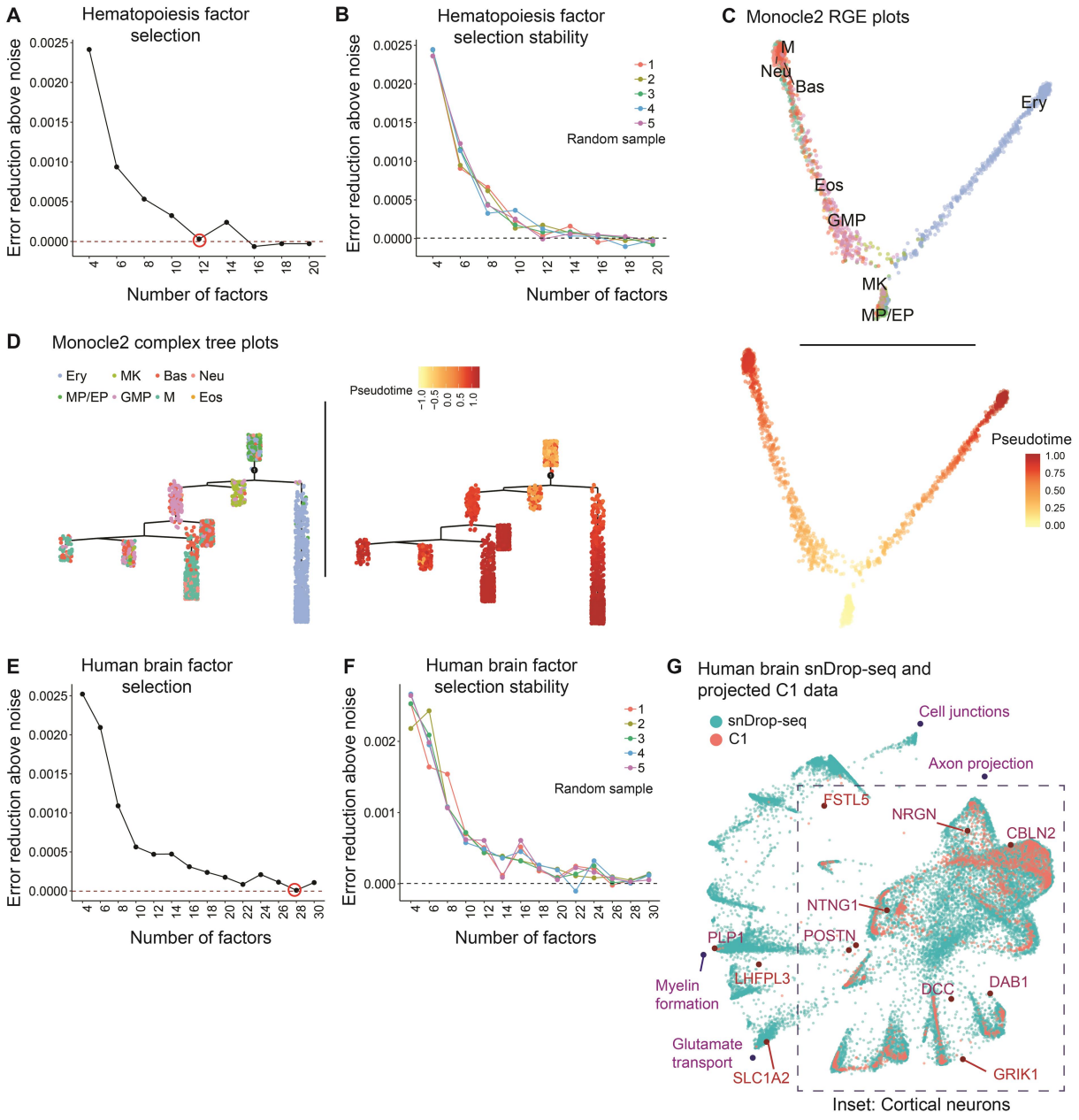
E Discrete: quantitative evaluation of factor selection



F Trajectory: quantitative evaluation of factor selection



Supplementary Figure 2: Factor selection plots and additional visualizations of the hematopoiesis and human brain datasets. Related to Figures 2 and 3. (a) Factor selection for the hematopoiesis dataset **(Figure 2)**. The optimal number of factors is when the decrease in reconstruction error above noise falls below zero **(Methods) (b)** Hematopoiesis factor selection plot across five randomizations to demonstrate stability **(Figure 2)**. **(c)** Monocle2 reversed graph embedding (RGE) plots of two RGE components with cells labeled by cell types and pseudotime **(Figure 2)**. **(d)** Monocle2 reversed graph embedding complex tree plots generated from ten RGE components with cells labeled by cell types and pseudotime **(Figure 2)**. **(e)** Factor selection for the human brain dataset **(Figure 3, Methods)**. **(f)** Hematopoiesis factor selection plot across five randomizations to demonstrate stability **(Figure 3)**. **(g)** Cortical neurons generated using the Fluidigm C1 system projected onto human brain data generated from single nucleus Drop-Seq (snDropSeq), labeled by the technology used to generate the cells **(Figure 3)**.


A
Hematopoiesis factor selection
Error reduction above noise
Number of factors
B
Hematopoiesis factor selection stability
Random sample
C
Monocle2 RGE plots
M
Neu
Bas
Ery
Eos
GMP
MK
MP/EP
D
Monocle2 complex tree plots
Ery
MK
Bas
Neu
MP/EP
GMP
M
Eos
Pseudotime
E
Human brain factor selection
F
Human brain factor selection stability
G
Human brain snDrop-seq and projected C1 data
snDrop-seq
C1
Cell junctions
Axon projection
FSTL5
NRGN
CBLN2
NTNG1
PLP1
POSTN
LHFPL3
Myelin formation
DCC
DAB1
Glutamate transport
SLC1A2
GRIK1
Inset: Cortical neurons

1.5 Materials and Methods

1.5.1 Normalization, variance adjustment, and scaling

We normalize the gene expression matrix by dividing each column (sample) by the column sum and multiplying by a scaling factor. Batch effects were normalized by a simple model, adapted from Pagoda2 (Barkas et al. 2018; Fan et al. 2016), that subtracts any batch specific expression from each gene. We used the variance adjustment method from Pagoda (Fan et al. 2016) to adjust the variance of features, an important step when dealing with RNA-seq data. Briefly, a mean-variance relationship for each feature is fit using a generalized additive model (GAM) and each feature is multiplied by a variance scaling factor calculated from the GAM fit. Feature scaling is also performed using either a log-transform, or the Freeman-Tukey transform.

1.5.2 Feature Selection

We recommend using feature selection to identify biologically relevant features/genes before running SWNE, as the NMF algorithm scales poorly with the number of features. Both Pagoda2 and Seurat offer feature selection methods that select overdispersed, and we have included an SWNE function for feature selection based off of the Pagoda2 method.

1.5.3 Nonnegative Matrix Factorization

We use the NNLM package (X. Lin and Paul C Boutros 2016) to run the Nonnegative Matrix Factorization (NMF). **Equation 1** shows the NMF decomposition:

$$A = WH \quad (1)$$

Where A is the (features x samples) data matrix, W is the (features x factors) feature loading matrix, and H is the (factors x samples) low dimensional representation of the data. The NMF initialization method can affect the embedding, and we offer an Independent

Component Analysis (ICA) initialization, a Nonnegative-SVD (NNSVD) initialization, and a purely random initialization. We have found that ICA initialization works well with most datasets, and is set as the default option. For datasets with a large number of features, ICA can be fairly slow so we use SVD as a pre-processing step for the ICA initialization.

1.5.4 Model Selection

To select the number of factors for NMF, we use the method developed by Frigyesi et al where we compare the decrease in reconstruction error for the input matrix with the decrease in reconstruction error for a randomized matrix. We take the highest number of factors such that the decrease in reconstruction error for the input matrix is still higher than the decrease in error for the randomized matrix (Frigyesi and Höglund 2008). Specifically, we calculate the reconstruction error for both the input matrix and the randomized input matrix for a range of factors. We then compute decrease in reconstruction error with an increasing number of factors (*k*) for both matrices, and subtract the decrease in error for the randomized matrix from the decrease in error for the input matrix to create an error reduction above noise metric. We select the maximum number of factors before this error reduction above noise falls below zero **(Figure S2a-b, S2e-f)**.

1.5.5 Generating the SNN matrix

In order to ensure that samples which are close to each other in the high dimensional space are close in the 2d embedding, we smooth the NMF embeddings with a Shared Nearest-Neighbors (SNN) matrix, calculated using code adapted from the Seurat package (Satija, Butler, and Hoffman 2018; Butler et al. 2018). Briefly, we calculate the approximate k-nearest neighbors for each sample using the Euclidean distance metric (in the Principal Component space. We then calculate the fraction of shared nearest neighbors between that

sample and its neighbors. We can then raise the SNN matrix, denoted here as S, to the exponent β: $S' = S^{\beta}$. If $\beta > 1$, then the effects of neighbors on the cell embedding coordinates will be decreased, and if $\beta < 1$, then the effects will be increased. Finally we normalize the SNN matrix so that each row sums up to one.

1.5.6 Weighted Factor Projection

We adapt the Onco-GPS (J. W. Kim et al. 2017) methodology to embed the NMF factors onto a two dimensional visualization. First, we smooth the H matrix with the SNN matrix using **Equation 2**:

$$H_{smooth} = H * S \quad (2)$$

We then calculate the pairwise similarities between the factors (rows of the H_{smooth} matrix) using either cosine similarity, or mutual information (J. W. Kim et al. 2016). The similarity is converted into a distance with **equation 3**:

$$D = \sqrt{2(1 - R)} \quad (3)$$

Here, R is the pairwise similarity. We use Sammon mapping(Sammon 1969) to project the distance matrix into two dimensions, which represent the x and y coordinates for each factor. The factor coordinates are rescaled to be within the range zero to one.

1.5.7 Weighted Sample Embedding

Let F_{ix}, F_{iy} represent the x and y coordinates for factor i. To embed the samples, we use the sample loadings from the *unsmoothed* H matrix via **equations 4 & 5**:

$$L_{jx} = \frac{\sum_i (H_{ij} F_{ix})^{\alpha}}{\sum_i {H_{ij}}^{\alpha}} \quad (4)$$

$$L_{jy} = \frac{\sum_i (H_{ij} F_{iy})^{\alpha}}{\sum_i {H_{ij}}^{\alpha}} \quad (5)$$

Here, j is the sample index and i is iterating over the number of factors in the decomposition (number of rows in the H matrix). The exponent α can be used to increase the "pull" of the NMF components to improve separation between sample clusters, at the cost of distorting the data. Additionally, we can choose to sum over a subset of the top factors by magnitude for a given sample, which can sometimes help reduce noise. We end up with a $2\ x\ N$ matrix of sample coordinates, L.

To weight the effects of the SNN matrix on the samples, the sample coordinates L are smoothed using **equation 6**:

$$L_{smo} = S * L \quad (6)$$

The smoothed sample coordinates (L_{smoo}) are then visualized. While we have found that an SNN matrix works well in improving the local accuracy of the embedding, other similarity matrices, such as those generated by scRNA-seq specific methods like SIMLR, could also work. In general, you should use whichever similarity or distance matrix you used for clustering.

1.5.8 Embedding features

In addition to embedding factors directly on the SWNE visualization, we can also use the gene loadings matrix (W) to embed genes onto the visualization. We simply use the W matrix to embed a gene relative to each factor, using the same method we used to embed the cells in the H matrix. If a gene has a high loading for a factor, then it will be very close to that factor in the plot, and far from factors for which the gene has zero loadings. To ensure that embedded features have both cluster specificity and contain relevant spatial information in the SWNE embedding, we plot the top cluster log fold-change against the top factor loading log fold-change for each feature, highlighting the embedded features **(Figure 4a-b)**. Any features that fall below the cluster log fold-change cutoff or the factor loading log fold-

change cutoff may not be good candidates for embedding, and SWNE will warn users if they attempt to embed those features.

1.5.9 Constructing the SNN matrix from different dimensional reductions

The SNN matrix can be constructed from either the original gene expression matrix (A), or on some type of dimensional reduction. We have found that constructing the SNN matrix from a PCA reduction tends to work well, especially in datasets where that follow a trajectory or trajectories. We believe this is due to PCA's ability to capture the axes of maximum variance, while NMF looks for a parts-based representation (Abdi and Williams 2010; Lee and Seung 1999). For datasets where there are discrete cell types, constructing the SNN matrix from the NMF factors is often similar to constructing the SNN matrix from PCA components. Thus, we default to building the SNN matrix from principal components.

1.5.10 Interpreting NMF components

In order to interpret the low dimensional factors, we look at the gene loadings matrix (W). We can find the top genes associated with each factor, in a manner similar to finding marker genes for cell clusters. Since we oftentimes only run the NMF decomposition on a subset of the overdispersed features, we can use a nonnegative linear model to project the all the genes onto the low dimensional factor matrix. One can also run Geneset Enrichment Analysis (Subramanian et al. 2005) on the gene loadings for each factor to find the top genesets associated with that factor.

1.5.11 Projecting New Data

To project new data onto an existing SWNE embedding, we first have to project the new gene expression matrix onto an existing NMF decomposition, which we can do using a simple nonnegative linear model. The new decomposition looks like **equation 7**:

$$A' = WH' \quad (7)$$

Here, A' is the new gene expression matrix, and W is the original gene loadings matrix, which are both known. Thus, we can simply solve for H'. The next step is to project the new samples onto the existing SNN matrix. We project the new samples onto the existing principal components, and then for each test sample, we calculate the *k* closest training samples. Since we already have the kNN graph for the training samples, we can calculate, for each test sample, the fraction of Shared Nearest Neighbors between the test sample and every training sample. With the test factor matrix H', and the test SNN matrix, we can run the SWNE embedding as previously described to project the new samples onto the existing SWNE visualization.

1.5.12 Generating Simulated Datasets

We used the Splatter (Zappia, Phipson, and Oshlack 2017) R package to generate a discrete dataset with five different clusters, estimating parameters from the 3k PBMC dataset published by 10X genomics. We generated five distinct clusters (groups), where Groups 1 and 5 had a differential expressed gene (DEG) probability of 0.3, while Groups 2 – 4 had a DEG probability of 0.15. Group 5 contains 1215 cells, Groups 2 – 4 contain 405 cells each, and Group 1 contains 270 cells. Thus, Groups 1 & 5 should be relatively distant and Groups 2 – 4 should be relatively close. To simulate a branching trajectory dataset, we estimated parameters from the hematopoiesis dataset from Paul et al. We generated four paths, where each path is parameterized by the number of cells in that path and the number of "time-

steps", which essentially controls how long the path is. Path 1 branches into Paths 2 and Paths 3, and Path 3 continues onto Path 4. Paths 1 & 2 contained 819 cells each, and Paths 3 & 4 contained 546 cells each. Path 1 had 100 steps, Path 2 was the "longest" path with 200 steps, and Paths 3 & 4 had 50 steps each. Each cell is assigned to a path, and a time-step. For example, Cell2522 might belong to Path1 and time-step 68.

1.5.13 Evaluating Embedding Performance

To evaluate how well each embedding maintained the global structure of the discrete simulation, we correlated the pairwise cluster distances in the 2D embedding with the pairwise cluster distances in the original gene expression space. We then calculated the average Silhouette score for each embedding, evaluating how well the visualization separates the clusters. For the trajectory simulation, we divided each path into "chunks" of five time-steps. We correlated the pairwise distances of each "path-time-chunk" in the embedding space with the pairwise distances in the gene expression space to evaluate how well the embeddings maintained the global structure. To evaluate the local structure, we constructed a "ground-truth" neighborhood graph by adding an edge between every cell in each path-time-step, and every cell in each neighboring path-time-step. For example, we would connect all the cells in Path1 at time-step 23, with all the cells in Path1 and time-step 24. We then created a nearest neighbor graph for each embedding, and took the Jaccard similarity between each cell's neighborhood in the embedding and the true neighborhood. We used the average Jaccard similarity as our "neighborhood score".

We adopted a similar approach to evaluate the hematopoiesis dataset. To quantitatively evaluate how well each embedding captured the global structure, we divided each annotated cluster into "chunks" of 50 cells by pseudotime calculated using Monocle2. We then correlated the pairwise distances of each cluster-time-chunk in the embedding

space with the pairwise distances in the gene expression space. To evaluate the local structure, we compute the overlap in the 30 nearest neighbors for each cell in the embedding space with the nearest neighbors in the gene expression space using the Jaccard similarity. We average the Jaccard similarities across all cells as our “neighborhood fidelity score”.

1.5.14 Running UMAP, t-SNE and other dimensional reduction methods

UMAP and t-SNE were run through the Seurat R package (Butler et al. 2018). We first reduced the dimensionality of the gene expression matrix with PCA, and used a variance explained elbow plot to select the number of principal components to keep. The principal components were used as inputs to UMAP and t-SNE.

Diffusion maps, Isomap, Locally Linear Embedding (LLE), and Multidimensional Scaling (MDS) were run directly on the normalized gene expression matrix. Diffusion maps was run using the Destiny R package (Angerer et al. 2015), Isomap and LLE were run with the RDRToolbox R package, while MDS was run using the cmdscale function in R. Default parameters were used in all cases unless otherwise specified.

1.5.15 Data and Software Availability

The SWNE package is available at https://github.com/yanwu2014/swne. The scripts used for this manuscript are under the Scripts directory. The data needed to recreate the figures can be found here:

- http://genome-tech.ucsd.edu/public/SWNE/hemato_data.tar.gz (Hematopoiesis data)
- http://genome-tech.ucsd.edu/public/SWNE/neuronal_data.tar.gz (Neuronal data)

The raw data for the hematopoietic and neuronal cells can be found at the GEO accessions GSE72857 and GSE97930, respectively. The PBMC dataset can be found at the 10X genomics website: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k. The simulated datasets can be found at: http://genome-tech.ucsd.edu/public/SWNE/splatter_simulated_data.tar.gz

1.6 Acknowledgement for Chapter 1

Chapter 1, in full, is a reprint of the material as it appears in Cell Systems (Wu, Yan; Tamayo, Pablo; Zhang, Kun. 2018.). The dissertation author was a primary author of this paper.

CHAPTER 2. Assessing the role of developmental regulatory genes using CRISPR knockout screens in a novel multi-lineage model of human development

2.1 Introduction

Current understanding of early human development heavily relies on inference from animal models. Experimental embryology approaches and genetic tools in model systems such as frogs(Vastag et al. 2011), fish(Farrell et al. 2018), and mice(Pijuan-Sala et al. 2019; Cao et al. 2019) have been instrumental in modeling development and have demonstrated that many features of early embryogenesis are evolutionarily conserved across species(Peter and Davidson 2011; Royo et al. 2011; Y. Lin et al. 2009). However, it is also notable that several aspects are highly species-specific(Richard et al. 2000; Richardson et al. 1997; Raff 1996). While there have been studies on human embryonic development(Y. Zhu et al. 2018; J. A. Miller et al. 2014), such studies are limited by a scarcity of relevant biological material and key ethical constraints. Thus, there has been a push to establish models specific to human development.

Human embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs), as well as other cell lines have been used as developmental models by directing differentiation of ESCs or iPSCs into various cell types. These studies have shed light on processes such as lineage bifurcation(Yao et al. 2017) and heterogeneity(J. Wang et al. 2017) during human neuronal development, as well as the presence of discrete cell states during early ESC differentiation(Jang et al. 2017). Additionally, systematic perturbation screens in these cell culture models have looked at the key regulators of differentiation(Parekh et al. 2018) and reprogramming(Tsunemoto et al. 2018). However, human development takes place in 3-

dimensions, which is difficult to capture with a 2-dimensional monolayer(Liu et al. 2018; Brown, Quadrato, and Arlotta 2018).

Newer methods for modeling human development use organoid systems. Organoids are 3D “mini-organs” derived from pluripotent stem cells (PSCs) or adult progenitors in which the cells spontaneously self-assemble into differentiated functional cell types which somewhat mimic their *in vivo* counterpart structurally and functionally(M. Huch and Koo 2015; Yin et al. 2016; Clevers 2016; Dutta, Heo, and Clevers 2017; Fligor et al. 2018; Capowski et al. 2019; Collin et al. 2019). The use of organoids has enabled researchers to model human specific development, which is especially beneficial for modeling rare genetic diseases or cancers(Bigorgne et al. 2014; Dekkers et al. 2013; D. Gao et al. 2014; Bartfeld et al. 2015; Boj et al. 2015; van de Wetering et al. 2015; M. Huch and Koo 2015; Yin et al. 2016). However, this system has limitations and challenges. Often tissue types derived are immature(Aurora and Spence 2016; Chambers, Tchieu, and Studer 2013) and limited in thickness and scale due to the absence of abundant vasculature. Additionally, most organoid models can only generate a single or few developmental lineages(Jabaudon and Lancaster 2018; Yin et al. 2016). For instance, intestinal organoids do not fully recapitulate the intestinal epithelium due to a lack of BMP signaling gradients. In addition, there is a lack of cellular diversity due to the addition of Wnt proteins, Nicotinamide, and Tgf-β inhibitors preventing differentiation of stem cells *in vitro*(Sato et al. 2009, 2011; Jung et al. 2011; Yin et al. 2016). To date, mature endocrine pancreatic tissue has been very difficult to derive properly *in vitro* due the heavy reliance on vasculature and surrounding 3D architecture for proper differentiation, which in turn also affects the ability to maintain these in culture for extended periods of time(Jacobson and Tzanakakis 2017; M. Huch and Koo 2015).

We propose here human PSC-derived teratomas as a facile yet powerful model for studying human development(Lensch et al. 2007) as: it displays extensive multi-lineage

differentiation to all germ layers; is an intrinsically vascularized 3D differentiation system; bears regions of complex tissue-like organization; and, its implementation is simple and accessible. Early teratoma research revealed that these derive from pluripotent germ cells which have a resemblance to cells embryonic in origin(L. Stevens 1967; L. C. Stevens and Pierce. 1975; L. Stevens 1962; THURLBECK, WLLIAM M. 1973). These tumors form in human patients presumably due to misguided primordial germ cell migration during embryogenesis(Nikolic et al. 2016; Saffman and Lasko 1999). PSC-derived teratomas in turn are generated by directly injecting PSCs subcutaneously into immunodeficient mice, where the cells will attach and differentiate in a semi-random fashion into all three germ layers(Willis 1934, 1935; THURLBECK, WLLIAM M. 1973; Bocker 2002), and are now the gold standard to validate pluripotency and developmental potential of ESC and iPSC lines(Smith, Luong, and Stein 2009; Avior, Biancotti, and Benvenisty 2015).

Leveraging their inherent differentiation potential, there has been some progress in the field utilizing teratomas to derive highly sought-after cell types. For instance, recently these were utilized to derive skeletal myogenic progenitors by injecting PSCs into the *tibialis anterior* muscle of mice to enrich for muscle cell types in the teratomas that formed in those muscles. Thus derived cells were then isolated and utilized in transplant and regenerative studies in mice with Duchenne muscular dystrophy(Chan et al. 2018). Additionally, some groups have successfully isolated hematopoietic stem cells (HSCs) from teratomas utilizing strategies such as HSC enrichment via human umbilical vein endothelial cell (HUVEC) pooling(Suzuki et al. 2013; Tsukada et al. 2017; Philipp et al. 2018; Amabile et al. 2019). However, the semi-random nature of teratoma development has previously made characterization of teratomas difficult, as the different lineages can often be found in close spatial proximity.

We surmised that with the advent of high-throughput single cell gene expression profiling via droplet-based methods(Rosenberg et al. 2018; Cao et al. 2017b; Zheng et al. 2017; Rosenberg et al. 2017; Macosko et al. 2015; Klein, Mazutis, Weitz, et al. 2015; Ding et al. 2019) and facile genetic perturbation toolsets such as CRISPR-Cas9 could enable us to address this challenge by enabling systematic analysis and modulation of teratomas at the single cell level(Dixit, Parnas, Li, Chen, et al. 2016; Adamson et al. 2016; Datlinger et al. 2017; Qi et al. 2013; M. Chen and Qi 2017; Akcakaya et al. 2018; Black et al. 2016; Dijk et al. 2018). Coupled with histology, and RNA *in situ* hybridization, we established a comprehensive experimental and computational framework to systematically analyze, perturb and engineer human PSC-derived teratomas.

2.2 Results

2.2.1 Teratoma Characterization

Initially, we wanted to characterize the teratoma to better understand its kinetics, architecture, and constituent cell types. Towards this we generated seven teratomas using H1 ESCs and characterized their cell types using both single cell RNA-seq and histology, with RNA FISH validation. To generate a teratoma, a subcutaneous injection of 5-10 million ESCs in a slurry of Matrigel and embryonic stem cell medium was made in the right flank of anesthetized $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice (**Figure 5A, Methods**). Weekly monitoring of teratoma growth was made by quantifying the approximate elliptical area of the tumor (mm^2) (Methods). Kinetic trajectories show an average time point of around 37 days when we can begin to outwardly see and measure tumor size. Growth continued for up to 70 days until the tumors were of a sufficient size for extraction and downstream analyses (~820 mm^2, **Figure 5B**). Post-extraction, tumors were assayed for external heterogeneity (i.e. presence of dark

pigmented regions, white tough areas, connective tissue, and vasculature) before being cut and frozen for sectioning and H&E staining (**Figure 5C, Methods**). The presence of all 3 germ layers (ectoderm, mesoderm, endoderm) was validated to confirm pluripotency and developmental potential (**Figure 5D, Methods**). Specific structures were consistently seen such as developing airways, retina, fetal cartilage and bone, muscle, vasculature, GI tract, and a predominance of connective tissue and neuroectoderm (**Figure S3A-J**). Remaining tissue was dissociated down to the single cell level for single cell RNA sequencing with the droplet-based 10X Genomics Chromium platform(Zheng et al. 2017).

Towards computational analyses, we generated a combined single cell gene expression matrix across the 7 teratomas for both human and mouse cells using the cellranger(Zheng et al. 2017) pipeline from 10X genomics (**Methods, Figure 5A**). We removed any teratoma specific batch effects by using the Seurat data integration pipeline(Stuart et al. 2018), which uses mutual nearest neighbors and canonical correlation analysis to correct for batch specific effects, while retaining any batch specific cell types (**Methods**). With this batch-corrected matrix, we clustered the cells using a shared nearest neighbors (SNN) community detection algorithm(Houle et al. 2010b), and generated a rough biological annotation of the clusters using a k-nearest neighbors classifier(Tarlow et al. 2013) trained on the Mouse Cell Atlas(Han et al. 2018). For the human clusters, we further refined the cluster annotations manually using canonical cell type markers (**Supplementary Files**). We then visualized both the human and mouse cells with Uniform Manifold Approximation and Projection (UMAP)(Becht et al. 2018) scatterplot (**Figure 5E**). In the human cells, we were able to detect 23 cell types across all three germ layers, including endodermal cell types (foregut, mid/hindgut), and an abundance of mesodermal cell types (**Figure 5E, Figure S3K, Supplementary Files**). We validated each human cell type by assessing the expression of key marker genes using a heatmap (**Figure S3L**). We also further validated

the cell types by correlating the expression of each of the teratoma cell types with the expression of cell types from the Mouse Organogenesis Cell Atlas(Cao et al. 2019), demonstrating that each teratoma cell generally correlates with at least one fetal mouse cell type (**Figure S3M**). Chondrogenic MSC, Smooth Muscle, and Kidney cells don't correlate well with any fetal mouse cell type while Lens, Hepatocytes, and Notochord cells don't correlate well with any human teratoma cell types (**Figure S3M**). In the mouse cells, we primarily observed invading immune cells, endothelial cells, and stromal cells (**Figure S3N, Supplementary Files**).

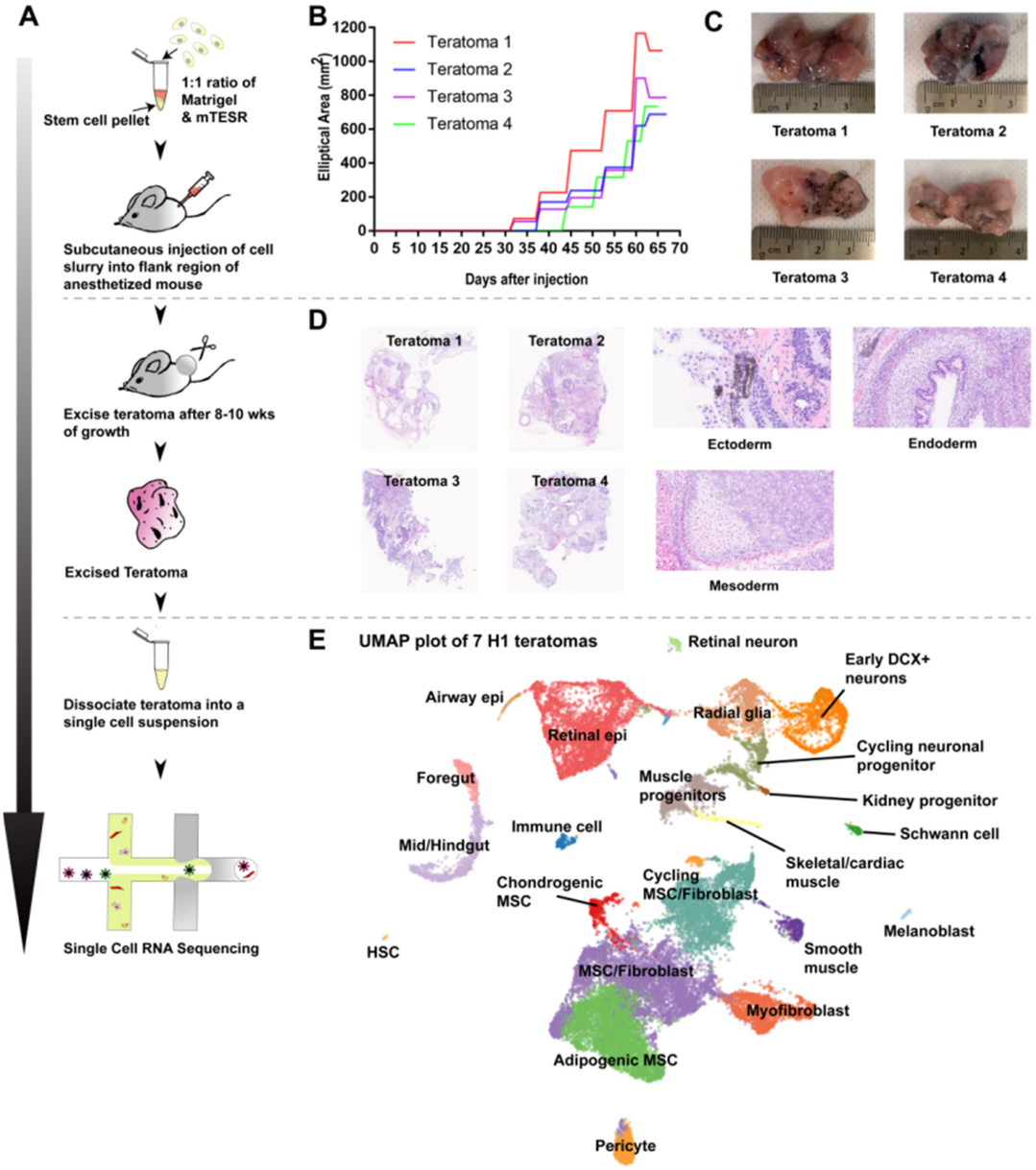


Figure 5: Comprehensive teratoma characterization. (A) Schematic of general workflow. Subcutaneous injection of H1 PSCs in a slurry of Matrigel® and embryonic stem cell medium was made in the right flank of Rag2$^{-/-}$;γc$^{-/-}$ immunodeficient mice. Weekly monitoring of teratoma growth was quantified by approximating elliptical area (mm^2). Tumors were then extracted after 8-10 wks of growth and observed for external heterogeneity before small sections were frozen for H&E staining. Remaining tumor dissociated into a single cell suspension via standard GentleMACS protocols. Single cell suspension used for scRNA-seq (10x Genomics). **(B)** Growth kinetics of four H1 teratomas. **(C)** Images of four teratomas generated from H1 cells. **(D)** H&E stains of the four teratoma histology sections. The presence of ectoderm, mesoderm, and endoderm confirmed for pluripotency and developmental potential. **(E)** UMAP visualization of cell types identified from single cell RNA-sequencing of the seven H1 teratomas

2.2.2 Teratoma Heterogeneity

Assessing heterogeneity present in and between teratomas (especially between different stem cell lines) is critical to determine repeatability and utility of this model. We generated additional teratomas (per **Figure 5A**) with the H9 and HUES62 ESC lines and the PGP1 iPSC line, and then identified how each individual cell line contributed to the cell types present (**Figure 6A**). We visualized all cell types present across all 3 additional cell lines with a UMAP scatterplot (**Figure 6B**) in addition to showing the relative contribution of each cell line to the UMAP embedding (**Figure S4A**). We also assessed the distribution of cell types represented in each individual H1 teratoma as well as the individual H9, HUES62, and PGP1 teratomas (**Figure 6C, Figure S4B**). We then compared the germ layer representation between all teratomas and used the zebrafish model for reference(Wagner et al. 2018) (**Figure 6D**). These data show that in general, most teratomas show a large amount of mesoderm and neuroectoderm, with very little endoderm (**Figure 6D**). This mesodermal predominance derives primarily from the MSC/Fibroblast contribution. This holds true with individual teratomas from the H1cell line, while teratomas from different cell lines show more variability in terms of the MSC/Fibroblast fraction (**Figure 6D**, **Figure S3K**). There is a relatively low fraction of endodermal cells in both the teratomas as well as the zebrafish and mouse embryos, suggesting that endodermal cells are simply less prevalent during development (**Figure 6D**). Qualitatively, while there is variability in cell type representation among the different teratomas, every teratoma contains most of the major cell types (**Figure 6C**). By computing the scaled mutual information between cell type assignments and teratoma assignments, we can compute a quantitative metric of heterogeneity across teratomas (**Figure 6E**)(J. W. Kim et al. 2016; Velasco et al. 2019). We can see that the cell type heterogeneity across the H1 teratomas is similar to that of the patterned brain organoids, while the teratomas generated from different cell lines have a much higher level of

heterogeneity (Figure 6E). There was only one replicate per cell line teratoma as our main goal was to assess the heterogeneity across cell lines versus the heterogeneity within the H1 cell line, while also demonstrating that we could generate teratomas using multiple cell lines (**Figure 6E, Methods**). Interestingly, line-specific kinetics were present in regard to teratoma growth with PGP1 teratomas growing the fastest and HUES62 the slowest (**Figure S4C**).

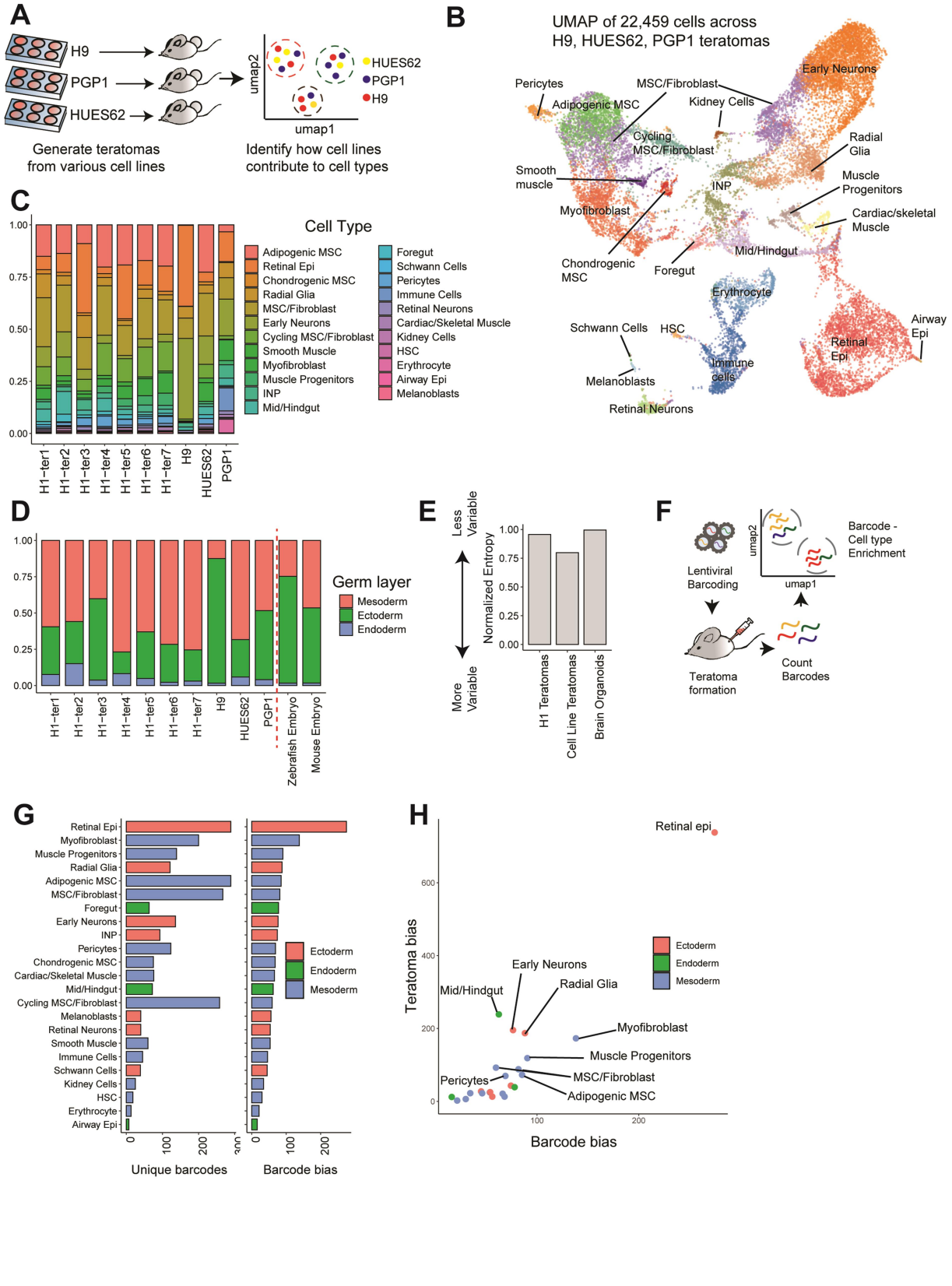
Another key question in teratoma formation is how many cells engraft into the teratoma after stem cell injection. Towards this we used lentiviral barcoding(Guo et al. 2018) to assess the fraction of injected PSCs that engraft and go on to form the teratoma. For 3 out of the 7 H1 ESC teratomas, prior to PSC injection, cells were transduced with an integrating lentiviral ORF barcode (**Figure 6F**). The barcode consisted of a 25 random base pair sequence proximal to the lentiviral 3' long terminal repeat (LTR) region, and can thus, be detected by scRNA-seq (**Figure S4D**). In this manner cells can be individually labeled prior to teratoma formation and teratoma cells that descend from these cells can be later captured via scRNA-seq. Transduced PSCs were evenly split: half for teratoma formation and half were frozen down for DNA sequencing. By comparing unique barcodes extracted from genomic DNA in these two cell populations we can calculate the proportion of cells that engraft. Results show that across the three teratomas, over 25% of cells engraft, out of a total of 10 million injected cells, which suggests that no major bottlenecking occurs during teratoma formation (**Figure S4E**). This is especially important in the context of using teratomas in multiplex screens, as one must ensure that there are enough cells contributing to the final tumor to enable an adequately powered high throughput assay.

We next also tracked barcodes in individual cells by amplifying the expressed barcode from scRNA-seq. Since cells from the teratoma with the same barcode originated from the same PSC, we were able to track whether certain PSCs were primed to develop into certain lineages. For each cell type, we computed a barcode bias score, which reflects

the level to which barcodes tend to be enriched or depleted in that cell type and plotted this barcode bias, alongside the total number of barcodes detected in each cell type (**Figure 6G, Methods**). We also computed a teratoma bias score for each cell type, which reflects how much the proportion of that cell type varies across teratomas and plotted the correlation of the teratoma bias score with the barcode bias score (**Figure 6H, Methods**). We can see that retinal epithelium is an outlier with both a high teratoma bias, and a high barcode bias (**Figure 6H**). Myofibroblast cells also have a relatively high barcode and teratoma bias score while Early Neurons, Radial Glia, Mid/Hindgut have high teratoma bias score (**Figure 6H**). Both the barcode bias and teratoma bias scores are scaled by the number of cells in each cell type (**Methods**). .

Taken together, we found teratomas derived from the same and different cell lines to generally contain the same major cell types at 10 weeks of growth. In general, teratomas contain a large fraction of MSC/Fibroblast and neuronal cell types, with a fairly small fraction of endodermal cell types. Retinal epithelium shows both a high degree of heterogeneity across teratomas and a high level of lineage priming as determined by lentiviral barcoding assays. The level of heterogeneity between teratomas generated from H1 stem cells is comparable to that observed in organoids(Quadrato et al. 2017; de Souza 2017; Velasco et al. 2019). There is a much higher level of heterogeneity between teratomas derived from different pluripotent cell lines, which reflects known variability across those lines(Ortmann and Vallier 2017).

Figure 6: Assaying teratoma heterogeneity. (A) Schematic portraying how to generate teratomas from multiple cell lines and identifying how lines contribute to cell types **(B)** UMAP scatterplot of all cell types present across 3 PSC lines (H9, HUES62, and PGP1) **(C)** Distribution of cell types represented in each individual teratoma **(D)** Distribution of germ layer representation in each individual teratoma (along with zebrafish and mouse comparison). **(E)** The Normalized Entropy represents how well cell type assignments are mixed with teratoma/organoid/cell line identities. A higher Normalized Entropy implies less cell type variation between teratomas/organoids/cell lines. The Cell Line Teratomas include one teratoma generated from each of HUES62, H9, and PGP1 lines. **(F)** H1 cells were uniquely barcoded at low MOI with lentiviral vectors before teratoma formation. The barcodes were counted and assessed for lineage/cell type priming of cells. **(G)** Number of unique barcodes detected in each cell type plotted alongside the cell type bias for specific barcodes (computed using the KL divergence of cell type identities with barcode identities scaled by the number of cells in each cell type) **(H)** Teratoma bias for each cell type plotted against barcode bias.


A
H9
PGP1
HUES62
Generate teratomas from various cell lines
umap2
umap1
HUES62
PGP1
H9
Identify how cell lines contribute to cell types
B
UMAP of 22,459 cells across H9, HUES62, PGP1 teratomas
Pericytes
Adipogenic MSC
MSC/Fibroblast
Kidney Cells
Early Neurons
Cycling MSC/Fibroblast
Radial Glia
Smooth muscle
INP
Muscle Progenitors
Myofibroblast
Cardiac/skeletal Muscle
Chondrogenic MSC
Foregut
Mid/Hindgut
Erythrocyte
Schwann Cells
HSC
Airway Epi
Retinal Epi
Immune cells
Melanoblasts
Retinal Neurons
C
Cell Type
Adipogenic MSC
Retinal Epi
Chondrogenic MSC
Radial Glia
MSC/Fibroblast
Early Neurons
Cycling MSC/Fibroblast
Smooth Muscle
Myofibroblast
Muscle Progenitors
INP
Mid/Hindgut
Foregut
Schwann Cells
Pericytes
Immune Cells
Retinal Neurons
Cardiac/Skeletal Muscle
Kidney Cells
HSC
Erythrocyte
Airway Epi
Melanoblasts
1.00
0.75
0.50
0.25
0.00
H1−ter1
H1−ter2
H1−ter3
H1−ter4
H1−ter5
H1−ter6
H1−ter7
H9
HUES62
PGP1
D
Germ layer
Mesoderm
Ectoderm
Endoderm
Zebrafish Embryo
Mouse Embryo
E
Less Variable
More Variable
Normalized Entropy
H1 Teratomas
Cell Line Teratomas
Brain Organoids
F
Lentiviral Barcoding
Teratoma formation
Count Barcodes
Barcode - Cell type Enrichment
G
Unique barcodes
Barcode bias
Ectoderm
Endoderm
Mesoderm
H
Teratoma bias
Barcode bias
Retinal epi
Early Neurons
Radial Glia
Mid/Hindgut
Myofibroblast
Muscle Progenitors
MSC/Fibroblast
Adipogenic MSC
Pericytes

2.2.3 Teratoma Maturity

We next assessed the transcriptional similarity of the teratoma cell types to human fetal cell types, specifically from the human cortex and gut, demonstrating the teratoma's relevance as a tool for modeling human development. We specifically looked at which which human embryonic stage the 10 week teratoma cell types most resemble, projected the teratoma data onto the fetal data, and assessed the expression of key cell type marker genes (**Figure 7A**). These results may vary depending on time and/or size allowed for growth and size/species of animal used to form the human teratoma. For our analyses we thus focused on a specific set of parameters i.e. 10-week old teratomas grown in $Rag2^{-/-}\gamma c^{-/-}$ immunodeficient mice.

Due to the semi-random nature of teratoma differentiation, it is possible that different cell types will resemble different stages of embryonic development. Thus, we analyzed individual tissue types separately, looking at the neuro-ectoderm and gut cell types in-depth. We first used the cosine similarity metric to compare the average expression of all cells belonging to neural subtypes with the average expression of the same subtypes in a (2,300 cell) fetal brain dataset at different stages of development(Zhong et al. 2018) (**Figure 7A, Figure 7B**). We found that the teratoma neuronal cells had high similarity scores to the human prefrontal cortex at weeks 13 – 17 with the highest score for weeks 16 – 17 (**Figure 7B**). We sub-clustered the neuro-ectoderm cells and identified additional subtypes, including a cluster of early interneurons (**Figure 7C**, Table 2). Due to the high similarity with week 16 – 17 human data, we identified those subtypes (Radial Glia, Cycling Progenitors, Early Neurons, Early Interneurons) that matched with the cell types seen in a larger 40,000+ cell week 17 – 18 dataset also from the human prefrontal cortex for futher analysis(Polioudakis et al. 2019) (**Figure 7A, Figure 7C**).

We then generated a Similarity Weighted Nonnegative Embedding (SWNE) of the week 17 – 18 human prefrontal cortex cells and projected the teratoma cells from the matching subtypes onto the fetal human SWNE (**Figure 7A, Figure 7D**)(Wu, Tamayo, and Zhang 2018). Briefly, SWNE embeds single cell gene expression data in two dimensions, similar to t-SNE and UMAP, while preserving more of the global structure and enabling genes to be visualized alongside the cells. The closer an embedded gene is to a group of cells, the higher the expression level of that gene in those cells(Wu, Tamayo, and Zhang 2018). We found similar cell types map to similar spatial positions in the SWNE embedding, although the teratoma SWNE embedding shows some overlap between cycling progenitors and radial glia as well as early interneurons and excitatory neurons (**Figure 7D**). Additionally, the teratoma radial glia cells project onto the fetal intermediate progenitors, suggesting that intermediate progenitors might be currently clustering with radial glia in the teratomas due to undersampling of cells (**Figure 7D**).

To assess the similarity of the teratoma neuro-ectoderm cell types to the fetal prefrontal cortex cell types, we defined a panel of neuronal cell type marker genes: DCX, NEUROD1, HES5, SOX2, HMGB2, VIM, DLX1 and then correlated the expression of these marker genes between the teratoma cells and fetal brain cells for every matched cell type (**Figure 7A, Figure 7E**). We found a fairly high correlation overall, with R = 0.82 for Radial Glia, R = 0.93 for Cycling Progenitors, R = 0.84 for Interneurons, and R = 0.77 for Early Neurons (**Figure 7E**). We also looked at the cell type proportions in the fetal prefrontal cortex versus the teratoma, showing the matched cell types have similar transcriptional profiles, the teratoma has far more progenitor cells such as Radial Glia, and fewer early neurons with no detectable mature neurons (**Figure 7F**). We also ran a differential expression as well as a geneset enrichment analysis between the matched teratoma and fetal prefrontal cortex cell types to assess the differences between the teratoma and fetal cells (**Figure S5A, S5B**). All

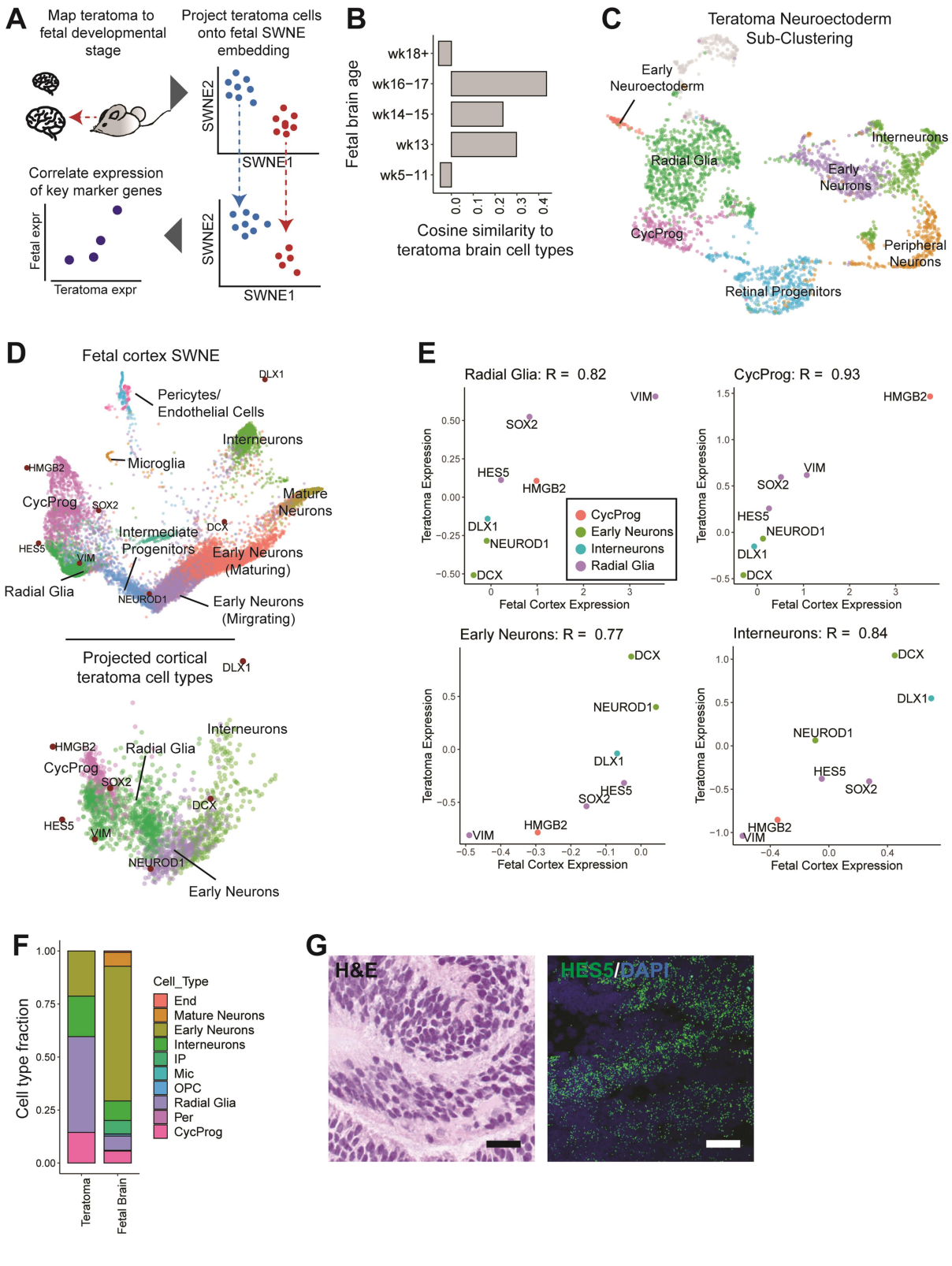
four cell types showed similar top differentially expressed genes as well as genesets, suggesting that the main differences between the teratoma and fetal cells are global and not cell type specific (**Figure S5A, S5B**). The teratoma cells have a higher expression of genes related to organ morphogenesis while the fetal cells express genes related to methylation, suggesting the teratoma cells may not have the same epigenetic signatures as fetal cells (**Figure S5A, S5B**).

This analysis was repeated with teratoma gut subtypes using a published fetal gut dataset as reference(S. Gao et al. 2018). The teratoma gut cells were most similar to week 8-11 fetal gut age (**Figure S5C**). We compared marker genes for gut cell types (CDX1, CDX2, HHEX, FOXJ1, PAX9, SOX2) between teratoma and fetal cells and found a high overall correlation, with an R = 0.96 for foregut and R = 0.97 for mid/hindgut (**Figure S5D**). The projection of fetal gut data onto the teratoma SWNE again shows relatively similar spatial positioning (**Figure S5E**). We see that the teratoma produces less foregut and more mid/hindgut than the fetal gut (**Figure S5F**). When looking at the differences between the teratoma and fetal gut cells, we again see that the fetal cells express more methylation related genes (**Figure S5A, S5B**). In this case, the teratoma cells express more genes related RNA/DNA metabolism (**Figure S5A, S5B**).

To additionally validate these results, the presence of key cell subtypes was further validated utilizing RNAScope ISH technology. We probed for the radial glia marker HES5 which showed high abundance in regions of neuroectoderm in fixed teratoma tissue sections (**Figure 7F**). Additionally, we probed for FOXJ1 (marker for cilia) and by imaging a speculative airway in a teratoma tissue section, visualized high abundance of FOXJ1 lining the airway, using POLR2A, PPIB, and UBC as positive controls (**Figure S5E, Figure S5F**). Overall, we were able to show that the teratoma neuro-ectoderm and gut cell types are

transcriptionally similar to their fetal counterparts, while also identifying the developmental stage of the teratoma cells.

Figure 7: Assaying teratoma maturity. (A) Teratoma neuro-ectoderm cell types were mapped to fetal cortical cell types and the corresponding teratoma cell types were projected onto SWNE embeddings of fetal cells. Key marker genes were correlated across matching teratoma/fetal cell types, and average expression of teratoma cell types were correlated with fetal cell types from different stages of development. **(B)** Cosine similarity of teratoma brain cells with fetal brain cells of different ages. **(C)** UMAP embedding of teratoma neuro-ectoderm sub-clusters. **(C)** Projection of teratoma neuro-ectoderm cell types onto the SWNE embedding of fetal cortical cells. **(D)** Correlation of the scaled expression of key marker genes across Radial Glia, Intermediate Neuronal Progenitors, and Early Neurons. **(E)** Fraction of brain related cell types in the teratoma and fetal cortex. **(G)** H&E stain (left) and RNAScope image (right) of HES5 (radial glia marker) expression. DAPI is a nuclear stain. 4-10 punctate dots/cell is a positive result. Scalebar = 50μM.


A
Map teratoma to fetal developmental stage
Project teratoma cells onto fetal SWNE embedding
SWNE2
SWNE1
Correlate expression of key marker genes
Fetal expr
Teratoma expr
B
Fetal brain age
wk18+
wk16−17
wk14−15
wk13
wk5−11
Cosine similarity to teratoma brain cell types
C
Teratoma Neuroectoderm Sub-Clustering
Early Neuroectoderm
Radial Glia
Interneurons
Early Neurons
CycProg
Peripheral Neurons
Retinal Progenitors
D
Fetal cortex SWNE
Pericytes/ Endothelial Cells
Microglia
Interneurons
Mature Neurons
Intermediate Progenitors
Early Neurons (Maturing)
Early Neurons (Mirgrating)
Radial Glia
Projected cortical teratoma cell types
E
Radial Glia: R = 0.82
CycProg: R = 0.93
Early Neurons: R = 0.77
Interneurons: R = 0.84
Teratoma Expression
Fetal Cortex Expression
CycProg
Early Neurons
Interneurons
Radial Glia
F
Cell type fraction
Cell_Type
End
Mature Neurons
Early Neurons
Interneurons
IP
Mic
OPC
Radial Glia
Per
CycProg
Teratoma
Fetal Brain
G
H&E
HES5/DAPI

2.2.4 Dissecting the Multi-Lineage Effects of Developmental Regulators using a CRISPR Knockout Screen in Teratomas

To establish the utility of the teratoma system for modeling developmental cell fate specification and lineage permissibility, we next performed a single-cell genetic perturbation screen utilizing CRISPR-Cas9. Towards this we compiled a list of 24 major organ/lineage specification genes that are embryonically lethal upon knockout in mice via literature review (**Table 1**). Studying the effects of these genes using cell lines or organoid models would require different experiments and different models for each cell lineage. With the teratoma model, we can screen the effects of these genetic perturbations in all major lineages in the same experiment. Utilizing the CROPseq-Guide-Puro vector backbone, we cloned in 48 individual single guide RNAs (sgRNAs) directed at each developmental gene (2 sgRNAs per gene)(Datlinger et al. 2017) **(Figure 8A)**. In order to perform the CRISPR-Cas9 perturbation screen we designed a stable Cas9-expressing iPSC line (PGP1). This line was created via knockin of a CAG-spCas9-P2A-EGFP cassette with an upstream T2A linked blasticidin resistance gene all into the AAVS1 locus of the PGP1 genome (**Figure S6A, Methods**). Proper integration of our cassette was confirmed via PCR amplification of the left and right arm utilizing primers that amplified regions spanning both the PGP1 AAVS1 endogenous locus and the engineered cassette (**Figure S6B**). This was further validated by direct sanger sequencing of the arms (**Figure S6A**). After creating a pooled lentiviral library with our sgRNAs, we transduced our engineered PGP1-Cas9 line at a MOI of 0.1 so that each cell received approximately one perturbation **(Figure 8A)**. After selection these cells were injected subcutaneously into 3 $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice for teratoma formation, extraction, and downstream scRNA-seq processing with 10X Genomics (**Figure 8A**).

We validated the editing efficiencies of all our guide RNAs using PCR amplification of the expected cut site and looking for mutations and indels with CRISPResso **(Table 2,**

Methods). We then selected the top guide targeting each gene with at least a 60% editing efficiency which resulted in a total of 16 guides (**Table 2, Methods)**. We then only used these top guides for further computational analysis. We also reran the CRISPR-KO screen by repooling these top guides and generated 3 additional teratomas **(Figure 8A, Methods)**. We successfully captured a median of 118 cells per gene/guide in the original screen and 1280 cells per gene/guide in the replicate screen (**Figure S6C**). We were able to capture more cells per guide in the replicate screen since we only pooled the top 16 guides, while the original screen had a total of 48 guides **(Methods)**.

In order to ensure consistent cell types across teratomas, we integrated all six teratomas across both the original and replicate screen using Seruat v3. We then called cell types in the PGP1 teratoma cells using Seurat label transfer with the 7 H1 teratomas serving as the reference dataset (**Methods**). Cell types with fewer than 100 cells were collapsed into their closest neighboring cell type, and we visualized the resulting cell types using a UMAP plot (**Figure S6D, Methods**). To determine the total effect of each knockout, we computed the Earth Mover's Distance (EMD) between all cells in each gene knockout with all cells belonging to the NTC separately for each screen(W. S. Chen et al. 2020) **(Figure 8A, Methods)**. EMD computes the difference in cell type composition between two groups of cells, weighted by how transcriptionally distinct the cell types are(W. S. Chen et al. 2020) **(Methods)**. We correlated the EMD between each gene and NTC across the original and replicate screens, showing that both TWIST1 and RUNX1 have high EMD in with an overall pearson correlation of 0.37 **(Figure S6E)**. In order to assess the effects of each gene knockout on a given cell type or developmental lineage, we used a ridge regression model to separately analyze the original and replicate screens(Dixit, Parnas, Li, Weissman, et al. 2016) (**Figure 8A, Methods**). The ridge regression model generated regression coefficients that represent the effects of each gene knockout on cell type enrichment/depletion **(Figure**

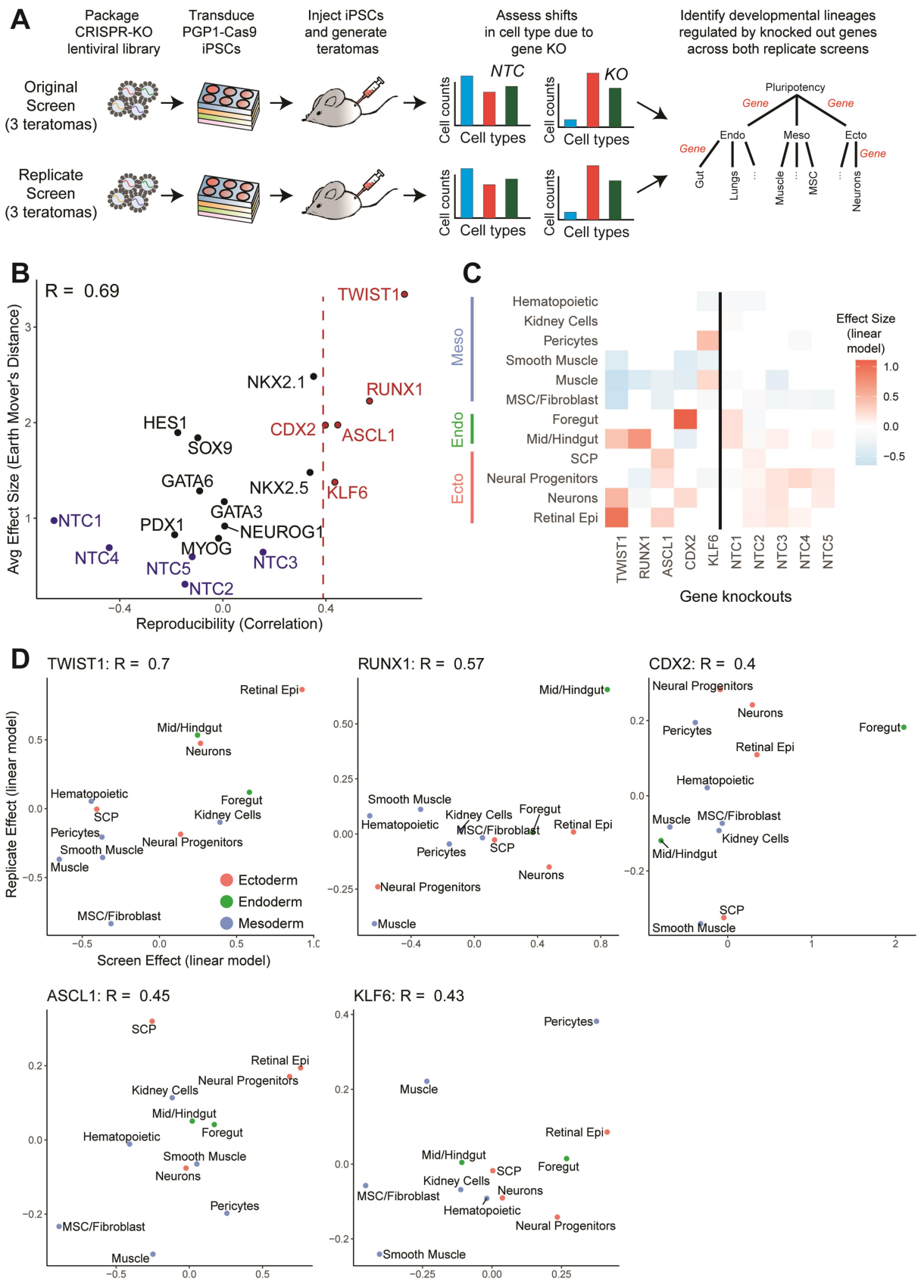
8A, Methods). For each gene, we plotted the Earth Mover's Distance between that gene's cells and the NTC cells, as well as the Pearson correlation of the regression coefficients for the both the original screen and the replicate screen, giving us a sense of both the effect size and reproducibility of each gene knockout. Plotting the average effect size the gRNA correlation shows that gene knockouts with strong effect sizes tend to be more reproducible (R = 0.69) (**Figure 8B, Methods**). We highlighted genes with a Pearson correlation of greater than 0.4 between the original and replicate screen for further analysis (**Figure 8B**).

For the highlighted genes, specifically TWIST1, RUNX1, CDX2, KLF6, and ASCL1, we wanted to identify the statistically significant effect of each gene knockout on each cluster. We merged the cells from both screens and ran a merged ridge regression analysis **(Methods)**. P-values for each gene-cluster coefficient were computed by shuffling gene assignments and generating a null distribution, and we computed false discovery rates using the Benjamani-Hochberg correction **(Methods)**. We set a significance threshold of FDR = 0.1 and all gene-cluster coefficients with a larger p-value were set to zero (**Methods**).

CDX2 is a known major organ specification gene for the development of the midgut and hindgut(N. Gao, White, and Kaestner 2009; Silberg et al. 2002). Interestingly, our data shows that cells containing a CDX2 knockout are shifting away from midgut/hindgut tissues with enrichment in foregut (**Figure 8C, 8D**). This has been shown in past literature that CDX2 knockout shifts the differentiation pathway away from intestine and instead promotes gastric activation(T. Kim and Shivdasani 2016; Simmini et al. 2014). TWIST1 also showed a large effect size in our screen and is a known transcription factor for epithelial-to-mesenchymal transition important in development as well as disease such as metastatic cancers **(Figure 8B)**(Kalluri and Weinberg 2009; Yang et al. 2004). Interestingly, our screen validates such findings as cells containing a TWIST1 knockout are shifting away from mesodermal cell types (muscle, smooth muscle, pericytes, and mesenchymal stem cell/fibroblasts) and

enriching for neuro-ectoderm (retinal epithelium, neurons) (**Figure 8C, 8D**). Studies have shown the importance of TWIST1 for mesodermal specification and differentiation(Qin et al. 2011). We see that RUNX1 knockout results in a depletion of neurons and muscle cell types and an enrichment in mid/hindgut, which is consistent with previous mouse and stem cell studies that show RUNX1 to be critical for neural crest formation, signaling in gut epithelium stem cells, and myoblast proliferation(Umansky et al. 2015; Marmigère et al. 2006; Sarper et al. 2018) **(Figure 8C, 8D)**. We showed that KLF6 knockout resulted in a depletion of perictyes, which is possibly consistent with its role in promoting endothelial activation during vascular repair(Garrido-Martín et al. 2012) **(Figure 8C, 8D)**. ASCL1 interestingly results in an increase in proportion of retinal epithlium and neural progenitors **(Figure 8C, 8D)**. Since ASCL1 is key to cell cycle exit and neuronal differentiation, it could be that knocking out ASCL1 slows down neurogenesis and results in a buildup of neural progenitors(Castro et al. 2011). With this CRISPR knockout screen of key developmental regulators, we were able to assay the multi-lineage functions of these genes in a human-specific model, something that to our knowledge, no other human specific developmental model can currently accomplish.

Figure 8: Genetic perturbations. **(A)** PGP1-Cas9 iPSCs were induced with a CRISPR library targeting a panel of 16 key developmental genes with 1 gRNAs per gene. After generating 3 teratomas with the PGP1-iPSCs, scRNA-seq was used to identify shifts in cell type formation as a result of gene knockouts. We repeated this process with 3 additional teratomas to serve as a replicate screen. (**B)** Average effect of gene knockout on cell type enrichment/depletion versus the correlation of cell type enrichment between the original screen and replicate screen. Genes with a reproducibility greater than 0.4 (**Methods**) were selected for further analysis. (**C)** A heatmap of the effect size (regression coefficient) of gene knockout enrichment for cell types and germ layers. **(D)** Scatterplot of individual guide RNA effects on cell type abundance for selected genes *TWIST1*, *RUNX1*, *CDX2, KLF6, ASCL1*.


A
Package CRISPR-KO lentiviral library
Transduce PGP1-Cas9 iPSCs
Inject iPSCs and generate teratomas
Assess shifts in cell type due to gene KO
Identify developmental lineages regulated by knocked out genes across both replicate screens
Original Screen (3 teratomas)
Replicate Screen (3 teratomas)
NTC
KO
Cell counts
Cell types
Pluripotency
Gene
Endo
Meso
Ecto
Gut
Lungs
Muscle
MSC
Neurons
B
R = 0.69
Avg Effect Size (Earth Mover's Distance)
Reproducibility (Correlation)
TWIST1
NKX2.1
RUNX1
HES1
CDX2
ASCL1
SOX9
GATA6
NKX2.5
KLF6
GATA3
NTC1
PDX1
NEUROG1
MYOG
NTC4
NTC5
NTC3
NTC2
C
Meso
Endo
Ecto
Hematopoietic
Kidney Cells
Pericytes
Smooth Muscle
Muscle
MSC/Fibroblast
Foregut
Mid/Hindgut
SCP
Neural Progenitors
Neurons
Retinal Epi
Effect Size (linear model)
TWIST1
RUNX1
ASCL1
CDX2
KLF6
NTC1
NTC2
NTC3
NTC4
NTC5
Gene knockouts
D
TWIST1: R = 0.7
RUNX1: R = 0.57
CDX2: R = 0.4
ASCL1: R = 0.45
KLF6: R = 0.43
Replicate Effect (linear model)
Screen Effect (linear model)
Ectoderm
Endoderm
Mesoderm

2.3 Discussion

Developmental biology has used *in vitro* systems such as cell line models and organoids, as well as mouse and fly models to elucidate key properties of human development. However, these models have limitations, such as a lack of tissue maturity, thickness, as well as issues with scalability, and efficiency(Meritxell Huch et al. 2017; M. Huch and Koo 2015; Yin et al. 2016). The teratoma has the potential to be a fully vascularized, multilineage model for human development. While a few studies have opted to use the teratoma as a tool to derive rare cell types such as muscle progenitors and hematopoietic stem cells(Suzuki et al. 2013; Tsukada et al. 2017; Philipp et al. 2018; Amabile et al. 2019; Chan et al. 2018), no study to our knowledge has delved into deeper characterization of this potential model for human development.

In this study, we have shown using scRNA-seq analysis, histological H&E staining, and RNA-FISH that the teratoma can give rise to over 20 cell types, ranging from radial glia to ciliated respiratory epithelium. We assessed the heterogeneity between teratomas derived from the H1 stem cell line, showing that the teratomas have a similar level of heterogeneity to that of organoids. We also generated teratomas from H9, HUES62, and PGP1 cell lines, and showed that teratoma heterogeneity is significantly higher across different cell lines. We then assessed lineage priming in H1 cells through lentiviral barcoding strategies. Additionally, we were able to assess the transcriptional similarity of our cell types to fetal mouse single cell datasets, and in the case of neuro-ectoderm/gut cell types, fetal human datasets. We also performed a developmental CRISPR-Cas9 knockout screen using the teratoma, resulting in depletion in midgut/hindgut and enrichment in foregut with the CDX2 knockout, as well as depletion in muscle cell types with the TWIST1 knockout. This ability to assess the developmental impact of key developmental regulators across a wide variety of cell types is unique to the teratoma system.

The teratoma's major advantages are that it can grow to a large size due to its vascularization, and it can produce a wide array of cell types from all major developmental lineages. We have been able to capture cell types such as Schwann cells that are difficult to make with neuronal organoid systems(Philipp et al. 2018; Suzuki et al. 2013; Tsukada et al. 2017; Amabile et al. 2019; Chambers, Tchieu, and Studer 2013). As we demonstrated with our CRISPR-Cas9 knockout screen, the teratoma's ability to generate cells from all lineages enables a comprehensive assessment of the effect of perturbations on development.

It is important to keep in mind that our observed variability may also be an overestimation of the true variability, since we sample such a small sub-population of cells from the entire teratoma for scRNA-seq. Additionally, there may be some cell types that are more robust in handling standard dissociation and collection protocols.

Overall, the cost of profiling a single teratoma with the 10X RNA-seq system runs at about $1,300 including sequencing costs for ~8000 cells at a sequencing depth of 50,000 reads per cell. Mouse husbandry and reagents related to teratoma formation (cells, Matrigel, media) are relatively cheap in comparison. During teratoma growth, the researcher needs to only monitor the mice for health concerns, weights, and tumor measurements if desired. It is also possible to inject both flanks of the mouse to generate 2 teratomas per animal. With the availability of easy to use analysis tools such as Seurat/PAGODA2, as well as methods for integrating datasets (such as CONOS), running a basic clustering and cell type annotation of scRNA-seq data is fairly straightforward.

One limitation of the teratoma system that it shares with organoids is the degree of heterogeneity between teratomas, especially for specific cell types like the retinal epithelium(de Souza 2017; Phipson et al. 2019; Capowski et al. 2019; Quadrato et al. 2017). Thus, the use of internal controls is extremely helpful. For example, in our CRISPR-Cas9

screen, each teratoma contained both the knockout guides and the negative controls, enabling us to compare cell type proportion shifts within each teratoma without having to worry about heterogeneity between teratomas. Additionally, while the teratoma has regions of organization and maturity, it still develops in a semi-random manner. This may prove to be a barrier in accessing certain mature cell types that need a highly ordered cellular context to develop. This semi-random development results in cell types maturing at different rates, which may cause issues when studying temporally sensitive outputs. Since the teratoma contains cell types from all lineages, finding a single dissociation protocol that captures as many cell types as possible still remains a challenge. The choice of dissociation method can drastically change the cell types profiled in single cell RNA-seq, and it is likely that certain cell types are not detected in our single cell assays due to our dissociation protocol(Denisenko et al. 2019). Additionally, we may still be underpowered in detecting rare cell types or cell subtypes and additional single cell RNA-seq could enable us to resolve some missing cell types. For some cell types that we would expect to see in the teratoma, such as hepatocytes, it is unclear if they are dropping out due to our dissociation protocol or if they are being collapsed into a larger cell type due to undersampling.

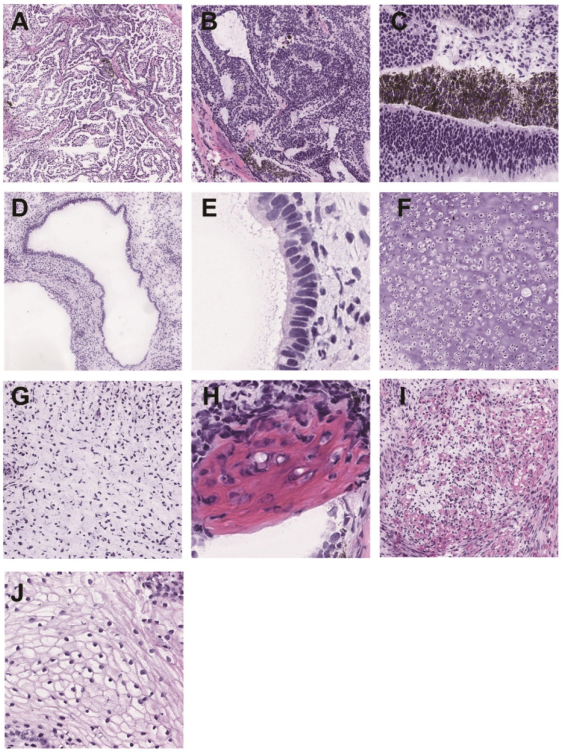
Interestingly, we see different embryonic stages in development depending on tissue type analyzed, suggesting teratoma development is asynchronous. We postulate that since our teratoma dataset is largely neural in origin, this tissue type may be permitted longer time for development and maturity. Conversely, the gut subtypes may appear later in teratoma development and thus, in smaller proportions with less maturation.

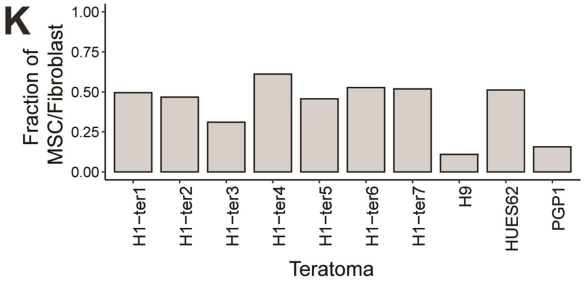
One critical future study is assessing the impact of different dissociation methods on teratoma cell type proportion. It may be the case that no single dissociation method can capture all cell types, and it will be necessary to design specific dissociation protocols to capture specific tissues, such as hepatocytes or lung epithelium. Additionally, the ability to

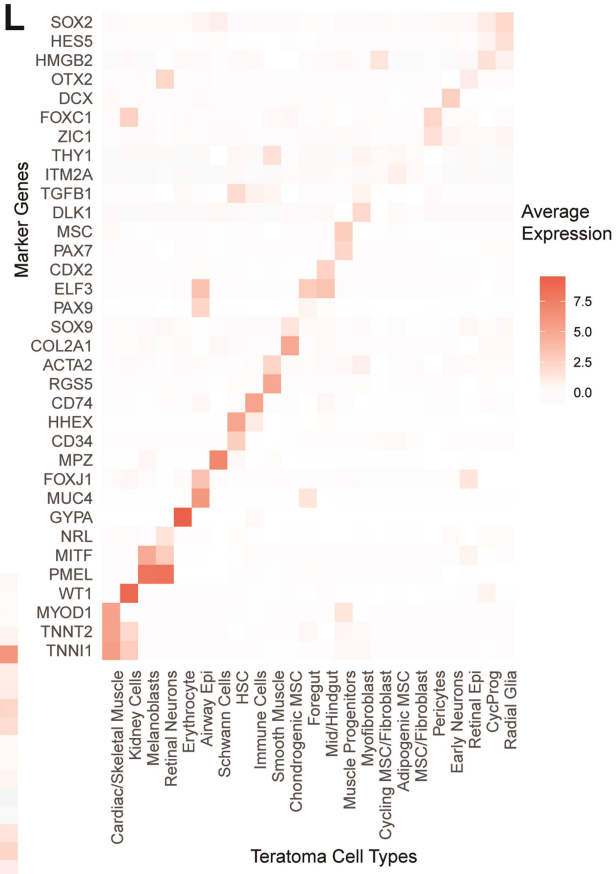
achieve greater cell numbers with the most current single cell RNA sequencing protocols, such as SPLiT-seq(Rosenberg et al. 2018, 2017) and sci-CAR(Cao et al. 2017b), will be vital for identifying additional cell types. A time series analysis of teratomas at multiple stages of maturity could help elucidate the developmental pathways that these cell types follow, and enable us to compare those pathways with the ones found in human development. Additionally, pooling different cell types together with PSCs prior to injection may help aid in cellular enrichment/maturity in the teratoma (i.e. HUVECs to enrich for HSC populations)(Philipp et al. 2018). Growing patient-specific teratomas could benefit disease research through isogenic iPSC lines. This could aid in understanding the disease state in various tissues that otherwise may be difficult to recapitulate. Finally, further optimization is necessary on the miRNA molecular sculpting technology, specifically generating stable miRNA cell lines and and optimizing the timing of GCV administration. Overall, the teratoma has potential to further developmental biology and human disease research and span tissue engineering applications.

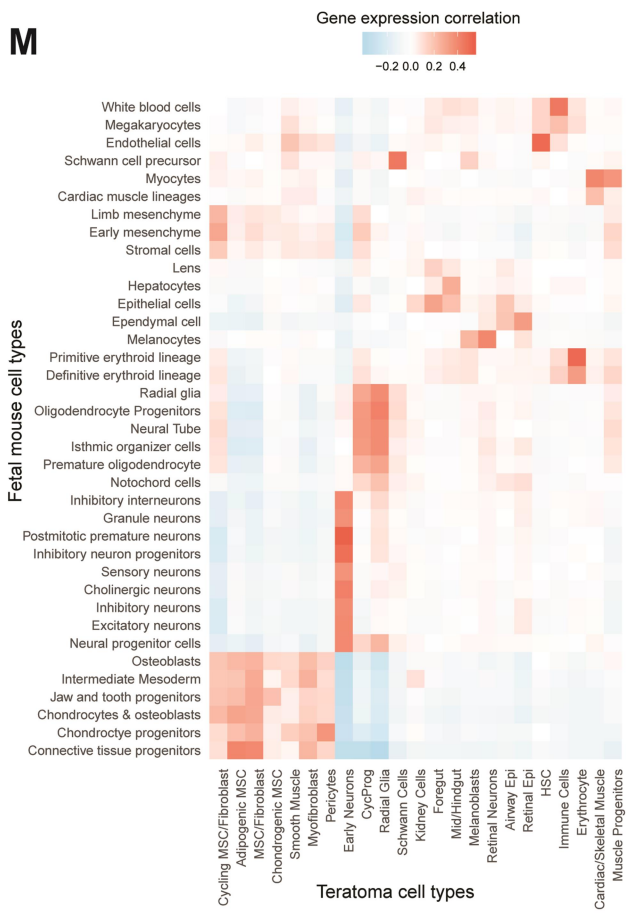
2.4 Supplemental Figures

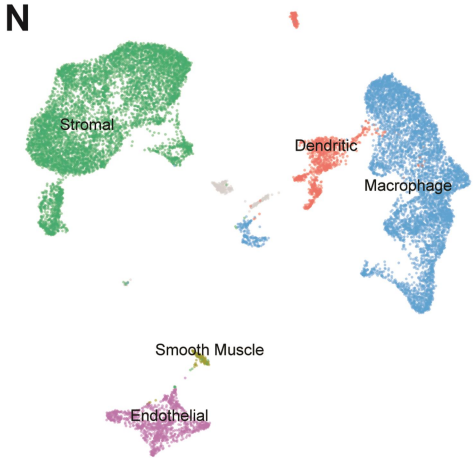
Supplementary Figure 3: Comprehensive teratoma characterization. H&E stains of the **(A)** Choroid plexus **(B)** Fetal neuroectoderm **(C)** Retinal pigmented epithelium **(D)** Developing airway **(E)** Ciliated respiratory epithelium **(F)** Fetal cartilage. **(G)** Mesenchyme **(H)** Bone **(I)** Developing cardiac muscle / skeletal muscle **(J)** Squamous epithelium (**K)** The fraction of cells that are classified as MSC/Fibroblast across each teratoma. **(L)** Heatmap of top marker genes for each cell type. **(M)** Correlation of the average expression of each human teratoma cell type with the average expression of each fetal mouse cell type. **(N)** UMAP plot of mouse cell types in the H1 teratomas

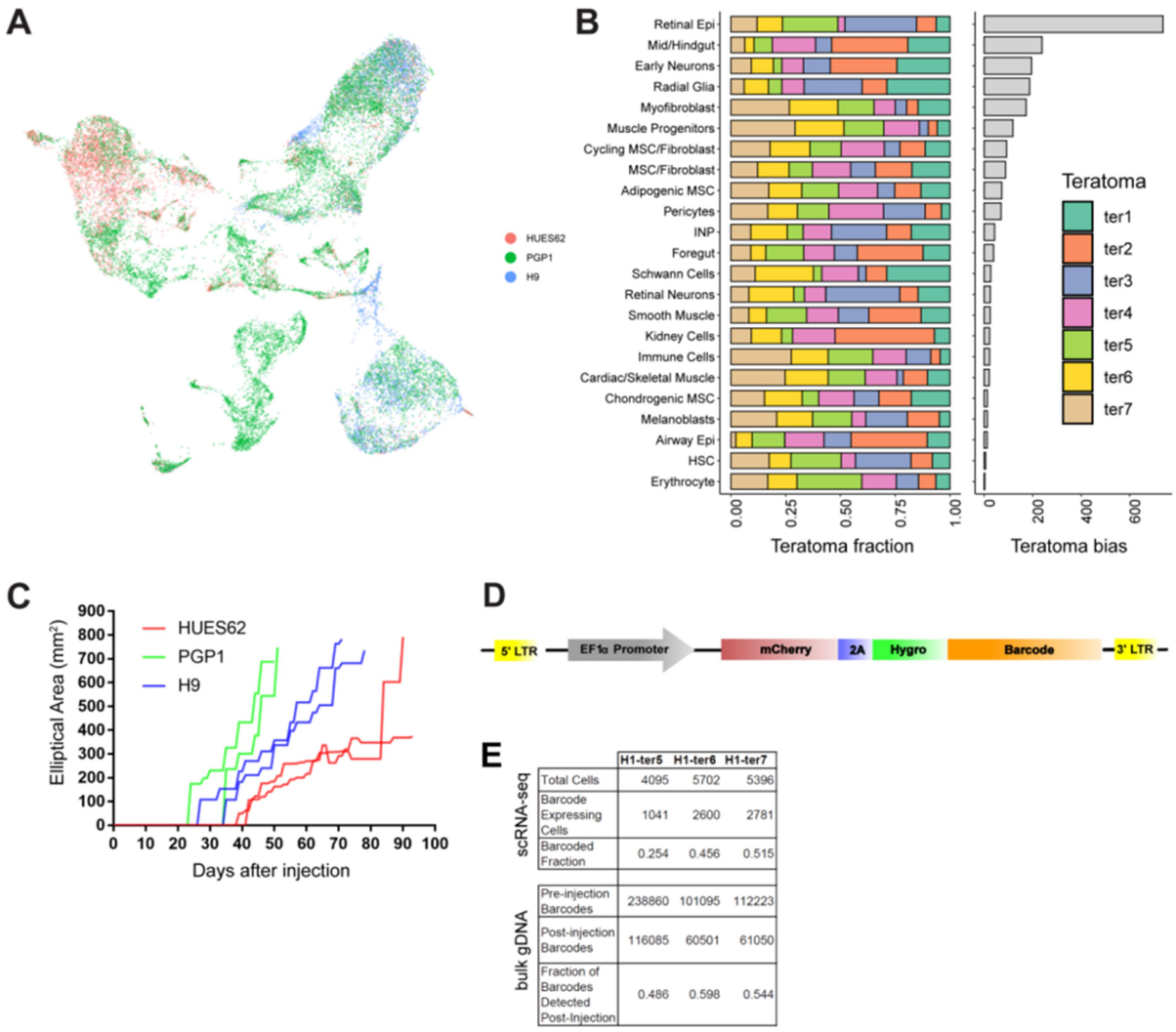

A
B
C
D
E
F
G
H
I
J


K
Fraction of MSC/Fibroblast
1.00
0.75
0.50
0.25
0.00
H1−ter1
H1−ter2
H1−ter3
H1−ter4
H1−ter5
H1−ter6
H1−ter7
H9
HUES62
PGP1
Teratoma


L
Marker Genes
SOX2
HES5
HMGB2
OTX2
DCX
FOXC1
ZIC1
THY1
ITM2A
TGFB1
DLK1
MSC
PAX7
CDX2
ELF3
PAX9
SOX9
COL2A1
ACTA2
RGS5
CD74
HHEX
CD34
MPZ
FOXJ1
MUC4
GYPA
NRL
MITF
PMEL
WT1
MYOD1
TNNT2
TNNI1
Average Expression
7.5
5.0
2.5
0.0
Cardiac/Skeletal Muscle
Kidney Cells
Melanoblasts
Retinal Neurons
Erythrocyte
Airway Epi
Schwann Cells
HSC
Immune Cells
Smooth Muscle
Chondrogenic MSC
Foregut
Mid/Hindgut
Muscle Progenitors
Myofibroblast
Cycling MSC/Fibroblast
Adipogenic MSC
MSC/Fibroblast
Pericytes
Early Neurons
Retinal Epi
CycProg
Radial Glia
Teratoma Cell Types


M
Gene expression correlation
−0.2 0.0 0.2 0.4
Fetal mouse cell types
White blood cells
Megakaryocytes
Endothelial cells
Schwann cell precursor
Myocytes
Cardiac muscle lineages
Limb mesenchyme
Early mesenchyme
Stromal cells
Lens
Hepatocytes
Epithelial cells
Ependymal cell
Melanocytes
Primitive erythroid lineage
Definitive erythroid lineage
Radial glia
Oligodendrocyte Progenitors
Neural Tube
Isthmic organizer cells
Premature oligodendrocyte
Notochord cells
Inhibitory interneurons
Granule neurons
Postmitotic premature neurons
Inhibitory neuron progenitors
Sensory neurons
Cholinergic neurons
Inhibitory neurons
Excitatory neurons
Neural progenitor cells
Osteoblasts
Intermediate Mesoderm
Jaw and tooth progenitors
Chondrocytes & osteoblasts
Chondrocyte progenitors
Connective tissue progenitors
Cycling MSC/Fibroblast
Adipogenic MSC
MSC/Fibroblast
Chondrogenic MSC
Smooth Muscle
Myofibroblast
Pericytes
Early Neurons
CycProg
Radial Glia
Schwann Cells
Kidney Cells
Foregut
Mid/Hindgut
Melanoblasts
Retinal Neurons
Airway Epi
Retinal Epi
HSC
Immune Cells
Erythrocyte
Cardiac/Skeletal Muscle
Muscle Progenitors
Teratoma cell types

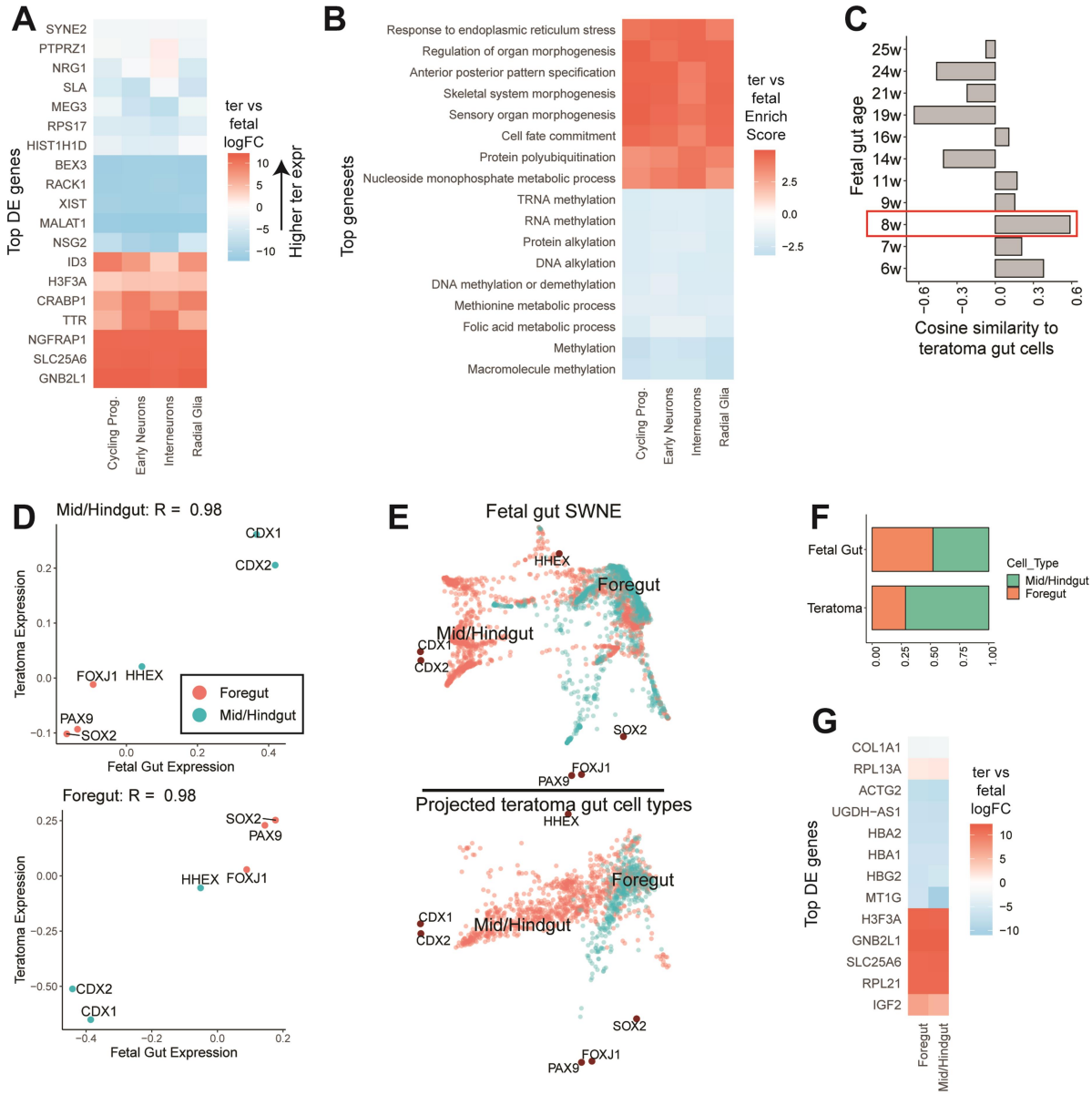

N
Stromal
Dendritic
Macrophage
Smooth Muscle
Endothelial

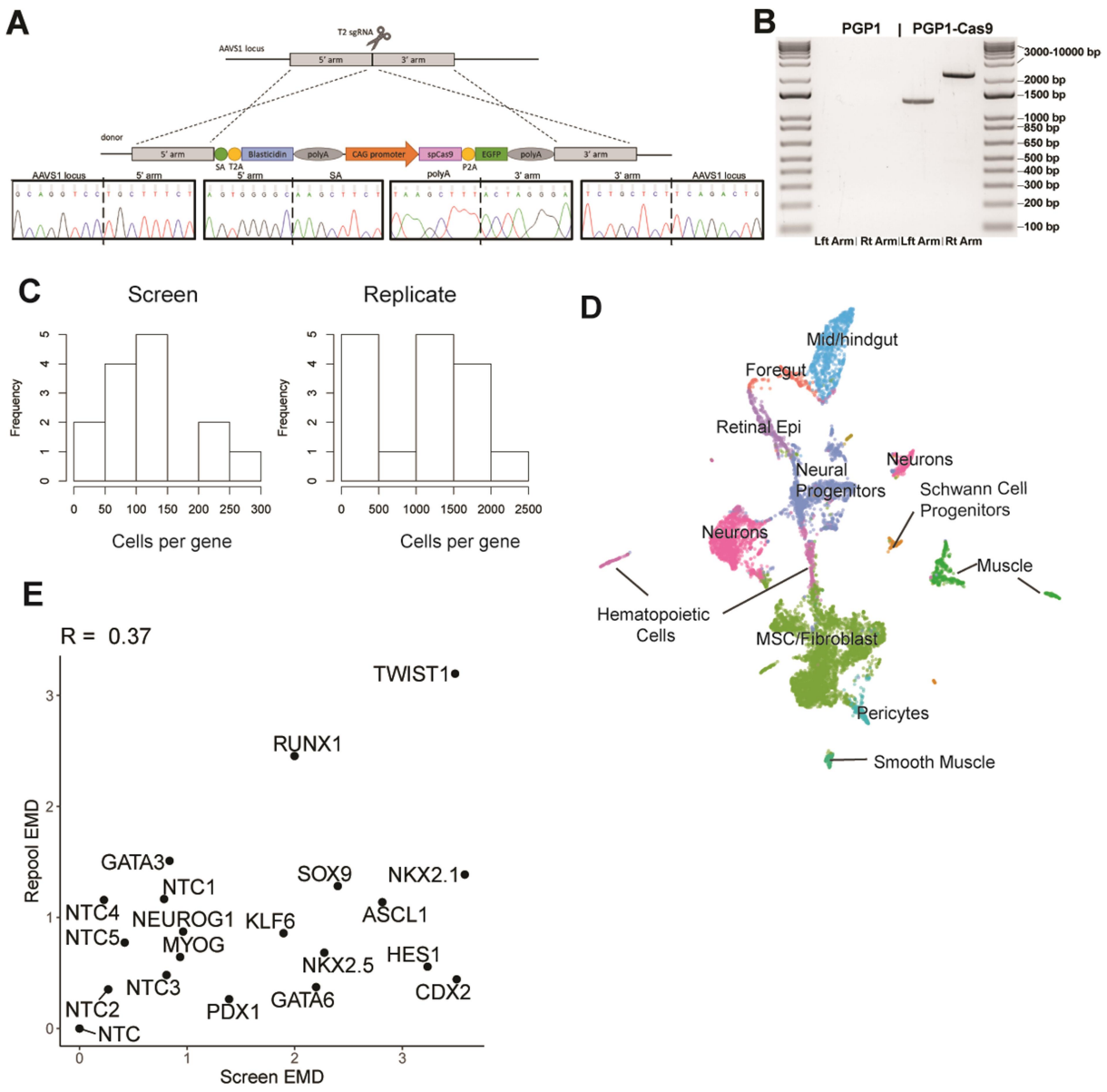


		H1-ter5	H1-ter6	H1-ter7
scRNA-seq	Total Cells	4095	5702	5396
	Barcode Expressing Cells	1041	2600	2781
	Barcoded Fraction	0.254	0.456	0.515
bulk gDNA	Pre-injection Barcodes	238860	101095	112223
	Post-injection Barcodes	116085	60501	61050
	Fraction of Barcodes Detected Post-Injection	0.486	0.598	0.544

Supplementary Figure 4: Assaying teratoma heterogeneity. (A) UMAP scatterplot showing how each line (HUES62, PGP1, and H9) contributes to the various cell type clusters. **(B)** Left: the normalized proportion of each teratoma in every cell type. Right: the bias each cell type shows towards specific teratomas. A low bias score means the cell type is well mixed across all 7 teratomas. **(C)** Growth kinetics of 6 teratomas based on cell line (HUES62, PGP1, and H9). **(D)** Lentiviral barcode construct map. **(E)** Barcoding summary statistics for both bulk and single cell assays across the three barcoded teratomas.

Supplementary Figure 5: Assaying teratoma maturity. (A) A heatmap of log fold-changes for the top differentially expressed genes between matched teratoma neuro-ectoderm and fetal cortical cell types. **(B)** A heatmap of the enrichment scores for top differential genesets (via Geneset Enrichment Analysis) between matched teratoma neuro-ectoderm and fetal cortical cell types. **(C)** Cosine similarity of teratoma gut cells with fetal gut cells of different ages. **(D)** Projection of fetal gut cell types onto a teratoma gut SWNE embedding. **(E)** Correlation of the scaled expression of key marker genes across mid/hindgut and foregut between teratoma and fetal cell types. **(F)** Proportion of foregut and mid/hindgut cells in the teratoma and fetal gut. **(G)** A heatmap of log fold-changes for the top differentially expressed genes between matched teratoma gut and fetal gut cell types. **(H)** A heatmap of the enrichment scores for top differential genesets (via Geneset Enrichment Analysis) between matched teratoma gut and fetal gut cell types. **(I)** H&E stain (left) as well as FOXJ1 staining of ciliated respiratory epithelium at 20x and 63x (right). Scalebar = 50μM (20x) and 20μM (63x). **(J)** Positive (left) and negative (right) RNAScope® control staining. DAPI is a nuclear stain. 4-10 punctate dots/cell is a positive result. Scalebar = 50μM.


A
Top DE genes
SYNE2
PTPRZ1
NRG1
SLA
MEG3
RPS17
HIST1H1D
BEX3
RACK1
XIST
MALAT1
NSG2
ID3
H3F3A
CRABP1
TTR
NGFRAP1
SLC25A6
GNB2L1
Cycling Prog.
Early Neurons
Interneurons
Radial Glia
ter vs fetal logFC
10
5
0
−5
−10
Higher ter expr
B
Top genesets
Response to endoplasmic reticulum stress
Regulation of organ morphogenesis
Anterior posterior pattern specification
Skeletal system morphogenesis
Sensory organ morphogenesis
Cell fate commitment
Protein polyubiquitination
Nucleoside monophosphate metabolic process
TRNA methylation
RNA methylation
Protein alkylation
DNA alkylation
DNA methylation or demethylation
Methionine metabolic process
Folic acid metabolic process
Methylation
Macromolecule methylation
Cycling Prog.
Early Neurons
Interneurons
Radial Glia
ter vs fetal Enrich Score
2.5
0.0
−2.5
C
Fetal gut age
25w
24w
21w
19w
16w
14w
11w
9w
8w
7w
6w
−0.6
−0.3
0.0
0.3
0.6
Cosine similarity to teratoma gut cells
D
Mid/Hindgut: R = 0.98
Teratoma Expression
Fetal Gut Expression
CDX1
CDX2
FOXJ1
HHEX
PAX9
SOX2
Foregut
Mid/Hindgut
Foregut: R = 0.98
Teratoma Expression
Fetal Gut Expression
SOX2
PAX9
HHEX
FOXJ1
CDX2
CDX1
E
Fetal gut SWNE
HHEX
Foregut
Mid/Hindgut
CDX1
CDX2
SOX2
FOXJ1
PAX9
Projected teratoma gut cell types
HHEX
Foregut
CDX1
Mid/Hindgut
CDX2
SOX2
FOXJ1
PAX9
F
Fetal Gut
Teratoma
Cell_Type
Mid/Hindgut
Foregut
0.00
0.25
0.50
0.75
1.00
G
Top DE genes
COL1A1
RPL13A
ACTG2
UGDH−AS1
HBA2
HBA1
HBG2
MT1G
H3F3A
GNB2L1
SLC25A6
RPL21
IGF2
Foregut
Mid/Hindgut
ter vs fetal logFC
10
5
0
−5
−10
H
Top genesets
DNA methylation or demethylation
DNA alkylation
RNA methylation
Protein alkylation
Cellular response to zinc ion
Macromolecule methylation
RRNA metabolic process
Nuclear export
RNA localization
DNA replication
RNA catabolic process
Protein polyubiquitination
Nucleobase compound transport
Foregut
Mid/Hindgut
ter vs fetal Enrich Score
2
0
−2
I
H&E
FOXJ1/DAPI
FOXJ1/DAPI
J
POLR2A/PPIB/UBC/DAPI
DapB/DapB/DapB/DAPI



Supplementary Figure 6: CRISPR-KO Screen in Teratomas. (A) Schematic showing knock-in of the CAG-spCas9-P2A-EGFP cassette with an upstream T2A linked blasticidin resistance gene into the AAVS1 locus thus, creating the Cas9-expressing PGP1 line (above). Accompanying validated trace sequences of the left and right arms (below). **(B)** 2% agarose gel confirming proper integration of the CAG-spCas9-P2A-EGFP cassette into the AAVS1 locus of the PGP1 line via PCR amplification of the left and right arm spanning the endogenous locus and the engineered cassette compared to a PGP1 negative control. **(C)** Cells per gRNA and cells per gene for the screen. **(D)** UMAP projection of PGP1 cell types classified using the H1 cell types as a reference. **(E)** Correlation of the Earth Mover's Distance (EMD) between each gene and the NTC across the original screen and the replicate screen.

Table 1: Developmental Genes for Screen

Human Gene ID	Target Gene Symbol	Lineage	KO phenotype	References (PMID)
64321	Sox17	all endoderm	die at E10.5, deficient of gut endoderm. In the chimeras, few in foregut and completely excluded from the mid- and hindgut	11973269
1045	Cdx2	intestines (also VE)	rescue the embryos to E11.5 in tetraploid chimera, defect in yolk sac circulation	15136723
3172	Hnf4a	liver, pancreas, colon (also VE)	rescue the embryos to midgestation stages (E12.5) in tetraploid chimera	10691738
2626	Gata4	heart (also VE)	rescue the embryos to E9.5 in tetraploid chimera	15310850
2627	Gata6	liver (also VE)	rescue the embryos to E10.5 in tetraploid chimera	15767668
861	Runx1	hematopoiesis	die at E11.5	8565077
3170	Foxa2	notochord, all germ layer	die at E11.5. Absence of head process and notochord	8069909
3651	Pdx1	pancreas	die within a few days after birth	7935793
7080	Nkx2-1	lung	die at birth	10706142
1482	Nkx2-5	heart	die at E9-10, heart looping morphogenesis defect	7628699
6662	Sox9	ductal system, chondrocyte	die at E11.5. In chimeras, excluded from chondrogenic mesenchymal	10319868
5629	Prox1	lymphatic endothelial, liver, lens	die at E14.5	10080188, 10499794
6615	Snai1	EMT	no mesoderm, die at E8.5	11689706
7291	Twist1	EMT	die at E11.5, defects in head mesenchyme	7729687
429	Ascl1	neural	delayed neuronal differentiation	16677628
4762	Neurog 1	neural	neonatal lethal, fail to generate the proximal subset of cranial sensory neurons	9539122
1316	Klf6	hematopoiesis, yolk sac, liver	die at E12.5	16234353
10365	Klf2	endothelia	loss of vessel tone, die at E9.5; Tie2-cre Klf fl/fl die E14.5	17141159
3280	Hes1	brain	die at E12-birth, lethal due to severe neural tube defects	8543157
2290	Foxg1	brain	die at birth, excess of Cajal-Retzius neuron, repression of cortical fate	14704420
7289	Tulp3	neural	die at E14.5, betaIII-tubulin positive cells is significantly decreased in the hindbrain	11406614
4656	MyoG	muscle	die immediately after birth with severe skeletal muscle deficiency	8393145
2625	Gata3	T cell development, endothelial lineage	die by 11 days post coitum (d.p.c.)	10835639
2263	Fgfr2	limb formation, skin, kidney, bone	lethality at E10–11 because of failures in the formation of functional placenta. Fail to form limb buds.	26273516

Table 2: Editing Efficiencies of sgRNAs 1-week post transduction in PGP1-PSCs

sgRNA	Editing Rate	Number of Reads	Selected for analysis and repool
ASCL1-1	88.82%	32222	Yes
ASCL1-2	100.00%	193295	
CDX2-1	95.78%	477555	Yes
CDX2-2	84.09%	380163	
FGFR2-1	4.50%	299260	
FGFR2-2	11.89%	351448	
FOXA2-1	47.55%	283036	
FOXA2-2	55.73%	373681	
FOXG1-1	64.73%	350961	
FOXG1-2	46.34%	308417	
GATA3-1	70.51%	242985	Yes
GATA3-2	55.86%	288202	
GATA4-1	61.46%	188486	
GATA4-2	33.37%	172848	
GATA6-1	68.34%	252259	
GATA6-2	100.00%	4056	Yes
HES1-1	99.97%	32556	Yes
HES1-2	97.99%	2535	
HNF4A-1	ND	ND	
HNF4A-2	35.98%	369691	
KLF2-1	92.00%	120075	Yes
KLF2-1	47.75%	143242	
KLF6-1	86.19%	354299	Yes
KLF6-2	71.22%	314585	
MYOG-1	92.11%	368909	Yes
MYOG-2	83.86%	462448	
NEUROG1-1	40.20%	456686	
NEUROG1-2	95.34%	120075	Yes
NKX2-1-1	58.44%	323414	
NKX2-1-2	75.11%	353478	Yes
NKX2-5-1	92.35%	43639	
NKX2-5-2	100.00%	13884	Yes
PDX1-1	100.00%	21621	
PDX1-2	94.55%	2201	Yes
PROX1-1	53.05%	381734	
PROX1-2	66.88%	449641	

Table 2: Editing Efficiencies of sgRNAs 1-week post transduction in PGP1-PSCs, continued

sgRNA	Editing Rate	Number of Reads	Selected for analysis and repool
RUNX1-1	69.57%	406677	
RUNX1-2	99.99%	286718	Yes
SNAI1-1	36.09%	306442	
SNAI1-2	62.92%	353088	
SOX17-1	34.28%	354167	
SOX17-2	36.82%	455174	
SOX9-1	69.91%	290480	Yes
SOX9-2	90.23%	334831	
TULP3-1	66.87%	317228	
TULP3-2	98.94%	2457	Yes
TWIST1-1	80.20%	155792	
TWIST1-2	70.21%	324104	Yes

2.5 Materials and Methods

2.5.1 Cell Culture

The H1, H9, PGP1, and HUES62 hESC cell line was maintained under feeder-free conditions in mTeSR medium (Stem Cell Technologies). Prior to passaging, tissue-culture plates were coated with growth factor-reduced Matrigel (Corning) diluted in DMEM/F-12/Glutamax medium (Thermo Fisher Scientific), and incubated for 30 minutes at 37ºC, 5% CO2. Cells were dissociated and passaged using the dissociation reagent Versene (Thermo Fisher Scientific). HEK 293T and HeLa were maintained in high glucose DMEM supplemented with 10% fetal bovine serum (FBS) and passaged every couple days upon confluency with .05% Trypsin-EDTA (Gibco). HUVECs were maintained in EGM-2 (Lonza).

2.5.2 PGP1-Cas9 Clone Generation

The PGP1 human induced pluripotent stem cell line was a kind gift of Dr. George Church at Harvard Medical School. The sgRNA targeting AAVS1 locus of the human genome (spacer sequence GGGCCACTAGGGACAGGAT) was cloned into the Lenti-guide-puro plasmid (Addgene #52963). To generate the knockin donor plasmid, we cloned the CAG promoter followed by a cassette of co-expression of spCas9 and EGFP splitting via the P2A sequence into the pCR4-Blunt-TOPO vector (Thermo Fisher Scientific). Two homology arms were amplified from upstream (804 bp) and downstream (837 bp) of the sgRNA targeting site in AAVS1 genomic locus and constructed into the donor plasmid flanking the CAG-spCas9-P2A-EGFP cassette. Between the upstream homology arm and the CAG promoter, we inserted a splice acceptor sequence following by a T2A linked blasticidin resistance gene.

Human iPSC PGP1 cells were electroporated using 4D-Nucleofector system and P3 Primary Cell X kit (Lonza) according to the manufacturer's instruction. Briefly, the PGP1 cells were dissociated into single cells. $1x10^6$ cells were mixed with 100 µl nucleofection reagents and 10 µg DNA (5 µg Cas9 donor + 5 µg sgRNA) and electroporated. The cells were recovered with pre-warmed medium and then cultured on inactivated MEF feeders in 10 cm dishes with mTeSR medium supplemented with 0.5 µM ROCK-inhibitor. Afterward, the mTeSR medium without ROCK-inhibitor was refreshed daily. 2 µg/ml blasticidin were added into the culture medium 7 days after electroporation. The cells were cultured without passage until clones emerged on the plate. The clones were checked under the microscope and those with EGFP expression were picked up and expanded individually.

To detect genomic integration, the genomic DNA from cultured cells was extracted using DNeasy Blood & Tissue Kits (Qiagen). Approximately 500 ng of genomic DNA was used for each PCR reaction using KAPA HiFi HotStart Ready Mix (Kapa Biosystems). The primer sequences are listed below.

Left_arm_forward	ACTTCCCCTCTTCCGATGTTG
Left_arm_reverse	ATTGTAGCCGTTGCTCTTTCA
Right_arm_forward	GAGCAAAGACCCCAACGAGAAGC
Right_arm_reverse	CTGCCTGGAGAAGGATGCAGGA

The activity of Cas9 in the PGP1-Cas9 cells was further validated by the generation of indels at the expected position when guide RNAs were introduced.

2.5.3 sgRNA Design

The CRISPR-KO sgRNA sequences targeting transcription factor genes were obtained from the GPP sgRNA Designer web tool (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design, accessed February

2018) as follows. The 24 gene symbols in the table below were converted to Entrez gene IDs using Bioconductor package org.Hs.eg.db_3.5.0, and the resulting IDs were submitted together with the following parameters: enzyme Sp, taxon human, quota 50, include unpicked. From the resulting output, the two guide sequences with the highest “pick order” were selected for each target gene. To check the validity of each guide sequence, the corresponding context sequence was compared to the human reference genome at the predicted cut location using Bioconductor package BSgenome.Hsapiens.UCSC.hg38_1.4.1, and the cut location was confirmed to be fully within the target gene coding sequence determined using Bioconductor package TxDb.Hsapiens.UCSC.hg38.knownGene_3.4.0.

Gene symbol	**Entrez ID**	**sgRNA-1**	**sgRNA-2**
SOX17	64321	GGCAACGGGTAGCCGTCGAG	AGGGCGAGTCCCGTATCCGG
CDX2	1045	CCGCAGTACCCGGACTACGG	CAAATATCGAGTGGTGTACA
HNF4A	3172	GGGACCGGATCAGCACTCGA	GCAATGACTACATTGTCCCT
GATA4	2626	TGTGGGCACGTAGACTGGCG	CCGGCTTACATGGCCGACGT
GATA6	2627	CGGGACGCCTCAGCTCGACA	GCCGACAGCGAGCTGTACTG
RUNX1	861	CTGATCGTAGGACCACGGTG	TGCTCCCCACAATAGGACAT
FOXA2	3170	ATGAACATGTCGTCGTACGT	TCCGTGAGCAACATGAACGC
PDX1	3651	GGAGAACAAGCGGACGCGCA	TATTCAACAAGTACATCTCA
NKX2-1	7080	GCGAGCGGCATGAACATGAG	GGTTGGCGCCGTACCATCCG
NKX2-5	1482	GTAGGCACGTGGATAGAAGG	GAAGACAGAGGCGGACAACG
SOX9	6662	ACGTCGCGGAAGTCGATAGG	TTCACCGACTTCCTCCGCCG
PROX1	5629	AGTGTCCACAACTTGCGACA	CGGGTTGAGAATATAATTCG
SNAI1	6615	GGGACTCTCCTGGAGCCGAA	TGTAGTTAGGCTTCCGATTG
TWIST1	7291	CGGGAGTCCGCAGTCTTACG	AGCGGGTCATGGCCAACGTG
ASCL1	429	CCAGGTTGACCAACTTGACG	AAACGCCGGCTCAACTTCAG
NEUROG1	4762	CCGCATGCACAACTTGAACG	TTGGTGTCGTCGGGGAACGA
KLF6	1316	TCTGAGGCTGAAACATAGCA	GCTGACCAAAACTTCGCCAA
KLF2	10365	GGTTCGGGGTAATAGAACGC	CTTCGGTCTCTTCGACGACG
HES1	3280	GTGCGAGGGCGTTAATACCG	AGCCAGTGTCAACACGACAC
FOXG1	2290	AGCGCGTTGTAGCTGAACGG	CCGCGCCACTACGACGACCC
TULP3	7289	GGAGTATGACAGTTCACCAA	TGAAAGTGTGAACTTCGATG

MYOG	4656	TTACACACCTTACACGCCCA	TCGAACCACCAGGCTACGAG
GATA3	2625	TCCAAGACGTCCATCCACCA	CAGGGAGTGTGTGAACTGTG
FGFR2	2263	CTTAGTCCAACTGATCACGG	TGACCAAACGTATCCCCCTG

2.5.4 Library Preparation

The lentiviral backbone plasmid for the barcode vector was constructed containing the EF1α promoter, mCherry transgene flanked by BamHI restriction sites, followed by a P2A peptide and hygromycin resistance enzyme gene immediately downstream (ECIH). The backbone was digested with HpaI, and a pool of 20 bp long barcodes with flanking sequences compatible with the HpaI site, was inserted immediately downstream of the hygromycin resistance gene by Gibson assembly. The vector was constructed such that the barcodes were located only 200 bp upstream of the 3'-LTR region. This design enabled the barcodes to be transcribed near the poly-adenylation tail of the transcripts and a high fraction of barcodes to be captured during sample processing for scRNA-seq.

The lentiviral backbone plasmid for the sgRNAs was the CROPseq-Guide-Puro vector (Addgene #86708). To create the sgRNA library, individual sgRNAs were PCR amplified utilizing overlapping forward and reverse primers custom designed with flanking sequences compatible with the BSMBI restriction sites. The lentiviral backbone was digested with BSMBI (New England Biolabs) at 55°C for 3 hours in a reaction consisting of: CROPseq-Guide-Puro backbone, 5 µg, Buffer NEB 3.1, 5 µl, BSMBI, 5 µl, H20 up to 50 µl. After digestion, the vector was purified using a QIAquick PCR Purification Kit (Qiagen). Each sgRNA was then individually assembled via Gibson assembly.

The Gibson assembly reactions were set up as follows: 1:10 molar ratio of digested backbone to sgRNA insert, 2X Gibson assembly master mix (New England Biolabs), H20 up to 20 µl. After incubation at 50°C for 1 h, the product was transformed into One Shot Stbl3 chemically competent Escherichia coli (Invitrogen). A fraction (150 µL) of cultures was

spread on carbenicillin (50 µg/ml) LB plates and incubated overnight at 37°C for 15-18hrs (miRNA constructs required longer incubation times). Individual colonies were picked, introduced into 5 ml of carbenicillin (50 µg/ml) LB medium and incubated overnight in a shaker at 37°C. The plasmid DNA was then extracted with a QIAprep Spin Miniprep Kit (Qiagen), and Sanger sequenced to verify correct assembly of the vector and to extract barcode sequences.

To assemble the library, individual sgRNA vectors were pooled together in an equal mass ratio along with 5 non-targeting control (NTC) sgRNAs which constituted 50% of the final pool.

2.5.5 Viral Production

HEK 293T cells were maintained in high glucose DMEM supplemented with 10% fetal bovine serum (FBS). Cells were seeded in a 15 cm dish 1 day prior to transfection, such that they were 60-70% confluent at the time of transfection. For each 15 cm dish 36 µl of Lipofectamine 2000 (Life Technologies) was added to 1.5 ml of Opti-MEM (Life Technologies). Separately 3 µg of pMD2.G (Addgene #12259), 12 µg of pCMV delta R8.2 (Addgene #12263) and 9 µg of an individual vector or pooled vector library was added to 1.5 ml of Opti-MEM. After 5 minutes of incubation at room temperature, the Lipofectamine 2000 and DNA solutions were mixed and incubated at room temperature for 30 minutes. Medium in each 15 cm dish was replenished with 25 ml of fresh medium. After the incubation period, the mixture was added dropwise to each dish of HEK 293T cells. Supernatant containing the viral particles was harvested after 48 and 72 hours, filtered with 0.45 µm filters (Steriflip, Millipore), and further concentrated using Amicon Ultra-15 centrifugal ultrafilters with a 100,000 NMWL cutoff (Millipore) to a final volume of 600-800 µl, divided into aliquots and frozen at -80°C.

2.5.6 Viral Transduction

For viral transduction, virus was added at a low MOI (ensuring a single barcode/cell or a single sgRNA/cell) to stem cells at 20% confluency alongside polybrene (5 μg/ml, Millipore) in fresh mTeSR medium. The following day, medium was replaced with fresh mTeSR. Appropriate selection reagent was added 48 hrs after transduction (hygromycin [50μg/μL] for barcode / puromycin [0.75μg/μL] for CRISPR KO screen / miRNA-HSV-tk-GFP) (Thermo Fisher Scientific) and was replaced daily. For miRNA-HSV-tk-GFP transduced cells puromycin selection did not begin until 5-7 days after transduction to allow for enough GFP positive cells. For editing in CRISPR KO screen, selection was continued for 5 days prior to use for teratoma formation in mice.

2.5.7 sgRNA Editing Rate Validation

We individually transduced each sgRNA into our PGP-Cas9 cell line in an arrayed format and selected with puromycin after 48 hrs and allowed editing to occur for an additional 5 days (7 days total). From there we retrieved the cell pellets from each individual sgRNA and extracted gDNA. We then designed primers upstream and downstream of the expected cut site for each individual sgRNA and amplified that region utilizing standard PCR on the gDNA extracted from each cell pellet transduced with each individual sgRNA. Each amplicon for each sgRNA was then sent out for deep sequencing. The reads we got back were then compared to a reference genome and the reads that showed a proper indel or SNP were counted as a proper cut and edit. This number was compared to overall reads and the percentage was calculated as the effificnecy of edits made for that single sgRNA onto our PGP-Cas9 line.

2.5.8 Animals

Housing, husbandry and all procedures involving animals used in this study were performed in compliance with protocols (#S16003) approved by the University of California San Diego Institutional Animal Care and Use Committee (UCSD IACUC). Mice were group housed (up to 4 animals per cage) on a 12:12 hr light-dark cycle, with free access to food and water in individually ventilated specific pathogen free (SPF) cages. All mice used were healthy and were not involved in any previous procedures nor drug treatment unless indicated otherwise. All studies performed in NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG) mice and maintained in autoclaved cages.

2.5.9 Teratoma Formation

A subcutaneous injection of 5-10 million PSCs in a slurry of Matrigel® and mTeSR medium (1:1) was made in the right flank of anesthetized $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice. Weekly monitoring of teratoma growth was made by quantifying approximate elliptical area (mm^2) with the use of calipers measuring outward width and height.

2.5.10 Teratoma Processing

After growth for 70 days on average mice were euthanized by slow release of CO2 followed by secondary means via cervical dislocation. Tumor area was shaved, sprayed with 70% ethanol, and then extracted via surgical excision using scissors and forceps. Tumor was rinsed with PBS, weighed, and photographed. Tumor was then cut into small pieces in a semi-random fashion and frozen in OCT for sectioning and H&E staining courtesy of the Moore's Cancer Center Histology Core. Remaining tumor was cut into small pieces 1-2mm in diameter and subjected to standard GentlaMACS™ protocols: Human Tumor Dissociation Kit (medium tumor settings), Red Blood Cell Lysis Kit, and Dead Cell Removal Kit. Single cells

were then resuspended in .04% BSA for 10X genomics chromium(Zheng et al. 2017) platform.

2.5.11 Histology and RNAScope®

H&E sections from teratomas were confirmed to have the presence of all 3 germ layers: ectoderm, mesoderm, and endoderm via microscopy identification. Fresh frozen sections were subjected to standard RNAScope® Fluorescent Multiplex Reagent Kit protocols following fresh frozen tissue requirements.

2.5.12 Microscopy

Following 24 hrs of incubation with RNAScope® probes in 4ºC, slides were imaged using Zeiss 880 Airyscan Confocal microscope with special thanks to Michael Hu for image processing utilizing the UC San Diego Microscopy Core.

2.5.13 Single cell RNA-seq Processing and Genotype Deconvolution

Using the 10X genomics CellRanger (v2.01) pipeline(Zheng et al. 2017), we aligned Fastq files to a combined hg19 and mm10 reference, counted UMIs to generate human and mouse gene-expression counts matrices, and aggregated samples across 10X runs with the cellranger aggr command. All cellranger commands were run using default settings.

2.5.14 Seurat Data Integration

Data integration was performed on the aggregated counts matrices for each of the following datasets: the 7 H1 teratomas, the 3 PGP1 CRISPR-KO screen teratomas, and the 4 chimeric teratomas. We used the Seurat v3 data integration pipeline(Stuart et al. 2018; Butler et al. 2018). Briefly, we first filtered the counts matrix for genes that are expressed in

at least 0.1% of cells, and cells that express at least 200 genes. We then normalized the counts matrix using total-counts normalization, and log-transformed the result. For each teratoma, we identified highly variable genes, and selected the top 4000 genes that appeared as overdispersed across the most teratomas. We then identified anchor cells, and integrated the teratomas to create a batch-corrected gene expression matrix. After batch correction, we used a linear model to regress away library depth, and mitochondrial gene fraction, and ran Principal Components Analysis (PCA)(Abdi and Williams 2010), keeping the first 30 principal components. We then used the PCs to generate a k Nearest Neighbors (kNN) graph, setting k = 10, and then used the kNN graph to calculate a shared nearest neighbors (SNN) graph(Houle et al. 2010b). We ran modularity optimization algorithm with a resolution of 0.4 on the SNN graph to find clusters(Butler et al. 2018).

2.5.15 H1 Teratoma Clustering and Validation

H1 clusters were assigned to cell types using a two-stage strategy. First, we trained a kNN classifier on the Mouse Cell Atlas dataset using k = 40(Tarlow et al. 2013), mapping mouse genes to their human orthologs. We projected each cell in the teratoma dataset onto the first 40 Principal Components (PCs) of the Mouse Cell Atlas and classified each cell in the H1 teratoma dataset using this kNN classifier to generate a rough set of cell type assignments for each cluster. We then manually inspected the marker genes for each cluster and adjusted the cell type based on the expression of canonical markers (**Table 2**). Clusters that mapped to the same MCA cell type, and expressed similar marker genes were merged. Finally, we ran UMAP on the first 30 PCs as input in order to visualize the results(McInnes and Healy 2018; Becht et al. 2018). We validated each annotated cell type by computing the pearson correlation between the average expression of each cell type and the average expression of each broad cell type in the Mouse Organogenesis Cell Atlas(Cao et al. 2019).

We used the union of all marker genes for the teratoma cell types and Mouse Organogenesis Cell Atlas cell types to perform the correlation analysis.

In some cases, it was necessary to sub-cluster the cells to achieve greater cell type resolution. Specifically we sub-clustered the all cells mapping to ciliated epithelium in order to separate retinal epithelium and airway epithelium. Additionally, we sub-clustered the neuro-ectoderm in order to identify interneurons, peripheral neurons, retinal progenitors, and early neuro-ectoderm. In both cases we simply subsetted the gene expression matrix with the cells of interest and reran the Seurat analysis pipeline, identifying sub-clusters using known marker genes (Supplemental Table 2).

2.5.16 Quantitative Assessment of Teratoma Heterogeneity and Cell Type Bias

In order to quantify the level of heterogeneity between teratomas we used the Normalized Relative Entropy metric from CONOS(Barkas et al. 2019).

$$1 - \frac{\sum_{k=1}^{n_{clu}} s_k \times KL(\boldsymbol{f_k}, \boldsymbol{F})}{\log(n_{teratomas}) \sum_{k=1}^{n_{clust}} s_k}$$

Where $\boldsymbol{f_k}$ is a vector with the number of cells in each teratoma from cluster *k,* $KL(\boldsymbol{f_k}, \boldsymbol{F})$ is the empirical KL divergence between $\boldsymbol{f_k}$ and the total number of cells in each teratoma, $\boldsymbol{F}$. Higher Normalized Relative Entropy means the cell types are more mixed across the teratomas and thus the teratomas are less heterogeneous.

To quantify the heterogeneity/bias of individual cell types across teratomas we simply take the KL divergence of the number of cells in each teratoma from that cell type/cluster and the total number of cells in each teratoma and then scale by the number of cells in each cell type. For each cell type *k*:

$$s_k \times KL(\boldsymbol{f_k}, \boldsymbol{F})$$

2.5.17 Lentiviral Barcode and CRISPR Guide Assignment

To assign one or more lentiviral/gRNA barcode to each cell, we extracted each barcode by identifying its flanking sequences, resulting in reads that contain cell, UMI, and barcode tags. To remove potential chimeric reads, we used a two-step filtering process. First, we only kept barcodes that made up at least 0.5% of the total amount of reads for each cell. We then counted the number of UMIs and reads for each plasmid barcode within each cell, and only assigned that cell any barcode that contained at least 10% of the cell's read and UMI counts. The code for assigning barcodes to each cell can be found on GitHub at: https://github.com/yanwu2014/genotyping-matrices.

2.5.18 H1 Cell Barcoding Analysis

We extracted lentiviral barcodes from the genomic DNA fastq files before and after teratoma formation for the 3 barcoded H1 teratomas. We counted the number of unique barcodes that were supported by at least 10 reads (the reads requirement is to mitigate overcounting unique barcodes due to minor sequencing errors) and then computed the fraction of unique barcodes that remain after teratoma formation to assess the approximate number of cells that are involved in the teratoma formation process.

We also identified lentiviral barcodes at the single cell level, using the barcode assignment strategy described in the Lentiviral Barcode and CRISPR Guide Assignment section. For each cell type, we computed its bias for specific barcodes using the same relative entropy metric we used to compute teratoma bias.

$$s_k \times KL(\boldsymbol{b_k}, \boldsymbol{B})$$

Where $\boldsymbol{b_k}$ is a vector with the number of cells in each barcode from cluster *k,* $KL(\boldsymbol{b_k}, \boldsymbol{B})$ is the empirical KL divergence between $\boldsymbol{b_k}$ and the total number of cells in each barcode, $\boldsymbol{B}$.

2.5.19 Developmental Staging Analysis

In order to assess the developmental maturity of the teratoma cell types, we computed the average expression of all cells related to neuro-ectoderm (Radial Glia, Intermediate Neuronal Progenitors, Early Neurons) and gut (Oral/Esophageal, Stomach, Intestine) cell types and calculated the cosine similarity of the teratoma average expression to the average expression of fetal human cells across different time points. We used all genes that were detected in both the fetal and teratoma data.

For the neuro-ectoderm cells, we then sub-clustered those cells and identified additional cell types using canonical marker genes (Supplemental Table 2). We then matched those neuro-ectoderm sub-clustered cell types to cell types in a larger fetal week 17-18 single cell prefrontal cortex dataset.

We next generated Similarity Weighted Nonnegative Embeddings (SWNE)(Wu, Tamayo, and Zhang 2018) for the neuronal and gut cell types using the top 3000 overdispersed genes in each tissue type. Briefly, SWNE uses nonnegative matrix factorization (NMF)(Lee and Seung 1999) to decompose a gene expression matrix into component factors, embeds the factors in 2D using sammon mapping(Sammon 1969), and embeds the cells and key genes in the 2D space relative to the factors. The cell positions are smoothed using a shared nearest neighbors (SNN) network. For the neuronal SWNE embedding, we used 30 NMF factors and 20 nearest neighbors when computing the SNN. For the gut SWNE embedding, we used 20 NMF factors and 30 nearest neighbors. We projected teratoma data onto the fetal SWNE, by first projecting the teratoma data onto the fetal NMF factors, and generating embedding coordinates. We then smooth the projected coordinates by projecting the teratoma data onto the fetal SNN.

We then compared the expression of key neuronal/gut marker genes in each neuronal and gut cell type by correlating the expression of those markers between the teratoma data and the fetal human data. We used the scaled gene expression for both the teratoma and fetal data, which involves subtracting the average expression and dividing by the standard deviation.

2.5.20 PGP1 Teratoma Screen Analysis

We first generated a batch corrected gene expression matrix using the method described in the Seurat Data Integration section. We projected each cell in the PGP1 dataset onto the first 30 principal components of the H1 teratoma dataset, and classified each cell using Seurat Label Transfer, collapsing cell types that had fewer than 100 cells into their closest cell type(Stuart et al. 2019). Specifically, Airway Epithelium was merged into Foregut (Airway epithelium is derived from the foregut epithelium during development), Schwann Cells and Melanoblasts were grouped as Schwann Cell Progenitors (SCP), Immune Cells, Erythrocytes, and Hematopoietic Stem Cells (HSCs) were grouped as Hematopoietic cells, Muscle Progenitors and Cardiac/Skeletal Muscle were grouped as Muscle, all MSC/Fibroblast populations were merged, Pericytes and Smooth Muscle were grouped as Perivascular Cells, Intermediate Neuronal Progenitors (INP) and Radial Glia were grouped as Neuronal Progenitors, and Retinal Neurons were merged into Early Neurons. In order to visualize the PGP1 data, we ran UMAP on the projected PCs.

We assigned CRISPR-KO gene perturbations using the barcode assignment strategy described in the Lentiviral Barcode and CRISPR Guide Assignment section. To assess the effect of gene knockouts on cell type proportions, we used a ridge regression model with the R glmnet package. For each CRISPR gRNA and gene knockout, this resulted in regression coefficients for each cell type describing the enrichment or depletion of that gRNA/gene-KO

in that cell type. We permuted the gRNA/gene assignments to assign p-values to each coefficient representing the probability that coefficient is truly non-zero. For each gene, we computed the cell type shift effect size as the sum of the absolute value of each coefficient across all cell types with an unadjusted p-value of less than 0.005. The reproducibility of the gene knockout was assessed by correlating the cell type effects of the two gRNAs associated with that gene. We also used bootstrap resampling with 1000 samples of 90% of the cells in each sample in order to compute a bootstrap standard deviation for the effect size of each gene.

2.6 Acknowledgement for Chapter 2

Chapter 2, in full, is a reprint of unpublished material being prepared for submission (McDonald, Daniella*; Wu, Yan*; Dailamy, Amir; Tat, Justin; Parekh, Udit; Zhao, Dongxin; Hu, Michael; Tipps, Ann; Zhang, Kun; Mali, Prashant. 2020) The dissertation author was a primary author of this paper. Chapter 2, in full, is also a reprint of the material as it appears in Cell Systems (Parekh, Udit*; Wu, Yan*; Zhao, Dongxin; Worlikar, Atharv; Shah, Neha; Zhang, Kun; Mali, Prashant. 2018)

*These authors contributed equally

CHAPTER 3. Assessing the cell-type specific function of noncoding regulatory regions linked to human evolution using single-cell accessibility and gene expression analysis

3.1 Introduction

Modern humans have a remarkable set of complex social and cognitive behaviors, which is at least partly driven by human specific neuro-development that results in a brain that is exceptional among primates in terms of size, cortical organization, and connectivity(J. a Miller et al. 2014; Rilling 2014; Levchenko et al. 2018; Pääbo 2014). One approach to understanding these human specific developmental changes is to search for “Human Accelerated Regions” (HARs), which are genomic regions that are highly conserved among vertebrates but have higher than expected substitution rates on the human specific lineage of the evolutionary tree(Pollard et al. 2006; Prabhakar et al. 2006; Bird et al. 2007; Bush and Lahn 2008; Lindblad-Toh et al. 2011; Hawkins et al. 2015). The fact that these regions are conserved in vertebrates suggests that they are functionally important regions (such as regulatory regions), and the high substitution rate in the human lineage suggests that their function has been modified specifically in humans(Pääbo 2014).

Massively Parallel Reporter Assays (MPRAs) have shown that many of these HARs function as enhancer regions, with many validated enhancers showing activity during embryonic development(Capra et al. 2013; Ryu et al. 2018). Additionally, HAR specific sequencing in patients with Autism Spectrum Disorder (ASD) coupled with chromatin interaction sequencing (Hi-C) and MPRAs has shown that mutations in HARs that serve as active enhancers can disrupt social and cognitive function(Doan et al. 2016). However, understanding the role of these HARs across the different cell types present during neuronal development remains a challenge. Integrating chromatin interaction maps in the fetal human brain with single-cell RNA-seq profiling of human brain development enabled the

identification of neuro-developmental risk genes expressed in specific developmental cell types that are regulated by HARs(Won et al. 2019). Nevertheless, to our knowledge, there is still no study that directly assays the activity of HARs and expression of HAR-linked genes with developmental cell-type resolution.

We performed single-cell chromatin accessibility and gene expression profiling on the week 16 and week 18 fetal human prefrontal cortex in order to assess both cell type specific activity of HARs and expression of HAR-linked genes. Assaying chromatin accessibility at the single-cell level enables us to determine which HARs and HAR-linked regulatory elements are active in which developmental cell types, while the single-cell RNA sequencing enables us to determine HAR-linked gene expression in those same cell types. Additionally, the single-cell chromatin accessibility dataset enables us to generate a map of co-accessible regions, supplementing existing chromatin interaction maps of the human fetal cortex and enabling us to link HARs to downstream genes with greater accuracy(Pliner et al. 2018; de la Torre-Ubieta et al. 2018).

3.2 Results

3.2.1 Identifying chromatin and RNA cell types

We first profiled the chromatin accessibility of the developing human cortex by processing 40,678 single nuclei from the week 16 and week 18 human fetal human cortex using scTHS-seq(B. B. Lake et al. 2017) **(Figure 9A, Methods)**. We profiled 7,865 nuclei from the same samples using snDropSeq to assess RNA expression, and identified RNA and chromatin cell types and trajectories(B. B. Lake et al. 2017; Macosko et al. 2015) **(Figure 9A, Methods)**. With the chromatin data, we found cell-type specific accessible sites and overlapped them with Human Accelerated Regions mutations to assess the role of these

developmental cell types in human evolution(Pollard et al. 2006; Prabhakar et al. 2006; Bird et al. 2007; Bush and Lahn 2008; Lindblad-Toh et al. 2011; Hawkins et al. 2015) **(Figure 9A)**.

To process the chromatin data, we first used SnapATAC to identify a consensus set of 434,899 peaks and generate a matrix of peaks by cells(Fang et al. 2019) **(Figure 9A, Methods)**. We used LDA to reduce the dimensionality of the peaks matrix via CisTopic and clustered the cells using Seurat(Bravo González-Blas et al. 2019; Stuart et al. 2019) **(Figure 9A, Methods)**. We then identified co-accessible peaks and generated an estimated gene activity matrix using Cicero(Pliner et al. 2018) **(Methods)**. To process the snDropSeq RNA data, we used the DropSeq computational pipeline to generate the gene expression matrix, Pagoda2 for clustering, and Seurat for visualization and marker gene identification(Macosko et al. 2015; Barkas et al. 2018; Stuart et al. 2019) **(Figure 9A, Methods)**.

We identified cell types in the RNA dataset by matching cluster specific marker genes to canonical markers from previous single-cell RNA-seq studies of the human fetal cortex, and correlated the RNA cell type with the chromatin gene activity clusters to generate a rough map of chromatin cell type identity(Polioudakis et al. 2019) (**Figure S7A, Methods)**. We validated the chromatin cell types by visualizing the activity of the same canonical markers(Polioudakis et al. 2019) **(Figure S7B, Methods)**. We further validated the cell types by correlating the average gene activities of the finalized chromatin cell types with the average gene expression of the finalized RNA cell types **(Figure S7C)**.

We then generated UMAP embeddings of the chromatin dataset using the LDA topics and the RNA dataset using the top 50 principal components(McInnes and Healy 2018; Becht et al. 2018) **(Figure 9B, 9C)**. In order to visually separate outer and ventral Radial Glia (oRG and vRG), we subsetted the binary accessibility matrix to just oRG, vRG, and CycProg cells and reran LDA and UMAP using CisTopic **(Figure 9D)**. Both the chromatin and RNA

datasets show distinct trajectories for excitatory neurons, starting from the Radial Glia cell types (oRG, vRG) and moving through intermediate progenitors (ipEx) and early born excitatory neurons (ebEx) to the more mature layer specific excitatory neurons (ExL2/3, ExL4, ExL5/6, ExL6b) **(Figure 9B, 9C)**. Other cell types include inhibitory neurons, which migrate into the cortex from the Medial Ganglionic Eminence (InMGE) or the Caudal Ganglionic Eminence (InCGE) as well as Oligodendrocyte progenitors (OPC) and Microglia (Mic)(Polioudakis et al. 2019) **(Figure 9B, 9C)**. Endothelial cells (End) are detected in the RNA dataset but not the chromatin dataset **(Figure 9B, 9C)**.

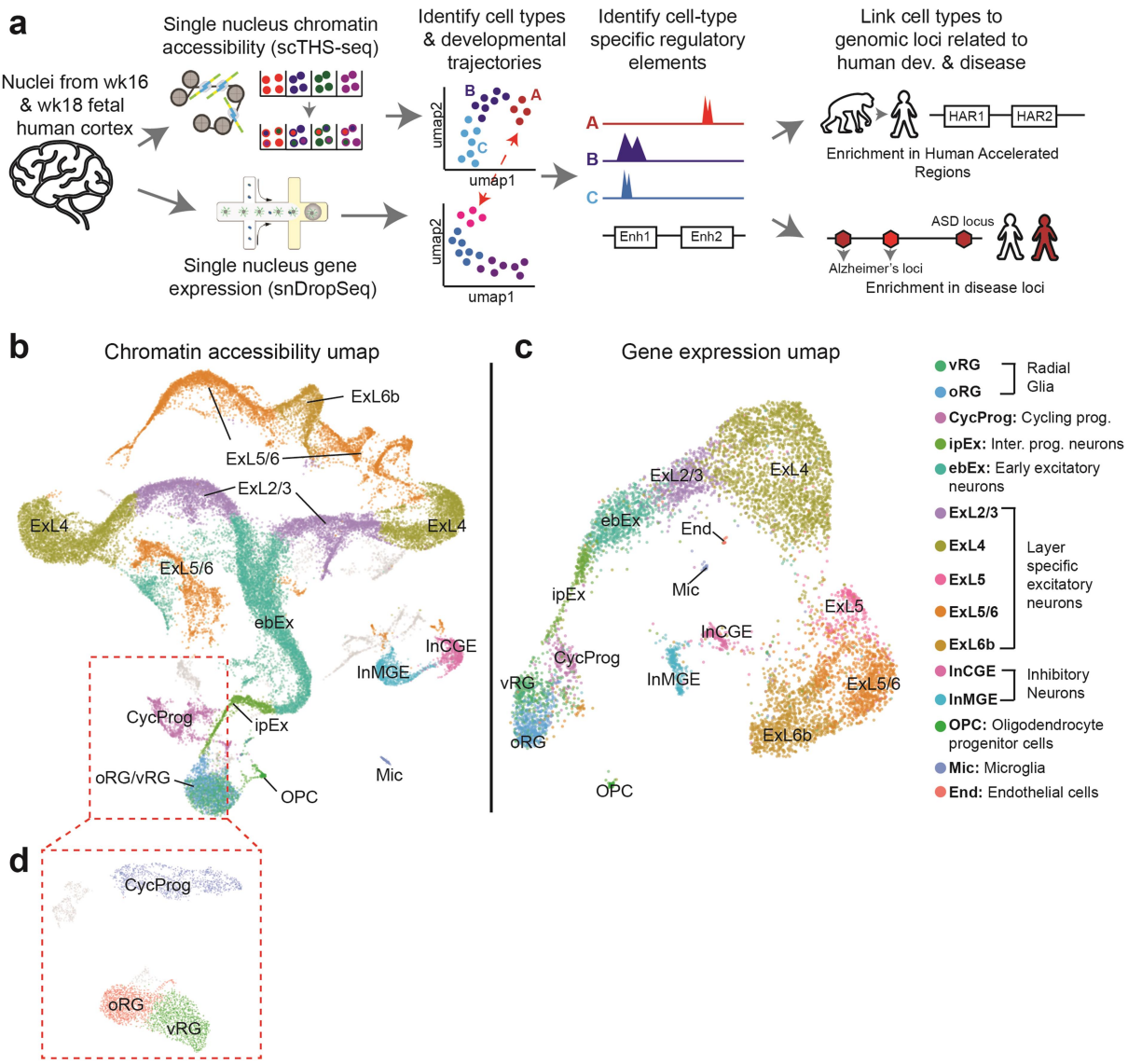


Figure 9: A single-cell map of human cortical development. **(A)** We isolated nuclei from week 16 and week 18 human pre-frontal cortex and ran single-nuclei THS-seq and single-nuclei Drop-Seq in order to assess both the transcriptional and chromatin accessibility states of cells through development. We identified cell types for both the RNA and chromatin datasets, and then computed cell-type specific accessible sites. We then used overlapped those cell type specific accessible sites with Human Accelerated Regions (HARs) and disease-linked SNPs to study the roles these developmental cell types play in human evolution and disease. **(b)** UMAP embedding of the chromatin accessibility (scTHS-seq) dataset. **(c)** UMAP embedding of the RNA (snDrop-Seq) dataset. **(d)** Sub-clustering of cycling progenitors (CycProg), outer Radial Glia (oRG), and ventral Radial Glia (vRG) for the chromatin dataset.

3.2.2 Chromatin Trajectory Inference and GWAS Phenotype Analysis

In order to more closely analyze the developmental progression from Radial Glia progenitors (oRG, vRG) to early excitatory neurons (ebEx), we subsetted the chromatin accessibility dataset to only include oRG, vRG, ipEx, and ebEx cell types **(Methods)**. We then reran LDA using cisTopic and generated a new gene activity matrix using Cicero **(Methods)**. We embedded the developmental cell types using SWNE, a visualization method that enables simultaneous embedding of cells and key features(Wu, Tamayo, and Zhang 2018) **(Figure 10A, Methods)**. We also embedded the promoter and binding site accessibility of NPAS3, NEUROG2, FOXA1, and EOMES, TFs known to play a role in brain development(Elsen et al. 2013; de la Torre-Ubieta et al. 2018; Stott et al. 2013; Aydin et al. 2019; Kamm et al. 2013) **(Figure 10A)**. The closer a group of cells is to the promoter or binding site, the greater the accessibility of that promoter/binding site is in that group of cells **(Figure 10A, Methods)**. Finally, we fit a trajectory using a principal curve, and estimated the pseudotime for each cell as the arc-length of that cell projected onto the principal curve(Hastie and Stuetzle 1989) **(Figure 10A, S8A, Methods)**.

In the SWNE embedding, we can see that for NEUROG2 and FOXA1, the promoter becomes accessible far earlier in the developmental trajectory than their binding sites, which potentially reflects their roles as pioneer transcription factors(Zaret and Mango 2016; Mayran and Drouin 2018) **(Figure 10A, S8A)**. On the other hand, EOMES and NPAS3's promoters and binding sites become accessible at around the same developmental time **(Figure 10A, S8A)**. We confirmed this trend by plotting the relative accessibility of the promoters and binding sites of EOMES, FOXA1, NEUROG2, and NPAS3 across pseudotime **(Figure 10B, S8B)**. Overall, our SWNE embedding of the excitatory neuron developmental trajectory enabled the visualization of transcription factor promoter and binding accessibility dynamics.

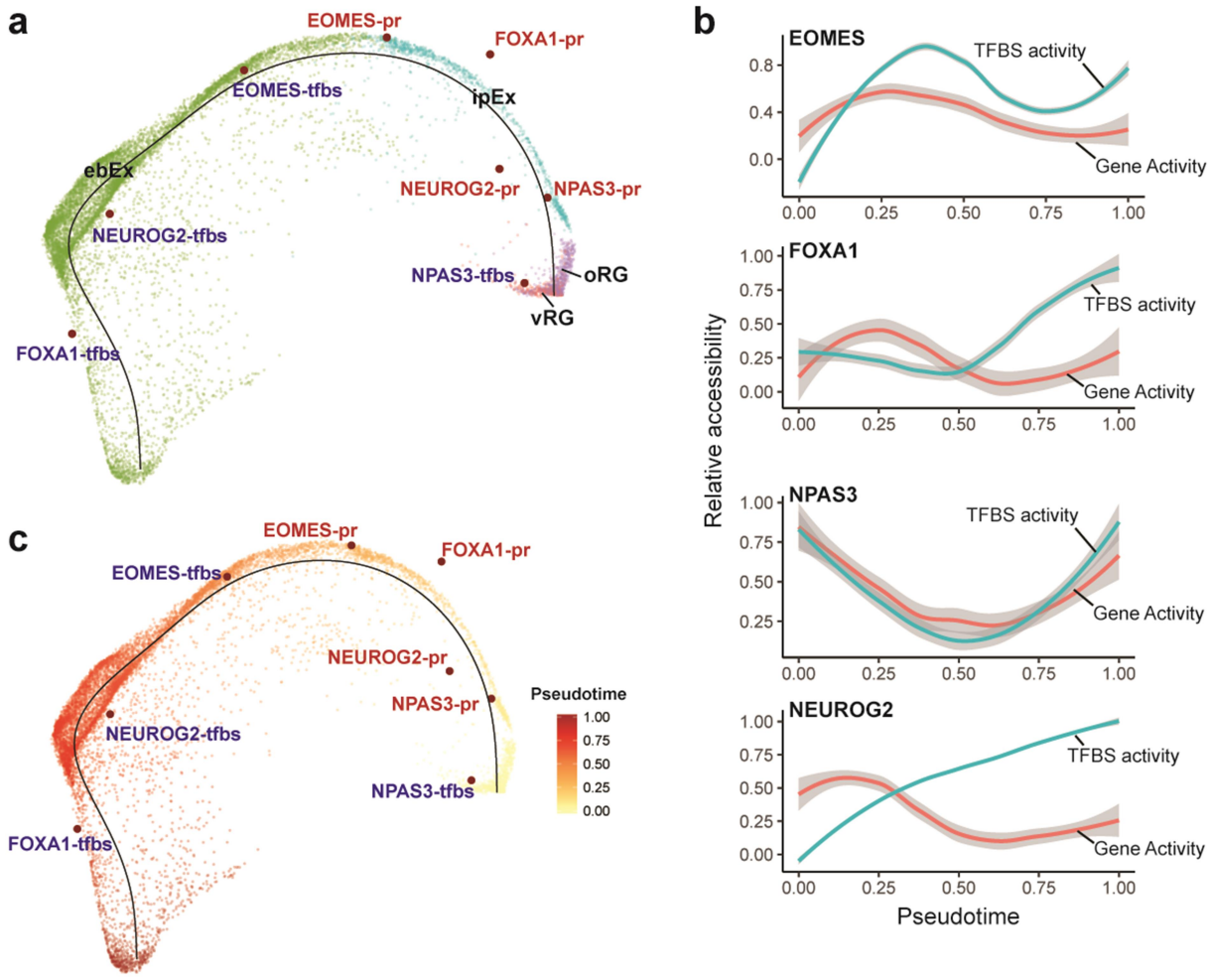


Figure 10: Trajectory Inference. **(a)** We used SWNE to embed oRG, vRG, ipEx, and ebEx cells, as well as the promoter and binding site accessibility of key TFs, and computed a developmental trajectory. The closer a cell is to a TF promoter or binding site, the more accessible that promoter/binding site is in that cell. **(b)** Plot of promoter accessibility vs binding site accessibility for EOMES, FOXA1, NPAS3, NEUROG2 across pseudotime. **(c)** SWNE embedding of oRG, vRG, ipEx, and ebEx developing human progenitor cell types colored by pseudotime

3.2.3 Understanding the Function of Human Accelerated Regions in Brain Development

Human Accelerated Regions (HARs) are regions of the human genome that are thought to play a role in human specific evolution **(Figure 11A)**. HARs are relatively small, with a median width of 234 base pairs (bp), and most HARs reside in noncoding regions of the genome **(Figure S8A, 11B)**. Thus, our goal was to use our developmental chromatin accessibility dataset to understand the cell types these HARs are most active in **(Figure 11A, 11B)**. We then used Cicero co-accessibility, which is an estimate of the likelihood that two genomic regions interact in some way (such as an enhancer region interacting with a promoter), to identify distal genes that HARs may be regulating, alongside identifying HARs in the promoter regions of key genes(Pliner et al. 2018) **(Figure 11A)**. We integrated our Cicero co-accessibility data with previous work mapping genes whose promoters have physical contact with HAR regions (via Hi-C) in the fetal human prefrontal cortex to generate a combined HAR regulatory network(Won et al. 2019, 2016) **(Figure 11A)**.

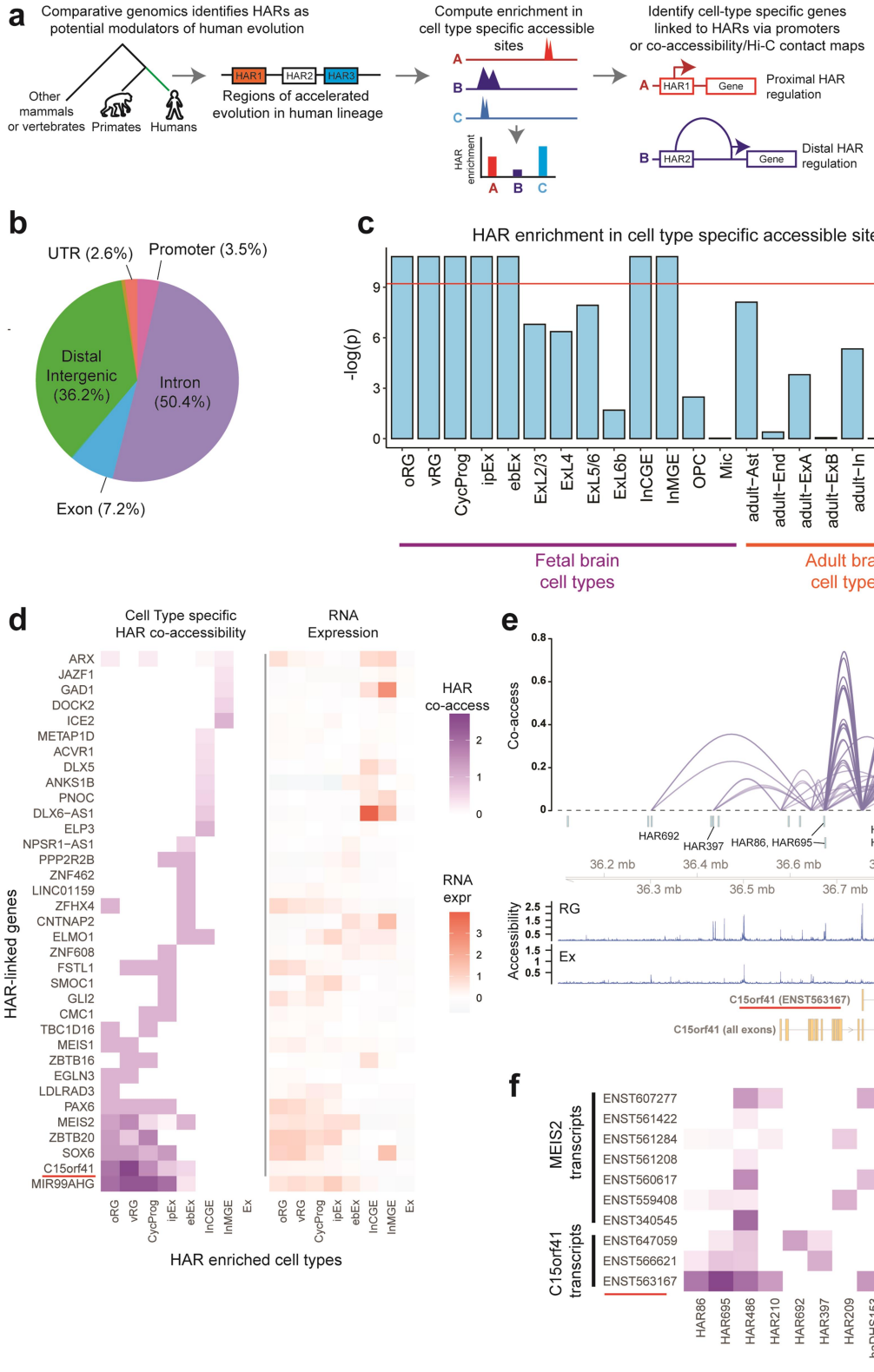
We compiled 3,168 HARs from five different studies and overlapped those HARs with both fetal and adult cell-type specific accessible regions, allowing for a 1kb gap to account for any uncertainty in HAR boundaries(Pollard et al. 2006; Prabhakar et al. 2006; Bird et al. 2007; Bush and Lahn 2008; Lindblad-Toh et al. 2011; Hawkins et al. 2015) **(Figure 11A, 11C, Methods)**. In order to compute p-value for the enrichment of HARs in the accessible regions of each cell type, we established a null distribution by sampling from the union of all accessible regions in the developing fetal and adult brains, and all ENCODE DNaseI hypersensitive sites(B. B. Lake et al. 2017; Neph et al. 2012) **(Figure 11C, Methods)**. We found that developmental progenitors, including developing inhibitory neurons, were highly enriched (p-value < 1e-4) for HARs, while layer specific excitatory neurons and adult cell types were not enriched **(Figure 11C)**. This suggests that HARs act primarily during human development, and not during adulthood. Additionally, HARs primarily act on progenitor cell

types such as radial glia (oRG, vRG), intermediate progenitors (ipEx), and early born excitatory neurons (ebEx) and not on more mature layer specific excitatory neurons in the fetal cortex.

While these HARs are primarily located in distal intergenic regions and introns, we found that the majority of HARs that intersect a cell-type specific accessible region are linked to at least one gene promoter either via co-accessibility or chromatin contact mapping **(Figure S8B)**. We identified genes that were linked with these cell-type specific HARs that also showed expression in the given cell-type and identified enrichment in Gene Ontology (GO) terms via a hypergeometric test(Liberzon et al. 2015) **(Figure S8C)**. This geneset enrichment analysis showed enrichment in GO terms related to brain development, such as “Glial Cell Differentiation” (P = 1.1E-3) and “GABA-ergic Neuron Differentiation”, and “Pallium Differentiation”, as well as more broad terms such as “Single Organism Behavior” **(Figure S8C)**. Interestingly the HAR-linked genes specific to fetal inhibitory neuron cell-types showed the strong geneset enrichment, suggesting a role for inhibitory neurons in human specific development **(Figure S8C)**. We then wanted to look specifically at the cell-type specific HAR-linked genes, and visualized co-accessibility and expression of the top 8 genes (as ranked by co-accessibility with cell-type specific HARs) for each HAR-enriched cell type, with all layer specific excitatory neurons (Ex) acting as a negative control **(Figure 11D)**. We found a number of genes known to regulate neurogenesis and brain development, such as ARX, and PAX6, and MEIS2(Kitamura et al. 2002; Machon et al. 2015; Zhang et al. 2010) **(Figure 11D)**. However, the gene with the highest HAR co-accessibility was C15orf41, a protein coding gene of unknown function that had been previously implicated as linked to anemia but to our knowledge has no known role in human brain development(Babbs et al. 2013; Russo et al. 2019) **(Figure 11D)**.

We generated a co-accessibility plot for the C15orf41 locus, where each arc represents the co-accessibility between an HAR and a gene promoter **(Figure 11E)**. We also plotted the accessibility in Radial Glia (RG) cell types as well as all in layer specific excitatory neurons (Ex) **(Figure 11E)**. We can see that the promoter region for the C15orf41 transcript that contains only the last 3 exons of the gene is a major HAR co-accessibility hub, and is also highly accessible in Radial Glia, suggesting that this specific short splice variant of C15orf41 (ENST0000563167) is driving the co-accessibility **(Figure 11E)**. We quantified the co-accessibility of all C15orf41 and MEIS2 transcripts with all HARs in the region (that were co-accessible with at least one promoter region), demonstrating that the C15orf41-short splice variant has by far the greatest total HAR co-accessibility **(Figure 11F)**.

Figure 11: Understanding the cell-type specific role of HARs. **(a)** Human Accelerated Regions (HARs) are regions of the human genome that have higher than expected mutation rates in the human specific branch of the evolutionary tree, suggesting a key role in human evolution. We identify cell types where HARs are more accessible than expected and link HARs with gene promoters using co-accessibility to identify cell type specific genes regulated by HARs. **(b)** A piechart of the genomic annotation of HARs **(c)** A log p-value barplot showing enrichment of HAR accessibility in human developmental progenitor cell types **(d)** A heatmap showing expression and HAR co-accessibility of the top 8 HAR-regulated genes for each HAR-enriched cell type, with all mature excitatory neurons (Ex) as a control. **(e)** A Cicero co-accessibility of the C15orf41 locus (the gene with the highest HAR co-accessibility). Each arc represents a connection between an HAR and a gene promoter. Each tick represents the location of an HAR and the accessibility of all Radial Glia (RG) and excitatory neurons (Ex) is displayed. **(f)** Heatmap quantifying the co-accessibility between each C15orf41 and MEIS2 transcript and each HAR in the locus


a
Comparative genomics identifies HARs as potential modulators of human evolution
Other mammals or vertebrates
Primates
Humans
HAR1
HAR2
HAR3
Regions of accelerated evolution in human lineage
Compute enrichment in cell type specific accessible sites
A
B
C
HAR enrichment
Identify cell-type specific genes linked to HARs via promoters or co-accessibility/Hi-C contact maps
Gene
Proximal HAR regulation
Distal HAR regulation
b
UTR (2.6%)
Promoter (3.5%)
Distal Intergenic (36.2%)
Intron (50.4%)
Exon (7.2%)
c
HAR enrichment in cell type specific accessible sites
-log(p)
oRG
vRG
CycProg
ipEx
ebEx
ExL2/3
ExL4
ExL5/6
ExL6b
InCGE
InMGE
OPC
Mic
adult-Ast
adult-End
adult-ExA
adult-ExB
adult-In
adult-Mic
adult-Oli
adult-Opc
Fetal brain cell types
Adult brain cell types
d
Cell Type specific HAR co-accessibility
RNA Expression
HAR co-access
RNA expr
HAR-linked genes
ARX
JAZF1
GAD1
DOCK2
ICE2
METAP1D
ACVR1
DLX5
ANKS1B
PNOC
DLX6-AS1
ELP3
NPSR1-AS1
PPP2R2B
ZNF462
LINC01159
ZFHX4
CNTNAP2
ELMO1
ZNF608
FSTL1
SMOC1
GLI2
CMC1
TBC1D16
MEIS1
ZBTB16
EGLN3
LDLRAD3
PAX6
MEIS2
ZBTB20
SOX6
C15orf41
MIR99AHG
HAR enriched cell types
e
Co-access
HAR692
HAR397
HAR86, HAR695
HAR486
HAR209
haDHS153
HAR210
Accessibility
RG
Ex
C15orf41 (ENST563167)
C15orf41 (all exons)
MEIS2
f
MEIS2 transcripts
C15orf41 transcripts
ENST607277
ENST561422
ENST561284
ENST561208
ENST560617
ENST559408
ENST340545
ENST647059
ENST566621
ENST563167
co-access
HAR86
HAR695
HAR486
HAR210
HAR692
HAR397
HAR209
haDHS153
Total HAR co-access

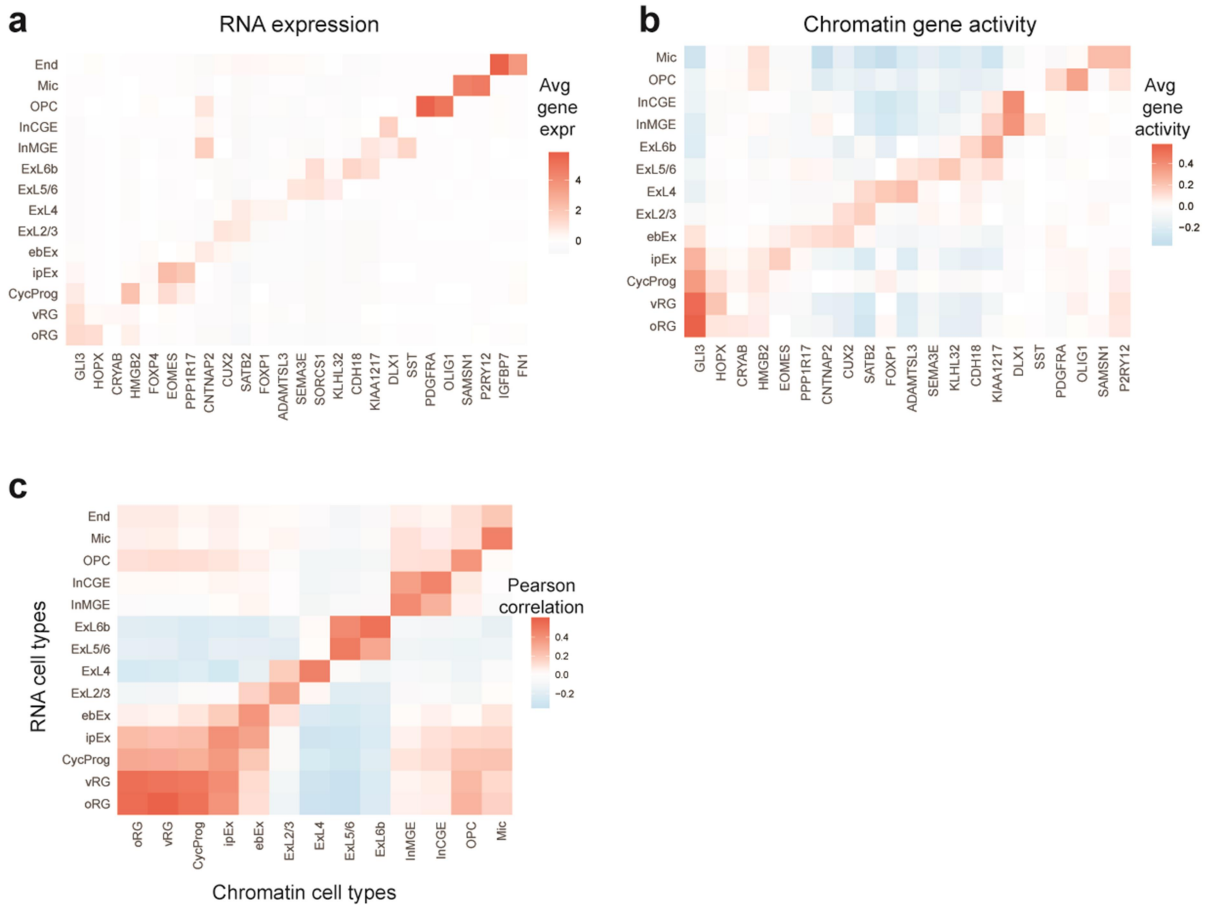
3.3 Discussion and Future Work

In summary, we generated 40,678 single-cell chromatin accessibility profiles and 7,865 single-cell gene expression profiles from the week 16 and week 18 fetal human cortex using scTHS-seq. We used SWNE to model the developmental trajectory of cortical excitatory neuron progenitors, revealing some transcription factor binding dynamics in the process. We then computed cell-type specific accessible sites and used them to identify developmental cell types that are enriched in HAR activity. For each HAR-enriched cell type, we used Cicero co-accessibility to identify genes regulated by the cell-type specific HARs, and found that a specific splice variant of the uncharacterized protein coding gene C15orf41 had the highest co-accessibility.

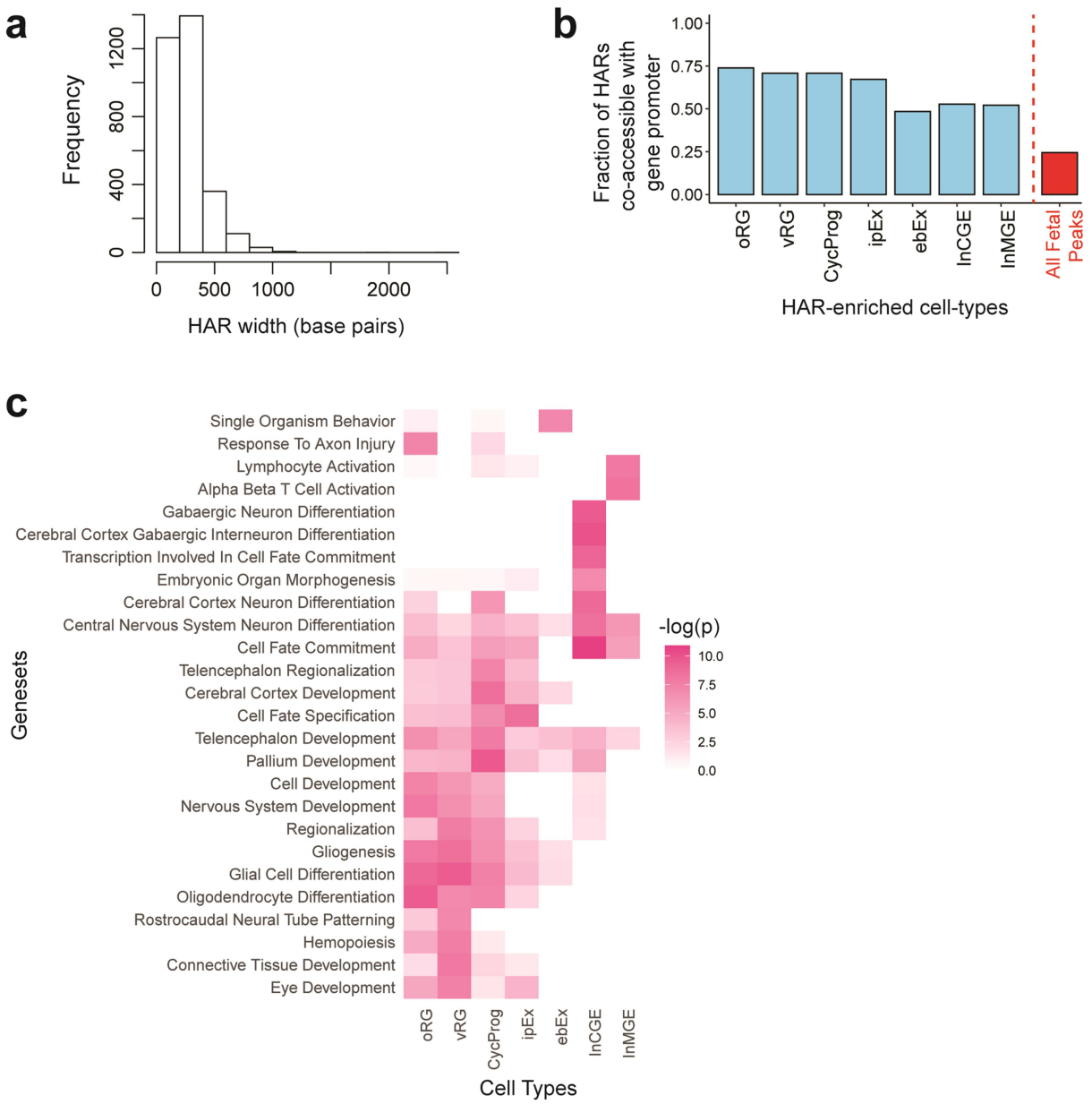
Our ability to profile chromatin accessibility at the single-cell level was critical to understanding the cell-type specific activity of HARs, and computing Cicero co-accessibility was extremely helpful in identifying HAR-linked genes. Additionally, coupling chromatin accessibility with single-cell RNA-seq enables us to validate the expression of HAR-linked genes in the same developmental cell types. This combined accessibility and gene expression map of human cortical development led us to identifying a splice variant of C15orf41, a protein coding gene of unknown function, as potentially involved in human brain development.

We are planning on validating the role of C15orf41 in human brain development by generating CRISPR knockouts of the short C15orf41 splice variant alongside knockouts of the full length variant of C15orf41 in a PGP1 iPSC line with Cas9 knocked in, and generating human cortical organoids to assess the functional impact on brain development(Lancaster and Knoblich 2014). Additionally, we are planning on validating the activity of HARs co-accessible with the C15orf41 short splice variant using an enhancer reporter assay in human cortical organoids.

3.4 Supplemental Figures and Tables



Supplementary Figure 7. (a) Gene activity heatmap of cell type specific marker genes. **(b)** RNA expression heatmap of cell-type specific marker genes. **(c)** Correlation of average chromatin gene activities and average RNA gene expression for each cell type.



Supplementary Figure 8. (a) Histogram of HAR sizes in base pairs **(b)** Fraction of HARs co-accessible with at least one gene promoter. **(c)** Heatmap of the hypergeometric enrichment of cell-type specific HAR-linked genes in Gene Ontology genesets.

3.5 Materials and Methods

3.5.1 snDropSeq nuclei preparation

All human tissue protocols were approved by the Office for Human Research Protection at Sanford Burnham Prebys Medical Discovery Institute and conformed to National Institutes of Health guidelines. Nuclei were prepared with nuclear extraction buffer (NEB) as described previously(B. B. Lake et al. 2017). Briefly, fresh-frozen postmortem brain tissue was sectioned at 50 μm with a cryostat and placed in 1 ml of ice-cold NEB for 10 min. Nuclei were extracted with 10–12 up-and-down strokes of a glass Dounce homogenizer with a Teflon pestle in 1 ml of NEB. Samples were passed through a 50-μm filter (Sysmex Partec) and then incubated on ice for 10 min. Samples were spun for 5 min at 250–300g, washed in PBS + 2 mM EGTA, and resuspended in PBS + 2 mM EGTA supplemented with 1% fatty-acid-free BSA (Gemini) containing 4′,6-diamidino-2-phenylindole (DAPI). DAPI+ sin- gle nuclei were purified by flow cytometry with a MoFlo Astrios (Beckman Coulter) or FACSAria Fusion (Becton Dickinson), concentrated at 900g for 10 min, and then used directly for droplet encapsulation

3.5.2 snDropSeq Library Preparation and Sequencing

Drop-seq with modifications optimized for nuclei processing was performed as described previously(B. B. Lake et al. 2017). Before droplet generation, connecting tubing and syringes were coated with 1% BSA to prevent nonspecific binding of nuclei to the surface, and then rinsed with PBS. To reduce nuclei settling, Ficoll PM-400 was added to the nuclei suspension buffer, rather than the lysis buffer. Nuclei were loaded at a concentration of 100 nuclei/μl and coencapsulated in drop- lets with barcoded beads purchased from ChemGenes Corporation (cat. no. Macosko201110). When encapsulation was complete, the contents of the droplet-collecting Falcon tubes were overlaid with a layer of mineral oil and

then transferred to a 72 °C water bath for 5 minutes to lyse the nuclear membranes. We then proceeded to reverse-transcription (RT) and PCR amplification of cDNA as previously described. A total of 12 snDropSeq libraries were prepared, and cDNA from each replicate was tagmented by Nextera XT and indexed with different Nextera index 1 primers. cDNA libraries were pooled and sequenced on an Illumina HiSeq 2500 with Read1CustSeqB17 for priming of read 1 (read 1 was 30 bp; read 2 (paired end) was 120 bp).

3.5.3 snDropSeq Data Processing and Clustering

Paired-end sequencing reads were processed largely as described (http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf), with additional correction steps as previously described(B. B. Lake et al. 2017). Briefly, reads were filtered to ensure the presence of a polyT and to remove reads with low sequencing quality bases. The right mate of each read pair was trimmed to remove any portion of the SMART adapter sequence or polyT tails. The trimmed reads were then aligned to the human genome (GENCODE GRCH38) with STAR v2.5 with the default parameter settings. Reads that mapped to intronic or exonic regions of genes as per the GENCODE gene annotation were recorded. We applied one further correction step to fix barcode synthesis errors by inserting N at the last base of the cell barcode for reads in which the first 11 bases of the cell barcode were identical and the last T base of UMI was the same. The digital expression matrix was then generated with genes as rows and cells as columns. We assigned UMI counts for each gene of each cell by collapsing UMI reads that had only 1 edit distance.

UMI matrix cell barcodes were tagged by their associated sequencing library batch ID and combined across independent experiments. Mitochondrial genes not expressed in nuclei were excluded, and only UMI counts associated with protein-coding genes were used for

clustering analyses. Nuclei with fewer than 300 molecules or more than 5,000 molecules (outliers) were omitted. We normalized molecular counts by using the total number of UMIs as the estimated library size for each cell. Variance normalization and clustering were done with the PAGODA2 package (https://github.com/hms-dbmi/pagoda2) as described previously. Briefly, we selected 2000 overdispersed genes and computed the top 50 principal components (PCs). We generated clusters using PAGODA2 and then imported the gene expression matrix and the PCs into Seurat for UMAP visualization and marker gene visualization. We identified cell types using previously described marker genes for fetal human cortical cell types(Polioudakis et al. 2019) **(Figure S7A)**.

3.5.4 scTHS-seq Nuclei Isolation

We prepared nuclei for the single-cell THS-seq chromatin accessibility assay using the previously described protocol(B. B. Lake et al. 2017). After flow cytometry, nuclei were kept on ice and spun down at 500g for 5 min at 4 °C, after which supernatant was removed and the pellet was resuspended in lysis buffer. Then nuclei were spun down at 500g for 5 min at 4 °C, supernatant was removed, and the pellet was resuspended in tagmentation buffer. At that point the nuclei sample was ready for nuclei counting A nuclei concentration of ~2.4 million nuclei/mL was obtained for each sample.

3.5.5 scTHS-seq Transposome Generation

We prepared transposons for scTHS-seq as previously described(B. B. Lake et al. 2017). Briefly each transposon consisted of two oligos: the 74-bp barcoded transposon and the 19-bp universal 5′ phosphorylated mosaic end. In total, there were 384 barcoded r5 transposons, each with a unique 6-bp barcode. For the generation of annealed transposons,

10 µL of each 100 µM oligo was added to each well of a 384-well plate (final concentration: 50 µM), incubated at 95 °C for 2 min, cooled to 14 °C at 0.1 °C/s, diluted to 8.4 µM in TE buffer with a final concentration of 50% glycerol, and then stored at −20 °C.

Tn5059 was generated and normalized for activity at Illumina. Transposome complexes were generated freshly for each scTHS-seq run and used within a few days. First, Tn5059 was diluted to 4.2 µM in standard storage buffer (Illumina), and 1 µL was added to each well of 384-well plate. Next, 1 µL of 8.4 µM annealed barcoded r5 transposon was added to each well, and the 384-well plate was incubated at room temperature for 30 min. For custom nXTv2_i7 Tn5059 transposome generation, the annealed nXTv2_i7 transposon (50 µM) was generated and we incubated 7 µM Tn5059 with 10 µM annealed transposon for 30 min at room temperature and then diluted to 0.7 µM Tn5059 transposome complex with standard storage buffer (Illumina).

3.5.6 scTHS-Seq Tagmentation, Barcoding and Library Preparation

We ran the scTHS-seq assay as previously described. Briefly, we added 4 µL of cell sample to each well of the 384 well-plate with the loaded transposome complex for a total of ~960 nuclei per well and a final concentration of 0.7 µM Tn5059 r5 transposome complex. To stop the reaction, we added 4.0 µL of 50 mM EDTA to each well and mixed gently five times with the electronic pipettor, and then incubated the mixture at 37 °C for 15 min. We added one volume of cold 2X FACS buffer (2 mM EDTA, 1% BSA in PBS) to each well, and samples were mixed gently three times with the electronic pipettor and pooled into one tube on ice. We spun down the tube, and resuspended the cells in 1X FACS buffer. Next, 75 µL of propidium iodide (PI; eBioscience) was added, and nuclei were sorted by flow cytometry into 96-well plates containing 10 µL of PBS per well at 100 nuclei per well and kept on ice.

Doublets were removed on the basis of forward and side scatter plots, and PI-staining events were selected.

Each 96-well plate of nuclei was then processed individually as previously described. Briefly, 11 µL of guanidine hydrochloride was added to each well and mixed by light vortexing. Reactions were purified with AMPure SPRI beads. 10 µL of 1× NEB Taq polymerase was added to each reaction, and the plate was lightly vortexed to resuspend the beads (SPRI beads left in the reaction), after which the reactions were run at 72 °C for 3 min for end fill-in. For in vitro transcription (IVT) amplification, we used the NEB HiScribe T7 high-yield synthesis kit. For reverse transcription, we added 2.5 µL of 20 µM random hexamers to each reaction, and used the Clontech SMART MMLV reverse transcriptase kit. Reactions were incubated at 22 °C for 10 min, then 42 °C for 60 min, and terminated at 70 °C for 10 min. To degrade RNA in cDNA–RNA hybrids, we added 1 µL of 0.5 units Enzymatics RNase H to each reaction, vortexed the plate lightly, and incubated the plate at 37 °C for 20 min. For second-strand synthesis, we added the first 2.5 µL of 20 µM sss_scnXTv2 (to each reaction and lightly vortexed it, then incubated it for 2 min at 65 °C and immediately cooled it on ice. Then we added 5.9 µL of NEB taq5X to each reaction and incubated it at 72 °C for 8 min. Double-stranded cDNA frag- ments then underwent simultaneous fragmentation and 3′ adaptor addition with a custom nXTv2_i7 Tn5059 transposome. To 7-µL volumes of each sample, we added 2 µL of 5× tagmentation buffer, followed by 2 µL of prepared 0.7 µM custom nXTv2_i7 Tn5059 transposomes (final transposome concentration of 0.14 µM). We added 19 µL of 6.32 M guanidine hydrochloride, for a final guanidine hydrochloride concentration of 4 M, to each reaction and briefly vortexed the sample. We eluted sample off SPRI beads held by the magnetic plate and transferred it to a qPCR plate. Standard Illumina Nextera XT v2 barcoding in an 8 × 12 (i5 × i7) format was performed with qPCR, using custom scTHS-seq i5 indexes and standard Illumina i7 indexes.

For pooling, 2 µL (4 µL or 6 µL if yields were low) of each uniquely barcoded qPCR reaction was combined and size-selection was performed as described9. Resultant size-selected libraries were quantified with Qubit and sequenced on an Illumina MiSeq system (50 + 32 + 32 single-end reads) for validation, then on the high-throughput Illumina HiSeq 2500 (50 + 8 + 32 single-end reads) for data generation.

3.5.7 scTHS-Seq Data Processing

We generated Fastq files for Read1, Index1, and Index2 and identified the reads that map to each unique barcode combination using deindexer (https://github.com/ws6/deindexer) with a zero mismatch stringency. This resulted in a single fastq file per cell barcode combination. After deindexing, we appended the cell barcode combination for each read to the read name, and then re-merged all fastq files for alignment. We aligned the merged fastq file to an hg38 reference genome (GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set) using BWA.

We then used the snaptools snap-pre command (with –keep-single = TRUE) to generate a snap file, and the snap-add-bmat command to generate 5 kb bins across the entire genome and create a bins by cells matrix(Fang et al. 2019). We then used the SnapATAC processing pipeline to run dimensionality reduction and clustering on the bins by cells matrix, and then use MACS2 to call peaks on the reads from each cluster separately. We merged the peaks from all clusters to generate a consensus list of peaks, and then generated a binary peaks by cells matrix(Fang et al. 2019).

3.5.8 scTHS-seq Dimensionality Reduction, Clustering, and Cell Type Identification

We first filtered the peaks by cells matrix for cells with at least 500 accessible sites to remove potential empty barcode combinations, and less than 20,000 accessible sites to remove potential multiplets. After filtering cells, we used cisTopic to run Latent Dirichlet Allocation (LDA) on the peaks matrix with 30 topics, with the optimal number of topics selected cisTopic's model selection functionality(Bravo González-Blas et al. 2019). We ran UMAP on the LDA topics to generate a 2D visualization of the data and then imported the LDA topics into Seurat v3 and clustered the cells using default parameters(Stuart et al. 2019). To help with cell type annotation and downstream analysis, we generated a gene activity matrix using Cicero, with a cell bin size of 80 and a minimum coaccessibility cutoff of 0.1(Pliner et al. 2018). We then correlated the average gene activities of each scTHS-seq cluster with the average gene expression of each snDropSeq cluster to identify rough cell types for each scTHS-seq cluster. scTHS-seq clusters that mapped to the same snDropSeq cluster were merged. We then validated scTHS-seq cell types by using Seurat to find gene activity markers for each scTHS-seq cell type, merging any cell types that lacked distinct markers.

3.5.9 scTHS-seq SWNE and Trajectory Analysis

In order to model the differentiation of progenitor cell types from Radial Glia to early born excitatory neurons, we subsetted the peaks matrix to only include Radial Glia (oRG, vRG), intermediate progenitors (ipEx), and early born excitatory neurons (ebEx) and then reran cisTopic with 20 topics and Cicero (with the same parameters). We embedded the LDA topics and cells in 2D visualization using Similarity Weighted Nonnegative Embedding(Wu, Tamayo, and Zhang 2018). To embed gene promoter accessibility alongside the cells, we

first identified each peak that is within 3 kb of a transcription start site for that gene using ChIPSeeker and then took an average of the topic loadings across the promoter peaks for each topic(Yu, Wang, and He 2015). Thus for each gene promoter, its 2D embedding position was defined by a weighted average of the position of each topic, with the weights being normalized loadings. This is identical to how genes are embedded in RNA-seq SWNE embeddings. To embed transcription factor (TF) binding sites, we used ChromVar to compute a TF binding activity matrix, and then correlated each TF's binding activities with each LDA topic using Pearson Correlation(Schep et al. 2017). The position of each TF binding site was defined by a weighted average of the position of each topic, with the weights corresponding to the correlation with that topic.

To compute the developmental pseudotime for each cell, we fit a principal curve to the SWNE cell embeddings using princurve, using each cell's projected arc-length on the curve as the estimated pseudotime(Hastie and Stuetzle 1989). We plotted gene activity and TF accessibility (as estimated using ChromVar) over pseudotime by segmenting the cells into bins of 80 cells along the pseudotime trajectory, and then computing the average gene activity or TF accessibility within each bin. We generated a smoothed line plot for both gene activity and TF using the loess package in R, as implemented in the geom_smooth function in ggplot2.

3.5.10 HAR Cell-Type Enrichment

We annotated the genomic location of each Human Accelerated Regions (HARs) using ChIPSeeker, with promoters defined as being within 3 kb of a Transcription Start Site (TSS). We identified differentially accessible peaks for each fetal and adult cell type using a modified Fisher's exact test as previously described(B. B. Lake et al. 2017). To get a sense

of the effect size of each differentially accessible peak, we computed the log-fold change for each peak in each cell type, defined here as the log2 transform of the fraction of cells in a cell type that are accessible at that peak divided by the fraction of all other cells that are accessible at that peak. We also computed a specificity score as previously described to assess how specific the accessibility of a peak is to a cell type(Cusanovich et al. 2018). We defined a peak as differentially accessible for a cell type if it had an adjusted p-value of less than 0.05, a log fold-change of greater than 0.5, and a specificity score of greater than 0.0001.

To compute the enrichment of HARs in differentially accessible we first computed the number of HARs overlapping a differentially accessible peak for each cell type, allowing for a gap size of up to 1000 base pairs. We then permuted the differentially accessible peaks for all cell types by sampling 20,000 times from the union of all fetal and adult cortex accessible peaks as well as all DNaseI hypersensitive sites from ENCODE, a total of 2,332,378 peaks. We computed empirical p-values by taking the number of samples where the number of overlapping HARs was greater than the true number of overlapping HARs and divided it by 20,000. We defined a cell-type as enriched for HARs if it had an empirical p-value of less than 1E-4.

3.5.11 HAR-gene Co-accessibility Analysis

We identified HARs overlapping differentially accessible peaks for all HAR-enriched cell types, and used Cicero co-accessibility to link each HAR to genes. A HAR can be linked to one or more genes if the HAR is the promoter region of that gene (within 3kb of a TSS) one or more peaks that the HAR overlaps is co-accessible with a peak in the promoter region of that gene, with a minimum co-accessibility of 0.1. We filtered for genes that were

expressed in that cell-type by only keeping HAR-linked genes that had an RNA expression log fold-change of greater than 0.25. After finding HAR-linked genes for each HAR-enriched cell type, and computed geneset enrichment using Fisher's exact test with Gene Ontology genesets [insert GO ref]. We visualized the negative log p-value of all genesets that were enriched for a cell type with a p-value of less than 2.5E-3.

We then visualized the co-accessibility and average RNA expression of the 8 HAR-linked genes that the highest HAR co-accessibility for each HAR-enriched cell-type. An HAR that overlapped a promoter region of a gene was given a co-accessibility score of 1 with that gene. We visualized the HAR co-accessibility of the C15orf41/MEIS2 locus using a Cicero co-accessibility plot, only showing the co-accessibilities between HARs and promoter regions. We generated accessibility tracks by making bigwigs for Radial Glia (RG) and layer specific excitatory neuron (Ex) by taking all reads belonging to cells in each cell type and writing them to a separate bam file and using the bamCoverage command from bedtools with parameters "-bs 50 --normalizeUsing CPM –skipNAs" [insert bedtools ref]. We also generated a custom track for HARs using the HAR bed file. Finally, we visualized the co-accessibility between each HAR in the locus and each transcript promoter.

3.6 Acknowledgement for Chapter 3

Chapter 3, in full, is a reprint of material being prepared for submission (Wu, Yan; Sos, Brandon; Chen, Song; Lake, Blue; Duong, Elizabeth; Dong, Weixiu; Limaye, Sid; Mali, Prashant; Zhang, Kun). The dissertation author was the primary author on this paper.

REFERENCES

Abdi, Hervé, and Lynne J. Williams. 2010. “Principal Component Analysis.” *Chemometrics and Intelligent Laboratory Systems* 2: 433–59. https://doi.org/10.1002/wics.101.

Adamson, Britt, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S. Weissman. 2016. “A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response.” *Cell* 167 (7): 1867-1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

Akcakaya, Pinar, Maggie L Bobbin, Jimmy A Guo, Jose M Lopez, M Kendell Clement, Sara P Garcia, Mick D Fellows, Michelle J Porritt, Mike A Firth, Alba Carreras, Tania Baccega, Frank Seeliger, Mikael Bjursell, Shengdar Q Tsai, Nhu T Nguyen, Roberto Nitsch, Lorenz M Mayr, Luca Pinello, Mohammad Bohlooly-Y, Martin J Aryee, Marcello Maresca, and J Keith Joung. 2018. “In Vivo CRISPR-Cas Gene Editing with No Detectable Genome-Wide off-Target Mutations.” *BioRxiv*, January, 272724. https://doi.org/10.1101/272724.

Amabile, Giovanni, Robert S Welner, Cesar Nombela-arrieta, Anna Morena D Alise, Annalisa Di Ruscio, Alexander K Ebralidze, Yevgenya Kraytsberg, Min Ye, Olivier Kocher, Donna S Neuberg, Konstantin Khrapko, Leslie E Silberstein, and Daniel G Tenen. 2019. “In Vivo Generation of Transplantable Human Hematopoietic Cells from Induced Pluripotent Stem Cells” 121 (8): 1–3. https://doi.org/110.1182/blood-2012-06-434407.There.

Angerer, Philipp, Laleh Haghverdi, Maren Büttner, Fabian J Theis, Carsten Marr, and Florian Buettner. 2015. “Destiny - Diffusion Maps for Large-Scale Single-Cell Data in R.” *Bioinformatics*, btv715-. https://doi.org/10.1093/bioinformatics/btv715.

Arthur, Wallace. 2002. “Evolutionary Developmental Biology.” *Nature*, 342–60. https://doi.org/10.1017/CCOL9780521851282.018.

Aurora, Megan, and Jason R Spence. 2016. “HPSC-Derived Lung and Intestinal Organoids as Models of Human Fetal Tissue.” *Developmental Biology* 420 (2): 230–38. https://doi.org/10.1016/j.ydbio.2016.06.006.

Avior, Yishai, Juan Carlos Biancotti, and Nissim Benvenisty. 2015. “TeratoScore: Assessing the Differentiation Potential of Human Pluripotent Stem Cells by Quantitative Expression Analysis of Teratomas.” *Stem Cell Reports* 4 (6): 967–74. https://doi.org/10.1016/j.stemcr.2015.05.006.

Aydin, Begüm, Akshay Kakumanu, Mary Rossillo, Mireia Moreno-Estellés, Görkem Garipler, Niels Ringstad, Nuria Flames, Shaun Mahony, and Esteban O. Mazzoni. 2019. “Proneural Factors Ascl1 and Neurog2 Contribute to Neuronal Subtype Identities by Establishing Distinct Chromatin Landscapes.” *Nature Neuroscience* 22 (6): 897–908. https://doi.org/10.1038/s41593-019-0399-y.

Babbs, Christian, Nigel A. Roberts, Luis Sanchez-Pulido, Simon J. McGowan, Momin R. Ahmed, Jill M. Brown, Mohamed A. Sabry, David R. Bentley, Gil A. McVean, Peter

Donnelly, Opher Gileadi, Chris P. Ponting, Douglas R. Higgs, and Veronica J. Buckle. 2013. “Homozygous Mutations in a Predicted Endonuclease Are a Novel Cause of Congenital Dyserythropoietic Anemia Type I.” *Haematologica* 98 (9): 1383–87. https://doi.org/10.3324/haematol.2013.089490.

Barkas, Nikolas, Brendan Joyce, Peter Kharchenko, Simon Steiger, Jean Fan, and Kamil Slowikowski. 2018. “Pagoda2: A Package for Analyzing and Interactively Exploring Large Single-Cell RNA-Seq Datasets.” https://github.com/hms-dbmi/pagoda2.

Barkas, Nikolas, Viktor Petukhov, Daria Nikolaeva, Yaroslav Lozinsky, Samuel Demharter, Konstantin Khodosevich, and Peter V. Kharchenko. 2019. “Joint Analysis of Heterogeneous Single-Cell RNA-Seq Dataset Collections.” *Nature Methods* 16 (8): 695–98. https://doi.org/10.1038/s41592-019-0466-z.

Bartfeld, Sina, Tülay Bayram, Marc van de Wetering, Meritxell Huch, Harry Begthel, Pekka Kujala, Robert Vries, Peter J Peters, and Hans Clevers. 2015. “In Vitro Expansion of Human Gastric Epithelial Stem Cells and Their Responses to Bacterial Infection.” *Gastroenterology* 148 (1): 126-136.e6. https://doi.org/https://doi.org/10.1053/j.gastro.2014.09.042.

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2018. “Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP.” *Nature Biotechnology* 37 (1). https://doi.org/10.1038/nbt.4314.

Bigorgne, Amélie E., Henner F. Farin, Roxane Lemoine, Nizar Mahlaoui, Nathalie Lambert, Marine Gil, Ansgar Schulz, Pierre Philippet, Patrick Schlesser, Tore G. Abrahamsen, Knut Oymar, E. Graham Davies, Christian Lycke Ellingsen, Emmanuelle Leteurtre, Brigitte Moreau-Massart, Dominique Berrebi, Christine Bole-Feysot, Patrick Nischke, Nicole Brousse, Alain Fischer, Hans Clevers, and Geneviève De Saint Basile. 2014. “TTC7A Mutations Disrupt Intestinal Epithelial Apicobasal Polarity.” *Journal of Clinical Investigation* 124 (1): 328–37. https://doi.org/10.1172/JCI71471.

Bird, Christine P., Barbara E. Stranger, Maureen Liu, Daryl J. Thomas, Catherine E. Ingle, Claude Beazley, Webb Miller, Matthew E. Hurles, and Emmanouil T. Dermitzakis. 2007. “Fast-Evolving Noncoding Sequences in the Human Genome.” *Genome Biology* 8 (6): 1–12. https://doi.org/10.1186/gb-2007-8-6-r118.

Black, Joshua B., Andrew F. Adler, Hong Gang Wang, Anthony M. D’Ippolito, Hunter A. Hutchinson, Timothy E. Reddy, Geoffrey S. Pitt, Kam W. Leong, and Charles A. Gersbach. 2016. “Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells.” *Cell Stem Cell* 19 (3): 406–14. https://doi.org/10.1016/j.stem.2016.07.001.

Bocker, W. 2002. “WHO classification of breast tumors and tumors of the female genital organs: pathology and genetics.” *Verhandlungen der Deutschen Gesellschaft fur Pathologie* 86: 116–19.

Boj, Sylvia F., Chang-Il Hwang, Lindsey A. Baker, Iok In Christine Chio, Dannielle D. Engle,

Vincenzo Corbo, Myrthe Jager, Mariano Ponz-Sarvise, Hervé Tiriac, Mona S. Spector, Ana Gracanin, Tobiloba Oni, Kenneth H. Yu, Ruben van Boxtel, Meritxell Huch, Keith D. Rivera, John P. Wilson, Michael E. Feigin, Daniel Öhlund, Abram Handly-Santana, Christine M. Ardito-Abraham, Michael Ludwig, Ela Elyada, Brinda Alagesan, Giulia Biffi, Georgi N. Yordanov, Bethany Delcuze, Brianna Creighton, Kevin Wright, Youngkyu Park, Folkert H.M. Morsink, I. Quintus Molenaar, Inne H. Borel Rinkes, Edwin Cuppen, Yuan Hao, Ying Jin, Isaac J. Nijman, Christine Iacobuzio-Donahue, Steven D. Leach, Darryl J. Pappin, Molly Hammell, David S. Klimstra, Olca Basturk, Ralph H. Hruban, George Johan Offerhaus, Robert G.J. Vries, Hans Clevers, and David A. Tuveson. 2015. “Organoid Models of Human and Mouse Ductal Pancreatic Cancer.” *Cell* 160 (1): 324–38. https://doi.org/https://doi.org/10.1016/j.cell.2014.12.021.

Bravo González-Blas, Carmen, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. 2019. “CisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-Seq Data.” *Nature Methods*. https://doi.org/10.1038/s41592-019-0367-1.

Brown, Juliana, Giorgia Quadrato, and Paola Arlotta. 2018. *Studying the Brain in a Dish : 3D Cell Culture Models of Human Brain Development and Disease*. *Human Embryonic Stem Cells in Development*. 1st ed. Vol. 129. Elsevier Inc. https://doi.org/10.1016/bs.ctdb.2018.03.002.

Buenrostro, Jason D, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. 2015. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” *Nature* 523 (7561): 486–90. https://doi.org/10.1038/nature14590.

Buettner, Florian, Naruemon Pratanwanich, Davis J. McCarthy, John C. Marioni, and Oliver Stegle. 2017. “F-ScLVM: Scalable and Versatile Factor Analysis for Single-Cell RNA-Seq.” *Genome Biology* 18 (1): 212. https://doi.org/10.1186/s13059-017-1334-8.

Bunge, R P. 1968. “Glial Cells and the Central Myelin Sheath.” *Physiological Reviews* 48 (1): 197 LP – 251. http://physrev.physiology.org/content/48/1/197.abstract.

Bush, Eliot C., and Bruce T. Lahn. 2008. “A Genome-Wide Screen for Noncoding Elements Important in Primate Evolution.” *BMC Evolutionary Biology* 8 (1): 1–10. https://doi.org/10.1186/1471-2148-8-17.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. “Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species.” *Nature Biotechnology*, no. February. https://doi.org/10.1038/nbt.4096.

Camp, J. Gray, Keisuke Sekine, Tobias Gerber, Henry Loeffler-Wirth, Hans Binder, Malgorzata Gac, Sabina Kanton, Jorge Kageyama, Georg Damm, Daniel Seehofer, Lenka Belicova, Marc Bickle, Rico Barsacchi, Ryo Okuda, Emi Yoshizawa, Masaki Kimura, Hiroaki Ayabe, Hideki Taniguchi, Takanori Takebe, and Barbara Treutlein. 2017. “Multilineage Communication Regulates Human Liver Bud Development from Pluripotency.” *Nature*. https://doi.org/10.1038/nature22796.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza

Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. 2017a. “Comprehensive Single Cell Transcriptional Profiling of a Multicellular Organism by Combinatorial Indexing.” *Science* 667 (August): 1–35. https://doi.org/http://dx.doi.org/10.1101/104844.

Cao, Junyue, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. 2017b. “Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism.” *Science (New York, N.Y.)* 357 (6352): 661–67. https://doi.org/10.1126/science.aam8940.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2019. “The Single-Cell Transcriptional Landscape of Mammalian Organogenesis.” *Nature*. https://doi.org/10.1038/s41586-019-0969-x.

Capowski, Elizabeth E, Kayvan Samimi, Steven J Mayerl, M Joseph Phillips, Isabel Pinilla, Sara E Howden, Jishnu Saha, Alex D Jansen, Kimberly L Edwards, Lindsey D Jager, Katherine Barlow, Rasa Valiauga, Zachary Erlichman, Anna Hagstrom, Divya Sinha, Valentin M Sluch, Xitiz Chamling, Donald J Zack, Melissa C Skala, and David M Gamm. 2019. “Reproducibility and Staging of 3D Human Retinal Organoids across Multiple Pluripotent Stem Cell Lines.” *Development* 146 (1): dev171686. https://doi.org/10.1242/dev.171686.

Capra, John a, Katherine S Pollard, Genevieve D Erwin, Gabriel Mckinsey, and John L R Rubenstein. 2013. “Many Human Accelerated Regions Are Developmental Enhancers.” *Philosophical Transactions of The Royal Society B*. https://doi.org/10.1098/rstb.2013.0025.

Castro, Diogo S., Ben Martynoga, Carlos Parras, Vidya Ramesh, Emilie Pacary, Caroline Johnston, Daniela Drechsel, Mélanie Lebel-Potter, Laura Galinanes Garcia, Charles Hunt, Dirk Dolle, Angela Bithell, Laurence Ettwiller, Noel Buckley, and Francxois Guillemot. 2011. “A Novel Function of the Proneural Factor Ascl1 in Progenitor Proliferation Identified by Genome-Wide Characterization of Its Targets.” *Genes and Development* 25 (9): 930–45. https://doi.org/10.1101/gad.627811.

Chambers, Stuart M., Jason Tchieu, and Lorenz Studer. 2013. “Build-a-Brain.” *Cell Stem Cell* 13 (4): 377–78. https://doi.org/10.1016/j.stem.2013.09.010.

Chan, Sunny Sun Kin, Robert W. Arpke, Antonio Filareto, Ning Xie, Matthew P. Pappas, Jacqueline S. Penaloza, Rita C.R. Perlingeiro, and Michael Kyba. 2018. “Skeletal Muscle Stem Cells from PSC-Derived Teratomas Have Functional Regenerative Capacity.” *Cell Stem Cell* 23 (1): 74-85.e6. https://doi.org/10.1016/j.stem.2018.06.010.

Chen, Meng, and Lei Stanley Qi. 2017. “Repurposing CRISPR System for Transcriptional Activation.” In *RNA Activation*, edited by Long-Cheng Li, 147–57. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-4310-9_10.

Chen, William S, Nevena Zivanovic, David Van Dijk, Guy Wolf, Bernd Bodenmiller, and Smita Krishnaswamy. 2020. “Uncovering Axes of Variation among Single-Cell Cancer

Specimens.” *Nature Methods*. https://doi.org/10.1038/s41592-019-0689-z.

Clevers, Hans. 2016. “Review Modeling Development and Disease with Organoids.” *Cell* 165 (7): 1586–97. https://doi.org/10.1016/j.cell.2016.05.082.

Collin, Joseph, Rachel Queen, Darin Zerti, Birthe Dorgau, Rafiqul Hussain, Jonathan Coxhead, Simon Cockell, and Majlinda Lako. 2019. “Deconstructing Retinal Organoids: Single Cell RNA-Seq Reveals the Cellular Components of Human Pluripotent Stem Cell-Derived Retina.” *Stem Cells* 37 (5): 593–98. https://doi.org/10.1002/stem.2963.

Combes, Alexander N., Luke Zappia, Pei Xuan Er, Alicia Oshlack, and Melissa H. Little. 2019. “Single-Cell Analysis Reveals Congruence between Kidney Organoids and Human Fetal Kidney.” *Genome Medicine*, 1–15. https://doi.org/10.1186/s13073-019-0615-0.

Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, Xingfan Huang, Lena Christiansen, William S. DeWitt, Choli Lee, Samuel G. Regalado, David F. Read, Frank J. Steemers, Christine M. Disteche, Cole Trapnell, and Jay Shendure. 2018. “A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.” *Cell*, 1–16. https://doi.org/10.1016/j.cell.2018.06.052.

Datlinger, Paul, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. 2017. “Pooled CRISPR Screening with Single-Cell Transcriptome Readout.” *Nature Methods* 14 (3): 297–301. https://doi.org/10.1038/nmeth.4177.

Dekkers, Johanna F., Caroline L. Wiegerinck, Hugo R. De Jonge, Inez Bronsveld, Hettie M. Janssens, Karin M. De Winter-De Groot, Arianne M. Brandsma, Nienke W.M. De Jong, Marcel J.C. Bijvelds, Bob J. Scholte, Edward E.S. Nieuwenhuis, Stieneke Van Den Brink, Hans Clevers, Cornelis K. Van Der Ent, Sabine Middendorp, and Jeffrey M. Beekman. 2013. “A Functional CFTR Assay Using Primary Cystic Fibrosis Intestinal Organoids.” *Nature Medicine* 19 (7): 939–45. https://doi.org/10.1038/nm.3201.

Denisenko, Elena, Belinda B. Guo, Matthew Jones, Rui Hou, Leanne de Kock, Timo Lassmann, Daniel Poppe, Olivier Clement, Rebecca K. Simmons, Ryan Lister, and Alistair R. R. Forrest. 2019. “Systematic Bias Assessment in Solid Tissue 10x ScRNA-Seq Workflows.” *BioRxiv*, 832444. https://doi.org/10.1101/832444.

DiCicco-Bloom, Emanuel, Catherine Lord, Lonnie Zwaigenbaum, Eric Courchesne, Stephen R. Dager, Christoph Schmitz, Robert T. Schultz, Jacqueline Crawley, and Larry J. Young. 2006. “The Developmental Neurobiology of Autism Spectrum Disorder.” *Journal of Neuroscience* 26 (26): 6897–6906. https://doi.org/10.1523/JNEUROSCI.1712-06.2006.

Dijk, David Van, Roshan Sharma, Juozas Nainys, Guy Wolf, Smita Krishnaswamy, Dana Pe, David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, and Ambrose J Carr. 2018. “Recovering Gene Interactions from Single-Cell Data Resource Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.” *Cell* 174 (3): 716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

Ding, Jiarui, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, John Y H Kwon, Boaz Barak, William Ge, Amanda J Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z Levin. 2019. “Systematic Comparative Analysis of Single Cell RNA-Sequencing Methods.” *BioRxiv*, January, 632216. https://doi.org/10.1101/632216.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” *Cell* 167 (7): 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

Dixit, Atray, Oren Parnas, Biyu Li, Jonathan S Weissman, Nir Friedman, Aviv Regev, Correspondence Aregev@broadinstitute Org, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, and Eric S Lander. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” *Cell* 167 (7): 1853-1857.e17. https://doi.org/10.1016/j.cell.2016.11.038.

Doan, Ryan N., Byoung-Il Bae, Beatriz Cubelos, Cindy Chang, Amer A. Hossain, Samira Al-Saad, Nahit M. Mukaddes, Ozgur Oner, Muna Al-Saffar, Soher Balkhy, Generoso G. Gascon, Marta Nieto, and Christopher A. Walsh. 2016. “Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior.” *Cell* 167 (2): 341-354.e12. https://doi.org/10.1016/j.cell.2016.08.071.

Dulken, Ben W., Dena S. Leeman, Stéphane C. Boutet, Katja Hebestreit, and Anne Brunet. 2017. “Single-Cell Transcriptomic Analysis Defines Heterogeneity and Transcriptional Dynamics in the Adult Neural Stem Cell Lineage.” *Cell Reports* 18 (3): 777–90. https://doi.org/10.1016/j.celrep.2016.12.060.

Dutta, Devanjali, Inha Heo, and Hans Clevers. 2017. “Disease Modeling in Stem Cell-Derived 3D Organoid Systems.” *Trends in Molecular Medicine* 23 (5): 393–410. https://doi.org/10.1016/j.molmed.2017.02.007.

Elsen, Gina E., Rebecca D. Hodge, Francesco Bedogni, Ray A.M. Daza, Branden R. Nelson, Naoko Shiba, Steven L. Reiner, and Robert F. Hevner. 2013. “The Protomap Is Propagated to Cortical Plate Neurons through an Eomes-Dependent Intermediate Map.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (10): 4081–86. https://doi.org/10.1073/pnas.1209076110.

Fan, Jean, Neeraj Salathia, Rui Liu, Gwendolyn E. Kaeser, Yun C. Yung, Joseph L. Herman, Fiona Kaper, Jian Bing Fan, Kun Zhang, Jerold Chun, and Peter V. Kharchenko. 2016. “Characterizing Transcriptional Heterogeneity through Pathway and Gene Set Overdispersion Analysis.” *Nature Methods* 13 (3): 241–44. https://doi.org/10.1038/nmeth.3734.

Fang, Rongxin, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, and Bing Ren. 2019. “Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types.” *BioRxiv*.

Farrell, Jeffrey A., Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. 2018. “Single-Cell Reconstruction of Developmental Trajectories during Zebrafish Embryogenesis.” *Science* 3131 (April): eaar3131. https://doi.org/10.1126/science.aar3131.

Fligor, Clarisse M, Kirstin B Langer, Akshayalakshmi Sridhar, Yuan Ren, Priya K Shields, Michael C Edler, Sarah K Ohlemacher, Valentin M Sluch, Donald J Zack, Chi Zhang, Daniel M Suter, and Jason S Meyer. 2018. “Three-Dimensional Retinal Organoids Facilitate the Investigation of Retinal Ganglion Cell Development, Organization and Neurite Outgrowth from Human Pluripotent Stem Cells.” *Scientific Reports* 8 (1): 14520. https://doi.org/10.1038/s41598-018-32871-8.

Franc, Vojtěch, Václav Hlaváč, and Mirko Navara. 2005. “Sequential Coordinate-Wise Algorithm for the Non-Negative Least Squares Problem.” *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3691 LNCS (i): 407–14. https://doi.org/10.1007/11556121_50.

Frigyesi, Attila, and Mattias Höglund. 2008. “Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes.” *Cancer Informatics* 6 (2003): 275–92.

Ganglberger, Florian, Joanna Kaczanowska, Josef M. Penninger, Andreas Hess, Katja Bühler, and Wulf Haubensak. 2018. “Predicting Functional Neuroanatomical Maps from Fusing Brain Networks with Genetic Information.” *NeuroImage* 170: 113–20. https://doi.org/10.1016/j.neuroimage.2017.08.070.

Gao, Dong, Ian Vela, Andrea Sboner, Phillip J. Iaquinta, Wouter R. Karthaus, Anuradha Gopalan, Catherine Dowling, Jackline N. Wanjala, Eva A. Undvall, Vivek K. Arora, John Wongvipat, Myriam Kossai, Sinan Ramazanoglu, Luendreo P. Barboza, Wei Di, Zhen Cao, Qi Fan Zhang, Inna Sirota, Leili Ran, Theresa Y. MacDonald, Himisha Beltran, Juan-Miguel Mosquera, Karim A. Touijer, Peter T. Scardino, Vincent P. Laudone, Kristen R. Curtis, Dana E. Rathkopf, Michael J. Morris, Daniel C. Danila, Susan F. Slovin, Stephen B. Solomon, James A. Eastham, Ping Chi, Brett Carver, Mark A. Rubin, Howard I. Scher, Hans Clevers, Charles L. Sawyers, and Yu Chen. 2014. “Organoid Cultures Derived from Patients with Advanced Prostate Cancer.” *Cell* 159 (1): 176–87. https://doi.org/https://doi.org/10.1016/j.cell.2014.08.016.

Gao, Nan, Peter White, and Klaus H Kaestner. 2009. “Establishment of Intestinal Identity and Epithelial-Mesenchymal Signaling by Cdx2.” *Developmental Cell* 16 (4): 588–99. https://doi.org/10.1016/j.devcel.2009.02.010.

Gao, Shuai, Liying Yan, Rui Wang, Jingyun Li, Jun Yong, Xin Zhou, Yuan Wei, Xinglong Wu, Xiaoye Wang, Xiaoying Fan, Jie Yan, Xu Zhi, Yun Gao, Hongshan Guo, Xiao Jin, Wendong Wang, Yunuo Mao, Fengchao Wang, Lu Wen, Wei Fu, Hao Ge, Jie Qiao, and Fuchou Tang. 2018. “Tracing the Temporal-Spatial Transcriptome Landscapes of the

Human Fetal Digestive Tract Using Single-Cell RNA-Sequencing." *Nature Cell Biology* 20 (6): 721–34. https://doi.org/10.1038/s41556-018-0105-4.

Garrido-Martín, Eva M., Francisco J. Blanco, Mercé Roquè, Laura Novensà, Mirko Tarocchi, Ursula E. Lang, Toru Suzuki, Scott L. Friedman, Luisa M. Botella, and Carmelo Bernabéu. 2012. "Vascular Injury Triggers Krüppel-like Factor 6 Mobilization and Cooperation with Specificity Protein 1 to Promote Endothelial Activation through Upregulation of the Activin Receptor-like Kinase 1 Gene." *Circulation Research* 112 (1): 113–27. https://doi.org/10.1161/circresaha.112.275586.

Griffiths, Jonathan A, Antonio Scialdone, and John C Marioni. 2018. "Using Single□cell Genomics to Understand Developmental Processes and Cell Fate Decisions." *Molecular Systems Biology* 14 (4): 1–12. https://doi.org/10.15252/msb.20178046.

Guo, Chuner, Brent A Biddy, Kenji Kamimoto, Wenjun Kong, and Samantha A Morris. 2018. "CellTag Indexing: A Genetic Barcode-Based Multiplexing Tool for Single-Cell Technologies." *BioRxiv*, January, 335547. https://doi.org/10.1101/335547.

Haghverdi, Laleh, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. 2018. "Batch Effects in Single-Cell RNA-Sequencing Data Are Corrected by Matching Mutual Nearest Neighbors." *Nature Biotechnology* 36 (5): 421–27. https://doi.org/10.1038/nbt.4091.

Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Zimin Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, Yibin Wang, Rui Yue, Tiefeng Li, He Huang, Stuart H. Orkin, Guo-Cheng Yuan, Ming Chen, and Guoji Guo. 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.

Hastie, Trevor, and Werner Stuetzle. 1989. "Principal Curves." *Journal of the American Statistical Association* 84 (406): 502–16. https://doi.org/10.1080/01621459.1989.10478797.

Hawkins, R. David, Joshua M. Akey, Rachel M. Gittelman, William S. Noble, Ferhat Ay, Len Pennacchio, Jennifer Madeoy, and Enna Hun. 2015. "Comprehensive Identification and Analysis of Human Accelerated Regulatory DNA." *Genome Research* 25 (9): 1245–55. https://doi.org/10.1101/gr.192591.115.

Hodge, Rebecca D, Trygve E Bakken, Jeremy A Miller, Kimberly A Smith, Eliza R Barkan, Lucas T Graybuck, Jennie L Close, Brian Long, Osnat Penn, Zizhen Yao, Jeroen Eggermont, Thomas Hollt, Boaz P Levi, Soraya I Shehata, Brian Aevermann, Allison Beller, Darren Bertagnolli, Krissy Brouner, Tamara Casper, Charles Cobbs, Rachel Dalley, Nick Dee, Song-Lin Ding, Richard G Ellenbogen, Olivia Fong, Emma Garren, Jeff Goldy, Ryder P Gwinn, Daniel Hirschstein, C Dirk Keene, Mohamed Keshk, Andrew L Ko, Kanan Lathia, Ahmed Mahfouz, Zoe Maltzer, Medea McGraw, Thuc Nghi Nguyen, Julie Nyhus, Jeffrey G Ojemann, Aaron Oldre, Sheana Parry, Shannon Reynolds, Christine Rimorin, Nadiya V Shapovalova, Saroja Somasundaram, Aaron Szafer, Elliot R Thomsen, Michael Tieu, Richard H Scheuermann, Rafael Yuste, Susan M Sunkin, Boudewijn Lelieveldt, David Feng, Lydia Ng, Amy Bernard, Michael Hawrylycz, John Phillips, Bosiljka Tasic, Hongkui Zeng, Allan R Jones, Christof Koch, and Ed S Lein.

2019. “Conserved Cell Types with Divergent Features between Human and Mouse Cortex.” *Nature*, 384826. https://doi.org/10.1101/384826.

Houle, Michael E., Hans Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. 2010a. “Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?” *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6187 LNCS: 482–500. https://doi.org/10.1007/978-3-642-13818-8_34.

Houle, Michael E, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. 2010b. “Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?” In *Scientific and Statistical Database Management*, edited by Michael Gertz and Bertram Ludäscher, 482–500. Berlin, Heidelberg: Springer Berlin Heidelberg.

Huch, M., and B.-K. Koo. 2015. “Modeling Mouse and Human Development Using Organoid Cultures.” *Development* 142 (18): 3113–25. https://doi.org/10.1242/dev.118570.

Huch, Meritxell, Juergen A Knoblich, Matthias P Lutolf, and Alfonso Martinez-arias. 2017. “The Hope and the Hype of Organoid Research,” 938–41. https://doi.org/10.1242/dev.150201.

Huisman, Sjoerd M.H., Baldur Van Lew, Ahmed Mahfouz, Nicola Pezzotti, Thomas Hollt, Lieke Michielsen, Anna Vilanova, Marcel J.T. Reinders, and Boudewijn P.F. Lelieveldt. 2017. “BrainScope: Interactive Visual Exploration of the Spatial and Temporal Human Brain Transcriptome.” *Nucleic Acids Research* 45 (10). https://doi.org/10.1093/nar/gkx046.

Jabaudon, Denis, and Madeline Lancaster. 2018. “Exploring Landscapes of Brain Morphogenesis with Organoids,” 2016–19. https://doi.org/10.1242/dev.172049.

Jacobson, Elena F., and Emmanuel S. Tzanakakis. 2017. “Human Pluripotent Stem Cell Differentiation to Functional Pancreatic Cells for Diabetes Therapies: Innovations, Challenges and Future Directions.” *Journal of Biological Engineering* 11 (1): 21. https://doi.org/10.1186/s13036-017-0066-3.

Jang, Sumin, Sandeep Choubey, Leon Furchtgott, Ling Nan Zou, Adele Doyle, Vilas Menon, Ethan B. Loew, Anne Rachel Krostag, Refugio A. Martinez, Linda Madisen, Boaz P. Levi, and Sharad Ramanathan. 2017. “Dynamics of Embryonic Stem Cell Differentiation Inferred from Single-Cell Transcriptomics Show a Series of Transitions through Discrete Cell States.” *ELife* 6: 1–28. https://doi.org/10.7554/eLife.20487.

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. 2013. “DNA-Binding Specificities of Human Transcription Factors.” *Cell* 152 (1–2): 327–39. https://doi.org/10.1016/j.cell.2012.12.009.

Jung, Peter, Toshiro Sato, Anna Merlos-Suárez, Francisco M. Barriga, Mar Iglesias, David Rossell, Herbert Auer, Mercedes Gallardo, Maria A. Blasco, Elena Sancho, Hans

Clevers, and Eduard Batlle. 2011. “Isolation and in Vitro Expansion of Human Colonic Stem Cells.” *Nature Medicine* 17 (10): 1225–27. https://doi.org/10.1038/nm.2470.

Kalluri, Raghu, and Robert A Weinberg. 2009. “The Basics of Epithelial-Mesenchymal Transition” 119 (6). https://doi.org/10.1172/JCI39104.1420.

Kamm, Gretel B., Francisco Pisciottano, Rafi Kliger, and Lucía F. Franchini. 2013. “The Developmental Brain Gene NPAS3 Contains the Largest Number of Accelerated Regulatory Sequences in the Human Genome.” *Molecular Biology and Evolution* 30 (5): 1088–1102. https://doi.org/10.1093/molbev/mst023.

Kanton, Sabina, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, Zhengzong Qian, Michael Heide, Wieland B Huttner, Philipp Khaitovich, Svante Pääbo, Barbara Treutlein, and J Gray Camp. 2019. “Organoid Single-Cell Genomic Atlas Uncovers Human-Specific Features of Brain Development.” *Nature*.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden. 2014. “Bayesian Approach to Single-Cell Differential Expression Analysis.” *Nature Methods* 11 (7): 740–42. https://doi.org/10.1038/nmeth.2967.

Kim, Jong Wook, Omar O. Abudayyeh, Huwate Yeerna, Chen Hsiang Yeang, Michelle Stewart, Russell W. Jenkins, Shunsuke Kitajima, David J. Konieczkowski, Kate Medetgul-Ernar, Taylor Cavazos, Clarence Mah, Stephanie Ting, Eliezer M. Van Allen, Ofir Cohen, John Mcdermott, Emily Damato, Andrew J. Aguirre, Jonathan Liang, Arthur Liberzon, Gabriella Alexe, John Doench, Mahmoud Ghandi, Francisca Vazquez, Barbara A. Weir, Aviad Tsherniak, Aravind Subramanian, Karina Meneses-Cime, Jason Park, Paul Clemons, Levi A. Garraway, David Thomas, Jesse S. Boehm, David A. Barbie, William C. Hahn, Jill P. Mesirov, and Pablo Tamayo. 2017. “Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States.” *Cell Systems* 5 (2): 105-118.e9. https://doi.org/10.1016/j.cels.2017.08.002.

Kim, Jong Wook, Olga B Botvinnik, Omar Abudayyeh, Chet Birger, Joseph Rosenbluh, Yashaswi Shrestha, Mohamed E Abazeed, Peter S Hammerman, Daniel DiCara, David J Konieczkowski, Cory M Johannessen, Arthur Liberzon, Amir Reza Alizad-Rahvar, Gabriela Alexe, Andrew Aguirre, Mahmoud Ghandi, Heidi Greulich, Francisca Vazquez, Barbara A Weir, Eliezer M Van Allen, Aviad Tsherniak, Diane D Shao, Travis I Zack, Michael Noble, Gad Getz, Rameen Beroukhim, Levi A Garraway, Masoud Ardakani, Chiara Romualdi, Gabriele Sales, David A Barbie, Jesse S Boehm, William C Hahn, Jill P Mesirov, and Pablo Tamayo. 2016. “Characterizing Genomic Alterations in Cancer by Complementary Functional Associations.” *Nature Biotechnology* 34 (5): 3–5. https://doi.org/10.1038/nbt.3527.

Kim, Tae-hee, and Ramesh A Shivdasani. 2016. “Stomach Development , Stem Cells and Disease,” 554–65. https://doi.org/10.1242/dev.124891.

Kitamura, Kunio, Masako Yanazawa, Noriyuki Sugiyama, Hirohito Miura, Akiko Iizuka-Kogo, Masatomo Kusaka, Kayo Omichi, Rika Suzuki, Yuko Kato-Fukui, Kyoko Kamiirisa, Mina Matsuo, Shin Ichi Kamijo, Megumi Kasahara, Hidefumi Yoshioka, Tsutomu Ogata, Takayuki Fukuda, Ikuko Kondo, Mitsuhiro Kato, William B. Dobyns, Minesuke

Yokoyama, and Ken Ichirou Morohashi. 2002. “Mutation of ARX Causes Abnormal Development of Forebrain and Testes in Mice and X-Linked Lissencephaly with Abnormal Genitalia in Humans.” *Nature Genetics* 32 (3): 359–69. https://doi.org/10.1038/ng1009.

Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 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.

Klein, Allon M, Linas Mazutis, David A Weitz, Marc W Kirschner, Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, and Leonid Peshkin. 2015. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells Resource 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.

Klimovskaia, Anna, David Lopez-Paz, Léon Bottou, and Maximilian Nickel. 2019. “Poincare Maps for Analyzing Complex Hierarchies in Single-Cell Data.” *BioRxiv*, 689547. https://doi.org/10.1101/689547.

Kobak, Dmitry, and Philipp Berens. 2019. “The Art of Using T-SNE for Single-Cell Transcriptomics.” *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-13056-x.

Kruskal, J. B. 1964. “Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis.” *Psychometrika* 29 (1): 1–27. https://doi.org/10.1007/BF02289565.

la Torre-Ubieta, Luis de, Jason L. Stein, Hyejung Won, Carli K. Opland, Dan Liang, Daning Lu, and Daniel H. Geschwind. 2018. “The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis.” *Cell* 172 (1–2): 289-304.e18. https://doi.org/10.1016/j.cell.2017.12.014.

Lake, Blue, Rizi Ai, Gwen E. Kaeser, Neeraj Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Jian-Bing Fan Ronaghi, Wei Wang, Jerold Chun, and Kun Zhang. 2016. “Neuronal Subtypes and Diversity Revealed by Single-Nucleus RNA Sequencing of the Human Brain.” *Science* 357 (2013): 352–57.

Lake, Blue B., Song Chen, Brandon C. Sos, Jean. Fan, Yun Yung, Gwendolyn E. Kaeser, Thu E. Duong, Derek Gao, Jerold Chun, Peter Kharchenko, and Kun Zhang. 2017. “Integrative Single-Cell Analysis by Transcriptional and Epigenetic States in Human Adult Brain.” *Nature Biotechnology*, no. December: 1–3. https://doi.org/10.1101/128520.

Lancaster, Madeline, and Juergen A. Knoblich. 2014. “Generation of Cerebral Organoids from Human Pluripotent Stem Cells.” *Nature Protocols* 148 (5): 29–30. https://doi.org/10.1038/nprot.2014.158.

Lee, Daniel D., and H. Sebastian Seung. 1999. “Learning the Parts of Objects by Non-Negative Matrix Factorization.” *Nature* 401 (6755): 788–91.

https://doi.org/10.1038/44565.

Lensch, M. William, Thorsten M. Schlaeger, Leonard I. Zon, and George Q. Daley. 2007. "Teratoma Formation Assays with Human Embryonic Stem Cells: A Rationale for One Type of Human-Animal Chimera." *Cell Stem Cell* 1 (3): 253–58. https://doi.org/10.1016/j.stem.2007.07.019.

Levchenko, Anastasia, Alexander Kanapin, Anastasia Samsonova, and Raul R. Gainetdinov. 2018. "Human Accelerated Regions and Other Human-Specific Sequence Variations in the Context of Evolution and Their Relevance for Brain Development." *Genome Biology and Evolution* 10 (1): 166–88. https://doi.org/10.1093/gbe/evx240.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." *Cell Systems* 1 (6): 417–25. https://doi.org/10.1016/j.cels.2015.12.004.

Lin, Xihui, and Paul C Boutros. 2016. "NNLM: Fast and Versatile Non-Negative Matrix Factorization." https://cran.r-project.org/package=NNLM.

Lin, YuShuang, DongYan Chen, QiuSheng Fan, and HongWei Zhang. 2009. "Characterization of SoxB2 and SoxC Genes in Amphioxus (Branchiostoma Belcheri): Implications for Their Evolutionary Conservation." *Science in China. Series C, Life Sciences* 52 (9): 813–22. https://doi.org/10.1007/s11427-009-0111-7.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke, Brian J. Raney, Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, Manolis Kellis, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Andrew Cree, Huyen H. Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R. Lewis, Lynne V. Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, and Erica Sodergren. 2011. "A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals." *Nature* 478 (7370): 476–82. https://doi.org/10.1038/nature10530.

Liu, Chun, Angelos Oikonomopoulos, Nazish Sayed, and Joseph C Wu. 2018. "Modeling Human Diseases with Induced Pluripotent Stem Cells : From 2D to 3D and Beyond," 1–6. https://doi.org/10.1242/dev.156166.

Lopez, Romain, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. "Deep Generative Modeling for Single-Cell Transcriptomics." *Nature Methods* 15 (12):

1053–58. https://doi.org/10.1038/s41592-018-0229-2.

Maaten, Laurens van der. 2014. “Accelerating T-Sne Using Tree-Based Algorithms.” *The Journal of Machine Learning Research* 15 (1): 3221–45. https://doi.org/10.1007/s10479-011-0841-3.

Maaten, Laurens Van Der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research* 9: 2579–2605. https://doi.org/10.1007/s10479-011-0841-3.

Machon, Ondrej, Jan Masek, Olga Machonova, Stefan Krauss, and Zbynek Kozmik. 2015. “Meis2 Is Essential for Cranial and Cardiac Neural Crest Development.” *BMC Developmental Biology* 15 (1): 1–16. https://doi.org/10.1186/s12861-015-0093-6.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Mao, Weiguang, Maziyar Baran Pouyan, Dennis Kostka, and Maria Chikina. 2020. “Non-Negative Independent Factor Analysis for Single Cell RNA-Seq.” *BioRxiv*, 2020.01.31.927921. https://doi.org/10.1101/2020.01.31.927921.

Marioni, John C., and Detlev Arendt. 2017. “How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology.” *Annual Review of Cell and Developmental Biology* 33 (1): 537–53. https://doi.org/10.1146/annurev-cellbio-100616-060818.

Marmigère, Frédéric, Andreas Montelius, Michael Wegner, Yoram Groner, Louis F. Reichardt, and Patrik Ernfors. 2006. “The Runx1/AML1 Transcription Factor Selectively Regulates Development and Survival of TrkA Nociceptive Sensory Neurons.” *Nature Neuroscience* 9 (2): 180–87. https://doi.org/10.1038/nn1631.

Mayran, Alexandre, and Jacques Drouin. 2018. “Pioneer Transcription Factors Shape the Epigenetic Landscape.” *Journal of Biological Chemistry* 293 (36): 13795–804. https://doi.org/10.1074/jbc.R117.001232.

McInnes, Leland, and John Healy. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *ArXiv*, 1–18. http://arxiv.org/abs/1802.03426.

Miller, Jeremy A., Song Lin Ding, Susan M. Sunkin, Kimberly A. Smith, Lydia Ng, Aaron Szafer, Amanda Ebbert, Zackery L. Riley, Joshua J. Royall, Kaylynn Aiona, James M. Arnold, Crissa Bennet, Darren Bertagnolli, Krissy Brouner, Stephanie Butler, Shiella Caldejon, Anita Carey, Christine Cuhaciyan, Rachel A. Dalley, Nick Dee, Tim A. Dolbeare, Benjamin A.C. Facer, David Feng, Tim P. Fliss, Garrett Gee, Jeff Goldy, Lindsey Gourley, Benjamin W. Gregor, Guangyu Gu, Robert E. Howard, Jayson M. Jochim, Chihchau L. Kuan, Christopher Lau, Chang Kyu Lee, Felix Lee, Tracy A. Lemon, Phil Lesnar, Bergen McMurray, Naveed Mastan, Nerick Mosqueda, Theresa Naluai-Cecchini, Nhan Kiet Ngo, Julie Nyhus, Aaron Oldre, Eric Olson, Jody Parente,

Patrick D. Parker, Sheana E. Parry, Allison Stevens, Mihovil Pletikos, Melissa Reding, Kate Roll, David Sandman, Melaine Sarreal, Sheila Shapouri, Nadiya V. Shapovalova, Elaine H. Shen, Nathan Sjoquist, Clifford R. Slaughterbeck, Michael Smith, Andy J. Sodt, Derric Williams, Lilla Zöllei, Bruce Fischl, Mark B. Gerstein, Daniel H. Geschwind, Ian A. Glass, Michael J. Hawrylycz, Robert F. Hevner, Hao Huang, Allan R. Jones, James A. Knowles, Pat Levitt, John W. Phillips, Nenad Šestan, Paul Wohnoutka, Chinh Dang, Amy Bernard, John G. Hohmann, and Ed S. Lein. 2014. “Transcriptional Landscape of the Prenatal Human Brain.” *Nature* 508 (7495): 199–206. https://doi.org/10.1038/nature13185.

Miller, Jeremy a, Song-Lin Ding, Susan M Sunkin, Kimberly a Smith, Lydia Ng, Aaron Szafer, Amanda Ebbert, Zackery L Riley, Joshua J Royall, Kaylynn Aiona, James M Arnold, Crissa Bennet, Darren Bertagnolli, Krissy Brouner, Stephanie Butler, Shiella Caldejon, Anita Carey, Christine Cuhaciyan, Rachel a Dalley, Nick Dee, Tim a Dolbeare, Benjamin a C Facer, David Feng, Tim P Fliss, Garrett Gee, Jeff Goldy, Lindsey Gourley, Benjamin W Gregor, Guangyu Gu, Robert E Howard, Jayson M Jochim, Chihchau L Kuan, Christopher Lau, Chang-Kyu Lee, Felix Lee, Tracy a Lemon, Phil Lesnar, Bergen McMurray, Naveed Mastan, Nerick Mosqueda, Theresa Naluai-Cecchini, Nhan-Kiet Ngo, Julie Nyhus, Aaron Oldre, Eric Olson, Jody Parente, Patrick D Parker, Sheana E Parry, Allison Stevens, Mihovil Pletikos, Melissa Reding, Kate Roll, David Sandman, Melaine Sarreal, Sheila Shapouri, Nadiya V Shapovalova, Elaine H Shen, Nathan Sjoquist, Clifford R Slaughterbeck, Michael Smith, Andy J Sodt, Derric Williams, Lilla Zöllei, Bruce Fischl, Mark B Gerstein, Daniel H Geschwind, Ian a Glass, Michael J Hawrylycz, Robert F Hevner, Hao Huang, Allan R Jones, James a Knowles, Pat Levitt, John W Phillips, Nenad Sestan, Paul Wohnoutka, Chinh Dang, Amy Bernard, John G Hohmann, and Ed S Lein. 2014. “Transcriptional Landscape of the Prenatal Human Brain.” *Nature* 508 (7495): 199–206. https://doi.org/10.1038/nature13185.

Moon, Kevin R., David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. 2019. “Visualizing Structure and Transitions in High-Dimensional Biological Data.” *Nature Biotechnology* 37 (12): 1482–92. https://doi.org/10.1038/s41587-019-0336-3.

Neph, Shane, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, Matthew T Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J Maccoss, Joshua M Akey, M A Bender, Mark Groudine, Rajinder Kaul, and John A Stamatoyannopoulos. 2012. “An Expansive Human Regulatory Lexicon Encoded in Transcription Factor Footprints.” *Nature* 488 (7414): 83–90. https://doi.org/10.1038/nature11212.

Nikolic, Aleksandar, Vladislav Volarevic, Lyle Armstrong, Majlinda Lako, and Miodrag Stojkovic. 2016. “Primordial Germ Cells : Current Knowledge and Perspectives” 2016. https://doi.org/10.1155/2016/1741072.

Ortmann, Daniel, and Ludovic Vallier. 2017. “Variability of Human Pluripotent Stem Cell

Lines." *Current Opinion in Genetics & Development* 46: 179–85. https://doi.org/10.1016/j.gde.2017.07.004.

Pääbo, Svante. 2014. "The Human Condition - A Molecular Approach." *Cell* 157 (1): 216–26. https://doi.org/10.1016/j.cell.2013.12.036.

Parekh, Udit, Yan Wu, Dongxin Zhao, Atharv Worlikar, Neha Shah, Kun Zhang, and Prashant Mali. 2018. "Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout." *Cell Systems*, 1–8. https://doi.org/10.1016/j.cels.2018.10.008.

Paul, Franziska, Ya'Ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. 2015. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors." *Cell* 163 (7): 1663–77. https://doi.org/10.1016/j.cell.2015.11.013.

Peter, Isabelle S, and Eric H Davidson. 2011. "Evolution of Gene Regulatory Networks Controlling Body Plan Development." *Cell* 144 (6): 970–85. https://doi.org/10.1016/j.cell.2011.02.017.

Philipp, Friederike, Anton Selich, Michael Rothe, Dirk Hoffmann, Susanne Rittinghausen, Michael A Morgan, Denise Klatt, Silke Glage, Stefan Lienenklaus, Vanessa Neuhaus, Katherina Sewald, Armin Braun, and Axel Schambach. 2018. "Human Teratoma-Derived Hematopoiesis Is a Highly Polyclonal Process Supported by Human Umbilical Vein Endothelial Cells." *Stem Cell Reports* 11 (5): 1051–60. https://doi.org/10.1016/j.stemcr.2018.09.010.

Phipson, Belinda, Pei X Er, Alexander N Combes, Thomas A Forbes, Sara E Howden, Luke Zappia, Hsan-Jan Yen, Kynan T Lawlor, Lorna J Hale, Jane Sun, Ernst Wolvetang, Minoru Takasato, Alicia Oshlack, and Melissa H Little. 2019. "Evaluation of Variability in Human Kidney Organoids." *Nature Methods* 16 (1): 79–87. https://doi.org/10.1038/s41592-018-0253-2.

Pijuan-Sala, Blanca, Jonathan A. Griffiths, Carolina Guibentif, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C. V. Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D. Simons, Jennifer Nichols, John C. Marioni, and Berthold Göttgens. 2019. "A Single-Cell Molecular Map of Mouse Gastrulation and Early Organogenesis." *Nature*. https://doi.org/10.1038/s41586-019-0933-9.

Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, Andrew C. Adey, Frank J. Steemers, Jay Shendure, and Cole Trapnell. 2018. "Cicero Predicts Cis -Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data." *Molecular Cell*, 1–14. https://doi.org/10.1016/j.molcel.2018.06.044.

Polioudakis, Damon, Luis de la Torre-Ubieta, Justin Langerman, Andrew G. Elkins, Xu Shi,

Jason L. Stein, Celine K. Vuong, Susanne Nichterwitz, Melinda Gevorgian, Carli K. Opland, Daning Lu, William Connell, Elizabeth K. Ruzzo, Jennifer K. Lowe, Tarik Hadzic, Flora I. Hinz, Shan Sabri, William E. Lowry, Mark B. Gerstein, Kathrin Plath, and Daniel H. Geschwind. 2019. “A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-Gestation.” *Neuron*, 1–17. https://doi.org/10.1016/j.neuron.2019.06.011.

Pollard, Katherine S., Sofie R. Salama, Bryan King, Andrew D. Kern, Tim Dreszer, Sol Katzman, Adam Siepel, Jakob S. Pedersen, Gill Bejerano, Robert Baertsch, Kate R. Rosenbloom, Jim Kent, and David Haussler. 2006. “Forces Shaping the Fastest Evolving Regions in the Human Genome.” *PLoS Genetics* 2 (10): 1599–1611. https://doi.org/10.1371/journal.pgen.0020168.

Prabhakar, Shyam, James P. Noonan, Svante Pääbo, and Edward M. Rubin. 2006. “Accelerated Evolution of Conserved Noncoding Sequences in Humans.” *Science* 314 (5800): 786. https://doi.org/10.1126/science.1130738.

Puram, Sidharth V, Itay Tirosh, Anuraag S Parikh, Derrick T Lin, Aviv Regev, Bradley E Bernstein Correspondence, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, Daniel G Deschler, Mark A Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James W Rocco, William C Faquin, and Bradley E Bernstein. 2017. “Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer.” *Cell* 172: 1–14. https://doi.org/10.1016/j.cell.2017.10.044.

Qi, Lei S., Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, Wendell A. Lim, R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero, P. Horvath, R.R. Beerli, C.F. Barbas, R.E. Campbell, O. Tour, A.E. Palmer, P.A. Steinbach, G.S. Baird, D.A. Zacharias, R.Y. Tsien, S.W. Cho, S. Kim, J.M. Kim, J.-S. Kim, L.S. Churchman, J.S. Weissman, L. Cong, F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, F. Zhang, E. Deltcheva, K. Chylinski, C.M. Sharma, K. Gonzales, Y. Chao, Z.A. Pirzada, M.R. Eckert, J. Vogel, E. Charpentier, G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, G.J. Hannon, W.Y. Hwang, Y. Fu, D. Reyon, M.L. Maeder, S.Q. Tsai, J.D. Sander, R.T. Peterson, J.-R.J. Yeh, J.K. Joung, W. Jiang, D. Bikard, D. Cox, F. Zhang, L.A. Marraffini, M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, E. Charpentier, M. Jinek, A. East, A. Cheng, S. Lin, E. Ma, J. Doudna, A. Klug, M. Lewis, J.B. Lucks, S.A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G.P. Schroth, L. Pachter, J.A. Doudna, A.P. Arkin, R. Lutz, H. Bujard, K.S. Makarova, D.H. Haft, R. Barrangou, S.J.J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F.J.M. Mojica, Y.I. Wolf, A.F. Yakunin, et al., P. Mali, L. Yang, K.M. Esvelt, J. Aach, M. Guell, J.E. Dicarlo, J.E. Norville, G.M. Church, L.A. Marraffini, E.J. Sontheimer, A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, E. Nudler, A. Goldfarb, M. Kashlev, J.-D. Pédelacq, S. Cabantous, T. Tran, T.C. Terwilliger, G.S. Waldo, L. Qi, R.E. Haurwitz, W. Shao, J.A. Doudna, A.P. Arkin, H.H. Wang, F.J. Isaacs, P.A. Carr, Z.Z. Sun, G. Xu, C.R. Forest, G.M. Church, B. Wiedenheft, G.C. Lander, K. Zhou, M.M. Jore, S.J.J. Brouns, J. van der Oost, J.A. Doudna, E. Nogales, B. Wiedenheft, S.H. Sternberg, J.A. Doudna, P.D. Zamore, T. Tuschl, P.A. Sharp, D.P. Bartel, F. Zhang, L. Cong, S. Lodato, S. Kosuri, G.M. Church, and P. Arlotta. 2013. “Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.” *Cell* 152 (5): 1173–83.

https://doi.org/10.1016/j.cell.2013.02.022.

Qin, Qian, Young Xu, Tao He, Chunlin Qin, and Jianming Xu. 2011. “Normal and Disease-Related Biological Functions of Twist1 and Underlying Molecular Mechanisms.” *Nature Publishing Group* 22 (1): 90–106. https://doi.org/10.1038/cr.2011.144.

Qiu, Xiaojie, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. 2017. “Reversed Graph Embedding Resolves Complex Single-Cell Trajectories.” *Nature Methods* 14 (10): 979–82. https://doi.org/10.1038/nmeth.4402.

Quadrato, Giorgia, Tuan Nguyen, Evan Z. Macosko, John L. Sherwood, Sung Min Yang, Daniel R. Berger, Natalie Maria, Jorg Scholvin, Melissa Goldman, Justin P. Kinney, Edward S. Boyden, Jeff W. Lichtman, Ziv M. Williams, Steven A. McCarroll, and Paola Arlotta. 2017. “Cell Diversity and Network Dynamics in Photosensitive Human Brain Organoids.” *Nature* 545 (7652): 48–53. https://doi.org/10.1038/nature22047.

Raff, R.A. 1996. *The Shape of Life; Genes, Development and the Evolution of Animal Form.* Chicago, IL: University of Chicago Press.

Richard, Isabelle, Marc Abitbol, David Wilson, Francoise Fougerousse, Jacques S Beckmann, Laurence Suel, Muriel Durand, Muriel Herasse, Philip Bullen, Steve Robson, Susan Lindsay, and Tom Strachan. 2000. “Human–Mouse Differences in the Embryonic Expression Patterns of Developmental Control Genes and Disease Genes.” *Human Molecular Genetics* 9 (2): 165–73. https://doi.org/10.1093/hmg/9.2.165.

Richardson, M K, J Hanken, M L Gooneratne, C Pieau, A Raynaud, L Selwood, and G M Wright. 1997. “There Is No Highly Conserved Embryonic Stage in the Vertebrates: Implications for Current Theories of Evolution and Development.” *Anatomy and Embryology* 196 (2): 91–106.

Rilling, James K. 2014. “Comparative Primate Neuroimaging: Insights into Human Brain Evolution.” *Trends in Cognitive Sciences* 18 (1): 46–55. https://doi.org/10.1016/j.tics.2013.09.013.

Rosenberg, Alexander B., Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas Gray, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. 2018. “Single-Cell Profiling of the Developing Mouse Brain and Spinal Cord with Split-Pool Barcoding.” *Science* 12 (April): eaam8999. https://doi.org/10.1126/science.aam8999.

Rosenberg, Alexander B, Charles M Roco, Richard A Muscat, Anna Kuchina, Wei Chen, David J Peeler, Zizhen Yao, Bosiljka Tasic, Drew L Sellers, H Pun, and Georg Seelig. 2017. “Scaling Single Cell Transcriptomics through Split Pool Barcoding.” *Bioarxiv*. https://doi.org/10.1101/105163.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20 (C): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Royo, Jose Luis, Ignacio Maeso, Manuel Irimia, Feng Gao, Isabelle S Peter, Carla S Lopes,

Salvatore D'Aniello, Fernando Casares, Eric H Davidson, Jordi Garcia-Fernandez, and Jose Luis Gomez-Skarmeta. 2011. "Transphyletic Conservation of Developmental Regulatory State in Animal Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 108 (34): 14186–91. https://doi.org/10.1073/pnas.1109037108.

Russo, Roberta, Roberta Marra, Immacolata Andolfo, Gianluca De Rosa, Barbara Eleni Rosato, Francesco Manna, Antonella Gambale, Maddalena Raia, Sule Unal, Susanna Barella, and Achille Iolascon. 2019. "Characterization of Two Cases of Congenital Dyserythropoietic Anemia Type I Shed Light on the Uncharacterized C15orf41 Protein." *Frontiers in Physiology* 10 (MAY): 1–10. https://doi.org/10.3389/fphys.2019.00621.

Ryu, Hane, Fumitaka Inoue, Sean Whalen, Alex Williams, Martin Kircher, Beth Martin, Beatriz Alvarado, Abul Hassan Samee, Kathleen Keough, and Sean Thomas. 2018. "Massively Parallel Dissection of Human Accelerated Regions in Human and Chimpanzee Neural Progenitors." *BioRxiv*.

Saffman, E E, and P Lasko. 1999. "Germline Development in Vertebrates and Invertebrates." *Cellular and Molecular Life Sciences : CMLS* 55 (8–9): 1141–63.

Sammon, John W. 1969. "A Nonlinear Mapping for Data Structure Analysis." *IEEE Transactions on Computers* C–18 (5): 401–9. https://doi.org/10.1109/T-C.1969.222678.

Sander, Thomas. 1997. "Allelic Association of Juvenile Absence Epilepsy with a GluR5 Kainate Receptor Gene (GRIK1) Polymorphism." *American Journal of Medical Genetics - Neuropsychiatric Genetics* 74 (4): 416–21. https://doi.org/10.1002/(SICI)1096-8628(19970725)74:4<416::AID-AJMG13>3.0.CO;2-L.

Sarper, Safiye E., Toshihiro Inubushi, Hiroshi Kurosaka, Hitomi Ono Minagi, Koh ichi Kuremoto, Takayoshi Sakai, Ichiro Taniuchi, and Takashi Yamashiro. 2018. "Runx1-Stat3 Signaling Regulates the Epithelial Stem Cells in Continuously Growing Incisors." *Scientific Reports* 8 (1): 1–12. https://doi.org/10.1038/s41598-018-29317-6.

Satija, Rahul, Andrew Butler, and Paul Hoffman. 2018. "Seurat: Tools for Single Cell Genomics." https://cran.r-project.org/package=Seurat.

Sato, Toshiro, Daniel E. Stange, Marc Ferrante, Robert G.J. Vries, Johan H. Van Es, Stieneke Van Den Brink, Winan J. Van Houdt, Apollo Pronk, Joost Van Gorp, Peter D. Siersema, and Hans Clevers. 2011. "Long-Term Expansion of Epithelial Organoids from Human Colon, Adenoma, Adenocarcinoma, and Barrett's Epithelium." *Gastroenterology* 141 (5): 1762–72. https://doi.org/10.1053/j.gastro.2011.07.050.

Sato, Toshiro, Robert G. Vries, Hugo J. Snippert, Marc Van De Wetering, Nick Barker, Daniel E. Stange, Johan H. Van Es, Arie Abo, Pekka Kujala, Peter J. Peters, and Hans Clevers. 2009. "Single Lgr5 Stem Cells Build Crypt-Villus Structures in Vitro without a Mesenchymal Niche." *Nature* 459 (7244): 262–65. https://doi.org/10.1038/nature07935.

Schep, Alicia N., Beijing Wu, Jason D. Buenrostro, and William J. Greenleaf. 2017. "ChromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data." *Nature Methods* 14 (10): 975–78.

https://doi.org/10.1038/nmeth.4401.

Seigneur, Erica, and Thomas C. Sudhof. 2017. “Cerebellins Are Differentially Expressed in Selective Subsets of Neurons Throughout the Brain.” *J Comp Neurol* 525 (15): 3286–3311. https://doi.org/10.1126/science.1249098.Sleep.

Setty, Manu, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe’er. 2016. “Wishbone Identifies Bifurcating Developmental Trajectories from Single-Cell Data.” *Nature Biotechnology* 34 (April): 1–14. https://doi.org/10.1038/nbt.3569.

Silberg, Debra G, Jessica Sullivan, Eugene Kang, Gary P Swain, Jennifer Moffett, Newman J Sund, Sara D Sackett, and Klaus H Kaestner. 2002. “Cdx2 Ectopic Expression Induces Gastric Intestinal Metaplasia,” 689–96. https://doi.org/10.1053/gast.2002.31902.

Simmini, Salvatore, Monika Bialecka, Meritxell Huch, Lennart Kester, Marc Van De Wetering, Toshiro Sato, Felix Beck, Alexander Van Oudenaarden, Hans Clevers, and Jacqueline Deschamps. 2014. “Transformation of Intestinal Stem Cells into Gastric Stem Cells on Loss of Transcription Factor Cdx2.” *Nature Communications* 5: 1–10. https://doi.org/10.1038/ncomms6728.

Smith, Kelly P., Mai X. Luong, and Gary S. Stein. 2009. “Pluripotency: Toward a Gold Standard for Human ES and IPS Cells.” *Journal of Cellular Physiology* 220 (1): 21–29. https://doi.org/10.1002/jcp.21681.

Souza, Natalie de. 2017. “Organoid Variability Examined.” *Nature Methods* 14 (June): 655.

Stevens, LC. 1962. “The Biology of Teratomas Including Evidence Indicating Their Origin Form Primordial Germ Cells.” *Annee Biol.* 1: 585–610.

———. 1967. “THE BIOLOGY OF TERATOMAS.” *Adv Morphog* 6: 1–31.

Stevens, Leroy C., and G. Barry Pierce. 1975. “Teratomas: Definitions and Terminology.” *Teratomas and Differentiation*, 13–14.

Stott, Simon R.W., Emmanouil Metzakopian, Wei Lin, Klaus H. Kaestner, Rene Hen, and Siew Lan Ang. 2013. “Foxa1 and Foxa2 Are Required for the Maintenance of Dopaminergic Properties in Ventral Midbrain Neurons at Late Embryonic Stages.” *Journal of Neuroscience* 33 (18): 8022–34. https://doi.org/10.1523/JNEUROSCI.4774-12.2013.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177 (7): 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William Mauck, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2018. “Comprehensive Integration of Single Cell Data.” *BioRxiv*, 1–34. https://doi.org/10.1101/460147.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Suzuki, Nao, Satoshi Yamazaki, Tomoyuki Yamaguchi, Motohito Okabe, Hideki Masaki, Satoshi Takaki, Makoto Otsu, and Hiromitsu Nakauchi. 2013. “Generation of Engraftable Hematopoietic Stem Cells From Induced Pluripotent Stem Cells by Way of Teratoma Formation.” *Molecular Therapy* 21 (7): 1424–31. https://doi.org/10.1038/mt.2013.71.

Tarlow, Daniel, Kevin Swersky, Laurent Charlin, Ilya Sutskever, and Rich Zemel. 2013. “Stochastic K-Neighborhood Selection for Supervised and Unsupervised Learning.” *Proceedings of the 30th International Conference on Machine Learning* 28 (3): 199–207. http://proceedings.mlr.press/v28/tarlow13.html.

THURLBECK, WLLIAM M., ROBERT E. SCULLY. 1973. “Solid Teratoma of the Ovary: A Clinicopatlzological Analysis of 9 Cases,” no. January 1960: 2563–71.

Tirosh, I., B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. K. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A.-C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jane-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, and L. A. Garraway. 2016. “Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq.” *Science* 352 (6282): 189–96. https://doi.org/10.1126/science.aad0501.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86. https://doi.org/10.1038/nbt.2859.

Tsukada, Masao, Yasunori Ota, Adam C Wilkinson, Hans J Becker, Motomi Osato, Hiromitsu Nakauchi, and Satoshi Yamazaki. 2017. “In Vivo Generation of Engraftable Murine Hematopoietic Stem Cells by Gfi1b, c-Fos, and Gata2 Overexpression within Teratoma.” *Stem Cell Reports* 9 (4): 1024–33. https://doi.org/10.1016/j.stemcr.2017.08.010.

Tsunemoto, Rachel, Sohyon Lee, Attila Szucs, Pavel Chubukov, Irina Sokolova, Joel W. Blanchard, Kevin T. Eade, Jacob Bruggemann, Chunlei Wu, Ali Torkamani, Pietro Paolo Sanna, and Kristin K. Baldwin. 2018. “Diverse Reprogramming Codes for Neuronal Identity.” *Nature* 557 (7705): 375–80. https://doi.org/10.1038/s41586-018-0103-5.

Umansky, Kfir Baruch, Yael Gruenbaum-Cohen, Michael Tsoory, Ester Feldmesser, Dalia Goldenberg, Ori Brenner, and Yoram Groner. 2015. “Runx1 Transcription Factor Is

Required for Myoblasts Proliferation during Muscle Regeneration." *PLoS Genetics* 11 (8): 1–31. https://doi.org/10.1371/journal.pgen.1005457.

van de Wetering, Marc, Hayley E. Francies, Joshua M. Francis, Gergana Bounova, Francesco Iorio, Apollo Pronk, Winan van Houdt, Joost van Gorp, Amaro Taylor-Weiner, Lennart Kester, Anne McLaren-Douglas, Joyce Blokker, Sridevi Jaksani, Sina Bartfeld, Richard Volckman, Peter van Sluis, Vivian S.W. Li, Sara Seepo, Chandra Sekhar Pedamallu, Kristian Cibulskis, Scott L. Carter, Aaron McKenna, Michael S. Lawrence, Lee Lichtenstein, Chip Stewart, Jan Koster, Rogier Versteeg, Alexander van Oudenaarden, Julio Saez-Rodriguez, Robert G.J. Vries, Gad Getz, Lodewyk Wessels, Michael R. Stratton, Ultan McDermott, Matthew Meyerson, Mathew J. Garnett, and Hans Clevers. 2015. "Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients." *Cell* 161 (4): 933–45. https://doi.org/https://doi.org/10.1016/j.cell.2015.03.053.

Vastag, Livia, Paul Jorgensen, Leonid Peshkin, Ru Wei, Joshua D. Rabinowitz, and Marc W. Kirschner. 2011. "Remodeling of the Metabolome during Early Frog Development." *PLoS ONE* 6 (2). https://doi.org/10.1371/journal.pone.0016881.

Velasco, Silvia, Amanda J. Kedaigle, Sean K. Simmons, Allison Nash, Marina Rocha, Giorgia Quadrato, Bruna Paulsen, Lan Nguyen, Xian Adiconis, Aviv Regev, Joshua Z. Levin, and Paola Arlotta. 2019. "Individual Brain Organoids Reproducibly Form Cell Diversity of the Human Cerebral Cortex." *Nature*. https://doi.org/10.1038/s41586-019-1289-x.

Wagner, Daniel E., Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, and Allon M. Klein. 2018. "Single-Cell Mapping of Gene Expression Landscapes and Lineage in the Zebrafish Embryo Daniel." *Science* 25 (3): 289–313. https://doi.org/10.1007/s11065-015-9294-9.Functional.

Wang, Bo, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. 2017. "Visualization and Analysis of Single-Cell RNA-Seq Data by Kernel-Based Similarity Learning." *Nature Methods*, no. June 2016: 1–6. https://doi.org/10.1038/nmeth.4207.

Wang, Jiaxu, Vladimir Espinosa Angarica, Akshay Bhinge, Piroon Jenjaroenpun, Antonio Del Sol, Vladimir A. Kuznetsov, Intawat Nookaew, and Lawrence W. Stanton. 2017. "Single-Cell Gene Expression Analysis Reveals Regulators of Distinct Cell Subpopulations among Developing Human Neurons." *Genome Research* 27 (11): 1783–94. https://doi.org/10.1101/gr.223313.117.

Willis, Rupert A. 1934. "The Structure of Teratoma." *The Journal of Pathology and Bacteriology* XL (I).

———. 1935. "THE HISTOGENESIS OF NEURAL TISSUE IN TERATOMAS . (PLATES." *The Journal of Pathology and Bacteriology*.

Won, Hyejung, Jerry Huang, Carli K. Opland, Chris L. Hartl, and Daniel H. Geschwind. 2019. "Human Evolved Regulatory Elements Modulate Genes Involved in Cortical Expansion and Neurodevelopmental Disease Susceptibility." *Nature Communications* 10 (1): 1–11. https://doi.org/10.1038/s41467-019-10248-3.

Won, Hyejung, Luis De La Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Jerry Huang, Carli K. Opland, Michael J. Gandal, Gavin J. Sutton, Farhad Hormozdiari, Daning Lu, Changhoon Lee, Eleazar Eskin, Irina Voineagu, Jason Ernst, and Daniel H. Geschwind. 2016. “Chromosome Conformation Elucidates Regulatory Relationships in Developing Human Brain.” *Nature* 538 (7626): 523–27. https://doi.org/10.1038/nature19847.

Wu, Yan, Pablo Tamayo, and Kun Zhang. 2018. “Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding.” *Cell Systems* 7 (6): 656-666.e4. https://doi.org/10.1101/276261.

Xin, Mei, Eric N. Olson, and Rhonda Bassel-Duby. 2013. “Mending Broken Hearts: Cardiac Development as a Basis for Adult Heart Regeneration and Repair.” *Nature Reviews Molecular Cell Biology* 14 (8): 529–41. https://doi.org/10.1038/nrm3619.

Yang, Jing, Sendurai A Mani, Joana Liu Donaher, Sridhar Ramaswamy, Raphael A Itzykson, Christophe Come, Pierre Savagner, Inna Gitelman, Andrea Richardson, Robert A Weinberg, Crlc Val, and Aurelle-paul Lamarque. 2004. “Twist , a Master Regulator of Morphogenesis , Plays an Essential Role in Tumor Metastasis” 117: 927–39.

Yao, Zizhen, John K. Mich, Sherman Ku, Vilas Menon, Anne Rachel Krostag, Refugio A. Martinez, Leon Furchtgott, Heather Mulholland, Susan Bort, Margaret A. Fuqua, Ben W. Gregor, Rebecca D. Hodge, Anu Jayabalu, Ryan C. May, Samuel Melton, Angelique M. Nelson, N. Kiet Ngo, Nadiya V. Shapovalova, Soraya I. Shehata, Michael W. Smith, Leah J. Tait, Carol L. Thompson, Elliot R. Thomsen, Chaoyang Ye, Ian A. Glass, Ajamete Kaykas, Shuyuan Yao, John W. Phillips, Joshua S. Grimley, Boaz P. Levi, Yanling Wang, and Sharad Ramanathan. 2017. “A Single-Cell Roadmap of Lineage Bifurcation in Human ESC Models of Embryonic Brain Development.” *Cell Stem Cell* 20 (1): 120–34. https://doi.org/10.1016/j.stem.2016.09.011.

Yin, Xiaolei, Benjamin E. Mead, Helia Safaee, Robert Langer, Jeffrey M. Karp, and Oren Levy. 2016. “Engineering Stem Cell Organoids.” *Cell Stem Cell* 18 (1): 25–38. https://doi.org/10.1016/j.stem.2015.12.005.

Yu, Guangchuang, Li Gen Wang, and Qing Yu He. 2015. “ChIP Seeker: An R/Bioconductor Package for ChIP Peak Annotation, Comparison and Visualization.” *Bioinformatics* 31 (14): 2382–83. https://doi.org/10.1093/bioinformatics/btv145.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack. 2017. “Splatter: Simulation of Single-Cell RNA Sequencing Data.” *Genome Biology* 18 (1): 174. https://doi.org/10.1186/s13059-017-1305-0.

Zaret, Kenneth S, and Susan E Mango. 2016. “Pioneer Transcription Factors, Chromatin Dynamics, and Cell Fate Control.” *Current Opinion in Genetics & Development* 37: 76–81. https://doi.org/10.1016/j.gde.2015.12.003.

Zhang, Xiaoqing, Cindy T. Huang, Jing Chen, Matthew T. Pankratz, Jiajie Xi, Jin Li, Ying Yang, Timothy M. LaVaute, Xue Jun Li, Melvin Ayala, Gennadiy I. Bondarenko, Zhong Wei Du, Ying Jin, Thaddeus G. Golos, and Su Chun Zhang. 2010. “Pax6 Is a Human Neuroectoderm Cell Fate Determinant.” *Cell Stem Cell* 7 (1): 90–100.

https://doi.org/10.1016/j.stem.2010.04.017.

Zheng, Grace X.Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8: 1–12. https://doi.org/10.1038/ncomms14049.

Zhong, Suijuan, Shu Zhang, Xiaoying Fan, Qian Wu, Liying Yan, Ji Dong, Haofeng Zhang, Long Li, Le Sun, Na Pan, Xiaohui Xu, Fuchou Tang, Jun Zhang, Jie Qiao, and Xiaoqun Wang. 2018. “A Single-Cell RNA-Seq Survey of the Developmental Landscape of the Human Prefrontal Cortex.” *Nature*. https://doi.org/10.1038/nature25980.

Zhu, Y. Y., E. M. Machleder, a. Chenchik, R. Li, and P. D. Siebert. 2001. “Reverse Transcriptase Template Switching: A SMART??? Approach for Full-Length CDNA Library Construction.” *BioTechniques* 30 (4): 892–97.

Zhu, Ying, André M. M. Sousa, Tianliuyun Gao, Mario Skarica, Mingfeng Li, Gabriel Santpere, Paula Esteller-Cucala, David Juan, Luis Ferrández-Peral, Forrest O. Gulden, Mo Yang, Daniel J. Miller, Tomas Marques-Bonet, Yuka Imamura Kawasawa, Hongyu Zhao, and Nenad Sestan. 2018. “Spatiotemporal Transcriptomic Divergence across Human and Macaque Brain Development.” *Science* 362 (6420): eaat8077. https://doi.org/10.1126/science.aat8077.