



PhD in biophysics

**Computational tools for identifying
and relating cell-types in development
and stem-cell engineering**

Alexander Valentin Nielsen

Supervised by Ala Trusina

September 2022

Alexander Valentin Nielsen

Computational tools for identifying and relating cell-types in development and stem-cell engineering

PhD in biophysics, September 2022

Supervisor: Ala Trusina

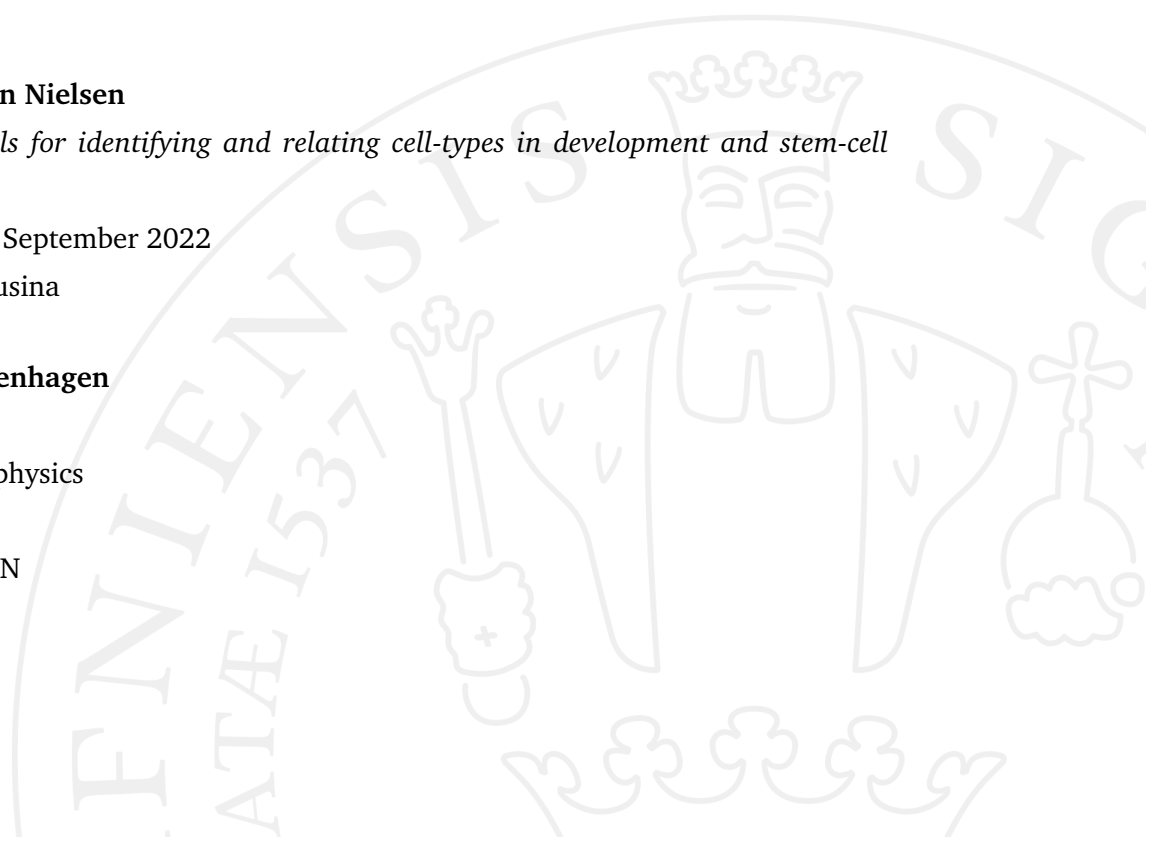
University of Copenhagen

Niels Bohr Institute

PhD Degree in biophysics

Blegdamsvej 17

2200 Copenhagen N



Abstract

The ability of stem-cells to differentiate into various cell-types is the cornerstone underpinning the marvellous diversity of tissues that make up animals. Despite their absolutely integral role in the formation of all multi-cellular life, the mechanisms that underlie stem-cells ability to orchestrate and coordinate the choices surrounding differentiation are not fully understood. In this thesis I describe the work from 3 articles: For the first article, we investigated the changes stem-cells undergo when they first start forming the mammalian gut. We did this by examining the RNA content of thousands of individual cells obtained from mouse embryos using single-cell RNA sequencing (scRNAseq) throughout the early days of development. To aid in this investigation, I developed a computational algorithm to compare and quantify the relationship between noisy high-dimensional data, such as scRNAseq data. Analysing the scRNAseq data, we find and characterise a group of cells previously thought to be confined to the extra-embryonic membranes surrounding the embryo, that seemingly defy their differentiation decision, and end up becoming part of the embryo's gut. In the 2 last articles, we investigated the differentiation decisions cells undergo when being chemically reprogrammed into different cell-types. We find a set of genes that are critical to activate for the reprogramming of somatic fibroblast cells to XEN-like cells to be successful. We further show that a set of chemicals can be used as a starting point for making new reprogramming protocols for multiple types of cells, showing that it works for generating both neuronal and muscle cells from fibroblast.

Resumé (Abstract in Danish)

Stamcellers evne til at differentiere til forskellige celle-typer er hjørnesteinen bag den fantastiske diversitet af vævstyper der udgør et dyr. På trods af deres fuldstændig afgørende rolle i formationen af alt multi-cellulært liv er mekanismen bag stam-cellers differentieringsevne ikke fuldt ud forstået. I denne afhandling vil jeg beskrive det arbejde der står 3 artikler: Den første artikel omhandler de ændringer stam-celler tager mens du udvikler tarmsystemet i pattedyr. Vi undersøgte dette ved at kigge på RNA indholdet af tusindvis af celler fra tidlige musefostre ved hjælp af single cell RNA sekventering (scRNAseq). Til at hjælpe i denne afsøgning udviklede jeg en algoritme til at sammenligne og kvantifiere forholdet mellem høj-dimensionel data, så som scRNAseq data. Ved analyser af denne data fandt vi en gruppe af celler som normalt antages at være del af de membraner som ligger uden om fostret, men som viser sig at ændre deres celle-type og bliver til en del af tarmsystemet. I de 2 andre artikler arbejdede vi med celle differentiering i kontekst af kemisk omprogrammering af celle-typer. Vi viser hvordan et sæt af gener er helt centrale i forhold til omprogrammeringen af fibroblast til XEN-lignende celler. Vi viser ydermere hvordan et sæt af kemikalier kan bruges til at lave omprogrammings protokoller til mange forskellige celle-type, herunder hjerne og muskel celler.

Acknowledgements

I would like to sincerely thank my supervisor, Ala Trusina, for inspiring me to become a better researcher. I deeply appreciate the support, help and guidance she has given me over the years. I could not have asked for a better supervisor.

I would also like to thank my unofficial co-supervisor Joshua Brickman and all the members of his lab, for generously sharing their biology knowledge and for patiently answering all my questions about embryo development. Particularly Michaela Rothova and Martin Proks for their invaluable insight, humor and good company.

Thanks to Bente Markussen for making all the practical details and paperwork related to my studies so much easier.

Thanks to Xiaojing Yang for her hospitality and help during me stay abroad at Peking University in Chao Tang's lab. Xiao Chan and Huixia Ren have my utmost appreciation and respect for going above and beyond to make me feel welcome in their lab and country.

Thanks to the members of the B-floor of the biophysics building at the Niels Bohr Institute, for providing an amazing environment to be in, both for scientific discussions and socializing. I am going to miss being a part of it.

And most importantly, thanks to my wife, Yue He, who makes every part of my life better. I am thankful for her encouragement during the difficult times of my study and her celebration of the good.

Contents

1	Introduction to the thesis	1
2	CAT - A tool providing quantitative comparison between cell-types	3
2.1	Introduction	4
2.1.1	Datasets	4
2.1.2	Standard processing of scRNAseq data	6
2.2	Cluster Alignment Tool (CAT) - Principles and functionalities	25
2.2.1	How the algorithm works	25
2.2.2	Example CAT usage	30
2.2.3	Limitations and strengths of CAT	33
2.3	Conclusion	42
3	Analysing data obtained on embryogenesis and gastrulation	44
3.1	Introduction	44
3.2	Methods and results	48
3.2.1	Identifying the embryonic cell types from experimental <i>in vivo</i> data	48
3.2.2	Determining the cell-types made using <i>in vitro</i> differentiation protocols	52
3.3	Conclusion	58
4	Chemical reprogramming of somatic mouse cells	60
4.1	Introduction: Cell reprogramming	60
4.2	Article 1	62
4.3	Article 2	64
5	Bibliography	66
6	Supplementary information	81
6.1	CAT tables	83
7	Appendix	93

List of abbreviations

CAT	Cluster Alignment Tool
CiPSC	Chemically induced pluripotent stem cells
DE	Definitive endoderm
DEG	Differentially expressed genes
EmVE	Embryonic visceral endoderm
ESC	Embryonic stem cells
ExVE	Extraembryonic visceral endoderm
HVG	High variable gene
ICM	Inner cell mass
InterVE	Intermediate visceral endoderm
iPSC	Induced pluripotent stem-cells
MDS	Multidimensional scaling
MEF	Mouse embryonic fibroblast
MNIST	Modified National Institute of Standards and Technology
nEnd	Naïve extra-embryonic endoderm
NLDR	Non-linear dimensionality reduction
PCA	Principal component analysis
PS	Primitive streak
scRNAseq	Single cell RNA sequencing
TF	Transcription factor
VE	Visceral endoderm
XEN	Extra-embryonic endoderm

Introduction to the thesis

This PhD project was structured as a 4-year integrated MSc/PhD programme, a programme structure offered at Danish universities. The programme consists of 2 parts, each lasting 2 years. The first two years consist of MSc studies performed alongside the PhD and conclude with a thesis that serves as a qualifying exam for the final two years of the PhD studies. This thesis is the conclusion of the second part of the programme, and the work presented here will therefore focus on the work done in the second half of the PhD programme.

The main recurring questions throughout this thesis are the following:

- 1) What cell-types are there? (In a given single-cell RNA sequencing dataset)
- 2) How are the cell-types related to each other? (E.g. what are the developmental trajectories between the types? How do we make sure we make the right cell-types in vitro?)
- 3) How to engineer cell-types in vitro?

The thesis is structured into three parts, each with their own introduction, results and conclusion:

Chapter 1: Introduces CAT, an algorithm I developed to compare cell-types from single-cell RNA sequencing data. This chapter deals with questions 1) and 2).

Chapter 2: Briefly introduces early embryo formation and describes our study of the cell-types that develop into the gut. This chapter deals with questions 1), 2), and 3).

Chapter 3: Discusses how to chemically reprogram somatic cells into other somatic cell-types. This chapter deals mainly with question 3).

During my PhD, I co-authored four articles, which are attached in the appendix. One of these articles was published in the first part of the PhD programme, so this thesis will only cover the work of the three most recent articles, along with additional results that are not yet published.

Publications covered in the thesis

Rothová, M. M.¹, Nielsen, A. V.¹, Proks, M.¹, Wong, Y. F., Riveiro, A. R., Linneberg-Agerholm, M., David, E., Amit, I., Trusina, A., Brickman, J. M. Identification of the central intermediate in the extra-embryonic to embryonic endoderm transition through single-cell transcriptomics. *Nature Cell Biology*, 1-12. (2022).

Yang, Z., Xu, X., Gu, C., Nielsen, A. V., Chen, G., Guo, F., Tang, C., and Zhao, Y. Chemical pretreatment activated a plastic state amenable to direct lineage reprogramming. *Frontiers in cell and developmental biology*, 10. (2022).

Yang, Z., Xu, X., Gu, C., Li, J., Wu, Q., Ye, C., Nielsen, A. V., Mao, L., Ye, J., Bai, K., et al. Chemicals orchestrate reprogramming with hierarchical activation of master transcription factors primed by endogenous sox17 activation. *Communications biology*, 3(1):1–10. (2020).

Additional publication

Larsen, H. L., Martín-Coll, L., Nielsen, A. V., Wright, C. V., Trusina, A., Kim, Y. H., & Grapin-Botton, A. Stochastic priming and spatial cues orchestrate heterogeneous clonal contribution to mouse pancreas organogenesis. *Nature communications*, 8(1), 1-13. (2017).

¹These authors contributed equally.

CAT - A tool providing quantitative comparison between cell-types

This chapter presents the computational tool CAT, which I developed as part of the work that resulted in the following article:

Rothová, M. M.¹, Nielsen, A. V.¹, Proks, M.¹, Wong, Y. F., Riveiro, A. R., Linneberg-Agerholm, M., David, E., Amit, I., Trusina, A., Brickman, J. M. (2022). Identification of the central intermediate in the extra-embryonic to embryonic endoderm transition through single-cell transcriptomics. *Nature Cell Biology*, 1-12.

In the article, we present a single-cell RNA sequencing (scRNAseq) dataset covering the early stages of mouse embryo development, which serves as the foundation of the investigation. CAT was designed to make analysis of such data easier. Specifically, the purpose of the algorithm is to reveal the relationship between groups of cells in scRNAseq data. However, as we will see, it also works for other types of high-dimensional data. Our article's primary focus is the biology surrounding the specification and development of the gut, and while CAT is introduced as a method and its results displayed, the method itself is not thoroughly discussed in the article. In this chapter, I will therefore go into more detail about why it was necessary to develop CAT, describe more precisely how it works and discuss results based on examples. In the next chapter, I will introduce the main biology results from the article.

The main question that drove the development of CAT is the second question stated in the "Introduction to the thesis": How are the cell-types related to each other? To answer this, however, we also need to answer question 1: How do we find and identify the cell-types?

¹These authors contributed equally.

In the following introduction I will introduce the datasets used in the chapter and cover how the two above questions are normally approached in scRNAseq studies.

2.1 Introduction

2.1.1 Datasets

Throughout this chapter, I am going to refer to three datasets: The scRNAseq dataset we published along with the paper [Rothová et al., 2022], the Modified National Institute of Standards and Technology (MNIST) digits dataset [LeCun et al., 1998] and fashion-MNIST [Xiao et al., 2017].

The exact format and content of the Rothova2022 dataset is covered in the paper, but I will quickly review the essential points. The dataset consists of measurements of the RNA sequences inside 6282 individual single cells taken from mouse embryos in the early stages of development (between days 6.5 - 9.5 after fertilization) and 4003 cells from *in vitro* experiments. The RNA of individual cells were sequenced using the MARS-seq protocol [Jaitin et al., 2014]. The resulting data characterize each cell with a vector, where each entry denotes the number of RNA sequences expressed corresponding to a specific and unique gene.

After collecting all the vectors representing the individual cells, the overall result can be represented as an M by N matrix, where M is the number of cells (10285), and N is the number of genes (24262). The rows in this matrix represent RNA levels for a given cell, and columns represent the expression of a given RNA across all cells. This matrix is commonly referred to as a count matrix and represents the whole transcriptome profile of the cell population. In conventional data-science terms, each cell is a sample, and each gene is a feature. The count matrix can therefore be interpreted as a standard "sample x feature" matrix common in the field of machine learning and scientific computing. An illustration of the matrix format can be seen in Figure 2.1.

ScRNAseq datasets are by nature high-dimensional, noisy and often contain various data defects, making it hard to analyze, interpret, and intuit about the data [Eraslan et al., 2019, Xu et al., 2021].

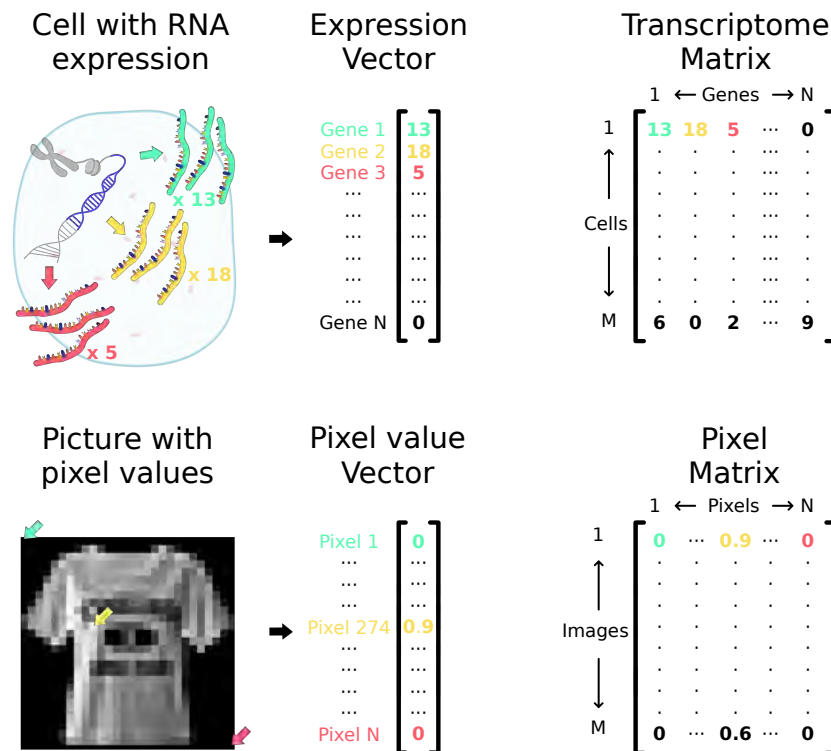


Figure 2.1: Data layout for scRNAseq datasets and fashion-MNIST. Using scRNAseq, the strands of RNA inside a cell can be measured. Adding up the counts corresponding to the various genes results in a vector summarizing the gene expression for the cell. Combining many such expression vectors yields a count matrix. The fashion-MNIST dataset is constructed similarly, but is composed of pictures instead of cells. Each image is vectorized by simply reading the pixel intensities. Combining many vectorized images yields a matrix representing the dataset of pictures. (DNA and RNA clip-art edited from IGI Glossary Icon Collection by Christine Liu).

In our article, [Rothová et al., 2022], CAT is only run on scRNAseq datasets. So in this chapter, in order to evaluate and illustrate the functionality of CAT, I will additionally make use of the famous MNIST digits and fashion-MNIST datasets (Figure 2.1), that are commonly used for benchmarking machine learning algorithms. In the MNIST digits dataset, each sample is a digitized 28 by 28-pixel photo of a handwritten digit, ranging from 0 to 9. The features are the intensity of each pixel. The MNIST digits dataset has the convenient property that each of its 70.000 samples conforms neatly into archetypes of 0,1,2... etc. (unlike scRNAseq), making it easy to judge the performance of algorithms run on it. The fashion-MNIST dataset is structured like the MNIST digit dataset but consists of more complex photos of fashion items, such as shoes and t-shirts, instead of digits. MNIST digits is one of the more simple datasets a machine-learning algorithm can be benchmarked against, and if an algorithm fails here, it can hardly be expected to

work for more complicated data. When MNIST digits is "too easy", fashion-MNIST is often used instead [Song et al., 2017, Xiao et al., 2017].

2.1.2 Standard processing of scRNAseq data

The raw sequencing data produced by scRNAseq protocols are not immediately ready for analysis, it needs to be preprocessed before becoming a usable count matrix (like the Rothova2022 dataset described above). The steps involved in a preprocessing pipeline affect all subsequent analyses, like finding and comparing cell-types. A typical scRNAseq processing/analysis pipeline is illustrated in Figure 2.2.

The raw data contains the sequences for each piece of RNA that was detected inside each cells of a sample. The sequencing reads are matched against a reference genome to determine which gene each sequence belongs to. The number of RNA sequences corresponding to each gene is counted for each cell. The steps involved in this quantification process are called trimming and alignment, and results in a raw (un-normalized) count matrix [Du et al., 2020].

The sequencing process is sensitive to the concentration of reagents, which vary slightly from cell to cell and sequencing batch to batch. For example, in the PCR amplification step of a standard scRNAseq protocol, this can lead to big differences in the total number of RNA reads between cells [Jaitin et al., 2014]. Due to this type of technical noise, the raw count matrix from the quantification step needs to be normalized [Brennecke et al., 2013]. The simplest normalization is normalizing each gene's count by the total RNA count for each cell, but many other methods exist [Cole et al., 2019]. Technical batch-to-batch variation can be normalized using specialized batch correction methods [Haghverdi et al., 2018]. Normalization has been extensively studied and is the preprocessing step that influences the subsequent analysis the most [Vieth et al., 2019].

Quality control is performed between each preprocessing step. This ensures the removal of cells that failed to be properly sequenced, e.g. cells with too few detected genes or a too low total number of reads. The end result of the preprocessing pipeline is a normalized count matrix with associated metadata (see section 2.1.1), which can then be used for the downstream analysis.

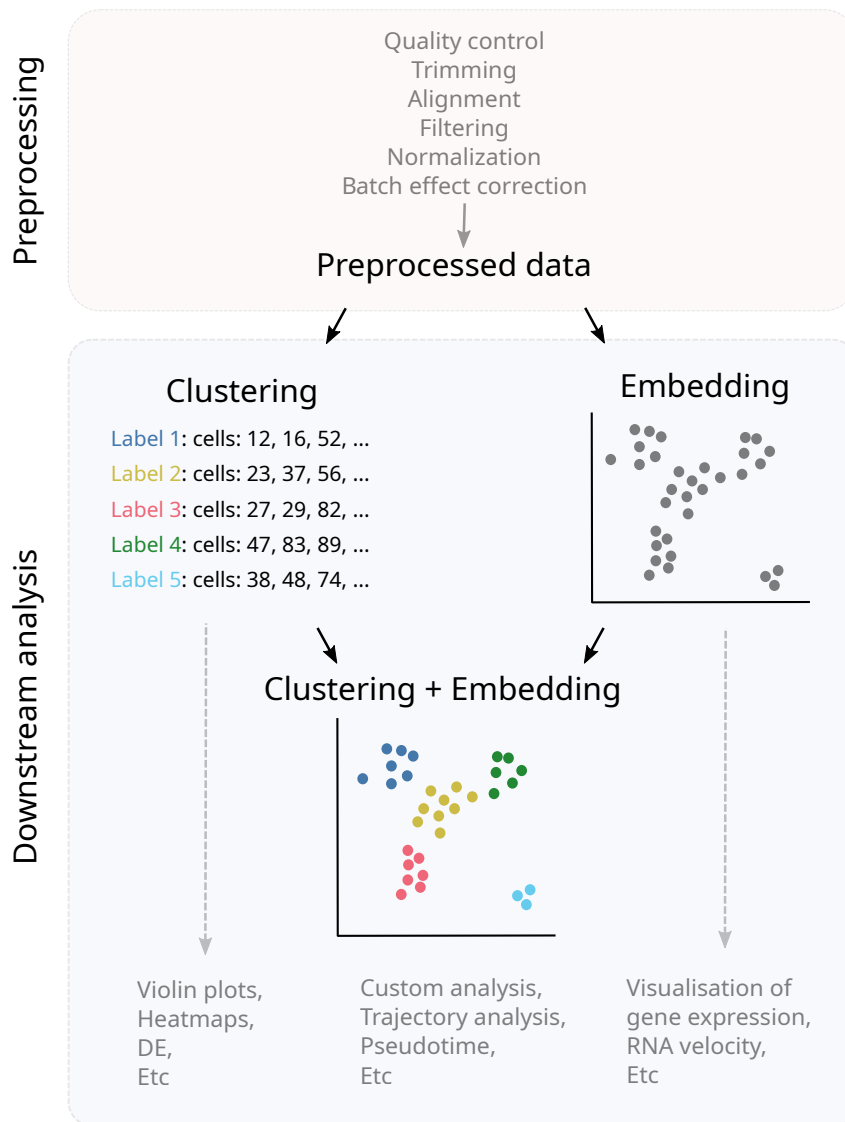


Figure 2.2: Illustration of a typical scRNAseq data-processing workflow. The raw data first undergoes preprocessing to get a normalized count-matrix with relevant metadata (preprocessed data). This count-matrix is then typically clustered and embedded. These steps are usually iterated a few times before converging on a set of good parameters for the preprocessing, clustering and embedding. Afterwards, there will be further downstream analysis.

Usually, the first task for downstream analysis is exploration and examination of the dataset's content, i.e. identifying cell types, figuring out their function and how they relate to each other and contribute to the various organs of the studied tissue.

2.1.2.1 Clustering

To address how the cell-types in a dataset are related, we first need to find out which cell-types there even are. The detection of cell-types (also called lineages) is typically done using clustering to label similar data points together, using some measure of "distance" for similarity [Kim et al., 2019]. Clustering is sensible for scRNAseq data since we expect cells of the same cell-type to have relatively similar RNA expression profiles compared to cells of a different cell-type. Cells of the same lineage should therefore have a small distance between their expression vectors and should cluster together. Once the cells have been clustered, the clusters need to be examined so we can identify the lineages they represent. This is typically done using known marker genes or gene ontology enrichment analysis on genes that are differentially expressed between the clusters (more on identification in section 3.2.1).

In principle, clustering could be done in a supervised manner, manually assigning labels to each individual cell in a dataset (assuming the researcher has enough prior knowledge and patience). In practise, due to the overgrowing datasets in the field, researchers turn to unsupervised clustering algorithms such as Louvain [Blondel et al., 2008], SNN-Cliq [Xu and Su, 2015], PhenoGraph [Levine et al., 2015], SC3 [Kiselev et al., 2017] or Leiden [Traag et al., 2019].

While the exact methods these algorithms use to determine clusters differ, they all have in common that they use some distance metric, e.g. euclidean or cosine distance, to measure similarity between cells for this decision. Consequently, they are all prone to the phenomenon dubbed "the curse of dimensionality" [Bellman et al., 1957]. The curse of dimensionality is technically an umbrella term that covers various phenomena that makes it difficult to analyze high-dimensional real-world data, but I will just cover the following aspect of the curse that is relevant to this topic: When the dimensionality of a dataset's feature space grows, the distances between all pairs of data points tend to converge, rendering any measure of distance meaningless [Ertöz et al., 2003]. This is naturally a problem for algorithms relying on distances, like clustering. The more noise a real-world dataset has, the worse the problem becomes. ScRNAseq data is notoriously noisy [Wagner et al., 2018, Eraslan et al., 2019].

A toy example illustrating the curse of dimensionality is shown in Figure 2.3. The figure contains two rows of plots; the first row shows points drawn from

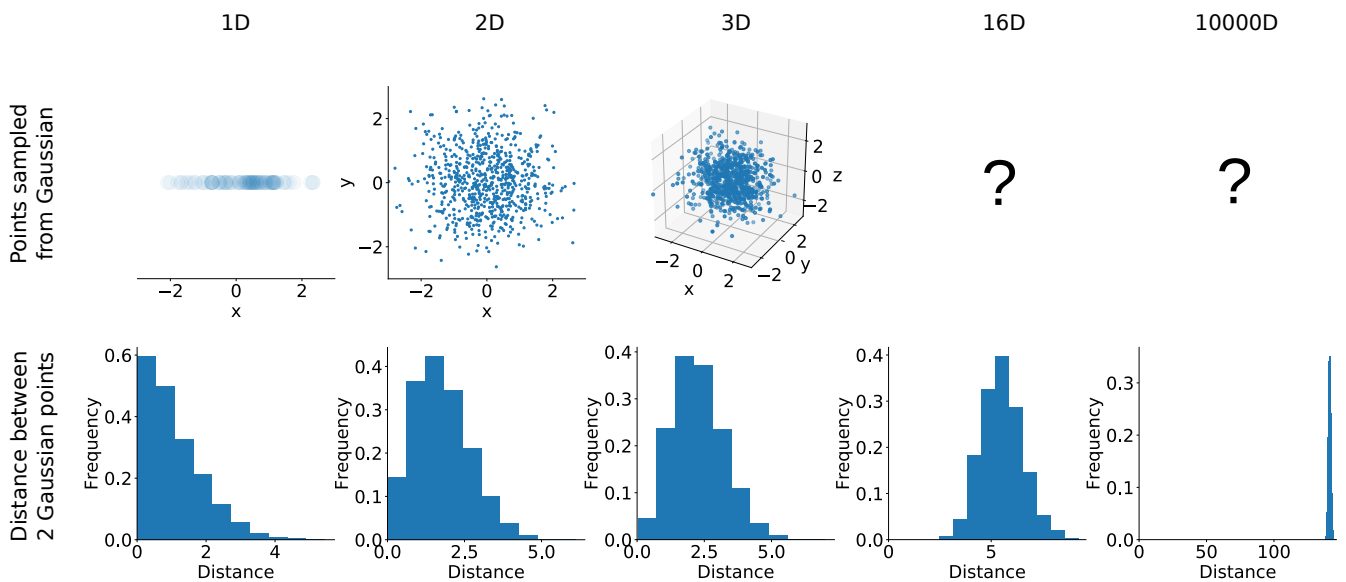


Figure 2.3: Figure illustrating the behaviour of noise in various dimensions. The top row shows points sampled from Gaussian distributions with 0 mean and unit variance. The bottom row shows the pairwise distance between points of the top row. As the number of dimensions grows, the pairwise distance between points grows as well, making it harder and harder to use the distance to distinguish between pairs of points that were sampled closely together and points that are sampled further apart. In the 10.000D case, the distance between all pairs of points are basically the same.

Gaussian distributions (with zero mean and unit variance) for different numbers of dimensions, and the second row shows the pairwise distance between these points. In the 1-dimensional case, the chance of sampling 2 points from the centre of the normal distribution is quite high, so the pairwise distance between 2 random points is often very close to 0. Occasionally, as we would expect, a point is sampled from further out in the Gaussian’s tail, which is reflected in a few high pairwise distances. Based on the distance in the 1D case, it is easy to distinguish between two points that were sampled closely together and two points they were sampled from further apart. Once the dimension increases to 2D, the expected pairwise distance shift to a higher value. The points are still most likely to be sampled close to the centre of the distribution, but the added degree of freedom means the points can be offset at an angle to each other and, as a result, will be slightly further apart on average. In the 2D case, it is still easy to distinguish between points that are close together and points that are further apart based on the pairwise distances. When the number of dimensions increases, the pairwise distances keep shifting higher. In the 10.000D case, the vastness of the high dimensional space further inflates the distances between the pairwise points and the degrees of freedom all but guarantee that the sampled points will be orthogonal to one another. As

a result, the difference between the highest possible pairwise distance and the smallest are virtually the same, meaning that all the points are basically equidistant. The examples shown in Figure 2.3, illustrate that noise in high dimensions always pushes points apart, never together. If the variance of the Gaussians had been 0, all the points would be located at origo and the pairwise distance between points would be a delta function at zero (regardless of the number of dimensions). The fact that distances increase with a larger number of dimensions can therefore be attributed solely to the noise. These are idealized examples to showcase the effect of the curse of dimensionality, but the effect of noise in real-world data is the same. Put in equation form:

$$\lim_{N_{dim} \rightarrow \infty} \frac{d_{max} - d_{min}}{d_{min}} = 0 \quad (2.1)$$

where N_{dim} is the dimensionality of the data and d_{max} and d_{min} denoted the largest and smallest possible pairwise distances between points in the given dataset. As the dimensionality of the data tends towards infinity, any contrast between distances vanishes, leaving every point equidistant not only to origo but also to each other. Fortunately, a typical scRNAseq dataset spans only 10.000 to 30.000 genes and current state-of-the-art clustering methods, like the ones mentioned above, have proven capable of successfully finding meaningful clusters, given proper tuning of clustering parameters. But, when dealing with 10.000 to 30.000 dimensions, the effect of the curse of dimensionality is a genuine concern, and it is crucial to verify the clustering's results. Having a lot of gene expression information for single cells is both a blessing and a curse. More genes in a dataset should in principle enable us to resolve more rare and specific cell-types that might only differ slightly in expression of only a few genes. On the other hand, having more genes makes it harder to do proper unsupervised clustering, embeddings and many other types of analysis that directly or indirectly rely on distances between the cells in the gene-expression space.

2.1.2.2 Dimensional embedding / reduction.

To visualize how the identified cell-lineages (clusters) relate to one another, the clustering is typically done alongside a dimensional embedding. Before explaining what an embedding is, let us take a step back.

To understand and convey data, we create graphical representations of it. This could be a pie-chart of the composition of cell-types in a dataset (1 variable) [Wang et al., 2019], a bar-plot of a gene's expression across experiment replicates (2 variables) [Zhao et al., 2021] or a scatter-plot of the sepal width versus the petal length, color-coded for multiple types of iris flowers (3 variables) [Fisher, 1936]. The purpose of the illustrations in these examples is to show how datapoints relate to each other in terms of a varying number of variables, i.e. which datapoints are similar/dissimilar and to reveal trends. How to best plot different kinds of data has been studied at length, and it is generally agreed that only a handful of variables can be plotted in one figure at a time [Bertin, 1983, Wilkinson, 2012]. This leaves high dimensional data, like scRNAseq data, in a difficult spot. To convey and intuitively understand the full picture of the relationships between high dimensional datapoints (such as cells in scRNAseq data), we cannot plot all the different combinations of 3 genes (variables) at a time. To overcome this difficulty researchers, use dimensional embedding techniques [Han et al., 2022, Kadur Lakshminarasimha Murthy et al., 2022, Bandler et al., 2022]. Dimensional embedding enables the otherwise unintuitive and high-dimensional scRNAseq datapoints (cells) to be plotted as a scatter-plot in just 2 (or sometimes 3) dimensions. The distances between the cells in the scatter-plot are computed so they capture the relationship/similarity across all the genes (variables) at once. This low-dimensional representation is called the "embedding" of the data. The goal of the dimensional embedding, when used to explore scRNAseq datasets, is twofold:

1. Check that the clusters are "reasonable" in the sense that they actually appear to be separate clusters that do not intermix.
2. Compare the relationship of clusters in the embedding, revealing which cells and cell-types are similar and which are not (Figure 2.2).

While both goals here can serve as a validity check for the clustering, an embedding is often done in its own right to gain novel insight into the structure of the dataset. The embedding can reveal possible trends between types and be used for plotting other variables on top of (e.g. color-coding the scatter points to specific genes) [Yao et al., 2021, Konstantinides et al., 2022, Melenhorst et al., 2022].

There is a multitude of dimensional embedding techniques, each with a different method for calculating an appropriate distance between data-points (cells) in the scatter-plot. Broadly speaking, the techniques can be put into two categories:

Linear and non-linear² [Sumithra and Surendran, 2015, Udell et al., 2016, Wang et al., 2020].

Linear

The linear class of techniques work by using low-rank matrix approximation: An input matrix is factorized to be approximately expressed as the product of 2 smaller matrices, as illustrated in Figure 2.4a. The two smaller matrices can be thought of as a representation matrix and an archetype matrix. A row in the representation matrix contains the weights of a single sample for a linear combination of signatures (rows in the archetype matrix) needed to approximately reconstruct a row of the original data. In the domain of scRNAseq, a row of the original data would represent a cell. Linear methods include examples like principal component analysis (PCA), single value decomposition (SVD), multidimensional scaling (MDS) and non-negative matrix factorization (NMF). The signatures in the archetype matrix are the directions in the case of PCA, and the representation rows are the placement of a datapoint along these directions. An example of PCA is shown in Figure 2.4c. The way these linear methods accomplish a dimensionality reduction, down to 2 dimensions for example, is by using the first 2 entries in the rows of the representation matrix as the coordinates for the cells in the embedding. In the case of PCA, this corresponds to plotting the two leading PCA components and ignoring the contribution of the rest of the components (the ones that account for the least variance in the dataset). The main difference between these linear methods comes down to how they weigh the minimization of the difference between the original data and the approximation, as well as the constraints they put on the problem [Udell et al., 2016]. Linear dimensionality reduction techniques are blazing fast and are successfully applied across many fields, but as the name would also suggest, they do not capture non-linear relationships well (like the ones typically present in biological data) [Van Der Maaten et al., 2009]. For this reason, non-linear DR have overtaken methods like PCA and are now the undisputed standard for visual exploration in the field of bio-informatics [Luecken and Theis, 2019].

Non-linear

Non-linear dimensionality reduction (NLDR), also sometimes called manifold learning, is a slightly broader category of algorithms. The idea behind NLDR is that high-dimensional data is usually not distributed randomly throughout all

²Technically, there are also autoencoders and other types of neural network approaches, but these are not common in the scRNAseq field (or data science in general) so I will not cover these.

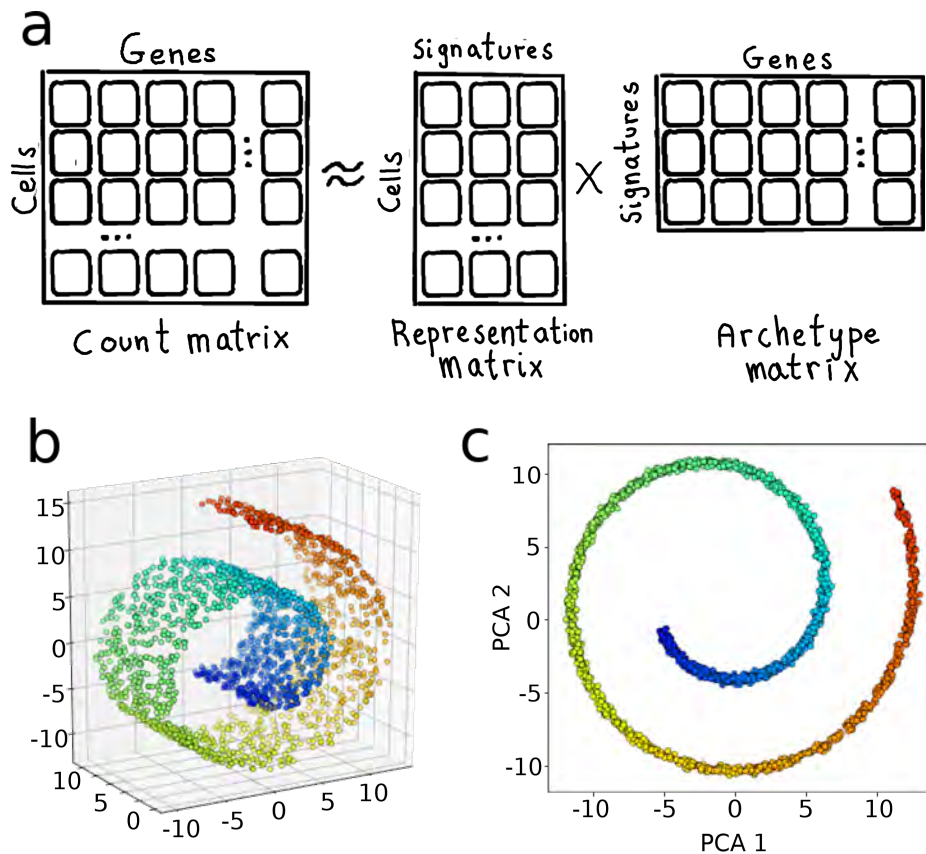


Figure 2.4: Linear dimensionality reduction example. a) Using matrix factorization a count matrix can be approximated by 2 smaller matrices. A dimensionality reduction can be achieved by dropping the least significant signatures, e.g. the signatures accounting for the least variance in the case of PCA. b) High dimensional data (3d) lying on a lower dimensional (2d) "swiss roll" manifold. c) 2d dimensional embedding of b) using the 2 leading PCA components (i.e. 2 columns of the representation matrix). PCA fails to capture the manifold nature of the data and instead simply removes the depth of the roll, which contains the least relative variance. The swiss-roll data was generated using scikit-learn [Pedregosa et al., 2011b].

its dimensions but tends to lie close to a lower-dimensional manifold within the feature (gene) space. A visual example can be seen in Figure 2.5a, illustrating a 3d dataset consisting of data-points lying on a 2d manifold in the shape of the classic "swiss roll" from [Tenenbaum et al., 2000]. The structure of the lower-dimensional manifold can be captured in the form of a weighted neighbour graph, which in turn can be used to embed data-points into a lower dimension in such a way that the features of the graph are preserved, even if the exact relative distances between all points cannot. Older methods like isomap [Tenenbaum et al., 2000] and Laplacian Eigenmaps (spectral embedding) [Belkin and Niyogi, 2001] uses the neighbour graph to construct a geodesic distance matrix which is then processed using MDS or other matrix factorization methods to create an embedding (exactly like the linear methods) [Whiteley et al., 2021]. Newer and more popular methods, like T-SNE, UMAP and PaCMAP, forgo this step and instead embed the neighbour graph itself directly into the lower dimension using force-directed layout algorithms: The datapoints (nodes of the neighbour graph) are placed in the desired lower dimension, sometimes at random, and the edges and nodes of the network are assigned attractive and repulsive forces. The NLDR algorithm then simulates the forces of the network, relaxing it iteratively into a shape that (ideally) recaptures the weights of the neighbour graph from the higher dimension and conserves thereby conserves the manifold structure. This process is illustrated in Figure 2.5 using UMAP on the swiss roll dataset. It is worth noting that the axis units on the final embedding are arbitrary, which is the reason NLDR embeddings are often displayed without their axis altogether, see e.g. [Cao et al., 2019]. The exact method for constructing the graph and the weights of the forces depend on the algorithm and will often vary slightly upon implementation [Wang et al., 2020].

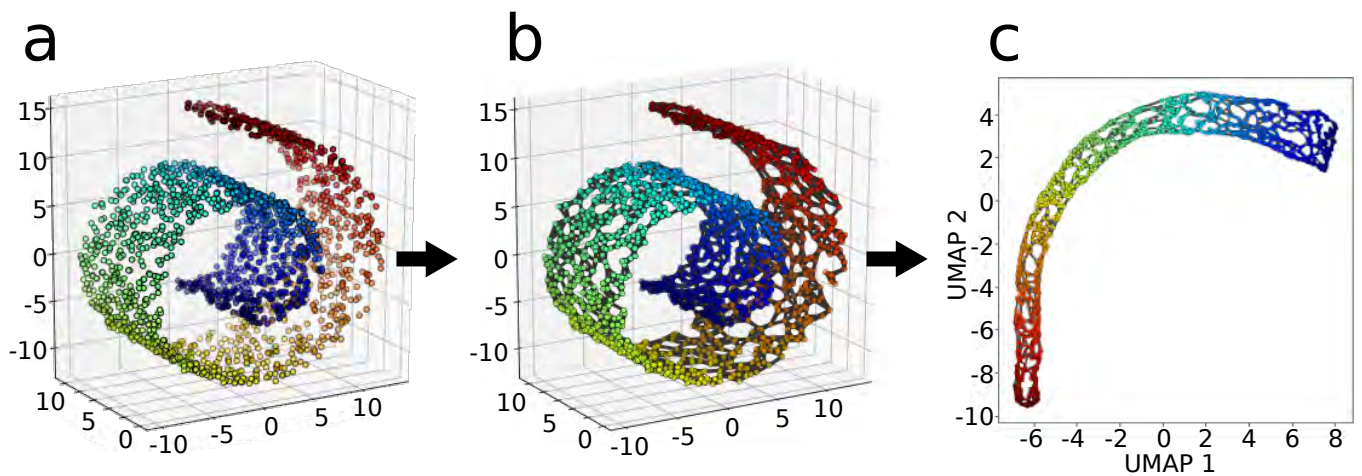


Figure 2.5: Non-linear dimensionality reduction example. a) High dimensional data lying close to a lower dimensional manifold. The example shows the 3d "swiss roll" dataset. b) To capture the structure of the lower-dimensional manifold, points are connected into a nearest-neighbour graph. The graph is then embedded in a lower dimension, for example, using a force-directed graph layout. c) Once the graph has been embedded, the dimensionality reduction is complete. This example shows a UMAP embedding in 2d plotted together with the neighbour graph. UMAP captures the manifold and "unrolls" the data along it. Axis units are arbitrary.

2.1.2.3 Challenges of using non-linear dimensionality reduction

Modern non-linear DR methods have proven capable of producing impressive visual results on many types of real-world datasets, ranging from high-throughput 'omics' technologies [Argelaguet et al., 2019] to climate modelling [Franch et al., 2020], astronomy [Jespersen et al., 2020], cybersecurity research [Bozkir et al., 2021] and computer vision [Väisänen et al., 2021]. An example of how embeddings are used in practice in the scRNAseq field could be; accessing similarities between cell-types after dataset integration. When datasets co-localise in an embedding, e.g. T-SNE or UMAP, they are judged to be similar, thereby proving success data integration [Butler et al., 2018, Chen et al., 2022]. Researchers also use the continuity of data-points in an embedding to infer relationship between cell-types (E.g. the clusters corresponding to two cell-types mix with each other at the border between them) [Krivanek et al., 2020].

However, despite their well-deserved and widespread use, NLDR methods have some challenges that limit their usefulness for interpretation of the data. Particularly for the types of scRNAseq-related questions outlined at the beginning of the "Dimensional embedding/reduction" section: verifying the clustering of cells

and finding relationship between cell-types. In practice, whether or not NLDR produce helpful embeddings, depends strongly on the quality of the data and the parameters used. Below I will discuss a few of the problems with modern NLDR algorithms, such as T-SNE and UMAP, and afterwards introduce my own algorithm, CAT, to address these issues.

Non-linear DRs are sensitive to parameter tuning

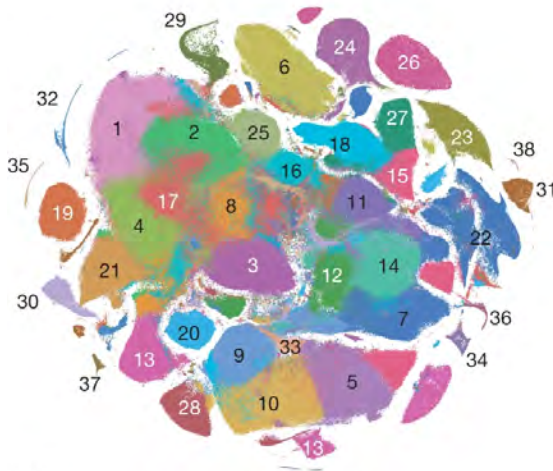
Firstly; non-linear DRs are known to be sensitive to parameter tuning [Wattenberg et al., 2016, Coenen and Pearce, 2019, Wang et al., 2020, Huang et al., 2022].

For example, algorithms like T-SNE and UMAP have a parameter that balances the algorithm's emphasis on local versus global structure in the data, usually called perplexity or simply the number of nearest neighbours. The parameter determines the number of neighbours that the algorithm will try to strictly preserve the distances between. A suitable value for this parameter depends on the number of datapoints, so the same value cannot be assumed to give similar results across different datasets or even for the same dataset if more samples are added to it later. Fortunately, tuning the parameter is easy, making it possible to quickly find reasonable-looking embeddings in most cases. Effectively perplexity controls the "tightness" or "clumpy-ness" of datapoints in the embedding, meaning higher values will group the points together in fewer larger clumps. Tuning perplexity up or down can therefore mean the difference between having a single spatial group of cells split in two or having separate groups adjacent or entirely apart in the embedding. These differences could lead to different interpretations of the data, so in order to get an intuition about the underlying data, it is necessary to run the non-linear DR algorithm with multiple perplexity values.

An example of just how different embeddings of the same data can look with different embedding parameters is provided in Figure 2.6.

Other parameters also profoundly influence the result of embedding algorithms. The exact parameters and their names depend on the respective algorithm [McInnes et al., 2018, Poličar et al., 2019, Wang et al., 2020]. To give a few examples using T-SNE terminology: The degree of freedom in the kernel shape for finding nearest neighbours, early exaggeration (negative sampling rate in UMAP) and the learning step size used in the descent algorithm. These parameters influence the way

a Cao et al.



Cell types (Both plots)

- 1-Connective tissue progenitors
- 2-Chondrocytes and osteoblasts
- 3-Intermediate mesoderm
- 4-Jaw and tooth progenitors
- 5-Excitatory neurons
- 6-Epithelial cells
- 7-Radial glia
- 8-Early mesenchyme
- 9-Neural progenitor cells
- 10-Postmitotic premature neurons
- 11-Oligodendrocyte progenitors
- 12-Isthmic organizer cells
- 13-Myocytes
- 14-Dorsal neural tube cells
- 15-Inhibitory neurons
- 16-Stromal cells
- 17-Osteoblasts
- 18-Inhibitory neuron progenitors
- 19-Premature oligodendrocytes
- 20-Endothelial cells
- 21-Chondrocyte progenitors
- 22-Definitive erythrocyte lineage
- 23-Schwann cell precursors
- 24-Sensory neurons
- 25-Limb mesenchyme
- 26-Primitive erythroid lineage
- 27-Inhibitory interneurons
- 28-Granule neurons
- 29-Hepatocytes
- 30-Notochord and floor plate cells
- 31-White blood cells
- 32-Ependymal cells
- 33-Cholinergic neurons
- 34-Cardiac muscle lineage
- 35-Megakaryocytes
- 36-Melanocytes
- 37-Lens
- 38-Neutrophils

b Kobek et al.



Figure 2.6: The same data is embedded in a) and b) by different authors using T-SNE with different parameters. Despite portraying the same data, the two embeddings look starkly different. a) Shows a T-SNE embedding of a mammalian organogenesis scRNAseq dataset. Dataset and plot produced by Cao et al. [Cao et al., 2019]. Each of the 2.026.641 tiny colored dots corresponds to a sequenced mouse cell. The authors have colored each cell according to cell-type (identified from clusters found using unsupervised Louvain clustering). b) The same dataset embedded using T-SNE by Kobek et al. [Kobak and Berens, 2019], but with different T-SNE parameters. colors and labels are the same between a) and b). Besides the obvious global structure difference, notice how for example, cluster 15 appears split up in a) but not in b). If a researcher uses co-localisation or mixing between cell-types to guide their interpretation of the data, they will reach different conclusions, depending on which of the 2 embeddings they are looking at.

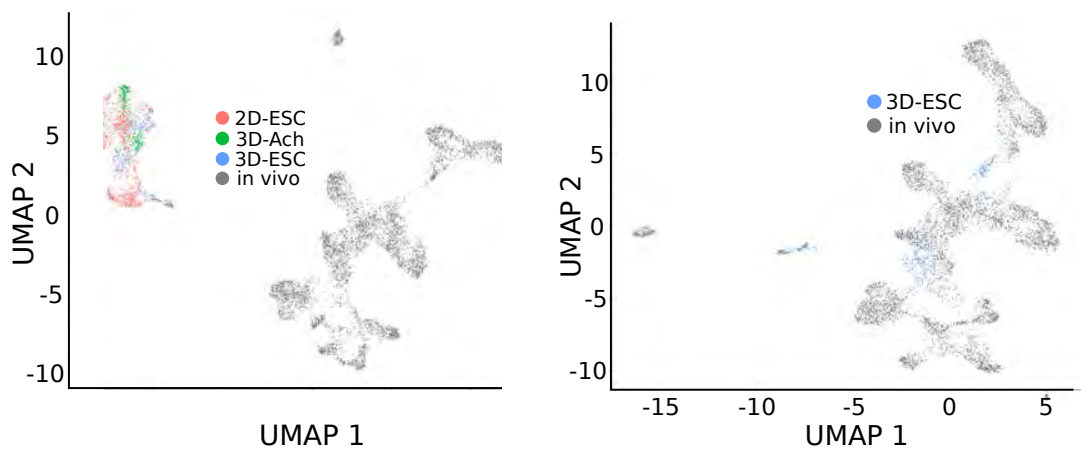
the manifold is reconstructed [Kobak et al., 2019], the likelihood of the layout getting stuck in a poor local minimum [Linderman and Steinerberger, 2019], the distance between clusters [Kobak and Berens, 2019] and how discrete/continuous the borders between clusters get presented [Böhm et al., 2020]. Similar to the case of perplexity, good values for these parameters depend on the question the researcher is trying to answer, as well as the intricacies of the dataset. To get a full understanding of a dataset, multiple embeddings should ideally be created varying these parameters as well. All in all, optimizing the parameters of non-linear DR is a decidedly non-trivial task, so much so that software has been written to automate the task [Belkina et al., 2019].

Results of NLDR are sensitive to heterogeneity of the dataset

Unlike most statistics tools, introducing more data-points does not necessarily make the embedding more reliable or easier to analyse. By adding more data-points to a dataset, NLDR can make the already present data-points squeeze closer together or fall in completely different regions on the embedding, giving the illusion of lower/higher similarity. An example of this is shown in Figure 2.7. The figure shows that after adding new clusters to a dataset and running the typical analysis pipeline steps required for an embedding (Figure 2.2), the distances between previous clusters in the embedding will change. Some researchers acknowledge that the distances between disconnected clusters do not carry any significance [Coenen and Pearce, 2019, Wattenberg et al., 2016], but this example shows that even the intermixing of clusters can be quite arbitrary. This problem is particularly relevant when researchers use co-localisation on an embedding to support conclusions of similarity between different cells, like in the cases of data integration mentioned above. The effect can be partially avoided if the embedding is run on the full gene-set instead of highly variable genes (HVGs) or PCA. Feature selection, however, is often used due to the performance scaling issues of NLDR and the ever-growing dataset sizes [UMAP, 2022, Wang et al., 2020] (see section "How to deal with the curse of dimensionality for manifold learning" below).

Stochasticity

Stochasticity itself is by no means a negative feature of an algorithm; randomness helps optimization algorithms find useful minima, speed up calculations and



(a) *In vivo* together with many *in vitro* cell (b) *In vivo* together with few *in vitro* cells

Figure 2.7: UMAP embedding using embryonic stem cells of *in vivo* (grey) and *in vitro* origin (colored) from Rothova2022. The embeddings on a) and b) portray the exact same *in vivo* cells but a different number of *in vitro* cells (3 *in vitro* clusters in a) versus 1 in b)). This difference influences whether or not the *in vivo* and *in vitro* mix together in the embedding (notice the blue population). The names of the *in vitro* populations are not important for this figure but are explained in section 3.2.2 of the next chapter, where these populations are more thoroughly introduced.

help discover solutions that would otherwise be hard or impossible to derive analytically. A modern example of the success of stochastic algorithms is the optimizers behind neural networks [Kingma and Ba, 2014, Loshchilov and Hutter, 2017]. All modern NLDR algorithms currently applied in the field of bioinformatics are also stochastic in nature, and can in fact be thought of as unsupervised machine learning³. NLDR "learns" a high-dimensional manifold that fits the data, which it then tries to project lower-dimensional data onto [Wang et al., 2020]. Like other machine learning algorithms, the network layout part of NLDR methods works like a minimizer, using some form of stochastic gradient descent to arrive at their objective function's minima. Using this approach, they are able to non-linearly embed high-dimensional data in a way that can take both large-scale and local features into account, something linear embedding methods inherently cannot. This feature comes at a cost; unlike linear methods, non-linear DR is not guaranteed to arrive at the global minima for the objective function, meaning that every time the algorithm is rerun, even with the same data and parameters, it produces a different embedding. The resulting embeddings can differ significantly, and with it, the possible interpretations we draw about the data. When commenting on

³Typically unsupervised. UMAP, for example, has been updated to support supervised learning using labelled data and neural networks under the hood.

the reproducibility in UMAP's official documentation, Vito Zanotelli puts it quite eloquently:

[...] setting a random seed is like signing a waiver "I am aware that this is a stochastic algorithm and I have done sufficient tests to confirm that my main conclusions are not affected by this randomness".

Since non-linear DR algorithms rely on stochastic iterations to converge to their result, they are not only affected by the random number generator but also by their initial configuration. Recent papers have investigated the choice of initialisation and shown that it can have a substantial impact on the ability of these algorithms to preserve global structure [Kobak and Berens, 2019, Kobak and Linderman, 2019, Wang et al., 2020, Kobak and Linderman, 2021].

To illustrate how non-linear DR can fail to produce consistent results, I have run UMAP 6 times with random initialisation using different random seeds, but otherwise identical default parameters. The test was run using the simple MNIST digits benchmark dataset (see section 2.1.1). Four of these UMAP embeddings are shown in Figure 2.8. While the four embeddings look comparable at first glance, upon further inspection, it can be seen how both subplot 2.8a and 2.8d, contain more than the expected 10 groups of datapoints (one for each digit). This could indicate that some of the sampled digits are not as consistently similar as we would expect (maybe people draw the same digit in two different ways?), but this conclusion is contradicted by the embeddings on subfigures 2.8b and 2.8c, that correctly show 10 groups on the embeddings. More likely than not, the embeddings got stuck in sub-optimal minima for subplot 2.8a and 2.8d, but this would have been impossible to guess without re-running the embedding multiple times or having a ground truth (in this case we know that there should be 10 types of digits). The embedding on subplot 2.8c, that correctly shows 10 groups, also shows how the groups corresponding to digits 9 and 7 are neighbours and slightly intermix at their borders, indicating a possible similarity between the shapes of these hand-drawn digits. With prior knowledge of how the digits look like, this similarity will probably make sense to most people. The "closeness" between 7 and 9 is not present in subplot 2.8a and 2.8b, and since there is no quantitative measure to judge with, it is hard to definitely say how related two digits are compared to other digits. The MNIST digits dataset is supposed to be one of the easier datasets to benchmark [Wang and Deng, 2022, Jiang, 2020], and while UMAP successfully groups most of the digits together according to their type, it does not

do so consistently. For datasets with less well-defined archetypes, more noise, and for which we have less prior knowledge, UMAP and other NLDL algorithms' lack of consistency should be a real concern.

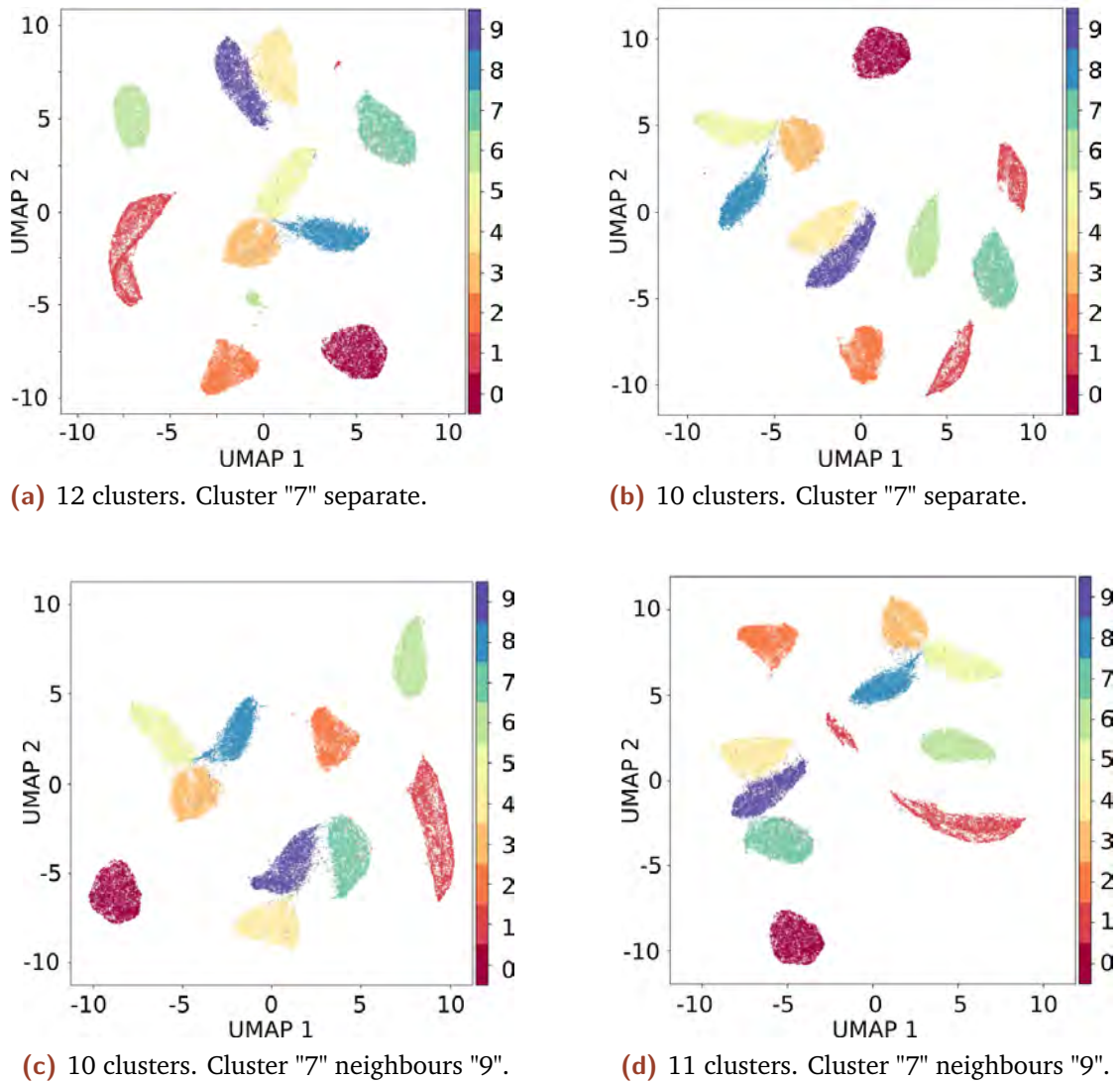


Figure 2.8: UMAP embeddings of the famous MNIST digits dataset. The 70.000 small dots on each embedding corresponds to an image of a handwritten digit between 0 and 9 and are color-coded according to the digit it represents. Ideally, the embeddings should consistently group datapoints corresponding to the same digits together, but as seen in subplots (a) and (d), this is not always the case since these have more than 10 spatial clusters. Subplots (b) and (c) correctly shows 10 spatial clusters but disagree on the neighbourhood of cluster "7". Embeddings were generated using default UMAP parameters (official python implementation, random initialization) for 6 random seeds, with 4 being cherry-picked to show here.

Curse of dimensionality and manifold learning

Because the goal of dimensionality reduction is to embed the relative distances between datapoints in high dimensions into a lower one, it is inherently sensitive to the curse of dimensionality. If the data is noisy and high dimensional enough, it will lose its contrast between similar and dissimilar points, the points becoming equidistant (Figure 2.3). This means that there will be no structure or manifold in the high dimension to embed in the first place. In practice, as mentioned earlier, datasets are rarely so high-dimensional or noisy as to be useless, but the higher dimensional the data is, the harder it will also be to do insightful dimensionality reduction on it [Van Der Maaten et al., 2009]. An often used method to limit the effect of the curse of dimensionality is to simply limit the number of dimensions by selecting a subset of features prior to doing dimensional embedding. The most common approaches include limiting the analysis to the most highly-variable genes (HVGs), i.e. the genes that vary most between the cells, or to perform PCA on the data and only use a limited set of the top PCA components [Johnstone and Lu, 2009, Wolf et al., 2019, Hao et al., 2021].

The logic of limiting the features to HVGs relies on the argument that the genes which do not vary much between cells also will not enable us to distinguish between samples and therefore effectively only contribute with more dimensions of potential noise [Luecken and Theis, 2019]. On the other hand, the genes that vary most between the cells are likely also the ones that are best at distinguishing cell-types. It is the same argument that can be made for PCA, simply using PCA directions in the gene space instead of genes. Feature selection is so common, that some researchers even recommend running PCA on top of HVGs before downstream analysis like NLDR [Weber, 2022, Amezquita et al., 2020]. Using a sub-set of features has the added benefit of significantly improving the runtime and memory usage of NLDR, since these tend to scale poorly with an increased number of features [McInnes et al., 2018, Wang et al., 2020].

The choice of pre-reduction of dimensions becomes clear a trade-off: better performance for the down-stream analysis (e.g. clustering and dimensionality reduction) versus the loss of signal that comes with throwing data away. An additional downside of using feature selection is that the results of the downstream analysis will depend on the details of input data in unpredictable ways. Imagine two scRNAseq datasets, dataset A and dataset B, that are entirely identical, with the exception

that A contains an additional population of cells that B do not. These additional cells may have significantly different genes expression compared to the average of the dataset, which means that the HVGS and PCA computed on dataset A will be different from dataset B. When downstream analysis of the datasets is concluded, the portion of cells that were otherwise identical between the two datasets will cluster differently and show different embedding relationships. This effect is similar to the one shown in Figure 2.7, but extends to all downstream analyses that depend on the explicitly data-derived feature selection.

Implications of the DR limitations

Linear dimensionality reduction struggle to capture the non-linear relationships between cells in scRNAseq data, which makes non-linear dimensionality reduction the natural alternative and de-facto standard [Liu et al., 2016, Verma and Engelhardt, 2020, Wu et al., 2021].

However, given the challenges with current state-of-the-art NLDR embedding methods, it can be risky to judge or draw conclusions from individual embedding. In my own research, I find that I often create multiple embeddings, changing parameters and rerunning the same parameters over and over (with different seeds) to gain an intuition and to find a plot I feel best illustrates the data. It is a tedious process, and even with multiple embeddings of the same data, it can be difficult to pick one since, in principle, each embedding is equally "correct" when there is no qualitative measurement to judge them by. In practice, the vast majority of publications include only a single embedding.

With the replication crisis in science, [Nissen et al., 2016], I think that it is fair to point out that without a ground truth to hold NLDR against, "finding the embedding that best illustrates the data" can be awfully close to "finding the embedding that best supports the desired conclusion".

In "*Eleven grand challenges in single-cell data science*", a widely cited review collaborated by many of the leading researchers in single-cell data science, the authors list 3 overarching themes among the challenges that are pervasive in the entire field [Lähnemann et al., 2020]:

1. A challenge of navigating varying levels of resolution.

2. A challenge of quantifying measurement uncertainty.
3. A challenge of scaling to higher dimensionalities.

While not intended to solve all issues in the field, I would still argue that non-linear dimensionality reduction, in its current state, falls short on all 3 accounts.

2.2 Cluster Alignment Tool (CAT) - Principles and functionalities

In an effort to sidestep the challenges of using and interpreting non-linear DR when applied to the use case, "verify clusters and examine the relationship between cell-type", I developed Cluster Alignment Tool (CAT). With CAT I wanted to develop a method that deals with the curse of dimensionality without limiting the number of input dimensions (genes) but instead by reducing noise. This makes CAT a complementary approach to how cell-types are currently compared with NLDR.

CAT is a tool for determining, with meaningful uncertainties, which cell-types in a dataset are most similar. This can be used to determine whether or not *in vitro* cultures produce cells corresponding to *in vivo* cells, it can assist in figuring out how cell-types within an embryo relate to each other, and it helps us ascertain that the results we find are actually consistent with previously published data. The result of CAT can be understood visually, with graphics like UMAP/T-SNE, but it is also easily interpretable via the statistics it produces. In the section below, I will cover how the algorithm works, discuss some of its shortcomings and show how it applies to the easily interpretable fashion-MNIST dataset.

2.2.1 How the algorithm works

By operating at a level of clusters rather than single cells, CAT reduces noise by averaging the gene expression among the cells in a cluster. Since clusters correspond to cell-types, this resolution is also fitting for testing how cell-types are related. For CAT to run, the input data therefore needs to be pre-clustered. Instead of finding distances between cells, we are interested in distances between clusters (cluster averaged gene expression). CAT uses the simple euclidean distance. Unfortunately, highly expressed genes will have a disproportionate influence on these distances, as illustrated in Figure 2.9. To avoid this, we normalize the expression of all genes by their median (excluding zeros), also illustrated in Figure 2.9.

For each cluster, the shortest distance will indicate which other cluster it is most similar to. To meaningfully distinguish between the shortest and next-shortest

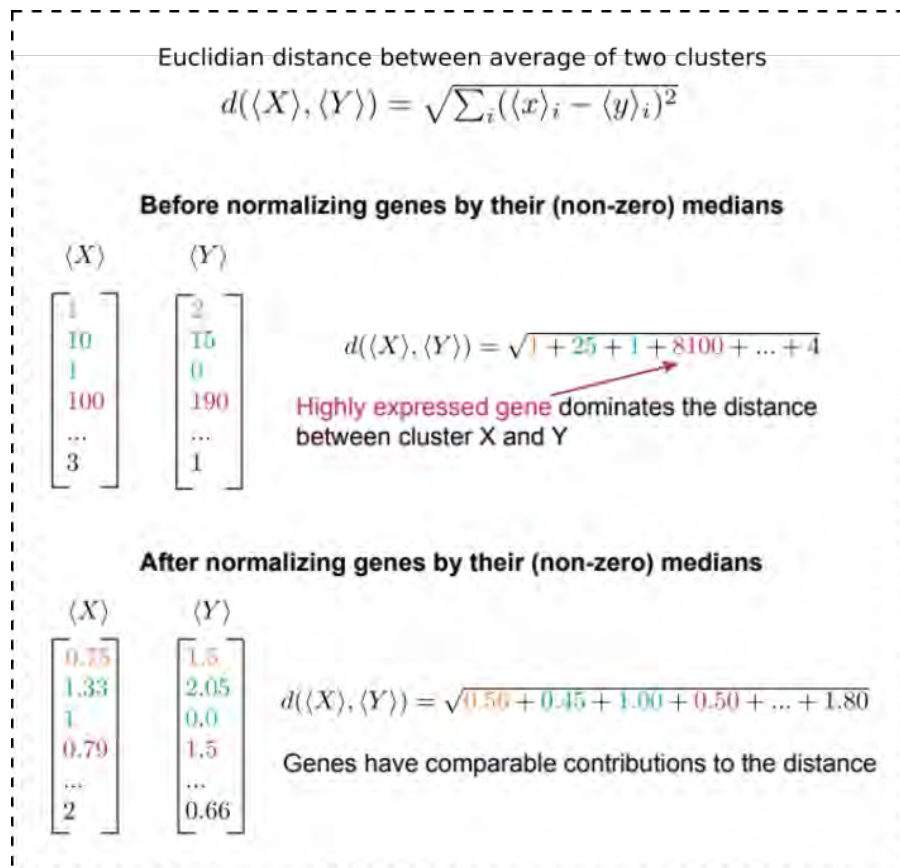


Figure 2.9: Illustration of the non-zero median normalization and its effect on highly expressed genes. Each gene, x_i , is normalized by the median for all cells expressing the gene (I.e. the median excluding zeroes). This normalization ensures that highly expressed genes (compare first and fourth gene) contribute to the distance calculation at the same order of magnitude as lowly expressed genes (compare contributions of the first (orange) and the fourth (red) gene to the distance before and after the normalization). $\langle X \rangle$ and $\langle Y \rangle$ are vectors of the average gene expression across cells in some cluster X and Y , respectively.

distances for each cluster, we need to quantify the uncertainty of the distance measurements.

Since clusters are ensembles of datapoints, it is possible to use a sampling method to estimate the distance uncertainties: Assuming that the cells faithfully represent the possible variety within a cluster, then if we take random samples from within the cluster (with replacement, i.e. allowing to take the same samples multiple time), the resulting collection of samples should be representative of what the cluster could potentially look like if we re-ran of the experiment. Performing tests on samplings with replacement is a type of statistical analysis known as bootstrapping [Efron, 1992]. Using this bootstrapping method, $N_{iteration}$ number of representations of each cluster is generated and then averaged. The Euclidean distances between the clusters are calculated at each bootstrap iteration. Thus, instead of getting one distance between pairs of clusters, there are $N_{iteration}$ distances, each with a bit of variation, representing the heterogeneity of clusters. Calculating the average and standard deviation of these distances allows the algorithm to estimate the confidence of similarity between clusters before finding the most similar nearest neighbours.

A parameter, σ , controls the confidence with which CAT can tell two distances apart when finding the nearest neighbours between clusters (constructing a neighbourhood graph). For example: Consider a cluster "a" with distances $d_b \pm \sigma_{d_b}$, $d_c \pm \sigma_{d_c}$ to some clusters "b" and "c" respectively. If d_b is the smallest of the two distances, we say that "b" is the nearest neighbour to "a". If the difference between d_b and d_c is very small (i.e. cannot be distinguished from 0, with a confidence higher than σ), we also count "c" as a nearest neighbour. The specific criteria is $\frac{|d_b - d_c|}{\sqrt{(d_b)^2 + (d_c)^2}} < \sigma$. This parameter is similar to the perplexity of T-SNE or the "n_neighbours" parameter of UMAP, since it controls the connectedness of the neighbourhood graph. As with UMAP and T-SNE it effectively controls global versus local structure; too low a sigma and the graph will not be connected, too high a sigma and everything will be connected.

If a cluster is ill-defined, for example consisting of many dissimilar cell types, the exact choices of sampled cells at each bootstrapping iteration will matter more compared to a completely homogeneous cluster. This variation between iterations will show up as high uncertainties on the distances from this cluster to other clusters, revealing that the cluster in question should be sub-clustered further or

that the cluster simply consists of heterogeneous cells (usually this would be cells that were half dead or not sequenced properly).

A pseudocode outline of CAT is given below:

Algorithm 1 Implementation of Cluster Alignment Tool (CAT)

Require:

- X - high dimensional data matrix with annotated cluster labels.
- $N_{iterations}$ - Number of iterations for the bootstrap process. Defaults to 1000.
- σ - Significance cutoff for nearest neighbours. Defaults to 1.6.

Pseudocode:

- Normalise each gene (column) of X by the median expression for all cells (rows) expressing the gene.

for $i=0$ **to** $N_{iterations}$ **do**

- Sample each cluster to its original size, with replacement.
- Calculate the average of each subsampled cluster.
- Calculate and record the euclidean distances between every combination of sampled cluster averages.

end for

- Calculate the mean and variance for the $N_{iterations}$ distances found for each combination of pairs of clusters.
- Find the nearest neighbour for each cluster based on the mean distances. For a given cluster, there might be multiple clusters with a low distance. If the difference between the distance of the nearest neighbour and another cluster is close to 0. (I.e. with less than σ levels of uncertainty), then we say we cannot determine which is actually the nearest neighbour, and they will both count as nearest neighbours.
- Construct a nearest neighbour graph and lay it out with `forceatlas2` for visualisation. There is an edge between clusters if they are nearest neighbours, with edge strength inversely proportional to distance (scaled).

Return:

- Table with distances with uncertainties and neighbour status for each cluster to all other clusters.
 - Graph visualisation.
 - Sankey visualisation.
-

2.2.1.1 Visualisation of CAT alignments

CAT can produce visualisations in the form of nearest neighbour graphs, Sankeys and tables. The visualisations show, each in their own way, which clusters in a dataset are most similar.

Nearest neighbour graph

Forceatlas2 was chosen for laying out the neighbourhood graph for CAT to visualise its results in a manner comparable to algorithms like UMAP. ForceAtlas2 [Jacomy et al., 2014] is the default layout algorithm of the network visualisation software Gephi [Bastian et al., 2009], and is available in many packages for easy use in Python [Chippada, 2022] and R [analyxcompany, 2022]. ForceAtlas2 works under the same principles as the layout algorithm in UMAP and the other algorithms like it, using repulsive and attractive forces between points to iteratively converge to a 2-dimensional graph representation. The embeddings produced by ForceAtlas2 have been shown to be on par with those of other embedding algorithms, even if its speed cannot compare to the newer algorithms like UMAP [Böhm et al., 2020].

For the purpose of CAT, the efficiency of the layout algorithm is of little concern since the plotted number of nodes (clusters) is typically less than 50, many orders of magnitude less than what is expected to be handled by this class of layout algorithm. In general, the choice of layout algorithm is not very important for CAT, almost any network layout algorithm could manage to preserve the graph structure when less than 50 points are involved. In cases where CAT only has a handful of nodes in its graph, it is even viable to draw the edges between neighbours, and as such, even a circular graph layout or linear DR like MDS could be a meaningful and interpretable representation of CAT's result. Nearest neighbour graphs are most useful when visualising the relationship between clusters within a single dataset.

Sankey

Sankey diagrams are typically used for visualising flows. CAT uses it for visualising how different clusters align to their targets, taking account of multiple nearest neighbours and the relative number of cells in each cluster. Sankeys are my preferred methods to visualise the relationship of clusters from one dataset to another, like the examples in Figures 2, 5 and 6 of [Rothová et al., 2022].

Tables

CAT produce a table for each cluster listing its distances to all other clusters, sorted by their respective distances. The nearest neighbours are highlighted in green. Examples of such tables can be seen in section 2.2.2 on Table 2.1 and in the supplementary data of [Rothová et al., 2022]. The advantage of the tables is that the uncertainties and statistics that go into the calculations of finding nearest neighbours can be displayed.

2.2.2 Example CAT usage

To help the reader get an intuitive understanding of how CAT performs and to probe its generality to problems other than scRNAseq, this section shows a usage example with the fashion-MNIST dataset, which is both easier to understand and interpret than the average scRNAseq dataset. At the same time, it is high-dimensional and complex enough to be a challenging dataset to work with.

The fashion-MNIST dataset consists of 70.000 greyscale 28x28 images, with examples from 10 types of clothing articles. The dataset is widely used, especially for classification and embedding benchmarks. For this example, we assume that the cluster classification has already been done and that the objective is to identify similarities between clusters, as we would for a typical scRNAseq dataset (like the outlined goals in section 2.1.2.2). Each image in the dataset is represented by a vector containing the individual 784 pixel values. Compared to a scRNAseq dataset, an image's pixel intensity corresponds to a cell's gene-expression value, and instead of clusters of cell-types, the dataset contains classes of clothing-types.



Figure 2.10: Fashion-MNIST dataset: The 10 categories are shown with 9 examples from each. The pictures are grayscale 28 by 28 pixels. The dataset consists of 70.000 clothing articles.

I have shown 9 images of each class in the dataset in Figure 2.10. Looking at these, you probably already have an idea about which clothing classes are going to be similar. Keep these similarities in mind as we see how CAT would align these.

Running CAT on the dataset is trivial. The dataset is available through the python

library sklearn [Pedregosa et al., 2011a] or github [Xiao et al., 2017] and is already formatted as a count matrix. Without any processing or normalisation, CAT produces the 2d network embedding, table and Sankey, that can be seen in Figure 2.11, 2.12 and Table 2.1.

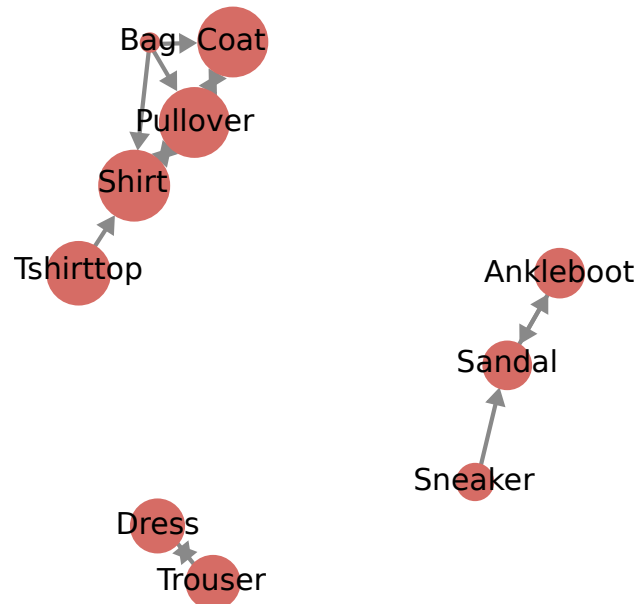


Figure 2.11: Result of CAT when run on the fashion-MNIST dataset and embedded as a network in 2d. The graph is directed, with each node pointing to its closest neighbour(s). The size of each node is scaled to reflect the average magnitude of its outgoing similarities. The plot shows the clothing articles falling into 3 categories. Notice how the (dis)similarity of the “bag” class results in many nearest neighbours for the class, each with a low similarity score.

The easiest way to get a quick overview of the results from CAT is to look at the 2d nearest-neighbours graph embedding shown in Figure 2.11. This figure that the graph separates into 3 disconnected neighbourhoods; footwear, upper-body wear and dresses/trousers. In my opinion, this corresponds well to how my human intuition would align these clusters. A possible outlier is the “Bag” class, which aligns to the upper-body wear. All classes have nearest neighbours, even if the alignment is not mutual. For potentially “bad” alignments such as the “Bag” class, there is no other class that is really a good match, which results in multiple nearest neighbours, all with a relatively high distance. This kind of sub-optimal alignment is more evident when looking at the alignments quantitatively, as in Table 2.1.

Looking at the "Bag" cluster, we can see that it has the highest numbers of nearest neighbours and that the quantitative distances to these nearest neighbours are the

	Tshirttop		Trouser		Pullover		Dress		Coat
Shirt	7,49 ± 0,11	Dress	9,89 ± 0,04	Coat	5,64 ± 0,12	Trouser	9,89 ± 0,04	Pullover	5,64 ± 0,12
Pullover	11,23 ± 0,17	Coat	16,01 ± 0,05	Shirt	5,85 ± 0,28	Coat	11,91 ± 0,07	Shirt	8,37 ± 0,29
Dress	12,88 ± 0,21	Tshirttop	16,74 ± 0,17	Tshirttop	11,23 ± 0,17	Tshirttop	12,88 ± 0,21	Dress	11,91 ± 0,07
Coat	13,30 ± 0,21	Shirt	17,07 ± 0,14	Dress	13,95 ± 0,07	Shirt	12,95 ± 0,18	Tshirttop	13,30 ± 0,21
Trouser	16,74 ± 0,17	Pullover	17,54 ± 0,06	Bag	15,56 ± 0,17	Pullover	13,95 ± 0,07	Bag	15,62 ± 0,18
Bag	18,68 ± 0,17	Bag	24,00 ± 0,12	Trouser	17,54 ± 0,06	Bag	20,09 ± 0,15	Trouser	16,01 ± 0,05
Ankleboot	21,71 ± 0,17	Ankleboot	26,32 ± 0,17	Sandal	18,01 ± 0,19	Sandal	22,92 ± 0,16	Sandal	18,18 ± 0,21
Sandal	22,00 ± 0,13	Sandal	26,39 ± 0,14	Ankleboot	19,10 ± 0,22	Ankleboot	22,99 ± 0,20	Ankleboot	19,42 ± 0,24
Sneaker	25,43 ± 0,12	Sneaker	27,26 ± 0,04	Sneaker	21,12 ± 0,06	Sneaker	24,13 ± 0,05	Sneaker	20,41 ± 0,05
	Sandal		Shirt		Sneaker		Bag		Ankleboot
Ankleboot	11,14 ± 0,21	Pullover	5,85 ± 0,28	Sandal	13,63 ± 0,32	Shirt	15,44 ± 0,19	Sandal	11,14 ± 0,21
Sneaker	13,63 ± 0,32	Tshirttop	7,49 ± 0,11	Coat	20,41 ± 0,05	Pullover	15,56 ± 0,17	Pullover	19,10 ± 0,22
Bag	17,77 ± 0,22	Coat	8,37 ± 0,29	Bag	20,44 ± 0,13	Coat	15,62 ± 0,18	Shirt	19,21 ± 0,19
Pullover	18,01 ± 0,19	Dress	12,95 ± 0,18	Ankleboot	20,78 ± 0,27	Sandal	17,77 ± 0,22	Coat	19,42 ± 0,24
Coat	18,18 ± 0,21	Bag	15,44 ± 0,19	Pullover	21,12 ± 0,06	Tshirttop	18,68 ± 0,17	Bag	19,81 ± 0,25
Shirt	18,66 ± 0,15	Trouser	17,07 ± 0,14	Shirt	22,37 ± 0,12	Ankleboot	19,81 ± 0,25	Sneaker	20,78 ± 0,27
Tshirttop	22,00 ± 0,13	Sandal	18,66 ± 0,15	Dress	24,13 ± 0,05	Dress	20,09 ± 0,15	Tshirttop	21,71 ± 0,17
Dress	22,92 ± 0,16	Ankleboot	19,21 ± 0,19	Tshirttop	25,43 ± 0,12	Sneaker	20,44 ± 0,13	Dress	22,99 ± 0,20
Trouser	26,39 ± 0,14	Sneaker	22,37 ± 0,12	Trouser	27,26 ± 0,04	Trouser	24,00 ± 0,12	Trouser	26,32 ± 0,17

Table 2.1: A table containing a simplified overview of CAT's quantitative result, showing the distance/similarity for each class of clothing. Green indicated the nearest neighbour(s). The "±" shows the standard deviation on the distance calculated from the bootstrap. This is a simplified overview; a full table would also include the p-values.

highest among all clusters. The difference between the nearest and farthest cluster is also the smallest for the "Bag" cluster. Distant alignments and small near/far ratios can occur for different reasons; 1) that the cluster is simply dissimilar to the other clusters or 2) that the cluster is ill-defined, consisting of heterogeneous samples. If a cluster is ill-defined, its distances after each bootstrap iteration are more likely to be influenced by noise and therefore become larger and more equidistant. This heterogeneity, however, should show up in the uncertainties of the distances since each bootstrap will contain a different sampling of the cluster, giving a larger spread in the distance. The "Bag" cluster has approximately the same size of uncertainties as the other cluster, so it is safe to conclude that the cluster is not ill-defined, it is simply just the most different from the other clusters in the dataset. To illustrate another use of the table, one might wonder how the footwear clusters are all related. Focusing on "Sandal", we can see that while "Ankleboots" is the only nearest neighbour at distance 11.14, "Sneakers" is not far off at 13.63. Both "Ankleboot" and "Sneaker" are much closer than the next cluster "Bag" at 17.77. This clearly suggests a relatively high similarity between "Sneaker" and "Sandal", even if "Sneaker" does not quite meet the criteria for being a nearest neighbour.

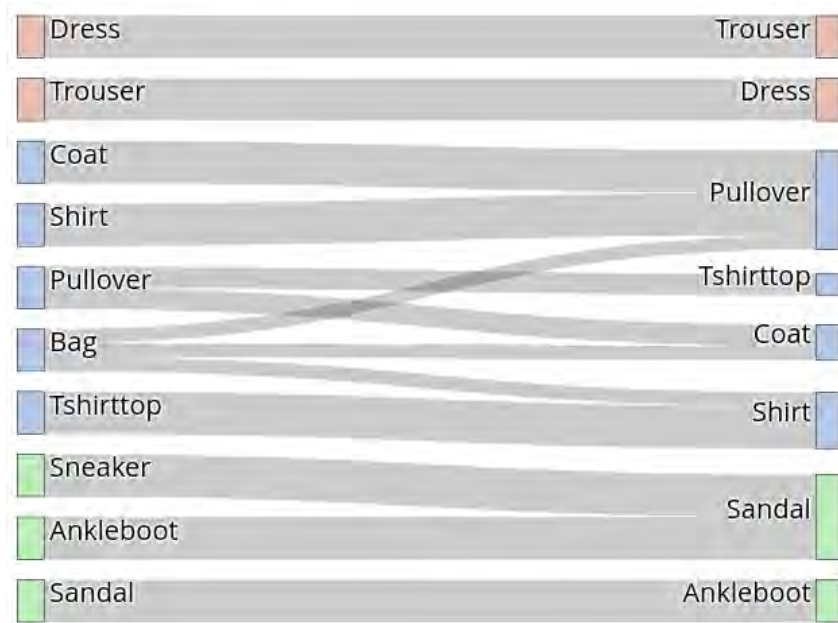


Figure 2.12: Snakey representation of CAT on the fashion-MNIST dataset. Connections indicate a nearest neighbour relationship from the nodes on the left to the nodes on the right. colors corresponds to the groups that can be seen on the network embedding in figure (2.11)

2.2.3 Limitations and strengths of CAT

The following section will cover some of the limitations and strengths of CAT, hopefully giving the reader a more thorough understanding of CAT in the process.

Limitations:

2.2.3.1 All clusters have a nearest neighbour.

Like other non-linear embedding algorithms, CAT relies on a nearest neighbour search, which means that each cluster will always have a “nearest” neighbour, no matter how far this might be. The consequence of this is that clusters that should intuitively be isolated will always “attach” themselves to other clusters and visually connect with these in the graph. Algorithms like T-SNE and UMAP have a parameter for tuning a cut-off for neighbours if the similarities get too distant. This feature could, in principle, also be implemented in CAT, but as of the moment of writing, this is not implemented, so potential users should beware of clusters that align towards many nearest neighbours with big distances.

2.2.3.2 How to interpret CAT distances

In the coverage of CAT so far, I have only addressed rankings of the distances CAT computes between clusters, i.e. which are closest and which are not. The main reason to focus on the rankings (the nearest neighbours) rather than the actual value of the distance to judge similarity is that the distances change from experiment to experiment (dataset to dataset).

Technical differences between sequencing methods influence, e.g. the number of genes that get observed in an experiment and can bias the relative detection between genes (sequence length bias, guanine-cytosine-content bias, etc) [Liu et al., 2012, Ding et al., 2020, Dabney and Meyer, 2012]. Technical differences such as these make it hard to say if a concrete distance value is short or long because it very much depends on the data. For the same reason, it is hard to directly compare the internal distances between clusters in one dataset to the internal distances between clusters in another dataset. The distances have meaning, but only relative to the other distances within the same experiment. Distances in a dataset seem to have a lower bound. Because of the noise (natural variance or otherwise) that is inevitably present in data, the curse of dimensionality dictates that datapoints will always be pushed at least a certain distance apart from each other (Figure 2.3). This lower bound can be estimated by comparing a cluster to itself (using randomly sampled sets of the same cluster with replacement). The only component that should contribute to this distance is noise/variation. Empirically, we find that the shortest distances between clusters seem to be comparable to this self-to-self distance.

The cell-types (hindgut1, hindgut2, midgut) will be discussed in the next chapter, for this examples the specific names of the clusters are not important.

Here is an example using the Rothova2022 dataset. (The specific names of the clusters are not important for this example, only the distances. The names are explained in the next chapter if the reader is curious). The distance from the hindgut1 cluster to itself (random re-sample to random re-sample) is on average $12.01 (\pm 0.36)$, while the distance from hindgut1 to its nearest neighbour, hindgut2, is 13.15 ± 0.28 . So the fact that the distances between clusters in the Rothova2022 dataset are never less than 10, does not mean that the clusters are actually far apart (i.e. dissimilar), it just means there is a minimum noise-driven (and therefore dataset dependant) contribution to all distances. The distance from hindgut1 to

midgut is 16.42 ± 0.24 ; at first glance, this distance seem comparable to its hindgut2 distance (13.15), but when the minimum distance is considered, the difference between these 2 distance can be more clearly appreciated.

It would be interesting to explore distances between datasets further, perhaps with the help of overlapping reference cell-types sequenced in both datasets. For now, we simply use nearest neighbours.

2.2.3.3 CAT requires a minimum number of cells per cluster

A requirement for CAT to work is the pre-clustering of the data it uses as input. In order to provide proper uncertainties on the distances between clusters, each cluster needs to contain enough cells to average over at each step of the bootstrap iteration to 1) average out the noise and 2) properly reflect the heterogeneity of the cells in the cluster. The number of cells (data points) that is required for these 2 criteria depends on the dimensionality of the data as well as the distributions of data within the dataset, but should, in principle, always converge given enough data points and bootstrap iterations (see section 2.2.3.5).

The question then becomes, how many cells per cluster is needed in practice? To get an estimate of what this number of cells might look like for an average scRNAseq dataset, I will use our own dataset, Rothhova2022, as an example. In this context, a cluster contains "enough" cells if the sub-sampled cluster averages used in the bootstrap process do not substantially change when including more cells. Example: if a sub-sampled cluster contains only 5 cells, then their average will look substantially different from the average of another 5 sub-sampled cells from the same cluster. The specific noise/variation of the selected 5 cells would dominate the signal. One way to measure if the sub-sampled cluster average has reached convergence towards the true cluster average is to look at the distance between the sub-sampled cluster average and origo. As the number of cells included in the sub-sample, its distance to origo will converge towards the true value. Figure 2.13 shows sub-sampled cluster averages for 3 randomly chosen clusters from the Rothova2022 dataset and how their distance to origo changes as a function of the number of sampled cells. Judging from this figure, I would estimate that about 50 cells per cluster are enough to get an accurate enough representation of the cluster to use with CAT.

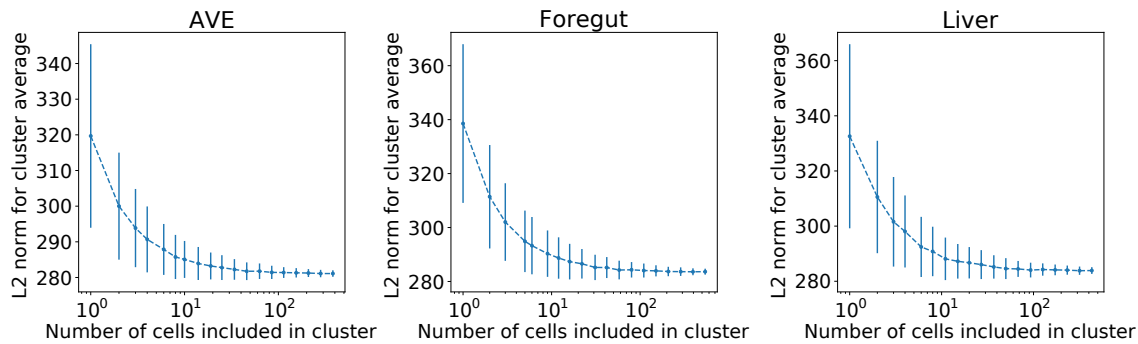


Figure 2.13: Cluster-averages all converge to a specific euclidean distance from origo ($d = \sqrt{(\sum_i x_i^2)}$), if enough cells are included when calculating the average. This figure shows 3 example clusters from the Rothova2022 dataset. Dashes lines are guides for the eye, while the vertical bars show the spread over 500 independent runs. X-axis show how many cells were sampled (with replacement) to calculate the cluster average.

Strengths:

2.2.3.4 Computational runtime

With the number of bootstrap iterations ($N_{\text{iterations}}$), ideally being larger than 100, the bootstrap for loop in the CAT algorithm takes up the vast majority of the computation. Conveniently the loops are independent and can therefore easily be parallelized. With a parallelized CAT implementation, it took about 5 minutes (or approximately 1 coffee break) to finish 1000 bootstrap iterations on the Rothova2022 dataset (10.285 cells x 24262 genes). This was done on an AMD Ryzen threadripper 2990WX CPU using 10 threads. As can be seen in Figure 2.14, CAT scales linearly with the number of cells in a dataset, enabling it to be used on even large datasets. An implementation using the GPU's superior number of parallel compute cores and its ability to rapidly calculate distances between vectors could potentially speed up CAT.

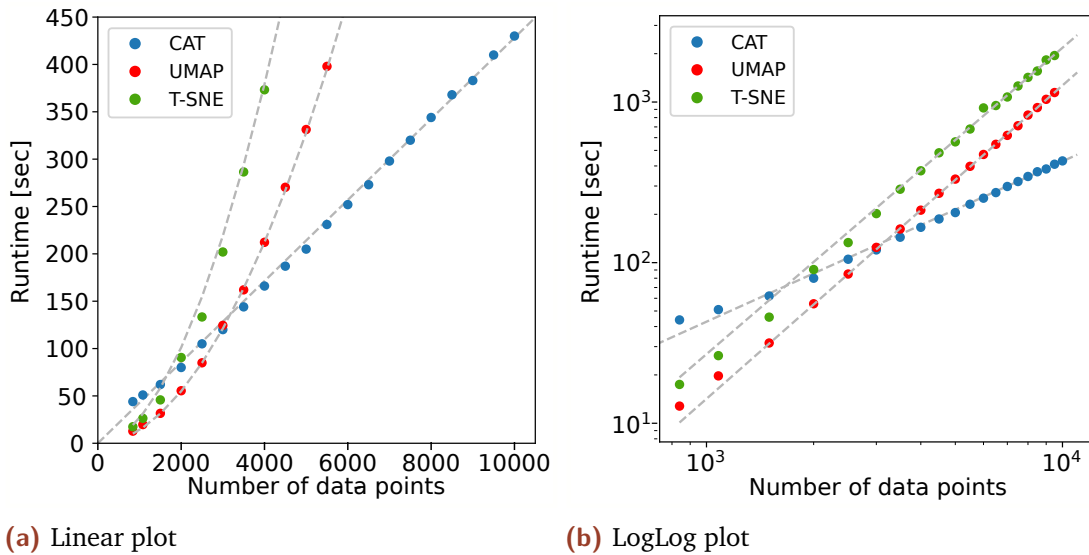


Figure 2.14: Comparison of the scaling of CAT, T-SNE and UMAP, using our own mouse embryonic scRNAseq dataset (Rothova2022) with 24262 genes and a variable number of cells. T-SNE and UMAP were run with default parameters and their exact solvers. The figure shows the ideal linear scaling of CAT compared to the scaling of competing non-linear DR methods.

2.2.3.5 CAT is not sensitive to random number seed

Although CAT uses bootstrapping to arrive at its results, making it a stochastic algorithm just like UMAP and T-SNE, it is not prone to the same fluctuation in its output. The bootstrapping in CAT is used to calculate distances between re-sampled clusters. In the following, I will argue that these distances follow a non-central χ^2 distribution. The distribution is important because we want to show that the average of these distances over many runs of the bootstrap will, therefore, always converge and that CAT's results are entirely reproducible across runs, regardless of random seeds.

Each cluster in a scRNAseq dataset represents a population of cells that can be thought of as a distribution of cells. These distributions have some quantifiable mean (the cluster average) and a finite variance. If we take multiple samples with a sufficiently large number of cells in each from these distributions with replacement, then the averages (mean or median) of these samples should be approximately normally distributed, given the central limit theorem [Bentkus, 2005].

If $\{C_1, C_2, \dots, C_n\}$ is a set of independent and identically distributed random variables (cells) being drawn from the same distribution (cluster), then the formula for the sampled cluster average, \bar{X} , is trivially:

$$\bar{X} = \frac{C_1 + C_2 + \dots + C_n}{n} \quad (2.2)$$

Each cell is a vector of independent gene-counts (features); $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$. To calculate the cluster average, \bar{X} , the different genes can, therefore, be added independently, meaning that the individual gene counts of \bar{X} , (\bar{x}_i) , are themselves sums of independent and identically distributed random variables, each subject to the central limit theorem.

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_m \end{bmatrix} = \begin{bmatrix} (c_{11} + c_{21} + c_{31} + \dots + c_{n1})/n \\ (c_{12} + c_{22} + c_{32} + \dots + c_{n2})/n \\ \vdots \\ (c_{1m} + c_{2m} + c_{3m} + \dots + c_{nm})/n \end{bmatrix} \quad (2.3)$$

To calculate the squared Euclidean distance between 2 cluster averages, \bar{X} and \bar{Y} , we get the following terms:

$$d(\bar{X}, \bar{Y})^2 = (\bar{x}_1 - \bar{y}_1)^2 + (\bar{x}_2 - \bar{y}_2)^2 + \dots + (\bar{x}_m - \bar{y}_m)^2 \quad (2.4)$$

Since the individual gene counts of the cluster averages, \bar{x}_i and \bar{y}_i , are normally distributed random variables, it follows that the difference between any two pairs of features for two cluster averages, $\bar{x}_i - \bar{y}_i$, should also be a normal distributed random variable, each with their own mean and variance; $\bar{z}_i = \bar{x}_i - \bar{y}_i$. This means that the squared difference for each term in the distance sum becomes a chi² distributed random variables.

$$d(\bar{X}, \bar{Y})^2 = \sum_i^m (\bar{z}_i)^2 \quad (2.5)$$

Because the different \bar{z}_i are normally distributed, but each has unequal means μ_i and unequal variances σ_i , the resulting distribution for the distance between 2 clusters does unfortunately not have an exact analytic solution. However, recent advances approximating sums of chi-square distributed random variables show that sums like these can be approximated by a non-central chi-squared distribution [Imhof, 1961, Castaño and López, 2005, Bodenham and Adams, 2016].

In the case of the CAT algorithm, we can make one simple approximation that will trivially show the same. In the first step of the algorithm, the genes are normalised against their median count (zeros excluded). This is done to avoid the distances being dominated by a single or few individual genes that happen to have larger copy numbers. Because of this normalisation, we can expect the random variables \bar{x}_i , \bar{y}_i and therefore z_i to have approximately the same order of variance. If we assume the variances to be the same, then the expected distance (squared) between clusters becomes exactly non-central chi-squared distributed. [Abramowitz and Stegun, 1972, p. 942]. The non-central chi-squared distribution has the convenient property that as the number of dimensions (genes), i.e. number of terms in distance sum, tends towards infinity, then the excess kurtosis (or “tailedness”), as well as skewness of the distribution tends to 0.

Using a bootstrap process, CAT repeatedly samples the distances between cluster averages from this distribution. Due to the above-mentioned properties of the distribution, we can expect the average of the bootstrapped distances between 2 clusters to, therefore, always converge. This means that CAT’s results should be entirely reproducible across random seeds, unlike its competitors.

I strongly suspect that similar approximations of convergence could be found for other distance measures, such as cosine similarity or fraction distances [Aggarwal et al., 2001]. Having run CAT multiple times with other distance measures, at least it seems to hold empirically. Perhaps an entirely analytical solution without the need for bootstrapping could even be derived.

2.2.3.6 Complexity

The entirety of the algorithm can be explained in a page or 2, and requires very little mathematics to understand. This is in stark contrast to T-SNE, UMAP and PaCMAP, whose family of algorithms seems to grow more complicated with each iteration,

as evidenced by the many papers and blog posts devoted to analysing their inner workings and trying to explain their performance [Wang et al., 2020, Chari et al., 2021, Oskolkov, 2019]. For algorithms designed to be used by people that do not necessarily fully understand them, I personally feel like simplicity is an important feature.

CAT only has a single tunable parameter, σ .

2.2.3.7 CAT results do not appear to depend on distance metric

As mentioned above, the concept of proximity and distance begin to lose qualitative meaning when operating in high-dimensional spaces. The choice of appropriate and best performing distance metrics has therefore been studied both in the setting of scRNAseq [Kim et al., 2019] and machine learning more generally [Aggarwal et al., 2001, Shirikhorsidi et al., 2015, Smets et al., 2019]. By default, CAT uses Euclidean distance under the hood due to its simplicity. In order to evaluate CAT's robustness to the choice of distance metrics, we have run it with cosine and Pearson distance, which both tend to perform better in high-dimensional spaces. Cosine distance measures the similarity between two vectors as the angle between them, making it invariant to the scaling of the vectors. Pearson distance is the same as cosine distance after a centering of the vectors, making it invariant to both scale and location. It measures the correlation between the features of the vectors. The formulas for these are in the table below:

Distance measure	Formula
Euclidean	$D = \sqrt{\sum_1^n (x_i - y_i)^2}$
Cosine	$D = \frac{\sum_1^n x_i y_i}{\sqrt{\sum_1^n x_i^2} \sqrt{\sum_1^n y_i^2}}$
Pearson correlation	$D = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2} \sqrt{\sum_1^n (y_i - \bar{y})^2}}$

Testing these on the Rothova2022 dataset, the results of CAT can be seen in Figure 2.15 in the form of network plots. Tables and Sankey diagrams reveal the same, i.e. that CAT is not particularly sensitive to the choice of distance metric. Switching between Euclidean, cosine, and Pearson does not change a single nearest neighbour found in this test.

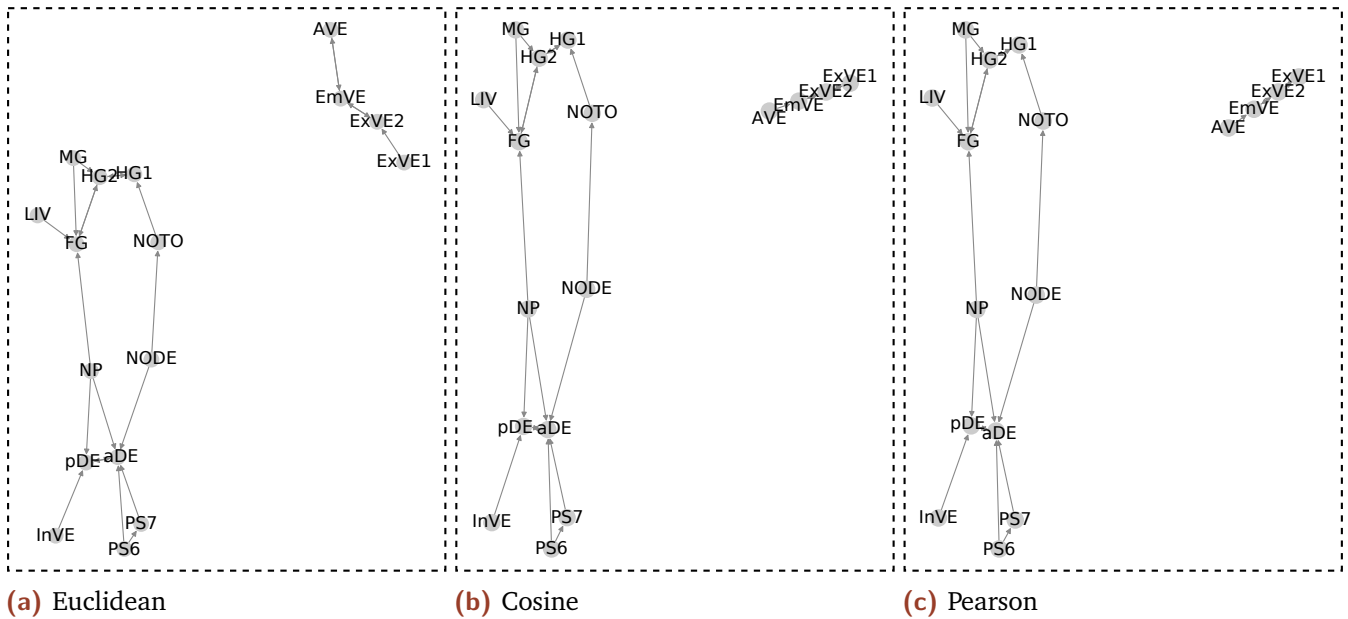


Figure 2.15: Comparison of nearest neighbour graphs produced by using CAT on our mouse embryonic scRNAseq dataset (Rothova2022) using different distance metrics to determine the similarity between clusters. The plots show that all nearest neighbours between clusters are identical regardless of metric. Each node represents an *in vivo* cluster of cells from the Rothova2022 dataset, labelled with its identified cell-type. The specific labels are not important for this comparison, but are explained in the introduction of chapter 3, if the reader is curious.

2.2.3.8 CATs sensitivity to number of genes

To test how sensitive CAT is to the selection of specific genes and the number of genes included, I ran CAT varying the percentage of genes included in the analysis and compared the outcome to CAT run on the full dataset. The results can be seen in Figure 2.16. Even with a random 80% of the geneset excluded, about 80% of the nearest neighbours are still correctly found. Including only 1% (243 genes) still conserve the majority of the nearest neighbours. When the number of genes included in the analysis goes down, CAT also begins to find nearest neighbours that were not present when using the full geneset. The number of these presumably wrong nearest neighbours is still small when using more than 50% of the geneset.

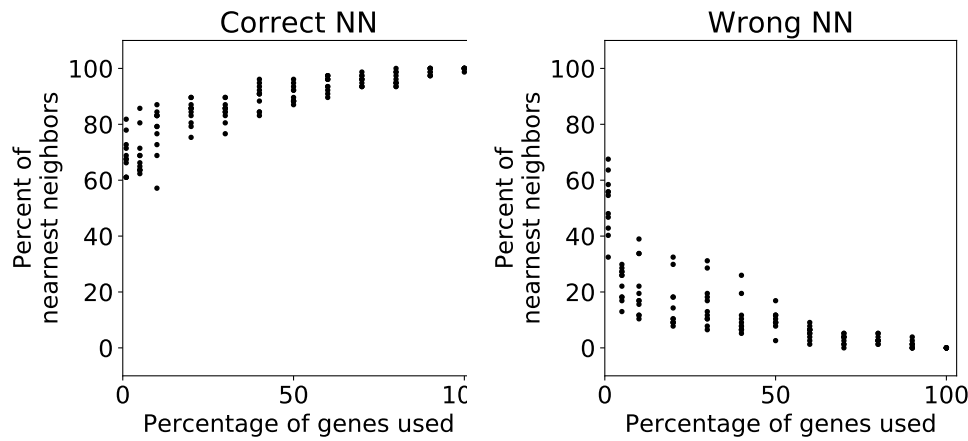


Figure 2.16: Test of CAT’s dependence on the percentage of genes included in the Rothova2022 dataset (both *in vivo* and *in vitro* cells). X-axis shows how many genes were included in 10 tests compared to the full geneset of 24262 genes. Genes were sampled at random. Y-axis on the left plot shows the percentage of nearest neighbours (compared to CAT run using the full geneset) that was also found with CAT using less genes. Y-axis on the plot on the right shows the percentage of nearest neighbours found in a CAT run, that was not present compared to CAT run on the full geneset.

2.3 Conclusion

CAT was developed to fill the need for a tool that could compare cell-types without the difficulties of conventional NLDR methods. CAT provides this by taking a different approach. By focusing on noise reduction as the main way to overcome the curse of dimensionality, CAT ends up providing answers in a way that is complementary to the current techniques. This makes it hard to directly benchmark CAT’s performance in cell-type comparison against NLDR.

I criticise current NLDR for failing to address the 3 main challenges described in "*Eleven grand challenges in single-cell data science*" [Lähnemann et al., 2020]. I will now argue how CAT meets these challenges.

1) The challenge of scaling to higher dimensionalities:

Because of the noise reduction, it is possible for CAT to utilize all dimensions (genes) in a dataset for its analysis without the need for prior feature selection with e.g. PCA. CAT furthermore provides the best possible runtime scaling (linear), allowing the number of cells in a dataset to scale as well.

2) The challenge of quantifying measurement uncertainty:

Due to the design choice of CAT to focus on clusters rather than single cells, CAT

can use bootstrapping to quantify the uncertainties of the results that it produces. This allows the researcher, if not to trust the results, then at least to know how strongly to doubt them. The fact that CAT converges to the same result every time is definitely a bonus.

3) The challenge of navigating varying levels of resolution:

CAT is inherently dependent on an external and predefined scale in the form of clusters. For NLDR, the problem is that its scaling parameter (perplexity) does not clearly indicate what resolution the algorithm draws out of the data to display (it is also dependent on the data). Since CAT uses clusters as a basis for its analysis, the resolution is always very explicit. To navigate to different levels of resolution, the user can easily re-cluster the data to the desired scale. Unfortunately, CAT requires a minimum number of cells to be included in each cluster. Even if this limit is fairly low (approximately 50 cells), it means that there is a lower bound on the resolution.

If I get more time to work on CAT, I would like to investigate how the values of the distances could be further used. It would be interesting to see how scRNAseq experiments using different sequencing methods but containing the same cells could be used to calibrate and investigate the behaviour of the distances. I would also like to address the shortcoming that all clusters in CAT have a nearest neighbour. It should be possible to define a cut-off for clusters that are obviously only a nearest neighbour from the perspective from one of the two clusters.

Analysing data obtained on embryogenesis and gastrulation

"It is not birth, marriage, or death, but gastrulation which is truly the most important time in your life."

Lewis Wolpert

This chapter presents the main analysis results that I helped contribute to the article introduced in the previous chapter: [Rothová et al., 2022].

The main question that the article tries to answer is how the cells surrounding the embryo contribute to the formation of the gut in the early stages of development. The article furthermore attempts to recreate the early gut development *in vitro*. The results I present in this chapter will be slightly more focused on the computational analysis angle than the article.

The introduction contains a quick overview of the formation of the early embryo, specifically the time period around the formation of the gut. The "Methods and results" section covers how the cell-types in our dataset are identified before exploring their relationship to one another. I also show how we can map the differences that make the *in vitro* cultures fail to fully mimic the early *in vivo* gut development. Lastly, I discuss the usefulness of the computation methods used.

3.1 Introduction

The embryo is, by nature, a diverse and changing organism. After conception, the nascent mammalian embryo expands orders of magnitudes in size, while cells

rearrange to change the embryo's shape, before maturing into an animal ready for birth. Despite undergoing this dynamic process, the embryonic growth spurt is incredibly robust, reaching developmental goals with remarkable precision along a predefined trajectory.

In the early embryogenesis, one of the first and most critical development milestones is gastrulation. Gastrulation occurs in all multi-cellular organisms (except for sponges [Nakanishi et al., 2014]) and is the process where the basic body plan is first laid down [Tyser et al., 2021]. For this thesis, the focus will be on mammalian gastrulation, specifically mouse. While the process happens at slightly different timescales for different species; around day 6-9 for mouse [Bardot and Hadjantonakis, 2020, Tam et al., 1993], and around day 14-21 for human [O'rahilly and Müller, 2010], it is otherwise remarkably similar between these two species, making mouse a good model organism for understanding our own human development.

Around day 3-4 (prior to gastrulation), the mouse embryo is approximately spherical in structure (a structure called the blastocyst), consisting of trophoblast, the outer layer of cells that will go on to form part of the placenta and the inner cell mass (ICM), which consists of the cells that will go on to form the embryo proper [Edgar et al., 2013]. The blastocyst stage is the starting point of the developmental map in Figure 3.1, where the abbreviations for each cell-type are also shown.

During gastrulation, the still pluripotent stem-cells of the ICM differentiate into one of the three germ layers; endoderm (inner layer), mesoderm (middle layer) and ectoderm (outer layer), in the process limiting their potency and locking their cell-fates onto one of these developmental paths [Riveiro and Brickman, 2020, Edgar et al., 2013]. In layman's terms, this is the point of development where the cells that actually form the animal proper first decide which part of the body they will become. The developmental paths and the tissues/body-parts associated with the three germ layers are shown in Figure 3.1. While the cells internally commit to these decisions, they globally migrate, reshaping the embryo from an elongated sphere to the classic fetus 'C'-shape, with a clearly defined left-right, back-front and head-tail axis. The ICM also gives rise to the hypoblast, a group of cells that, among other things, form the yolk sac and visceral endoderm (VE), membranes that envelop the embryo and help pattern and orient its growth [Riveiro and Brickman, 2020, Srinivas, 2006]. It has normally been understood that cells of the hypoblast are an extra-embryonic lineage that does not become

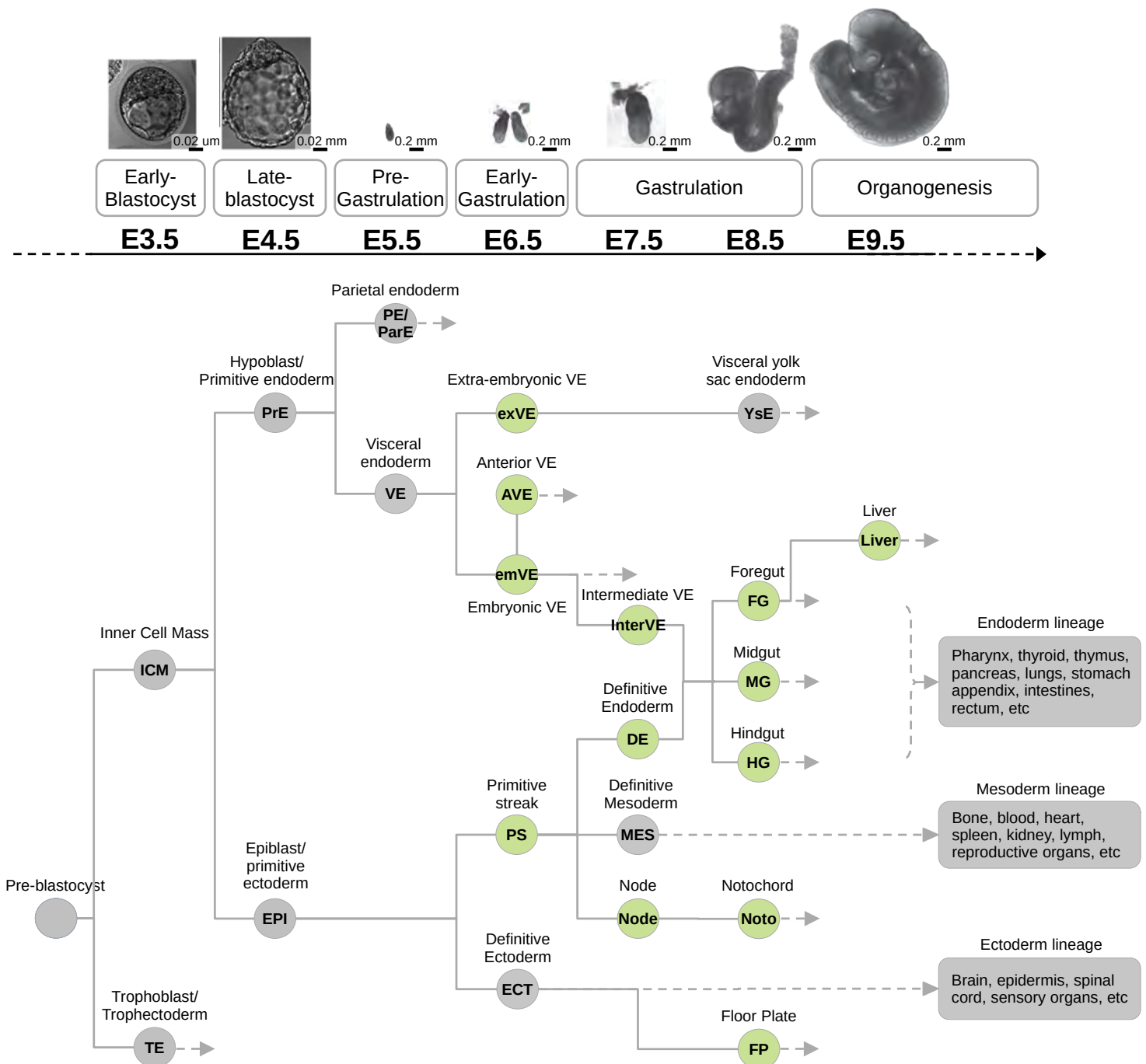


Figure 3.1: A developmental map of the early mouse embryo, from blastocyst stage to the end of gastrulation. The map shows the lineage structure and earliest occurrence of each cell type. The green color indicates the cell types present in the Rothova2022 dataset. Lineage tree idea inspired by [Nowotschin et al., 2018]. Images of mouse embryo stages from [Rothová et al., 2022, Xue et al., 2013, Saiz and Plusa, 2013]

part of the animal, only serving as a transient support structure for the embryo [Hogan and Zaret, 2002]. Recent research has cast doubt on this picture.

Using genetic labelling with fluorescent markers and high-resolution imaging of individual embryos at the gastrulation stage, it has been shown that a portion of VE lineage cells intermixes with the DE lineage and end up contributing to the formation of early gut tube [Kwon et al., 2008, Viotti et al., 2014, Scheibner et al., 2021]. Transcriptomic experiments support this notion [Pijuan-Sala et al., 2019, Nowotschin et al., 2019].

This behavior is curious, suggesting that the distinction between the cell lineages that diverge at the late-blastocyst stage might not be so clear-cut. In the classical view of development, as exemplified by the Waddington landscape metaphor, the cells become less potent as they mature, following a path along an ever more branching tree structure to more specific and narrow cell-types [Waddington, 1957]. While some plasticity of the cell-types at the branching points of the tree is expected, the notion that branches of the tree merge is not.

If VE contributes significantly to DE, then it should be able to form gut organoids on its own. This is hard to test in an in-vivo setting, given the presence of DE, but it could potentially be tested in in-vitro.

To study this extra-embryonic VE to embryonic endoderm transition in more detail, we generated an in-vivo single-cell RNA sequencing dataset (using the MARS-seq protocol [Jaitin et al., 2014] as mentioned in section 2.1.1), focusing specifically on the gastrulation period, day 6.5 - 9.5. Analysing the Rothova2022 data, we find and trace the canonical lineages and furthermore pinpoint the exact population of VE-derived cells with an expression profile suggesting that they are transitioning to become more "DE gut"-like. We call these the intermediate visceral endoderm (InterVE). As part of the Rothova2022 dataset, we also generated in-vitro samples comprised of embryonic stem cells (i.e. in-vitro epiblast-like cells) and differentiated them towards the gut lineages under different conditions. We determine the efficiency and precision of the in-vitro differentiation protocols by using CAT to compare their outcome to the in-vivo part of Rothova2022. Additionally, we performed the above hypothetical experiment, sequencing naïve extra-embryonic endoderm (nEnd) stem cells (i.e. in-vitro hypoblast-like cells), showing that it is possible to differentiate these to produce gut-like cells. Our results suggest that the developmental trajectory of these cells goes through the InterVE path by comparing

to our in-vivo dataset. The comparisons between our in-vivo and in-vitro datasets, as well as an external 3rd party dataset, were made possible using CAT.

3.2 Methods and results

3.2.1 Identifying the embryonic cell types from experimental *in vivo* data

As the fertilized egg grows and divides from 1 cell to 2, 4, 8 and so on, the cells change their gene expression, differentiating into different cell-types (as illustrated in Figure 3.1). The genes' expression and changes in their regulation are what determine the cells' function and, thereby, identity.

To study and map the origin of the gut lineages in our *in vivo* scRNAseq datasets (see section 2.1.1), we first need to determine exactly which cell types are present.

Using the transcriptomes we have gathered for each cell, we sort the cells into clusters of similar expression profiles (see clustering in chapter 2.1.2.1). The differences in expression patterns between the clusters enable us to assign identities to the cell types. We did this in three ways: 1) Using known marker genes, 2) using differential gene expression with enrichment analysis and 3) using CAT to compare clusters to annotations in a 3rd party dataset.

1) Using known marker genes.

Embryogenesis has been studied for at least 2 thousand years [Wallingford, 2021] and through experiments like dissections [Wallingford, 2021], electrophoretic mobility shift assay [Okamoto et al., 1990, Schöler, 1991], immunochemical staining [Herrmann, 1991], northern and southern blots [Pruitt, 1994, Chambers et al., 2003], gene modification with fluorescence microscopy [Botchkarev et al., 1999], reverse transcription polymerase chain reaction [Bao et al., 2011], RNAseq [Irie et al., 2015] and recently scRNAseq [Lun et al., 2016], many cell-types and their associated marker genes are already well established. Simply quantifying and inspecting the expressions of these known genes across the clusters of our dataset enables us to assign identities to the clusters. This approach, however, is not without its flaws; unfortunately, markers are often shared between multiple lineages [Zhao et al., 2012] and are typically manually selected, requiring a survey of the

current literature and databases, leaving a bit of wriggle room depending on which markers are chosen or how much weight/trust each marker is given. This approach becomes extra tricky once applied to cells grown *in vitro* since specific protocols might up/down-regulate genes in unexpected and unknown ways, skewing the significance of the marker genes. Since marker genes are based on current literature, they are also less useful for rare sub-types or new cell-types that have not been extensively studied previously, such as InterVE. Relying on the expertise of colleagues and the databases: Mouse Genome Informatics [Bult et al., 2019], Gene Expression Database [Smith et al., 2019], and Mouse Models of Human Cancer database [Krupke et al., 2017], we compiled a list of marker genes for different known cell lineages, which is shown in Table 6.1 in the supplementary of this thesis. Visualizing these genes in different combinations on a 2d embedding plot (UMAP) helped guide us in annotating the various clusters in our dataset.

2) Differential gene expression and enrichment analysis.

Once the cells have been clustered, it is common to find the differentially expressed genes (DEG) among the clusters. Typically one focuses only on the genes that are more highly expressed within a cluster when compared to other clusters. Some natural variation in the gene expression is expected; to account for this, only genes where this variation cannot statistically explain the observed difference between clusters are considered. For the Rothova2022 data I used the `rank_genes_groups` function from Scanpy (a scRNAseq python toolkit) [Wolf et al., 2018] to perform this DEG analysis. The function is a wrapper for a t-test [Student, 1908] between the clusters that correct for multiple comparisons (one comparison for each gene) using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995].

The list of statistically significant DEGs for each cluster can then be "text-mined". In databases like the Mouse Genome Informatics and the others mentioned above, genes are annotated with text terms describing which biological functions they are involved in and in which part of the anatomy they are expressed. We used MouseMine [Motenko et al., 2015] (a tool for interacting with the Mouse Genome Informatics database) to count the number of occurrences for each text term associated with the DEGs found in the previous step and to perform an enrichment analysis to determine the statistical likelihood of getting those exact counts if the genes had been chosen at random (among all possible genes in the dataset).

If the observed counts for those text terms were unlikely to occur by random chance, we assume that our cluster does indeed have something to do with the features described by the text terms associated with its DEGs.

For example: At the time we did this analysis, we had not definitively assigned identities to each cluster, and the (then unnamed) InterVE cluster did not have any unique known marker genes pinpointing its identity as either VE or DE derived in origin for certain. Applying the process above we get the following DEGs for InterVE: *2900073G15Rik, 5730469M10Rik, Abhd2, Actn1, Apoc1, Apoe, Arfgap3, Car2, Cbx7, Cd8a, Cdh1, Cited2, Cldn19, Cldn6, Cldn7, Cnn2, Colec12, Cpm, Cst3, Cstb, Ctsc, Ctsh, Degs1, Egr1, Emb, Emp2, Epcam, Eras, Fam107b, Fgfbp1, Fhl2, Fmr1nb, Fos, Foxa3, Gas6, Glrx, Gm6030, Gprc5a, Gpx2, Gsn, Hdac6, Itm2b, Jun, Jup, Klf6, Krt18, Krt19, Krt8, Laptm4b, Lhfpl2, Lima1, Lpar1, Mogat2, Myl6b, Nid2, Nptx2, Parm1, Pdzk1ip1, Peg10, Perp, Phlda1, Pkdcc, Plat, Polg, Prkch, Prss12, Pvrl2, Rbp1, S100a10, Sat1, Sepp1, Serpinb6a, Slc16a1, Slc2a1, Slc2a3, Slc39a4, Slc39a8, Sox17, Sp5, Stard10, Stard8, Sul2, Tagln2, Tfpi, Tmprss2, Trap1a, Trh, Tspan7, Txndc12*. Parsing these InterVE DEGs to MouseMine platform [Motenko et al., 2015] for anatomy enrichment analysis, we find anatomy terms (with a likelihood less than $1e-9$ to have occurred by random chance) that are displayed in Table 3.1 below.

Anatomy Terms	P-value
Endoderm	2.761674e-18
Embryo endoderm	7.175610e-16
Extraembryonic component	5.546667e-11
Extraembryonic endoderm	7.076719e-11
Hindgut	7.496198e-9
Stomach	9.151894e-9

Table 3.1: The anatomical terms associated with the InterVE. The terms are found by anatomy enrichment analysis based on the genes in InterVE that are up-regulated compared to the other cell-types in the dataset.

As we can see from the table, the InterVE cluster has anatomy features that look both like embryonic *and* extra-embryonic endoderm, supporting the hypothesis that this cluster is made up of cells of VE (an extra-embryonic lineage) origin that is in the process of intercalating into the embryonic endoderm as discussed in the introduction.

We performed similar enrichment analyses for all clusters in our dataset to help

guide our cell-type annotation.

3) Using CAT to find similar clusters in a 3rd party dataset.

To further confirm the cluster identities from the 2 previous steps, we used CAT to compare Rothova2022 to Nowotschin’s dataset [Nowotschin et al., 2019], a previously published scRNAseq dataset that also covers the gastrulation period of the mouse embryo. The Nowotschin dataset covers days 3.5 to 8.75 of development, providing a significant but not complete overlap. The comparison result between the Nowotschin and Rothova2022 datasets are included as tables in the supplementary section 6.1. These are the tables that correspond to the Sankey that can be seen in Figure 2C in [Rothová et al., 2022].

End result of identification

Combining the results from the three methods, we label the in-vivo clusters from Rothova2022. The labeling can be seen on Table 3.2 below, as well as in Figure 3.1, where the clusters are highlighted in green. Some clusters share the same identity (we postfix them 1 and 2), and appear to be the same overall population, but are captured at different time-points along their differentiation. DE1 and DE2 contain markers and enrichment terms that seem to suggest that the early DE1 sits more anterior, while the later DE2 sits relatively posterior within the embryo. To ensure the identified cell types are placed correctly in the developmental tree in relation to each other, we apply RNA velocity [Bergen et al., 2020]. The directionality of differentiation discernible by RNA velocity confirms that InterVE is differentiating from a more VE-like state towards a more DE-like state, and not the other way around.

Label	Cells	Label	Cells
Anterior visceral endoderm (AVE)	391	Hindgut 2 (HG2)	281
Definitive endoderm 1 (DE1)	319	InterVE	193
Definitive endoderm 2 (DE2)	506	Liver	441
Embryonic visceral endoderm (EmVE)	209	Midgut (MG)	356
Extra-embryonic visceral endoderm 1 (ExVE1)	319	Node	430
Extra-embryonic visceral endoderm 2 (ExVE2)	253	Notochord	269
Foregut (FG)	553	Parietal endoderm (PE)	162
Floor plate (FP)	809	Primitive streak (PS1)	181
Hindgut 1 (HG1)	379	Primitive streak (PS2)	231

Table 3.2: The cell-types label assigned to each cluster in the Rothova2022 dataset, together with the number of cells in each cluster.

3.2.1.1 Functional similarities of the intermediate visceral endoderm

By running CAT on a subset of genes belonging to functional categories, we can investigate how InterVE compares to other clusters in relation to only these features, thereby learning what it is InterVE is changing first to become more gut-like. Functional projection of the distances between clusters was run separately for 23 lists of genes corresponding to biologically relevant processes retrieved from the gene ontology databases mentioned in section 3.2.1. The exact lists are supplied in the supplementary of our paper [Rothová et al., 2022]. The result is shown in the colored tables in Figure 3.2 along with 3 examples of functional projections reveal how the nearest neighbour of InterVE changes between either the DE lineages, VE lineages or a mix, depending on the projection used.

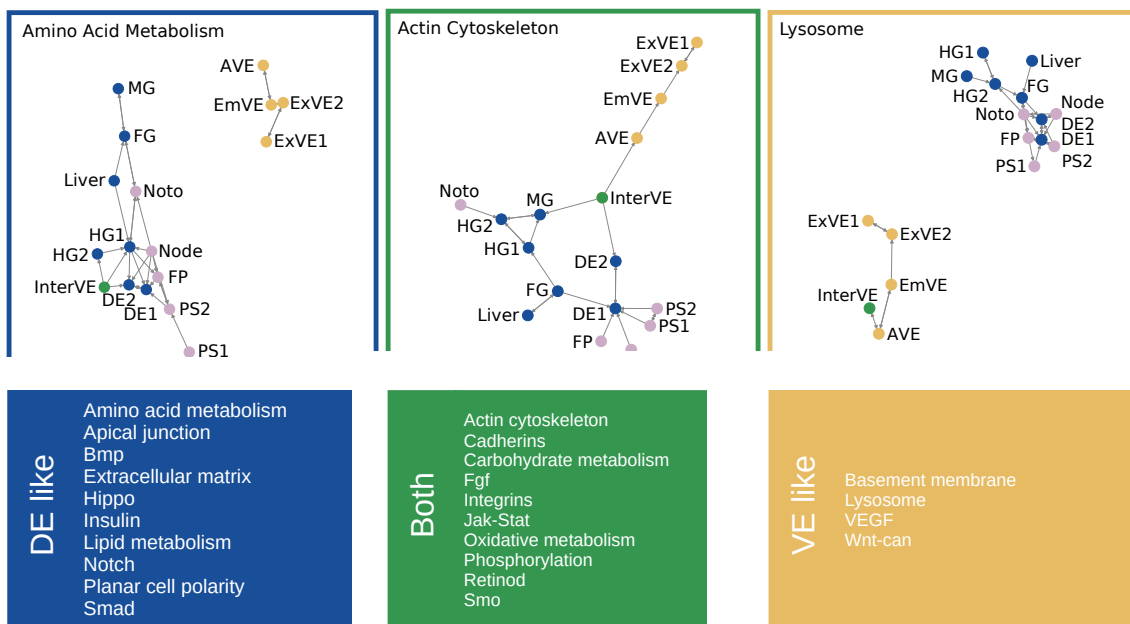


Figure 3.2: CAT projections on functionally related subset of genes. Depending on the functional projection, InterVE will be more similar to the DE lineages (shown in blue), the VE lineages (shown in yellow) or somewhere in between (shown in green). One CAT nearest neighbour graph is shown for each category. Figure modified from [Rothová et al., 2022].

3.2.2 Determining the cell-types made using *in vitro* differentiation protocols

To mimic *in vivo* conditions and guide the development of *in vitro* grown mouse embryonic stem cells (ESC), a mixture of chemical compounds is typically added [Whitten, 1957, Gonzalez et al., 2016, Linneberg-Agerholm and Brickman, 2022]. Depending on the specific protocol used, the ESC cells can be inhibited from

developing into certain lineages while other paths remain open. [Riveiro and Brickman, 2020].

It is said that imitation is the most sincere form of flattery. When studying embryo development, it might also be the most useful. If we can design a protocol for *in vitro* culturing of ESC in such a way that they will grow and differentiate to exactly mimic cell populations found *in vivo*, it implies that our understanding of the mechanics that governs the differentiation is sound. The opposite is likewise true. If we can show that a culture of ESC grown *in vitro* does not reflect the expected *in vivo* counterpart, we know we are missing something; a discrepancy that can be used to pinpoint a lapse in our understanding.

As part of the Rothova2022 dataset, we tested 4 established *in vitro* protocols, sequencing the cells at different time-points along the differentiation. We tested two 2D protocols (naming these 2D-ESC and 2D-PI3Ki) based on the same starting population of an ESC-derived cell line with the same formulation of culturing media with and without phosphoinositide 3-kinase inhibitor. We furthermore tested a 3D protocol (3D-ESC) using the same starting population, and lastly, we tried an alternate 3D protocol (3D-AChir) using a different starting cell-line. For more information about the exact protocols see [Rothová et al., 2022], particularly supplementary Data 8.

Two commonly used ways to check whether the *in vitro* samples match *in vivo* cell types include correlation between a subset of genes [Krenkel et al., 2019, Yan et al., 2013, Kim et al., 2020] or co-localisation on an embedding map [Chen et al., 2017, Tyser et al., 2021]. As mentioned above, using a limited set of genes for identification/comparison is sensitive to the specific selection of genes (see section 3.2.1), and relying on co-localisation on an embedding map is risky for a variety of reasons (see section 2.1.2.3). But since we have already identified the populations of the *in vivo* part of the Rothova2022 (and these cells were sequenced using the same technology as the *in vitro* part), we can use CAT to find the closest corresponding cell-types for the clusters found *in vitro* and map these onto the *in vivo* cell-types.

CAT alignments for the clusters from these 4 protocols can be seen in Figure 5 in [Rothová et al., 2022]. The data that was used to generate these figures can be seen in Table 3.3 for the 2D-ESC protocol, Table 3.4 for the 2D-PI3Ki protocol and supplementary Table 6.2 for 3D-ESC and 3D-AChir.

D4a (171 cells)		D4b (128 cells)		D4-D6a (316 cells)		D4-D6b (92 cells)	
DE2	18,17 ± 0,17	PS2	17,11 ± 0,27	DE2	14,87 ± 0,22	Foregut	22,64 ± 0,46
PS2	19,10 ± 0,24	DE2	18,95 ± 0,18	DE1	14,98 ± 0,21	DE2	22,72 ± 0,53
DE1	19,53 ± 0,21	DE1	19,50 ± 0,22	Foregut	17,74 ± 0,19	Hindgut2	23,10 ± 0,47
NP	20,12 ± 0,19	PS1	19,66 ± 0,29	Hindgut2	18,52 ± 0,22	NP	23,19 ± 0,49
Foregut	20,48 ± 0,21	NP	20,10 ± 0,21	NP	19,07 ± 0,23	Hindgut1	23,41 ± 0,49
Hindgut2	20,54 ± 0,23	Hindgut1	21,64 ± 0,26	PS2	19,13 ± 0,25	DE1	23,47 ± 0,53
PS1	20,67 ± 0,26	Foregut	21,78 ± 0,22	Hindgut1	19,19 ± 0,21	PS2	23,67 ± 0,54
Hindgut1	21,00 ± 0,24	Hindgut2	22,02 ± 0,25	PS1	19,90 ± 0,30	Notochord	24,74 ± 0,48
Notochord	23,16 ± 0,29	Notochord	24,01 ± 0,30	Notochord	21,86 ± 0,28	PS1	24,93 ± 0,52
Liver	26,21 ± 0,27	Liver	26,95 ± 0,27	Midgut	24,44 ± 0,62	Liver	27,47 ± 0,43
Midgut	27,33 ± 0,58	Midgut	28,65 ± 0,55	AVE	24,84 ± 0,28	Midgut	28,63 ± 0,65
AVE	28,92 ± 0,28	AVE	30,65 ± 0,27	InterVE	24,98 ± 0,54	InterVE	30,90 ± 0,57
InterVE	29,67 ± 0,53	InterVE	31,30 ± 0,51	Liver	25,36 ± 0,31	AVE	31,07 ± 0,40
Node	33,21 ± 0,47	Node	33,06 ± 0,48	Node	30,21 ± 0,49	Node	34,17 ± 0,52
EmVE	37,10 ± 0,52	EmVE	38,56 ± 0,51	EmVE	34,54 ± 0,52	EmVE	38,51 ± 0,55
ExVE2	48,89 ± 0,58	ExVE2	49,99 ± 0,57	ExVE2	47,31 ± 0,58	ExVE2	50,00 ± 0,59
ExVE1	61,31 ± 0,66	ExVE1	62,20 ± 0,66	ExVE1	60,44 ± 0,67	ExVE1	62,42 ± 0,66
PE	87,44 ± 1,82	PE	88,04 ± 1,82	PE	86,83 ± 1,82	PE	88,03 ± 1,80

D5-D6a (256 cells)		D5-D6b (173 cells)		D5-D6c (67 cells)		D6 (185 cells)	
DE2	20,30 ± 0,20	DE2	18,34 ± 0,31	DE2	27,56 ± 1,03	DE2	24,80 ± 0,40
DE1	20,63 ± 0,21	DE1	18,91 ± 0,31	Foregut	28,19 ± 0,88	DE1	25,64 ± 0,40
Hindgut2	23,22 ± 0,20	Hindgut2	20,41 ± 0,22	InterVE	28,87 ± 0,99	Hindgut2	25,88 ± 0,34
Foregut	23,55 ± 0,21	Foregut	20,64 ± 0,20	Hindgut2	28,90 ± 0,94	Foregut	26,66 ± 0,35
Hindgut1	23,70 ± 0,20	Hindgut1	20,90 ± 0,24	DE1	29,11 ± 1,04	Hindgut1	27,20 ± 0,35
NP	24,84 ± 0,22	NP	21,70 ± 0,26	Hindgut1	30,16 ± 0,93	Notochord	27,90 ± 0,32
PS2	25,07 ± 0,22	PS2	22,06 ± 0,29	Midgut	31,04 ± 0,91	InterVE	28,80 ± 0,52
Notochord	25,48 ± 0,24	Notochord	22,46 ± 0,27	Notochord	31,45 ± 0,90	Midgut	28,94 ± 0,56
AVE	25,79 ± 0,25	PS1	22,97 ± 0,29	AVE	32,33 ± 0,97	NP	29,00 ± 0,39
InterVE	25,80 ± 0,45	AVE	25,43 ± 0,32	NP	32,48 ± 0,95	PS2	29,15 ± 0,39
PS1	25,91 ± 0,25	Midgut	25,61 ± 0,60	Liver	33,01 ± 0,85	AVE	30,20 ± 0,39
Midgut	27,22 ± 0,57	InterVE	25,78 ± 0,54	PS2	33,26 ± 0,96	PS1	30,92 ± 0,43
Liver	28,75 ± 0,29	Liver	26,32 ± 0,29	PS1	34,70 ± 0,98	Liver	32,18 ± 0,39
Node	31,98 ± 0,41	Node	30,52 ± 0,49	Node	36,82 ± 0,93	Node	34,68 ± 0,42
EmVE	35,26 ± 0,47	EmVE	34,39 ± 0,56	EmVE	37,68 ± 0,91	EmVE	38,38 ± 0,50
ExVE2	47,59 ± 0,56	ExVE2	46,86 ± 0,61	ExVE2	49,05 ± 0,86	ExVE2	50,06 ± 0,57
ExVE1	60,76 ± 0,65	ExVE1	60,06 ± 0,69	ExVE1	61,58 ± 0,85	ExVE1	62,59 ± 0,65
PE	86,87 ± 1,81	PE	87,00 ± 1,81	PE	89,24 ± 1,79	PE	87,29 ± 1,79

Table 3.3: CAT distance table from aligning the 2D-ESC *in vitro* protocol to the *in vivo* cell-types within Rothova2022. Green indicated the nearest neighbour(s). The “±” denotes the standard deviation on the distance calculated from the bootstrap. The names on top of each sub-table, e.g. D4a, is the label for a cluster from the *in vitro* experiments. The D followed by a number denotes the day along the differentiation of the cells that make up the cluster. The clusters were obtained using unsupervised clustering.

To measure the success of the differentiation, we define what targets *in vivo* cell-types that we expect the *in vitro* cells to mimic: *In vitro* clusters composed of earlier cells should ideally align to PS and/or DE, and later cells should align to any of DE, foregut, midgut or hindgut, while clusters that consist of a mix of early and late cells should align to PS, DE, foregut, midgut or hindgut (these alignments corresponds to one arm of the developmental tree in Figure 3.1).

The results of the alignments are roughly as expected: All 4 protocols produce cells along the DE lineage tree with different efficiencies. Taking 2D-ESC as an

PI3Ki-D4-D6b (111 cells)		PI3Ki-D4-D6a (156 cells)		PI3Ki-D4-D6c (129 cells)	
DE2	19,64 ± 0,25	DE2	19,80 ± 0,17	DE2	22,32 ± 0,31
DE1	21,03 ± 0,27	PS2	21,02 ± 0,25	Hindgut2	22,73 ± 0,26
Hindgut2	21,26 ± 0,23	DE1	21,05 ± 0,22	Hindgut1	22,91 ± 0,26
Hindgut1	21,60 ± 0,24	Hindgut2	21,15 ± 0,20	Foregut	22,93 ± 0,25
Foregut	21,65 ± 0,23	Foregut	21,19 ± 0,19	DE1	23,37 ± 0,31
PS2	22,18 ± 0,25	Hindgut1	21,60 ± 0,23	PS2	23,59 ± 0,35
FP	22,23 ± 0,24	FP	21,95 ± 0,21	FP	23,68 ± 0,28
Notochord	23,04 ± 0,27	Notochord	23,08 ± 0,28	Notochord	23,91 ± 0,30
PS1	23,85 ± 0,29	PS1	23,44 ± 0,26	PS1	25,89 ± 0,38
Midgut	27,07 ± 0,59	Midgut	27,42 ± 0,57	Midgut	28,28 ± 0,58
Liver	27,73 ± 0,30	Liver	27,83 ± 0,27	Liver	28,68 ± 0,33
AVE	28,29 ± 0,28	InterVE	29,69 ± 0,52	InterVE	30,58 ± 0,50
InterVE	28,29 ± 0,52	AVE	30,18 ± 0,25	AVE	31,18 ± 0,27
Node	32,35 ± 0,45	Node	33,29 ± 0,46	Node	33,45 ± 0,44
EmVE	36,61 ± 0,52	EmVE	38,23 ± 0,51	EmVE	38,92 ± 0,52
ExVE2	48,88 ± 0,60	ExVE2	50,18 ± 0,59	ExVE2	50,57 ± 0,58
ExVE1	61,76 ± 0,72	ExVE1	62,68 ± 0,72	ExVE1	63,06 ± 0,70
PE	88,10 ± 1,81	PE	88,39 ± 1,82	PE	88,76 ± 1,80

PI3Ki-D5-D6a (196 cells)		PI3Ki-D5-D6b (54 cells)		PI3Ki-D6 (90 cells)	
DE2	22,41 ± 0,18	DE2	21,67 ± 0,42	DE2	25,94 ± 0,35
Hindgut2	23,43 ± 0,19	Foregut	22,07 ± 0,39	DE1	26,67 ± 0,34
DE1	23,72 ± 0,21	DE1	### ± 0,41	Hindgut2	26,83 ± 0,35
Foregut	24,04 ± 0,18	Hindgut2	22,77 ± 0,40	Foregut	26,95 ± 0,34
Hindgut1	24,08 ± 0,20	Hindgut1	23,36 ± 0,38	Hindgut1	27,42 ± 0,34
Notochord	24,92 ± 0,25	FP	23,65 ± 0,39	Notochord	28,09 ± 0,35
FP	25,52 ± 0,21	PS2	24,04 ± 0,42	FP	29,17 ± 0,36
PS2	25,88 ± 0,23	Notochord	24,81 ± 0,41	InterVE	29,23 ± 0,49
PS1	27,48 ± 0,23	PS1	25,58 ± 0,43	AVE	29,77 ± 0,32
Midgut	27,90 ± 0,56	Midgut	27,90 ± 0,60	Midgut	29,89 ± 0,59
InterVE	27,94 ± 0,48	Liver	28,95 ± 0,39	PS2	30,03 ± 0,35
AVE	28,73 ± 0,23	AVE	29,40 ± 0,37	PS1	31,31 ± 0,36
Liver	29,63 ± 0,28	InterVE	29,42 ± 0,55	Liver	32,59 ± 0,37
Node	32,82 ± 0,40	Node	33,44 ± 0,50	Node	34,71 ± 0,42
EmVE	36,58 ± 0,50	EmVE	37,50 ± 0,54	EmVE	38,12 ± 0,51
ExVE2	48,93 ± 0,59	ExVE2	49,91 ± 0,62	ExVE2	50,33 ± 0,58
ExVE1	61,94 ± 0,71	ExVE1	62,70 ± 0,73	ExVE1	63,24 ± 0,69
PE	88,00 ± 1,80	PE	88,72 ± 1,80	PE	87,98 ± 1,77

Table 3.4: CAT distance table from aligning the 2D-PI3Ki *in vitro* protocol to the *in vivo* cell-types within Rothova2022. Green indicated the nearest neighbour(s). The “±” denotes the standard deviation on the distance calculated from the bootstrap. The names on top of each sub-table, e.g. D4a, is the label for a cluster from the *in vitro* experiments. The D followed by a number denotes the day along the differentiation of the cells that make up the cluster. The clusters were obtained using unsupervised clustering.

example, the cell from this experiment clustered into 8 distinct populations that roughly coincide with time. The alignments show that earlier stages of the *in vitro* differentiation (day 4) align to PS and DE2. As time progresses towards day 6, the cells align almost purely towards DE1 and DE2, with only the 2 smallest clusters (clusters D4-D6b and D5-D6c) additionally aligning to the gut lineages and others.

Comparing the results of the 2D-ESC alignments to those of 2D-PI3Ki, the effects of the phosphoinositide 3-kinase inhibitor become clear. All clusters from the 2D-PI3Ki experiment align to DE2, and a significantly larger percentage of the cells align to hindgut. This indicates that the inhibitor pushes the cells towards a more posterior identity. The fact that none of the 2D-PI3Ki clusters align to PS also seems to suggest that the inhibition does not delay or halt the differentiation process, which has previously been suggested [Villegas et al., 2013].

Curiously it seems like the distances of the alignments for all protocols grow larger the longer the cells differentiate *in vitro*, suggesting they are somehow diverging from the *in vivo* cell-types.

3.2.2.1 Functional differences between *in vitro* differentiations and *in vivo* cell-types

To figure out how well the *in vitro* differentiation protocols captures different biological processes compared to their *in vivo* counterpart, we run CAT using functional projections. As mentioned above, we expect the different *in vitro* clusters to align to different *in vivo* cell-types, depending on how long they have been differentiating along their respective protocols.

However, since the different *in vivo* clusters might look similar under certain functional projections, it is fine for an *in vitro* cluster to also align to other *in vivo* cell-types than the target ones, as long as a target *in vivo* cell-type also align to these other *in vivo* clusters. A visual illustration of correct and incorrect alignments is shown in Figure 3.3. There can also be cases where the alignments are only partially successful. For example; an *in vitro* cluster aligns to DE, EmVE and ExVE while DE under the same projection aligns itself to EmVE but not ExVE. In this case, we judge the alignments from the *in vitro* cluster to be only partially correct. We formalize this logic using the following similarity score:

$$S = 1 - \frac{n_{vitro \rightarrow vivo}}{n_{max}} \cdot \left(1 - \frac{n_{vivo \rightarrow vivo}}{n_{vitro \rightarrow vivo} \cdot (n_{vitro \rightarrow vivo} - 1)}\right) \quad (3.1)$$

$n_{vitro \rightarrow vivo}$ is the number of alignments for the *in vitro* cluster, corresponding to the red arrows in the figure. n_{max} is the number of *in vivo* clusters and, therefore, the highest possible number of alignments for an *in vitro* cluster. $n_{vivo \rightarrow vivo}$ is the

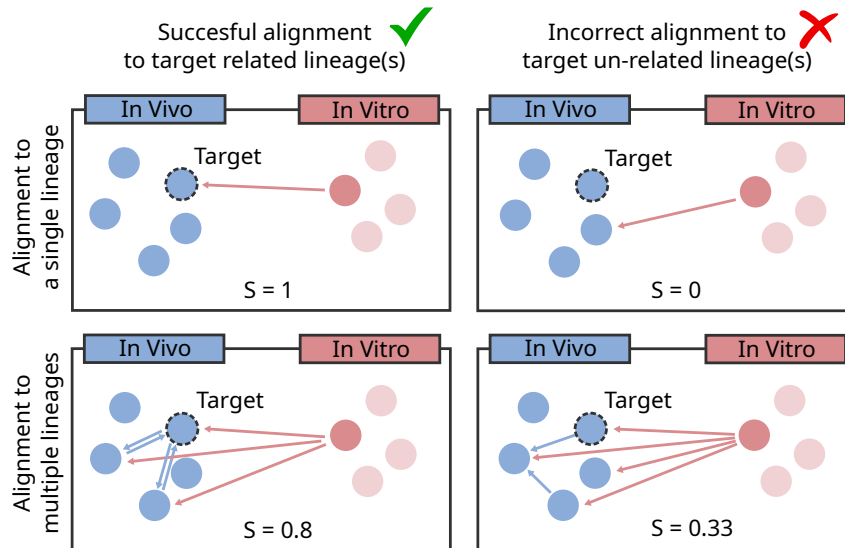


Figure 3.3: Illustration showing different scenarios for the alignment of an *in vitro* cluster (red dot) to the *in vivo* lineages (blue dots). Blue arrows symbolise alignments internally in the *in vivo* dataset, while red arrows symbolize the alignments from clusters in an *in vitro* experiment to the *in vivo* lineages. S is the similarity score in each of the 4 scenarios. Figure modified from [Rothová et al., 2022].

number of alignments between the *in vivo* clusters internally in the *in vivo* dataset (between clusters that the *in vitro* cluster aligns to), corresponding to the blue arrows in the figure. The similarity is always set to 0 if the *in vitro* cluster aligns to none of the target *in vivo* clusters. The first fraction in the formula accounts for the specificity of the alignments. In the case where an *in vitro* cluster aligns to every *in vivo* cluster in the dataset, this fraction will equal 1, and the similarity score will therefore most likely be low (depending on the second term). If the *in vitro* cluster aligns to only a single or few *in vivo* clusters, the similarity score will be high. The second fraction in the similarity formula calculates how connected the *in vivo* alignments are compared to a fully connected graph between them. If the *in vivo* targets are fully connected, the similarity score will be 1. Otherwise, if the targets are less well connected, the term inside the parenthesis will be high, and the similarity score will therefore be smaller. A more thorough derivation of the formula is discussed in the methods of [Rothová et al., 2022].

We set a heuristic value of $S = 0.75$ as the criteria that the an *in vitro* cluster successfully mimics *in vivo* for the given biological functions.

For each protocol and functional projection, we calculated the median S score across all its clusters. If the median is larger than 0.75, the protocol successfully mimics the biological process. The result for each biological function is shown in

Figure 3.4. This analysis indicates that the protocols produce cells that correctly mimic many of the biological functions of *in vivo*, but not all of them. All protocols fail with regard to Hippo signalling pathway, suggesting that this pathway could be a point of focus for future *in vitro* differentiation protocols.

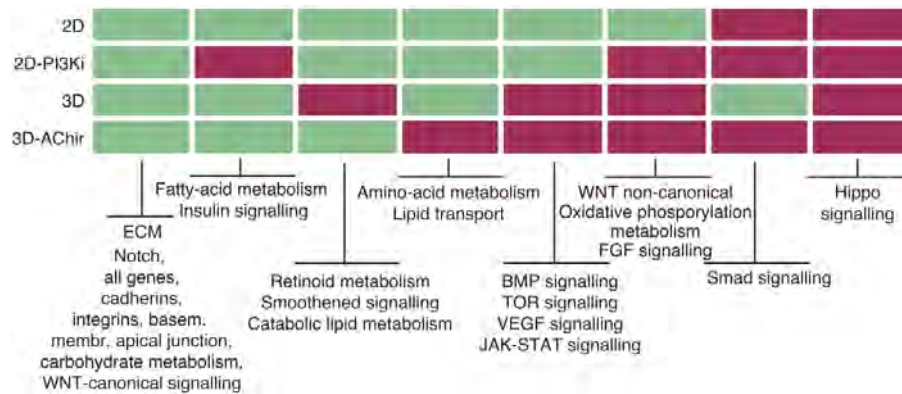


Figure 3.4: Matrix showing which *in vitro* differentiation protocol that produces cells that are similar to specific *in vivo* lineages (DE-like lineages, see text) with regard to various biological functions. Green indicates a median similarity score, S , higher than 0.75 for the clusters of each protocol, while red indicates a lower average score. Figure modified from [Rothová et al., 2022].

3.3 Conclusion

We still don't fully understand the cell-types of the early (mouse) embryo or the relationship between them. This is evident by the number of papers that keep getting published on the subject¹ [Qiu et al., 2022, Lohoff et al., 2022, Meistermann et al., 2021]. InterVE is an example of a cell-type we have only recently begun understanding [Nowotschin et al., 2019].

In addition to classical tools, we use CAT to investigate the relationship between cell-types inside the embryos at the stage of gastrulation. We further the understanding of the InterVE cell population by finding its marker genes and pinpointing its position in the developmental tree. We compare the results we find to a 3rd party dataset, ensuring the validity of our findings. Furthermore, we compare cells grown *in vitro* to their supposed *in vivo* counterpart, and find that they correctly mimic most of the biological functions as *in vivo*, but not all.

¹These 3 cited papers all use UMAP by the way. It is hard to find a recent embryo scRNAseq paper that does not.

The use of functional projections to find similarities between cell-types is a kind of dimensionality reduction, not unlike that of PCA or HVGs. The functional categories are defined apriori however, externally from the dataset itself. This means that the use of functional categories is at least reproducible, even if it is still a biased approach. In practice it turned out to be very useful.

Similarity scores, based on functional projections, allow us to see which pathways the *in vitro* protocols have trouble mimicking. We can use this information to suggest how to improve *in vitro* protocols. To the best of my knowledge, people normally use methods like DEGs to identify discrepancies between *in vivo* and *in vitro* [Ye et al., 2020, Noguchi et al., 2020, Wells and Patrizio, 2008]. CAT offers a new more systematic approach, that allows for an overview at the resolution of pathways. We can rank functional projections by their distances to find the projections where the cells mimic their expected target the worst. Having the functional projections, it is easy to identify the individual genes that contribute most to each distance. By focusing on these genes, we can quickly suggest gene perturbation targets for experimentalists to improve in *in vitro* protocols, in a way that wasn't possible before. I am happy to say that the members of the Brickman Lab² have already incorporated CAT into their workflows, even before we have officially published a separate methods paper for CAT.

²The Brickman Lab is the lab of the last author of our paper [Rothová et al., 2022], and the lab where the scRNAseq experiments were conducted. The lab is part of the "Novo Nordisk Foundation Center for Stem Cell Medicine, reNEW" at the University of Copenhagen.

Chemical reprogramming of somatic mouse cells

This chapter is based on the following 2 articles concerning reprogramming of somatic cells (differentiated body cells) into stem-cells.

Yang, Z., Xu, X., Gu, C., Li, J., Wu, Q., Ye, C., Nielsen, A. V., Mao, L., Ye, J., Bai, K., et al. (2020). Chemicals orchestrate reprogramming with hierarchical activation of master transcription factors primed by endogenous sox17 activation. *Communications biology*, 3(1):1–10.

Yang, Z., Xu, X., Gu, C., Nielsen, A. V., Chen, G., Guo, F., Tang, C., and Zhao, Y. (2022). Chemical pretreatment activated a plastic state amenable to direct lineage reprogramming. *Frontiers in cell and developmental biology*, 10

I will introduce the main concepts of stem-cell reprogramming and give a short presentation of the background for each of the two papers, along with a summary of their methods and results. Afterwards, the findings of the papers will be discussed in the context of current research. The articles are included in full length in the end of the thesis. My contribution to the work behind both articles includes dimensionality reduction analysis, clustering of the data, processing of the ATAC-seq data, handling of the single cell sequencing pipeline and various data visualisations.

4.1 Introduction: Cell reprogramming

Stem-cells are useful for research in a variety of ways. By differentiating stem-cells into other cell types (e.g muscle, retina, neurons), scientists can gain insight into genetic diseases [Halevy and Urbach, 2014], more easily test drugs [Jensen et al., 2009, Shi et al., 2017], study development (see chapters above), or perform cell replacement therapies [Davila et al., 2004, Ronaghi et al., 2010]. Unfortunately, stem-cells are not easy to acquire in large quantities. Stem-cells can painstakingly

be obtained from embryos, but this practice remains somewhat controversial, especially for human embryonic stem-cells (ESC) [Parkin, 2010], and since ESCs are specific to individual patients and can only be obtained at the embryo stage they have limited potential in clinical settings. Other avenues for obtaining stem-cells are therefore an active field of study.

Perhaps the most important breakthrough in this field came in 2006, when a team of researchers, led by Shinya Yamanaka, showed that stem-cells could be created from somatic cells (fibroblast cells at first) [Takahashi and Yamanaka, 2006]. The 2006 paper demonstrates that by infecting somatic cells with a virus engineered to contain certain transcription factors (TFs), the cells can be reprogrammed to a stem-cell-like state. TFs are proteins that bind to regions of the DNA to either promote or decrease the expression of specific genes, thereby changing the state of the cell. After this discovery of induced pluripotent stem-cells (iPSC), several advances followed: In 2013 Hou. et al. produced a protocol that could chemically induce pluripotent stem cells (CiPSC) from mouse fibroblast cells, using a set of molecules without the need for exogenous transcription factors [Hou et al., 2013]. Chemical induction has several advantages over the use of transgenic TFs; There is no risk of the TF-delivering virus integrating permanently into the genome of the host (potentially leading to tumorigenesis), it has a lower cost of reagents and it is easier to control temporally [Xie et al., 2017]. By tweaking the chemical cocktail of small molecules proposed by Hou. et al, researchers have since improved the "fibroblast to CiPSC" conversion rate a 1.000 fold over the original protocol (which was comparable to the TF-mediated method) [Zhao et al., 2015]. I will refer to this improved protocol as the Zhao protocol. A schematic of the molecules used for CiPSC reprogramming in the Zhao protocol is shown on Figure 4.1.

Compared to induction via exogenous transcription factors, chemical induction has one big drawback though; the mechanism behind its gene regulation is not as well understood [Cao et al., 2018]. There exist databases for TFs and their respective binding sites on the DNA, describing which genes they activate or inhibit, and for TFs not already in these databases, established experiments like ChIP-Seq can reveal this [Sandelin et al., 2004, Lachmann et al., 2010, Kulakovskiy et al., 2016]. As a consequence, it is relatively easy to explain what TFs are changing at a genetic level to facilitate the conversion of cell-type, as well as pick out candidate TFs for conversion of new somatic cell-types to iPSC [Deng et al., 2021].

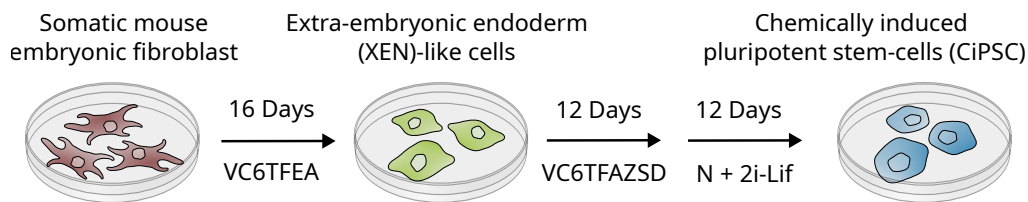


Figure 4.1: Schematic showing the original Zhao protocol for chemical reprogramming of fibroblast cells into pluripotent stem-cells via the XEN-like state. The protocol consists of 3 stages of differing lengths with different small molecules. The molecules are written as abbreviations: VPA (V), CHIR99021 (C), E-616542 (6), Tranylcypromine (T), Forskolin (F), AM580 (A), EPZ004777 (E), DZNep (Z), SGC0946 (S), 5-aza-dC (D), N2B27 (N). In the first stage (16 days), fibroblast cells are converted into an intermediate XEN-like cell-type. Once the cells are in the XEN-like stage they are then pushed to become CiPSC over the course of the last two stages.

Since the small molecules used in chemical induction do not generally bind directly to DNA, similar experiments are not readily available to explain the efficiency of CiPSC protocols, making it significantly harder to quantify exactly which genes the molecules regulate and how [Xie et al., 2017, Cao et al., 2018]. The molecules currently used for chemically induced reprogramming have therefore largely been found through trial and error, where researchers screen vast libraries of chemicals (up to 10.000 at a time) to find suitable ones [Shi et al., 2008, Zhu et al., 2010, Hou et al., 2013, Dai et al., 2014, Li et al., 2015]. There are many outstanding questions about the mechanisms behind CiPSC reprogramming and how the process could be optimized. These questions are the topic of the 2 articles.

4.2 Article 1 - Chemicals orchestrate reprogramming with hierarchical activation of master transcription factors primed by endogenous *sox17* activation

Background

Unlike earlier reprogramming protocols, the type of protocol pioneered by Zhao et al. (Figure 4.1) starts by pushing the somatic fibroblast cells not toward the CiPSC state but a distinct intermediate XEN-like state, from which the cells are then subsequently reprogrammed to become CiPSC. By designing the protocol to include this extra stop on that way to become pluripotent, the efficiency of

the reprogramming could be vastly improved [Zhao et al., 2015]. In the original paper outlining this popular protocol, the authors note that once the cells arrive at the XEN-like state they express 3 genes commonly associated with the XEN state, namely: Sall4, Gata4 and Sox17, at the same level as embryo-derived XEN cells. Knocking out any of these genes severely hampers the number of fibroblast cells converting to the XEN state and thereby the total number of cells converting successfully to CiPSC. We set out to further pinpoint exactly how the chemicals used in this type of protocol regulate the XEN "master-genes" (Sall4, Gata4 and Sox17) and facilitate the activation of the XEN-like state.

Methods and results

To test the regulation of XEN-related genes throughout the reprogramming process, we first subjected mouse embryonic fibroblast (MEF) cells to 20 days of VC6FEA (see caption of Figure 4.1 for chemical abbreviations). We gathered gene expression data from immunofluorescence staining (taken every day along the experiment) and single-cell sequencing (every second day) to quantify the temporal activation pattern of the XEN master-genes, confirming the results from the Zhao et al. paper, with higher precision. Combining the above with knockout experiments, we conclude that the XEN master-genes are activated in a hierarchical fashion and propose a small gene-regulatory network explaining this. It was only in the last part of the reprogramming (from MEF to XEN-like state) that fibroblast-related genes were substantially down-regulated. To investigate the specific role of the individual chemicals, we performed experiments with different combinations of them, for varying amounts of time, while gathering the gene expression of the XEN-master genes with immunostaining. We find that different chemical affects different parts of the regulatory master-gene network and work in concert to produce their desired effect. Since the regulatory network of the XEN master-genes gets activated in a hierarchical and time-dependent manner, we show that the efficiency of the protocol can be further improved by restricting some of the chemicals to specific periods of the reprogramming.

Conclusion

The paper adds to the current knowledge in the field by revealing what effects the small molecules have through the MEF to XEN-like reprogramming. The first part of the reprogramming up-regulates Sox17, priming the cells for activation of other XEN-like genes such as Gata4 and Sall4. After a period of up-regulation

of XEN-related genes, the cells begin to repress their fibroblast identity, transiting to the new XEN-like state. This "prime–specify–transit" model of reprogramming could help explain why the use of the chemicals CHIR99021, 616452 and forskolin are so common among different reprogramming protocols (e.g. fibroblast → neural stem-cells [Zhang et al., 2016]). These chemicals were crucial, particularly in the priming state, and we speculate that they might serve the same priming role in other protocols.

4.3 Article 2 - Chemical pretreatment activated a plastic state amenable to direct lineage reprogramming

Background

Fibroblast cells have successfully been reprogrammed directly into a multitude of other cell-types, e.g. adipocytes [Takeda et al., 2017], cardiomyocyte [Fu et al., 2015], neurons [Li et al., 2015], photoreceptors [Mahato et al., 2020], skeletal muscle [Bansal et al., 2019], and more. Curiously the chemicals CHIR99021, 616452 and forskolin are often used in these protocols, despite their different end goals. In "article 1", we speculated that these molecules might serve to prime the fibroblast cells, somehow making them more plastic and receptive to reprogramming, explaining why they are useful in reprogramming of different cell-types. In this article we seek to test this hypothesis.

Methods and results

We treated MEF cells with the chemicals C6FAE for 4,8,12, and 16 days before bulk-RNA-sequencing the resulting cells. This expression data revealed the up-regulation of a wide range of TFs associated with multiple lineages (not just the XEN master-gene Sox17), detectable as soon as day 4. Revisiting the gene expression data we obtained using single-cell RNA sequencing in article 1, we determined that the TFs were not up-regulated consistently in every cell but heterogeneously and seemingly at random throughout the cell population. Using ATAC-seq, a popular method for measuring chromatin accessibility (whether a gene is open to be transcribed), we found that after 4 days of treatment with C6FAE the MEF cells had a significantly more open chromatin pattern. The open places in the chromatin correlated with the TFs found using RNA seq, explaining their up-regulation. Together these results indicate that 4 days of treatment pushes the MEF cells to a primed state that could potentially be used to induce a multitude

of lineages. Encouraged by this finding, we then showed that this is indeed the case. By changing the culturing media of cells in the primed state to one favored by neuronal cells with just 3 added chemicals (CHIR99021, Forskolin, and ISX9), the MEF cells were successfully reprogrammed into neuron-like cells. The same procedure, using a different media, could also directly reprogram the primed MEF into skeletal muscle cells.

Conclusion

The results of the paper further reveal why splitting up the chemicals used in reprogramming of MEF to XEN into distinct stages worked for article 1. We show that a primed state, achievable after only 4 days of C6FAE, can be used as a stepping stone to directly induce multiple different cell-types.

Bibliography

- [Abromowitz and Stegun, 1972] Abromowitz, M. and Stegun, I. A. (1972). Handbook of mathematical functions.
- [Aggarwal et al., 2001] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.
- [Amezquita et al., 2020] Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Sonesson, C., et al. (2020). Orchestrating single-cell analysis with bioconductor. *Nature methods*, 17(2):137–145.
- [analyxcompany, 2022] analyxcompany (2022). Forceatlas2 - r. <https://github.com/analyxcompany/ForceAtlas2>.
- [Argelaguet et al., 2019] Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., et al. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491.
- [Bandler et al., 2022] Bandler, R. C., Vitali, I., Delgado, R. N., Ho, M. C., Dvoretzkova, E., Ibarra Molinas, J. S., Frazel, P. W., Mohammadkhani, M., Machold, R., Maedler, S., et al. (2022). Single-cell delineation of lineage and genetic identity in the mouse brain. *Nature*, 601(7893):404–409.
- [Bansal et al., 2019] Bansal, V., De, D., An, J., Kang, T. M., Jeong, H.-J., Kang, J.-S., and Kim, K. K. (2019). Chemical induced conversion of mouse fibroblasts and human adipose-derived stem cells into skeletal muscle-like cells. *Biomaterials*, 193:30–46.
- [Bao et al., 2011] Bao, L., He, L., Chen, J., Wu, Z., Liao, J., Rao, L., Ren, J., Li, H., Zhu, H., Qian, L., et al. (2011). Reprogramming of ovine adult fibroblasts to pluripotency via drug-inducible expression of defined factors. *Cell research*, 21(4):600–608.

- [Bardot and Hadjantonakis, 2020] Bardot, E. S. and Hadjantonakis, A.-K. (2020). Mouse gastrulation: coordination of tissue patterning, specification and diversification of cell fate. *Mechanisms of development*, 163:103617.
- [Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.
- [Belkin and Niyogi, 2001] Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14.
- [Belkina et al., 2019] Belkina, A. C., Ciccolella, C. O., Anno, R., Halpert, R., Spidlen, J., and Snyder-Cappione, J. E. (2019). Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):1–12.
- [Bellman et al., 1957] Bellman, R., Bellman, R., and Corporation, R. (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.
- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- [Bentkus, 2005] Bentkus, V. (2005). A lyapunov-type bound in rd. *Theory of Probability & Its Applications*, 49(2):311–323.
- [Bergen et al., 2020] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12):1408–1414.
- [Bertin, 1983] Bertin, J. (1983). *Semiology of graphics*. University of Wisconsin press.
- [Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- [Bodenham and Adams, 2016] Bodenham, D. A. and Adams, N. M. (2016). A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, 26(4):917–928.
- [Böhm et al., 2020] Böhm, J. N., Berens, P., and Kobak, D. (2020). Attraction-repulsion spectrum in neighbor embeddings. *arXiv preprint arXiv:2007.08902*.

- [Botchkarev et al., 1999] Botchkarev, V. A., Botchkareva, N. V., Roth, W., Nakamura, M., Chen, L.-H., Herzog, W., Lindner, G., McMahon, J. A., Peters, C., Lauster, R., et al. (1999). Noggin is a mesenchymally derived stimulator of hair-follicle induction. *Nature cell biology*, 1(3):158–164.
- [Bozkir et al., 2021] Bozkir, A. S., Tahillioglu, E., Aydos, M., and Kara, I. (2021). Catch them alive: A malware detection approach through memory forensics, manifold learning and computer vision. *Computers & Security*, 103:102166.
- [Brennecke et al., 2013] Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baving, B., Benes, V., Teichmann, S. A., Marioni, J. C., et al. (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11):1093–1095.
- [Bult et al., 2019] Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., and Richardson, J. E. (2019). Mouse genome database (mgd) 2019. *Nucleic acids research*, 47(D1):D801–D806. Mouse Genome Database (MGD) at the Mouse Genome Informatics website (URL: <http://www.informatics.jax.org>). [Data was retrieved 01/2020].
- [Butler et al., 2018] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420.
- [Cao et al., 2019] Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502.
- [Cao et al., 2018] Cao, S., Yu, S., Li, D., Ye, J., Yang, X., Li, C., Wang, X., Mai, Y., Qin, Y., Wu, J., et al. (2018). Chromatin accessibility dynamics during chemical induction of pluripotency. *Cell stem cell*, 22(4):529–542.
- [Castaño and López, 2005] Castaño, M. and López, B. (2005). Distribution of a sum of weighted central chi squared variables. *Communications in Statistics Theory and Methods*, 34:515.
- [Chambers et al., 2003] Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, 113(5):643–655.
- [Chari et al., 2021] Chari, T., Banerjee, J., and Pachter, L. (2021). The specious art of single-cell genomics. *bioRxiv*.
- [Chen et al., 2022] Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., and Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*.

- [Chen et al., 2017] Chen, Y.-J. J., Friedman, B. A., Ha, C., Durinck, S., Liu, J., Rubenstein, J. L., Seshagiri, S., and Modrusan, Z. (2017). Single-cell rna sequencing identifies distinct mouse medial ganglionic eminence cell types. *Scientific reports*, 7(1):1–11.
- [Chippada, 2022] Chippada, B. (2022). forceatlas2 - python. <https://github.com/bhargavchippada/forceatlas2>.
- [Coenen and Pearce, 2019] Coenen, A. and Pearce, A. (2019). Understanding umap; 2019. <https://pair-code.github.io/understanding-umap>.
- [Cole et al., 2019] Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., and Yosef, N. (2019). Performance assessment and selection of normalization procedures for single-cell rna-seq. *Cell systems*, 8(4):315–328.
- [Dabney and Meyer, 2012] Dabney, J. and Meyer, M. (2012). Length and gc-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern dna sequencing libraries. *Biotechniques*, 52(2):87–94.
- [Dai et al., 2014] Dai, R., Rossello, R., Chen, C.-c., Kessler, J., Davison, I., Hochgeschwender, U., and Jarvis, E. D. (2014). Maintenance and neuronal differentiation of chicken induced pluripotent stem-like cells. *Stem cells international*, 2014.
- [Davila et al., 2004] Davila, J. C., Cezar, G. G., Thiede, M., Strom, S., Miki, T., and Trosko, J. (2004). Use and application of stem cells in toxicology. *Toxicological Sciences*, 79(2):214–223.
- [Deng et al., 2021] Deng, W., Jacobson, E. C., Collier, A. J., and Plath, K. (2021). The transcription factor code in ipsc reprogramming. *Current Opinion in Genetics & Development*, 70:89–96.
- [Ding et al., 2020] Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, M. H., Burks, T., Nguyen, L. T., et al. (2020). Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature biotechnology*, 38(6):737–746.
- [Du et al., 2020] Du, Y., Huang, Q., Arisdakessian, C., and Garmire, L. X. (2020). Evaluation of star and kallisto on single cell rna-seq data alignment. *G3: Genes, Genomes, Genetics*, 10(5):1775–1783.
- [Edgar et al., 2013] Edgar, R., Mazor, Y., Rinon, A., Blumenthal, J., Golan, Y., Buzhor, E., Livnat, I., Ben-Ari, S., Lieder, I., Shitrit, A., et al. (2013). Lifemap discovery™: the embryonic development, stem cells, and regenerative medicine research portal. *PloS one*, 8(7):e66629.
- [Efron, 1992] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

- [Eraslan et al., 2019] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14.
- [Ertöz et al., 2003] Ertöz, L., Steinbach, M., and Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 47–58. SIAM.
- [Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.
- [Franch et al., 2020] Franch, G., Maggio, V., Coviello, L., Pendesini, M., Jurman, G., and Furlanello, C. (2020). Taasrad19, a high-resolution weather radar reflectivity dataset for precipitation nowcasting. *Scientific Data*, 7(1):1–13.
- [Fu et al., 2015] Fu, Y., Huang, C., Xu, X., Gu, H., Ye, Y., Jiang, C., Qiu, Z., and Xie, X. (2015). Direct reprogramming of mouse fibroblasts into cardiomyocytes with chemical cocktails. *Cell research*, 25(9):1013–1024.
- [Gonzalez et al., 2016] Gonzalez, J. M., Morgani, S. M., Bone, R. A., Bonderup, K., Abelchian, S., Brakebusch, C., and Brickman, J. M. (2016). Embryonic stem cell culture conditions support distinct states associated with different developmental stages and potency. *Stem cell reports*, 7(2):177–191.
- [Haghverdi et al., 2018] Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427.
- [Halevy and Urbach, 2014] Halevy, T. and Urbach, A. (2014). Comparing esc and ipsc—based models for human genetic disorders. *Journal of clinical medicine*, 3(4):1146–1162.
- [Han et al., 2022] Han, L., Wei, X., Liu, C., Volpe, G., Zhuang, Z., Zou, X., Wang, Z., Pan, T., Yuan, Y., Zhang, X., et al. (2022). Cell transcriptomic atlas of the non-human primate macaca fascicularis. *Nature*, 604(7907):723–731.
- [Hao et al., 2021] Hao, Y., Hao, S., Andersen-Nissen, E., III, W. M. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*.
- [Herrmann, 1991] Herrmann, B. G. (1991). Expression pattern of the brachyury gene in whole-mount twis/twis mutant embryos. *Development*, 113(3):913–917.

- [Hogan and Zaret, 2002] Hogan, B. L. and Zaret, K. S. (2002). 15 - development of the endoderm and its tissue derivatives. In Rossant, J. and Tam, P. P., editors, *Mouse Development*, pages 301–330. Academic Press, San Diego.
- [Hou et al., 2013] Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., et al. (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science*, 341(6146):651–654.
- [Huang et al., 2022] Huang, H., Wang, Y., Rudin, C., and Browne, E. P. (2022). Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications biology*, 5(1):1–11.
- [Imhof, 1961] Imhof, J.-P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426.
- [Irie et al., 2015] Irie, N., Weinberger, L., Tang, W. W., Kobayashi, T., Viukov, S., Manor, Y. S., Dietmann, S., Hanna, J. H., and Surani, M. A. (2015). Sox17 is a critical specifier of human primordial germ cell fate. *Cell*, 160(1-2):253–268.
- [Jacomy et al., 2014] Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679.
- [Jaitin et al., 2014] Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.
- [Jensen et al., 2009] Jensen, J., Hyllner, J., and Björquist, P. (2009). Human embryonic stem cell technologies and drug discovery. *Journal of cellular physiology*, 219(3):513–519.
- [Jespersen et al., 2020] Jespersen, C. K., Severin, J. B., Steinhardt, C. L., Vinther, J., Fynbo, J. P., Selsing, J., and Watson, D. (2020). An unambiguous separation of gamma-ray bursts into two classes from prompt emission alone. *The Astrophysical Journal Letters*, 896(2):L20.
- [Jiang, 2020] Jiang, W. (2020). Mnist-mix: a multi-language handwritten digit recognition dataset. *IOP SciNotes*, 1(2):025002.
- [Johnstone and Lu, 2009] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.
- [Kadur Lakshminarasimha Murthy et al., 2022] Kadur Lakshminarasimha Murthy, P., Sontake, V., Tata, A., Kobayashi, Y., Macadlo, L., Okuda, K., Conchola, A. S., Nakano, S., Gregory, S., Miller,

- L. A., et al. (2022). Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature*, 604(7904):111–119.
- [Kim et al., 2020] Kim, I. S., Wu, J., Rahme, G. J., Battaglia, S., Dixit, A., Gaskell, E., Chen, H., Pinello, L., and Bernstein, B. E. (2020). Parallel single-cell rna-seq and genetic recording reveals lineage decisions in developing embryoid bodies. *Cell reports*, 33(1):108222.
- [Kim et al., 2019] Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H., and Yang, P. (2019). Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326.
- [Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Kiselev et al., 2017] Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., et al. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486.
- [Kobak and Berens, 2019] Kobak, D. and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14.
- [Kobak et al., 2019] Kobak, D., Linderman, G., Steinerberger, S., Kluger, Y., and Berens, P. (2019). Heavy-tailed kernels reveal a finer cluster structure in t-sne visualisations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 124–139. Springer.
- [Kobak and Linderman, 2019] Kobak, D. and Linderman, G. C. (2019). Umap does not preserve global structure any better than t-sne when using the same initialization. *BioRxiv*.
- [Kobak and Linderman, 2021] Kobak, D. and Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-sne and umap. *Nature biotechnology*, 39(2):156–157.
- [Konstantinides et al., 2022] Konstantinides, N., Holguera, I., Rossi, A. M., Escobar, A., Dudragne, L., Chen, Y.-C., Tran, T. N., Martínez Jaimes, A. M., Özel, M. N., Simon, F., et al. (2022). A complete temporal transcription factor series in the fly visual system. *Nature*, 604(7905):316–322.
- [Krenkel et al., 2019] Krenkel, O., Hundertmark, J., Ritz, T. P., Weiskirchen, R., and Tacke, F. (2019). Single cell rna sequencing identifies subsets of hepatic stellate cells and myofibroblasts in liver fibrosis. *Cells*, 8(5):503.
- [Krivanek et al., 2020] Krivanek, J., Soldatov, R. A., Kastriti, M. E., Chontorotzea, T., Herdina, A. N., Petersen, J., Szarowska, B., Landova, M., Matejova, V. K., Holla, L. I., et al. (2020). Dental

cell type atlas reveals stem and differentiated cell types in mouse and human teeth. *Nature communications*, 11(1):1–18.

[Krupke et al., 2017] Krupke, D. M., Begley, D. A., Sundberg, J. P., Richardson, J. E., Neuhauser, S. B., and Bult, C. J. (2017). The mouse tumor biology database: a comprehensive resource for mouse models of human cancer. *Cancer research*, 77(21):e67–e70. [Data was retrieved 01/2020].

[Kulakovskiy et al., 2016] Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-Alawi, W., Bajic, V. B., Medvedeva, Y. A., Kolpakov, F. A., et al. (2016). Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research*, 44(D1):D116–D125.

[Kwon et al., 2008] Kwon, G. S., Viotti, M., and Hadjantonakis, A.-K. (2008). The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Developmental cell*, 15(4):509–520.

[Lachmann et al., 2010] Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., and Ma’ayan, A. (2010). Chea: transcription factor regulation inferred from integrating genome-wide chip-x experiments. *Bioinformatics*, 26(19):2438–2444.

[Lähnemann et al., 2020] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Levine et al., 2015] Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. (2015). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197.

[Li et al., 2015] Li, X., Zuo, X., Jing, J., Ma, Y., Wang, J., Liu, D., Zhu, J., Du, X., Xiong, L., Du, Y., et al. (2015). Small-molecule-driven direct reprogramming of mouse fibroblasts into functional neurons. *Cell stem cell*, 17(2):195–203.

[Linderman and Steinerberger, 2019] Linderman, G. C. and Steinerberger, S. (2019). Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332.

[Linneberg-Agerholm and Brickman, 2022] Linneberg-Agerholm, M. and Brickman, J. M. (2022). Differentiation and expansion of human extra-embryonic endoderm cell lines from naive pluripotent stem cells. In *Human Naïve Pluripotent Stem Cells*, pages 105–116. Springer.

- [Liu et al., 2016] Liu, H., Li, P., Zhu, M., Wang, X., Lu, J., and Yu, T. (2016). Nonlinear network reconstruction from gene expression data using marginal dependencies measured by dcol. *PLoS one*, 11(7):e0158247.
- [Liu et al., 2012] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012(251364):251364.
- [Lohoff et al., 2022] Lohoff, T., Ghazanfar, S., Missarova, A., Koulena, N., Pierson, N., Griffiths, J., Bardot, E., Eng, C.-H., Tyser, R., Argelaguet, R., et al. (2022). Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nature biotechnology*, 40(1):74–85.
- [Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- [Luecken and Theis, 2019] Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746.
- [Lun et al., 2016] Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5.
- [Mahato et al., 2020] Mahato, B., Kaya, K. D., Fan, Y., Sumien, N., Shetty, R. A., Zhang, W., Davis, D., Mock, T., Batabyal, S., Ni, A., et al. (2020). Pharmacologic fibroblast reprogramming into photoreceptors restores vision. *Nature*, 581(7806):83–88.
- [McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- [Meistermann et al., 2021] Meistermann, D., Bruneau, A., Loubersac, S., Reignier, A., Firmin, J., François-Campion, V., Kilens, S., Lelièvre, Y., Lammers, J., Feyeux, M., et al. (2021). Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. *Cell stem cell*, 28(9):1625–1640.
- [Melenhorst et al., 2022] Melenhorst, J. J., Chen, G. M., Wang, M., Porter, D. L., Chen, C., Collins, M. A., Gao, P., Bandyopadhyay, S., Sun, H., Zhao, Z., et al. (2022). Decade-long leukaemia remissions with persistence of cd4+ car t cells. *Nature*, 602(7897):503–509.
- [Motenko et al., 2015] Motenko, H., Neuhauser, S. B., O’keefe, M., and Richardson, J. E. (2015). Mousemine: a new data warehouse for mgi. *Mammalian Genome*, 26(7):325–330.
- [Nakanishi et al., 2014] Nakanishi, N., Sogabe, S., and Degnan, B. M. (2014). Evolutionary origin of gastrulation: insights from sponge development. *BMC biology*, 12(1):1–9.

- [Nissen et al., 2016] Nissen, S. B., Magidson, T., Gross, K., and Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *Elife*, 5:e21451.
- [Noguchi et al., 2020] Noguchi, T., Aizawa, T., Munakata, Y., and Iwata, H. (2020). Comparison of gene expression and mitochondria number between bovine blastocysts obtained in vitro and in vivo. *Journal of Reproduction and Development*, 66(1):35–39.
- [Nowotschin et al., 2019] Nowotschin, S., Setty, M., Kuo, Y.-Y., Liu, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., Boutet, S. C., et al. (2019). The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*, 569(7756):361–367.
- [Nowotschin et al., 2018] Nowotschin, S., Setty, M., Kuo, Y.-Y., Lui, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., Boutet, S. C., et al. (2018). Charting the emergent organotypic landscape of the mammalian gut endoderm at single-cell resolution. *BioRxiv*, page 471078.
- [Okamoto et al., 1990] Okamoto, K., Okazawa, H., Okuda, A., Sakai, M., Muramatsu, M., and Hamada, H. (1990). A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells. *cell*, 60(3):461–472.
- [O’rahilly and Müller, 2010] O’rahilly, R. and Müller, F. (2010). Developmental stages in human embryos: revised and new measurements. *Cells Tissues Organs*, 192(2):73–84.
- [Oskolkov, 2019] Oskolkov, N. (2019). How exactly umap works. <https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668>.
- [Parkin, 2010] Parkin, A. (2010). George w. bush and the stem cell research funding ban.
- [Pedregosa et al., 2011a] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011a). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- [Pedregosa et al., 2011b] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [Pijuan-Sala et al., 2019] Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C., Ho, D. L. L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495.
- [Poličar et al., 2019] Poličar, P. G., Stražar, M., and Zupan, B. (2019). opentsne: a modular python library for t-sne dimensionality reduction and embedding. *BioRxiv*, page 731877.

- [Pruitt, 1994] Pruitt, S. C. (1994). Primitive streak mesoderm-like cell lines expressing pax-3 and hox gene autoinducing activities. *Development*, 120(1):37–47.
- [Qiu et al., 2022] Qiu, C., Cao, J., Martin, B. K., Li, T., Welsh, I. C., Srivatsan, S., Huang, X., Calderon, D., Noble, W. S., Disteche, C. M., et al. (2022). Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nature genetics*, 54(3):328–341.
- [Riveiro and Brickman, 2020] Riveiro, A. R. and Brickman, J. M. (2020). From pluripotency to totipotency: an experimentalist’s guide to cellular potency. *Development*, 147(16):dev189845.
- [Ronaghi et al., 2010] Ronaghi, M., Erceg, S., Moreno-Manzano, V., and Stojkovic, M. (2010). Challenges of stem cell therapy for spinal cord injury: human embryonic stem cells, endogenous neural stem cells, or induced pluripotent stem cells? *Stem cells*, 28(1):93–99.
- [Rothová et al., 2022] Rothová, M. M., Nielsen, A. V., Proks, M., Wong, Y. F., Riveiro, A. R., Linneberg-Agerholm, M., David, E., Amit, I., Trusina, A., and Brickman, J. M. (2022). Identification of the central intermediate in the extra-embryonic to embryonic endoderm transition through single-cell transcriptomics. *Nature Cell Biology*, pages 1–12.
- [Saiz and Plusa, 2013] Saiz, N. and Plusa, B. (2013). Early cell fate decisions in the mouse embryo. *Reproduction*, 145(3):R65–R80.
- [Sandelin et al., 2004] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94.
- [Scheibner et al., 2021] Scheibner, K., Schirge, S., Burtscher, I., Büttner, M., Sterr, M., Yang, D., Böttcher, A., Irmeler, M., Beckers, J., Cernilogar, F. M., et al. (2021). Epithelial cell plasticity drives endoderm formation during gastrulation. *Nature Cell Biology*, 23(7):692–703.
- [Schöler, 1991] Schöler, H. R. (1991). Octamania: the pou factors in murine development. *Trends in genetics: TIG*, 7(10):323–329.
- [Shi et al., 2008] Shi, Y., Desponts, C., Do, J. T., Hahm, H. S., Schöler, H. R., and Ding, S. (2008). Induction of pluripotent stem cells from mouse embryonic fibroblasts by oct4 and klf4 with small-molecule compounds. *Cell stem cell*, 3(5):568–574.
- [Shi et al., 2017] Shi, Y., Inoue, H., Wu, J. C., and Yamanaka, S. (2017). Induced pluripotent stem cell technology: a decade of progress. *Nature reviews Drug discovery*, 16(2):115–130.
- [Shirkhorshidi et al., 2015] Shirkhorshidi, A. S., Aghabozorgi, S., and Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12):e0144059.

- [Smets et al., 2019] Smets, T., Verbeeck, N., Claesen, M., Asperger, A., Griffioen, G., Tousseyn, T., Waelput, W., Waelkens, E., and De Moor, B. (2019). Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. *Analytical chemistry*, 91(9):5706–5714.
- [Smith et al., 2019] Smith, C. M., Hayamizu, T. F., Finger, J. H., Bello, S. M., McCright, I. J., Xu, J., Baldarelli, R. M., Beal, J. S., Campbell, J., Corbani, L. E., et al. (2019). The mouse gene expression database (gxd): 2019 update. *Nucleic acids research*, 47(D1):D774–D779. Mouse Genome Informatics Web Site (URL: <http://www.informatics.jax.org>). [Data was retrieved 01/2020].
- [Song et al., 2017] Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. (2017). Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*.
- [Srinivas, 2006] Srinivas, S. (2006). The anterior visceral endoderm—turning heads. *Genesis*, 44(11):565–572.
- [Student, 1908] Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.
- [Sumithra and Surendran, 2015] Sumithra, V. and Surendran, S. (2015). A review of various linear and non linear dimensionality reduction techniques. *Int J Comput Sci Inf Technol*, 6:2354–2360.
- [Takahashi and Yamanaka, 2006] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676.
- [Takeda et al., 2017] Takeda, Y., Harada, Y., Yoshikawa, T., and Dai, P. (2017). Direct conversion of human fibroblasts to brown adipocytes by small chemical compounds. *Scientific reports*, 7(1):1–11.
- [Tam et al., 1993] Tam, P. P., Williams, E. A., and Chan, W. (1993). Gastrulation in the mouse embryo: ultrastructural and molecular aspects of germ layer morphogenesis. *Microscopy research and technique*, 26(4):301–328.
- [Tenenbaum et al., 2000] Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.
- [Traag et al., 2019] Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

- [Tyser et al., 2021] Tyser, R. C., Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A., and Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, 600(7888):285–289.
- [Udell et al., 2016] Udell, M., Horn, C., Zadeh, R., Boyd, S., et al. (2016). Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118.
- [UMAP, 2022] UMAP (2022). Umap benchmarking - umap 0.5 documentation. <https://umap-learn.readthedocs.io/en/latest/benchmarking.html>.
- [Väisänen et al., 2021] Väisänen, T., Heikinheimo, V., Hiippala, T., and Toivonen, T. (2021). Exploring human–nature interactions in national parks with social media photographs and computer vision. *Conservation Biology*, 35(2):424–436.
- [Van Der Maaten et al., 2009] Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.
- [Verma and Engelhardt, 2020] Verma, A. and Engelhardt, B. E. (2020). A robust nonlinear low-dimensional manifold for single cell rna-seq data. *BMC bioinformatics*, 21(1):1–15.
- [Vieth et al., 2019] Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, 10(1):1–11.
- [Villegas et al., 2013] Villegas, S. N., Rothová, M., Barrios-Llerena, M. E., Pulina, M., Hadjantonakis, A.-K., Le Bihan, T., Astrof, S., and Brickman, J. M. (2013). Pi3k/akt1 signalling specifies foregut precursors by generating regionalized extra-cellular matrix. *Elife*, 2:e00806.
- [Viotti et al., 2014] Viotti, M., Nowotschin, S., and Hadjantonakis, A.-K. (2014). Sox17 links gut endoderm morphogenesis and germ layer segregation. *Nature cell biology*, 16(12):1146–1156.
- [Waddington, 1957] Waddington, C. H. (1957). *The strategy of the genes*. Routledge.
- [Wagner et al., 2018] Wagner, F., Yan, Y., and Yanai, I. (2018). K-nearest neighbor smoothing for high-throughput single-cell rna-seq data. *BioRxiv*, page 217737.
- [Wallingford, 2021] Wallingford, J. B. (2021). Aristotle, buddhist scripture and embryology in ancient mexico: building inclusion by re-thinking what counts as the history of developmental biology. *Development*, 148(3):dev192062.
- [Wang and Deng, 2022] Wang, M. and Deng, W. (2022). Oracle-mnist: a realistic image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:2205.09442*.

- [Wang et al., 2019] Wang, Q., Tan, Y., Fang, C., Zhou, J., Wang, Y., Zhao, K., Jin, W., Wu, Y., Liu, X., Liu, X., et al. (2019). Single-cell rna-seq reveals rad51ap1 as a potent mediator of egfrviii in human glioblastomas. *Aging (Albany NY)*, 11(18):7707.
- [Wang et al., 2020] Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2020). Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *arXiv preprint arXiv:2012.04456*.
- [Wattenberg et al., 2016] Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*, 1(10):e2.
- [Weber, 2022] Weber, L. (2022). Orchestrating spatially resolved transcriptomics analysis with bioconductor. <https://lmweber.org/OSTA-book/dimensionality-reduction.html>.
- [Wells and Patrizio, 2008] Wells, D. and Patrizio, P. (2008). Gene expression profiling of human oocytes at different maturational stages and after in vitro maturation. *American journal of obstetrics and gynecology*, 198(4):455–e1.
- [Whiteley et al., 2021] Whiteley, N., Gray, A., and Rubin-Delanchy, P. (2021). Matrix factorisation and the interpretation of geodesic distance. *Advances in Neural Information Processing Systems*, 34:24–38.
- [Whitten, 1957] Whitten, W. K. (1957). Culture of tubal ova. *Nature*, 179(4569):1081–1082.
- [Wilkinson, 2012] Wilkinson, L. (2012). The grammar of graphics. In *Handbook of computational statistics*, pages 375–414. Springer.
- [Wolf et al., 2018] Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5. [version='1.4.4.post1'].
- [Wolf et al., 2019] Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1):1–9.
- [Wu et al., 2021] Wu, Z., Su, K., and Wu, H. (2021). Non-linear normalization for non-umi single cell rna-seq. *Frontiers in genetics*, 12:612670.
- [Xiao et al., 2017] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.
- [Xie et al., 2017] Xie, X., Fu, Y., and Liu, J. (2017). Chemical reprogramming and transdifferentiation. *Current Opinion in Genetics & Development*, 46:104–113.

- [Xu et al., 2021] Xu, C., Cai, L., and Gao, J. (2021). An efficient scrna-seq dropout imputation method using graph attention network. *BMC bioinformatics*, 22(1):1–18.
- [Xu and Su, 2015] Xu, C. and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980.
- [Xue et al., 2013] Xue, L., Cai, J.-Y., Ma, J., Huang, Z., Guo, M.-X., Fu, L.-Z., Shi, Y.-B., and Li, W.-X. (2013). Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis. *BMC genomics*, 14(1):1–17.
- [Yan et al., 2013] Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139.
- [Yao et al., 2021] Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R. S., Aldridge, A. I., Ament, S. A., Bartlett, A., Behrens, M. M., Van den Berge, K., et al. (2021). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110.
- [Ye et al., 2020] Ye, M., Yang, Z.-Y., Zhang, Y., Xing, Y.-X., Xie, Q.-G., Zhou, J.-H., Wang, L., Xie, W., Kee, K., and Chian, R.-C. (2020). Single-cell multiomic analysis of in vivo and in vitro matured human oocytes. *Human Reproduction*, 35(4):886–900.
- [Zhang et al., 2016] Zhang, M., Lin, Y.-H., Sun, Y. J., Zhu, S., Zheng, J., Liu, K., Cao, N., Li, K., Huang, Y., and Ding, S. (2016). Pharmacological reprogramming of fibroblasts into neural stem cells by signaling-directed transcriptional activation. *Cell stem cell*, 18(5):653–667.
- [Zhao et al., 2012] Zhao, W., Ji, X., Zhang, F., Li, L., and Ma, L. (2012). Embryonic stem cell markers. *Molecules*, 17(6):6196–6246.
- [Zhao et al., 2021] Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshov, J. H., and McShane, L. M. (2021). Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. *Journal of translational medicine*, 19(1):1–15.
- [Zhao et al., 2015] Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., Zhu, J., Meng, G., Ge, J., Yang, S., et al. (2015). A xen-like state bridges somatic cells to pluripotency during chemical reprogramming. *Cell*, 163(7):1678–1691.
- [Zhu et al., 2010] Zhu, S., Li, W., Zhou, H., Wei, W., Ambasudhan, R., Lin, T., Kim, J., Zhang, K., and Ding, S. (2010). Reprogramming of human primary somatic cells by oct4 and chemical compounds. *Cell stem cell*, 7(6).

Supplementary information

6

Table 6.1: Compilation of marker genes for early mouse embryonic cell-types

Anterior visceral endoderm	Aggf1, Amot, Celsr1, Cer1, Chst15, Cnbp, Dkk1, Emb, Eomes, Fgf8, Foxa2, Furin, Fzd5, Gpc4, Gsc, Hesx1, Hhex, Lefty1, Lhx1, Nuak, Otx2, Pkdcc, Sfrp5, Shisa2, T
Definitive endoderm	Cer1, Chst15, Cldn4, Col2a1, Cpm, Dkk1, Efna1, Emb, Eomes, Ersp1, Foxa2, Fzd5, Hesx1, Hhex, Hnf1b, Itga3, Kdm5b, Lama1, Lamb1, Lhx1, Nepn, Otx2, Pkdcc, Prdm1, Prdm4, Sdc1, Sfrp5, Shisa2, Smad3, Sox17, T, Tes, Tgif1, Tmprss2, Trh, Zic3
Embryonic visceral endoderm	Amot, Apoe, Cer1, Chst15, Dkk1, Eomes, Fgf5, Fgf8, Foxq1, Fzd5, Gpc4, Hesx1, Hhex, Lefty1, Lhx1, Pkdcc, Sfrp5, Shisa2
Extra embryonic visceral endoderm	Apln, Apoa2, Apoe, Cited1, Foxo4, Igf2, Msx1, Tbx3, Tdh, Ttr
Floor plate	Fgf8, Foxa1, Hspg2, Ntn1, Ptch1, Scrib, Tle3, Vangl2
Foregut	Apoa2, Apoe, Bambi, Cdkn1c, Gata4, Gata6, Hhex, Nepn, Phlda2, Pitx2, Sfrp5, Ttr
Gut endoderm	Afp, Ahnak, Ankrd17, Bmp4, Bmp7, Ccnd1, Ccnd2, Ccnd3, Cdh1, Cdx1, Cdx2, Chst15, Cldn4, Crabp2, Cxcr4, Cyp26a1, Dkk1, Dusp9, Efna1, Epha2, Ersp1, Fgf8, Fgfr1, Fgfr4, Fn1, Foxa1, Foxa2, Foxd4, Gsc, Gtf2ird1, Hesx1, Hhex, Hnf4a, Hoxb1, Hoxb6, Hs6st1, Ifitm3, Igf2, Isl1, Itga3, Kit, Klf5, Lemd3, Mcc, Nav2, Nck2, Nepn, Otx2, Prdm1, Prrx2, Ptpn12, Pyy, Rarb, Rhou, Ripk4, Sall1, Sfrp5, Shisa2, Snrk, Sox17, Sox2, T, Tjp1, Tle3, Tmprss2, Trh, Twsg1, Vegfa, Zmiz2
Hindgut	Cdx1, Cited1, Hoxa7, Hoxb1, Hoxb2, Hoxb6, Msx1, T, Zfp503
Midgut	Gata3, Gfpt2, Has2, Isl1, Otx2
Neural progenitors	Meis2, Ncam1, Otx2, Ptch1, Sox2
Node	Bicc1, Bmp7, Calca, Capsl, Car3, Ccdc151, Ccdc40, Cfc1, Chrd, Cobl, Cyb561, Dand5, Fam183b, Foxd4, Foxj1, Gal, Hoxb1, Ift57, Lhx1, Mlf1, Ndrgr1, Nog, Noto, Ppp1r1a, Prnp, Scara3, Slit2, Sox9, T, Ttc25, Vtn, Zic2
Node/notochord	Bicc1, Calca, Capsl, Ccdc40, Cfc1, Chrd, Cobl, Cyb561, Fam183b, Foxd4, Foxj1, Gal, Ift57, Mlf1, Ndrgr1, Nog, Noto, Ppp1r1a, Scara3, Slit2, Sox9, T
Notochord	Acvr1, Ahnak, Akap13, Anxa4, Arhgef16, Arl4a, Ate1, Atp9a, Bicc1, Bmp2, Bmp7, Calca, Car3, Cdh2, Cdh3, Cdo1, Cdx2, Clcn2, Clu, Cobl, Col2a1, Ctgf, Cthrc1, Dab2, Dlg1, Dvl2, Epha2, Etl4, Ezr, F5, Fam183b, Fgg, Fn1, Foxa1, Foxa2, Foxd4, Fzd6, Gal, Gdf1, Grik3, Gsn, Gtf2ird1, Hoxb6, Igfbp5, Ippk, Irx3, Itga3, Kdelr3, Krt7, Lhx1, Lrig1, Malat1, Mcc, Mixl1, Mmp15, Moxd1, Msx1, Mtrr, Ncam1, Nck2, Ndrgr1, Nodal, Nog, Noto, Otx2, Ppp1r1a, Prmt1, Ptch1, Rbp1, Rec8, Ripk4, Robo1, Sall3, Scara3, Sdsl, Sel1l3, Sept9, Slit2, Smad2, Smoc1, Sod1, Sox9, Sp5, Spred2, Sulf1, T, Tacc1, Tapbp, Tcf12, Tes, Tgm2, Tmem176a, Tmem176b, Tmprss2, Vangl1, Vtn, Zfp704, Zic3, Zmiz1, Zmiz2
Notochordal plate	Anxa4, Arhgef16, Bmp7, Cer1, Cfc1, Cobl, Col2a1, Epha2, Foxa2, Furin, Fzd7, Gal, Grik3, Lhx1, Nodal, Nog, Notch1, Sdc1, Shisa2, Sox9, T
Notochordal process	Acvr1, Arhgef16, Cfc1, Cobl, Foxa2, Foxd4, Lhx1, Otx2, T, Tdglf1, Zic2
Parietal endoderm	Aqp8, Cd9, Col4a1, Col4a2, Cryab, Fabp3, Lamb1, Lamc1, Serpinh1, Sparc
Primitive streak	Cer1, Chrd, Cyp26a1, Dnmt3b, Eomes, Fgf5, Fgf8, Fst, Gsc, Hhex, Lhx1, Mixl1, Otx2, T, Tdglf1, Zic2
Visceral endoderm	Afp, Aggf1, Ahnak, Amot, Apoe, Cdh1, Celsr1, Cer1, Chrd, Chst15, Cited1, Cldn6, Cnbp, Col2a1, Cttna1, Cttnb1, Dab2, Dkk1, Dll1, Dll3, Ecsit, Emb, Eomes, F3, Fgf3, Fgf8, Foxa1, Foxa2, Furin, Fzd5, Gata6, Gpc4, Gsc, Hesx1, Hhex, Ifitm3, Ihh, Ino80, Lefty1, Lhx1, Mixl1, Nodal, Notch1, Nuak1, Otx2, Pam16, Pkd2, Pkdcc, Prdm4, Racgap1, Sfrp5, Shisa2, Snai1, Stat3, T, Vps26a

6.1 CAT tables

Comparing clusters in Rothova2022 to Nowotchin2019

	AVE	Std	sigma	pvalue		EmVE	Std	sigma	pvalue
Now-emVE2	31.73	0.37	0.00	-	Now-exVE	32.62	0.29	0.00	-
Now-emVE0	32.28	0.36	1.07	0.143401	Now-emVE0	33.41	0.36	1.70	0.0448277
Now-mes5	32.89	0.31	2.40	0.00824523	Now-emVE2	34.98	0.50	4.08	2.24E-05
Now-DE	33.19	0.23	3.36	0.000394799	Now-mes5	39.29	0.51	11.35	3.64E-30
Now-VE-Lung	33.96	0.25	4.96	3.51E-07	Now-emVE1	40.03	0.45	13.90	3.24E-44
Now-DE-Lung	34.38	0.21	6.24	2.17E-10	Now-VE-Liver	40.26	0.42	14.84	4.08E-50
Now-DE-Pancreas	34.60	0.24	6.48	4.65E-11	Now-DE	40.33	0.45	14.40	2.42E-47
Now-DE-Thymus	34.71	0.22	6.92	2.29E-12	Now-VE-Lung	40.63	0.45	14.97	6.05E-51
Now-DE-Liver	34.90	0.22	7.31	1.29E-13	Now-DE-Pancreas	40.84	0.44	15.56	7.00E-55
Now-DE-Small int	34.97	0.21	7.58	1.73E-14	Now-DE-Liver	41.05	0.43	16.18	3.72E-59
Now-VE-Liver	35.00	0.22	7.59	1.57E-14	Now-Gut tube	41.19	0.44	16.30	4.74E-60
Now-Gut tube	35.45	0.26	8.22	1.05E-16	Now-DE-Lung	41.36	0.44	16.49	2.07E-61
Now-exVE	35.67	0.24	8.88	3.23E-19	Now-VE-Pancreas	41.48	0.42	17.18	2.02E-66
Now-mes4	35.69	0.23	9.11	3.94E-20	Now-DE-Small int	41.68	0.45	17.00	4.11E-65
Now-mes1	35.73	0.22	9.31	6.63E-21	Now-DE-Thymus	41.77	0.44	17.28	3.36E-67
Now-VE-Colon	35.81	0.21	9.54	6.86E-22	Now-VE-Small int	41.98	0.44	17.73	1.14E-70
Now-TE	35.96	0.56	6.32	1.27E-10	Now-TE	42.10	0.61	14.08	2.45E-45
Now-VE-Small int	35.97	0.23	9.77	7.60E-23	Now-VE-Colon	42.12	0.44	17.94	3.11E-72
Now-ExE	36.09	0.28	9.33	5.46E-21	Now-mes1	42.53	0.44	18.70	2.55E-78
Now-mes2	36.13	0.35	8.68	2.01E-18	Now-mes4	42.66	0.44	18.87	1.04E-79
Now-VE-Pancreas	36.24	0.25	10.10	2.77E-24	Now-ExE	42.69	0.46	18.37	1.09E-75
Now-VE-Thymus	36.32	0.32	9.39	3.03E-21	Now-DE-Colon	42.73	0.55	16.27	8.44E-60
Now-DE-Colon	36.36	0.42	8.26	7.12E-17	Now-DE-Thyroid	42.99	0.42	20.18	7.29E-91
Now-mes3	36.57	0.22	11.18	2.54E-29	Now-VE-Thymus	43.01	0.47	18.83	1.98E-79
Now-mes7	36.80	0.34	10.08	3.26E-24	Now-mes2	43.01	0.48	18.56	3.27E-77
Now-DE-Thyroid	36.83	0.23	11.75	3.60E-32	Now-mes3	43.15	0.44	20.03	1.49E-89
Now-emVE3	37.37	0.46	9.58	5.09E-22	Now-emVE3	43.35	0.53	17.79	4.22E-71
Now-mes0	37.52	0.23	13.24	2.65E-40	Now-mes7	43.41	0.51	18.37	1.10E-75
Now-mes6	37.75	0.49	9.83	4.28E-23	Now-mes6	44.33	0.56	18.59	2.12E-77
Now-EPI	39.11	0.31	15.26	6.74E-53	Now-mes0	44.33	0.42	22.99	2.90E-117
Now-Endothelial	39.16	0.31	15.30	3.93E-53	Now-emVE4	44.81	0.64	17.39	5.28E-68
Now-Midline1	39.17	0.33	15.07	1.29E-51	Now-Midline1	44.92	0.45	22.88	3.77E-116
Now-VE-Thyroid	39.28	1.41	5.17	1.18E-07	Now-VE	44.94	2.03	6.00	9.61E-10
Now-emVE4	39.30	0.57	11.10	5.95E-29	Now-VE-Thyroid	45.08	1.25	9.70	1.57E-22
Now-GermCells	40.61	0.30	18.59	2.12E-77	Now-Endothelial	45.44	0.44	24.12	8.47E-129
Now-VE	41.45	2.07	4.61	2.01E-06	Now-EPI	45.81	0.46	24.22	7.61E-130
Now-Midline3	41.86	0.39	18.80	4.06E-79	Now-GermCells	46.26	0.46	25.14	1.01E-139
Now-Midline0	42.32	0.54	16.23	1.45E-59	Now-Midline3	47.35	0.48	26.34	3.24E-153
Now-Midline2	42.43	0.55	16.08	1.83E-58	Now-Midline0	47.80	0.59	23.18	3.44E-119
Now-emVE1	44.96	0.50	21.28	9.57E-101	Now-Midline2	47.84	0.57	23.60	1.91E-123
Now-PrE	45.42	0.59	19.62	4.90E-86	Now-PrE	48.34	0.62	23.07	4.86E-118
Now-YsE	63.20	1.20	25.11	1.84E-139	Now-YsE	58.98	1.18	21.70	9.59E-105
Now-ICM	69.06	0.56	55.86	0	Now-ICM	72.68	0.58	61.91	0
Now-ParE	164.42	2.78	47.26	0	Now-ParE	165.77	2.76	47.89	0
Now-Blood	560.86	11.54	45.85	0	Now-Blood	561.34	11.52	45.87	0

	ExVE1	Std	sigma	pvalue		ExVE2	Std	sigma	pvalue
Now-exVE	47.60	0.65	0.00	-	Now-exVE	36.48	0.44	0.00	-
Now-emVE1	52.78	0.62	5.80	3.27E-09	Now-emVE0	41.07	0.52	6.80	5.25E-12
Now-emVE0	53.69	0.68	6.52	3.63E-11	Now-emVE1	42.53	0.46	9.59	4.42E-22
Now-emVE2	57.51	0.67	10.62	1.23E-26	Now-emVE2	45.10	0.55	12.33	3.20E-35
Now-mes5	60.45	0.70	13.52	5.69E-42	Now-mes5	49.00	0.58	17.17	2.13E-66
Now-VE-Liver	61.39	0.62	15.41	7.27E-54	Now-VE-Liver	50.07	0.49	20.74	8.35E-96
Now-TE	61.61	0.70	14.71	2.76E-49	Now-DE	50.84	0.51	21.32	3.59E-101
Now-DE	62.40	0.64	16.26	9.74E-60	Now-TE	50.91	0.63	18.92	4.18E-80
Now-DE-Liver	62.50	0.63	16.46	3.84E-61	Now-VE-Lung	51.04	0.51	21.66	2.55E-104
Now-DE-Pancreas	62.55	0.63	16.52	1.24E-61	Now-DE-Pancreas	51.14	0.51	21.97	2.75E-107
Now-ExE	62.60	0.66	16.25	1.04E-59	Now-DE-Liver	51.15	0.49	22.29	2.54E-110
Now-VE-Pancreas	62.68	0.61	16.96	8.31E-65	Now-VE-Pancreas	51.46	0.48	23.17	4.82E-119
Now-VE-Lung	62.68	0.64	16.60	3.34E-62	Now-Gut tube	51.54	0.50	22.75	6.94E-115
Now-Gut tube	62.95	0.62	17.19	1.45E-66	Now-DE-Lung	51.55	0.51	22.57	3.82E-113
Now-DE-Lung	62.97	0.64	16.95	9.22E-65	Now-ExE	51.76	0.53	22.29	2.29E-110
Now-DE-Small int	63.32	0.64	17.32	1.67E-67	Now-emVE3	51.85	0.60	20.71	1.33E-95
Now-DE-Thymus	63.33	0.64	17.35	1.07E-67	Now-DE-Small int	51.92	0.50	23.28	3.34E-120
Now-mes1	63.42	0.63	17.51	6.32E-69	Now-DE-Thymus	51.95	0.51	23.13	1.25E-118
Now-VE-Small int	63.53	0.62	17.72	1.41E-70	Now-VE	52.08	1.92	7.92	1.20E-15
Now-VE-Colon	63.55	0.63	17.72	1.54E-70	Now-VE-Small int	52.18	0.49	24.06	3.03E-128
Now-VE	63.63	1.71	8.76	9.96E-19	Now-VE-Colon	52.22	0.50	23.85	5.44E-126
Now-mes7	63.80	0.64	17.88	9.20E-72	Now-mes1	52.33	0.51	23.71	1.58E-124
Now-mes4	63.89	0.64	17.96	1.86E-72	Now-DE-Colon	52.66	0.57	22.46	5.28E-112
Now-DE-Colon	63.91	0.67	17.57	2.21E-69	Now-mes4	52.69	0.51	24.29	1.27E-130
Now-emVE3	63.91	0.72	16.90	2.29E-64	Now-emVE4	52.85	0.68	20.17	9.05E-91
Now-mes3	63.91	0.63	18.07	2.73E-73	Now-mes3	52.91	0.50	24.82	2.68E-136
Now-mes2	64.09	0.65	18.04	4.98E-73	Now-mes2	52.93	0.54	23.74	7.57E-125
Now-VE-Thymus	64.33	0.64	18.34	1.83E-75	Now-VE-Thymus	53.08	0.52	24.50	7.10E-133
Now-YsE	64.41	0.99	14.18	6.32E-46	Now-DE-Thyroid	53.09	0.49	25.41	9.68E-143
Now-DE-Thyroid	64.44	0.62	18.82	2.76E-79	Now-mes7	53.10	0.54	24.01	1.13E-127
Now-emVE4	64.92	0.74	17.58	1.67E-69	Now-mes6	54.00	0.57	24.52	4.58E-133
Now-mes6	64.99	0.66	18.77	6.95E-79	Now-mes0	54.13	0.49	26.82	9.98E-159
Now-mes0	65.24	0.62	19.65	2.79E-86	Now-Midline1	54.43	0.51	26.87	2.44E-159
Now-PrE	65.27	0.74	17.94	2.94E-72	Now-PrE	54.49	0.67	22.54	8.89E-113
Now-Endothelial	65.65	0.62	20.08	5.18E-90	Now-EPI	54.84	0.54	26.48	7.29E-155
Now-Midline1	65.67	0.63	20.00	2.50E-89	Now-VE-Thyroid	54.85	1.13	15.11	6.76E-52
Now-EPI	65.73	0.65	19.77	2.74E-87	Now-Endothelial	54.91	0.52	27.11	3.88E-162
Now-GermCells	65.75	0.61	20.36	2.01E-92	Now-GermCells	55.27	0.51	27.92	7.44E-172
Now-VE-Thyroid	65.88	1.04	14.93	1.12E-50	Now-Midline3	56.62	0.51	30.01	3.58E-198
Now-Midline2	67.52	0.66	21.54	3.44E-103	Now-Midline2	57.03	0.58	28.20	2.94E-175
Now-Midline0	67.57	0.66	21.69	1.41E-104	Now-Midline0	57.03	0.58	28.23	1.18E-175
Now-Midline3	67.60	0.63	22.20	1.64E-109	Now-YsE	59.23	1.09	19.45	1.34E-84
Now-ICM	85.47	0.63	41.82	0	Now-ICM	78.07	0.61	55.56	0
Now-ParE	173.02	2.66	45.75	0	Now-ParE	168.59	2.73	47.78	0
Now-Blood	563.24	11.48	44.83	0	Now-Blood	562.11	11.51	45.64	0

	Foregut	Std	sigma	pvalue		Hindgut1	Std	sigma	pvalue
Now-DE-Lung	22.20	0.19	0.00	-	Now-VE-Colon	23.97	0.20	0.00	-
Now-DE-Thymus	23.00	0.18	3.07	0.00108569	Now-DE-Colon	24.30	0.56	0.55	0.291005
Now-DE-Liver	23.14	0.22	3.23	0.000623116	Now-DE-Small int	24.54	0.27	1.69	0.0452877
Now-VE-Lung	23.76	0.31	4.27	9.65E-06	Now-DE-Lung	25.42	0.22	4.95	3.77E-07
Now-mes4	23.96	0.19	6.45	5.77E-11	Now-VE-Lung	25.80	0.26	5.57	1.27E-08
Now-DE-Pancreas	24.27	0.28	6.17	3.40E-10	Now-DE-Thymus	25.87	0.23	6.20	2.87E-10
Now-DE-Thyroid	24.39	0.22	7.61	1.36E-14	Now-mes4	26.70	0.24	8.73	1.30E-18
Now-mes2	25.01	0.40	6.31	1.35E-10	Now-DE-Pancreas	27.22	0.24	10.43	8.65E-26
Now-DE-Small int	25.06	0.20	10.39	1.38E-25	Now-DE-Liver	27.27	0.24	10.73	3.55E-27
Now-VE-Thymus	25.09	0.31	7.86	1.90E-15	Now-Gut tube	27.55	0.35	8.96	1.62E-19
Now-VE-Liver	25.19	0.24	9.73	1.09E-22	Now-mes2	27.68	0.43	7.77	4.05E-15
Now-Gut tube	25.54	0.28	9.91	1.92E-23	Now-VE-Thymus	28.00	0.37	9.53	7.92E-22
Now-mes0	26.14	0.21	13.89	3.41E-44	Now-DE	28.06	0.20	14.44	1.50E-47
Now-VE-Small int	26.94	0.23	15.94	1.78E-57	Now-VE-Small int	28.47	0.25	14.10	1.99E-45
Now-DE	27.11	0.17	19.16	4.29E-82	Now-DE-Thyroid	28.61	0.21	16.20	2.51E-59
Now-DE-Colon	27.16	0.51	9.03	8.51E-20	Now-mes1	28.64	0.22	15.76	2.75E-56
Now-VE-Colon	27.46	0.19	19.59	9.09E-86	Now-mes5	29.02	0.27	14.99	4.12E-51
Now-mes6	27.47	0.62	8.18	1.45E-16	Now-VE-Liver	29.05	0.24	16.38	1.32E-60
Now-mes1	27.78	0.18	21.09	5.24E-99	Now-mes0	29.17	0.25	16.48	2.37E-61
Now-VE-Pancreas	27.93	0.32	15.34	2.04E-53	Now-mes3	29.19	0.23	17.02	3.02E-65
Now-VE-Thyroid	28.17	1.74	3.41	0.000330622	Now-mes6	29.83	0.60	9.30	7.28E-21
Now-mes5	28.51	0.24	20.81	1.61E-96	Now-VE-Pancreas	30.22	0.29	17.97	1.72E-72
Now-mes3	28.55	0.19	23.85	4.72E-126	Now-mes7	30.98	0.36	17.16	2.61E-66
Now-mes7	30.01	0.32	20.81	1.93E-96	Now-Midline1	31.21	0.44	15.00	3.43E-51
Now-Midline1	30.77	0.39	19.94	9.14E-89	Now-VE-Thyroid	31.92	1.64	4.81	7.51E-07
Now-ExE	31.93	0.34	25.13	1.25E-139	Now-ExE	32.46	0.35	20.88	4.12E-97
Now-Endothelial	32.20	0.36	24.42	4.93E-132	Now-Endothelial	32.89	0.36	21.71	8.61E-105
Now-Midline3	32.93	0.45	21.94	6.01E-107	Now-Midline3	34.00	0.47	19.67	1.85E-86
Now-TE	34.02	0.60	18.85	1.53E-79	Now-TE	34.57	0.60	16.87	4.06E-64
Now-GermCells	34.23	0.34	31.14	3.19E-213	Now-GermCells	34.63	0.35	26.51	3.43E-155
Now-EPI	34.45	0.29	35.35	5.24E-274	Now-EPI	35.32	0.30	31.23	2.00E-214
Now-emVE2	35.92	0.41	30.61	4.03E-206	Now-emVE2	35.80	0.41	25.96	6.51E-149
Now-Midline0	36.37	0.61	22.17	3.10E-109	Now-Midline2	36.29	0.62	18.89	6.83E-80
Now-Midline2	36.49	0.63	21.62	6.40E-104	Now-Midline0	36.41	0.59	19.89	2.67E-88
Now-emVE0	41.49	0.40	43.13	0	Now-emVE0	41.70	0.40	39.51	0
Now-emVE3	44.47	0.41	49.58	0	Now-emVE3	45.09	0.41	46.28	0
Now-exVE	45.00	0.27	69.69	0	Now-exVE	45.12	0.26	64.69	0
Now-VE	45.33	1.80	12.77	1.23E-37	Now-VE	45.63	1.79	12.03	1.27E-33
Now-emVE4	46.71	0.51	44.91	0	Now-emVE4	47.42	0.52	42.19	0
Now-PrE	50.07	0.62	43.26	0	Now-PrE	49.96	0.60	41.02	0
Now-emVE1	55.05	0.51	60.06	0	Now-emVE1	55.12	0.51	57.03	0
Now-ICM	68.32	0.57	76.72	0	Now-ICM	68.28	0.57	73.36	0
Now-YsE	69.78	1.22	38.55	0	Now-YsE	70.19	1.21	37.84	0
Now-ParE	165.51	2.79	51.32	0	Now-ParE	165.55	2.78	50.72	0
Now-Blood	560.52	11.54	46.64	0	Now-Blood	560.55	11.54	46.48	0

	Hindgut2	Std	sigma	pvalue		InterVE	Std	sigma	pvalue
Now-DE-Small int	23.39	0.23	0.00	-	Now-emVE2	26.23	0.45	0.00	-
Now-DE-Lung	24.47	0.21	3.47	0.000258685	Now-VE-Lung	31.07	0.47	7.46	4.37E-14
Now-VE-Small int	24.75	0.24	4.16	1.58E-05	Now-DE	31.71	0.42	8.87	3.56E-19
Now-VE-Lung	24.86	0.26	4.23	1.17E-05	Now-DE-Pancreas	32.22	0.48	9.17	2.44E-20
Now-DE-Thymus	25.02	0.21	5.30	5.85E-08	Now-Gut tube	32.42	0.50	9.23	1.32E-20
Now-DE-Pancreas	25.30	0.24	5.77	3.90E-09	Now-VE-Small int	33.32	0.47	10.92	4.81E-28
Now-DE-Liver	25.37	0.21	6.40	7.68E-11	Now-DE-Small int	33.34	0.47	10.89	6.37E-28
Now-Gut tube	25.52	0.30	5.69	6.18E-09	Now-VE-Pancreas	33.35	0.48	10.88	7.01E-28
Now-VE-Colon	25.90	0.24	7.71	6.28E-15	Now-DE-Lung	33.38	0.47	11.01	1.80E-28
Now-mes4	25.91	0.22	8.02	5.25E-16	Now-VE-Colon	33.62	0.45	11.65	1.11E-31
Now-DE-Colon	26.29	0.54	4.92	4.28E-07	Now-DE-Liver	33.72	0.46	11.61	1.88E-31
Now-DE-Thyroid	26.63	0.22	10.35	2.20E-25	Now-DE-Thymus	33.85	0.45	11.93	4.03E-33
Now-mes2	26.67	0.40	7.18	3.49E-13	Now-VE-Liver	33.96	0.48	11.80	2.05E-32
Now-DE	27.01	0.19	12.21	1.43E-34	Now-DE-Thyroid	34.24	0.43	12.92	1.70E-38
Now-VE-Thymus	27.08	0.35	8.91	2.65E-19	Now-mes5	34.36	0.49	12.23	1.12E-34
Now-VE-Liver	27.12	0.23	11.60	2.09E-31	Now-VE-Thymus	34.51	0.48	12.61	8.80E-37
Now-VE-Pancreas	27.60	0.28	11.68	7.95E-32	Now-DE-Colon	34.75	0.57	11.75	3.34E-32
Now-mes1	28.37	0.21	16.05	2.84E-58	Now-mes4	35.19	0.46	13.90	3.37E-44
Now-mes0	28.40	0.22	15.75	3.55E-56	Now-mes2	35.78	0.50	14.17	7.39E-46
Now-mes5	28.79	0.26	15.84	8.28E-57	Now-mes1	35.93	0.44	15.35	1.91E-53
Now-mes6	28.91	0.58	8.88	3.37E-19	Now-Midline1	36.48	0.49	15.37	1.23E-53
Now-mes3	29.13	0.22	18.25	1.11E-74	Now-mes3	36.55	0.44	16.36	1.74E-60
Now-VE-Thyroid	30.26	1.67	4.07	2.31E-05	Now-VE-Thyroid	36.94	1.40	7.27	1.78E-13
Now-mes7	30.30	0.35	16.52	1.33E-61	Now-mes0	36.98	0.45	16.93	1.26E-64
Now-Midline1	31.06	0.38	17.32	1.74E-67	Now-mes7	37.30	0.52	16.17	4.44E-59
Now-ExE	32.25	0.33	21.97	2.69E-107	Now-mes6	37.31	0.59	14.96	6.53E-51
Now-Endothelial	32.94	0.35	22.77	4.08E-115	Now-emVE0	37.38	0.55	15.70	7.72E-56
Now-TE	33.81	0.57	16.90	2.20E-64	Now-TE	37.75	0.61	15.17	2.95E-52
Now-Midline3	34.14	0.46	21.11	3.09E-99	Now-ExE	37.76	0.47	17.69	2.52E-70
Now-GermCells	34.36	0.33	27.48	1.61E-166	Now-Endothelial	38.52	0.45	19.34	1.22E-83
Now-emVE2	34.72	0.41	24.07	2.82E-128	Now-Midline3	39.32	0.51	19.19	2.06E-82
Now-EPI	35.34	0.31	30.88	1.03E-209	Now-Midline0	40.12	0.60	18.60	1.60E-77
Now-Midline0	36.31	0.59	20.35	2.37E-92	Now-Midline2	40.13	0.61	18.38	1.01E-75
Now-Midline2	36.42	0.63	19.52	3.73E-85	Now-GermCells	40.49	0.47	21.98	2.05E-107
Now-emVE0	40.85	0.41	37.41	1.25E-306	Now-exVE	40.80	0.49	21.79	1.33E-105
Now-exVE	44.18	0.27	58.56	0	Now-EPI	41.14	0.46	23.26	6.11E-120
Now-emVE3	44.72	0.42	44.72	0	Now-emVE3	43.77	0.54	25.03	1.34E-138
Now-VE	45.25	1.82	11.95	3.17E-33	Now-VE	45.58	1.86	10.14	1.93E-24
Now-emVE4	47.08	0.53	41.25	0	Now-emVE4	46.03	0.62	25.73	2.55E-146
Now-PrE	49.48	0.62	39.73	0	Now-emVE1	48.53	0.65	28.17	7.80E-175
Now-emVE1	54.23	0.52	54.11	0	Now-PrE	49.81	0.64	30.24	4.06E-201
Now-ICM	68.05	0.57	72.42	0	Now-YsE	66.10	1.27	29.56	2.28E-192
Now-YsE	69.27	1.23	36.65	1.95E-294	Now-ICM	70.17	0.60	58.92	0
Now-ParE	165.70	2.79	50.87	0	Now-ParE	165.15	2.78	49.39	0
Now-Blood	560.67	11.54	46.56	0	Now-Blood	561.08	11.53	46.35	0

	Liver	Std	sigma	pvalue		Midgut	Std	sigma	pvalue
Now-VE-Liver	24.46	0.28	0.00	-	Now-DE-Pancreas	26.46	0.54	0.00	-
Now-DE-Liver	25.54	0.29	2.68	0.00367958	Now-Gut tube	27.63	0.53	1.55	0.0601567
Now-DE-Lung	28.80	0.27	11.23	1.40E-29	Now-VE-Pancreas	28.21	0.48	2.42	0.00784586
Now-mes4	29.41	0.27	12.71	2.44E-37	Now-VE-Lung	28.46	0.57	2.56	0.00518552
Now-DE-Thymus	29.42	0.27	12.80	8.00E-38	Now-DE-Lung	29.10	0.57	3.36	0.000393471
Now-mes2	29.97	0.42	10.88	7.10E-28	Now-DE-Liver	29.86	0.57	4.33	7.57E-06
Now-VE-Lung	30.10	0.38	11.91	4.98E-33	Now-DE-Small int	30.00	0.55	4.59	2.27E-06
Now-DE-Small int	30.50	0.26	15.79	1.95E-56	Now-DE-Thymus	30.02	0.56	4.60	2.09E-06
Now-DE-Pancreas	30.80	0.33	14.55	2.83E-48	Now-VE-Small int	30.09	0.52	4.85	6.10E-07
Now-DE-Thyroid	31.17	0.31	16.15	5.72E-59	Now-DE-Thyroid	30.36	0.55	5.07	1.98E-07
Now-mes0	31.31	0.28	17.29	2.86E-67	Now-VE-Thymus	30.78	0.56	5.55	1.47E-08
Now-Gut tube	31.48	0.34	16.02	4.42E-58	Now-VE-Liver	30.83	0.56	5.59	1.13E-08
Now-VE-Thymus	31.85	0.39	15.32	2.94E-53	Now-mes4	30.96	0.54	5.86	2.25E-09
Now-mes5	31.87	0.26	19.51	4.29E-85	Now-mes2	31.66	0.60	6.46	5.15E-11
Now-DE-Colon	31.89	0.50	12.90	2.29E-38	Now-VE-Colon	31.81	0.51	7.20	2.97E-13
Now-DE	31.92	0.23	20.49	1.46E-93	Now-DE-Colon	31.90	0.66	6.38	8.99E-11
Now-mes1	32.00	0.23	20.72	1.09E-95	Now-DE	32.29	0.52	7.80	3.08E-15
Now-VE-Colon	32.02	0.25	20.16	1.15E-90	Now-mes0	32.64	0.53	8.15	1.77E-16
Now-VE-Small int	32.16	0.28	19.45	1.50E-84	Now-VE-Thyroid	33.41	1.57	4.19	1.40E-05
Now-mes6	32.22	0.60	11.72	4.81E-32	Now-mes6	33.51	0.69	8.07	3.48E-16
Now-mes3	32.72	0.24	22.57	3.83E-113	Now-Midline1	33.69	0.56	9.29	7.43E-21
Now-VE-Pancreas	33.09	0.35	19.12	8.91E-82	Now-mes1	34.04	0.50	10.29	3.88E-25
Now-mes7	33.48	0.31	21.62	5.61E-104	Now-mes5	34.09	0.50	10.35	2.15E-25
Now-VE-Thyroid	34.18	1.55	6.19	3.07E-10	Now-mes3	34.58	0.49	11.15	3.49E-29
Now-ExE	34.78	0.34	23.41	1.74E-121	Now-mes7	35.96	0.54	12.40	1.31E-35
Now-Midline1	35.73	0.40	23.06	6.44E-118	Now-Midline3	36.21	0.59	12.24	9.22E-35
Now-Endothelial	36.36	0.38	25.07	5.02E-139	Now-Endothelial	36.37	0.53	13.07	2.30E-39
Now-TE	36.69	0.59	18.66	5.71E-78	Now-emVE2	37.09	0.55	13.73	3.34E-43
Now-GermCells	37.61	0.33	30.65	1.31E-206	Now-ExE	37.27	0.51	14.52	4.22E-48
Now-EPI	37.75	0.31	31.95	2.76E-224	Now-TE	38.26	0.64	14.14	1.07E-45
Now-Midline3	37.88	0.46	25.14	9.60E-140	Now-GermCells	39.19	0.51	17.07	1.17E-65
Now-emVE2	38.85	0.37	31.01	1.72E-211	Now-Midline2	39.65	0.67	15.28	4.97E-53
Now-Midline2	40.37	0.61	23.59	2.36E-123	Now-Midline0	39.70	0.65	15.63	2.14E-55
Now-Midline0	40.37	0.59	24.37	1.91E-131	Now-EPI	39.97	0.48	18.74	1.12E-78
Now-emVE0	42.57	0.37	39.39	0	Now-emVE0	42.94	0.52	21.97	2.65E-107
Now-exVE	44.99	0.26	54.33	0	Now-exVE	45.92	0.43	28.08	9.55E-174
Now-emVE3	47.33	0.42	45.24	0	Now-emVE3	46.60	0.50	27.38	2.57E-165
Now-VE	48.02	1.80	12.91	1.93E-38	Now-VE	46.86	1.65	11.73	4.31E-32
Now-emVE4	49.44	0.53	41.81	0	Now-emVE4	48.74	0.59	27.96	2.23E-172
Now-PrE	51.82	0.61	41.03	0	Now-PrE	51.20	0.62	30.03	1.88E-198
Now-emVE1	54.29	0.48	53.72	0	Now-emVE1	55.16	0.58	36.11	6.79E-286
Now-YsE	66.83	1.19	34.80	1.13E-265	Now-YsE	69.14	1.23	31.84	8.09E-223
Now-ICM	69.61	0.56	71.99	0	Now-ICM	70.35	0.61	54.01	0
Now-ParE	166.86	2.77	51.14	0	Now-ParE	164.01	2.78	48.61	0
Now-Blood	560.78	11.54	46.47	0	Now-Blood	561.00	11.53	46.31	0

	FP	Std	sigma	pvalue		Node	Std	sigma	pvalue
Now-DE-Lung	23.29	0.15	0.00	-	Now-Midline0	25.94	0.30	0.00	-
Now-mes4	23.48	0.20	0.80	0.213125	Now-Midline2	27.01	0.30	2.55	0.0054152
Now-DE-Thymus	23.83	0.15	2.51	0.00606712	Now-VE-Lung	33.85	0.36	16.92	1.60E-64
Now-mes2	24.70	0.46	2.92	0.00173821	Now-DE	34.33	0.42	16.38	1.23E-60
Now-DE-Liver	25.66	0.18	10.17	1.37E-24	Now-DE-Lung	34.36	0.40	17.03	2.58E-65
Now-DE-Small int	26.01	0.18	11.64	1.32E-31	Now-DE-Thymus	34.61	0.39	17.74	1.11E-70
Now-VE-Lung	26.07	0.34	7.48	3.75E-14	Now-DE-Colon	34.69	0.53	14.34	5.78E-47
Now-mes1	26.17	0.18	12.29	5.22E-35	Now-DE-Small int	34.86	0.39	18.27	7.02E-75
Now-VE-Thymus	26.71	0.36	8.78	8.49E-19	Now-DE-Pancreas	35.09	0.37	19.41	3.41E-84
Now-mes0	26.75	0.24	12.39	1.44E-35	Now-VE-Colon	35.41	0.36	20.27	1.09E-91
Now-DE	26.78	0.17	15.64	2.08E-55	Now-Midline1	35.46	0.43	18.12	1.02E-73
Now-mes3	26.83	0.18	15.23	1.06E-52	Now-mes4	35.46	0.41	18.93	3.48E-80
Now-mes6	27.41	0.68	5.93	1.49E-09	Now-mes1	35.59	0.40	19.27	5.06E-83
Now-DE-Colon	27.50	0.52	7.75	4.54E-15	Now-DE-Thyroid	35.63	0.33	21.65	3.28E-104
Now-mes5	27.72	0.24	15.58	5.32E-55	Now-DE-Liver	35.68	0.39	19.89	2.50E-88
Now-DE-Pancreas	27.92	0.30	13.61	1.69E-42	Now-VE-Thymus	35.71	0.41	19.16	4.29E-82
Now-DE-Thyroid	28.01	0.23	17.32	1.70E-67	Now-mes2	35.76	0.47	17.67	3.60E-70
Now-mes7	28.11	0.31	14.15	9.05E-46	Now-mes5	35.87	0.41	19.74	5.32E-87
Now-VE-Liver	28.22	0.21	18.90	5.75E-80	Now-Gut tube	35.92	0.40	19.96	6.29E-89
Now-VE-Colon	28.27	0.18	21.57	1.89E-103	Now-mes3	36.11	0.40	20.45	3.25E-93
Now-VE-Small int	29.50	0.22	23.57	3.52E-123	Now-mes7	36.73	0.49	18.74	1.14E-78
Now-Gut tube	29.58	0.31	18.03	5.75E-73	Now-mes0	37.05	0.39	22.76	5.25E-115
Now-ExE	31.13	0.39	18.80	3.90E-79	Now-VE-Small int	37.14	0.37	23.44	7.97E-122
Now-VE-Thyroid	31.22	1.76	4.49	3.49E-06	Now-VE-Liver	37.33	0.38	23.54	7.94E-123
Now-Midline1	31.75	0.41	19.56	1.82E-85	Now-mes6	37.40	0.54	18.53	6.00E-77
Now-VE-Pancreas	31.81	0.32	23.89	1.75E-126	Now-Midline3	37.78	0.48	21.08	5.91E-99
Now-Endothelial	32.18	0.39	21.44	3.07E-102	Now-VE-Pancreas	37.98	0.36	25.82	2.81E-147
Now-GermCells	33.37	0.32	28.17	7.62E-175	Now-VE-Thyroid	38.44	1.40	8.72	1.34E-18
Now-EPI	33.61	0.31	30.31	4.82E-202	Now-ExE	39.79	0.43	26.49	5.63E-155
Now-TE	33.75	0.63	16.07	2.20E-58	Now-GermCells	39.96	0.42	27.22	2.00E-163
Now-Midline3	34.30	0.46	22.68	3.80E-114	Now-Endothelial	40.29	0.42	27.75	9.47E-170
Now-Midline0	35.60	0.63	18.92	3.80E-80	Now-TE	40.69	0.56	23.28	3.39E-120
Now-Midline2	35.69	0.67	17.94	2.71E-72	Now-emVE2	41.14	0.46	27.79	2.67E-170
Now-emVE2	38.13	0.40	34.72	2.27E-264	Now-EPI	41.37	0.41	30.35	1.43E-202
Now-emVE0	42.46	0.40	44.75	0	Now-emVE0	46.25	0.45	37.90	0
Now-emVE3	45.80	0.42	50.89	0	Now-emVE3	49.79	0.45	44.25	0
Now-exVE	45.82	0.27	73.31	0	Now-exVE	50.15	0.32	55.15	0
Now-VE	46.36	1.84	12.51	3.45E-36	Now-VE	50.97	1.69	14.58	1.78E-48
Now-emVE4	48.16	0.53	45.14	0	Now-emVE4	52.12	0.54	42.36	0
Now-PrE	49.93	0.64	40.61	0	Now-PrE	54.70	0.62	41.66	0
Now-emVE1	55.99	0.51	60.98	0	Now-emVE1	59.08	0.51	55.96	0
Now-ICM	67.43	0.57	74.76	0	Now-ICM	71.21	0.58	69.88	0
Now-YsE	71.03	1.19	39.87	0	Now-YsE	73.87	1.16	40.06	0
Now-ParE	166.46	2.79	51.31	0	Now-ParE	167.54	2.76	51.00	0
Now-Blood	560.48	11.54	46.54	0	Now-Blood	561.19	11.53	46.41	0

	Notochord	Std	sigma	pvalue		PE	Std	sigma	pvalue
Now-Midline1	25.63	0.43	0.00	-	Now-PrE	76.11	1.69	0.00	-
Now-DE-Lung	26.99	0.24	2.74	0.003111148	Now-VE	77.47	2.95	0.40	0.345191
Now-VE-Lung	27.01	0.23	2.82	0.00237791	Now-DE-Pancreas	87.62	1.68	4.83	6.96E-07
Now-DE-Thymus	27.34	0.24	3.45	0.000280278	Now-VE-Lung	88.01	1.69	4.98	3.25E-07
Now-mes4	27.48	0.25	3.71	0.000104072	Now-Gut tube	88.17	1.67	5.08	1.89E-07
Now-DE-Small int	27.55	0.27	3.78	7.87E-05	Now-VE-Pancreas	88.44	1.67	5.19	1.04E-07
Now-DE-Colon	27.61	0.54	2.88	0.00201986	Now-DE-Lung	88.75	1.70	5.27	6.89E-08
Now-VE-Colon	27.95	0.23	4.74	1.09E-06	Now-mes5	88.82	1.72	5.28	6.52E-08
Now-mes2	28.26	0.41	4.41	5.18E-06	Now-DE	88.91	1.70	5.33	4.82E-08
Now-DE-Pancreas	28.44	0.25	5.60	1.10E-08	Now-DE-Thymus	88.97	1.70	5.36	4.08E-08
Now-DE-Liver	28.61	0.24	6.02	8.68E-10	Now-Endothelial	89.10	1.66	5.49	2.04E-08
Now-VE-Thymus	28.69	0.34	5.55	1.42E-08	Now-VE-Thymus	89.15	1.68	5.47	2.21E-08
Now-DE-Thyroid	28.94	0.21	6.87	3.25E-12	Now-mes4	89.16	1.70	5.44	2.66E-08
Now-Midline3	28.95	0.51	4.96	3.59E-07	Now-ExE	89.17	1.72	5.42	2.98E-08
Now-Gut tube	29.33	0.32	6.90	2.64E-12	Now-DE-Small int	89.26	1.70	5.49	2.00E-08
Now-mes0	29.62	0.26	7.91	1.31E-15	Now-DE-Thyroid	89.37	1.67	5.58	1.22E-08
Now-DE	29.79	0.24	8.44	1.59E-17	Now-VE-Liver	89.39	1.69	5.55	1.39E-08
Now-mes1	30.09	0.24	9.00	1.13E-19	Now-DE-Liver	89.41	1.70	5.56	1.38E-08
Now-VE-Liver	30.35	0.25	9.43	2.03E-21	Now-VE-Small int	89.45	1.69	5.59	1.16E-08
Now-mes6	30.36	0.58	6.55	2.90E-11	Now-mes2	89.47	1.70	5.57	1.25E-08
Now-VE-Small int	30.45	0.25	9.66	2.15E-22	Now-VE-Colon	89.49	1.69	5.60	1.09E-08
Now-mes3	30.77	0.25	10.24	6.54E-25	Now-mes1	89.52	1.71	5.58	1.18E-08
Now-mes5	30.85	0.29	10.08	3.31E-24	Now-TE	89.53	1.72	5.56	1.32E-08
Now-VE-Pancreas	31.64	0.29	11.51	6.10E-31	Now-Midline1	89.64	1.66	5.71	5.79E-09
Now-VE-Thyroid	32.42	1.60	4.09	2.16E-05	Now-DE-Colon	89.79	1.67	5.76	4.29E-09
Now-mes7	32.63	0.35	12.58	1.41E-36	Now-emVE2	89.81	1.66	5.78	3.84E-09
Now-Endothelial	33.16	0.37	13.30	1.20E-40	Now-mes0	89.83	1.68	5.76	4.33E-09
Now-ExE	34.49	0.35	16.01	5.46E-58	Now-mes3	89.92	1.70	5.76	4.09E-09
Now-Midline2	35.20	0.58	13.27	1.71E-40	Now-mes6	89.98	1.70	5.78	3.69E-09
Now-Midline0	35.42	0.56	13.80	1.22E-43	Now-mes7	90.53	1.70	6.01	9.09E-10
Now-GermCells	36.48	0.35	19.45	1.39E-84	Now-Midline3	90.76	1.63	6.24	2.21E-10
Now-TE	36.62	0.55	15.74	4.18E-56	Now-VE-Thyroid	90.87	1.76	6.06	6.99E-10
Now-EPI	37.37	0.31	21.98	2.02E-107	Now-emVE0	91.12	1.65	6.36	1.02E-10
Now-emVE2	37.46	0.41	19.94	8.97E-89	Now-exVE	91.75	1.64	6.63	1.63E-11
Now-emVE0	43.24	0.41	29.49	1.99E-191	Now-GermCells	91.79	1.67	6.60	1.99E-11
Now-emVE3	46.59	0.42	34.96	4.13E-268	Now-EPI	92.03	1.67	6.70	1.07E-11
Now-exVE	46.98	0.27	41.85	0	Now-emVE3	92.17	1.61	6.88	3.00E-12
Now-VE	47.17	1.71	12.23	1.11E-34	Now-emVE4	92.48	1.62	7.00	1.29E-12
Now-emVE4	48.78	0.51	34.49	5.06E-261	Now-Midline2	92.55	1.66	6.94	1.93E-12
Now-PrE	51.33	0.60	34.94	9.61E-268	Now-Midline0	92.60	1.65	6.99	1.39E-12
Now-emVE1	56.51	0.50	46.78	0	Now-emVE1	97.15	1.57	9.12	3.61E-20
Now-ICM	69.27	0.57	61.29	0	Now-ICM	105.40	1.51	12.93	1.43E-38
Now-YsE	71.37	1.19	36.03	1.39E-284	Now-YsE	105.78	1.58	12.83	5.69E-38
Now-ParE	165.75	2.78	49.81	0	Now-ParE	106.02	3.09	8.48	1.11E-17
Now-Blood	560.74	11.54	46.34	0	Now-Blood	566.75	11.41	42.53	0

	PS-e6.5	Std	sigma	pvalue		PS-e7.5	Std	sigma	pvalue
Now-mes1	25.12	0.25	0.00	-	Now-mes1	23.77	0.20	0.00	-
Now-mes3	25.69	0.22	1.71	0.0436155	Now-mes4	24.53	0.20	2.68	0.00370713
Now-mes5	25.97	0.27	2.32	0.0102676	Now-mes3	24.83	0.22	3.51	0.00022377
Now-mes7	26.14	0.39	2.21	0.0134301	Now-DE-Lung	25.00	0.17	4.63	1.83E-06
Now-DE-Lung	26.59	0.21	4.51	3.28E-06	Now-DE-Thymus	25.21	0.18	5.33	5.01E-08
Now-DE	26.75	0.21	4.96	3.48E-07	Now-mes5	25.44	0.26	5.06	2.06E-07
Now-EPI	26.89	0.30	4.50	3.33E-06	Now-mes2	25.64	0.43	3.90	4.90E-05
Now-DE-Thymus	26.89	0.21	5.42	2.95E-08	Now-DE	26.35	0.20	9.07	6.04E-20
Now-mes4	27.05	0.23	5.70	5.85E-09	Now-DE-Small int	26.38	0.20	9.23	1.40E-20
Now-DE-Liver	28.05	0.20	9.10	4.71E-20	Now-DE-Liver	26.91	0.19	11.24	1.30E-29
Now-DE-Small int	28.17	0.23	8.95	1.85E-19	Now-mes7	27.13	0.31	8.99	1.28E-19
Now-mes2	28.19	0.45	5.92	1.59E-09	Now-DE-Colon	27.15	0.54	5.87	2.15E-09
Now-ExE	28.70	0.41	7.46	4.27E-14	Now-mes0	27.35	0.24	11.45	1.21E-30
Now-VE-Lung	29.44	0.41	9.00	1.13E-19	Now-VE-Lung	27.65	0.37	9.26	1.01E-20
Now-DE-Colon	29.54	0.53	7.53	2.56E-14	Now-VE-Colon	27.79	0.18	14.81	6.38E-50
Now-VE-Colon	29.91	0.21	14.71	2.60E-49	Now-mes6	28.13	0.63	6.54	3.01E-11
Now-mes0	29.97	0.26	13.46	1.31E-41	Now-VE-Thymus	28.34	0.40	10.11	2.51E-24
Now-VE-Liver	30.14	0.24	14.47	9.52E-48	Now-VE-Liver	29.05	0.22	17.65	5.47E-70
Now-VE-Thymus	30.37	0.44	10.43	9.11E-26	Now-DE-Pancreas	29.49	0.31	15.44	4.17E-54
Now-mes6	30.56	0.65	7.79	3.33E-15	Now-ExE	29.68	0.40	13.13	1.11E-39
Now-DE-Pancreas	31.64	0.34	15.38	1.13E-53	Now-Gut tube	30.13	0.32	16.64	1.69E-62
Now-TE	31.80	0.70	9.04	8.13E-20	Now-DE-Thyroid	30.36	0.23	21.56	2.19E-103
Now-VE-Small int	32.09	0.25	19.70	1.13E-86	Now-VE-Small int	30.44	0.23	21.95	4.25E-107
Now-Gut tube	32.42	0.35	17.07	1.30E-65	Now-EPI	32.51	0.32	22.88	4.03E-116
Now-DE-Thyroid	32.52	0.25	20.72	1.19E-95	Now-GermCells	32.53	0.33	22.69	2.65E-114
Now-GermCells	32.53	0.32	18.07	2.71E-73	Now-Midline1	32.56	0.46	17.61	9.96E-70
Now-VE-Pancreas	35.01	0.33	23.67	3.34E-124	Now-Endothelial	32.72	0.39	20.26	1.35E-91
Now-VE-Thyroid	35.22	1.67	6.00	9.88E-10	Now-TE	32.95	0.66	13.31	1.04E-40
Now-Endothelial	35.36	0.40	21.84	4.96E-106	Now-VE-Pancreas	33.07	0.32	24.70	5.50E-135
Now-Midline1	35.73	0.43	21.15	1.25E-99	Now-VE-Thyroid	33.27	1.70	5.56	1.38E-08
Now-Midline0	36.52	0.68	15.63	2.12E-55	Now-Midline0	34.57	0.66	15.54	9.08E-55
Now-Midline2	36.75	0.72	15.26	6.56E-53	Now-Midline2	34.77	0.70	15.03	2.17E-51
Now-Midline3	38.04	0.47	24.37	1.61E-131	Now-Midline3	34.91	0.47	21.58	1.29E-103
Now-emVE2	39.05	0.41	29.16	3.22E-187	Now-emVE2	38.00	0.40	31.43	3.78E-217
Now-emVE0	42.27	0.40	36.01	3.16E-284	Now-emVE0	42.10	0.40	40.61	0
Now-emVE3	44.68	0.44	38.75	0	Now-emVE3	45.31	0.43	45.62	0
Now-VE	45.22	2.00	9.98	9.47E-24	Now-exVE	45.44	0.26	65.30	0
Now-exVE	45.46	0.28	54.41	0	Now-VE	45.61	1.84	11.81	1.75E-32
Now-emVE4	47.27	0.55	36.62	5.89E-294	Now-emVE4	47.68	0.54	41.45	0
Now-PrE	48.93	0.67	33.06	5.86E-240	Now-PrE	49.33	0.65	37.51	0.28E-307
Now-emVE1	56.11	0.53	53.13	0	Now-emVE1	55.77	0.52	57.20	0
Now-ICM	66.05	0.59	64.15	0	Now-ICM	67.25	0.58	70.82	0
Now-YsE	71.59	1.17	38.94	0	Now-YsE	71.08	1.18	39.47	0
Now-ParE	166.89	2.78	50.71	0	Now-ParE	166.08	2.79	50.90	0
Now-Blood	560.42	11.55	46.35	0	Now-Blood	560.47	11.54	46.48	0

	DE1	Std	sigma	pvalue		DE2	Std	sigma	pvalue
Now-DE	22.87	0.20	0.00	-	Now-DE	22.62	0.15	0.00	-
Now-DE-Lung	23.83	0.17	3.64	0.00013394	Now-DE-Lung	23.20	0.15	2.73	0.0031285
Now-DE-Thymus	23.88	0.18	3.75	8.93E-05	Now-DE-Thymus	23.66	0.16	4.71	1.22E-06
Now-mes4	24.60	0.20	6.18	3.12E-10	Now-DE-Small int	24.47	0.16	8.60	4.05E-18
Now-mes1	25.09	0.19	8.07	3.58E-16	Now-VE-Lung	24.55	0.31	5.59	1.16E-08
Now-DE-Small int	25.37	0.19	9.15	2.80E-20	Now-DE-Liver	24.57	0.18	8.48	1.14E-17
Now-DE-Liver	25.52	0.19	9.55	6.60E-22	Now-mes4	24.85	0.18	9.66	2.15E-22
Now-VE-Lung	25.60	0.32	7.17	3.63E-13	Now-mes1	25.42	0.16	13.03	3.90E-39
Now-mes5	25.64	0.25	8.72	1.41E-18	Now-mes5	25.66	0.23	11.25	1.10E-29
Now-mes2	25.82	0.44	6.12	4.56E-10	Now-mes2	25.93	0.41	7.56	2.06E-14
Now-mes3	25.99	0.19	11.34	4.21E-30	Now-mes3	26.18	0.17	15.85	7.36E-57
Now-VE-Thymus	26.67	0.36	9.18	2.18E-20	Now-DE-Pancreas	26.21	0.27	11.61	1.93E-31
Now-mes7	26.89	0.36	9.82	4.66E-23	Now-VE-Thymus	26.34	0.37	9.45	1.68E-21
Now-DE-Colon	27.10	0.53	7.43	5.53E-14	Now-VE-Liver	26.56	0.21	15.36	1.47E-53
Now-mes0	27.17	0.23	14.22	3.66E-46	Now-DE-Colon	26.58	0.53	7.20	2.99E-13
Now-DE-Pancreas	27.46	0.29	13.00	6.52E-39	Now-VE-Colon	26.71	0.17	18.40	6.50E-76
Now-VE-Colon	27.56	0.18	17.22	9.24E-67	Now-mes7	26.79	0.34	11.18	2.42E-29
Now-VE-Liver	27.73	0.22	16.33	2.87E-60	Now-VE-Small int	27.07	0.21	17.37	6.85E-68
Now-DE-Thyroid	28.00	0.23	16.79	1.44E-63	Now-Gut tube	27.24	0.29	14.09	2.31E-45
Now-Gut tube	28.12	0.30	14.48	8.33E-48	Now-DE-Thyroid	27.38	0.24	17.09	9.21E-66
Now-mes6	28.22	0.64	7.97	8.15E-16	Now-mes0	27.89	0.22	19.99	3.61E-89
Now-VE-Small int	28.57	0.23	18.79	4.43E-79	Now-mes6	28.26	0.62	8.84	5.00E-19
Now-ExE	29.60	0.37	15.89	3.45E-57	Now-VE-Pancreas	29.28	0.31	19.59	9.95E-86
Now-VE-Pancreas	30.92	0.31	21.72	6.25E-105	Now-ExE	29.29	0.35	17.42	2.75E-68
Now-Midline1	31.19	0.42	18.05	3.72E-73	Now-VE-Thyroid	30.60	1.73	4.59	2.23E-06
Now-VE-Thyroid	31.26	1.70	4.89	4.97E-07	Now-TE	31.23	0.64	13.12	1.21E-39
Now-EPI	31.68	0.31	23.80	1.64E-125	Now-Midline1	31.68	0.41	20.99	4.20E-98
Now-TE	31.97	0.64	13.52	5.61E-42	Now-GermCells	32.21	0.31	27.70	3.78E-169
Now-Endothelial	32.34	0.38	21.89	1.61E-106	Now-Endothelial	32.24	0.37	24.08	2.13E-128
Now-GermCells	32.42	0.33	24.70	5.24E-135	Now-EPI	32.30	0.31	28.01	6.99E-173
Now-Midline0	32.88	0.65	14.67	5.04E-49	Now-emVE2	32.43	0.41	22.47	4.12E-112
Now-Midline2	33.35	0.69	14.55	2.82E-48	Now-Midline3	34.58	0.46	24.68	9.55E-135
Now-Midline3	33.58	0.46	21.38	1.14E-101	Now-Midline0	34.74	0.65	18.25	1.12E-74
Now-emVE2	34.71	0.43	25.17	3.93E-140	Now-Midline2	34.86	0.68	17.62	9.39E-70
Now-emVE0	39.98	0.44	35.33	1.19E-273	Now-emVE0	38.75	0.42	36.23	1.08E-287
Now-emVE3	42.79	0.42	42.94	0	Now-exVE	42.31	0.26	65.29	0
Now-exVE	43.92	0.27	63.04	0	Now-emVE3	43.01	0.43	44.42	0
Now-VE	44.35	1.92	11.15	3.69E-29	Now-VE	43.79	1.92	11.00	2.00E-28
Now-emVE4	45.25	0.55	38.52	0	Now-emVE4	45.68	0.55	40.38	0
Now-PrE	48.49	0.65	37.82	0	Now-PrE	47.56	0.65	37.37	6.31E-306
Now-emVE1	54.33	0.52	56.22	0	Now-emVE1	52.58	0.53	54.08	0
Now-ICM	66.63	0.58	71.03	0	Now-ICM	66.59	0.57	74.16	0
Now-YsE	70.04	1.20	38.64	0	Now-YsE	68.63	1.22	37.45	3.35E-307
Now-ParE	166.09	2.79	51.20	0	Now-ParE	165.30	2.79	50.99	0
Now-Blood	560.48	11.54	46.56	0	Now-Blood	560.48	11.54	46.60	0

3D-ESC protocol				3D-AChir protocol				3D-AChir protocol			
	3D-Ach-D4	diff in sigma	sigma in pvalue		3D-D4	diff in sigma	sigma in pvalue		3D-D6	diff in sigma	sigma in pvalue
DE2	25,79 ± 0,26	0,00	0,50	DE2	19,52 ± 0,20	0,00	0,50	DE2	25,67 ± 0,38	0,00	0,50
DE1	27,10 ± 0,30	3,36	3,9E-04	DE1	20,03 ± 0,22	1,72	4,3E-02	DE1	26,92 ± 0,38	2,32	1,0E-02
Hindgut2	27,64 ± 0,26	5,06	2,1E-07	Foregut	22,46 ± 0,20	10,44	8,2E+00	Foregut	27,06 ± 0,31	2,85	2,2E-03
Foregut	27,74 ± 0,26	5,32	5,3E-08	PS1	22,52 ± 0,30	8,42	1,9E+00	Hindgut2	27,45 ± 0,34	3,50	2,3E-04
Hindgut1	28,06 ± 0,28	5,99	1,0E-09	Hindgut2	22,53 ± 0,21	10,45	7,4E+00	Hindgut1	28,74 ± 0,35	5,96	1,3E-09
NP	28,92 ± 0,29	8,15	1,9E+00	PS2	22,85 ± 0,27	9,98	9,1E+00	InterVE	28,77 ± 0,48	5,09	1,8E-07
PS2	29,52 ± 0,32	9,14	3,2E+00	Hindgut1	23,25 ± 0,24	12,09	6,1E+00	AVE	29,33 ± 0,38	6,83	4,2E-12
PS1	29,71 ± 0,34	9,30	6,9E+00	NP	23,69 ± 0,22	13,99	8,9E+00	NP	30,38 ± 0,36	9,05	7,3E+00
InterVE	30,15 ± 0,36	9,80	5,5E+00	Notochord	25,55 ± 0,29	17,13	4,4E+00	PS2	30,45 ± 0,39	8,83	5,3E+00
Notochord	30,25 ± 0,32	10,83	1,2E+00	Liver	27,58 ± 0,26	24,76	1,3E+00	Notochord	30,61 ± 0,35	9,63	2,9E+00
Liver	31,54 ± 0,32	14,15	8,9E+00	InterVE	28,23 ± 0,46	17,44	2,2E+00	PS1	30,80 ± 0,41	9,20	1,8E+00
Midgut	31,67 ± 0,53	10,02	6,2E+00	AVE	28,33 ± 0,28	25,72	3,2E+00	Liver	30,82 ± 0,31	10,52	3,5E+00
AVE	31,96 ± 0,29	###	1,5E+00	Midgut	28,39 ± 0,56	14,88	2,1E+00	Midgut	31,12 ± 0,56	8,11	2,5E+00
Node	36,22 ± 0,40	21,99	1,8E+00	Node	32,54 ± 0,44	26,84	5,8E+00	Node	36,10 ± 0,42	18,36	1,3E+00
EmVE	39,83 ± 0,50	24,85	1,4E+00	EmVE	37,35 ± 0,53	31,39	1,6E+00	EmVE	36,34 ± 0,55	16,05	2,9E+00
ExVE2	51,26 ± 0,57	41,04	0,0E+00	ExVE2	49,50 ± 0,61	46,95	0,0E+00	ExVE2	48,31 ± 0,62	31,15	2,4E+00
ExVE1	63,49 ± 0,68	52,04	0,0E+00	ExVE1	62,07 ± 0,72	57,16	0,0E+00	ExVE1	61,21 ± 0,73	43,26	0,0E+00
PE	88,95 ± 1,79	34,93	1,2E+00	PE	88,77 ± 1,82	37,73	0,0E+00	PE	88,86 ± 1,79	34,46	1,5E+00
	3D-Ach-D4+D6	diff in sigma	sigma in pvalue		3D-D4A	diff in sigma	sigma in pvalue		3D-D6A	diff in sigma	sigma in pvalue
DE2	24,09 ± 0,49	0,00	0,50	Foregut	23,56 ± 0,67	0,00	0,50	DE2	23,24 ± 0,59	0,00	0,50
DE1	24,89 ± 0,52	1,11	0,13	DE2	23,61 ± 0,81	0,04	0,48	DE1	24,11 ± 0,56	1,07	0,14
Foregut	24,93 ± 0,50	1,20	0,12	PS2	23,77 ± 0,93	0,19	0,43	Foregut	24,20 ± 0,50	1,25	0,11
Hindgut2	25,71 ± 0,47	2,38	8,6E-03	DE1	24,08 ± 0,83	0,49	0,31	Hindgut2	25,37 ± 0,54	2,66	3,9E-03
Hindgut1	26,27 ± 0,47	3,20	6,9E-04	Hindgut2	24,41 ± 0,67	0,90	0,18	Hindgut1	26,71 ± 0,54	4,35	6,8E-06
NP	26,60 ± 0,49	3,62	1,5E-04	Hindgut1	24,57 ± 0,66	1,07	0,14	PS2	27,42 ± 0,56	5,17	1,2E-07
PS2	27,32 ± 0,50	4,62	1,9E-06	NP	25,33 ± 0,78	1,72	4,3E-02	PS1	27,46 ± 0,59	5,09	1,8E-07
PS1	27,77 ± 0,52	5,13	1,4E-07	PS1	25,65 ± 0,97	1,78	3,8E-02	NP	27,60 ± 0,55	5,38	3,6E-08
Notochord	28,89 ± 0,49	6,90	2,6E-12	Notochord	26,25 ± 0,61	2,96	1,5E-03	AVE	28,55 ± 0,54	6,65	1,4E-11
Liver	29,73 ± 0,45	8,48	1,1E+00	Liver	28,22 ± 0,61	5,15	1,3E-07	InterVE	28,69 ± 0,62	6,39	8,5E-11
Midgut	30,14 ± 0,62	7,61	1,4E-14	Midgut	30,00 ± 0,71	6,61	1,9E-11	Liver	28,87 ± 0,48	7,44	4,9E-14
InterVE	30,17 ± 0,54	8,32	4,5E+00	InterVE	32,18 ± 0,65	9,23	1,3E+00	Notochord	29,31 ± 0,52	7,76	4,4E+00
AVE	30,57 ± 0,54	8,88	3,5E+00	AVE	32,94 ± 0,63	10,22	7,9E+00	Midgut	29,95 ± 0,67	7,52	2,8E-14
Node	35,62 ± 0,54	15,82	1,2E+00	Node	34,08 ± 0,63	11,40	2,2E+00	Node	35,35 ± 0,54	15,10	8,2E+00
EmVE	39,03 ± 0,61	19,13	7,2E+00	EmVE	40,24 ± 0,65	17,82	2,3E+00	EmVE	36,03 ± 0,69	14,14	1,1E+00
ExVE2	50,85 ± 0,63	33,64	2,1E+00	ExVE2	51,42 ± 0,67	29,32	3,1E+00	ExVE2	48,30 ± 0,71	27,20	2,9E+00
ExVE1	63,20 ± 0,72	44,96	0,0E+00	ExVE1	63,43 ± 0,75	39,66	0,0E+00	ExVE1	61,22 ± 0,78	38,79	0,0E+00
PE	88,91 ± 1,80	34,79	1,9E+00	PE	89,40 ± 1,81	34,15	7,2E+00	PE	88,59 ± 1,80	34,49	5,7E+00
	3D-Ach-D6	diff in sigma	sigma in pvalue								
Hindgut2	29,20 ± 0,30	0,00	0,50								
Foregut	29,44 ± 0,29	0,58	0,28								
DE2	29,49 ± 0,30	0,70	0,24								
Hindgut1	30,30 ± 0,29	2,62	4,4E-03								
DE1	30,34 ± 0,31	2,65	4,0E-03								
Notochord	31,53 ± 0,32	5,35	4,3E-08								
NP	31,79 ± 0,30	6,07	6,6E-10								
Midgut	31,79 ± 0,53	4,24	1,1E-05								
AVE	31,85 ± 0,30	6,22	2,5E-10								
InterVE	32,75 ± 0,45	6,56	2,6E-11								
PS2	33,09 ± 0,31	8,98	1,4E+00								
Liver	33,83 ± 0,33	10,36	1,8E+00								
PS1	34,50 ± 0,33	11,97	2,5E+00								
Node	37,58 ± 0,37	17,54	3,7E+00								
EmVE	39,61 ± 0,50	###	1,9E+00								
ExVE2	50,84 ± 0,59	32,62	9,4E+00								
ExVE1	63,16 ± 0,69	45,15	0,0E+00								
PE	88,96 ± 1,78	33,16	1,8E+00								

Table 6.2: CAT distance table from aligning the 3D-ESC and 3D-AChir *in vitro* protocol to the *in vivo* cell-types within Rothova2022. Green indicated the nearest neighbor(s). The “±” denotes the standard deviation on the distance, calculated from the bootstrap. The names on top of each sub-table, e.g. D4a, is the label for a cluster from the *in vitro* experiments. The D followed by a number denotes the day along the differentiation of the cells that make up the cluster. The clusters were obtained using unsupervised clustering.

Appendix



Identification of the central intermediate in the extra-embryonic to embryonic endoderm transition through single-cell transcriptomics

Michaela Mrugala Rothová^{1,4}, Alexander Valentin Nielsen^{2,4}, Martin Proks^{1,4}, Yan Fung Wong¹, Alba Redo Riveiro¹, Madeleine Linneberg-Agerholm¹, Eyal David³, Ido Amit³, Ala Trusina² and Joshua Mark Brickman¹

High-resolution maps of embryonic development suggest that acquisition of cell identity is not limited to canonical germ layers but proceeds via alternative routes. Despite evidence that visceral organs are formed via embryonic and extra-embryonic trajectories, the production of organ-specific cell types in vitro focuses on the embryonic one. Here we resolve these differentiation routes using massively parallel single-cell RNA sequencing to generate datasets from FOXA2^{Venus} reporter mouse embryos and embryonic stem cell differentiation towards endoderm. To relate cell types in these datasets, we develop a single-parameter computational approach and identify an intermediate en route from extra-embryonic identity to embryonic endoderm, which we localize spatially in embryos at embryonic day 7.5. While there is little evidence for this cell type in embryonic stem cell differentiation, by following the extra-embryonic trajectory starting with naïve extra-embryonic endoderm stem cells we can generate embryonic gut spheroids. Exploiting developmental plasticity therefore offers alternatives to pluripotent cells and opens alternative avenues for in vitro differentiation.

A key question in developmental biology is resolving cell lineages, their contribution to organ function and how to recapitulate this process in vitro. A large proportion of visceral organs are specified from the endoderm lineage. In eutherian mammals, endoderm is specified in two waves. The first occurs during pre-implantation development when the largely extra-embryonic primitive endoderm (PrE) forms from the inner cell mass and segregates from the embryonic epiblast. The second occurs at gastrulation with the specification of definitive endoderm (DE)^{1,2} from the epiblast, the major progenitor population of the visceral organs. Although initially identified as embryonic and extra-embryonic³, these routes are less divergent than previously thought^{4–8}, and despite their different trajectories, they are induced by similar signalling pathways and transcription factors (TFs)⁹.

At embryonic day (E) 4.5, the PrE further differentiates towards the parietal endoderm (PE) and visceral endoderm (VE)¹⁰. The PE provides mechanical protection and nutrient absorption, while the VE supports nutrient provision and has a role in patterning the embryo^{11,12}. Before gastrulation, the VE expands to cover the embryo surface, and at E5.5, it specializes into embryonic VE (EmVE, overlaying the embryonic epiblast) and extra-embryonic VE (ExVE, overlaying the extra-embryonic ectoderm)¹². The distal E5.5 EmVE then migrates anteriorly, forming the anterior visceral endoderm (AVE), a signalling centre that restricts primitive streak (PS) formation to the embryo's posterior region^{12,13}.

DE induction occurs in a temporal sequence that reflects anterior–posterior (A–P) identity. During early gastrulation, endoderm cells express markers associated with anterior identity^{14,15}. The first DE emerges along the midline from the embryo's distal tip, giving rise to anterior definitive endoderm (ADE), the anterior most axial

tissue migrating ahead of the axial mesoderm (node, notochord and prechordal plate)¹⁶. Subsequently, DE is recruited from the epiblast, emerging from the PS region up to E7.5 (ref. 17). While previously the DE was thought to displace the VE, recent lineage tracing⁴, imaging^{6,18} and transcriptomic^{7,8,18} data suggest that epiblast cells emerge from the PS to intercalate with the VE. The resulting endodermal epithelium contains both embryonic and extra-embryonic progenitors of the embryonic gut^{17,8,19}.

Naïve embryonic stem cells (ESCs) are pluripotent cell lines derived from the pre-implantation embryo. Previous attempts to generate in vitro cell types recapitulating organ function have focused on using pluripotent and, specifically, ESCs, to generate DE^{20–22}, but the definition of DE was historically based on a limited marker set^{22–24}. Moreover, cytokines known to promote DE can induce PrE in naïve ESC cultures⁹, calling for better characterization of in vitro differentiation trajectories.

In this Article, we address alternative differentiation routes towards endoderm by following FOXA2, a TF expressed in the embryonic and extra-embryonic endodermal primordia²⁵. We used a FOXA2^{Venus} reporter mouse²⁶ and performed fluorescence-activated cell sorting (FACS) of FOXA2^{POS} populations for single-cell transcriptomics using massively parallel single-cell RNA sequencing (MARS-seq)²⁷. With a modest number of cells and a simple computational tool, we identify intermediates and derive insights into endoderm specification in vivo and in vitro.

Results

FOXA2-based MARS-seq enhances lineage resolution in the endoderm. We used MARS-seq to determine the transcriptome of FOXA2^{POS} FACS-isolated embryonic cells to produce a detailed

¹Novo Nordisk Foundation Center for Stem Cell Medicine (reNEW), University of Copenhagen, Copenhagen, Denmark. ²Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. ³Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. ⁴These authors contributed equally: Michaela Mrugala Rothová, Alexander Valentin Nielsen, Martin Proks. ✉e-mail: trusina@nbi.ku.dk; joshua.brickman@sund.ku.dk

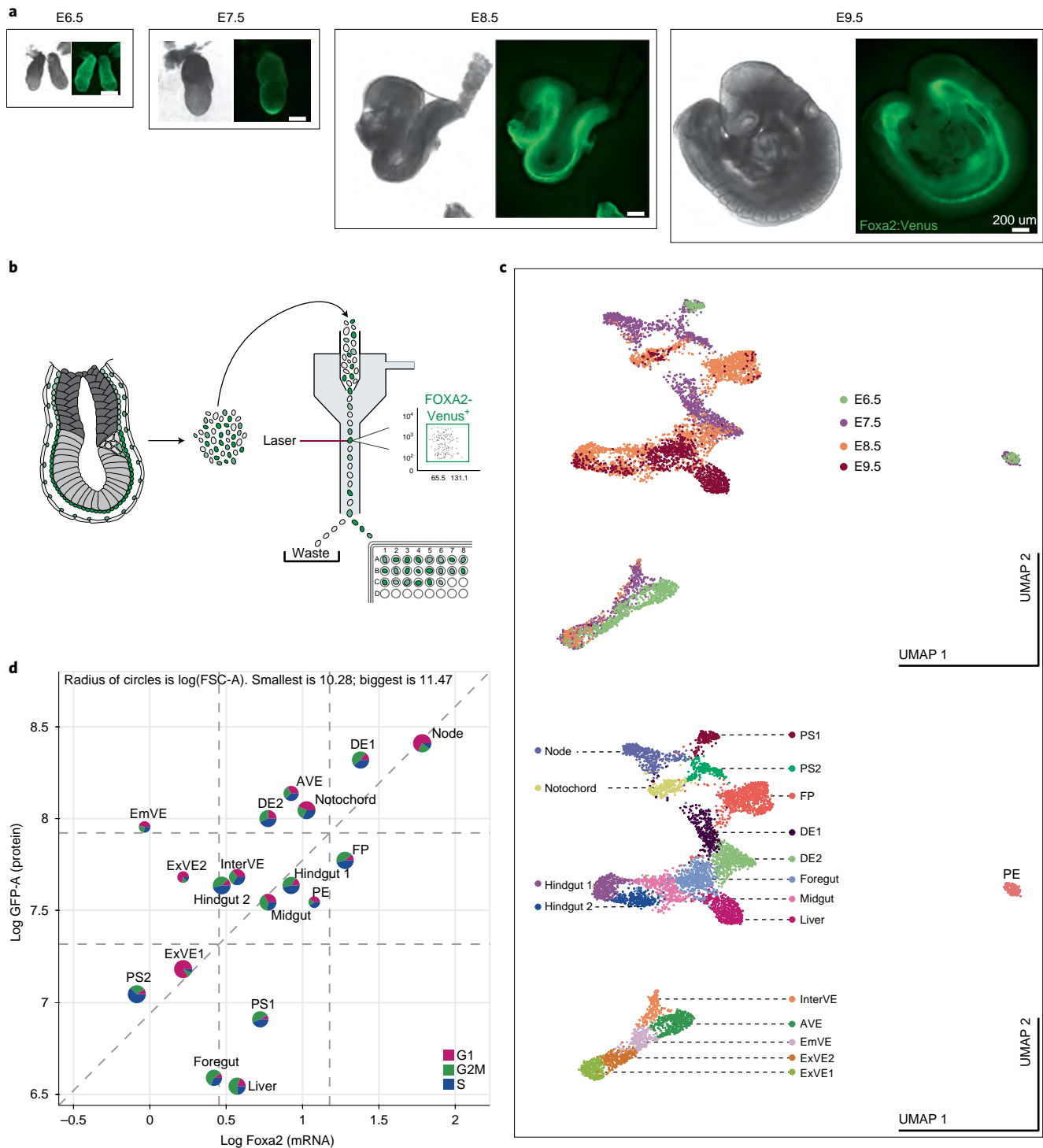


Fig. 1 | Single-cell sequencing and analysis of embryonic *Foxa2*^{POS} lineages. **a**, *FOXA2*^{POS} mouse embryos collected at stages between E6.5 and E9.5. **b**, Schematic of MARS-seq preparatory steps. Single *FOXA2*^{POS} cells from mouse embryos were isolated and sorted by FACS into wells for sequencing. Each sorted cell's specifications (such as forward scatter, proxy for cell size, and *FOXA2*-Venus intensity, a proxy for *FOXA2* protein levels) were recorded and linked to a specific well or cell. **c**, UMAP dimensional embedding of 6282 *FOXA2*^{POS} cells partitioned into 18 clusters. **d**, FACS and cell cycle chart representing the cell size (radius of circles), *Foxa2* messenger RNA (mRNA) level (x axis), *FOXA2* protein level (y axis) and cell cycle phase (colour of pie charts, G1 phase in red, G2M phase in green and S phase in blue) for specific clusters. Grey dashed lines correspond to 33% fractiles of all cells. FP, floor plate.

map of endodermal populations emerging during gastrulation (Fig. 1a–c). Homozygous *FOXA2*^{Foxa2-Venus/Foxa2-Venus} were crossed to C57BL/6 to ensure genetically uniform *FOXA2* expression (Fig. 1a). After stringent quality filtering, our in vivo dataset comprised 6282 *FOXA2*^{POS} cells, forming 18 distinct clusters spanning E6.5 to E9.

(Fig. 1c and Extended Data Fig. 1a). The cluster identities were assigned by expression of lineage-specific markers selected from differentially expressed genes (DEGs) (Supplementary Dataset 1; examples in Extended Data Fig. 1b) and confirmed with the Mouse Genome Informatics database²⁸. The identified clusters contain all

known FOXA2^{POS} progenitors between E6.5 and E9.5, including PS, node, notochord, floor plate (FP), DE, VE and PE (Fig. 1c and Extended Data Fig. 1b).

With MARS-seq, we can relate transcriptomes to FOXA2 protein levels, cell size, discrete embryonic regions and embryo age (Extended Data Fig. 2a,b and Supplementary Dataset 2) since embryos were imaged, E8.5 to E9.5 embryos dissected (Extended Data Fig. 2c) and every cell FACS-indexed (Fig. 1b–d). As FOXA2 is a recognized early node marker¹⁷, we explored this cluster as a proof of principle for our dataset. We found that the node population comprised E7.5 cells (Supplementary Dataset 2), consistent with its developmental emergence. The primary descendants of the node, the notochord, first appeared at E7.5, with the cells clustered together irrespective of their anterior or posterior origin (Supplementary Dataset 2). A second node-derived population contains the FP progenitors and is characterized by the expression of markers for both caudal (*Nkx2.6*, *Hoxa1*, *Hoxb1* and *Hoxa2* positive) and rostral (*Fezf1*) neural cells.

We identified clusters with high FOXA2 protein levels as those associated with patterning centres that are either node derived or exhibit similar patterning activity, including notochord, FP and AVE. Similarly, one of our DE clusters, early DE (DE1), that expresses anterior markers such as *Cer1*, *Fzd5* and *Lhx1* also expresses high levels of *Foxa2*. This analysis also suggested that node and notochord are some of the largest cells in our dataset (Fig. 1d).

To explore the nature of gastrulation-stage FOXA2^{POS} populations, we performed subclustering of early cell types, specifically PS, DE, node and VE (Extended Data Fig. 2d). Here we identified two node subpopulations, one largely in G1 and the other actively proliferating (referred to as Node-pr, Extended Data Fig. 2d,e; Supplementary Dataset 3). Node-pr may represent the founders of primed pluripotent epiblast stem cells, as these are thought to resemble anterior later PS (PS2)²⁹.

From global clustering (Fig. 1c) and gastrulation-stage subclustering (Extended Data Fig. 2d), we observed two PS clusters expressing lower levels of FOXA2 than midline (node and notochord) and AVE clusters (Fig. 1d). These PS clusters largely resolved on the basis of developmental timing. We identified both early and late cells in single embryos, suggesting that they represent phases of differentiation rather than embryonic stage (Supplementary Dataset 2). While cells in both PS clusters appear to proliferate rapidly³⁰ and spend little time in G1, the PS2 are predominantly in S phase (Fig. 1d).

In contrast to the previously published datasets^{7,8}, MARS-seq of FOXA2^{POS} populations resolved two clusters of DE cells (DE1 and later DE (DE2)) (Fig. 1c and Extended Data Fig. 2d) consistent with two stages of DE differentiation. These clusters are induced with different dynamics, with the majority (91%) of DE1 forming at E7.5 and DE2 between E7.5 and E8.5 (Extended Data Fig. 2b and Supplementary Dataset 2).

The five annotated VE clusters all grouped together by uniform manifold approximation and projection (UMAP)³¹ and separated from the embryonic clusters (Fig. 1c and Extended Data Fig. 2d). Four VE clusters corresponded to distinct regions of the VE: VE overlaying the putative extra-embryonic region (ExVE1 and ExVE2); the embryonic region (EmVE); and AVE. Both ExVE clusters are primarily in G1 and appear different in size (Fig. 1d), while lineages overlaying the embryonic region (EmVE and AVE) proliferate more actively (Fig. 1d).

The fifth VE cluster expressed both canonical VE markers, namely *Trap1a*, *Sepp1* and *Apoc1*, and DE markers *Trh*³², *Gpx2* and *Cpm* (Supplementary Dataset 1). The unique expression profile of this cluster suggested it could represent a transition state in the conversion of VE to DE, and this cluster is referred to as intermediate VE (InterVE). In support of this, we identified *Trap1a* as specifically upregulated in the InterVE population (Supplementary Dataset 1), which has previously been used as a marker gene for tracking VE contribution to gut endoderm *in vivo*¹⁹. The InterVE consisted of mainly E7.5 but ranged up to E9.0 cells, scattered in both anterior and posterior embryonic regions (Extended Data Fig. 2b and Supplementary Dataset 2), and like the EmVE, AVE, DE1 and DE2, these cells are also proliferating (Fig. 1d).

Mapping FOXA2 populations onto a global endoderm map. A central challenge in single-cell transcriptomics is performing reliable cross-dataset comparisons^{33,34}. Many current approaches identify cell-to-cell similarities across datasets, after which cells are combined and the joint dataset clustered, which requires limiting the comparison to a subset of genes and adjusting non-trivial heuristic parameters^{33–36}. To take all genes into account, we developed a single-parameter approach to compare cell clusters both within and between datasets called Cluster Alignment Tool (CAT). Rather than comparing datasets at the level of single cells, we ask which cell clusters are most similar to each other. We first identify the overall gene profile of a cluster by averaging gene expression across the cells within a given cluster (Fig. 2a). To find the best matching pair, here referred to as ‘alignment’, we perform nearest neighbour search, that is, calculate all pairwise Euclidean distances between the clusters and identify the cluster pair with the shortest distance (Extended Data Fig. 3a). A cluster can align to several other clusters if their distances do not differ significantly from the shortest, and to assess the significance, we perform bootstrapping (with $P > 0.05$ cut-off) (Fig. 2b and Extended Data Fig. 3b).

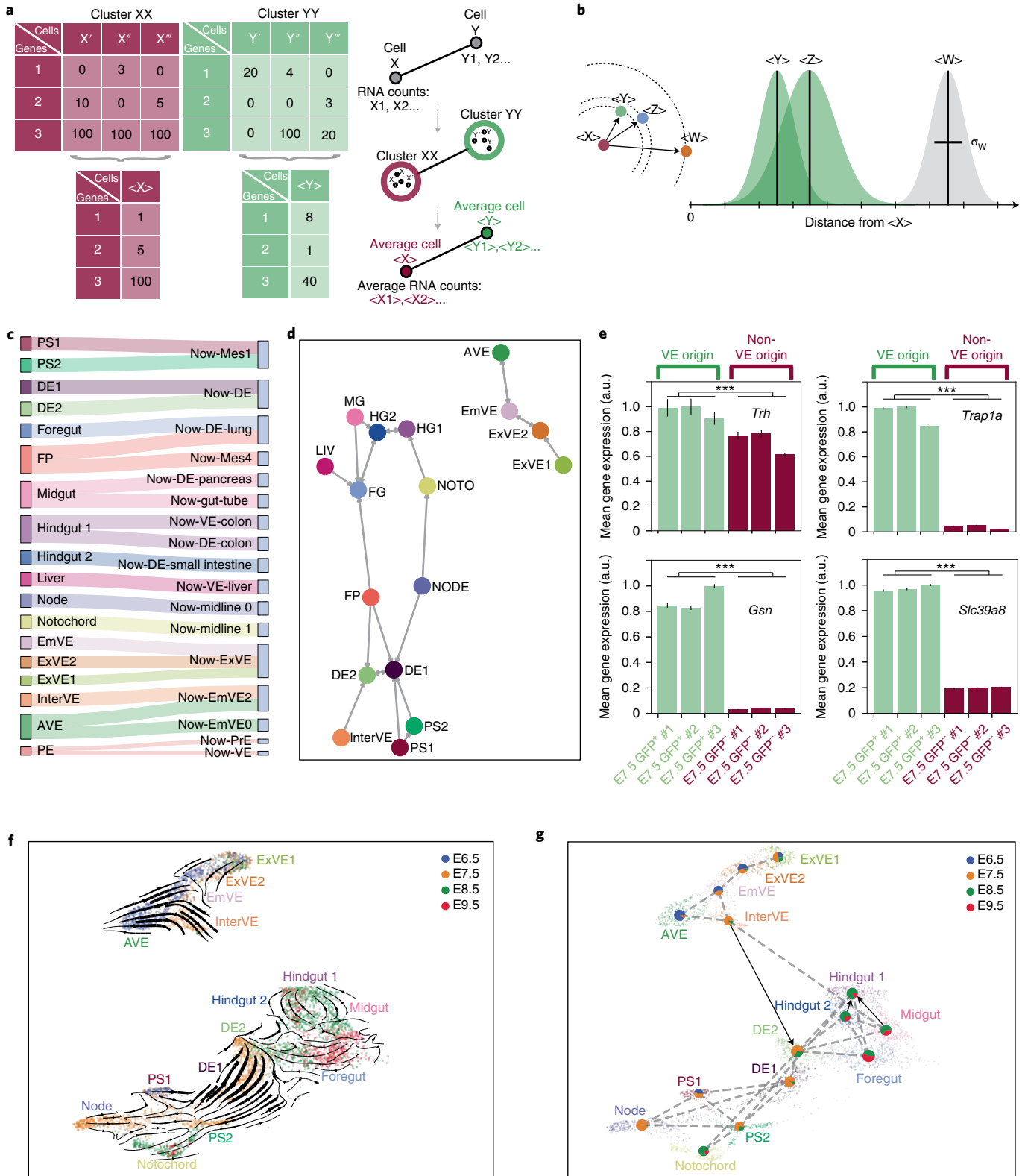
To validate CAT, we compared lineages defined by FOXA2 expression in our MARS-seq to publicly available 10x Genomics data⁷ (Nowotschin et al., 2019) referred to as Now-2019 (Fig. 2c and Extended Data Fig. 3c). CAT correctly aligned equivalent clusters, including our dataset’s foregut to Now-lung, node to Now-midline and hindgut to Now-colon (Extended Data Fig. 3c), with a few exceptions (Extended Data Fig. 3c). Subclustering of the Now-mesoderm,

Fig. 2 | CAT and identification of the InterVE cluster as an intermediate population between VE and DE. **a,b**, A schematic illustration of CAT (**a**) and the identification of alignments (**b**). The single-cell matrix is reduced to represent the cluster’s bulk RNA-seq values, and the Euclidean distances are calculated between all cluster pairs. In **b**, the distances between clusters are illustrated by arrows and dashed circles, histograms represent bootstraps of cluster-cluster distances and the significant closest pairs are highlighted in green. **c**, Sankey diagram visualizing the CAT alignments between our dataset and the subclustered Now-2019 dataset (Now). Thickness of the line is inversely proportional to the pairwise distance. **d**, Network projection of significant CAT alignments highlights the position of InterVE together with the embryonic but not the VE lineages. **e**, Expression of four InterVE markers (*Trh*, *Trap1a*, *Gsn* and *Slc39a8*) in E7.5 cells from Now-2019 dataset. In these cells, the VE origin is assessed by AFP-GFP lineage tracing. The InterVE markers are predominantly expressed in the VE-origin cells (AFP-GFP^{POS}). The bars represent standard error of the mean. The number of samples (n) is: E7.5 GFP⁺ #1: 4,085; E7.5 GFP⁺ #2: 5,007; E7.5 GFP⁺ #3: 8,717; E7.5 GFP⁻ #1: 3,426; E7.5 GFP⁻ #2: 3,793; E7.5 GFP⁻ #3: 6,723. Mann-Whitney U test was used to determine the significance between GFP⁺ and GFP⁻ samples; *** $P \leq 0.001$ ($P = 5 \times 10^{-130}$ for *Trh* and $P = 0$ for the rest of markers). **f**, Results of RNA velocity analysis on VE- and DE-related lineages showing that the InterVE is from the VE lineages. Arrows represent the direction and magnitude of changes in gene expression. Although a few arrows point in the opposite direction to major developmental trends (for example, foregut to DE2) these are all at the lowest level of significance and are likely due to known limitations of RNA velocity. **g**, Directed PAGA constructed on the basis of RNA velocity highlights the direction of InterVE differentiation to DE2. FG, foregut; FP, floor plate; HG, hindgut; LIV, liver; MG, midgut; NOTO, notochord.

Now-midline and Now-EmVE resulted in a better mapping onto the complete Now-2019 dataset (Fig. 2c), regardless of the direction of the comparison (Extended Data Fig. 3d). Particularly, the DE1 aligned specifically to Now-DE, InterVE to Now-EmVE2 and notochord to Now-midline1 (Fig. 2c). While the majority of alignments improved by subclustering, our FP cluster still stood out as aligning with Now-Mes and Now-lung clusters (Fig. 2c). We investigated what gene expression is shared between these three clusters

and identified the top five significant genes with $P < 10^{-10}$ (*Prtg*, *Apex1*, *Mdk*, *Fxyd6* and *Pcbp4*), all of which have known expression in neural lineages²⁸.

We then used CAT to map relationships between lineages within our dataset and observed alignments that fit with known lineage relationships, including node to DE1 and notochord; liver to foregut; DE1 to DE2; AVE to EmVE; ExVE to EmVE (Extended Data Fig. 3e), although the InterVE cluster aligned to DE2 but not the



other VE clusters. The CAT-inferred lineage alignments are summarized in a network (Fig. 2d) with the nodes representing lineages, the links representing alignments and directionality based on the orientation of the comparison. In the CAT network, VE and embryonic lineages are separated from each other, but the InterVE population is located at the edge of the embryonic subnetwork linked to DE2 (Fig. 2d). The unique position of InterVE between the embryonic and extra-embryonic endoderm identity was also reflected in principal component analysis (PCA) (Extended Data Fig. 3f), partition-based graph abstraction (PAGA)³⁷ and trajectory inference (Extended Data Fig. 4a,b). The visceral quality of InterVE is apparent from its alignment to the Now-EmVE2 cells of visceral origin, based on AFP-GFP lineage tracing, (ref. 7) (Supplementary Dataset 4 and Fig. 2c) that also express InterVE markers (Fig. 2e).

To resolve the directionality of differentiation, we implemented RNA velocity³⁸ (Fig. 2f,g) and this indicates InterVE is differentiating from VE lineages. The strongest link in directed PAGA projection (Fig. 2g) further confirms differentiation of InterVE to DE2 (Fig. 2d). Also indicated by our CAT analysis, the relationship between DE1 and DE2 is consistent with RNA velocity, suggesting that DE2 is initially formed from early PS (PS1) via DE1 and possibly later from InterVE (up to E8.5). While some DE1 cells appear related to PS2, these are limited to E7.5 and the E8.5 PS2 cells associated with notochord (Fig. 2f).

Taken together, this implies a continuum of DE differentiation that begins with the E6.5 PS and ends with E8.5 visceral endoderm. The contribution of InterVE to DE2 is also confirmed by DEG analysis showing InterVE markers progressively being downregulated in DE2 as it matures at E8.5 (Supplementary Dataset 5).

Functional properties of VE in transition to DE. InterVE markers are expressed individually in various definitive and visceral lineages, but are co-expressed within the InterVE (Fig. 3a and Extended Data Fig. 5a). While transdifferentiation of VE to embryonic endoderm has been described previously^{4,7,8,19}, the means of this transformation has not been defined. To understand the differences between InterVE, VE and DE populations, we performed Gene Ontology (GO) analysis of DEGs³⁹ (Fig. 3b and Supplementary Dataset 6). The GO terms downregulated in InterVE compared with VE were the same categories upregulated in InterVE compared with DE2 (the closest cluster pair for InterVE in our CAT analysis). Additionally, there were no significantly downregulated terms when InterVE was compared with DE2, suggesting that InterVE had acquired DE identity, but not yet lost the visceral signature (Fig. 3b). DEGs upregulated in InterVE relative to VE were linked to proliferation, actin cytoskeleton and migration, whereas those downregulated relative to VE, or upregulated with respect to DE2, included lipid, extracellular matrix (ECM) and Smad pathways (Fig. 3b).

We then assessed the behaviour of gene groups linked to specific GO functional categories^{39,40} (GO-FCs; Supplementary Dataset 7). First, we compared cumulative gene expression in a cell for each GO-FC, visualized by box plots, in the sequence of clusters leading

from VE to DE (Fig. 3c). These confirmed the GO analysis results (Fig. 3b) and highlighted the transitions in lysosome, Smad, retinoid metabolism and lipid-related genes that were specifically downregulated in the InterVE cells compared with VE (Fig. 3c). Only a few specific GO-FCs were upregulated (for example, Smoothened) as cells progress from VE through InterVE into DE (Extended Data Fig. 5b). VE and embryonic gut-tube derivatives share common basement membrane features, and this GO-FC is specifically reduced in both DE2 and DE1 (Fig. 3d). To further resolve functional trends, we used CAT to compare function-specific lineage expression states and uncovered candidate functions regulating the transition from VE to DE (Fig. 3d–f and Extended Data Fig. 5c,d). When applied to the entire set of functional categories, alignments formed three groups: those where InterVE had already acquired DE-like functional properties, those where InterVE retained VE-like properties, and a set of functional alignments where InterVE was still in transition and aligned to both VE and DE (Fig. 3e). When these are visualized as CAT networks (Fig. 3f), InterVE became disconnected from the other VE clusters and connected to the DE2 or hindgut clusters based on cell–cell junction, amino-acid metabolism and Hippo and Notch pathways GO-FCs (Fig. 3f and Extended Data Fig. 5c). In contrast, canonical Wnt signalling specifically connected the InterVE with AVE, but not with the other embryonic lineages (Fig. 3f). Consistent with the GO-FC box plot for basement membrane (Fig. 3d), CAT network for this category positions InterVE between VE and embryonic gut (Extended Data Fig. 5d), suggesting that the basement membrane composition of the VE is adopted by the prospective gut tube. Thus, RNA velocity, CAT and GO analyses together suggest that, during the transition from VE to DE, InterVE acquires definitive identity more efficiently than it loses its visceral identity.

Spatial localization of InterVE cells in E7.5 embryos. Historically, the localization of the cells expressing VE markers was used to conclude that VE cells are displaced⁴¹, although genetic lineage tracing suggests this model is incorrect. To localize InterVE cells in E7.5 embryos, we used Resolve Molecular Cartography (MC) for spatial transcriptomics (ST) implemented for quantitative multiplexed spatial mRNA analysis at single-molecule and single-cell resolution (Fig. 4 and Extended Data Fig. 6). Markers like *Afp* illustrate why displacement was an attractive hypothesis, as *Afp* expression at E7.5 is not detectable in VE regions overlapping with emergent DE (distal tip or prospective anterior region). However, in line with lineage tracing *Trap1a* persists in cells that have lost *Afp* expression (Fig. 4a).

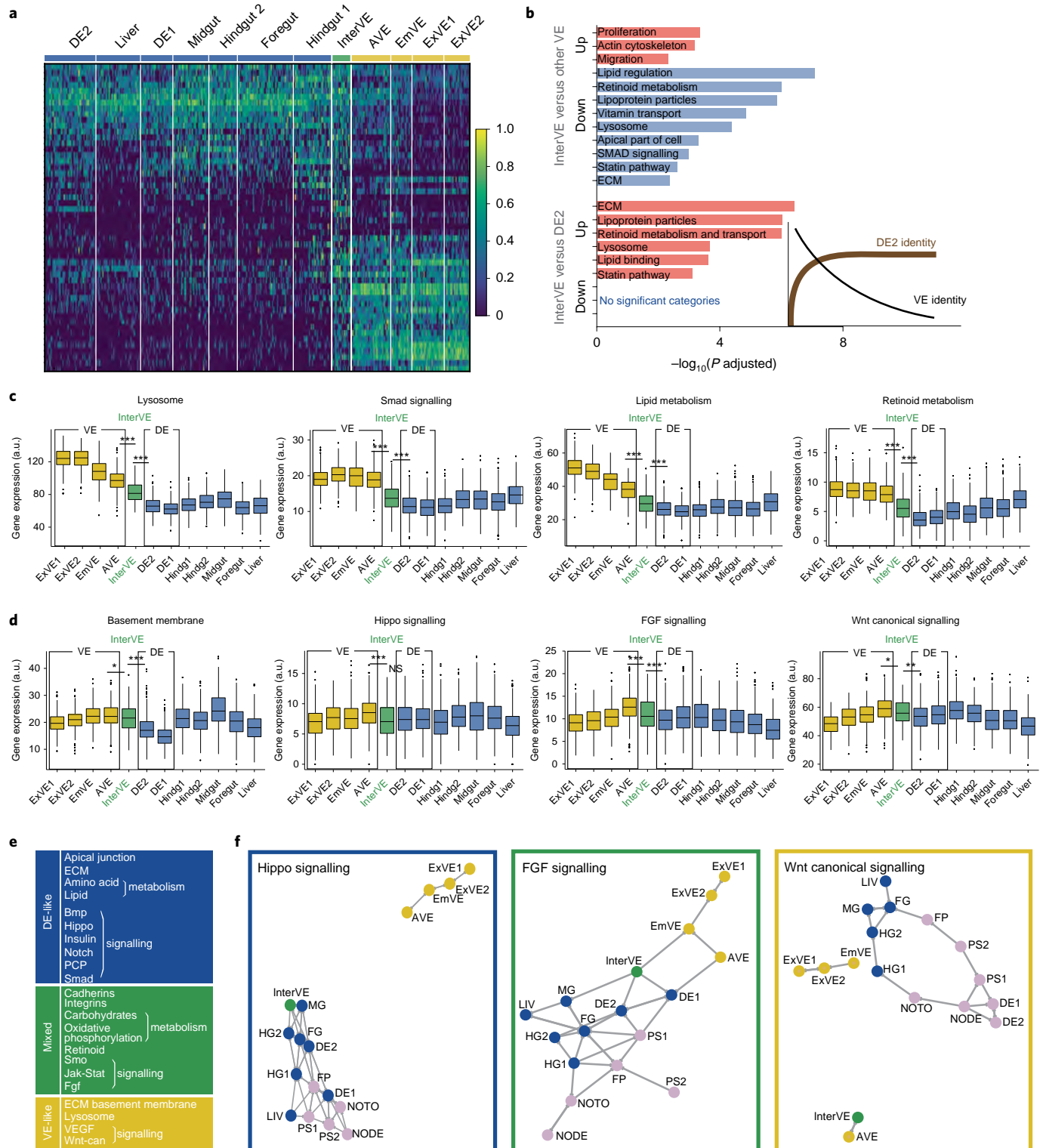
We used a decision tree algorithm trained on MARS-seq data to predict InterVE cells in ST (Fig. 4b,c and Extended Data Fig. 6a). From this, we compiled a list of key InterVE markers (Extended Data Fig. 6a) that could assign InterVE identity in ST data (Extended Data Fig. 6b–d). Identified InterVE cells are primarily located on the anterior surface of the embryo, although there are also posterior cells (Fig. 4d and Extended Data Fig. 6d). The ST-identified InterVE

Fig. 3 | Functional properties of VE in the transition to DE. **a**, Heat map representing selected DEGs in InterVE cluster. The DE-expressed genes (top of the heat map) and VE-expressed genes (bottom) are co-expressed in the InterVE cluster. **b**, GO-term analysis of genes down- and upregulated in InterVE versus other VE clusters or versus DE2. The scheme illustrates that InterVE acquires DE2 identity faster than it loses its VE signature. **c**, Cumulative gene expression of genes in three GO-FCs identified by GO analysis in **b**. **d**, Cumulative gene expression of genes in GO-FCs not revealed by GO analysis. Box plot in **c** and **d** shows median value, bordered by upper/lower (75/25) percentile, whiskers showing maximum/minimum value in 1.5 interquartile range (IQR) and outliers as black dots. Two-sided Mann–Whitney *U* test was used to determine the significance between AVE and InterVE as well as DE2 and InterVE; **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. NS, nonsignificant. The number of samples (cells), in each box plot are given in Supplementary Table 2. **e**, Table summarizing CAT alignments between InterVE and the rest of our in vivo dataset using specific GO-FC. The InterVE alignments are grouped into three categories based on the alignment to either DE lineages (blue), or to both VE and DE lineages (green), or to VE lineages only (yellow). **f**, Examples of CAT networks based on GO-FC in **e** representing pathways in **d**. In contrast to the cumulative gene expression in **d**, the CAT network for Hippo signalling GO-FC clearly illustrates that the InterVE aligns to DE but not VE lineages, while for the Wnt-canonical signalling GO-FC, the InterVE aligns to VE and not to DE lineages. The VE lineages are visualized in yellow, DE lineages in blue, InterVE in green and other lineages in pink.

expression profiles are comparable to MARS-seq InterVE, showing characteristic co-expression of DE and VE markers (Fig. 4d and Extended Data Fig. 6b,c). The abundance of anterior InterVE is consistent with our trajectory inference, where InterVE contributes to DE2, a cluster consisting of both anterior and posterior cells (Fig. 2f,g). To further probe the anterior origin of InterVE, we identified AVE clusters in Nowotschin-2019 dataset (Extended Data Fig. 6e,f) and find that E7.5 InterVE (Now-EmVE2) aligns to E6.5 AVE (Now-EmVE0) (Supplementary Dataset 4 and Extended Data

Fig. 6g). These data suggest that early InterVE is anterior in origin and later InterVE is posterior. In line with this, PAGA analysis performed on the Now-2019 InterVE cluster (Now-EmVE2) (Extended Data Fig. 6e) suggests that InterVE contribution spans the entire endodermal A–P axis.

In vitro differentiation of endoderm lineages. In vitro differentiation to DE has focused heavily on cell surface markers, such as CXCR4 or C-KIT^{22,42}, to identify DE precursors. We assessed the



expression of these canonical markers in our dataset and found that their RNAs are also expressed broadly throughout the VE (Extended Data Fig. 7a). As these markers may not be adequate to distinguish DE from VE, we asked whether established *in vitro* DE differentiation successfully mimics *in vivo* development.

We differentiated ESCs towards DE in a two-dimensional (2D) environment²¹ (Supplementary Dataset 8) using a double fluorescent reporter for *Gooseoid* (*Gsc*), a marker for early PS, node and later anterior PS, and, for *Hhex*, a marker for the later anterior endoderm (*Gsc*-GFP/*Hhex*-RedStar)⁴³. This enabled the sorting of specific differentiation stages, with *Gsc*-low cells representing PS1, *Gsc*-high anterior PS and *Gsc*/*Hhex* double-positive nascent DE⁴⁴ (Extended Data Fig. 7b). We performed MARS-seq to identify differentiating ESCs as they acquire an endodermal identity (Supplementary Dataset 9). In this dataset, referred to as 2D-ESC, we assessed the expression of PS, DE and VE markers in distinct clusters (Extended Data Fig. 7c,d).

Similar to comparisons of *in vivo* datasets across different platforms (Fig. 2c and Extended Data Fig. 3c), comparing *in vitro* data with *in vivo* data in high-dimensional space is nontrivial. Commonly used approaches to tackle this include correlational analyses on subsets of marker genes⁴⁵ or testing co-localization in a dimensional embedding^{46,47}. We combined our *in vivo* FOXA2^{POS} data together with the *in vitro*-derived DE cells and applied fastMNN³⁴ batch correction followed by UMAP visualization on the entire gene set. This resulted in segregation of cells based solely on their *in vitro* or *in vivo* origin (Fig. 5a). Since CAT does pairwise comparisons one at a time, we can directly ask which *in vitro* states best match *in vivo* counterparts without dealing with global patterns of cluster co-localization or batch-like segregations observed in dimensional embedding. On the basis of CAT, we found 2D-ESC differentiation recapitulates the expected developmental stages, with the intermediate stage of differentiation (day 4) aligning with *in vivo* PS (Fig. 5b) and early DE lineages (DE1 and DE2). While the clusters aligning to DE1 appear at later timepoints in differentiation, we presume this is because both early and anterior endoderm express common sets of markers (*Cer1*, *Lhx1*, *Otx2* and *Hesx1*)^{14,15} and this alignment reflects the anterior character acquired by these cultures in late stage differentiation. Overall, later stages of *in vitro* differentiation (day 5 and day 6) aligned to DE (Fig. 5b) with two small clusters aligning to foregut and hindgut. Each of these small clusters has additional alignments to InterVE and FP, respectively.

On the basis of a limited set of markers, we have previously shown that PI3K modulates DE differentiation and its inhibition results in posterior/pan-endodermal DE that retains plasticity, suggesting delayed differentiation⁴³. To distinguish between posteriorization and delay, we perturbed PI3K signalling with the antagonist LY294002 (PI3Ki) (Extended Data Fig. 7c,d and Extended Data Fig. 8a) and assessed differentiation by CAT. Rather than delaying differentiation and increasing alignment to PS, we observed an expansion in alignments to DE2 and hindgut (Fig. 5c). As DE2 no longer expresses early DE markers that remain expressed in the anterior endoderm, we conclude that the PI3K inhibitor homogenizes *in vitro* differentiation, while eliminating anterior gene expression associated with prospective foregut. The increase in alignments to DE2 in PI3Ki-treated cells is also confirmed, when data from these experiments are clustered together (Extended Data Fig. 8b,c).

Given the success of 3D culture systems as models for embryonic differentiation^{48,49}, we considered whether it alters differentiation trajectories. We guided wild-type (WT) E14 naive mouse ESCs towards epiblast-like cells (EpiLC)⁵⁰ and further differentiated using 3D protocols (Supplementary Dataset 8). We sorted DE from differentiation with E-CAD and CXCR4 and performed single-cell RNA-seq (Extended Data Fig. 7c,d). We found that these 3D-cultured cells clustered together with 2D-ESCs (Extended Data Fig. 8d), and while differentiation in 2D produced both DE1 and DE2, 3D differentiation enhanced the fraction of DE2 cells (Fig. 5b–d). To promote differentiation towards an anterior identity, we stimulated Nodal and Wnt signalling (increased Activin A and GSK3 inhibitor CHIR99021; Supplementary Dataset 8 and Extended Data Fig. 7c,d). This restored anterior clusters and accelerated differentiation, producing fore- and hindgut identities (Fig. 5e). These observations suggest that the ESC differentiation protocols tested here are biased towards the generation of DE2 and show little or no alignment of *in vitro* clusters to visceral endoderm.

To identify molecular differences between the *in vitro* differentiated cells and their *in vivo* counterparts, we introduced a similarity measure (Extended Data Fig. 8e). We defined expected *in vivo* target lineages for *in vitro* differentiation and compared *in vitro* clusters with these targets (Fig. 5f and Supplementary Dataset 10). The similarity score uses the set of CAT alignments for specific signalling pathways to quantify how well *in vitro* protocols recapitulate *in vivo* signalling. This analysis identified the Hippo, Smad, Wnt non-canonical and FGF pathways (Supplementary Dataset 10) as candidates for future manipulation in differentiation protocols.

Generating gut spheroids from naïve extra-embryonic endoderm.

Despite identifying InterVE as a key intermediate in embryonic gut specification, we did not detect a prominent visceral population in tested *in vitro* differentiation protocols (Fig. 5b–e). We therefore asked if we could generate *in vitro* embryonic gut starting with naïve extra-embryonic endoderm (nEnd) stem cells⁹. MARS-seq performed on nEnd showed that these express similar genes to *in vivo* PrE (Now-PrE) and parietal endoderm (Now-ParE). As a result, Now-PrE and nEnd form a hierarchical cluster that includes Now-ParE (Fig. 6a and Extended Data Fig. 9a). Consistent with this, nEnd aligns to Now-PrE and Now-VE clusters when analysed by CAT (Fig. 6b).

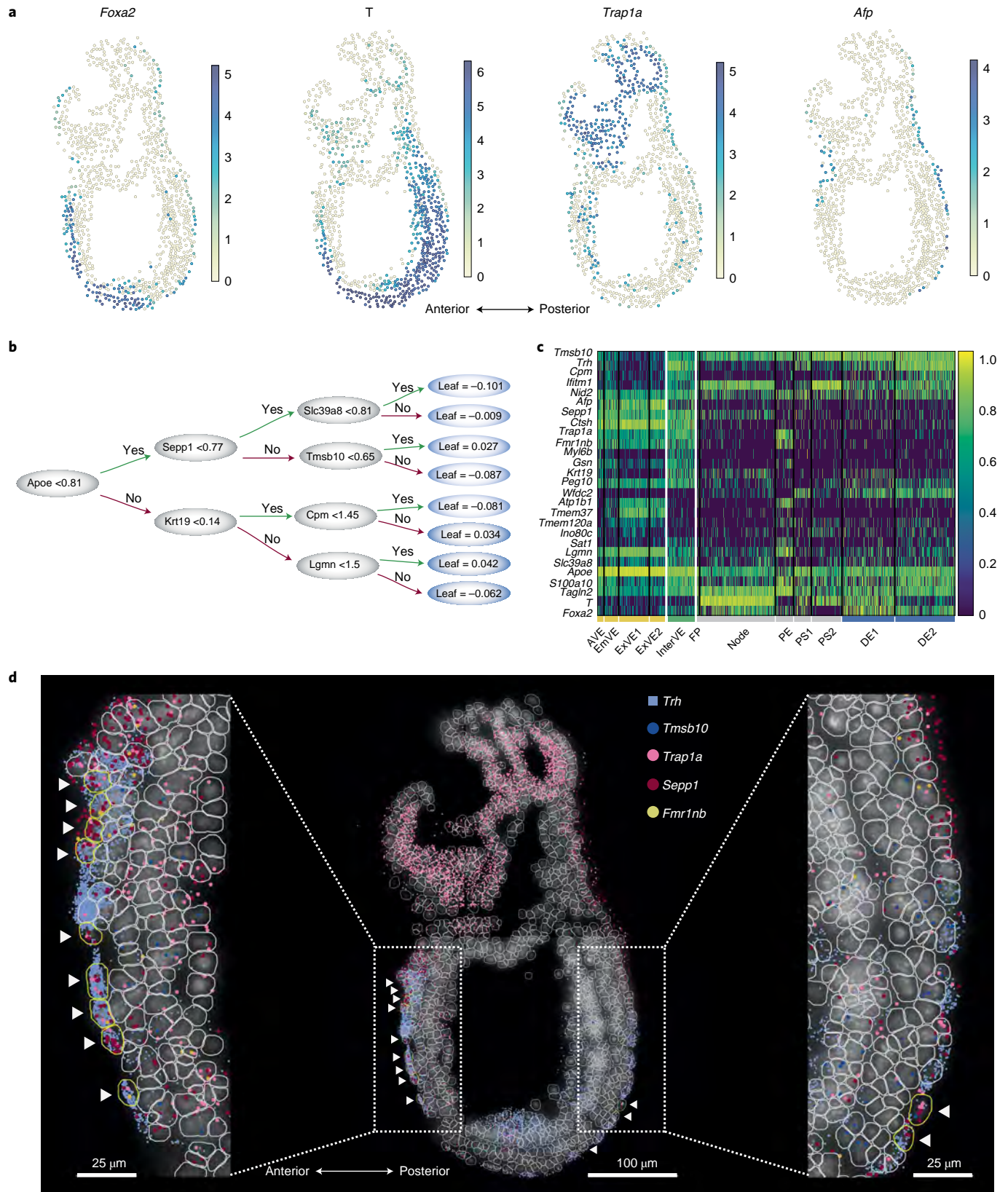
To determine whether nEnd could produce embryonic gut cell types via an InterVE intermediate, we differentiated nEnd into gut spheroids (nEnd spheroids), adapting an existing human ESC differentiation protocol⁵¹ (Fig. 6c–d). By the end of differentiation, we identified spheroids expressing the gut markers CDX2 and FOXA2 (Fig. 6d) in CDH1 positive epithelial cells (Extended Data Fig. 9b). To assess the trajectory of differentiation, we performed quantitative PCR with reverse transcription (RT-qPCR) (Extended Data Fig. 9c) and found InterVE markers to be dominant in the spheroid stage. To resolve distinct subpopulations, we performed MARS-seq (Fig. 6e and Extended Data Fig. 9d) and identified clusters expressing InterVE and gut markers in nEnd spheroids (Fig. 6e and Extended Data Fig. 9d). RNA velocity shows a clear trend of the population expressing InterVE markers differentiating into the gut-like and the other cell types found at this stage of differentiation (Fig. 6f and Extended Data Fig. 9c–e). To compare the two *in vitro* routes (DE

Fig. 4 | ST reveals InterVE in E7.5 embryos. **a**, Average mRNA counts per cell in a sagittal section of E7.5 embryo for selected genes (*Foxa2*, *Brachyury*, *Trap1a* and *Afp*). **b**, An example of the decision tree trained on MARS-seq data to predict if a cell is InterVE or not. Nodes (black ovals) represent the criteria for making the yes (green arrow) or no (red arrow) decisions. Leaves (light-blue ovals) are the ends of the decision branches. **c**, Heat map of gene expression patterns for genes used in the ST study. The genes, selected on the basis of the decision tree, were chosen to identify InterVE out of the related VE and DE genes. **d**, DAPI-stained sagittal section of E7.5 embryo. Each colour dot or square represents a single mRNA molecule. Cell boundaries, obtained in QuPATH, are outlined in grey, and the InterVE cells identified by decision tree are outlined in yellow and pointed by white arrowheads. The DE lineage markers (*Trh* and *Tmsb10*) are in shades of blue; the VE lineage markers are in shades of pink (*Trap1a* and *Sepp1*) and yellow (*Fmm1b*).

from 2D-ESCs and VE in nEnd spheroids), we assessed the direction of differentiation by RNA velocity (Extended Data Fig. 9e,f). We found that 2D-ESC progresses to DE via a PS-like stage, whereas the nEnd-spheroid route involves a cluster expressing Inter-VE markers.

We then used CAT to further assess nEnd differentiation (Extended Data Fig. 9g). As PrE is absent in our dataset and PE

is its closest analogue (Fig. 6a), nEnd aligns to PE. We also find alignments to PE later in differentiation, suggesting aspects of the nEnd signature persist throughout differentiation. During differentiation, the first cluster to align to both InterVE and regions of the embryonic gut, including DE2, is cluster 3 (nEnd'), the pre-population for spheroid formation. In agreement with marker



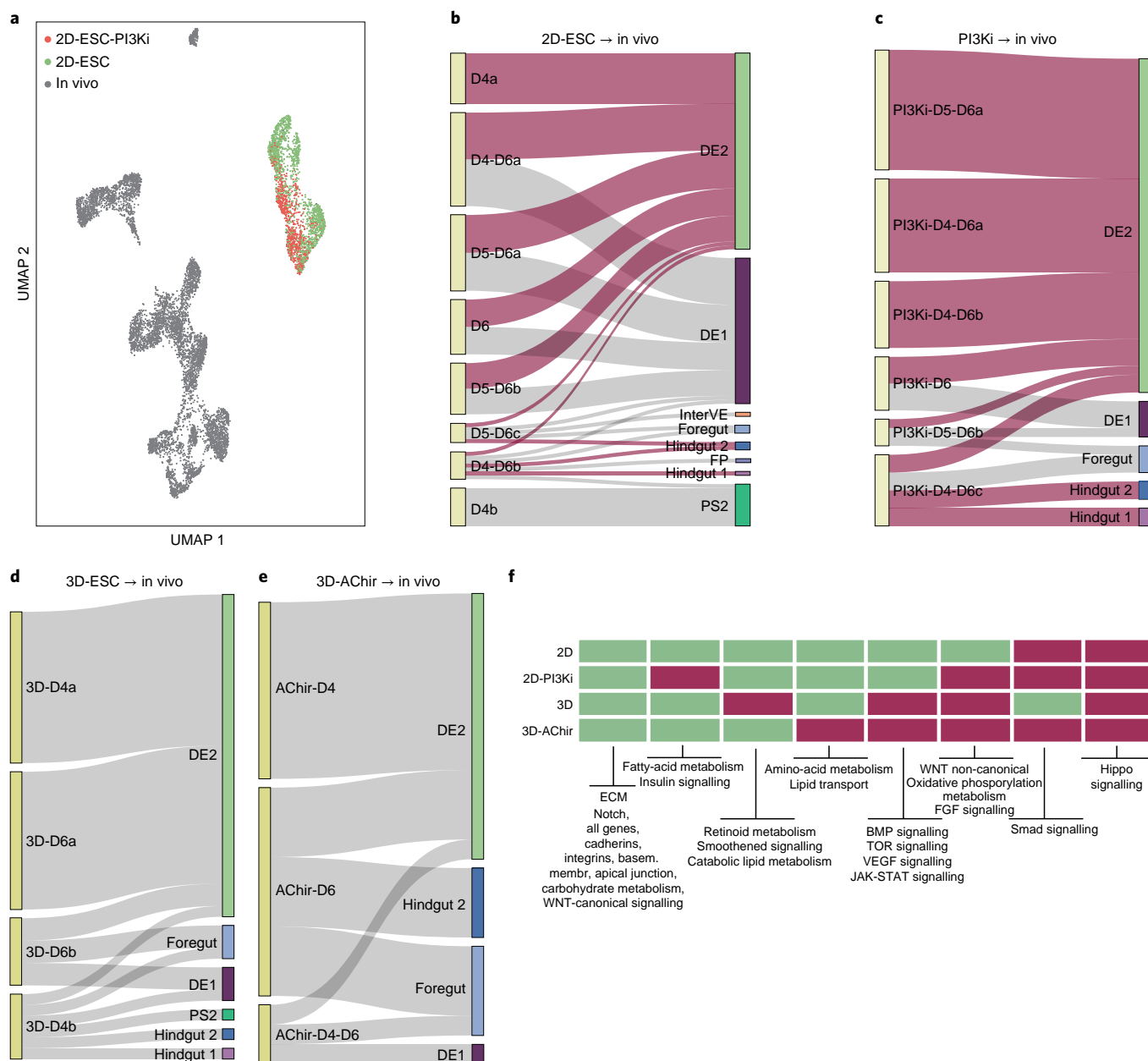


Fig. 5 | In vitro differentiation of endoderm lineages. **a**, UMAP dimensional embedding of our in vivo dataset together with 2D-ESC in vitro differentiation protocols. **b,c**, Sankey diagrams representing significant CAT alignments between clusters from 2D-ESC in vitro protocol (**b**), 2D-ESC-PI3Ki in vitro protocol (**c**) and our in vivo dataset. The alignments to DE2 and hindgut lineages are highlighted in red. The alignments to DE2 and hindgut lineages in **c** dominate over the alignments in **b**. **d,e**, Sankey diagrams representing significant CAT alignments between clusters from 3D-ESC in vitro protocol (**d**), from Activin A/Chiron 99021 (3D-AChir) in vitro protocol (**e**) and our in vivo dataset. In **b-e**, the length of the cluster bar on the left reflects the size of the cluster. The thickness of the lines is given by the cluster bar length divided by the number of alignments. The bar length on the right is given by the sum of the line widths and reflects the fraction of in vitro cells best matching the specific in vivo lineage. **f**, Table representing the similarity of the in vitro cells to the appropriate in vivo lineages. For each of the protocols (2D, 2D-PI3Ki, 3D and 3D-AChir) GO-FCs are coloured according to similarity measure, *S*, (median over similarities calculated for individual clusters in each protocol; Extended Data Fig. 8e) with high similarity in green and low similarity in red. All in vitro protocols have low similarity to in vivo DE lineages in Hippo signalling GO-FC. D4, day 4; D5, day 5; D6, day 6. The names for the in vitro clusters include the days from which they are derived and a letter label (**a-c**) to distinguish different clusters that occur at the same timepoint.

analysis (Fig. 6e and Extended Data Fig. 9c,d), the gut-like cluster 6 aligns to midgut. This cluster retains considerable visceral identity, consistent with the persistence of visceral signatures in several in vivo datasets^{7,19}. While key InterVE markers are most pronounced in cluster 2, the robust expression of these appears insufficient to drive alignment to InterVE and overall gene expression pulls this cluster towards PE.

Taken together, the combination of CAT, RNA velocity and marker analysis suggests that nEnd priming initiates a shift to InterVE/DE2 that is continued in nEnd spheroids as they give rise to gut derivatives. Although visceral signatures persist throughout differentiation and the exact sequence of differentiation events needs to be resolved, our findings suggest nEnd differentiation to gut-like derivatives proceeds via an InterVE-like intermediate.

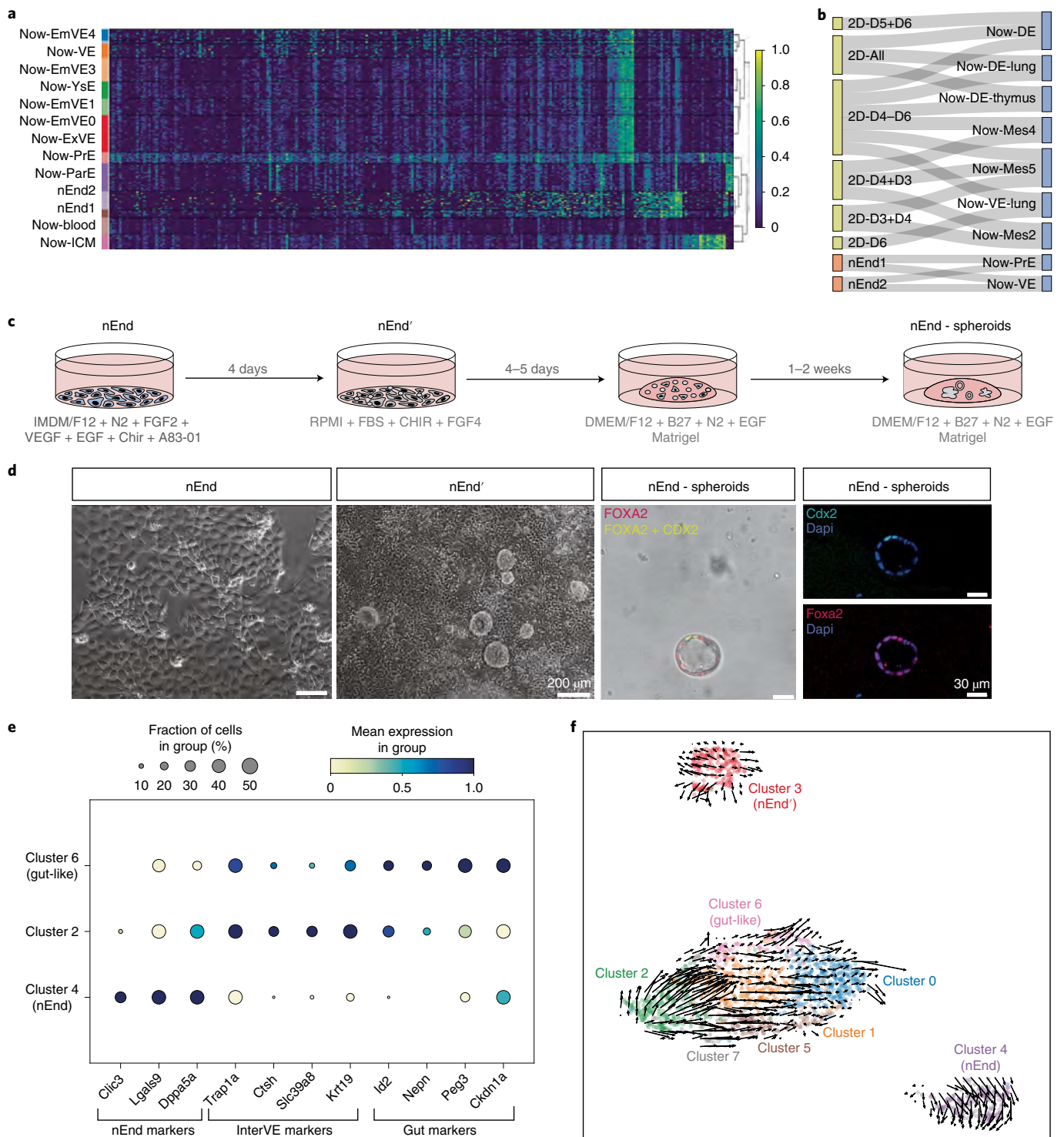


Fig. 6 | An in vitro model for generating gut spheroids (nEnd spheroids) from extra-embryonic endoderm. a, Heat map representing differentially expressed PrE genes from Now-2019 dataset. The two nEnd clusters, nEnd1 and nEnd2, form a hierarchical cluster together with Now-PrE and Now-ParE. **b**, Sankey diagram representing alignments between in vitro clusters from 2D-ESC differentiation and the Now-2019 dataset. nEnd aligns specifically to Now-PrE and Now-VE clusters. **c**, A scheme of the nEnd-spheroid differentiation protocol. **d**, Images of the cell cultures in the progression of nEnd-spheroid differentiation. Results are representative of four independent experiments. **e**, Dot plot showing nEnd, InterVE and gut markers in MARS-seq dataset consisting of nEnd cells differentiating into nEnd spheroids. The cluster 2 exhibits a pattern of InterVE markers while Cluster 6 (Gut-like) shows expression of gut markers. **(f)** RNA velocity showing the direction of differentiation from Cluster 2 that may represent a starting point for differentiation into the gut-like cluster and other nEnd-spheroid clusters.

Discussion

Empowered with diverse computational tools and MARS-seq, we uncovered the trajectories from visceral to definitive endoderm via

an intermediate cell type, InterVE. We placed in vitro differentiation in the context of in vivo development and assessed the sequence of developmental events leading to organ ontogeny in the endoderm.

Comparing single-cell RNA-seq datasets is plagued by a ‘curse of dimensionality’, as it requires the non-trivial assessment of $\sim 10^4$ gene-expression values. Paradoxically, the more genes characterized, the less meaningful cell–cell distances become with noise in gene expression further amplifying this effect^{52–54}. While most current methods reduce the number of dimensions by limiting data to a subset of genes, for example, those with highly variable gene expression^{33–35,45}, we focused on reducing noise as opposed to dimensions, performing comparisons with just one free parameter—the significance cut-off. CAT proved particularly useful when comparing *in vitro* datasets to *in vivo* lineages, where noise created by environmental differences tends to dominate co-localization in a dimensional embedding (for example t-SNE⁵⁵, UMAP³¹ or PCA⁵⁶). Using this tool we found PI3K activity sustains heterogeneity and anterior identity. Its inhibition homogenizes nascent DE towards a later endodermal cell type lacking regional identity (DE2), providing important developmental context for the common use of the PI3K inhibitor in human ESC differentiation towards different endodermal derivatives^{57,58}.

In vitro ESC differentiation has mostly been concerned with producing DE via a PS-like mesendoderm intermediate²³. While cells traversing the PS contribute to DE, recent studies suggest a substantial contribution of VE cells previously thought to be extra-embryonic in nature^{7,8,19}. In intestinal differentiation starting with extra-embryonic nEnd, cells embark on a visceral trajectory towards embryonic gut, suggesting that the inclusion of VE in ESC differentiation could help with the generation of more physiological cell types. Moreover, as CAT analysis suggests InterVE and DE2 identity are already induced in the second stage (nEnd’) of nEnd-spheroid differentiation, further optimization of the protocol described here could produce robust approaches towards the generation of visceral organs.

Our dataset relates levels of the central endoderm TF, FOXA2, to single-cell transcriptomes and suggests that the highest levels of FOXA2 are associated with embryonic signalling centres such as the node, AVE, ADE and FP, consistent with FOXA2 mutant phenotypes⁵⁹. A central feature of these organizing centres is the production of signalling antagonists, and our data suggest their induction requires the highest level of FOXA2 protein, implying they might represent low-affinity targets in canonical feedback inhibition circuits⁶⁰.

As FOXA2 expression delimited the progenitor set present in our data, it served to better resolve specific lineage relationships in the endoderm. Complemented by spatial transcriptomics, both CAT and RNA velocity suggest that different regions of VE contribute to the specific endoderm cluster, DE2, in a time-dependent fashion and these contributions span the entire A–P axis. The generation of embryonic gut via two distinct routes, InterVE and DE1, supports the continuum of DE differentiation in development. Early DE is derived from the epiblast via the PS starting at E6.5 and later endoderm recruitment via the VE occurs from E7.5, with both progenitor types converging on a more mature DE2 cell type. This biphasic recruitment to endoderm is consistent with the reported early lineage restriction event where epiblast-derived DE specification is finalized by E7.5 (ref. ⁶¹). The extent to which VE can produce additional endodermal cell types remains to be seen, but our work suggests that the use of both embryonic and ‘extra-embryonic lineages’ as starting endodermal cell types will benefit future stem cell applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41556-022-00923-x>.

Received: 15 January 2021; Accepted: 22 April 2022;
Published online: 9 June 2022

References

- Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* **25**, 221–251 (2009).
- Kraus, M. R. C. & Grapin-Botton, A. Patterning and shaping the endoderm *in vivo* and *in culture*. *Curr. Opin. Genet. Dev.* **22**, 347–353 (2012).
- Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Cell fate and cell lineage in the endoderm of the presomite mouse embryo, studied with an intracellular tracer. *Dev. Biol.* **115**, 325–339 (1986).
- Kwon, G. S., Viotti, M. & Hadjantonakis, A.-K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev. Cell* **15**, 509–520 (2008).
- Viotti, M., Nowotschin, S. & Hadjantonakis, A.-K. Afp::mCherry, a red fluorescent transgenic reporter of the mouse visceral endoderm. *Genesis* **49**, 124–133 (2011).
- Viotti, M., Nowotschin, S. & Hadjantonakis, A.-K. SOX17 links gut endoderm morphogenesis and germ layer segregation. *Nat. Cell Biol.* **16**, 1146–1156 (2014).
- Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
- Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
- Anderson, K. G. V. et al. Insulin fine-tunes self-renewal pathways governing naive pluripotency and extra-embryonic endoderm. *Nat. Cell Biol.* **19**, 1164–1177 (2017).
- Gardner, R. L. Investigation of cell lineage and differentiation in the extraembryonic endoderm of the mouse embryo. *Development* **68**, 175–198 (1982).
- Verheijen, M. H. & Defize, L. H. Signals governing extraembryonic endoderm formation in the mouse: involvement of the type 1 parathyroid hormone-related peptide (PTHrP) receptor, p21Ras and cell adhesion molecules. *Int. J. Dev. Biol.* **43**, 711–721 (1999).
- Lu, C. C., Brennan, J. & Robertson, E. J. From fertilization to gastrulation: axis formation in the mouse embryo. *Curr. Opin. Genet. Dev.* **11**, 384–392 (2001).
- Beddington, R. S. & Robertson, E. J. Axis development and early asymmetry in mammals. *Cell* **96**, 195–209 (1999).
- Belo, J. A. et al. Cerberus-like is a secreted factor with neuralizing activity expressed in the anterior primitive endoderm of the mouse gastrula. *Mech. Dev.* **68**, 45–57 (1997).
- Costello, I. et al. Lhx1 functions together with Otx2, Foxa2, and Ldb1 to govern anterior mesendoderm, node, and midline development. *Genes Dev.* **29**, 2108–2122 (2015).
- Robb, L. & Tam, P. P. L. Gastrula organiser and embryonic patterning in the mouse. *Semin. Cell Dev. Biol.* **15**, 543–554 (2004).
- Burtscher, I. & Lickert, H. Foxa2 regulates polarity and epithelialization in the endoderm germ layer of the mouse embryo. *Development* **136**, 1029–1038 (2009).
- Scheibner, K. et al. Epithelial cell plasticity drives endoderm formation during gastrulation. *Nat. Cell Biol.* **23**, 692–703 (2021).
- Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
- Chu, L.-F. et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17**, 173 (2016).
- Morrison, G. M. et al. Anterior definitive endoderm from ESCs reveals a role for FGF signaling. *Cell Stem Cell* **3**, 402–415 (2008).
- Yasunaga, M. et al. Induction and monitoring of definitive and visceral endoderm differentiation of mouse ES cells. *Nat. Biotechnol.* **23**, 1542–1550 (2005).
- Tada, S. et al. Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture. *Development* **132**, 4363–4374 (2005).
- Zhong, W. et al. Wnt and Nodal signaling simultaneously induces definitive endoderm differentiation of mouse embryonic stem cells. *Rom. J. Morphol. Embryol. Rev. Roum. Morphol. Embryol.* **58**, 527–535 (2017).
- Ang, S. L. et al. The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of HNF3/forkhead proteins. *Dev. Camb. Engl.* **119**, 1301–1315 (1993).
- Burtscher, I., Barkey, W. & Lickert, H. Foxa2-venus fusion reporter mouse line allows live-cell analysis of endoderm-derived organ formation. *Genesis* **51**, 596–604 (2013).
- Jaitin, D. A. et al. Massively parallel single cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
- MGI-Mouse Gene Expression Database (GXD) (MGI); (Acc. date 12/04/2019) <http://www.informatics.jax.org/expression.shtml>

29. Kojima, Y. et al. The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell Stem Cell* **14**, 107–120 (2014).
30. Snow, M. H. L. Gastrulation in the mouse: growth and regionalization of the epiblast. *Development* **42**, 293–303 (1977).
31. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
32. McKnight, K. D., Hou, J. & Hoodless, P. A. Dynamic expression of thyrotropin-releasing hormone in the mouse definitive endoderm. *Dev. Dyn.* **236**, 2909–2917 (2007).
33. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
34. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
35. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
36. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
37. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
38. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
39. g:Profiler – a web server for functional enrichment analysis and conversions of gene lists; (Acc. date 3/9/2029) <https://biit.cs.ut.ee/gprofiler/gost>
40. molecular_function Gene Ontology Term (GO:0003674); (Acc. date 3/9/2019) http://www.informatics.jax.org/vocab/gene_ontology
41. Duncan, S. A. et al. Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst. *Proc. Natl Acad. Sci. USA* **91**, 7598–7602 (1994).
42. Mfopou, J. K. et al. Efficient definitive endoderm induction from mouse embryonic stem cell adherent cultures: a rapid screening model for differentiation studies. *Stem Cell Res.* **12**, 166–177 (2014).
43. Villegas, S. N. et al. PI3K/Akt1 signalling specifies foregut precursors by generating regionalized extra-cellular matrix. *eLife* **2**, e00806 (2013).
44. Rothová, M., Hölzenspies, J. J., Livigni, A., Villegas, S. N. & Brickman, J. M. Differentiation of mouse embryonic stem cells into ventral foregut precursors. *Curr. Protoc. Stem Cell Biol.* **36**, 1G.3.1–1G.3.12 (2016).
45. Bhaduri, A. et al. Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578**, 1–7 (2020).
46. Chen, Y.-J. J. et al. Single-cell RNA sequencing identifies distinct mouse medial ganglionic eminence cell types. *Sci. Rep.* **7**, 45656 (2017).
47. Hwang, Y. S. et al. Reconstitution of prospermatogonial specification in vitro from human induced pluripotent stem cells. *Nat. Commun.* **11**, 5656 (2020).
48. Simunovic, M. & Brivanlou, A. H. Embryoids, organoids and gastruloids: new approaches to understanding embryogenesis. *Dev. Camb. Engl.* **144**, 976–985 (2017).
49. Amadei, G. et al. Inducible stem-cell-derived embryos capture mouse morphogenetic events in vitro. *Dev. Cell* **56**, 366–382 (2020).
50. Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011).
51. Zhang, R.-R. et al. Human iPSC-derived posterior gut progenitors are expandable and capable of forming gut and liver organoids. *Stem Cell Rep.* **10**, 780–793 (2018).
52. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. in *Database Theory — ICDT 2001* (eds. Van den Bussche, J. & Vianu, V.) 420–434 (Springer, 2001). https://doi.org/10.1007/3-540-44503-X_27
53. Vershynin, R. (ed.) *High-Dimensional Probability: An Introduction with Applications in Data Science* 38–69 (Cambridge University Press, 2018). <https://doi.org/10.1017/9781108231596.006>
54. Blum, A., Hopcroft, J. & Kannan, R. *Foundations of Data Science* (Cambridge University Press, 2020). <https://doi.org/10.1017/9781108755528>
55. Maaten, L. vander & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
56. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. <https://doi.org/10.1080/14786440109462720> (1901).
57. McLean, A. B. et al. Activin A efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3-kinase signaling is suppressed. *Stem Cells* **25**, 29–38 (2007).
58. Ortmann, D. et al. Naive pluripotent stem cells exhibit phenotypic variability that is driven by genetic variation. *Cell Stem Cell* **27**, 470–481.e6 (2020).
59. Ang, S. L. & Rossant, J. HNF-3 beta is essential for node and notochord formation in mouse development. *Cell* **78**, 561–574 (1994).
60. Ptashne, M. *A genetic switch: phage [lambda] and higher organisms* (Cell Press: Blackwell Scientific Publications, 1992).
61. Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Dev. Camb. Engl.* **113**, 891–911 (1991).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Methods

Statistical analysis and reproducibility. No statistical methods were used to pre-determine sample size. Data distribution was assumed to be normal, but this was not formally tested. The experiments were not randomized. Data collection and analysis were not performed blind to the conditions of the experiments.

Data collection was performed using Microsoft Office Excel (16.16.2), and statistical analysis was performed using GraphPad Prism (9) (Fig. 3b and Extended Data Fig. 9c). All statistical tests performed, sample sizes and *P* values are specified in figure legends.

Mouse maintenance and embryo collection. FOXA2:Venus mice²⁶ were generously provided by Heiko Lickert. Animal work was authorized by the Danish National Animal Experiments Inspectorate (Dyreforsøgstilsynet, license no. 2018-15-0201-01520) and performed according to national guidelines. Mice were kept in rooms at a temperature of 22°C (±2°C), with a humidity of 55% (±10%), air in the room was changed eight to ten times per hour, according to Danish regulations for animal experiments. Natural mating was set up in the evening. Mouse females (C57BL/6; 8–30 weeks old) were mated with homozygous FOXA2:Venus males (8–60 weeks old) to produce heterozygote FOXA2:Venus embryos and ensure uniform FOXA2:Venus expression. Females were checked for plugs the following morning, which was established as E0.5. Embryos were collected from E6.5 until E9.5. E8.5 embryos were sorted on the basis of somite number into more detailed stages (E8.0, E8.25, E8.5, E8.75 and E9.0) and divided into anterior and posterior parts. E9.5 embryos were dissected to minimize neural tube and increase the gut cell contribution (Extended Data Fig. 2c).

Mouse ESC culture and differentiation. Mouse ESCs used in this study are E14JU, with a 129/Ola background or Gsc-GFP/Hhex-RedStar double reporter line⁴³. Both lines were cultured in 0.1% gelatin-coated plates and maintained in complete ESC medium⁶²: GMEM (Sigma G5154) supplemented with 10% FBS (Gibco), 1× MEM non-essential amino acids (Thermo Fisher 11140-035), 2 mM L-glutamine (Thermo Fisher 25030-24), 1 mM sodium pyruvate (Thermo Fisher 11360-039), 100 μM 2-mercaptoethanol (Sigma) and 1,000 U ml⁻¹ leukaemia inhibitory factor (LIF) (made in house).

For naïve culture condition⁶³, ESCs were maintained in N2B27 supplemented with 3 μM GSK3βi (CHIR99021; Axon) and 1 μM MEKi (PD0325901; Sigma) and 1,000 U ml⁻¹ LIF.

DE, nEnd and nEnd-gut spheroid differentiations. Detailed overview of all DE differentiations^{43,44}, nEnd differentiation and culture^{9,64,65}, and gut spheroid protocols^{51,66} are summarized in Supplementary Dataset 8.

Immunofluorescence. The nEnd spheroids were washed and fixed in 4% PFA at room temperature for 10 min (Fisher Scientific, PI-28906). Samples were subsequently permeabilized in 0.5% Triton and blocked in 3% donkey serum. Primary antibodies were incubated in 1% donkey serum, 0.1% Triton in PBS at 4°C for 48 h, and subsequently incubated with the appropriate secondary antibody (AlexaFluor, Molecular probes) and DAPI at room temperature for 2 h. See Supplementary Dataset 11 for a list of antibodies and concentrations used. The samples were imaged using Leica SP8 confocal microscope with Las X software (3.5.7.23225) and processed in Imaris 9.6.

RT-qPCR. Total RNA was collected using the RNeasy Mini Kit (Qiagen). One micogram of total RNA was used for first-strand synthesis using SuperScript III reverse transcriptase according to the manufacturer's instructions. Complementary DNA corresponding to 10 ng of total RNA was used for RT-qPCR analysis using the Roche LC480 LightCycler (1.5.1.62.SP3), and target amplification was detected with the Universal Probe Library system. See Supplementary Dataset 12 for a list of primers and probes used.

Single-cell preparation. Embryos were dissociated with accutase (Sigma) into single cells immediately after collection. The collected embryos were mixed with mouse ESCs, which were counterstained with CellVue Maroon Cell Labeling Kit (Thermo Fisher, #88-0870-16) to enlarge the number of cells in the sample and to avoid loss of the scarce FOXA2^{POS} cells during centrifugation. Samples were then incubated with accutase and DNase (10 U μl⁻¹) at 37°C for 8–10 min and pipetted up and down multiple times to ensure a good single-cell suspension. The accutase was diluted by addition of FACS buffer (10% FBS in PBS). The cells were washed with FACS buffer twice and re-suspended in FACS buffer with DAPI in a round-bottom polystyrene tube.

The in vitro differentiated cells (Gsc-GFP/Hhex-RedStar double reporter line) were dissociated by 0.1% trypsin and transferred to a FACS-DAPI buffer in a FACS collection tube. The E14JU differentiated ESCs were stained with APC-conjugated CXCR4 (BD Bioscience, #558644, at dilution 1:400 in FACS buffer) and E-CAD antibodies (eBioscience, #50-3249, at dilution 1:400 in FACS buffer) for 20 min and washed three times in FACS buffer.

Flow cytometry and single-cell index sorting. Cells were sorted using a BD FACS Aria III (BD FACSDiva Software version 8) with a 100 μm nozzle and 20 psi

sheath pressure. FCS Express version 6 was used for post-acquisition analysis. The boundary between positive and negative populations was set on the basis of negative population of control ESCs. Forward scatter (FSC) and side scatter (SSC) were used to define a homogeneous population, FSC-H/FSC-W gates were used to exclude doublets and dead cells were excluded on the basis of DAPI staining. A gating strategy example can be found in Extended Data Fig. 2a.

Sorting speed was kept at 100–300 events per second to eliminate sorting two or more cells into one well. Single-cell sorting was verified colorimetrically⁶⁷. Cells were sorted directly into lysis buffer containing the first RT primer and RNase inhibitor, immediately frozen and later processed by the MARS-seq1 protocol²⁷.

Single-cell RNA-seq low-level processing and filtering. All single-cell MARS-seq libraries were sequenced using Illumina NextSeq500 (4.0.1) at a median sequencing depth of 225,000 reads per single cell. For detailed statistics in single-cell resolution on barcodes, reads, mapping and genes, see Supplementary Dataset 13. Sequences were mapped to mouse mm9 genome, de-multiplexed and filtered^{27,68}, extracting a set of unique molecular identifiers (UMIs) that define distinct transcripts in single cells for further processing. We estimated the level of spurious UMIs in the data using statistics on empty MARS-seq wells²⁷. Mapping of reads was done using HISAT (version 0.1.6) (ref. 69); reads with multiple mapping positions were excluded. Reads were associated with genes if they mapped to an exon. The pre-processing single-cell RNA-seq pipeline can be found at https://tanaylab.github.io/old_resources/pages/672.html.

Pre-processing of single-cell data for downstream analysis. To pre-process our data, we loaded the entire raw count matrix dataset (in vivo and in vitro) together with the metadata using scanpy⁷⁰. We filtered away cells with very low numbers of genes detected. Our cut-off for minimum number of genes was 1,436. This is the lower tail of the distribution, corresponding to the median of gene detected per cell, minus one standard deviation of gene detected per cell. When we performed CAT analysis, we also removed genes with no overall expression and genes present only in the cells with top 500 most read counts.

Seurat processing pipeline. The raw dataset was subset using filtered cells and converted to Seurat (v3.1.3) object together with metadata. Additionally, empty Zero stage was removed as well ERCC genes. As estimated mitochondrial content was below 2%, no extra filtering was necessary. The raw UMI counts were normalized to 10,000 and log-transformed using 'NormalizeData' followed by identifying 2,000 highly variable genes using 'FindVariableFeatures' with default settings. For batch correction, we used 'RunFastMNN' from batchelor package, which was integrated into Seurat's ecosystem. The corrected counts were used to construct a shared nearest neighbour graph using the first 20 dimensions. We used Louvain clustering with resolution of 1.2 to identify higher number of clusters in the dataset followed by UMAP visualization using again the first 20 dimensions. Lastly, for each cell in the dataset, we estimated cell cycle phase by converting Seurat's cell cycle genes from human to mouse⁷¹ and used these as input for the 'CellCycleScoring' function.

PAGA. To perform trajectory and pseudotime inference, we replicated the steps from the previous analysis using Python package scanpy. We initially de-noised the dataset using built-in function 'diffmap' with default settings followed by running 'paga' for previously annotated clusters (excluding PE), which construct a connected graph representation of the dataset. Finally, we set cluster PS1 as our initial starting point for inferring pseudotime using 'dpt' function.

CAT. Averaging over clusters to lower noise. Our method is based on exploiting pre-defined clusters that can be obtained by commonly used clustering algorithms⁷² or defined by, for example, meta-cell analysis⁷³.

In high dimensions, noise effectively pushes all datapoints away from each other, limiting the ability to meaningfully distinguish between nearest and furthest neighbours of cells. To overcome this, we averaged out the noise by calculating the mean gene expression of each cluster, thus increasing the contrast between cluster-cluster distances, with the cost of limiting us to cluster-scale resolution.

Normalizing gene expression to avoid bias towards highly expressed genes in distance calculation. All single-cell datasets were normalized to 10,000 transcripts per cell, except the data for organoid comparisons where we did log normalization. Log normalization compensated for lower read depth in the organoid data. Before calculating the distance between the cluster means, we normalized each gene, x_i , by the non-zero median, that is, median expression for all cells expressing the gene. Thus, a vector of genes, $X = \{x_i\}$ is normalized to $\bar{X} = \frac{X}{\text{NonZeroMedian}}$, where $\text{NonZeroMedian} = \text{median}(X = \{x_i\}, x_i > 0)$. This normalization ensured that high-expressed genes contribute to the distance calculation to a similar extent as low-expressed genes.

As the only adjustment to the Now-2019 dataset, we re-scaled the gene expression such that the distributions of median gene expression in both datasets were matching.

Calculating distance between clusters. The distance between all pairs of cluster means was calculated using the Euclidean norm for simplicity; however, different distance metrics also work.

Bootstrapping the distances to get uncertainties. To determine the uncertainties in the distance measurements, we bootstrapped each cluster 1,000 times with replacement, to its original size, and calculated the distances between averages of each re-sampled cluster. For each cluster pair, we thus obtained distribution of cluster–cluster distances and corresponding mean and standard deviation. The resulting distributions, based on sampled averages with replacement, are approximately Gaussian.

Significant nearest neighbours. The closest neighbour was always counted as a nearest neighbour. If the difference between the closest and another neighbour was less than 1.6 standard deviations (corresponding to $P > 0.05$, one-sided), we deemed that we cannot meaningfully distinguish which is the nearest, and both clusters were counted as the nearest neighbours. All CAT comparisons and alignments are directional: from A's perspective, B may be a closest match, although from B's perspective, some other cluster may match better.

Sankey diagram. To visualize the CAT alignments, we used Sankey plot from 'plotly' library in R, where the thickness of the line is inversely proportional to the pairwise distance (Figs. 2c and 6b and Extended Data Fig. 3c–e). In Fig. 5b–e and Extended Data Fig. 8b,c, thickness of the lines is given by the cluster height divided by the number of alignments.

Analysing the alignment of FP to Now-Mes and Now-lung. To find out why FP aligns to Now-Mes and Now-lung, we did pairwise DEG in the following way. First, we identified list of DEGs between (1) FP and the Now-lineages (excluding Now-Mes4 and Now-DE-lung), (2) Now-Mes4 and Now-lineages (excluding Now-DE-lung) and (3) Now-lung and Now-lineages (excluding Now-Mes4). We considered only upregulated genes with an adjusted (Benjamini–Hochberg) P value $< 10^{-10}$. The genes common to the three gene lists (overlap of gene lists in (1)–(3)) are the genes that are consistently upregulated among the three clusters compared with the rest and contribute to the particular alignment of the FP to Now-Mes4 and Now-DE-lung. Passing these genes through MGI-GXD²⁸ shows that they are expressed in neural lineages, thus explaining their presence in our dataset's FP cluster.

Networks. Networks were visualized by directionally connecting clusters to their nearest neighbours with the strength of the connection equal to 1 divided by the mean bootstrapped distance. The layout of the network was calculated using the Fruchterman–Reingold force-directed algorithm implemented in Python framework NetworkX.

Similarity score. Similarity score was calculated as follows:

First, for each cluster or condition, we pre-defined the set of in vivo lineage targets, that is, in vivo lineages that in vitro cells were expected to reproduce. We referred to these as targets, $T = \{T_1, T_2, \dots\}$, where, for example, $T_1 = \text{PS1}$ (Supplementary Dataset 10).

Second, for each in vitro cluster 'D', we compared the obtained n_{DE} CAT alignments with the in vivo lineages, $E = \{E_1, E_2, \dots, E_{n_{\text{DE}}}\}$ with the set of targets, T . (Extended Data Fig. S8e, for example $n_{\text{DE}} = 3$ in bottom left).

If E and T do not overlap, that is, have no lineages in common, we assume no similarity and set similarity score, $S = 0$ (Extended Data Fig. 8e, top right). If E and T overlap, that is, share at least one lineage, we evaluate to what extent the alignments to non-target lineages is erroneous and to what extent these reflect that fact that all lineages from CAT alignments share similar traits due to overlapping function (for example, a given signalling pathway).

We can quantify this functional similarity among any n_{DE} in vivo lineages as a fraction of identified vivo-to-vivo alignments, n_{EE} , to the number of all possible alignments among n_{DE} lineages, $n_{\text{DE}}(n_{\text{DE}} - 1)$ (number of links in a fully connected graph with n_{DE} nodes). Thus, the functional similarity $\text{FS} = \frac{n_{\text{EE}}}{n_{\text{DE}}(n_{\text{DE}} - 1)}$ is 0 when no in vivo lineages align to each other and $\text{FS} = 1$ when all align to all.

We can then quantify the unspecificity of a given alignment as $U = \frac{n_{\text{DE}}}{N_{\text{E}}} (1 - \text{FS})$. When in vivo lineages do not share traits in common, ($\text{FS} = 0$), the unspecificity is given by the fraction of possible vitro-to-vivo alignments divided by the number of all in vivo lineages, $N_{\text{E}}, \frac{n_{\text{DE}}}{N_{\text{E}}}$. On the other hand, when all lineages are completely similar to each other ($\text{FS} = 1$), the expression becomes 0, and the large number of vitro-to-vivo alignments does not count towards low specificity.

As it is more natural to think in terms of specificity (and similarity), rather than unspecificity, we define similarity score as $S = 1 - U = 1 - \frac{n_{\text{DE}}}{N_{\text{E}}} \left(1 - \frac{n_{\text{EE}}}{n_{\text{DE}}(n_{\text{DE}} - 1)} \right)$. Examples of how the similarity is calculated are shown in Extended Data Fig. 8e.

To estimate how in vitro conditions perform, we calculated similarity scores for each cluster and for each GO-FC (Supplementary Dataset 10). A similarity score per condition (Supplementary Dataset 10) was calculated by taking a median over cluster-specific similarity scores in each condition (Supplementary Dataset 10). To identify genes that differ most from the target lineages, we ranked gene contributions to the distances and listed those contributing with 0.1 or more as the genes causing low similarity.

FACS plots. The position of each cluster in the plot was calculated by first doing a kernel density estimation over the log of the protein fluorescent signal from the FACS (GFP-A) index sort data and the log of the processed single-cell transcript count of the same, for all cells in each cluster and then finding the maximum of the distribution. The calculation was done using the Gaussian kernel density estimation function from the `scipy.stats`⁷⁴ using a Python framework with default parameters.

The pie charts summarizing the cell-cycle composition of each cluster's cells was found using the 'CellCycleScoring' mentioned in the Seurat processing pipeline. The size of the pie charts represents the normalized average size of the cells in the cluster. The information about the size is the log of the FSC-A (area) signal from Cell Sorter. The sizes are normalized so the smallest cluster has a log (FSC-A) value of 10.28 and the biggest has one of 11.47.

Resolve BioSciences' MC technology. *Sample preparation and cryo-sectioning.* FOXA2^{Venus} embryos were collected at E7.5 and fixed in PAXgene fixative (QIAGEN) for 1 h, then stabilized in PAXgene stabilizer (QIAGEN) for 2 h and transferred to 30% sucrose in PBS for 30 min. The embryos were transferred into OCT (Sakura) and snap-frozen in liquid nitrogen. The samples were stored at -80°C until cryo-sectioning into 10 μm sections using a HM560 cryostat onto Resolve cover slips without any freeze–thaw cycle.

MC. Tissue sections were thawed and used for MC according to the manufacturer's instructions (protocol 3.0). Briefly, tissues were primed followed by overnight hybridization of all probes specific for the target transcripts. Samples were washed the next day and fluorescently tagged in a two-step colour development process. Regions of interest were imaged⁷⁵ and fluorescent signals removed during de-colourization. Colour development, imaging and de-colourization were repeated for a total of eight rounds to build a unique combinatorial code for every target transcript that was derived from raw images as described below.

Probe design. The probes for 27 genes were designed using Resolve's proprietary design algorithm⁷⁶. Supplementary Dataset 14 highlights the gene names and catalogue numbers for the specific probes designed by Resolve BioSciences.

Imaging, spot segmentation and decoding. Imaging and signal decoding were performed according to standard procedures^{76–77}. Briefly, samples were imaged on a Zeiss Celldiscover7, using the 50 \times Plan Achromat water immersion objective with an NA of 1.2 and the 0.5 \times magnification changer, resulting in a 25 \times final magnification. Imaging was automated with a custom Python script using the scripting API of the Zeiss ZEN 3.2 software (open application development). All images and transcript coordinates are based on a pixel size of 138 \times 138 nm.

Images were corrected for background fluorescence. The brightest maxima per plane were determined and the z -groups with the highest absolute brightness were stored as a 3D-point cloud. To align the raw data images from different imaging rounds, images had to be corrected. To do so, the extracted feature point clouds were used to find the transformation matrices with an iterative closest point cloud algorithm to minimize the error between two point clouds. On the basis of the transformation matrices, the corresponding images were aligned with rigid transformation using trilinear interpolation.

The aligned images were used to create a profile for each pixel that was used to decode signals and define transcript location in x , y and z . The fraction of false positives in the decoded signals was estimated by decoding signals for blank codes that were not assigned to a probe in the experiment.

Final signal segmentation and decoding. **Cell segmentation.** To assign RNA counts to individual cells, we have segmented cells using QuPath⁷⁸ on DAPI images with following adjusted settings: pixel width and height (0.138 μm), sigma (1 μm), minimum area (2 μm^2) and cell expansion (3 μm). Using proprietary ImageJ⁷⁹ plugin from Resolve BioSciences (PolyLux), we were able to import the cell segmentation borders and extract a final count matrix similar to that obtained in single-cell RNA-seq experiment. The spatial information derived from the image is in count matrix format, with rows representing single RNAs and columns representing its x and y positions of mRNA inside the identified cell. To visualize mRNA counts in single cells preserving spatial information, as in Fig. 4a, the count matrix was processed by Python package Squidpy⁸⁰ and for each segmented cell we computed centroid using shapely package.

Decision trees to identify InterVE features or genes from MARS-seq data. To find the most important genes for InterVE identification, we used a supervised decision tree implemented in the XGBoost python package⁸¹. The MARS-seq dataset was subset to E7.5 cells, non-zero median normalized and split into training (70%) and validation set (30%). Each cell was labelled as either InterVE or not InterVE. We used GridSearchCV⁷⁹ to estimate the best parameters for the binary model. Finally, tenfold cross-validation was used to estimate the overall prediction accuracy. We have re-trained the model ten more times with different random seed and stored features (genes important for InterVE classification) from each iteration. The final list of InterVE markers (Fig. 4c) was compiled manually, taking into account the consistency of feature/gene ranks in the ten repeats and

the distinct expression pattern across VE, DE and InterVE that made the most biological sense.

Decision trees to identify InterVE cells from MC data. To predict InterVE cells from MC data, we have trained a final model on MARS-seq data using only 25 genes (Extended Data Fig. 6b,c) following all the steps above. By running this final model on non-zero-median-normalized MC data, we were able to identify InterVE cells in individual sections. The selected single RNAs (Fig. 4d) as well as outlines of predicted InterVE cells were visualized using ImageJ and PolyLux.

RNA velocity. To our knowledge, MARS-seq reads are not supported by any RNA velocity⁸² tools at the moment. We have therefore created a custom in-house script that converts the reads into 10x format allowing us to run 'StarSolo' with the 'scVelo' feature described in refs.^{82,83}. The script can be found on our GitHub repository.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The single-cell RNA-seq data used in this study is in the Gene Expression Omnibus under accession number GSE164464. Previously published Nowotchin-2019 data that were re-analysed here are available under accession code GSE123046. Source data are provided with this study. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Code availability

All analyses, package versions as well as development environments for reproducibility purposes are publicly available at <https://github.com/brickmanlab/rothova-et-al-2022>. Implementation of CAT and the related pre-processed datasets are publicly available at <https://github.com/brickmanlab/CAT>. The interactive CAT app is available at <https://align-clusters.com>.

References

62. Riveiro, A. R. & Brickman, J. M. From pluripotency to totipotency: an experimentalist's guide to cellular potency. *Development* **147**, dev189845 (2020).
63. Ying, Q.-L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
64. Linneberg-Agerholm, M. et al. Naïve human pluripotent stem cells respond to Wnt, Nodal and LIF signalling to produce expandable naïve extra-embryonic endoderm. *Development* **146**, dev180620 (2019).
65. Linneberg-Agerholm, M. & Brickman, J. M. in *Human Naïve Pluripotent Stem Cells* (ed. Rugg-Gunn, P.) 105–116 (Springer US, 2022). https://doi.org/10.1007/978-1-0716-1908-7_8
66. McCracken, K. W., Howell, J. C., Wells, J. M. & Spence, J. R. Generating human intestinal tissue from pluripotent stem cells in vitro. *Nat. Protoc.* **6**, 1920–1928 (2011).
67. Rodrigues, O. R. & Monard, S. A rapid method to verify single-cell deposition setup for cell sorters. *Cytometry A* **89**, 594–600 (2016).
68. Keren-Shaul, H. et al. MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nat. Protoc.* **14**, 1841–1862 (2019).
69. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
70. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
71. cell-cycle genes (cc.genes) for mouse · Issue #2493 · satijalab/seurat (GitHub) <https://github.com/satijalab/seurat/issues/2493>
72. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
73. Giladi, A. et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).
74. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
75. Groiss, S. et al. Highly resolved spatial transcriptomics for detection of rare events in cells. <https://doi.org/10.1101/2021.10.11.463936> (2021).
76. Guillemins, M. et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* **185**, 379–396.e38 (2022).
77. D'Gama, P. P. et al. Diversity and function of motile ciliated cell types within ependymal lineages of the zebrafish brain. *Cell Rep.* **37**, 109775 (2021).
78. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
79. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
80. Palla, G. et al. Squidpy: a scalable framework for spatial single cell analysis. <https://doi.org/10.1101/2021.02.19.431994> (2021).
81. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>
82. Bergen, V. et al. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
83. Bergen, V. et al. RNA velocity—current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021).

Acknowledgements

We thank H. Lickert for the kind gift of the FOXA2^{Venus} reporter mice, the DanStem Genomics Platform (H. Neil, M. Michaut and H. Wollmann), DanStem Flow Cytometry Platform (G. dela Cruz and P. van Dieken), Stem Cell Culture Platform (A. Meligkova, E. F. Rebollo and M. Paulsen), DanStem Imaging Platform (J. Bulkescher and A. Shrestha) for technical expertise, support and the use of instruments, M.S. Paulsen for facilitating ST, Brickman laboratory members for critical discussions, K. J. Won for comments and T. Machet for graphical advice and help. Raw sequencing data were converted from bcl to fastq using the DeIC National Life Science Supercomputer at DTU (www.computerome.dk).

Work in our groups was supported by grants from the Lundbeck Foundation (R198-2015-412), J.M.B.; Independent Research Fund Denmark (8020-00100B and 6110-00009), J.M.B.; the Novo Nordisk Foundation (NNF17OC0028218), J.M.B.; the Danish National Research Foundation (DNRF116), J.M.B. and A.T. The Novo Nordisk Foundation Center for Stem Cell Medicine is supported by Novo Nordisk Foundation (NNF21CC0073729), and its predecessor The Novo Nordisk Foundation Center for Stem Cell Biology was also supported by the Novo Nordisk Foundation (NNF17CC0027852).

Author contributions

M.M.R., J.M.B. and A.T. conceived the study. M.M.R., J.M.B., A.T., A.V.N., A.R.R., M.L.-A. and Y.F.W. designed and interpreted experiments. M.M.R. performed all experiments, Y.F.W., A.R.R. and M.L.-A. contributed to the nEnd-spheroid experiments. The pre-processing and initial filtering of the raw sequencing data was done by E.D. All subsequent data analysis was carried out by A.V.N. and M.P. with input from I.A. M.M.R., J.M.B. and A.T. wrote the manuscript with input from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41556-022-00923-x>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41556-022-00923-x>.

Correspondence and requests for materials should be addressed to Ala Trusina or Joshua Mark Brickman.

Peer review information *Nature Cell Biology* thanks Heiko Lickert, Manu Setty, and Patrick Tam for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Chemicals orchestrate reprogramming with hierarchical activation of master transcription factors primed by endogenous *Sox17* activation

Zhenghao Yang^{1,9}, Xiaochan Xu^{2,9}, Chan Gu^{3,9}, Jun Li², Qihong Wu³, Can Ye⁴, Alexander Valentin Nielsen⁵, Lichao Mao¹, Junqing Ye⁶, Ke Bai¹, Fan Guo³✉, Chao Tang^{7,8}✉ & Yang Zhao^{1,2,4}✉

Mouse somatic cells can be chemically reprogrammed into pluripotent stem cells (CiPSCs) through an intermediate extraembryonic endoderm (XEN)-like state. However, it is elusive how the chemicals orchestrate the cell fate alteration. In this study, we analyze molecular dynamics in chemical reprogramming from fibroblasts to a XEN-like state. We find that *Sox17* is initially activated by the chemical cocktails, and XEN cell fate specialization is subsequently mediated by *Sox17* activated expression of other XEN master genes, such as *Sall4* and *Gata4*. Furthermore, this stepwise process is differentially regulated. The core reprogramming chemicals CHIR99021, 616452 and Forskolin are all necessary for *Sox17* activation, while differently required for *Gata4* and *Sall4* expression. The addition of chemical boosters in different phases further improves the generation efficiency of XEN-like cells. Taken together, our work demonstrates that chemical reprogramming is regulated in 3 distinct “prime–specify–transit” phases initiated with endogenous *Sox17* activation, providing a new framework to understand cell fate determination.

¹State Key Laboratory of Natural and Biomimetic Drugs, MOE Key Laboratory of Cell Proliferation and Differentiation, Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine, Peking University, 100871 Beijing, China. ²Peking-Tsinghua Center for Life Sciences, Peking University, 100871 Beijing, China. ³Center for Translational Medicine, Ministry of Education Key Laboratory of Birth Defects and Related Diseases of Women and Children, Department of Obstetrics and Gynecology, West China Second Hospital, Sichuan University, 610041 Chengdu, Sichuan, China. ⁴BioMed-X center, Peking University, 100871 Beijing, China. ⁵The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark. ⁶Boehringer Ingelheim International GmbH (China), 100027 Beijing, China. ⁷Center for Quantitative Biology, Peking University, 100871 Beijing, China. ⁸School of Physics, Peking University, 100871 Beijing, China. ⁹These authors contributed equally: Zhenghao Yang, Xiaochan Xu, Chan Gu. ✉email: guofan@scu.edu.cn; tangc@pku.edu.cn; yangzhao@pku.edu.cn

Somatic cells can be reprogrammed to become pluripotent by nuclear transfer into oocytes, by delivery of transcription factors or by treatment with a cocktail of chemicals^{1–3}. These somatic reprogramming techniques hold great promise in regenerative medicine for providing an unlimited source for functional cells. In comparison with the other two strategies, chemical reprogramming is attractive for future applications due to its non-integrative nature, ease to be standardized and temporally controlled, and lower tumorigenicity^{3,4}.

In recent years, the understanding of the cell dynamics and the molecular mechanisms of chemical reprogramming has gone deeper and broader. For instance, an extraembryonic endoderm (XEN)-like state bridges the chemical reprogramming towards chemically reprogrammed into pluripotent stem cells (CiPSCs) from different somatic cell types^{4,5}. Dynamic early-embryonic-like programs are found critical for the transition of XEN-like state into a pluripotent state⁶. In addition, the chemical reprogramming efficiency has been found greatly improved by additional chemical boosters, such as bromodeoxyuridine, retinoic acid agonists, Dolt1L inhibitors, and glycogen synthase kinase 3 inhibitors, and CiPSC can even be induced with a chemically defined medium^{4,7–9}.

Furthermore, chemical reprogramming strategies have been extended to inducing direct cell lineage conversion into functional cell types without an intermediate pluripotent state. For instance, neural progenitors¹⁰, functional neurons^{11,12}, cardiomyocytes^{13,14}, skeletal muscles¹⁵, brown adipocytes^{16,17}, astrocytes¹⁸, endoderm progenitor-like cells¹⁹, and photoreceptor-like cells²⁰ are reported to be induced from fibroblasts with chemicals alone. Besides, endoderm progenitor cells are induced from gut epithelium with pure chemicals²¹, and human fetal astrocytes are converted into functional neurons by chemical combinations²².

Intriguingly, the small molecules essential for XEN induction, CHIR99021 (a GSK3 inhibitor), 616452 (Repsox, an ALK5 inhibitor), and Forskolin (a cAMP agonist) have frequently been used for the direct induction of the many of different cell types noted above. Unlike the master genes used in transgenic reprogramming, which are associated with the target cell type, these chemicals always target signaling pathways that play roles in different cell types and are not associated with any specific cell lineage. Therefore, it is still unclear how the chemical cocktails determine the target cell type, and the molecular dynamics during chemically induced cell fate transition are still elusive²³.

Here, to better understand how chemically induced cell fate alteration and determination are orchestrated, we studied the chemical reprogramming process from fibroblasts to XEN-like cells in terms of the time-course and at the single-cell resolution. We revealed that cell fate transition was primed by endogenously expressed *Sox17*, which mediated further hierarchical activation of master transcription factors in chemical reprogramming. We further investigated the role of small molecules in various stages throughout the process.

Results

Chemically induced *Sox17* expression initiates XEN-like cell fate reprogramming. To investigate how the chemical cocktail determines XEN-like cell fate during C6FAE-mediated reprogramming (C, CHIR99021; 6, 616452; F, Forskolin; A, AM580; E, EPZ004777) (Fig. 1a), we analyzed the reprogramming process at 10 time points over a course of 20 days with single-cell RNA-sequencing (Fig. 1b). In comparison with the existing dataset⁶, our data detected more UMIs and genes, and the expression pattern of XEN and fibroblast master genes in various periods were comparable (Supplementary Fig. 1a–d). Importantly, MEF cells and XEN cells (day 20) merged perfectly with those in the

existing dataset (Supplementary Fig. 1e), indicating the fidelity of our single-cell RNA-seq data.

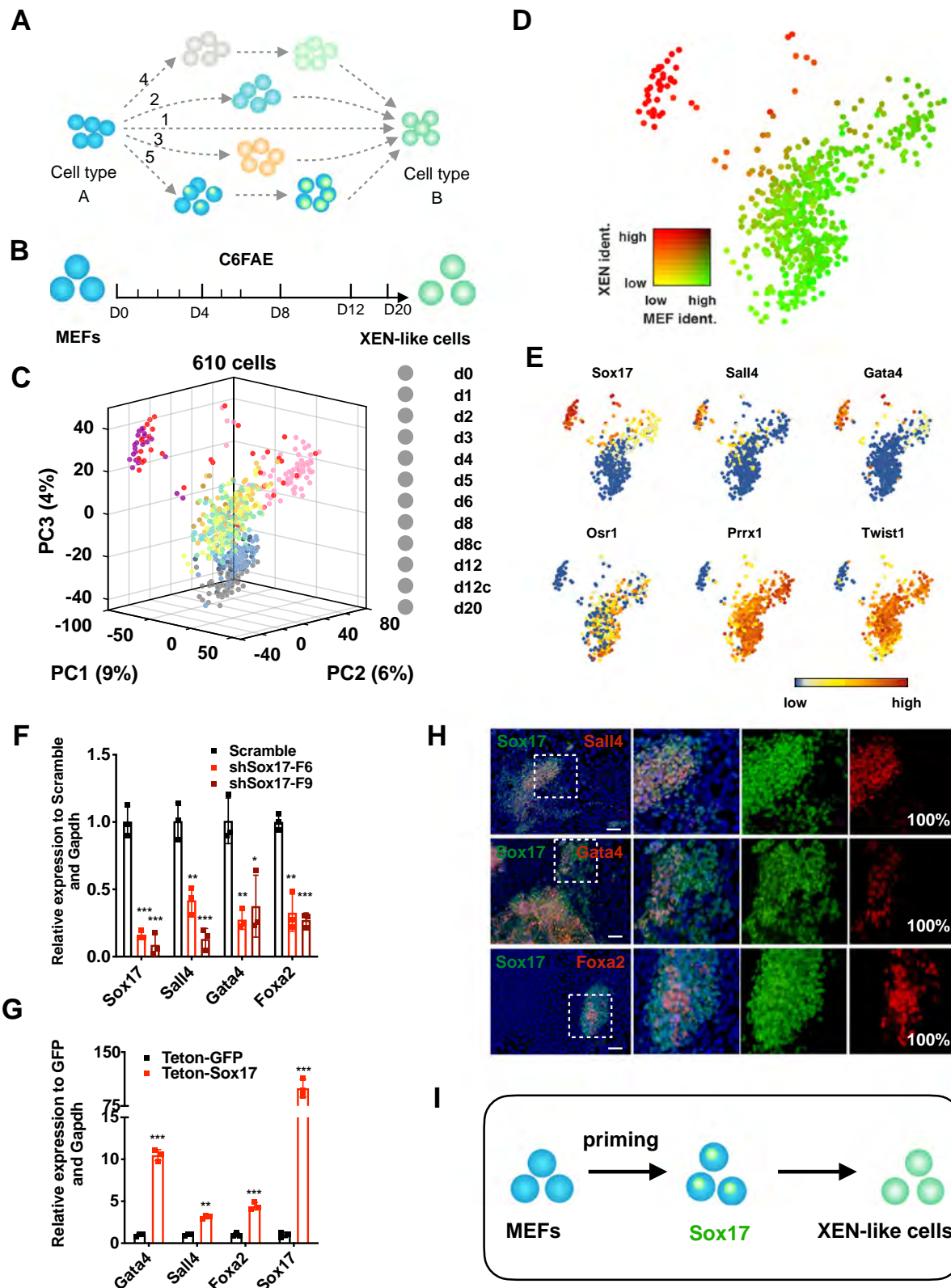
By principal component analysis (PCA) of single-cell RNA-seq data from days 0 to day 20, we found that cells in the earlier stage were quite close to each other and then separated gradually, and ultimately divided into two branches (Fig. 1c). This bifurcated reprogramming trajectory was also confirmed by pseudo-time analysis (Supplementary Fig. 1f). These two branches mainly consisted of cells from day 12 and day 20 and they belonged to different clusters in which cells were grouped using Louvain clustering with a resolution 0.95^{24–26} (Supplementary Fig. 1g). We evaluated cell identities by using the analytical technique based on quadratic programming with 100 genes representing for MEF cell identity and 100 genes representing for XEN cell identity²⁷. The left branch had established the major XEN identity without MEF identity (Fig. 1d and Supplementary Fig. 1h), indicating successful reprogramming into XEN-like cells and was further referred to as “the proceeding branch”. While, the right branch reserved the major profiles of MEF identity, without the establishment of XEN identity, which was further termed as “the trapped branch”. The cells located at the proceeding branch had a remarkable higher expression of the XEN master genes *Sox17*, *Sall4*, and *Gata4*, which were reported to promote the differentiation into XEN cells from mouse embryonic stem cells by forming a self-activation loop^{28–31} (Fig. 1e). The trapped branch included cells with a lower expression of *Sall4* and *Gata4* while still retained the high expression of fibroblast master genes, such as *Osr1*, *Prrx1*, and *Twist2* (Fig. 1e). Interestingly, very few cells had low scores of MEF identity before XEN-like cells were induced. This indicates no distinct de-differentiated, or other kinds of intermediate, cells during chemical reprogramming from MEFs to XEN-like cells (model 1–4 in Fig. 1a).

We noticed that the major differences between the proceeding and the trapped branches were the differential expression of XEN and fibroblast master genes. The expression of XEN master transcriptional factors (TFs), especially *Sox17*, were enriched in the proceeding branch (Fig. 1e and Supplementary Fig. 1i). Besides, the order of the activated expression of XEN master TFs was *Sox17*, *Sall4*, *Gata4*, and *Foxa2*, suggesting that *Sox17* was upstream of the other XEN TFs (Supplementary Fig. 1j–l).

In line with the above, we found that *Sox17* knockdown impaired the activation of most of the XEN TFs, *Gata4*, *Sall4*, and *Foxa2*, as well as XEN-like colony formation (Fig. 1f). Meanwhile, *Sox17* overexpression promoted the upregulation of *Sall4*, *Gata4*, and *Foxa2* (Fig. 1g). Then, we analyzed the co-expression of XEN master gene expression every day throughout the reprogramming process by immunofluorescence. We found that the expression of *Sox17* was detected as early as day 4. *Sall4*, *Gata4*, and *Foxa2*-expressing cells were all subpopulations of *Sox17*-expressing cells that emerged in day 5–8 (Fig. 1h), indicating *Sall4*, *Gata4*, and *Foxa2* were only activated in *Sox17*-positive cells.

These findings suggest that chemical-mediated XEN-like cell reprogramming is mediated by the endogenously activated *Sox17* in fibroblasts (Fig. 1i).

XEN-like cell fate specification and transition with the accumulated master TFs downstream of *Sox17*. To further investigate how the cell fate reached a XEN-like state after *Sox17* activation, we focused on the activation of *Gata4*, *Sall4*, and *Foxa2*. We found that *Gata4*-positive cells were a subset of *Sall4*-positive cells and no *Gata4*-positive/*Sall4*-negative cells appeared before day 6 by analyzing the co-expression of *Gata4* and *Sall4* using immunostaining. This suggests *Gata4* activation is only in *Sall4*-expressing cells (Fig. 2a, b). Afterward from day 7 to day 12,



the number of Gata4-positive colonies greatly increased while the number of Sall4-positive colonies declined (Fig. 2c). Finally, at day 12, Sall4-positive cells turned out to be a subpopulation of Gata4-positive cells (Fig. 2a). This was probably due to the self-repression function of Sall4 expression as reported³² or resulted from another wave of Gata4 activation without Sall4. Staining for Foxa2 revealed a subpopulation of Gata4 expressing cells, leading

us to believe that Gata4 might be upstream of Foxa2 in cell fate specification (Fig. 2d).

In a knockdown experiment of *Sall4* the expression of *Gata4* and *Foxa2* was severely disrupted, suggesting that *Sall4* is the upstream regulator of *Gata4* and *Foxa2*, which is consistent with the immunostaining data (Fig. 2e). Knockdown of *Gata4* decreased the expression of *Foxa2* but had no influence on the

Fig. 1 Chemically induced Sox17 expression initiates cell fate reprogramming towards XEN-like cells. **a** Potential models for cell fate reprogramming. The “directly switch model” proposes cell fate switch directly without any intermediate cell type (1); The “de-differentiate and re-differentiate model” indicates cell fate reprogramming mediated by a multipotent stem cell with the differentiation potential into both the initial and target cell types (2); The “discrete state transit model” assumes cell fate reprogramming process with gradual fading of the initial cell features and gradual formation of the target cell identities (3); The “reset and reconfigure model” refers to cell fate reprogramming with the erasing of initial cell identity before the establishment of target cell identity (4); The “prime, specify and transit model” indicates cell fate reprogramming with priming and specification state without substantial alteration of initial cell identities before cell fate transition into the target cell types (5). **b** Schematic diagram of chemical-mediated XEN-like cell reprogramming. **c** PCA projection of all individual cells during the reprogramming process. d8c, single cells picked from colonies of day 8; d12c, single cells picked from colonies of day 12; **d** MEF and XEN identity in the PCA projection. For each cell on the XEN reprogramming path, the similarity to bulk RNA-seq from either MEFs or XEN-like cells as calculated using quadratic programming. **e** Expression of XEN master genes (*Sox17*, *Sall4*, *Gata4*) and MEF master genes (*Osr1*, *Prrx1*, *Twist2*) in the PCA projection. **f** Relative mRNA levels of XEN genes induced by C6FAE on day 12 with the knockdown of *Sox17* ($n = 3$). **g** Relative expression of XEN genes induced by C6FAE on day 12 with the overexpression of *Sox17* ($n = 3$). **h** Co-staining of *Sox17* and other XEN master genes induced by C6FAE on day 12. Scale bar, 100 μm . The percentage of *Sall4*, *Gata4* and *Foxa2*-positive cells emerged in *Sox17*-positive cells were labeled in the lower right corner of each picture. **i** Schematic of the stepwise XEN induction mediated by *Sox17* activation. Significance was assessed compared with the controls using a one-tailed Student’s *t*-test. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

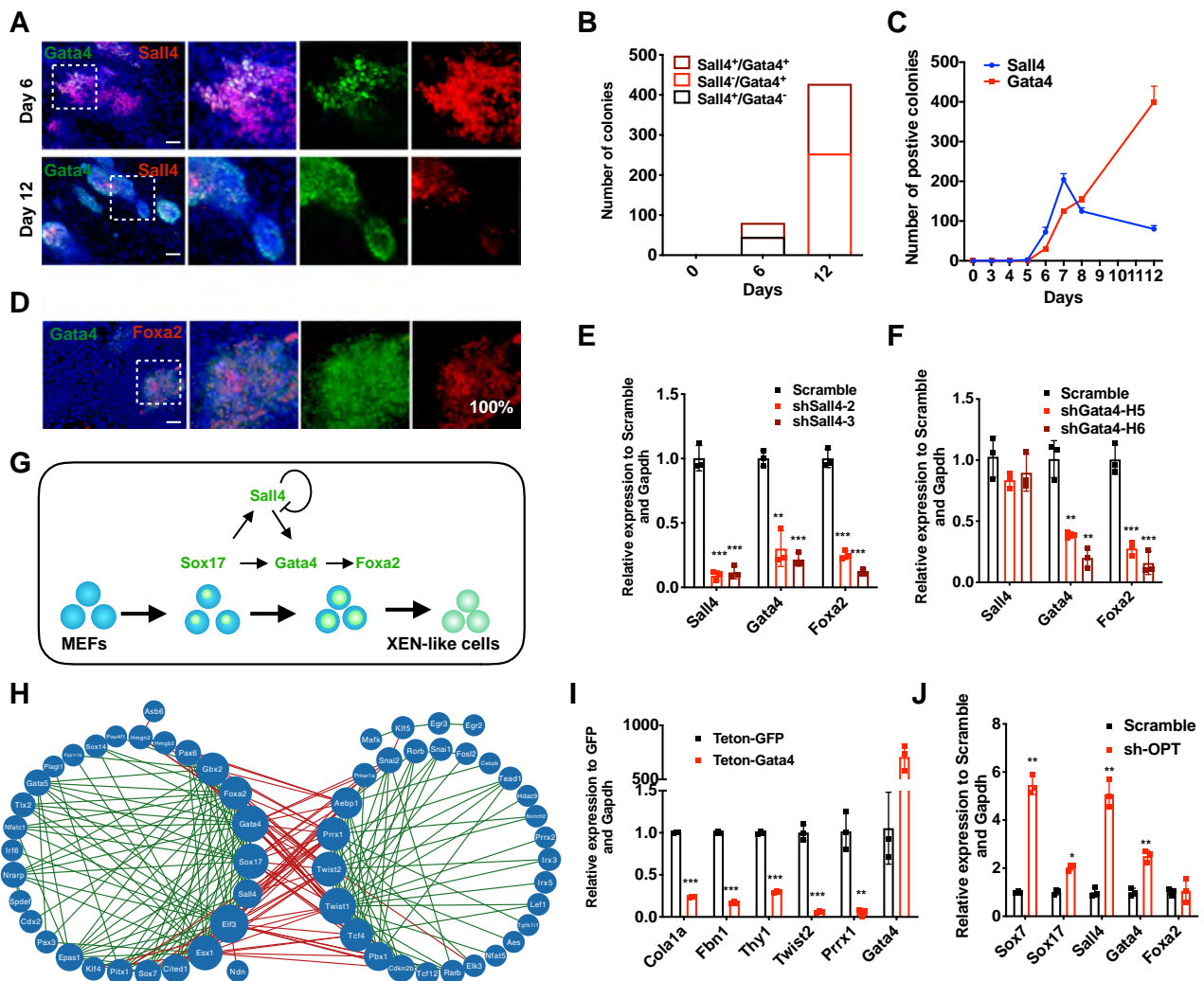


Fig. 2 Cell fate specification and transition into XEN-like cells with the accumulated expression of master TFs after Sox17 activation. **a** Co-staining of *Sall4* and *Gata4* induced by C6FAE on day 6 and day 12. Scale bar, 100 μm . **b** Quantitation of *Sall4*⁺/*Gata4*⁺, *Sall4*⁺/*Gata4*⁻, *Sall4*⁻/*Gata4*⁺ colonies per well of 12-well plate induced by C6FAE on day 0, 6, and 12. **c** Numbers of *Sall4* or *Gata4*-positive colonies per well of 12-well plate at different time points. **d** Co-staining of *Gata4* and *Foxa2* induced by C6FAE on day 12. Scale bar, 100 μm . The percentage of *Foxa2*-positive cells emerged in *Gata4*-positive cells were labeled in the lower right corner of the picture. **e** Relative mRNA levels of XEN genes induced by C6FAE on day 12 with the knockdown of *Sall4* ($n = 3$). **f** Relative mRNA levels of XEN genes induced by C6FAE on day 12 with the knockdown of *Gata4* ($n = 3$). **g** Schematic diagram of the hierarchical regulation circuitry among XEN master genes. **h** Transcription factor correlation network of XEN-like cells and MEFs. Green lines represent positive correlation and red lines represent the negative correlation. **i** Relative mRNA levels of MEF master genes with the overexpression of *Gata4* ($n = 3$). **j** Relative mRNA level of XEN genes with the knockdown of MEF genes. Sh-OPT stands for triple knockdown of MEF master genes, *Osr1*, *Prrx1*, and *Twist2* ($n = 3$). Significance was assessed compared with the controls using a one-tailed Student’s *t*-test. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

transcription of *Sall4* (Fig. 2f), which further supported that *Foxa2*, but not *Sall4*, is a downstream factor of *Gata4*. In summary, we found the activation of *Gata4* and *Sall4* was regulated differently, and the mutual regulation between them was dynamic. Figure 2g shows the hierarchical regulatory network of XEN master TFs.

The regulatory network for the cell fate specification and the transition was established after the core network of XEN master TFs was constructed (Fig. 2h). The cell fates of MEF and XEN-like were seen mutual antagonizing from the regulatory network. In the transition process, the up-regulation of XEN master TFs promoted the down-regulation of fibroblast master TFs, and vice versa (Fig. 2i, j and Supplementary Fig. 2). Such a positive feedback loop could account for the fast transition from fibroblasts to XEN-like cell fates in the final stage of reprogramming, which was exhibited in the single-cell analysis (Fig. 1f and Supplementary Fig. 1h).

Taken together, the core XEN transcriptional network, including *Sox17*, *Sall4*, *Gata4*, and *Foxa2*, was established consecutively and hierarchically, and thus completing the cell fate specification and transition process in the end. These findings supported the “prime, specify and transit” model that previously speculated (model 5 in Fig. 1a).

Chemicals are essential for *Sox17* expression while play different roles in the specification and transition process. Since CHIR99021, 616452 and Forskolin are pivotal to chemical reprogramming to XEN-like cells (Supplementary Fig. 3a), we further explored whether they were necessary to the entire process of reprogramming. We found that the chemical compounds worked with a stepwise approach.

Subtracting any one of CHIR99021, 616452 and Forskolin, respectively, from C6FAE from day 0 hampered the expression of XEN master TFs, especially the major TF, *Sox17* (Supplementary Fig. 3b). We then withdrew CHIR99021, 616452 and Forskolin after 4-day induction when *Sox17* was already activated albeit at a lower level (Fig. 3a). We found that the presence of CHIR99021 and Forskolin was essential for *Gata4* activation, while the addition of 616452 was critical for the up-regulation of *Sall4* (Fig. 3a). The expression of *Foxa2* was also greatly impaired when CHIR99021 or Forskolin was removed (Fig. 3a), which is consistent with our previous finding that *Foxa2* activation might be downstream of *Gata4* activation.

We further studied the requirement of CHIR99021, 616452, and Forskolin for the protein expression of *Sall4* and *Gata4*, by subtracting CHIR99021, 616452, and Forskolin after day 6, when *Sox17*-positive cell number was greatly increased. Similar to the transcriptional level, 616452 was essential for the expression of *Sall4* protein, and chemical cocktails containing 616452 after 6-day treatment of C6FAE were sufficient to induce the expression of *Sall4* protein (Fig. 3b, c). Moreover, we detected the expression of *Gata4* in *Sall4*-positive colonies when CHIR99021 or Forskolin was subtracted from the cocktail after day 6 in the presence of 616452 (Fig. 3b, c). It was consistent with our previous findings that *Sall4* activated *Gata4* expression. Interestingly, when 616452 was removed from the cocktail after day 6, in the presence of CHIR99021 and Forskolin, *Gata4* expression was still detected at a high level, and *Sall4* expression was substantially impaired. This indicates CHIR99021 and Forskolin were sufficient to induce *Gata4* expression after the activation of *Sox17*, which is independent of *Sall4* expression (Fig. 3d, e). This was also consistent with another wave of *Gata4* expression that was found after 6 days of C6FAE treatment (Fig. 2a–c).

In summary, it was the cooperation of CHIR99021, 616452 and Forskolin that activated *Sox17*; thereafter, CHIR99021/Forskolin

and 616452 activated the expression of *Gata4* and *Sall4*, respectively, in the specification stage, which further established the entire core regulatory network of XEN (Fig. 3f). After day 8, CHIR99021, 616452, and Forskolin were not essential for XEN gene expression (Supplementary Fig. 3c), suggesting that the transition phase was a self-organizing process by XEN master genes. These were also in line with the findings that the transduction of *Sall4* and *Gata4* was able to replace the function of CHIR99021, 616452 and Forskolin in inducing XEN-like colonies⁴. Overall, CHIR99021, 616452, and Forskolin played different roles in the reprogramming processes before and after *Sox17* expression although they were required for both of the two phases.

Endogenously activated BMP signaling is critical for *Sox17* activation and XEN induction. We further explored the upstream factors of *Sox17* after chemical treatment. Using bulk RNA-sequencing in the very early stage of XEN reprogramming, we found that *Bmp2* was one of the factors that were activated before *Sox17* expression (Fig. 4a and Supplementary Fig. 4a).

To investigate the roles of *Bmp2* in the activation of *Sox17* and the subsequent reprogramming into XEN-like cells, we inhibited Bmp signaling with small molecule inhibitors Dorsomorphin and DMH1. We found that both the transcription of *Sox17* and the number of *Sox17*-positive colonies remarkably decreased (Fig. 4b–d). Meanwhile, the overexpression of *Bmp2* promoted the activation of *Sox17* drastically (Supplementary Fig. 4b–e). Adding recombinant BMP2 or BMP4 in the reprogramming medium also improved the messenger RNA (mRNA) level of *Sox17* and the number of *Sox17*-positive colonies (Fig. 4e–g). Consistently, Dorsomorphin and DMH1 compromised the upregulation of *Sox17* expression by BMP2 or BMP4 (Supplementary Fig. 4f–i).

Importantly, the mRNA level of other XEN master genes (*Sall4*, *Gata4*, *Foxa2*) and the efficiency of XEN-like cell induction were hampered by Dorsomorphin and DMH1 (Fig. 4h, i), and were promoted by BMP2 and BMP4 (Fig. 4j, k). Also, we found that BMP4 notably promoted the up-regulation of *Sox17* in the iCD1 serum-free medium used in CiPSC induction⁹ (Supplementary Fig. 4j). However, the effects of BMP4 on the activation of *Sox17* relied on the presence of C6F. BMP4 could not replace the role of C6F on the activation of *Sox17* (Supplementary Fig. 4k).

These results indicate that the early activation of endogenous Bmp signaling by chemical cocktails promoted the expression of *Sox17* and thus facilitated the stepwise induction of XEN-like cells (Fig. 4l and Supplementary Fig. 4l).

The chemical boosters, CH55 and VPA, benefit *Sox17* activation and XEN specification differently. The two phases before and after *Sox17* expression revealed in our study, raised the possibility that the chemical boosters played different roles in the stepwise process from fibroblast to XEN-like cells. Thus, we compared the gene expression profiles induced with and without the previously reported chemical boosters, CH55 and valproic acid (VPA), in the presence of C6FAE^{3,5}.

Interestingly, CH55 promoted the expression of *Sox17* notably in the first 4 days, even on the basis of exogenously provided *Bmp4* (Supplementary Fig. 5a), while had nearly no function in further activation of other XEN genes from day 4 to 12 (Fig. 5a, c). VPA was found to promote the up-regulation of most XEN genes and XEN identity from day 4 to day 12 (Fig. 5b, c). However, VPA has no beneficial effect on *Sox17* expression in the first 4 days, suggesting that VPA improved XEN reprogramming efficiency by supporting the up-regulation of the XEN network after the activation of endogenous *Sox17*.

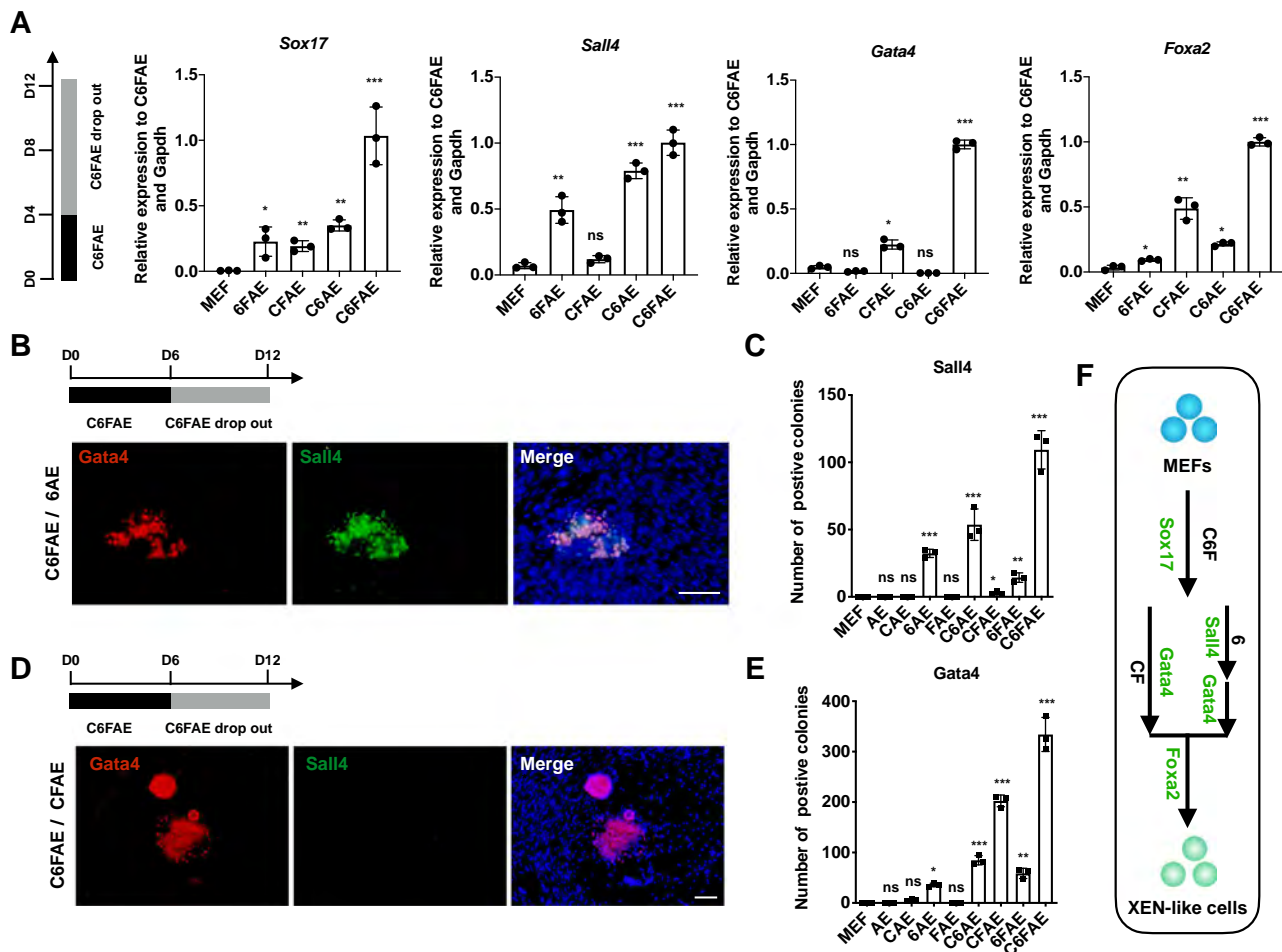


Fig. 3 CHIR99021, 616452, and Forskolin are essential for Sox17 expression while play different roles in the specification and transition process. **a** Relative mRNA levels of key XEN master genes on day 12 with removing C, 6, and F, respectively, from day 4 (n = 3). **b** Immunostaining of Sall4 and Gata4 on day 12 after removal of C, F from day 6. Scale bar, 100 μm. **c** Sall4-positive colony numbers on day 12 after removal of C, 6, and F, respectively, from day 6 (n = 3). **d** Immunostaining for Sall4 and Gata4 on day 12 after removal of 6 from day 6. Scale bar, 100 μm. **e** Gata4-positive colony numbers on day 12 after removal of C, 6, and F, respectively, from day 6 (n = 3). **f** Schematic of the roles of C, 6, and F in the regulation of XEN master genes. Significance was assessed compared with the controls using a one-tailed Student’s t-test. ***p < 0.001; **p < 0.01; *p < 0.05.

We also found that the cocktail mainly induced “smoothed” colonies co-expressing *Sox17*, *Gata4*, and *Sall4* in the presence of VPA (VC6FAE). Without VPA, it induced many “fuzzy” colonies with robust *Sox17* expression and very low expression of *Sall4* and *Gata4* (Supplementary Fig. 5b–d). The fuzzy colonies had higher mRNA levels of the fibroblast master genes, *Osr1*, *Prrx1* and *Twist2* (Supplementary Fig. 5e) and could rarely be induced into XEN-like cells (Supplementary Fig. 5f, g). Importantly, smoothed colonies, but not fuzzy colonies, could be induced into pOct4-GFP-positive CiPSCs (Supplementary Fig. 5h, i). VPA promoted the induction of pOct4-GFP-positive CiPSCs (Supplementary Fig. 5j). These results support that VPA improves the C6FAE-mediated XEN reprogramming by promoting the XEN specification process, which was previously reported to bridge chemical reprogramming from fibroblasts to pluripotency^{4,5}.

We further determined whether using chemical boosters, CH55 and VPA, in a stepwise manner could promote the XEN reprogramming efficiency. We found that treating the cells with CH55 only in the first 4 days was more efficient than treating for the entire process in promoting the expression of *Sox17* and *Sall4* (Fig. 5d). Also, VPA induced a higher level of *Gata4* and *Foxa2* mRNA when using in the last 8 days rather than in the entire process (Fig. 5e). Collectively, the sequential use of CH55 and

VPA in different steps reached the highest efficiency of XEN-like colony generation (Fig. 5f, g). Taken together, these findings not only suggested the “prime, specify and transit” model in chemical reprogramming but also revealed the roles of the chemicals on the stepwise processes (Fig. 1a, h).

Discussion

A major question in chemical reprogramming is “how does a set of chemicals, which bear no obvious relation to any genes or molecules that are directly associated with a specific cell type, enable the determination of a specific cell fate”. In this study, we made a significant conceptual leap towards an answer to this question. We demonstrated that the chemical reprogramming was a stepwise process by studying the molecular roadmap from fibroblasts to XEN-like cells. First, the chemicals orchestrated a priming state with the activated expression of *Sox17*, a master gene of XEN, without substantial cell fate alteration or determination. Afterward, the chemicals further guided hierarchical accumulation of endogenous master transcription factors for cell fate specification. Finally, cell fate was transitioned with the combination of those activated master transcription factors. In brief, chemicals used in reprogramming guided the hierarchical

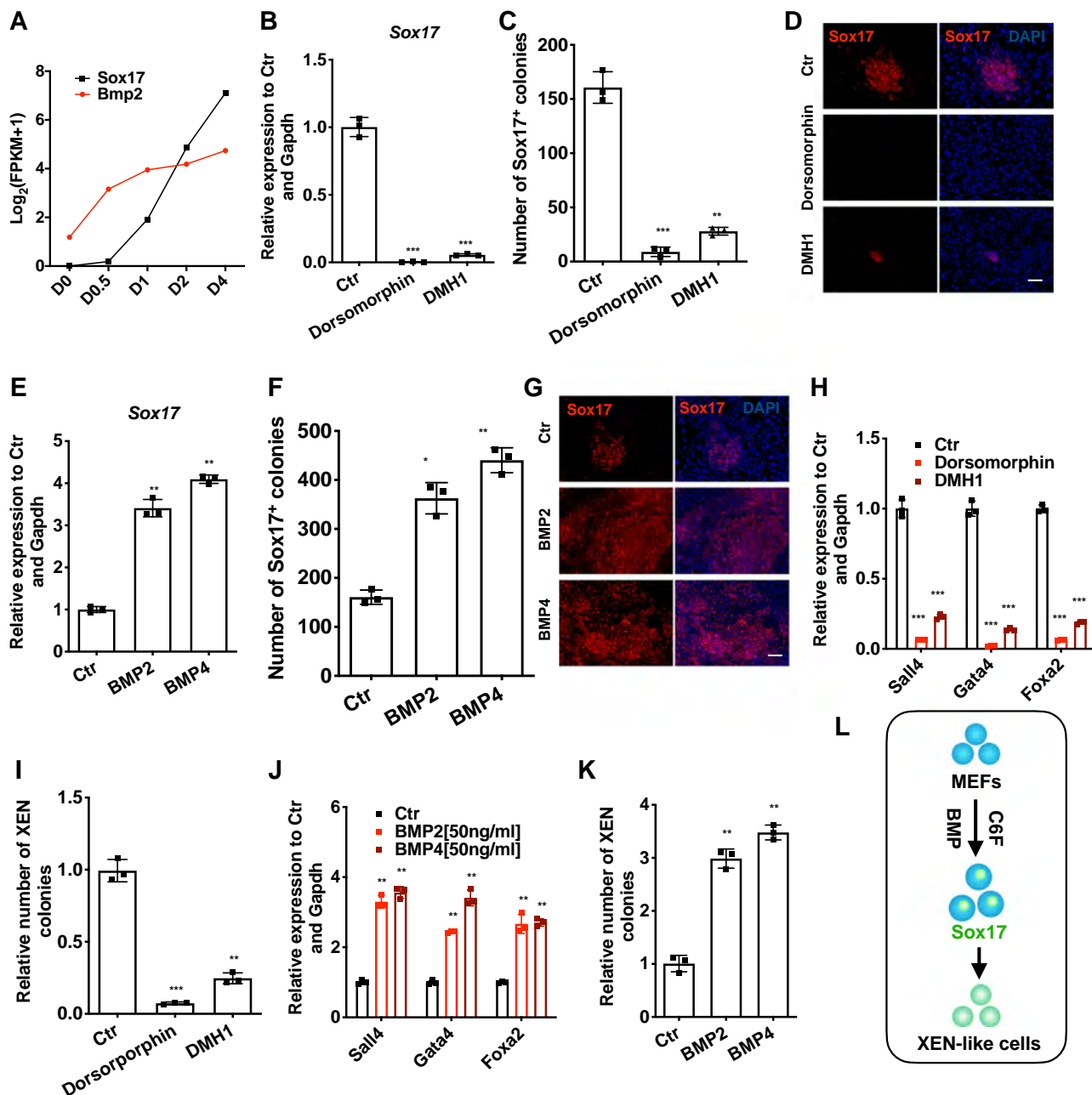


Fig. 4 Endogenously activated BMP signaling pathway is critical for the activation of Sox17 and XEN induction. **a** Expression of *Bmp2* and *Sox17* at different time points (analyzed by bulk RNA-sequencing). **b** The effects of Dorsomorphin and DMH1 on C6FAE-mediated *Sox17* activation (analyzed by RT-qPCR on day 4) ($n = 3$). **c** Number of *Sox17*-positive cells induced with Dorsomorphin and DMH1 on day 6 ($n = 3$). **d** *Sox17*-positive cells induced with Dorsomorphin and DMH1 on day 6. Scale bar, 100 μm . **e** Relative expression of *Sox17* after 4 days treatment with BMP2 and BMP4 at the dosage of 50 ng/ml ($n = 3$). **f** Number of *Sox17*-positive cells induced with BMP2 and BMP4 on day 6 ($n = 3$). **g** *Sox17*-positive cells induced with BMP2 and BMP4 on day 6. Scale bar, 100 μm . **h** Relative expression of *Sall4*, *Gata4*, *Foxa2* induced with Dorsomorphin and DMH1 on day 6 ($n = 3$). **i** Relative number of XEN colonies induced with Dorsomorphin and DMH1 on day 12 ($n = 3$). **j** Relative expression of *Sall4*, *Gata4*, *Foxa2* induced with BMP2 and BMP4 on day 6 ($n = 3$). **k** Relative number of XEN colonies induced with BMP2 and BMP4 on day 12 ($n = 3$). **l** Schematic of the roles of BMP signaling in the regulation of *Sox17* and XEN cell fate. Significance was assessed compared with the controls using a one-tailed Student's *t*-test. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

activation of master genes in the cell-fate-associated regulatory network in a stepwise manner.

Therefore, we indicate that the chemicals previously used in the entire process from fibroblasts to XEN-like cells had different functions in different phases and played different roles in activating different genes. The core chemicals CHIR99021, 616452, and Forskolin (C6F) were all essential to stimulate *Sox17* expression in the priming phase. Afterward, they supported the activation of other master genes, such as *Sall4* and *Gata4* for the

XEN cell fate specification in the *Sox17* expressing cells differently. CHIR99021 and Forskolin facilitated *Gata4* expression, while 616452 enabled the expression of *Sall4*. CH55 and BMP signaling functioned through elevating *Sox17* activation, while VPA functioned through activating the other XEN master genes in the *Sox17* expressing cells.

Importantly, the “prime-specify-transit” model may be extended to other chemical reprogramming systems according to the gene expression profiling data during the reprogramming

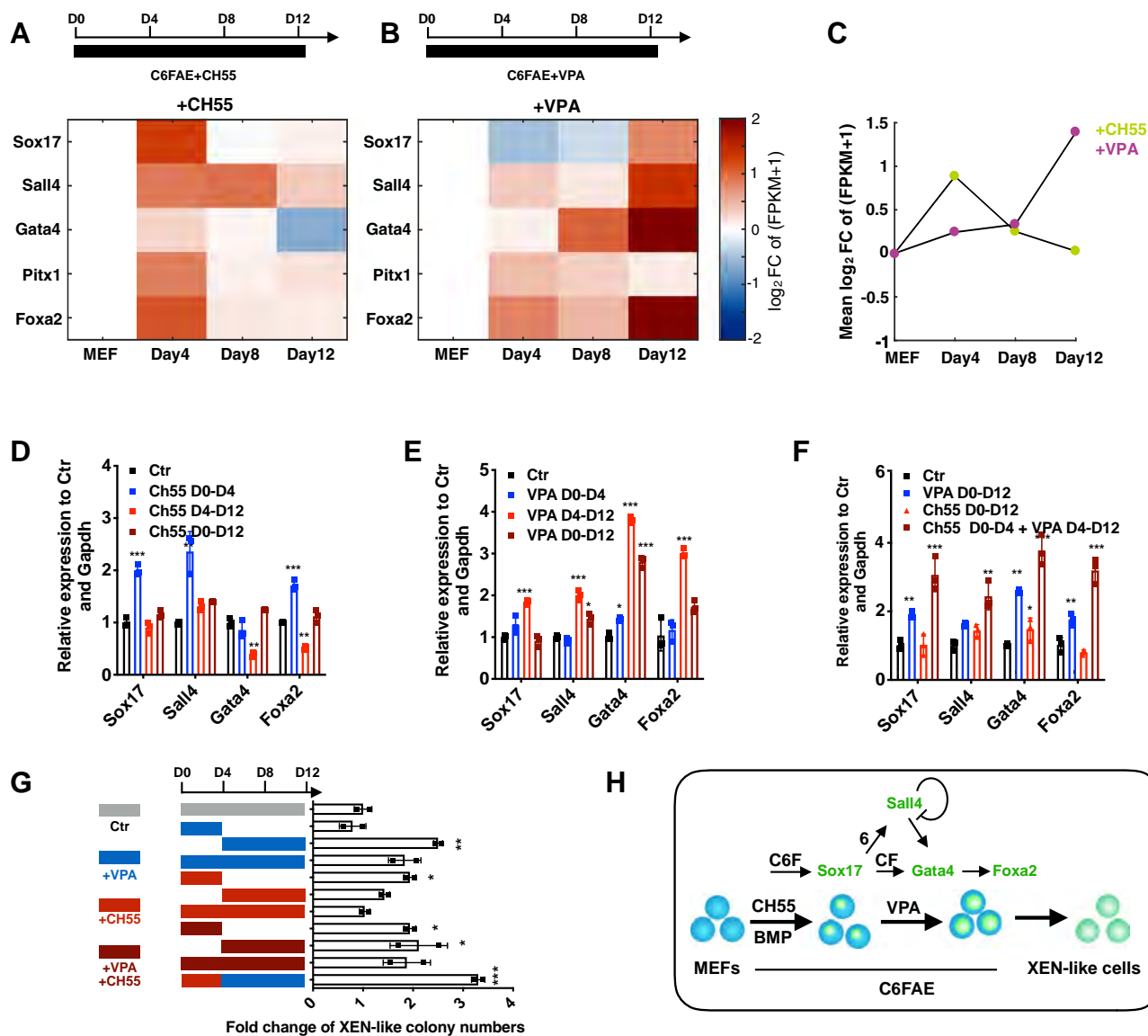


Fig. 5 Chemical reprogramming boosters benefit different steps of reprogramming. **a, b** Effects of CH55 (**a**) and VPA (**b**) on the expression of XEN genes at different time points. C6FAE with additional CH55 or VPA conditions were compared with only C6FAE condition. **c** Effects of CH55 and VPA on XEN identity at different time points. C6FAE with additional CH55 or VPA conditions were compared with only C6FAE condition. **d, e** Relative mRNA expression of XEN master genes in the presence of CH55 (**d**) or VPA (**e**) with different durations ($n=3$). **f** Relative fold change of the XEN genes expression in cells treated with CH55 and VPA by different duration ($n=3$). (analyzed on day 12 by RT-qPCR). **g** Relative numbers of XEN colonies in the presence of CH55 and VPA with different durations ($n=3$). **h** Summary of the functions of chemical boosters, CH55 and VPA, in XEN induction. Significance was assessed compared with the controls using a one-tailed Student's t -test. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

processes. For instance, in the chemical reprogramming process from fibroblast to neural stem cells, *Sox2* was activated very early by small molecules in the first 4 days and might initiate a priming state to neural stem cell³³. After that, neural stem cell core regulators network was built up, which was reminiscent of a specification process. Besides, the molecular dynamics during the chemical reprogramming from fibroblast to photoreceptor-like cells (CiPCs) was probably initiated by the activation of photoreceptor-specifying transcription factors, such as *Rorb*, *Ascl1* and *Pias3*, reminiscent of a priming phase²⁰. The stepwise manner in the activation of the master transcription factors suggested identifying the key molecular events in a chemical reprogramming process could help to optimize the protocol in a stepwise manner to achieve higher efficiency.

Unexpectedly, although cells in the priming phase had already expressed some of the XEN master TFs, they were still fibroblast-

like with a high level of the fibroblasts program, as well as the high expression of fibroblast master genes. This was rather different from an intermediate multipotent cell type that was presumed in previous reports^{34,35} and from the other possible intermediates that were speculated before this study (Fig. 1a). The “Disc model” for cell fate reprogramming matches these findings since cell fate priming helps the cell to escape the attractor of an initial cell type without cell fate determination, while the hierarchically accumulated expression of endogenous transcription factors provides the “guide rail” to determine a cell fate progressively, without entering into a multipotent attractor³⁶.

Our findings also highlight the importance of activating some or even one master transcription factor of the target cell type in developing a chemical reprogramming system. In the reprogramming process to XEN-like cells, the activation of *Sox17* was a molecular event that was not easy to be triggered. It required

most chemicals in the cocktail like C, 6, F and CH55, and even took advantage of the endogenously activated expression of BMP2 or other BMP signaling stimuli from serum or KSR. Thereafter, Sox17 expression made the subsequent molecular events possible and easier. Thus, the activation of one or more transcription factors in a cell type may represent the major molecular basis for cell plasticity. The cells initially express one or more transcription factors of another cell type may have superiority in cell fate transition in chemical reprogramming.

Moreover, since the chemicals C, 6, F, and their combinations have been widely used in different chemical reprogramming systems and generate many different cell fates^{8,12,13,15–17,19,20}, it is still unclear whether Sox17 activation is a specific outcome of the chemical treatment and whether these chemicals can prime the cells and facilitate cell fate conversion into other lineages simultaneously. These are some of the questions we intend to address in our future study.

Methods

MEF isolation. MEFs were isolated from E13.5 embryos of ICR mouse. After the removal of head, limbs, and viscera, embryos were minced with scissors and dissociated in trypsin-EDTA at 37 °C for 10 min. After centrifugation, MEF cells were collected and cultured in MEF medium, which included: high-glucose Dulbecco's modified Eagle's medium (DMEM) supplemented, 10% fetal bovine serum (FBS), 1% GlutaMAX, 1% nonessential amino-acids (NEAAs), and 1% penicillin-streptomycin. Oct4-EGFP mice were obtained from The Jackson Laboratory (004654). This study was performed under in accordance with protocols by Peking University laboratory animal research center.

Generation of XEN-like cells from fibroblasts. Twenty-thousand MEF cells were seeded into a well of 12-well plate. Twenty-four hours later, the medium was changed to XEN reprogramming medium, which included: KnockOut DMEM supplemented, 10% KnockOut Serum Replacement (KSR), 10% FBS, 1% GlutaMAX, 1% NEAAs, 0.055 mM 2-mercaptoethanol, 1% penicillin-streptomycin (Invitrogen), 50 ng/ml basic fibroblast growth factor (bFGF), and the small-molecule cocktail VC6FAE (0.5 mM valproic acid, 20 μM CHIR99021, 10 μM 616452, 50 μM Forskolin, 0.05 μM AM580, and 5 μM EPZ004777). XEN reprogramming medium was changed every 4 days for 12 to 20 days.

Immunofluorescence. Primary antibodies were those specific to rabbit anti-SALL4 (Abcam, 1:500), goat anti-SOX17 (R&D, 1:500), goat anti-GATA4 (Santa Cruz, 1:300), goat anti-GATA6 (R&D, 1:200), rabbit anti-Nanog (Sigma Aldrich, 1:200). The secondary antibodies used were FITC-conjugated secondary antibodies and TRITC-conjugated secondary antibodies (Jackson ImmunoResearch, 1:200).

Cells were fixed in 4% paraformaldehyde for 15 min at room temperature. Then, removing 4% paraformaldehyde and washing cells with PBS for two times, cells were permeabilized and blocked in PBS containing 0.2% Triton X-100 and 3% donkey serum for 1 h at room temperature. Then the cells were incubated with primary antibodies at 4 °C overnight. After washing three times with PBS, secondary antibodies (Jackson ImmunoResearch) were incubated at 37 °C for 1 h. The nuclei were stained with DAPI (Roche Life Science) for 5 min.

Quantitative reverse transcription PCR (RT-qPCR). RT-qPCR was performed according to protocols. Briefly, total RNA samples were extracted by using the EasyPure RNA Kit (TransGen Biotech) and were reverse transcribed into complementary DNA (cDNA) using TransScript One-step gDNA Removal and cDNA Synthesis SuperMix (TransGen Biotech); Real-time PCR was performed on a Quantagene q225 System (KUBO technology) using 2 × T5 Fast qPCR Mix (TSINGKE Biological Technology).

Single-cell RNA-seq. Individual cell at different time points was manually picked after digestion, lysed and subjected to cDNA synthesis^{37,38}. Single-cell cDNA was then amplified and fragmented as published steps^{37,38}. The sequencing library was constructed (New England Biolabs) and sequenced with paired-end 150-bp reads on an Illumina HiSeq X-Ten platform (Novogene). Raw reads were firstly separated by cell barcodes, then trimmed, and aligned to the mm9 mouse transcriptome and de-duplicated by UMIs information as described previously³⁹.

Pseudotime analysis. Monocle (v2.6.4) were adopted to perform the pseudotime analysis. Differentially expressed genes (DEGs) identified from each cell type were used as ordering genes. The whole workflow followed the recommended pipeline with default parameters.

Integration analysis of gene expression between data in this study and in Zhao et al.⁶ Cells from day 0 to day 20 in this study, and cells belong to MEFs, Stage I day 5, Stage I day 12 and XEN-like cells in Zhao et al.⁶ were used to perform the integration analysis. To integrate different data sets, CCA algorithm from Seurat was used.

Statistics and reproducibility. All experiments contain at least three independent biological replicates. No randomization or blinding was used. The statistical analysis in this paper uses Student's *t*-test. *p*-value of <0.05 is considered a significant difference.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The accession number for the RNA-seq and single-cell RNA-seq data reported in this paper is NCBI GEO: GSE144097. Source data underlying plots shown in figures are provided in Supplementary Data 1. Full blots are shown in Supplementary Information. All other data, if any, are available upon reasonable request.

Received: 31 December 2019; Accepted: 11 September 2020;

Published online: 30 October 2020

References

- Gurdon, J. B. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *J. Embryol. Exp. Morphol.* **10**, 622–640 (1962).
- Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
- Hou, P. et al. Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* **341**, 651–654 (2013).
- Zhao, Y. et al. A XEN-like State Bridges Somatic Cells to Pluripotency during Chemical Reprogramming. *Cell* **163**, 1678–1691 (2015).
- Ye, J. et al. Pluripotent stem cells induced from mouse neural stem cells and small intestinal epithelial cells by small molecule compounds. *Cell Res.* **26**, 34–45 (2016).
- Zhao, T. et al. Single-cell RNA-seq reveals dynamic early embryonic-like programs during chemical reprogramming. *Cell Stem Cell* **23**, 31–45 e37 (2018).
- Long, Y., Wang, M., Gu, H. & Xie, X. Bromodeoxyuridine promotes full-chemical induction of mouse pluripotent stem cells. *Cell Res.* **25**, 1171–1174 (2015).
- Li, X. et al. Small-molecule-driven direct reprogramming of mouse fibroblasts into functional neurons. *Cell Stem Cell* **17**, 195–203 (2015).
- Cao, S. et al. Chromatin accessibility dynamics during chemical induction of pluripotency. *Cell Stem Cell* **22**, 529–542 e525 (2018).
- Cheng, L. et al. Generation of neural progenitor cells by chemical cocktails and hypoxia. *Cell Res.* **24**, 665–679 (2014).
- Li, X. et al. Direct reprogramming of fibroblasts via a chemically induced XEN-like state. *Cell Stem Cell* **21**, 264–273 e267 (2017).
- Hu, W. et al. Direct conversion of normal and Alzheimer's disease human fibroblasts into neuronal cells by small molecules. *Cell Stem Cell* **17**, 204–212 (2015).
- Fu, Y. et al. Direct reprogramming of mouse fibroblasts into cardiomyocytes with chemical cocktails. *Cell Res.* **25**, 1013–1024 (2015).
- Cao, N. et al. Conversion of human fibroblasts into functional cardiomyocytes by small molecules. *Science* **352**, 1216–1220 (2016).
- Bansal, V. et al. Chemical induced conversion of mouse fibroblasts and human adipose-derived stem cells into skeletal muscle-like cells. *Biomaterials* **193**, 30–46 (2019).
- Nie, B. et al. Brown adipogenic reprogramming induced by a small molecule. *Cell Rep.* **18**, 624–635 (2017).
- Takeda, Y., Harada, Y., Yoshikawa, T. & Dai, P. Direct conversion of human fibroblasts to brown adipocytes by small chemical compounds. *Sci. Rep.* **7**, 4304 (2017).
- Tian, E. et al. Small-molecule-based lineage reprogramming creates functional astrocytes. *Cell Rep.* **16**, 781–792 (2016).
- Cao, S. et al. Chemical reprogramming of mouse embryonic and adult fibroblast into endoderm lineage. *J. Biol. Chem.* **292**, 19122–19132 (2017).
- Mahato, B. et al. Pharmacologic fibroblast reprogramming into photoreceptors restores vision. *Nature* **581**, 83–88 (2020).
- Wang, Y. et al. Conversion of human gastric epithelial cells to multipotent endodermal progenitors using defined small molecules. *Cell Stem Cell* **19**, 449–461 (2016).

22. Yin, J. C. et al. Chemical conversion of human fetal astrocytes into neurons through modulation of multiple signaling pathways. *Stem Cell Rep.* **12**, 488–501 (2019).
23. Zhao, Y. Chemically induced cell fate reprogramming and the acquisition of plasticity in somatic cells. *Curr. Opin. Chem. Biol.* **51**, 146–153 (2019).
24. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
25. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
26. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
27. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
28. Lim, C. Y. et al. Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages. *Cell Stem Cell* **3**, 543–554 (2008).
29. Niakan, K. K. et al. Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. *Genes Dev.* **24**, 312–326 (2016).
30. Shimosato, D. et al. Extra-embryonic endoderm cells derived from ES cells induced by GATA factors acquire the character of XEN cells. *BMC Dev. Biol.* **7**, 80 (2007).
31. Hwang, J. T. et al. GATA6 and FOXA2 regulate Wnt6 expression during extraembryonic endoderm formation. *Stem Cells Dev.* **21**, 3220–3232 (2012).
32. Yang, J., Gao, C., Chai, L. & Ma, Y. A novel SALL4/OCT4 transcriptional feedback network for pluripotency of embryonic stem cells. *PLoS ONE* **5**, e10766 (2010).
33. Zhang, M. et al. Pharmacological reprogramming of fibroblasts into neural stem cells by signaling-directed transcriptional activation. *Cell Stem Cell* **18**, 653–667 (2016).
34. Han, X. et al. A molecular roadmap for induced multi-lineage trans-differentiation of fibroblasts by chemical combinations. *Cell Res.* **27**, 386–401 (2017).
35. Xie, X., Fu, Y. & Liu, J. Chemical reprogramming and transdifferentiation. *Curr. Opin. Genet. Dev.* **46**, 104–113 (2017).
36. Maamar, H. et al. Noise in gene expression determines cell fate in *Bacillus subtilis*. *Science* **317**, 526–529 (2007).
37. Li, L. et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873 e854 (2017).
38. Gu, C., Liu, S., Wu, Q., Zhang, L. & Guo, F. Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes. *Cell Res.* **29**, 110–123 (2019).
39. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

Acknowledgements

We thank Chunyan Yang, Yang Liu, and Siyuan Zhang for technical assistance, and Jiayu Chen for providing pOct4-GFP mice. We thank Iain C. Bruce for editing the manuscript. This work was supported by the National Key Research and Development Program of China (2018YFA0800504), the National Natural Science Foundation of China (31771645, 31922020, 31821091 and 31771590), the Science and Technology Department of Sichuan Province (2018JZ0025).

Author contributions

Z.Y., J.L., Q.W., C.Y., L.M., J.Y., and K.B. performed experiments. X.X., C.G., and A.N. conducted the bioinformatics analyses. Y.Z., C.T., and F.G. supervised this project. Y.Z., Z.Y., and X.X. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s42003-020-01346-w>.

Correspondence and requests for materials should be addressed to F.G., C.T. or Y.Z.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020



Chemical Pretreatment Activated a Plastic State Amenable to Direct Lineage Reprogramming

Zhenghao Yang^{1†}, Xiaochan Xu^{2†}, Chan Gu³, Alexander Valentin Nielsen⁴, Guokai Chen⁵, Fan Guo^{3*}, Chao Tang^{2,6*} and Yang Zhao^{1,2*}

¹State Key Laboratory of Natural and Biomimetic Drugs, MOE Key Laboratory of Cell Proliferation and Differentiation, Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine, College of Future Technology, Peking University, Beijing, China, ²Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China, ³State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, ⁴The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark, ⁵Centre of Reproduction, Development and Aging, Faculty of Health Sciences, University of Macau, Macau, China, ⁶Center for Quantitative Biology, Peking University, Beijing, China

OPEN ACCESS

Edited by:

Tongbiao Zhao,
Institute of Zoology (CAS), China

Reviewed by:

Zhili Rong,
Southern Medical University, China
Wenxiang Hu,
University of Pennsylvania,
United States
Shijun Hu,
Soochow University, China

*Correspondence:

Yang Zhao
yangzhao@pku.edu.cn
Chao Tang
tangc@pku.edu.cn
Fan Guo
guofan@ioz.ac.cn

[†]These authors share first authorship

Specialty section:

This article was submitted to
Stem Cell Research,
a section of the journal
Frontiers in Cell and Developmental
Biology

Received: 29 January 2022

Accepted: 21 February 2022

Published: 25 March 2022

Citation:

Yang Z, Xu X, Gu C, Nielsen AV,
Chen G, Guo F, Tang C and Zhao Y
(2022) Chemical Pretreatment
Activated a Plastic State Amenable to
Direct Lineage Reprogramming.
Front. Cell Dev. Biol. 10:865038.
doi: 10.3389/fcell.2022.865038

Somatic cells can be chemically reprogrammed into a pluripotent stem cell (CiPSC) state, mediated by an extraembryonic endoderm- (XEN-) like state. We found that the chemical cocktail applied in CiPSC generation initially activated a plastic state in mouse fibroblasts before transitioning into XEN-like cells. The plastic state was characterized by broadly activated expression of development-associated transcription factors (TFs), such as *Sox17*, *Ascl1*, *Tbx3*, and *Nkx6-1*, with a more accessible chromatin state indicating an enhanced capability of cell fate conversion. Intriguingly, introducing such a plastic state remarkably improved the efficiency of chemical reprogramming from fibroblasts to functional neuron-like cells with electrophysiological activity or beating skeletal muscles. Furthermore, the generation of chemically induced neuron-like cells or skeletal muscles from mouse fibroblasts was independent of the intermediate XEN-like state or the pluripotency state. In summary, our findings revealed a plastic chemically activated multi-lineage priming (CaMP) state at the onset of chemical reprogramming. This state enhanced the cells' potential to adapt to other cell fates. It provides a general approach to empowering chemical reprogramming methods to obtain functional cell types bypassing inducing pluripotent stem cells.

Keywords: chemical reprogramming, cell plasticity, chromatin accessibility, cell fate transition, direct reprogramming

INTRODUCTION

Somatic cells can be chemically reprogrammed into functional cell types indirectly by first becoming pluripotent through a XEN-like state (Hou et al., 2013; Zhao et al., 2015) or directly without an intermediate pluripotent state. The application superiority of the chemical reprogramming strategy over the transgenic approach in inducing cell fate reprogramming is well established (Zhao, 2019). For example, small molecules are genetically non-integrative, easy to be manipulated, cell-culture standardized, and cost-effective. Chemical cocktails could also help increase efficiency in generating a defined cell type (Zhao et al., 2015). To date, chemical reprogramming has been a promising strategy for obtaining functional cell types in regenerative medicine. Fibroblasts were reported to be reprogrammed into many cell types, including neural progenitors (Cheng et al., 2014), neuron cells

(Hu et al., 2015; Li et al., 2015; Mahato et al., 2020; Yin et al., 2019), cardiomyocytes (Fu et al., 2015; Cao et al., 2016), skeletal muscles (Bansal et al., 2019), brown adipocytes (Nie et al., 2017; Takeda et al., 2017), astrocytes (Tian et al., 2016), and endoderm progenitor-like cells (Cao et al., 2017; Wang et al., 2016).

However, the roles of chemicals in reprogramming systems are still elusive, which hampered the development of chemical cocktails for an assigned cell type. In the transgenic approach, the reprogramming factors are always the target cell type enriched TFs. Those TFs are associated with the target cell type's development or differentiation. They have been intensively studied in somatic reprogramming into induced pluripotency stem cells (iPSCs) and direct lineage reprogramming (Takahashi and Yamanaka, 2006; Xu et al., 2015). They directly serve as the pioneer factors to shape specific cell type favored epigenetic states and activate the expression of other master TFs for cell fate reprogramming (Iwafuchi-Doi and Zaret, 2014). Unlike these reprogramming genes, the mechanisms of chemicals in reprogramming and determining a cell fate are far from known.

Notably, the small molecules essential for CiPSC induction, CHIR99021 (a GSK3 inhibitor), 616452 (Repsox, an ALK5 inhibitor), and Forskolin (a cAMP agonist) and their combinations have been frequently used for the direct induction of different cell types (Zhao, 2019). It suggests that some common mechanisms are underlying these chemical-induced cell-type reprogramming processes. Understanding the mechanisms is beneficial to developing additional chemical reprogramming systems based on the same rationale.

Herein, we found that mouse fibroblasts were initially induced into a plastic chemically activated multi-lineage priming (CaMP) phase in chemical reprogramming before further specification into specific lineages. The CaMP phase was characterized by heterogeneous expression of multiple developmental genes and a global gain of chromatin accessibility. It was induced concomitantly by core small molecules, CHIR99021, 616452, and Forskolin. Introducing the CaMP phase with a chemical pretreatment, we improved the chemical reprogramming systems from fibroblasts directly into functional neuron-like cells and beating myocytes.

MATERIALS AND METHODS

Mice

All procedures involving mice were approved by the Institutional Animal Care and Use Committee (IACUC) and Use Committee at the Peking University, Beijing. For lineage-tracing experiments, 12–16 weeks Col1a2-CreERT2 and Rosa26tdTom mice were used. For *in vivo* labeling, all pregnant female mice have received intraperitoneal injections of tamoxifen (20 mg/ml, Sigma, United Kingdom) at a dose of 4 mg/30 g body weight before the isolation of MEF.

MEF Isolation

MEF medium: high glucose Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1% GlutaMAX, 1% nonessential amino acids (NEAAs), 0.055 mM 2-

mercaptoethanol, and 1% penicillin-streptomycin. Mouse embryonic fibroblasts (MEFs) were isolated from ICR mouse embryos. Briefly, after removal of the head, limbs, and viscera, E13.5 embryos were minced with scissors and dissociated in trypsin-EDTA at 37°C for 10 min. After adding MEF medium and centrifugation, MEF cells were collected and cultured.

Generation of XEN-Like Cells From Fibroblasts

XEN reprogramming medium: KnockOut DMEM supplemented with 10% KnockOut Serum Replacement (KSR), 10% FBS, 1% GlutaMAX, 1% NEAAs, 0.055 mM 2-mercaptoethanol, 1% penicillin-streptomycin (Invitrogen), 50 ng/ml basic fibroblast growth factor (bFGF), and the small-molecule cocktail VC6FAE (0.5 mM valproic acid, 20 μM CHIR99021, 10 μM 616,452, 50 μM Forskolin, 0.05 μM AM580, and 5 μM EPZ004777). MEFs were seeded at 20,000 cells per well of a 12-well plate with an MEF culture medium. For XEN induction, the medium was changed to XEN reprogramming medium the next day, and it was changed every 4 days for 12–20 days.

Induction of Skeletal Muscle Cells From CaMP State

Skeletal muscle reprogramming medium: DMEM/M199 medium (4:1) supplemented with 10% KSR, 10% FBS, 1% GlutaMAX, 1% NEAAs, 1% penicillin-streptomycin (Invitrogen), and the small-molecule cocktail C6FS (20 μM CHIR99021, 10 μM 616,452, 50 μM Forskolin, and 3 μM SB431542). MEFs were seeded at 20,000 cells per well of a 12-well plate with an MEF culture medium. For skeletal muscle cell induction, the medium was changed to XEN reprogramming medium the next day (day 0), and the medium was switched to skeletal muscle reprogramming medium at day 4 to induce myocytes for 8–12 days. The skeletal muscle reprogramming medium was changed every 4 days.

Induction of Neuron-Like Cells From CaMP State

Neuron-like cells reprogramming medium: neurobasal plus medium supplemented with 2% B27-plus supplement, 1% GlutaMAX, 1% penicillin-streptomycin (Invitrogen), and the small-molecule cocktail CFI (3 μM CHIR99021, 10 μM Forskolin, and 10 μM ISX-9).

MEFs were seeded at 20,000 cells per well of a 12-well plate with an MEF culture medium. For neuron-like cells induction, the medium was changed to XEN reprogramming medium the next day (day 0), and the medium was switched to neuron-like cells reprogramming medium at day 4 to induce neuron-like cells for 8–12 days. For neuron-like cells maturation, cells were plated on astrocytes at day 12 or day 16, and further culture for 16 days was supplemented with BDNF (20 ng/ml) and GDNF (20 ng/ml). Neuron-like cells reprogramming medium was changed every 4 days.

Isolation of Astrocytes

Astrocyte medium: DMEM/F12 supplemented with 10% FBS, 1% GlutaMAX, 1% nonessential amino-acids (NEAAs), and 1% penicillin-streptomycin.

After the newborn mice were anesthetized on ice and sacrificed, they were disinfected with 75% alcohol for 5 min. Brain tissue was taken and cut into pieces of 4 mm³ with scissors. Pieces of brain tissue were collected and digested with 2 ml 0.25% trypsin and 0.1 ml DNase I (2 mg/ml) for 20 min at 37°C. The digestion was stopped with 2 ml astrocyte medium and centrifuged for 5 min at 1,500 rpm. About 600,000 cells were resuspended with 10 ml astrocyte medium and plated into 10 cm dish. The supernatant was taken into a new T75 flask after 30 min and cultured for 7–10 days. After that, cultured cells were shaken on a shaker for 16 h at 250 rpm. Adherent astrocytes were digested and plated for neuron-like cells maturation.

Isolation of Neuron Cells

Neuron culture medium: neurobasal plus medium supplemented with 2% B27-plus supplement, 1% GlutaMAX, 1% penicillin-streptomycin. The plating medium was prepared with a neuron culture medium supplement with 10% FBS.

After the newborn mice were anesthetized on ice and sacrificed, they were disinfected with 75% alcohol for 5 min. Brain tissue was taken and cut into pieces of 9 mm³ with scissors. Pieces of brain tissue were collected and digested with 3 ml 0.1% trypsin and 0.1 ml DNase I (2 mg/ml) for 9 min at 37°C. We discarded the supernatant and add 4 ml plating medium. After that, we discarded the supernatant, added 1.5 ml plating medium and 0.1 ml DNase I (2 mg/ml) and pipette 20 times, and collected the supernatant. We repeated this step one more time and centrifuged the collected supernatant for 5 min at 1,000 rpm. Cells were seeded at 25,000 cells per well of a poly-L-lysine pre-coated 12-well plate with a plating medium. We then gently shook the 12-well plate and switched the medium to neuron culture medium 6 h later, and cells were cultured at 37°C.

Isolation and Culture of Myocytes

After the newborn mice were anesthetized on ice and sacrificed, they were disinfected with 75% alcohol for 5 min. Limb tissue was taken and cut into pieces of 0.1 mm³ with scissors. Pieces of limb tissue were collected and digested with 6 ml 0.25% trypsin and 0.2 ml DNase I (2 mg/ml) for 20 min at 37°C, pipette tissue every 5 min. Cells were centrifuged for 10 min at 1,000 rpm and 500,000 cells were plated into 10 cm dish for 2 h. Cells in the supernatant were transferred into a new 10 cm dish for further culture. After 3 days, the medium was switched into skeletal muscle differentiation medium (DMEM medium supplement 2% horse serum) for further culture.

Immunofluorescence

Cells were washed with PBS and fixed in 4% paraformaldehyde for 15 min at room temperature. After washing twice with PBS, cells were permeabilized and blocked in PBS containing 0.2% Triton X-100 and 3% donkey serum for 1 h at room temperature. Then, the cells were incubated with primary antibodies at 4°C

overnight. After washing three times with PBS, secondary antibodies (Jackson ImmunoResearch) were incubated at 37°C for 1 h. The nuclei were stained with DAPI (Roche Life Science) for 5 min. Primary antibodies were those specific to rabbit anti-SALL4 (Abcam, 1:500), mouse anti-Tubb3 (Biolegend, 1:300), rabbit anti-Synapsin 1 (Abcam, 1:500), rabbit anti-Map2 (Millipore, 1:200), rabbit anti-Neun (Millipore 1:500), rabbit anti-GABA (Sigma, 1:300), anti-mouse neurofilament 200 (Millipore 1:300), rabbit anti-vGlut1 (Synaptic system, 1:300), mouse anti-myosin heavy chain (R&D, 1:300), mouse anti-MyoD (Thermo fisher, 1:200), mouse anti-myogenin (Thermo fisher, 1:200), and mouse anti- α -actinin (Sigma, 1:500). The secondary antibodies used were FITC-conjugated secondary antibodies and TRITC-conjugated secondary antibodies (Jackson ImmunoResearch, 1:200).

RT-qPCR

Total RNA was extracted using the EasyPure RNA Kit (TransGen Biotech) and was reverse-transcribed into cDNA using TransScript One-step gDNA Removal and cDNA Synthesis SuperMix (TransGen Biotech). Real-time PCR was performed on a Quantagene q225 System (KUBO technology) using 2 \times T5 Fast qPCR Mix (TSINGKE Biological Technology).

ATAC-Seq

ATAC-seq libraries were prepared using Trueprep DNA library Prep Kit V2 for Illumina (vazyme). Totally, 50,000 cells were used for every single reaction. Cells were washed in 100 μ l cold PBS and resuspended in 50 μ l lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl₂, 0.5% NP-40) for 10 min, and nuclei were spun at 500 g for 10 min using a refrigerated centrifuge. Then, the pellet was resuspended in 50 μ l transposase reaction mix and incubated at 37°C for 30 min. The samples were purified and purified, and then libraries were amplified by PCR for 13 cycles. The libraries were sequenced using an Illumina HiSeq 2,500 machine.

RNA-Seq

We treated MEF cells with C6FAE, 6FAE, CFAE, or C6AE from day 0 and extracted total RNA on day 4, day 8, and day 12. Total RNA was extracted using the EasyPure RNA Kit (TransGen Biotech). Library construction was completed by Novogene company. The libraries were sequenced using an Illumina HiSeq 2,500 machine.

Electrophysiology

Whole-cell patch-clamp recordings were performed by Scope Research Institute of Electrophysiology. All currents were recorded using a MultiClamp 700 A amplifier. For whole-cell patch-clamp recordings, the ACSF (artificial cerebrospinal fluid) extracellular solution contained 128 mM NaCl, 30 mM glucose, 25 mM HEPES, 5 mM KCl, 2 mM Ca²⁺, and 1 mM MgCl₂. The pH of the bath solution was adjusted to 7.3 with NaOH, and osmolality was 300–305 mOsm/L. The pipette solution consisted of 135 mM KCl, 5 mM Na-phosphocreatine, 10 mM HEPES, 2 mM EGTA, 4 mM Mg-ATP, and 0.5 mM Na₂-GTP. The pH of pipette solution was adjusted to 7.3 with KOH and osmolality

to 280–290 mOsm/L. Whole-cell patch-clamp recordings were carried out using a HEKA EPC10 amplifier with PatchMaster software (HEKA; Instrument Inc., Lambrecht/Pfalz, Germany). To record the sodium and potassium currents, cells were held at -80 mV and depolarized from -80 to $+80$ mV in 10 mV increments for 1 s. The sample and sweep intervals were 20 μ s and 2 s, respectively. To record spontaneous excitatory postsynaptic currents (EPSCs), induced neuron-like cells were held at -70 and 0 mV, respectively.

Data Processing and Analysis

The data used in the RNA-Seq analysis was combined with our previous data (Yang et al., 2020), and we additionally performed RNA-Seq on the C6F5UE treatment group. Gene expression levels were normalized as \log_2 (FPKM+1) in all bulk RNA-Seq data. For temporal bulk RNA-Seq data clustering in each cocktail treatment, genes that have detected FPKM >1 at least at one sample remained to perform *K*-means clustering and were grouped into 20 clusters. For the cell fate induction experiment, genes that vary >1 among all samples remained to perform heterarchical clustering. PCA was done with all normalized gene expression levels with scaling the normalized expression for each gene by *z*-score among samples. Day 16 and XEN data were adapted from GEO Datasets (GEO IDs: GSE73631). For dropout experiments, gene expression under C6FAE conditions was first compared in MEFs on day 4 and day 8. TFs that increased at least by 1.5 were identified as upregulated and retained for the follow-up analysis.

Single-cell data was adapted from the previously published dataset (GEO IDs: GSE144097). We used Seurat (Stuart et al., 2019) package to do t-distributed stochastic neighbor embedding (t-SNE) projection and visualization. The three germ layer transcriptional factors correlation was calculated with the spearman correlation coefficients between gene pairs with \log_2 (UMI count +1) and visualized with heatmap.2 function from ggplot2 package.

ATACseq analysis includes peak calling with MACS (version 2.1.2) (Zhang et al., 2008), differential peak detection with RPKM, and visualization with EnrichedHeatmap (Gu et al., 2018). Two biological replications of Control and CaMP samples were processed and CaMP enriched peak regions and Control enriched peak regions were labeled. Coverage tracks of the samples were generated with the alignment of reads (BAM file) with the bamCoverage function from deepTools (Ramirez et al., 2014). The number of reads per bin was calculated and normalized by reads per kilobase per million mapped reads (RPKM).

These analyses were done with MATLAB and R script.

RESULTS

Chemical Reprogramming Cocktails Initially Activated the Expression of a Broad Spectrum of Development-Associated Transcription Factors

To investigate how chemical compounds alter cell fate, we previously studied the chemical reprogramming process from

mouse embryonic fibroblasts (MEFs) to XEN-like cells. The study revealed a hierarchical activation of XEN cell master TFs primed by Sox17 and the different roles of essential chemicals during the process (Yang et al., 2020). In parallel to that study, we measured the global gene expression profiles by RNA sequencing during reprogramming. We treated the initial MEFs with chemical cocktails composed of CHIR99021 (C), 616452 (6), Forskolin (F), AM580 (A), and EPZ00477 (E) and collected the samples at days 0, 4, 8, and 12 (Figure 1A).

The gene expression induced by C6FAE showed various dynamics (Figure 1B). The genes were categorized into four major groups stepwise (Figures 1B,C): downregulated fibroblast genes in the first 4 days, upregulated genes from day 4 to day 12, upregulated XEN genes, and decreased master genes of fibroblast in the last period. Unexpectedly, those upregulated genes during fibroblast reprogramming to XEN-like cells included a broad spectrum of lineage-associated TFs, such as *Ascl1*, *Zic1*, *Hand2*, *Hey2*, *Nkx6-1*, and *Gata2*, which, respectively, regulate the development of ectoderm, mesoderm, and endoderm (Figures 1B,C). The top-ranked Gene Ontology terms of these lineage-associated genes included “regulation of developmental process,” “multicellular organismal development,” and “system development” (Figure 1D).

We also detected the activation of lineage-associated TFs in a chemical reprogramming system with additional chemical boosters for XEN generation, VPA, UNC0638 (Hou et al., 2013), and CH55 (Zhao et al., 2018) (Supplementary Figure S1). These different chemical cocktails activated lineage-associated TFs before XEN cell fate transition (Figure 1E). The activation timing for 56%–71% of these TFs can be as early as day 4 (Figure 1F). Interestingly, these upregulated lineage-associated TFs were highly overlapping in different chemical cocktails (Figure 1G). Thus, we refer to the induction phase with the activation of multi-lineage TFs, including Sox17 for XEN-like cell induction (Yang et al., 2020), as chemically activated multi-lineage priming (CaMP).

The XEN master genes *Sall4* and *Gata4* were significantly activated in the latest stage after the CaMP state (Figure 1C). It indicates that XEN cell fate was induced in a “plasticization and specification” manner rather than determined in the initial stage of chemical reprogramming (Figure 1H).

Heterogeneous Expression of Endogenous Development-Associated Transcription Factors in Single Cells

The upregulated developmental genes could be activated in 1 cell simultaneously or in different cells heterogeneously. To clarify these two scenarios, gene expression in individual cells needs to be investigated. Thus, we re-analyzed our single-cell RNA-sequencing data obtained with SMART-seq2 (Yang et al., 2020).

We first confirmed that the upregulated TFs expression profiles are consistent in bulk RNA-Seq and single-cell RNA-Seq (Figure 2A). Cells from early induction time points (d0–d8) were mixed on the dimensional reduction projection by t-distributed stochastic neighbor embedding (t-SNE) with

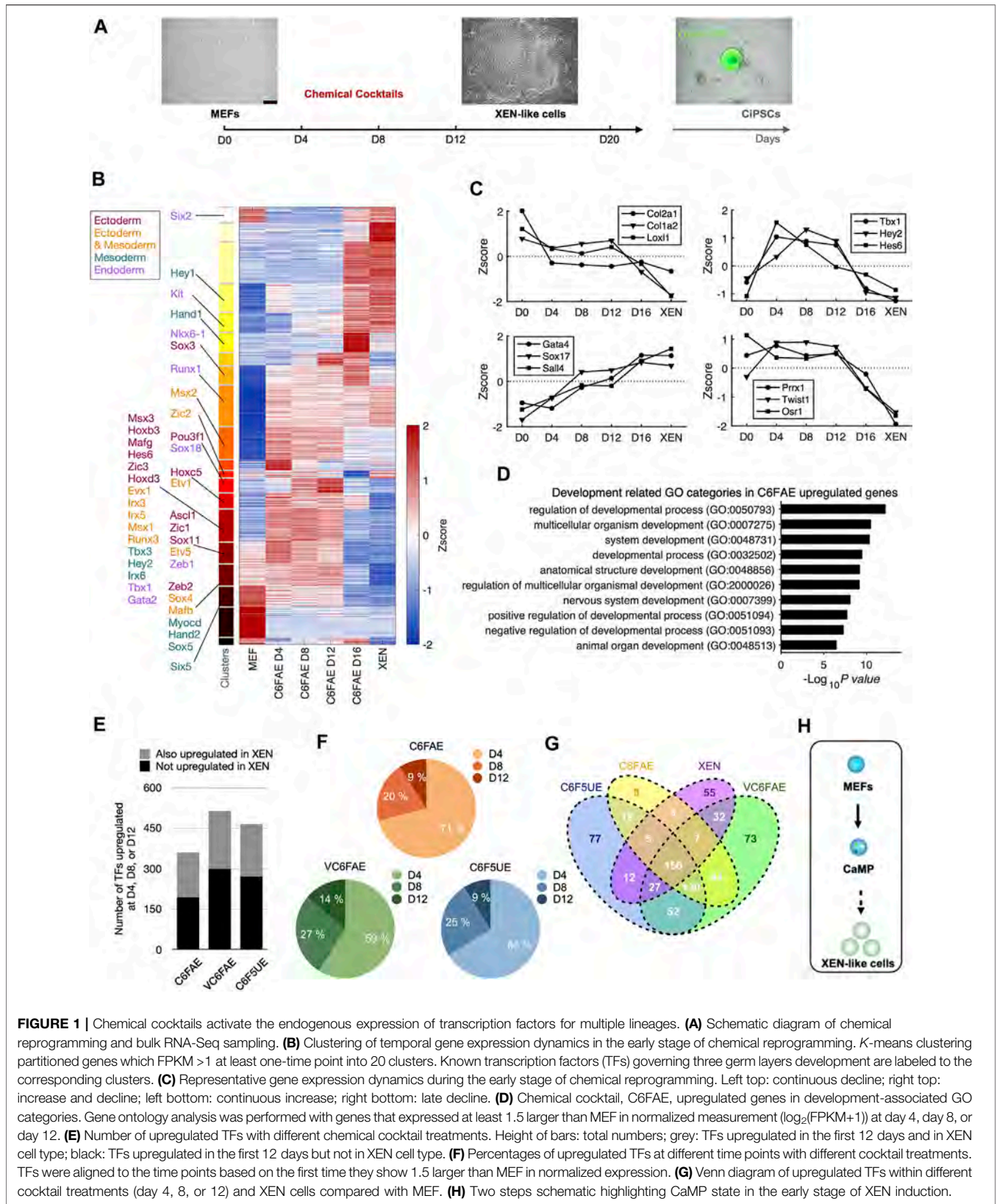
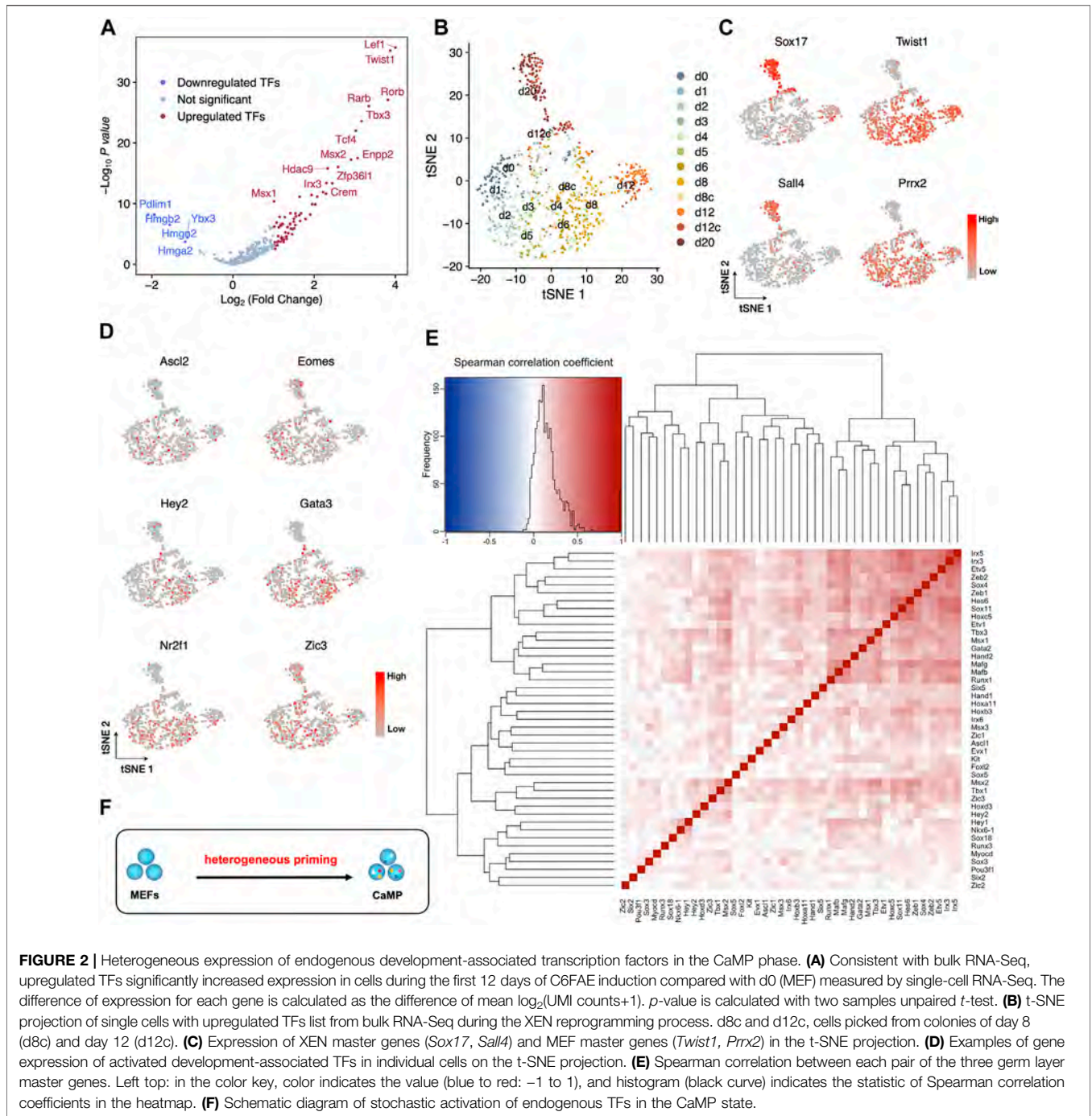


FIGURE 1 | Chemical cocktails activate the endogenous expression of transcription factors for multiple lineages. **(A)** Schematic diagram of chemical reprogramming and bulk RNA-Seq sampling. **(B)** Clustering of temporal gene expression dynamics in the early stage of chemical reprogramming. *K*-means clustering partitioned genes which FPKM > 1 at least one-time point into 20 clusters. Known transcription factors (TFs) governing three germ layers development are labeled to the corresponding clusters. **(C)** Representative gene expression dynamics during the early stage of chemical reprogramming. Left top: continuous decline; right top: increase and decline; left bottom: continuous increase; right bottom: late decline. **(D)** Chemical cocktail, C6FAE, upregulated genes in development-associated GO categories. Gene ontology analysis was performed with genes that expressed at least 1.5 larger than MEF in normalized measurement ($\log_2(\text{FPKM}+1)$) at day 4, day 8, or day 12. **(E)** Number of upregulated TFs with different chemical cocktail treatments. Height of bars: total numbers; grey: TFs upregulated in the first 12 days and in XEN cell type; black: TFs upregulated in the first 12 days but not in XEN cell type. **(F)** Percentages of upregulated TFs at different time points with different cocktail treatments. TFs were aligned to the time points based on the first time they show 1.5 larger than MEF in normalized expression. **(G)** Venn diagram of upregulated TFs within different cocktail treatments (day 4, 8, or 12) and XEN cells compared with MEF. **(H)** Two steps schematic highlighting CaMP state in the early stage of XEN induction.



upregulated TFs detected from bulk RNA-Seq (**Figure 2B**). Cells harvested at day 20 (d20) stretched out of the major population and highly expressed XEN lineage marker genes, *Sox17* and *Sall4*, indicating successfully adapted XEN cell fate (**Figure 2C**). Cells harvested at day 12 (d12) close to the main group kept the expression of MEF master genes (*Twist1* and *Prrx2*) and a low level of *Sox17* (**Figure 2C**). This result denied the hypothesis that upregulated TFs were expressed in a group manner while supporting the other one together with the scatter pattern of their expression on the t-SNE map (**Figure 2D**). Thus, lineage-

specific genes upregulated their mRNA levels in the early days heterogeneously.

Furthermore, correlation analysis shows that the expression of three germ layer master TFs also does not cluster the genes into groups (**Figure 2E**). The pairs of these TFs have low Spearman correlation coefficients except that a few pairs have slightly high coefficients around 0.5, such as *Irx3* and *Irx5*, *Mafb* and *Mafg*, and *Tbx3* and *Msx1* (**Figure 2F**). This may be due to their inherent co-expression patterns or regulation relationships during development (Reinke et al., 2010; Gaborit et al., 2012;

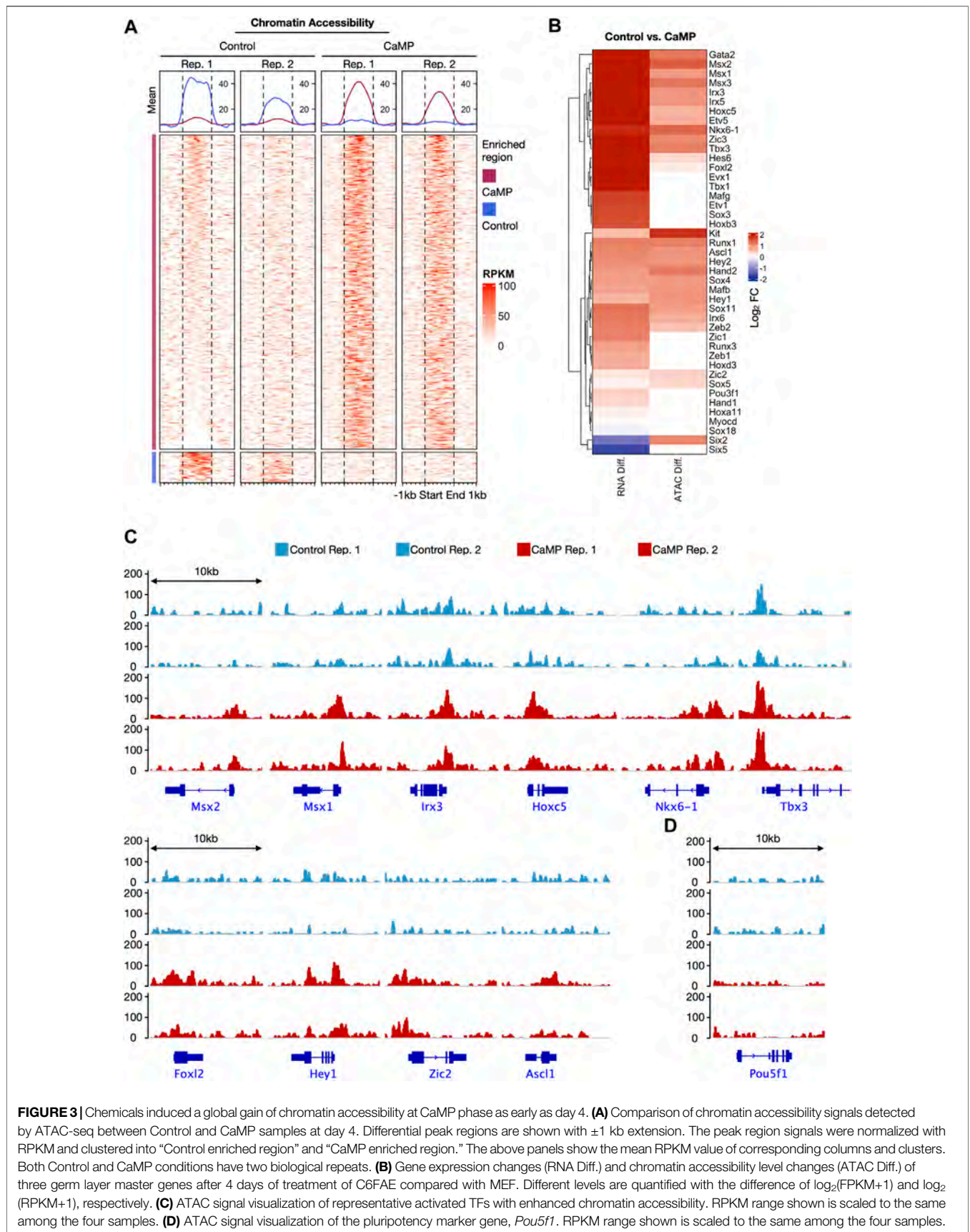


FIGURE 3 | Chemicals induced a global gain of chromatin accessibility at CaMP phase as early as day 4. **(A)** Comparison of chromatin accessibility signals detected by ATAC-seq between Control and CaMP samples at day 4. Differential peak regions are shown with ± 1 kb extension. The peak region signals were normalized with RPKM and clustered into “Control enriched region” and “CaMP enriched region.” The above panels show the mean RPKM value of corresponding columns and clusters. Both Control and CaMP conditions have two biological repeats. **(B)** Gene expression changes (RNA Diff.) and chromatin accessibility level changes (ATAC Diff.) of three germ layer master genes after 4 days of treatment of C6FAE compared with MEF. Different levels are quantified with the difference of $\log_2(\text{FPKM}+1)$ and $\log_2(\text{RPKM}+1)$, respectively. **(C)** ATAC signal visualization of representative activated TFs with enhanced chromatin accessibility. RPKM range shown is scaled to the same among the four samples. **(D)** ATAC signal visualization of the pluripotency marker gene, *Pou5f1*. RPKM range shown is scaled to the same among the four samples.

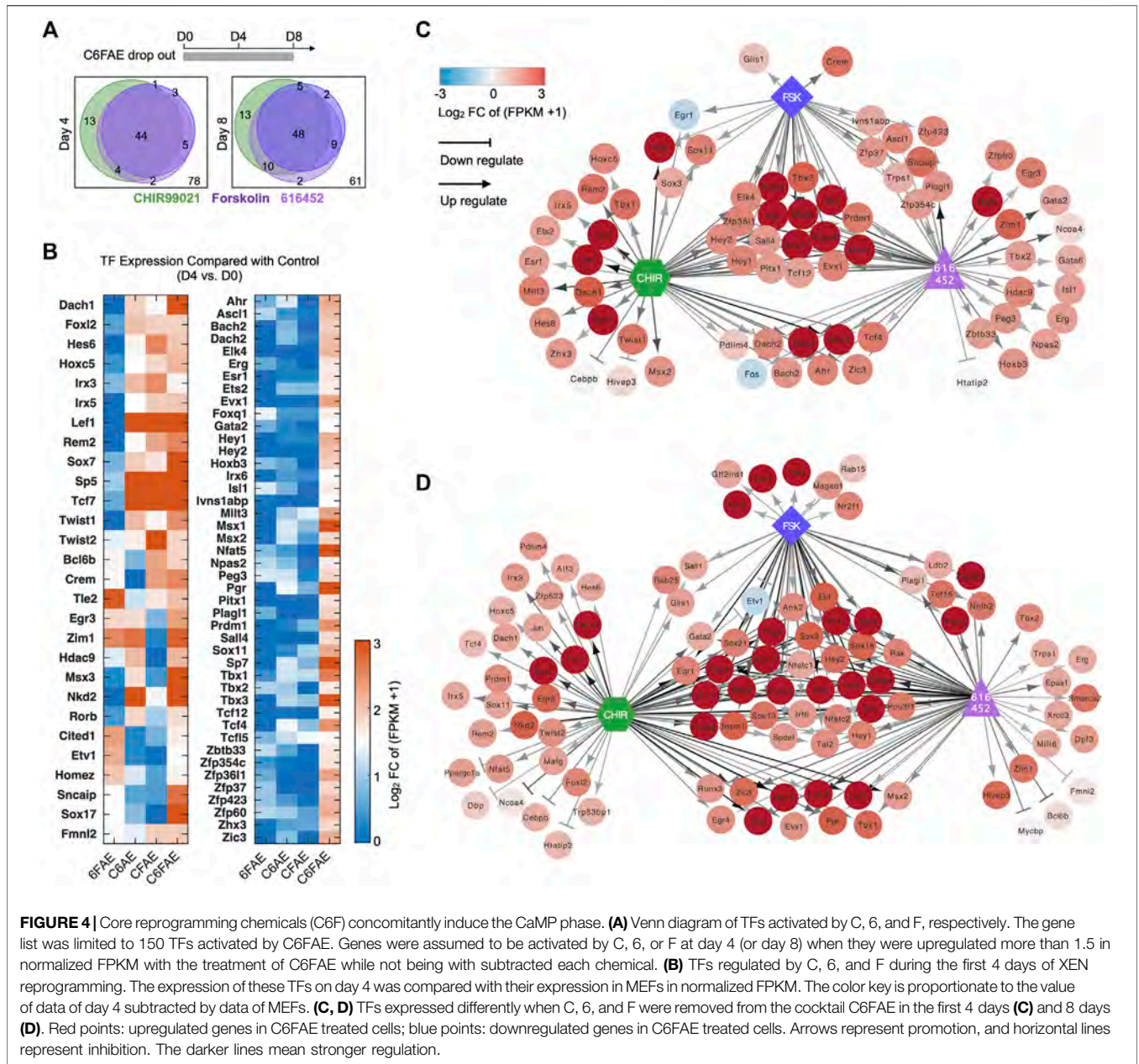


FIGURE 4 | Core reprogramming chemicals (C6F) concomitantly induce the CaMP phase. **(A)** Venn diagram of TFs activated by C, 6, and F, respectively. The gene list was limited to 150 TFs activated by C6FAE. Genes were assumed to be activated by C, 6, or F at day 4 (or day 8) when they were upregulated more than 1.5 in normalized FPKM with the treatment of C6FAE while not being with subtracted each chemical. **(B)** TFs regulated by C, 6, and F during the first 4 days of XEN reprogramming. The expression of these TFs on day 4 was compared with their expression in MEFs in normalized FPKM. The color key is proportionate to the value of data of day 4 subtracted by data of MEFs. **(C, D)** TFs expressed differently when C, 6, and F were removed from the cocktail C6FAE in the first 4 days **(C)** and 8 days **(D)**. Red points: upregulated genes in C6FAE treated cells; blue points: downregulated genes in C6FAE treated cells. Arrows represent promotion, and horizontal lines represent inhibition. The darker lines mean stronger regulation.

Munshi, 2012). Thus, the endogenous expression of TFs for multiple lineages was activated more stochastically. Therefore, the early initiated plastic state was formatted as heterogeneous priming (Figure 2F).

Chemicals Induced a Global Gain of Chromatin Accessibility at CaMP State as Early as Day 4

We further explored the chromatin accessibility change of the CaMP state by ATAC-seq. We detected elevation in chromatin accessibility of the upregulated TFs after 4 days of treatment with chemical compounds.

Interestingly, we found that the CaMP phase significantly opened chromatin accessibility for a large number of genes compared to Control, and only a small proportion of genes had their chromatin accessibility state closed (Figure 3A). This finding was consistent with a substantial upregulation of gene expression induced by chemical compounds. The upregulated level of the activated three germ layer TFs in expression detected by RNA-Seq was positively correlated with the chromatin accessibility change (Figure 3B). In particular, some of the activated lineage-specific TFs, such as *Msx1*, *Nkx6-1*, *Tbx3*, *Zic2*, and *Ascl1*, had strong upregulation of chromatin accessibility after C6FAE treatment (Figure 3C). A few TFs in this list do not show a significant increase in chromatin

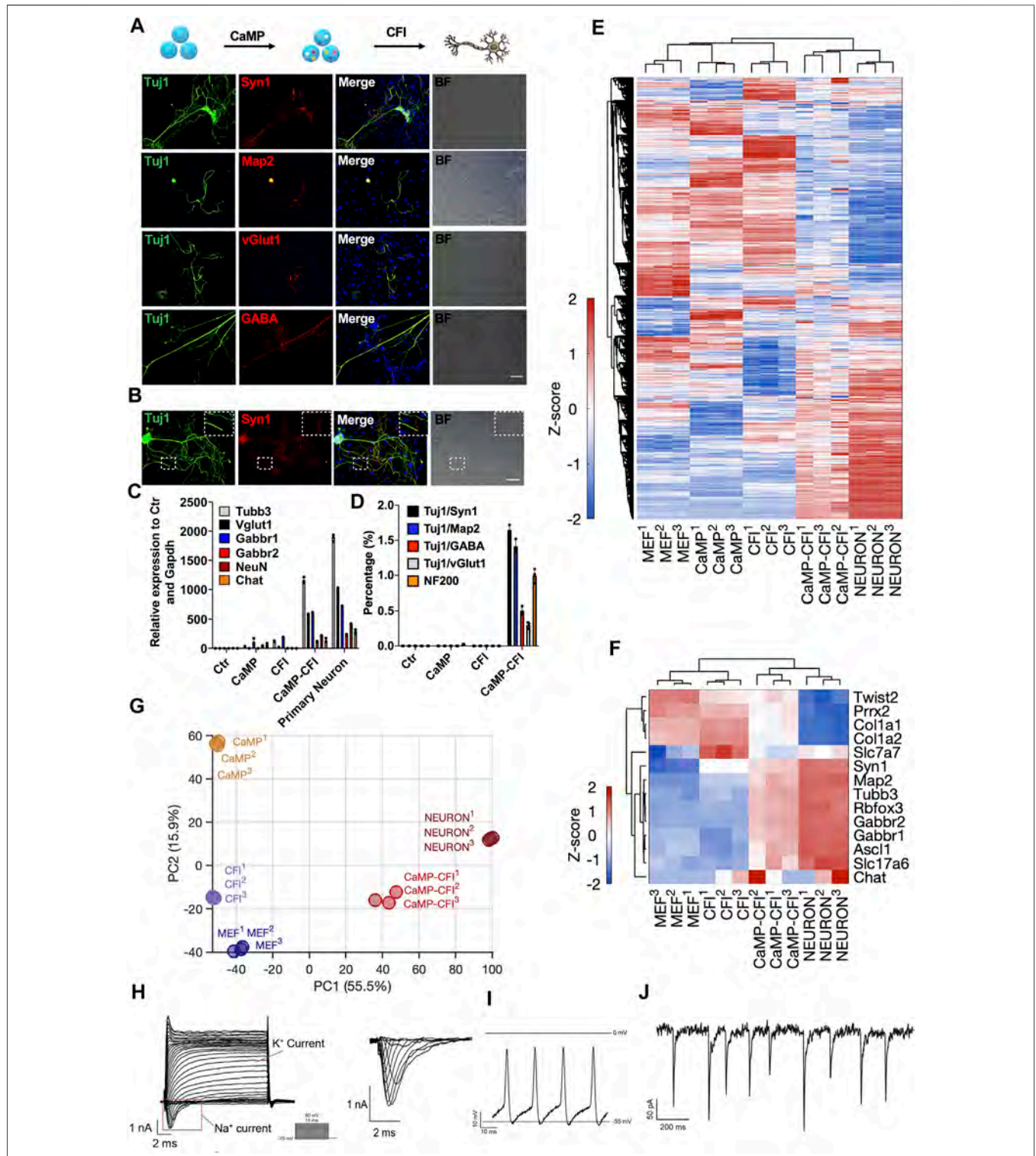


FIGURE 5 | Direct chemical reprogramming system into neuronal lineage was empowered with CaMP induction. **(A)** Immunostaining for the chemically induced neuron-like cells with the CaMP pretreatment. Scale bar, 100 μm; CFI, CHIR99021/616452/Forskolin/ISX-9. **(B)** Immunostaining of CaMP-CFI induced functional synapses identified by Syn1 immunostaining. Scale bar, 100 μm. **(C)** Relative mRNA expression of typical neuronal marker genes induced with or without CaMP pretreatment. **(D)** The efficiency statistics of matured neuron-like cells identified by immunostaining. **(E)** Gene expression heatmap of all differentially expressed genes (normalized FPKM changed more than 1.5 among the samples) in neuron-like cells induced with or without CaMP step. The color key, Z-score of normalized FPKM. **(F)** Gene expression heatmap of typical neuron genes in neuron-like cells induced with or without CaMP step. The color key, Z-score of normalized FPKM. **(G)** (Continued)

FIGURE 5 | PCA projection of neuron-like cells reprogramming processes analyzed with all differentially expressed genes. The coefficient of variation (CV) was calculated for each gene among samples, and the PCA was performed with genes with CV larger than 0.1. **(H)** Action potentials of CaMP-CFI induced neuron-like cells after co-culture with astrocytes. One exemplary action potential trace was highlighted. **(I)** Spontaneous excitement potential of CaMP-CFI induced neuron-like cells after co-culture with astrocytes. **(J)** Spontaneous excitatory postsynaptic currents of CaMP-CFI induced neuron-like cells after co-culture with astrocytes.

accessibility due to their original highly open chromatin state (**Supplementary Figure S2**).

Moreover, chromatin accessibility of pluripotent genes *Pou5f1* was still not opened after 4 days of treatment with C6FAE (**Figure 3D**), which meant the global open of chromatin accessibility was not caused by the activation of the pluripotent gene. Overall, the change of chromatin accessibility was further in line with the activation of the development-associated TFs in the CaMP state.

Core Reprogramming Chemicals (C6F) Concomitantly Induce the CaMP Phase

We further investigated the roles of the essential reprogramming chemicals, CHIR99021, 616452, and Forskolin (C6F) in inducing the plastic CaMP phase. We compared the bulk RNA expression profiles of samples with the treatment of partial cocktails and those with full cocktails. CHIR99021, 616452, and Forskolin were essential for the transcriptional activation of most lineage-specific TFs in the CaMP state. Most of the TFs activated by the cocktail of C6FAE could not be activated without any one of CHIR99021, 616452, and Forskolin (**Figures 4A,B**). Subtracting any one of CHIR99021, 616452, and Forskolin from day 0 also hampered the expression of the XEN master TFs (**Figures 4C,D**). Thus, The cooperation of the three core chemicals activated the expression of those TFs in the CaMP phase.

In summary, CHIR99021, 616452, and Forskolin concomitantly initiated the CaMP state. All of them contributed to transcriptional activation of development-associated TFs, which explained why most previous chemical reprogramming systems used these three small molecules or those targeting the same pathways.

Direct Chemical Reprogramming System Into Neuronal Lineage Was Empowered With CaMP Induction

Inspired by the molecular frameworks during cell fate specification in another study of us (Yang et al., 2020), the endogenously activated TFs of multiple lineages in the CaMP induction might be beneficial to induce cell types of other lineages. We found that the TFs of neurons, including *Ascl1*, a master gene for neuronal reprogramming, were also activated in the CaMP phase. It is possible to induce the neuron-like cells after CaMP induction more efficiently.

By initially introducing CaMP state and changing the culture medium to which favored neuronal maintenance in culture and fine-tuning the composition of chemical cocktails after the CaMP phase, we found that a cocktail, CHIR99021, Forskolin, and ISX9 (CFI), drastically induced the transition from the CaMP to neuron-like cells only after the pretreatment

of C6FAE for 4 days. The resulting induced neuron-like cells had more classic neuronal cell morphology and expressed typical neuronal markers *Tuj1*, *Map2*, *Syn1*, *neurofilam 200*, and functional markers *vGlut1*, *GABA*, *Rbfox3*, *Gabbr2*, and *Chat* (**Figures 5A–C**). The efficiency of matured neuron-like cells identified by functional synapses marker-Syn1 and *Tuj1* co-staining was about 1.7%, and the proportions of GABA or *vGlut1* positive cells were about 0.5% and 0.3%, respectively (**Figure 5D**). In comparison, cells induced without the CaMP pretreatment expressed nearly no mature neural markers after 28 days of induction (**Figure 5D**).

By RNA sequencing, we found that neuron-like cells induced through CaMP had activated the expression of neuron-specific genes and similar expression profiles to primary neurons. (**Figures 5E,F**). By the principal component analysis, we found that the neuron-like cells induced through CaMP induction had transcriptional states closer to primary functional cells than cells induced without CaMP priming (**Figure 5G**). Importantly, action potential, spontaneous excitement potential, and spontaneous excitatory postsynaptic currents (EPSCs) were recorded in induced neuron-like cells after co-culture with astrocytes (**Figures 5H–J** and **Supplementary Table S1**).

Direct Chemical Reprogramming System Into Myocytes Was Empowered With CaMP Induction

Similarly, we optimized the myocytes' induction through the CaMP state. By treating cells in the CaMP phase with myocyte culture medium containing CHIR99021, 616452, Forskolin, and SB431542 (C6FS), we induced contractile and multinucleated skeletal muscle cells expressing *MyHC*, *Myog*, *Myod1*, α -actinin, and *Tnnt3* in 8–12 days, more efficient and faster than induction with only C6FS (**Figures 6A,B** and **Supplementary video S1**). The efficiency of *MyHC* and α -actinin double-positive skeletal muscle cells was over 4% with P2 MEFs as starting cells. The efficiency of *MyHC* and α -actinin double-positive skeletal muscle cells was over 4%, and the efficiency could reach 30% with P1 MEFs as starting cells. In comparison, few skeletal muscles cells could be induced without CaMP induction or specification stage with myocyte medium (**Figure 6C**).

By RNA sequencing, we found that the skeletal muscle cells induced through CaMP had activated the expression of myocyte-specific genes (**Figures 6D,E**). By principal component analysis, we found that the myocytes induced through CaMP induction had more comparable transcriptional profiles to primary functional cells than the cells induced without the CaMP priming (**Figure 6F**).

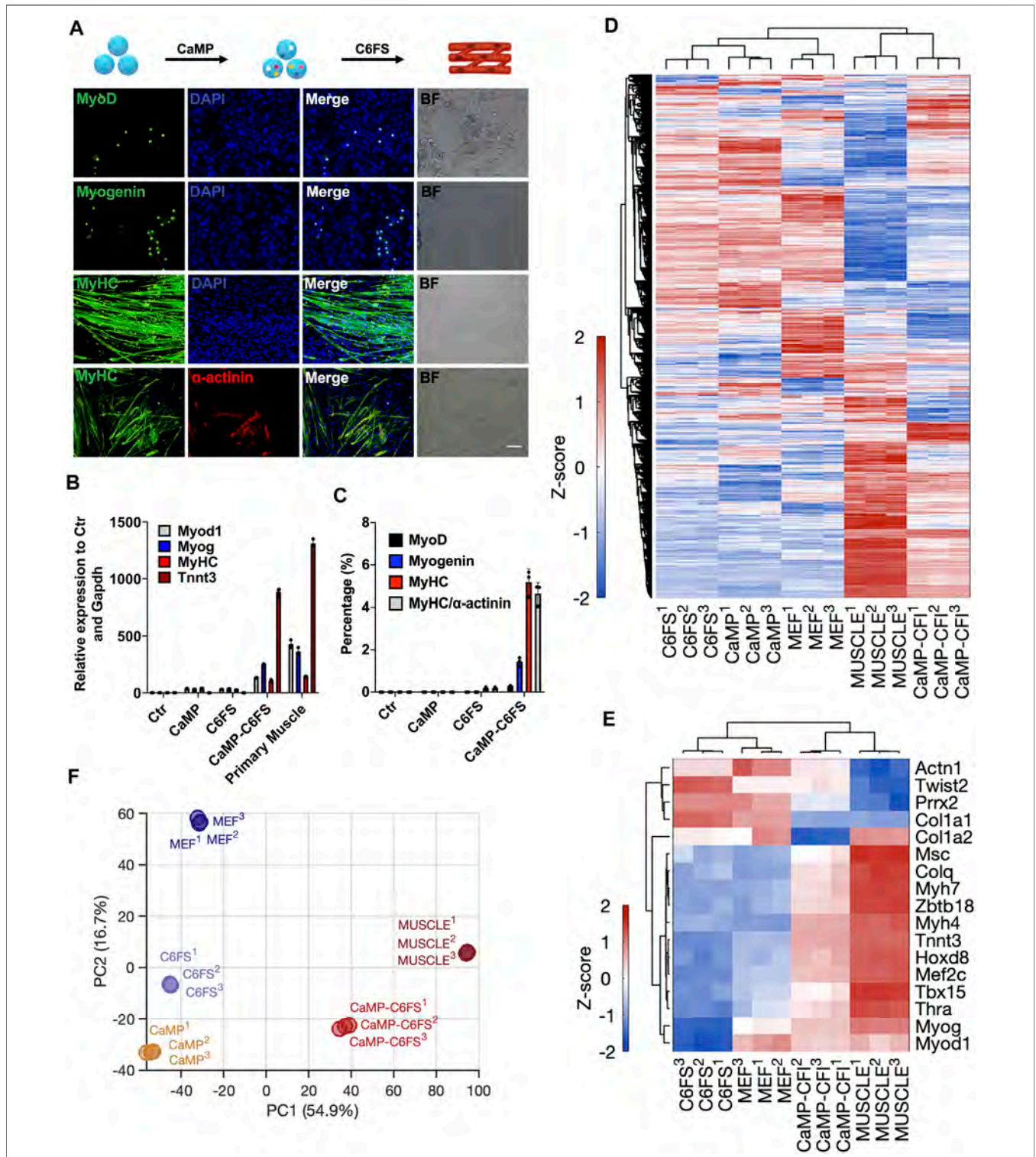


FIGURE 6 | Direct chemical reprogramming system into myocytes was empowered with CaMP induction. **(A)** Induction of skeletal muscle cells with the CaMP pretreatment. Scale bar, 100 μ m; C6FS, CHIR99021/616452/Forskolin/SB431542. **(B)** Relative mRNA expression of typical skeletal muscle marker genes induced with or without CaMP pretreatment. **(C)** The efficiency statistics of induced skeletal muscle cells identified by immunostaining. **(D)** Gene expression heatmap of all differentially expressed genes (normalized FPKM changed more than 1.5 among the samples) in skeletal muscle cells induced with or without CaMP step (analyzed by RNA-Seq of more than 30 clusters of induced skeletal muscle cells). The color key, Z-score of normalized FPKM. **(E)** Gene expression heatmap of typical skeletal muscle genes in skeletal muscle cells induced with or without CaMP step (analyzed by RNA-Seq of more than 30 clusters of induced skeletal muscle cells). The color key, Z-score of normalized FPKM. **(F)** PCA projection of myocytes reprogramming processes analyzed with all differentially expressed genes. The coefficient of variation (CV) was calculated for each gene among samples, and the PCA was performed with genes with CV larger than 0.1.

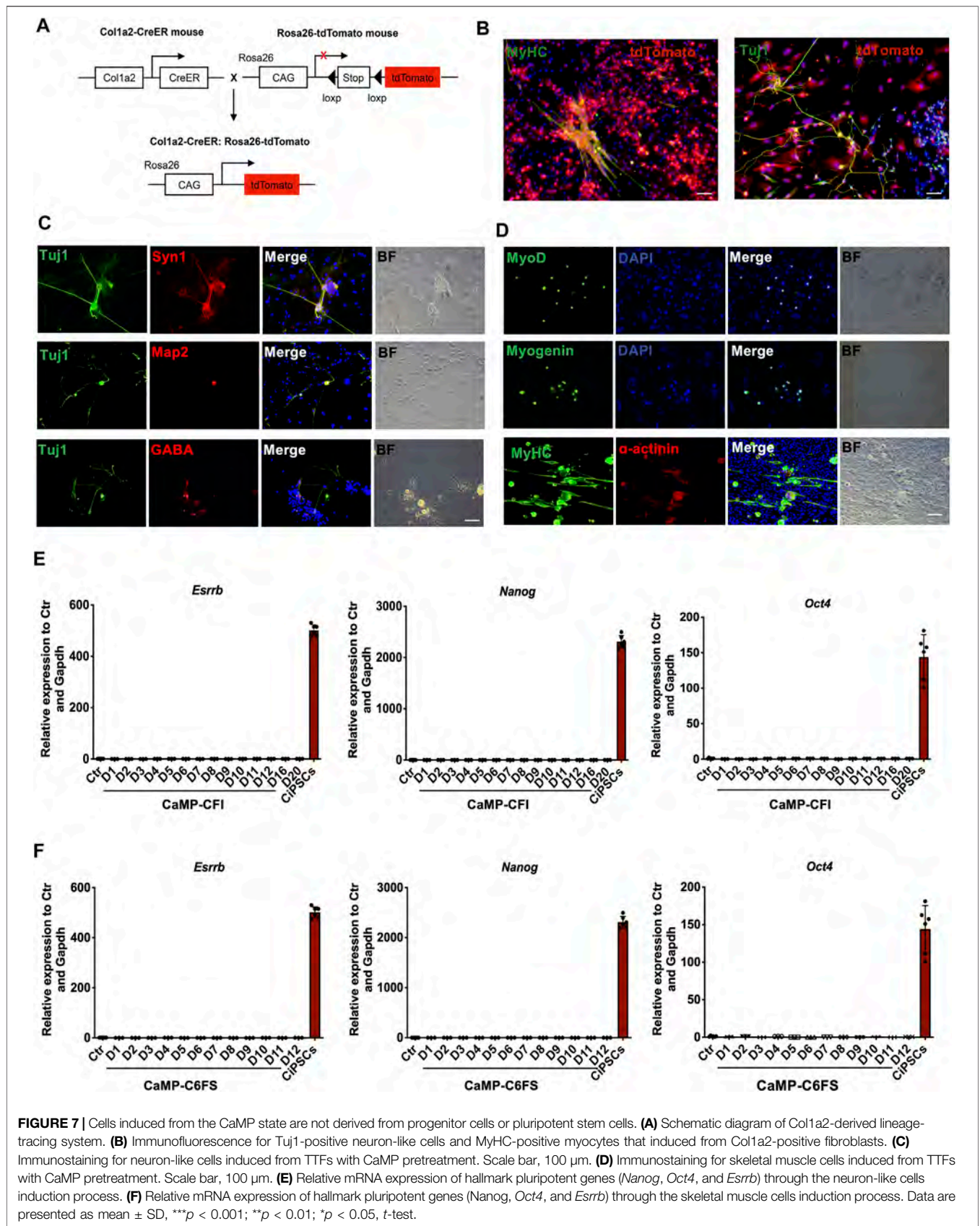
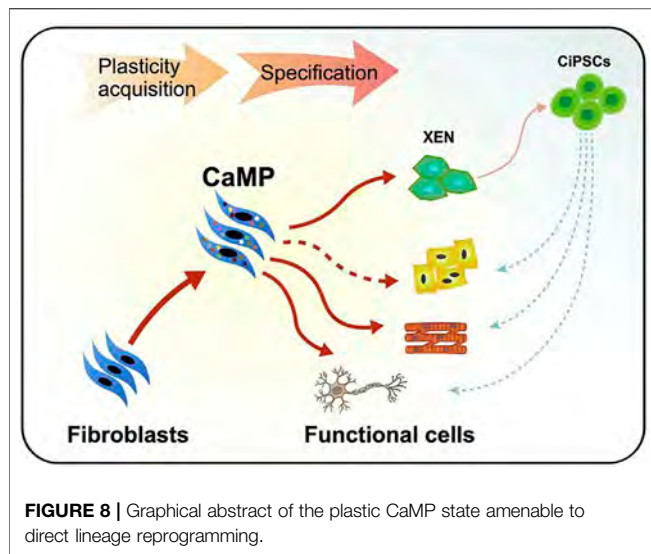


FIGURE 7 | Cells induced from the CaMP state are not derived from progenitor cells or pluripotent stem cells. **(A)** Schematic diagram of Col1a2-derived lineage-tracing system. **(B)** Immunofluorescence for Tuj1-positive neuron-like cells and MyHC-positive myocytes that induced from Col1a2-positive fibroblasts. **(C)** Immunostaining for neuron-like cells induced from TTFs with CaMP pretreatment. Scale bar, 100 μ m. **(D)** Immunostaining for skeletal muscle cells induced from TTFs with CaMP pretreatment. Scale bar, 100 μ m. **(E)** Relative mRNA expression of hallmark pluripotent genes (*Nanog*, *Oct4*, and *Esrrb*) through the neuron-like cells induction process. **(F)** Relative mRNA expression of hallmark pluripotent genes (*Nanog*, *Oct4*, and *Esrrb*) through the skeletal muscle cells induction process. Data are presented as mean \pm SD, *** p < 0.001; ** p < 0.01; * p < 0.05, t -test.



Cells Induced From the CaMP State Are Not Derived From Progenitor Cells or Pluripotent Stem Cells

To determine whether the induced neuron-like cells and myocytes were derived from fibroblasts and to rule out the contamination of progenitor cells in fibroblast culture, we applied the Col1a2 lineage-tracing system during the induction of neuron-like cells and myocytes. By immunostaining, we found that Tuj1 or MyHC expressing cells were mostly induced from Col1a2 expressing cells, suggesting that neuron-like cells were induced from fibroblasts rather than contaminated progenitors (**Figures 7A,B**). Furthermore, we induced adult mouse tail tip fibroblasts into neuron-like cells and beating myocytes expressing specific markers by introducing CaMP state (**Figures 7C,D**), which confirmed that CaMP induced plasticity in TTFs and further ruled out the contamination of neural progenitor cells in MEFs.

Besides, we had not detected the endogenous expression of pluripotent genes, such as *Nanog*, *Esrrb*, and *Oct4*, throughout the chemical reprogramming processes by RT-qPCR analysis (**Figures 7E,F**). The Oct4-GFP was not activated during the chemical reprogramming processes to neuron-like cells and myocytes by daily observation. These indicated that the neuron-like cells induction processes initiated by CaMP pretreatment did not activate the pluripotent genes.

Moreover, we found that the neuron-like cells and myocytes induced through the CaMP phase did not require the intermediate XEN-like states. The gene *Sall4*, a typically expressed master TF of XEN cells, had low expression in the process of reprogramming. The knockdown of XEN master genes, *Sall4* and *Gata4*, impaired the formation of XEN-like colonies but did not decrease the induction efficiency of neuron-like cells or myocytes (**Supplementary Figure S3**). It indicated a more direct cell fate conversion from fibroblasts to target cell types without establishing the XEN-like cell identity (**Figure 8**). These findings supported that cell plasticity with

neuron-like cells and myocyte lineage specification potential was induced during the CaMP process.

DISCUSSION

In this study, we found that the cell fate specification was not initially determined in the chemical reprogramming process. Instead, a plastic CaMP state was induced, with the heterogeneous expression of multiple developmental genes, and without the determination to any specific cell fate. The establishment of such cell plasticity may account for the common roles of these key chemicals used in the CaMP induction in inducing different cell types.

Our findings provide a new understanding of cell plasticity and stability. Although it has been reported that the master regulators of cell fates, such as tumorigenicity-related genes, are always regulated strictly by multiple epigenetic mechanisms (Graf and Enver, 2009; Li et al., 2016; Dhar et al., 2018), our findings suggest that a considerable number of developmental-associated TFs are not quite strictly regulated in fibroblasts. These suggest that somatic cells possess plasticity in response to exogenous stimuli, in terms of expressing master genes for another cell fate, which could be an initial step and a priming phase for cell fate conversion (Dhar et al., 2018). In addition, the CaMP state would be reminiscent of pluripotent stem cells (PSCs) regarding their specification potential into cells of differentiated germ layers, such as myocytes and neuron-like cells as indicated in this study, as well as extraembryonic cell types like XEN-like cells in the previous report (Zhao et al., 2015). Even PSCs have a priming state with stochastically low expression of developmental genes (Bernstein et al., 2006), similar to the stochastic activation of developmental genes in the CaMP cells revealed in this study. These may suggest similar molecular bases for cell plasticity in the CaMP phase and pluripotent cells. In contrast, cells in the CaMP phase differ from pluripotent stem cells regarding their different gene expression profiles, spontaneous differentiation potential, and development potential in a single cell.

Furthermore, cells in the CaMP phase may not have the potential to differentiate by nature, with the fact that the treatment of chemicals and culture medium after CaMP are also essential for determining the cell fate specification derivative. Our previous study found that the trigger for cell fate specification is also very critical to hierarchically activating all the essential TFs for cell fate determination and transition (Yang et al., 2020). The initial fibroblasts program could not be substantially impaired unless major master TFs for another cell type are all co-expressed in the final transition stage (Yang et al., 2020). These support the concept that cell fate is somehow stable and not easily reversible, although easily primed.

In comparison, the process of transgenes *OSKM*- (*Oct4*, *Sox2*, *Klf4*, and *c-Myc*) induced pluripotent stem cell (iPS) generation is also a similar biphasic process, with an early stochastic gene-activation stage induced mainly by *c-Myc* and a late, more determined process mainly orchestrated by OSK, downstream of *Sox2* expression (Sridharan et al., 2009; Buganim et al., 2012;

Polo et al., 2012). Recently, the heterogeneity of early-reprogramming cells expressing considerable development-associated genes induced by OSKM has also been reported (Schiebinger et al., 2019). These suggest that the biphasic “plasticization and specification” process revealed in our study could be a general principle for cell-fate reprogramming for both chemical and transgenic reprogramming.

Importantly, by harnessing the CaMP state induced in the initial stage of chemical reprogramming, we improved the reprogramming systems towards neuron-like cells and myocytes with pure chemicals by pretreating the cells with the CaMP inducing chemical cocktails. Moreover, the resting membrane potential of CaMP-induced neuron-like cells was -48.68 ± 2.43 mV (**Supplementary Table S1**), which was significantly lower than other reported chemical-induced neurons (-25 or -35 mV). It indicates that CaMP-induced neuron-like cells had more complete ion channels and were functionally closer to primary neurons. Besides, rather than using limited genes as biomarkers (typically done in other reported chemical reprogramming systems) (Hu et al., 2015; Li et al., 2015; Li et al., 2017), we compared all differentially expressed genes among samples. We found that CaMP-induced neuron-like cells had more similar transcriptional profiles to the primary functional cells. Overall, the CaMP-induced neuron-like cells were more mature in transcriptional profile and function than those previously reported. It would also be interesting to further determine whether chemical reprogramming through the CaMP state can be extended to obtain other functional cell types as a general strategy for developing chemical reprogramming systems and even be applied to human cells.

Similar to this strategy, a cell activation and signaling-directed (CASD) strategy, has been reported by transiently overexpressing Yamanaka factors, OSKM, to obtain different functional cell types, such as hepatocytes, pancreatic beta cells, and cardiomyocytes (Efe et al., 2011; Li et al., 2014; Wang et al., 2014; Zhu et al., 2014), although it was found that this strategy involved the transient acquisition of pluripotency (Bar-Nur et al., 2015; Maza et al., 2015). However, in chemical reprogramming through the CaMP state, it is not likely that chemicals induced a pluripotent state in the initial first 4 days of the 40 days long chemical reprogramming process towards CiPSCs. The chemicals rather established a more plastic state with a more active epigenetic state beneficial for transcriptional activation. Besides, it has been reported that the XEN-like cells induced during chemical reprogramming to CiPSCs are also plastic and can be further induced into other cell lineages, such as neurons or hepatocyte-like cells (Li et al., 2017). In comparison, our study showed that cell plasticity can be induced at the very beginning of chemical reprogramming for further cell specification, even before the establishment of XEN cell identity and without substantial silencing of core transcriptional networks of fibroblasts.

In comparison with cell differentiation from induced pluripotent stem cells or expandable XEN-like cells, cell fate lineage reprogramming systems through the CaMP state are more direct, bypassing the concerns of potential tumorigenicity resulting from uncontrolled cell expansion in *in vivo* applications and has the potential to be induced to cell types of all three germ layers. As a result, direct cell fate reprogramming through the CaMP state may be a new paradigm and a shortcut to obtaining functional cells for regenerative medicine (**Figure 8**).

CONCLUSION

This study enlightens the understanding of chemical reprogramming by dissecting the contribution of reprogramming chemicals to the activation of development-associated transcription factors. It proves a new approach to obtain functional cell types through a CaMP state in the future of regenerative medicine.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/geo/>, GSE155818.

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) and Use Committee at Peking University.

AUTHOR CONTRIBUTIONS

ZY performed the experiments. XX conducted the bioinformatics analyses. CG participated in the ATACseq analysis. AN and GC helped with data analysis and experiments. YZ, CT and FG supervised this project. ZY, XX and YZ wrote the paper. All authors reviewed the manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (2018YFA0800504), the National Natural Science Foundation of China (31922020 and 31771645) and National Key Research and Development Program of China (2018YFA0107701).

ACKNOWLEDGMENTS

We thank Xudong Wu, Qiushi Sun, Yi Zhang, Zhidan Wang, and Yang Liu for technical assistance. We thank Bin Zhou for the gift of Col1a2-CreERT2 mice. We thank Gongxin Wang for the electrophysiological assay. We thank Iain C. Bruce for critical reading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcell.2022.865038/full#supplementary-material>

REFERENCES

- Bansal, V., De, D., An, J., Kang, T. M., Jeong, H.-J., Kang, J.-S., et al. (2019). Chemical Induced Conversion of Mouse Fibroblasts and Human Adipose-Derived Stem Cells into Skeletal Muscle-like Cells. *Biomaterials* 193, 30–46. doi:10.1016/j.biomaterials.2018.11.037
- Bar-Nur, O., Verheul, C., Sommer, A. G., Brumbaugh, J., Schwarz, B. A., Lipchina, I., et al. (2015). Lineage Conversion Induced by Pluripotency Factors Involves Transient Passage through an iPSC Stage. *Nat. Biotechnol.* 33, 761–768. doi:10.1038/nbt.3247
- Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., et al. (2006). A Bivalent Chromatin Structure marks Key Developmental Genes in Embryonic Stem Cells. *Cell* 125, 315–326. doi:10.1016/j.cell.2006.02.041
- Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., et al. (2012). Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* 150, 1209–1222. doi:10.1016/j.cell.2012.08.023
- Cao, N., Huang, Y., Zheng, J., Spencer, C. I., Zhang, Y., Fu, J.-D., et al. (2016). Conversion of Human Fibroblasts into Functional Cardiomyocytes by Small Molecules. *Science* 352, 1216–1220. doi:10.1126/science.aaf1502
- Cao, S., Yu, S., Chen, Y., Wang, X., Zhou, C., Liu, Y., et al. (2017). Chemical Reprogramming of Mouse Embryonic and Adult Fibroblast into Endoderm Lineage. *J. Biol. Chem.* 292, 19122–19132. doi:10.1074/jbc.M117.812537
- Cheng, L., Hu, W., Qiu, B., Zhao, J., Yu, Y., Guan, W., et al. (2014). Generation of Neural Progenitor Cells by Chemical Cocktails and Hypoxia. *Cell Res* 24, 665–679. doi:10.1038/cr.2014.32
- Dhar, S. S., Zhao, D., Lin, T., Gu, B., Pal, K., Wu, S. J., et al. (2018). MLL4 Is Required to Maintain Broad H3K4me3 Peaks and Super-enhancers at Tumor Suppressor Genes. *Mol. Cell* 70, 825–841. e826. doi:10.1016/j.molcel.2018.04.028
- Efe, J. A., Hilcove, S., Kim, J., Zhou, H., Ouyang, K., Wang, G., et al. (2011). Conversion of Mouse Fibroblasts into Cardiomyocytes Using a Direct Reprogramming Strategy. *Nat. Cell Biol* 13, 215–222. doi:10.1038/ncb2164
- Fu, Y., Huang, C., Xu, X., Gu, H., Ye, Y., Jiang, C., et al. (2015). Direct Reprogramming of Mouse Fibroblasts into Cardiomyocytes with Chemical Cocktails. *Cel Res* 25, 1013–1024. doi:10.1038/cr.2015.99
- Gaborit, N., Sakuma, R., Wylie, J. N., Kim, K.-H., Zhang, S.-S., Hui, C.-C., et al. (2012). Cooperative and Antagonistic Roles for Irx3 and Irx5 in Cardiac Morphogenesis and Postnatal Physiology. *Development* 139, 4007–4019. doi:10.1242/dev.081703
- Graf, T., and Enver, T. (2009). Forcing Cells to Change Lineages. *Nature* 462, 587–594. doi:10.1038/nature08533
- Gu, Z., Eils, R., Schlesner, M., and Ishaque, N. (2018). EnrichedHeatmap: an R/Bioconductor Package for Comprehensive Visualization of Genomic Signal Associations. *BMC Genomics* 19, 234. doi:10.1186/s12864-018-4625-x
- Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., et al. (2013). Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. *Science* 341, 651–654. doi:10.1126/science.1239278
- Hu, W., Qiu, B., Guan, W., Wang, Q., Wang, M., Li, W., et al. (2015). Direct Conversion of Normal and Alzheimer's Disease Human Fibroblasts into Neuronal Cells by Small Molecules. *Cell Stem Cell* 17, 204–212. doi:10.1016/j.stem.2015.07.006
- Iwafuchi-Doi, M., and Zaret, K. S. (2014). Pioneer Transcription Factors in Cell Reprogramming. *Genes Dev.* 28, 2679–2692. doi:10.1101/gad.253443.114
- Li, K., Zhu, S., Russ, H. A., Xu, S., Xu, T., Zhang, Y., et al. (2014). Small Molecules Facilitate the Reprogramming of Mouse Fibroblasts into Pancreatic Lineages. *Cell Stem Cell* 14, 228–236. doi:10.1016/j.stem.2014.01.006
- Li, N., Li, Y., Lv, J., Zheng, X., Wen, H., Shen, H., et al. (2016). ZMYND8 Reads the Dual Histone Mark H3K4me1-H3K14ac to Antagonize the Expression of Metastasis-Linked Genes. *Mol. Cell* 63, 470–484. doi:10.1016/j.molcel.2016.06.035
- Li, X., Liu, D., Ma, Y., Du, X., Jing, J., Wang, L., et al. (2017). Direct Reprogramming of Fibroblasts via a Chemically Induced XEN-like State. *Cell Stem Cell* 21, 264–273. e267. doi:10.1016/j.stem.2017.05.019
- Li, X., Zuo, X., Jing, J., Ma, Y., Wang, J., Liu, D., et al. (2015). Small-Molecule-Driven Direct Reprogramming of Mouse Fibroblasts into Functional Neurons. *Cell Stem Cell* 17, 195–203. doi:10.1016/j.stem.2015.06.003
- Mahato, B., Kaya, K. D., Fan, Y., Sumien, N., Shetty, R. A., Zhang, W., et al. (2020). Pharmacologic Fibroblast Reprogramming into Photoreceptors Restores Vision. *Nature* 581, 83–88. doi:10.1038/s41586-020-2201-4
- Maza, I., Caspi, I., Zviran, A., Chomsky, E., Rais, Y., Viukov, S., et al. (2015). Transient Acquisition of Pluripotency during Somatic Cell Transdifferentiation with iPSC Reprogramming Factors. *Nat. Biotechnol.* 33, 769–774. doi:10.1038/nbt.3270
- Munshi, N. V. (2012). Gene Regulatory Networks in Cardiac Conduction System Development. *Circ. Res.* 110, 1525–1537. doi:10.1161/CIRCRESAHA.111.260026
- Nie, B., Nie, T., Hui, X., Gu, P., Mao, L., Li, K., et al. (2017). Brown Adipogenic Reprogramming Induced by a Small Molecule. *Cel Rep.* 18, 624–635. doi:10.1016/j.celrep.2016.12.062
- Polo, J. M., Anderssen, E., Walsh, R. M., Schwarz, B. A., Nefzger, C. M., Lim, S. M., et al. (2012). A Molecular Roadmap of Reprogramming Somatic Cells into iPSCs. *Cell* 151, 1617–1632. doi:10.1016/j.cell.2012.11.039
- Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deepTools: a Flexible Platform for Exploring Deep-Sequencing Data. *Nucleic Acids Res.* 42, W187–W191. doi:10.1093/nar/gku365
- Reinke, A. W., Grigoryan, G., and Keating, A. E. (2010). Identification of bZIP Interaction Partners of Viral Proteins HBZ, MEQ, BZLF1, and K-bZIP Using Coiled-Coil Arrays. *Biochemistry* 49, 1985–1997. doi:10.1021/bi902065k
- Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* 176, 928–943. e922. doi:10.1016/j.cell.2019.01.006
- Sridharan, R., Tchieu, J., Mason, M. J., Yachechko, R., Kuoy, E., Horvath, S., et al. (2009). Role of the Murine Reprogramming Factors in the Induction of Pluripotency. *Cell* 136, 364–377. doi:10.1016/j.cell.2009.01.001
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888–1902. doi:10.1016/j.cell.2019.05.031
- Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* 126, 663–676. doi:10.1016/j.cell.2006.07.024
- Takeda, Y., Harada, Y., Yoshikawa, T., and Dai, P. (2017). Direct Conversion of Human Fibroblasts to Brown Adipocytes by Small Chemical Compounds. *Sci. Rep.* 7, 4304. doi:10.1038/s41598-017-04665-x
- Tian, E., Sun, G., Sun, G., Chao, J., Ye, P., Warden, C., et al. (2016). Small-Molecule-Based Lineage Reprogramming Creates Functional Astrocytes. *Cel Rep.* 16, 781–792. doi:10.1016/j.celrep.2016.06.042
- Wang, H., Cao, N., Spencer, C. I., Nie, B., Ma, T., Xu, T., et al. (2014). Small Molecules Enable Cardiac Reprogramming of Mouse Fibroblasts with a Single Factor, Oct4. *Cel Rep.* 6, 951–960. doi:10.1016/j.celrep.2014.01.038
- Wang, Y., Qin, J., Wang, S., Zhang, W., Duan, J., Zhang, J., et al. (2016). Conversion of Human Gastric Epithelial Cells to Multipotent Endodermal Progenitors using Defined Small Molecules. *Cell Stem Cell* 19, 449–461. doi:10.1016/j.stem.2016.06.006
- Xu, J., Du, Y., and Deng, H. (2015). Direct Lineage Reprogramming: Strategies, Mechanisms, and Applications. *Cell Stem Cell* 16, 119–134. doi:10.1016/j.stem.2015.01.013
- Yang, Z., Xu, X., Gu, C., Li, J., Wu, Q., Ye, C., et al. (2020). Chemicals Orchestrate Reprogramming with Hierarchical Activation of Master Transcription Factors Primed by Endogenous Sox17 Activation. *Commun. Biol.* 3, 629. doi:10.1038/s42003-020-01346-w
- Yin, J. C., Zhang, L., Ma, N. X., Wang, Y., Lee, G., Hou, X. Y., et al. (2019). Chemical Conversion of Human Fetal Astrocytes into Neurons through Modulation of

- Multiple Signaling Pathways. *Stem Cell Rep.* 12, 488–501. doi:10.1016/j.stemcr.2019.01.003
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoutte, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi:10.1186/gb-2008-9-9-r137
- Zhao, T., Fu, Y., Zhu, J., Liu, Y., Zhang, Q., Yi, Z., et al. (2018). Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. *Cell Stem Cell* 23, 31–45. e37. doi:10.1016/j.stem.2018.05.025
- Zhao, Y. (2019). Chemically Induced Cell Fate Reprogramming and the Acquisition of Plasticity in Somatic Cells. *Curr. Opin. Chem. Biol.* 51, 146–153. doi:10.1016/j.cbpa.2019.04.025
- Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., et al. (2015). A XEN-like State Bridges Somatic Cells to Pluripotency during Chemical Reprogramming. *Cell* 163, 1678–1691. doi:10.1016/j.cell.2015.11.017
- Zhu, S., Rezvani, M., Harbell, J., Mattis, A. N., Wolfe, A. R., Benet, L. Z., et al. (2014). Mouse Liver Repopulation with Hepatocytes Generated from Human Fibroblasts. *Nature* 508, 93–97. doi:10.1038/nature13020

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.


Copyright © 2022 Yang, Xu, Gu, Nielsen, Chen, Guo, Tang and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

ARTICLE

DOI: 10.1038/s41467-017-00258-4

OPEN

Stochastic priming and spatial cues orchestrate heterogeneous clonal contribution to mouse pancreas organogenesis

Hjalte List Larsen¹, Laura Martín-Coll¹, Alexander Valentin Nielsen², Christopher V. E. Wright³, Ala Trusina², Yung Hae Kim¹ & Anne Grapin-Botton ¹

Spatiotemporal balancing of cellular proliferation and differentiation is crucial for postnatal tissue homeostasis and organogenesis. During embryonic development, pancreatic progenitors simultaneously proliferate and differentiate into the endocrine, ductal and acinar lineages. Using *in vivo* clonal analysis in the founder population of the pancreas here we reveal highly heterogeneous contribution of single progenitors to organ formation. While some progenitors are bona fide multipotent and contribute progeny to all major pancreatic cell lineages, we also identify numerous unipotent endocrine and ducto-endocrine bipotent clones. Single-cell transcriptional profiling at E9.5 reveals that endocrine-committed cells are molecularly distinct, whereas multipotent and bipotent progenitors do not exhibit different expression profiles. Clone size and composition support a probabilistic model of cell fate allocation and *in silico* simulations predict a transient wave of acinar differentiation around E11.5, while endocrine differentiation is proportionally decreased. Increased proliferative capacity of outer progenitors is further proposed to impact clonal expansion.

¹DanStem, University of Copenhagen, 3B Blegdamsvej, DK-2200 Copenhagen N, Denmark. ²Niels Bohr Institute, University of Copenhagen, 17 Blegdamsvej, DK-2200 Copenhagen N, Denmark. ³Department of Cell & Developmental Biology, Vanderbilt University, Nashville, TN 37232-0494, USA. Correspondence and requests for materials should be addressed to Y.H.K. (email: yung.kim@sund.ku.dk) or to A.G.-B. (email: anne.grapin-botton@sund.ku.dk)

Defining the rules governing embryonic organ development and postnatal tissue homeostasis is essential for understanding disease pathology and for the generation of functional cell types for regenerative medicine purposes. Seminal studies have demonstrated how rapidly proliferating postnatal tissues such as the skin and the intestine are homeostatically maintained by equipotent stem cells undergoing seemingly stochastic cell fate choices by neutral competition for limited niche signals^{1–4}. In contrast to postnatal tissue homeostasis, embryonic development of most organs occurs at a state of system disequilibrium, as a population of progenitors expands while simultaneously giving rise to differentiating progeny. Although optimality in the design of strategies ensuring rapid organ development has been proposed⁵, little is known regarding how global embryonic organogenesis is orchestrated when deconstructed into clonal units originating from single progenitors at the onset of organ bud formation. Studies of retinal development have provided compelling evidence for a stochastic process of cell fate choices using both *in vitro*⁶ and *in vivo* approaches⁷. However, a deterministic model of embryonic neocortical development was proposed⁸, based on the observation of similar behaviour of the two daughters of individual cells. These discrepancies in organ design emphasise the need for studies investigating individual cell progenies in other organ systems. Here we investigate how the allocation of endocrine and acinar fates is balanced with progenitor expansion from the beginning of pancreas formation using clonal analysis and single-cell molecular profiling.

Embryonic mouse pancreas development is initiated at around embryonic day (E)9.0 by the specification of pancreatic progenitors at the dorsal and ventral sides of the posterior foregut endoderm⁹. Though induced by different mechanisms, the two anlage are composed of expanding *Pdx1*⁺*Hnf1b*⁺*Sox9*⁺*Ptf1a*⁺ progenitors forming bud-like structures protruding into the surrounding mesenchyme¹⁰. A small number of *Neurog3*⁺ endocrine precursors giving rise to the endocrine lineage of the pancreas are also found in these early buds^{11, 12}. Morphogenetic processes occur concomitantly leading to the formation of lumens and their organisation into a plexus and subsequent tree-like branches^{13, 14}. While the distal tip-domain is comprised of *Ptf1a*⁺ unipotent acinar progenitors after E13.5^{15, 16}, the *Hnf1b*⁺ trunk domain is bipotent and gives rise to endocrine cells, as well as the ductal cells that will eventually line the epithelial network draining acinar digestive enzymes to the duodenum^{17–19}. Following specification towards the endocrine lineage, *Neurog3*⁺ endocrine precursors delaminate from the epithelial trunk domain to form immature islet clusters that will eventually mature into the endocrine Islets of Langerhans²⁰. Although population-based lineage tracing has demonstrated the multipotency of the early pancreatic progenitors by virtue of their capability to give rise to progeny in the three major pancreatic lineages^{12, 15–17, 21} (Fig. 1a), no study has addressed the clonal contribution of the proposed multipotent pancreatic progenitors (MPCs) to pancreas organogenesis. One previous clonal analysis indeed restricted its focus on the progeny of single endocrine precursors examining their postnatal expansion²². Recent studies have demonstrated that pancreatic trunk progenitors undergo stochastic priming towards the endocrine lineage at mid-gestation¹⁹, and thus we questioned whether there are subpopulations of pancreatic progenitors exhibiting restricted lineage potencies from the onset of embryonic pancreas development or whether progeny from equipotent progenitors undergo stochastic lineage commitment.

In this study, using clonal analysis of E9.5 pancreatic progenitors, when the pancreatic primordium has just been specified, we demonstrate that individual pancreatic progenitors contribute heterogeneously to pancreas organogenesis both in progeny size

and fate composition. While some progenitors are multipotent *per se*, giving rise to acinar, endocrine and ductal progeny, we also demonstrate the existence of bipotent ducto-endocrine and unipotent endocrine cells forming half of the primordium. This population represents cells at different stages of progression on the endocrine differentiation path, including proliferative endocrine-committed cells, and exhibits undetectable to low levels of PTF1A. In contrast, bipotent and multipotent clones do not exhibit different expression profiles, suggesting they are not molecularly distinct cell populations. We show that clonal expansion and fate heterogeneity are compatible with a simple model of probabilistic cell fate acquisition operating downstream of spatially controlled proliferative and fate-biasing patterning cues.

Results

Single E9.5 pancreatic cells produce heterogeneous progeny. To investigate how individual cells among the about 500 cells that have just been specified to found the pancreas contribute to organogenesis, we devised a lineage tracing strategy making use of the *Rosa26*^{CreER} driver (Fig. 1b). The ubiquitous activity of the *Rosa26* locus ensures *CreER* expression throughout the developing embryo and hence also enables non-biased labelling of pancreatic cells²³. We selected the *mT/mG*²⁴ reporter over other multicolour reporters to be able to mark the differentiation status of clonal progeny. This required the dosage of very low levels of the active tamoxifen metabolite 4-OH tamoxifen (4-OHTm) to reach labelling of only one cell per pancreatic primordium within the 24 h following injection²⁵. The labelling index of 11.8% (20 epithelial clones in 170 embryos) ensured a low risk of labelling two progenitors in the same pancreatic bud as of 1.4% (0.118×0.118). Whole-mount staining of E14.5 pancreata for endocrine (PAX6), progenitors lining the ducts (SOX9) and acinar (CPA1) markers enabled us to determine the fate of labelled GFP⁺ progeny at E14.5, a stage by which acinar cells are committed (Fig. 1c–h)^{12, 15–17}. We observed a large extent of clone size heterogeneity, ranging from single-cell clones to clones consisting of hundreds of cells (Fig. 1h). Single GFP⁺ cells belonged to the endocrine lineage based on immunoreactivity for PAX6, cell morphology and location outside the pancreatic epithelium in islet-like structures (Fig. 1d, f, g). These single cells are expected to result from labelling non-proliferative endocrine cells, their *Neurog3*-expressing precursors or pancreatic progenitors differentiating directly into the endocrine lineage without dividing. We also observed 2- and 3-cell endocrine clones, suggesting that a labelled endocrine-biased progenitor had undergone a single or two rounds of divisions. In line with the postulated existence of multipotent progenitors based on non-clonal analyses, multipotent clones of 40–250 cells were found, consisting of ductal, endocrine and acinar progeny (Fig. 1h). Moreover we did observe bipotent clones of 6–100 cells harbouring only ductal and endocrine progeny, indicating that not all E9.5 progenitors contribute to the acinar lineage during pancreas organogenesis. However, we did not observe unipotent acinar clones arising from E9.5 progenitors. While confirming the existence of MPCs at the single-cell level, our results reveal heterogeneity in potency and contribution to pancreas organogenesis from single pancreatic cells. Furthermore they uncover the existence of bipotent progenitors as early as E9.5 and that half of the cells in the early pancreatic anlage give rise to solely endocrine progeny, a surprising finding considering that the adult endocrine cells only account for about 1% of the adult organ²⁶.

Heterogeneous marker expression in E9.5 progenitors. Heterogeneity in the clonal progeny may be either due to an intrinsic lineage bias in sub-populations of E9.5 progenitors or

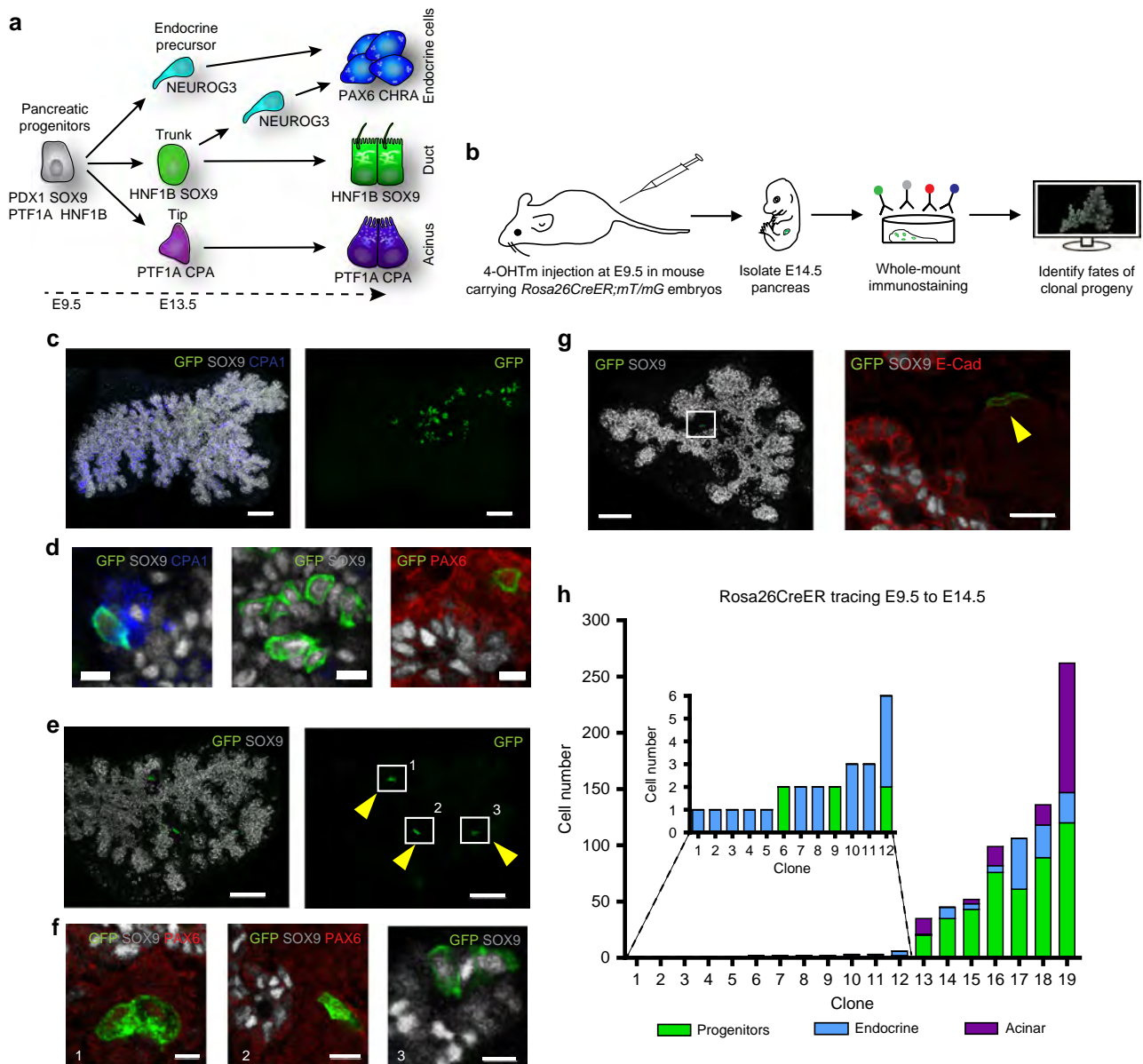


Fig. 1 *Rosa26^{CreER}*-mediated clonal analysis reveals heterogeneous contribution of E9.5 progenitors to pancreas organogenesis. **a** Schematic overview of lineage relationships based on previous global lineage tracing. **b** Schematic overview of strategy applied to identify fates of clonal progeny from E9.5 pancreatic progenitors. **c** 3D maximum intensity projection (MIP) of a large, multipotent clone containing acinar (CPA1), progenitors lining the ducts (SOX9) and endocrine progeny (PAX6, from other multipotent clone) **d**. Scale bars, 100 μ m **c** and 15 μ m **d**. **e** 3D MIP of a bipotent clone containing endocrine (insets 1 and 2) and ductal (inset 3) clonal progeny **f**. Scale bars, 100 μ m **e** and 10 μ m **f**. **g** 3D MIP and optical section (inset) showing a single-labelled endocrine cell after clonal analysis from E9.5 to E14.5. Note the localisation of the GFP⁺ cell in an E-CAD^{Low} endocrine cluster. Scale bars, 80 μ m and 15 μ m (inset). **h** Quantification of clone sizes and compositions following clonal analysis from E9.5 to E14.5 ($n = 170$ embryos, 20 with clones, 1 excluded due to poor immunocytochemistry-the images displayed show representative data from those)

due to the stochastic fate allocation of clonal progeny during progressive organisation and compartmentalisation of the pancreatic epithelium. To test the first hypothesis, we conducted single-cell qRT-PCR following FACS isolation of dorsal foregut progenitor cells at E9.5 (Fig. 2a). tSNE-mediated dimensionality reduction of single cells revealed the existence of three distinct populations (Fig. 2b). On the basis of the expression of known lineage markers, we classify these three clusters as pancreatic endocrine (*Neurog3*⁺ and *Glucagon*⁺), pancreatic progenitors (*Pdx1*⁺ and *Sox9*⁺) and duodenal progenitors (*Cdx2*⁺, absence of *Pdx1* and *Sox9*). Interestingly, cells characterised as belonging to the endocrine lineage organised on one projection axis forming a pseudo-temporal trajectory starting from *Neurog3*⁺ endocrine

precursors and progressing with the expression of markers associated with progressive endocrine maturation (Fig. 2c). These molecularly distinct cells are expected to contribute to the non-proliferative endocrine-committed cells observed in the lineage tracing. When focusing the dimensionality reduction on *Pdx1*⁺ pancreatic progenitors only, we observed marked heterogeneities in expression of various pancreas-associated transcription factors. Since at E14.5 *Ptf1a* marks acinar cells at the tip while *Nkx6-1*^{16, 27}, *Hnf1b*¹⁷ and *Hes1*¹⁸ mark bipotent progenitors in the trunk, we investigated whether cells expressing these markers at E9.5 already had specific molecular signatures suggestive of emerging tip and trunk fates. However, in spite of heterogeneous expression of these markers, no global gene

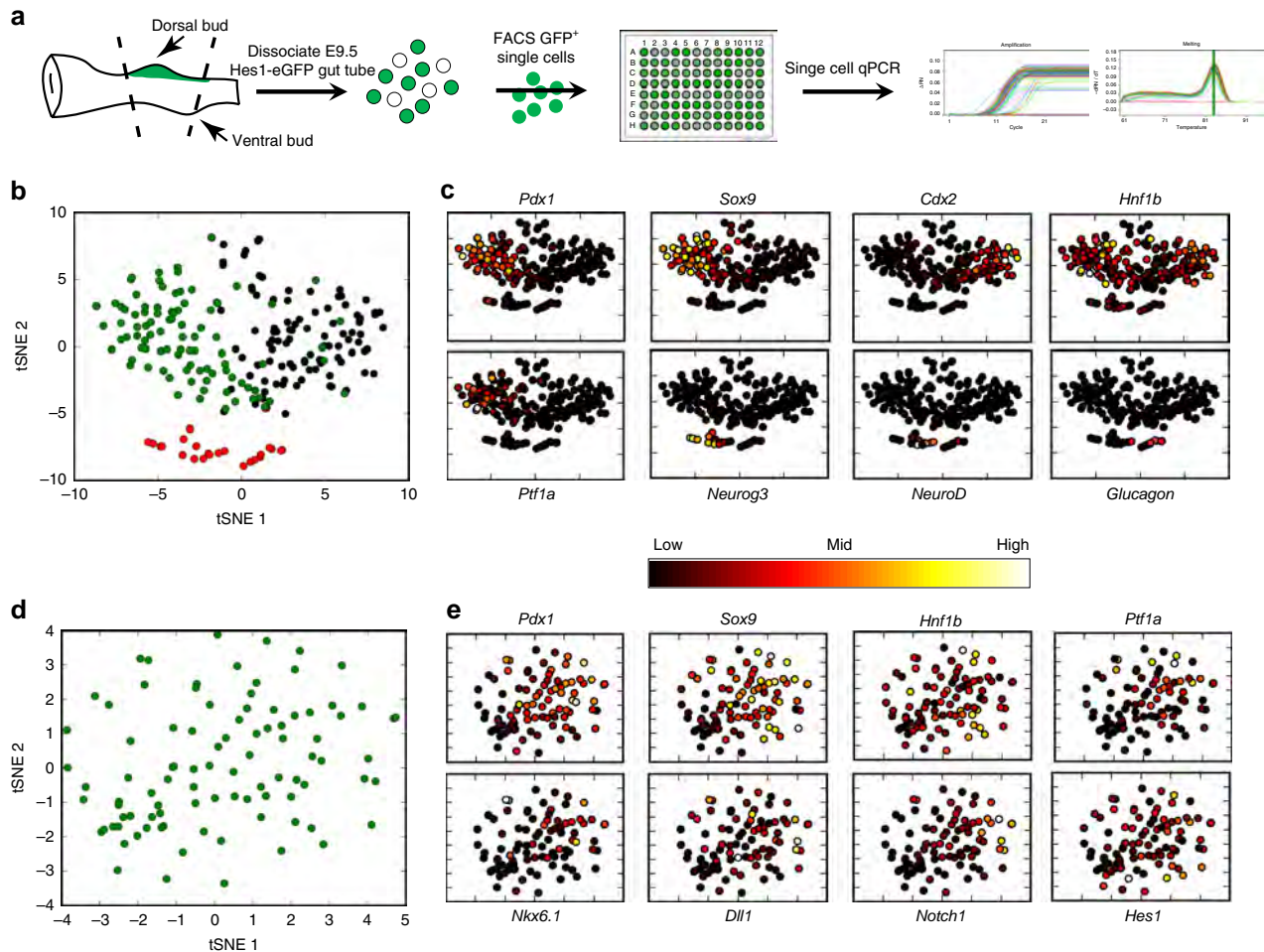


Fig. 2 Single-cell qRT-PCR of E9.5 pancreatic progenitors reveals endocrine differentiation progression but no apparent tip-trunk segregation signature. **a** Schematic outline of the strategy used to isolate single pancreatic progenitor cells for single-cell qRT-PCR analysis using the Hes1-eGFP strain⁴³. **b** tSNE dimensionality reduction of all single cells passing initial quality controls. Three cell clusters are displayed in *green*, *black* and *red*, corresponding to pancreatic progenitors, duodenal progenitors and pancreatic endocrine cells, respectively. **c** *Hnf1b* is expressed in both pancreatic and duodenal progenitors, whereas *Ptfla* expression is exclusively detected in pancreatic progenitors albeit in heterogeneous pattern. Cells in the endocrine population cluster organise on a pseudo-temporal differentiation pathway starting with *Neurog3*⁺ endocrine precursors at the left, progressing into glucagon-expressing terminally differentiated endocrine cells at the right. **d** tSNE dimensionality reduction of *Pdx1*⁺ pancreatic progenitors only. **e** Despite heterogeneous expression of transcription factors and signalling components involved in subsequent tip-trunk segregation, expression of these markers seems uncorrelated and does not partition *Pdx1*⁺ progenitors into tip- and trunk-biased clusters

signature was associated with *Nkx6-1*, *Hnf1b* or *Hes1*, and these three markers showed no cross-correlation (Fig. 2d, e; Supplementary Figs. 1–4). Although *Ptfla* expression did not correlate strongly with specific single markers, *Ptfla*⁺-cells clustered after tSNE-mediated dimensionality reduction, suggesting that they are more similar to each other than to other progenitor cells. This molecular analysis uncovers that cells committed to endocrine differentiation can be molecularly identified, whereas subpopulations of multipotent or bipotent progenitors identified by clonal analysis cannot be molecularly predicted with this set of markers.

Spatial patterns of progenitor marker expression. To further assess heterogeneity in markers at the protein level, we used whole-mount immunostaining of E9.5 gut tubes and quantitative image analysis (Fig. 3a, b). We observed heterogeneous expression levels of HES1, SOX9 and PTF1A, whereas HNF1B and PDX1 were more homogeneously expressed among progenitors (Fig. 3c, d). Using 3D Voronoi-Delaunay triangulation (Fig. 3a)

and measurements of the correlation in expression levels between neighbouring cells, we observed that HNF1B is expressed at higher levels towards the more posterior side of the pancreatic bud and in the gut tube and that PTF1A expressing cells appear clustered at a medial-bilateral location in E9.5 dorsal buds. Other transcription factors did not show any regionalisation of expression levels (Fig. 3d). These results demonstrate that the transcriptional profiles observed by single-cell qRT-PCR are translated into similar global profiles at the protein level, and that the levels of PTF1A and HNF1B display regionalised patterns in the E9.5 bud.

Distinct clonal progeny of *Hnf1b*- and *Ptfla*-expressing cells. The differential expression of PTF1A and HNF1B at E9.5 and the subsequent segregation of these markers to the tip and trunk domain, respectively, led us to investigate whether single progenitors expressing *Ptfla* or *Hnf1b* at E9.5 contribute differential progeny by clonal analysis using *Ptfla*^{CreER}- and *Hnf1b*^{CreER} drivers (Fig. 4a). The *Hnf1b*^{CreER}-driver (labelling

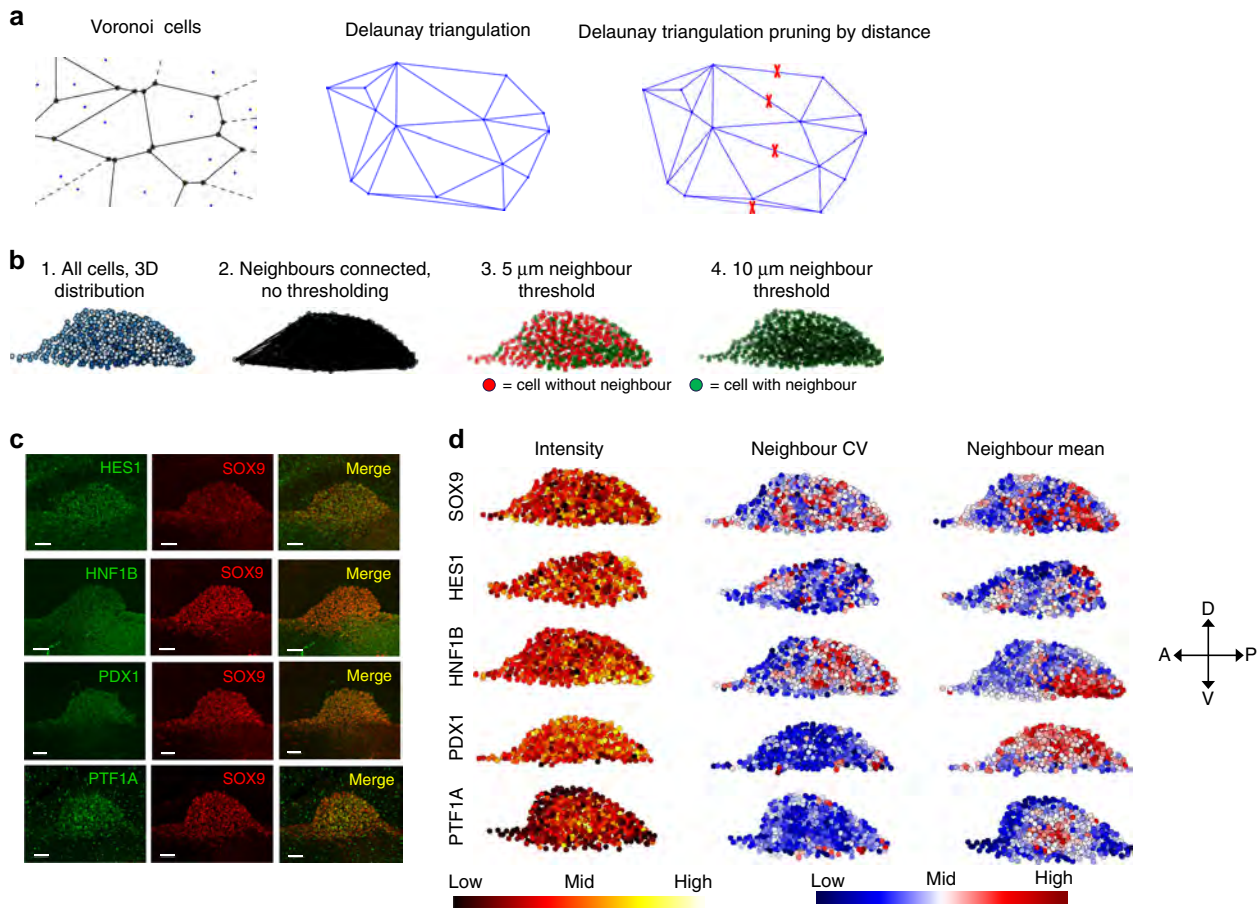


Fig. 3 E9.5 pancreatic progenitors display heterogeneous expression of pancreas-associated transcription factors and distinct patterns of neighbour cell expression correlation. **a** Schematic of Voronoi-Delaunay triangulation implemented to identify neighbour cells in 3D. Distance-based pruning was implemented in order to identify biologically meaningful neighbours in 3D space. **b** Example of Voronoi-Delaunay triangulation-mediated neighbour identification in E9.5 pancreatic bud. Cells colour-coded according to z-position (1) are subjected to Voronoi-Delaunay triangulation, identifying all neighbour relations in 3D (2). Implementation of a distance threshold of 5 μm (3) or 10 μm (4) generates different number of cells with and without 3D neighbours (green and red cells, respectively). For all downstream analyses, 10 μm was chosen as neighbour distance threshold. **c** Example of transcription factor expression pattern in E9.5 pancreatic buds following whole-mount staining. 3D MIP is displayed. Scale bars, 30 μm . **d** 3D plots of staining intensity, neighbour coefficient of variation and neighbour mean intensities from immunostaining against the indicated transcription factors. Note the heterogeneous expression patterns of HES1, SOX9 and PTF1A and the regionalised expression of HNF1B (posterior) and PTF1A (lateral) ($n = 2$ for all, except $n = 8$ for SOX9). The images displayed show representative data from those)

index: 35 clones in 120 pancreata, 29%; probability of double labelling, 8%) resulted in a similar pattern of clones as observed with *Rosa26^{CreER}*, that is unipotent endocrine, bipotent ducto-endocrine as well as multipotent clones (Fig. 4b, c and f). However, the endocrine-committed clones were less frequent, constituting 33% instead of 50% of the total clone repertoire, likely due to the fact that unlike *Rosa26*, *Hnf1b* is not expressed in mature endocrine cells²⁸. In addition, we detect HNF1B immunoreactivity in $67.7 \pm 3.8\%$ of the NEUROG3-expressing endocrine precursors at this stage, while *Rosa26* is expected to be expressed in all (Supplementary Fig. 5). A similar frequency of endocrine-committed precursors was observed when tracing *Hnf1b^{CreER}*-labelled cells from E9.5 to E10.5 (Supplementary Fig. 6). Interestingly, we observed a clone consisting solely of 6 endocrine cells. Combined with the two 3-cell clones seen in the *Rosa26* tracing, this suggests that some endocrine-biased progenitors can undergo multiple rounds of divisions (Fig. 4b, f; clone # 12), in line with the recent observation that cells with low levels of *Neurog3* transcription can proliferate²⁹. The low-differentiation rate towards the endocrine lineage ($p = 0.12 \pm 0.007$, Supplementary Fig. 7) makes independent probabilistic entry of progeny into the endocrine lineage highly

unlikely, suggesting that the E9.5 pancreatic bud contains progenitors biased towards multigenerational endocrine specification. On the other hand, we never observed unipotent acinar clones arising from E9.5 progenitors. Furthermore, acinar-containing clones always contained ductal and endocrine progeny, suggesting that acinar-lineage allocation has not yet occurred in any cell at E9.5. The proportion of multipotent clones was similar to what was observed using the *Rosa26^{CreER}* driver after correction for the absence of labelling of mature endocrine cells by *Hnf1b^{CreER}*. Though only five of the clones were found in the ventral pancreas, which is smaller than the dorsal, they were bi- or multipotent but not endocrine committed, possibly due to a delay in endocrine program onset in the ventral pancreas. This suggests that the assumed labelling of progenitors with high-expression levels of *Hnf1b* expression at E9.5 does not bias lineage contribution to the trunk domain (Fig. 4f). Similarly, *Ptf1a^{CreER}*-based lineage tracing did not bias progenitors towards the acinar lineage, either (Fig. 4d, e and g). However, cells traced by *Ptf1a^{CreER}* (labelling index: 13 clones in 30 dorsal pancreata, 44%; probability of double labelling, 19%) did not form endocrine-only clones, unlike what was seen with *Rosa26^{CreER}*- and *Hnf1b^{CreER}*-drivers. This suggests that cells exhibiting high

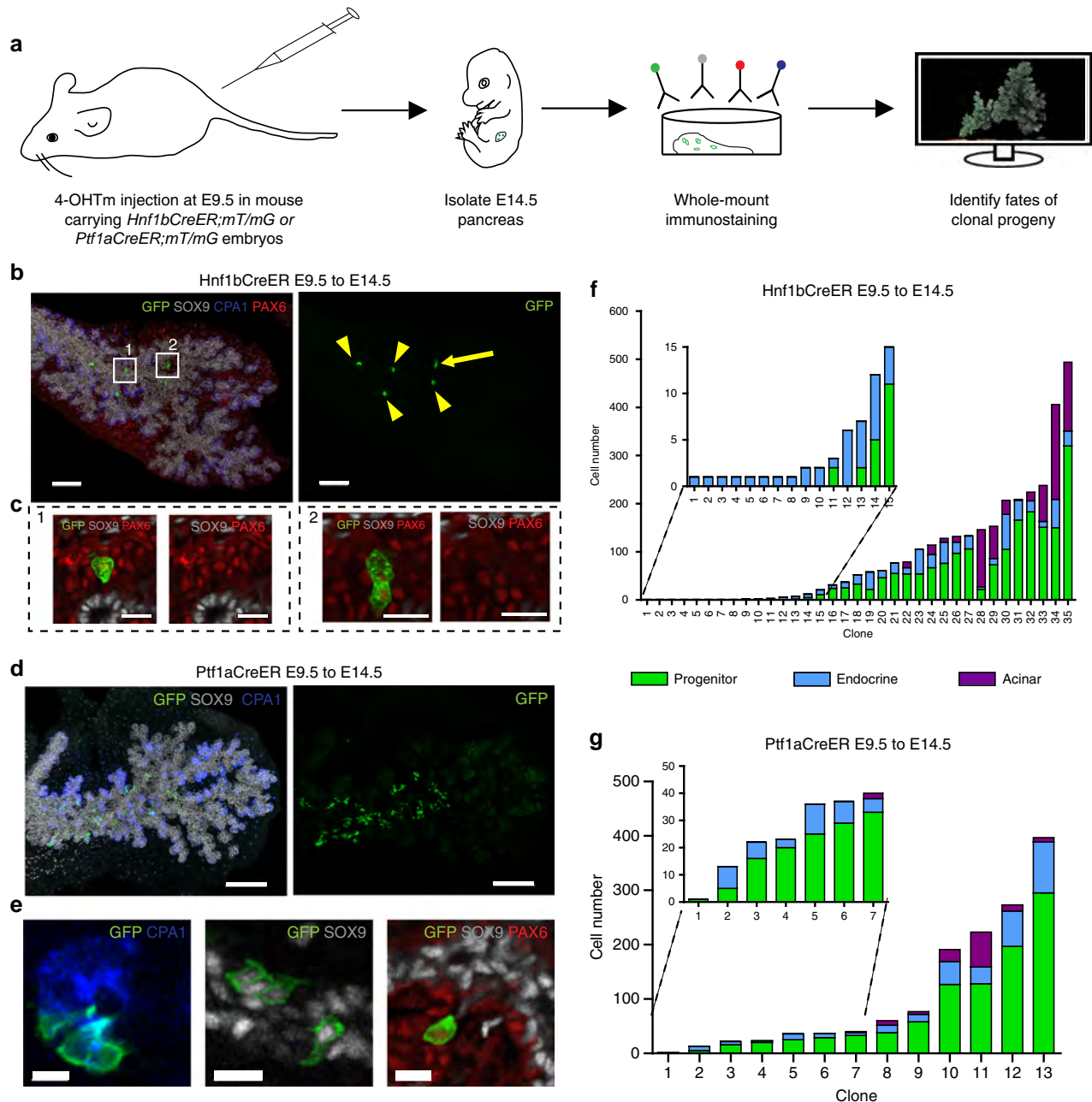


Fig. 4 *Hnf1b^{CreER}*- and *Ptf1a^{CreER}*-mediated clonal analysis confirms the existence of ducto-endocrine bipotent clones and the absence of unipotent acinar progenitors at E9.5. **a** Schematic overview of strategy applied to identify fates of clonal progeny from E9.5 pancreatic progenitors. **b** 3D MIP and high-magnification optical sections (**c**) of clonal progeny from a unipotent endocrine clone containing six labelled endocrine PAX6⁺ endocrine cells generated by *Hnf1b^{CreER}*-mediated clonal analysis. Arrow indicates two juxtaposed GFP⁺ cells, whereas arrowheads indicate single GFP-labelled cells. Scale bars, 100 μ m **b** and 15 μ m **c**. **d** 3D MIP and high-magnification optical sections (**e**) of a multipotent clone derived from *Ptf1a^{CreER}*-mediated lineage tracing. Scale bars, 150 μ m **d** and 15 μ m **e**. **f**, **g** Quantification of clone sizes and fate compositions following *Hnf1b^{CreER}*- and *Ptf1a^{CreER}*-mediated clonal analysis, respectively ($n = 120$ embryos for *Hnf1b^{CreER}* and 34 embryos for *Ptf1a^{CreER}*- the images displayed show representative data from those). Note that clone no. 1 contains only one cell and is only SOX9⁺. Most clones were found in the dorsal pancreas, except clones no. 14, 17, 18, 26, 33 in **f** and clones no. 3, 8, 9 in **g** which were found in the ventral pancreas

Ptf1a expression at around E9.5 do not immediately form endocrine cells, unlike progenitors traced by *Rosa26^{CreER}* and *Hnf1b^{CreER}*, though they retain endocrine differentiation capacity as these cells give rise to endocrine-containing clones later in their clonal evolution. This hypothesis is supported by *Ptf1a* anti-correlation with early markers of endocrine differentiation such as *Mfng* and *Neurog3* in our single-cell qRT-PCR analysis at E9.5 (Supplementary Fig. 3).

A probabilistic model of progenitor progeny fate allocation. The apparent lack of tip-trunk biased progenitors suggested by both single-cell analysis and tracing at E9.5 led us to investigate whether a model of probabilistic cell-fate choices could recapitulate the in vivo clonal distribution data. To this end, we constructed a mathematical model of in silico clonal growth by simulating cell divisions over a period spanning the in vivo clonal tracing. Every time a cell gave rise to progeny through cell division, clonal progeny were fate-allocated with a probability of

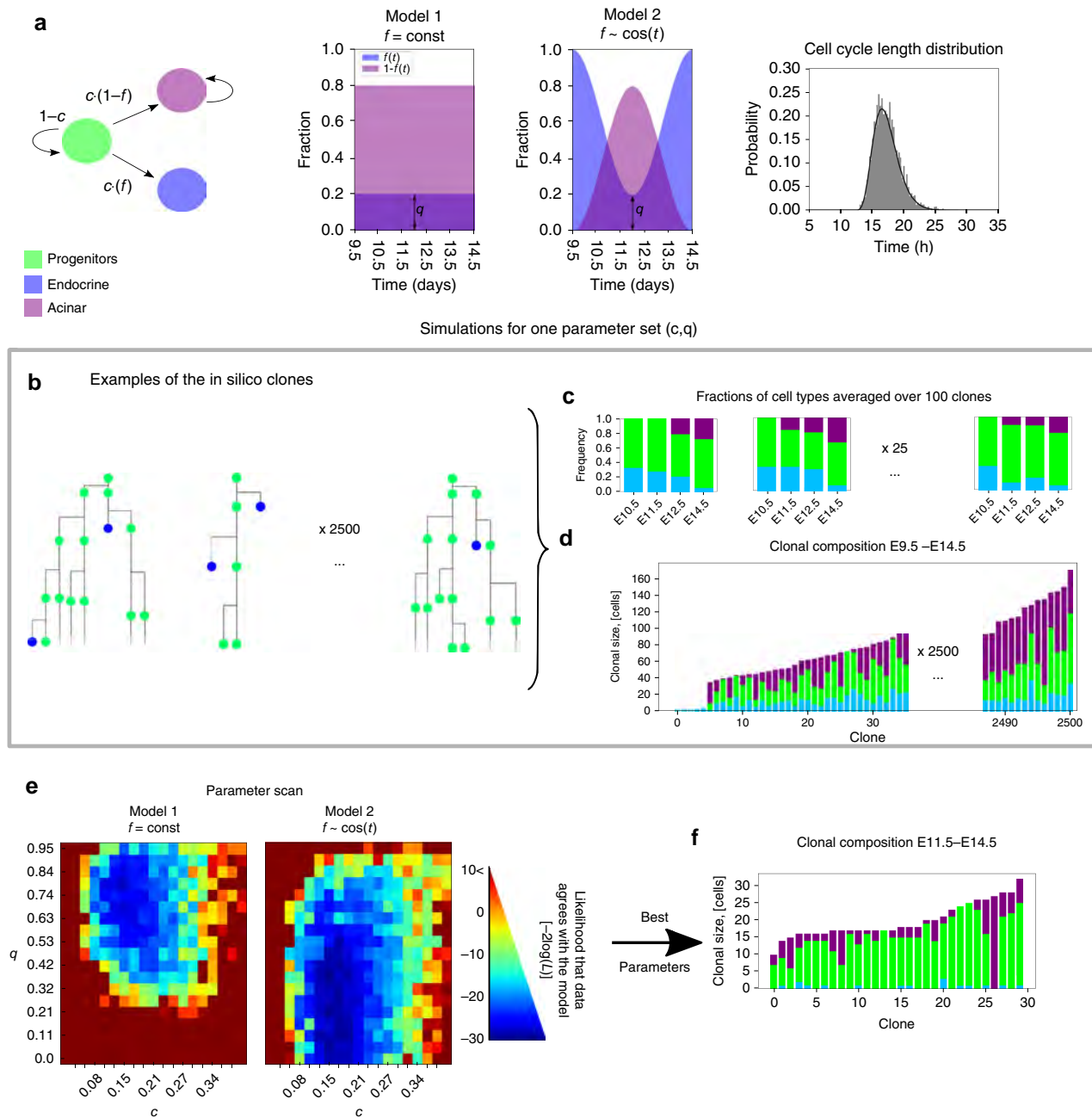


Fig. 5 Stochastic model of clonal expansion. **a** Probabilistic transitions among three states are presented in a state diagram: at the division progenitors (green) differentiate with probability c and maintain progenitor state with probability $1 - c$. When differentiating, they become endocrine (blue) with probability f and acinar (purple) with probability $1 - f$. We compare two models: in Model 1 the probability f is constant between E9.5 and E14.5, and in Model 2 f is time-dependent and has a minimum at around E12. The height of the minimum is characterised by parameter q . In Model 1 $f = g$. See Methods for the exact functional form of f for Model 2. While acinar cells continue replicating, endocrine cells are assumed not to replicate. For all replicating cells the cell-cycle lengths are drawn from the gamma distribution from¹⁹ (right panel). For every parameter set both models were simulated 2500 times. **b** Examples of the in silico clonal lineages. **c** Fraction of cell types from individual clones at E9.5, 11.5, 12.5 and 14.5 (corresponds to experimental data in Supplementary Fig. 7d). **d** Clonal composition at E14.5 (corresponds to Figs. 1h and 4f). **e** For each of the parameter sets c, q we estimate the likelihood of the model fitting the data (See Methods and Supplementary Figs. 8–10). The results of the parameter scan are quantified by the log-likelihood, $-2\log(L)$. Parameter scans show that Model 2, where the differentiation of acinar cells and endocrine cells change with time, is more likely. Using Akaike Information Criteria score, AIC, we find that Model 2 is better at describing the data ($AIC_1 = -14.5$ and $AIC_2 = -22.1$). The probability that the two models are equally good is $p = 0.02$. **f** The model predicts that if clonal analyses are started at E11.5–12 instead of E9.5, it becomes more likely to observe lineages fully committing to the acinar fate

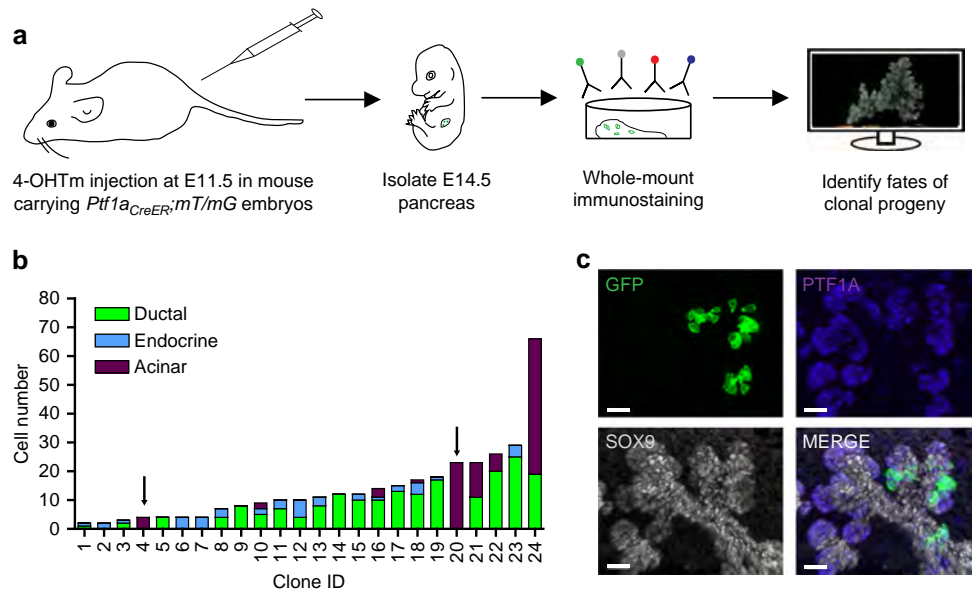


Fig. 6 *Ptf1a*^{CreER}-mediated clonal analysis identifies unipotent acinar progenitors at E11.5. **a** Schematic overview of strategy applied to identify fates of clonal progeny from E11.5 pancreatic progenitors. **b** Quantification of clone sizes and fate compositions following *Ptf1a*^{CreER}-mediated clonal analysis ($n = 24$ clones in 70 embryos analysed). Arrows indicate two acinar unipotent acinar clones, as predicted by our mathematical model. **c** A representative 3D MIP of one of the two acinar unipotent clones. Most clones were found in the dorsal pancreas, except clones no. 2, 9 and 16 which were found in the ventral pancreas. Scale bars, 20 μ m

differentiating c and a probability of becoming endocrine cf or acinar $c(1-f)$ (Fig. 5a). Here f is a bias towards endocrine fate; $f = 0.5$ means that cells are equally likely to be allocated exocrine or endocrine fates. Simulated cell cycle lengths were randomly drawn from a gamma-distribution based on our measurements in vitro¹⁹ and those of Bankaitis et al.³⁰ in vivo. Progeny committed to the endocrine lineage were approximated to be non-proliferative while acinar cells proliferated. For simplicity the acinar proliferation rate was approximated to be similar to progenitors, which is a small underestimation¹⁵. In total 2500 clones were simulated spanning a parameter space of probabilities for both c and f (Fig. 5b, d). Two models were compared, one with fixed probability of becoming endocrine rather than exocrine $f = q$, or one where f varied over time with a minimum q around E12 (Fig. 5a). The minimum around E11.75–E12 is suggested by the observation that the number of NEUROG3-expressing cells has a minimum at this time point³¹.

To quantitatively compare simulation results with the data, we focused on two types of datasets. First, histograms in Figs. 1h and 4f contain information about the clonal variance in fractions of acinar and endocrine cells at one time point E14.5. For simplicity we focused on the variance in fractions of acinar cells and to increase sample size combined the datasets in Figs. 1h and 4f into one. Second, the staining of pancreata at four time points in Supplementary Fig. 7d does not contain information about the clonal variance but represents the typical cell fractions. To compare our models with the first data set we recorded the acinar fraction from simulated lineages with at least one acinar cell for each parameter set (Supplementary Fig. 8a–e). We estimated the underlying probability density function (PDF), shown on top of the histogram in Supplementary Fig. 8e by Kernel Density estimation with bandwidth 0.5 (see Supplementary Fig. 9 and methods for details).

To compare our models with the second data set, we grouped together 100 clones to approximate the data from stained pancreata in Supplementary Fig. 7d. Here we used both acinar and endocrine fractions at E10.5, 11.5, 12.5 and 14.5 to estimate PDFs as described above. This allowed us to estimate the likelihood that the experimental data points came from the PDF

derived from the simulations. In other words we estimated how likely it is for simulations to produce exactly those fractions observed in vivo. The likelihood that both datasets agree with the model was a product of each of the two likelihoods (Methods, Model Implementation). Spanning a parameter space for c and q (Supplementary Fig. 10), we observed that both the model 1 with fixed endocrine/acinar probability and the model 2 displaying temporal variations in this ratio had a parameter space of good likelihood for c and q (Fig. 5e). However, the model 2 with a time-variable probability of becoming acinar peaking at around E12 was superior at describing the data according to the Akaike Information Criteria (AIC, see Methods). The shape of the optimal parameter space is also in support of model 2: once the probability to become acinar is set to peak around E12 (model 2), the performance of the model becomes less constrained by parameter q . The statistical approach used allows us to identify the best model, but a combination of limited amount of biological data and high stochasticity prevents us from statistically testing how well each model matches the data. Taken together, our mathematical modelling suggests that the clonal analysis data are compatible with a model of probabilistic cell fate choices and predicts that when the probability of becoming endocrine is low at around E12, the progenitors most efficiently commit to the acinar lineage at this time point.

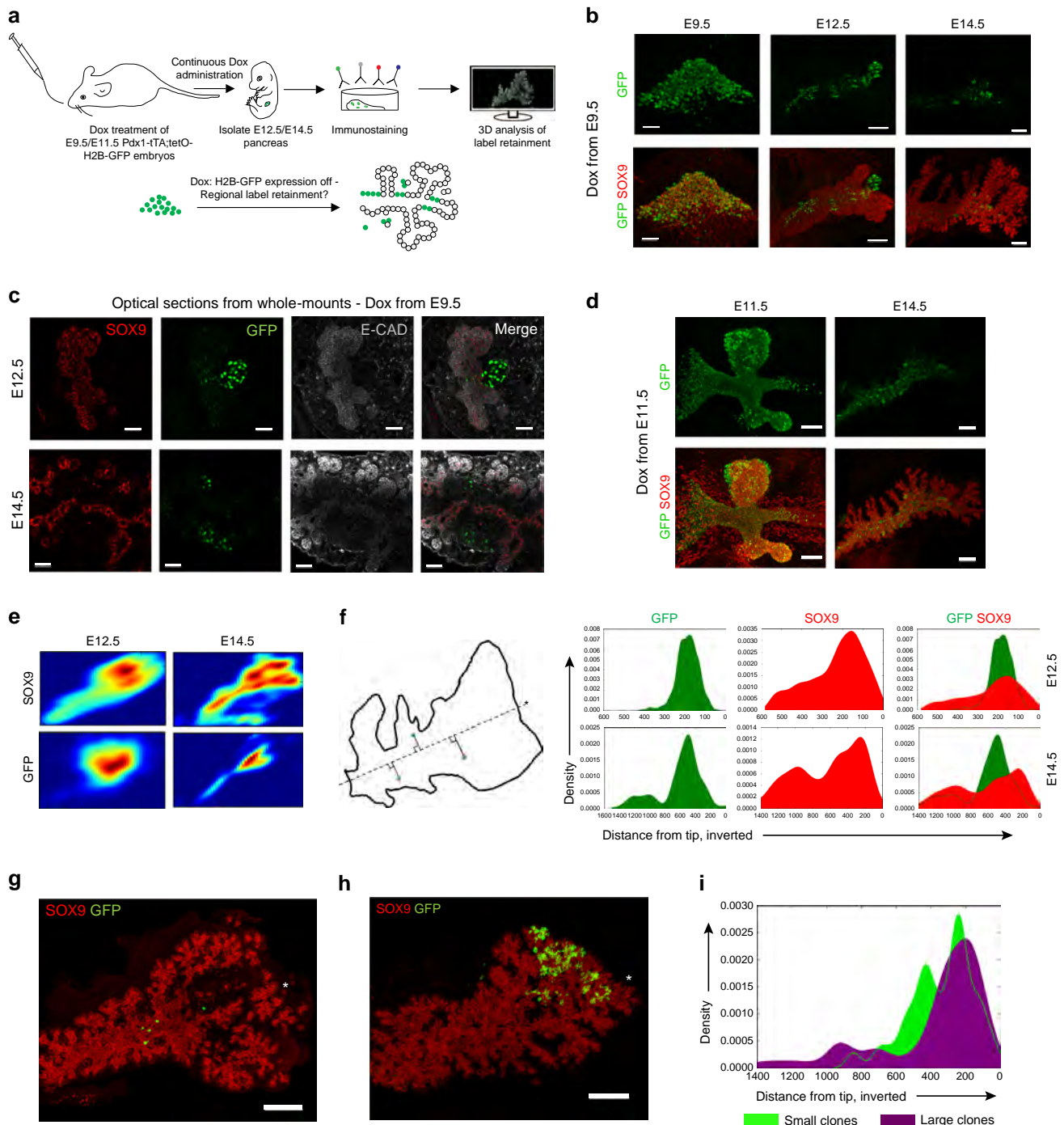
Acinar-committed cells are detected from E11.5 to 12.

According to the model prediction, acinar-committed cells should be undetectable at E9.5, as we have seen, but should be easily identified by clonal lineage tracing from E11.5 (Fig. 5f). Previous non-clonal tracing with *Ptf1a*^{CreER} suggested that all cells expressing *Ptf1a* were acinar-committed at E14.5–E15, whereas some were still multipotent at earlier time points¹⁵. However, previous observations did not address whether some cells may be acinar-committed earlier. Tracing using *Ptf1a*^{CreER};mT/mG mice injected with 4-OHTm at E11.5 revealed bipotent and tripotent clones, as at earlier time points and also showed that 8% of the PTF1A-traced cells were already acinar-committed (Fig. 6),

reinforcing the notion of heterogeneity in progenitor behaviours at the clonal level during pancreas organogenesis.

Spatial differences in proliferation impacts clonal growth. In addition to the compatibility of the in vivo clonal heterogeneity with a probabilistic model of cell cycle progression and cell fate allocation, we questioned whether spatial patterns of differential proliferation rates might also impact clone size. To interrogate the proliferative capacity of pancreatic epithelial subdomains we performed a label dilution experiment using *Pdx1-tTA;tetO-H2B-GFP* embryos (Fig. 7a). These embryos display ubiquitous H2B-GFP expression in *Pdx1*⁺ progenitors, however upon doxycycline (Dox) administration, H2B-GFP expression is suppressed and

will be linearly diluted by equal partition to daughter cells upon cell division³². By tracing the extent of label dilution from E9.5 to E12.5 and E14.5, we observed label retention in SOX9^{NeE}-CAD^{Low} cell clusters corresponding to non-proliferative endocrine cells derived early after suppression of H2B-GFP expression (Fig. 7b, c). Label dilution following continuous Dox administration from E9.5 was evident in SOX9⁺ progenitors compared to endocrine cells at E12.5, and this was even more apparent at E14.5 (Fig. 7b, c). At E14.5, cells located in the central portion of the SOX9⁺ epithelium still displayed retention of H2B-GFP signal, whereas SOX9⁺ progenitors in lateral branches and the more distal portion of the epithelium displayed label dilution. This was also apparent when administering Dox at E11.5 and tracing to E14.5 (Fig. 7d). These results suggest that pancreatic



SOX9⁺ ductal progenitors undergo preferential proliferation at the peripheral epithelial domains. Such preferential label retention within the central epithelial domain was additionally confirmed by plotting the 2D kernel density estimation of SOX9⁺ and GFP-retaining SOX9⁺ progenitors (Fig. 7e). To investigate whether the size of clones from our lineage tracing correlated with the spatial location of H2B-GFP retaining SOX9⁺ progenitors, we sought to map the spatial location of clones onto the domains of differential label retention. Because of the non-stereotypic macroscopic anatomy of the pancreata between embryos, we turned to a simplistic model of spatial mapping, where we projected the location of a cell or group of cells on an axis extending from the tip of the dorsal pancreas to the duodenal root of the dorsal pancreas. This method confirms the enrichment of H2B-GFP-retaining SOX9⁺ cells in a central domain of the pancreas epithelium at E14.5 (Fig. 7f) and indicates that the largest *Hnf1b*^{CreER}-derived clones tend to map to the tip (Fig. 7g–i). These results suggest that the spatial location impacts the proliferation of clonal progeny by dispersal to spatial niches with distinct proliferative capacity.

Discussion

In this study we aimed at uncovering whether the roughly 500 cells that found the mouse pancreas contribute homogeneously to the size of the final organ and to its different functional cell types. The multipotent state of the early pancreatic progenitor population has been inferred from population-based lineage tracing studies, masking potential heterogeneity in single-progenitor contribution to organ formation^{12, 15, 21}. We tested whether there are subpopulations with biases in proliferation or differentiation capacity, and whether they can be predicted by their molecular expression profile or by their initial location in the primordium.

We find that single E9.5 pancreatic cells exhibit heterogeneous contribution to organ formation, as we identify unipotent endocrine, bipotent ducto-endocrine and multipotent clones by lineage tracing at clonal density (Fig. 8). Among these categories, only the unipotent endocrine-committed cells can be predicted by single-cell molecular profiling at E9.5. These cells account for 50% of founder cells and encompass the already differentiated endocrine cells and *Neurog3*-expressing endocrine progenitors each accounting for about 12% of pancreatic cells at this stage (Supplementary Fig. 7). In addition, early endocrinogenesis encompasses other endocrine-biased cells, some of which may be replicative, possibly expressing low levels of *Neurog3*²⁹. The size of this population is estimated to about 25% of all pancreatic cells based on both *Rosa26*^{CreER} and *Hnf1b*^{CreER} lineage tracing.

Although we did not identify any positive predictor for such endocrine-biased progenitors, *Ptf1a* is a negative correlator based on the rarity of unipotent endocrine clones from *Ptf1a*^{CreER}-based lineage tracing, as well as the negative correlation with known endocrine specifiers from single-cell qRT-PCR (Supplementary Fig. 3). This is in agreement with the previous observation that early endocrine cells can form in the absence of *Ptf1a*^{33, 34}. The fact that 50% of the cells in the emerging pancreatic primordium are biased to the endocrine lineage is surprising, since the endocrine cells make only 1–2% of the adult pancreas²⁶. As the largely non-proliferative nature of endocrine-biased cells extends the time required to generate an organ of proper size, the generation of such a high fraction of endocrine cells at early stages of organogenesis contradicts expectations of optimal design theories⁵. These cells may thus carry important functions for the development of the mouse pancreas, perhaps by producing growth-stimulating components.

Despite heterogeneous and spatial differences in expression of pancreatic progenitor-associated transcription factors within the E9.5 bud, the bipotent ducto-endocrine and multipotent progenitors cannot be discriminated by single-cell qRT-PCR using our selected gene targets. Investigating more targets, protein expression or their modifications may however uncover subpopulations. Nevertheless, the heterogeneity in clone sizes and differentiation is compatible with a stochastic model of cell-fate allocation during clonal history. Comparison of several models shows that the model that best fits the data is one where cells have a probability of differentiation and where differentiation bias towards endocrine over acinar fates changes over time between E9.5 and 14.5. This would imply a double molecular gate, one controlled by the Notch pathway that controls differentiation, and a switch controlled by an unknown molecular cue that selects between endocrine and acinar fates. There is ample data supporting that Notch controls the differentiation of both acinar and endocrine lineages^{10, 35–38}. In the model displaying optimal fit with our experimental data, the progenitors are predicted to have a low probability of becoming endocrine at around E11.5–E12, as supported by the progressive decrease and subsequent reappearance of *NEUROG3* cells at this time point³¹. The model predicts that this corresponds to a wave of acinar cell commitment centred at around E11.5–E12 that we can experimentally capture (Fig. 6).

We also report spatial heterogeneity in progenitor proliferation which may underlie the observation of progenitors that divide only once to extreme progenies of hundreds of cells in 5 days. Recently it was demonstrated that the progeny of dividing E10.5 pancreatic progenitors in the central area of the pancreas tends to remain central but that this rule is not strict³⁹. The combined

Fig. 7 The pancreatic epithelium displays regional differential proliferation explaining impacting clone size. **a** Schematic overview of strategy implemented to identify spatial differences in proliferative capacities. E9.5 oral gavage and subsequent continuous administration of doxycycline (Dox) prevents expression of H2B-GFP in *Pdx1*;–*tTA*;*tetO-H2B-GFP* embryos, enabling proliferation-induced label dilution in pancreatic progenitors. **b** 3D MIP of whole-mount immunostainings of dorsal pancreata at various stages following Dox administration at E9.5. Note the gradual decrease in GFP signal in SOX9⁺ cells and the presence of strongly label-retaining endocrine clusters and low-retaining central progenitors, as well as the absence of label retention in the distal epithelium and in lateral branches by E14.5 ($n = 3$ at E9.5 and $n = 4$ each at E12.5 and E14.5). Representative images were extracted from those. Scale bars, 30 μ m (E9.5), 80 μ m (E12.5) and 150 μ m (E14.5). **c** Optical sections of E12.5 (*top*) and E14.5 (*bottom*) dorsal pancreata following Dox administration at E9.5. E-CAD^{low} endocrine clusters display strong label retention, whereas label-dilution is more pronounced in the proliferative SOX9⁺ progenitors. Distal lateral branches at E14.5 display complete absence of H2B-GFP retention. Scale bars, 50 μ m (E12.5) and 30 μ m (E14.5). **d** Following E11.5 Dox administration, the central portions of the E14.5 pancreas retains H2B-GFP signal, whereas lateral branches exhibit label dilution ($n = 3$ at E11.5 and $n = 1$ at E14.5, from which representative images were extracted). Scale bars, 70 μ m (E11.5) and 150 μ m (E14.5). **e** Kernel density estimation of SOX9⁺ progenitors and the density of the top 10% highest GFP-retaining SOX9⁺ cells. Note the central location of GFP-retaining cells. **f** One-dimensional projection of SOX9⁺ progenitors and the top 10% of GFP-retaining cells onto a diagonal line running along the length axis of the dorsal pancreas demonstrate enrichment of GFP-signal in distinct domains of the pancreatic epithelium. **g, h** 3D MIP showing the spatial distribution of clonal progeny in a small clone in the central, proximal epithelium and a large distal clone, respectively. Scale bars, 150 μ m. **i** Comparison of spatial distribution of smallest and the half of largest clones from *Hnf1b*^{CreER}-mediated E9.5–E14.5 clonal analysis ($n = 12$ clones in total)

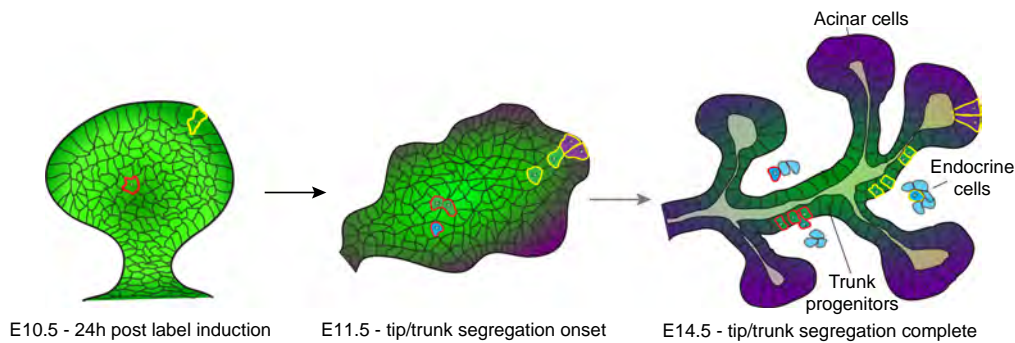


Fig. 8 Proposed model of orchestration of pancreas organogenesis by heterogeneous clones. Although initially possessing the same intrinsic potency, the spatial position of induced clones in the pancreatic bud influences expressed *in vivo* potency by dispersing clonal progeny to different niches. While peripherally labelled cells (yellow outline) will be exposed to acinar-inducing cues concomitant with generation of trunk-fated progeny during branching morphogenesis, centrally labelled cells (red outline) are less likely to experience acinar-inducing signals and thus produce ducto-endocrine bipotent clones. Stochastic priming of centrally located progenitors towards the endocrine lineage might further generate heterogeneity in clonal contribution to the ductal and endocrine lineage

effect of probabilistic cell fate choices operating downstream of spatially biased progenitor proliferation and differentiation thus ultimately determines the contribution of progeny from proliferating progenitors to pancreas organogenesis.

Similar to our observations, differential potency and lineage contribution of progenitors expressing early organ markers have recently been demonstrated during heart development⁴⁰. Our findings might facilitate the identification of niche-derived signals supporting *in vitro* generation of specific pancreatic cell types for regenerative medicine purposes and help elucidate the rules governing embryonic organogenesis by the concerted spatio-temporal orchestration of clones with variable contributions to organ formation.

Methods

Mice. Mice (*Mus musculus*) of mixed background were housed at the University of Copenhagen. All experiments were performed according to ethical guidelines approved by the Danish Animal Experiments Inspectorate (Dyreforsøgstilsynet). The following genetically modified mouse lines were used: *Pdx1-tTA*⁴¹, *tetO-HIST1HB*/*GFP(tetO-H2B-GFP)*⁴², *Hnf1b**CreER* (*Hnf1b*^{*CreER*})¹⁷, *Gt* (*ROSA*)26Sortm4(*ACTB-tdTomato*,-EGFP)*Luo*/*J* (*mT/mG*)²⁴, *Ptf1a**CreERTM* (*Ptf1a*^{*CreER*})¹⁶, *Tg*(*Hes1*-EGFP)^{Htri} (*Hes1*-eGFP)⁴³, *Gt*(*ROSA*)26Sor^{tm1(cre/ERT2)}*Tyj*/*J* (*Rosa26*^{*CreER*})²³. The data were collected on male and female embryos.

Whole-mount immunohistochemistry. Embryonic gut tubes or isolated pancreata were fixed in 4% paraformaldehyde (PFA) for 10–30 min on ice depending on tissue size. After washing in PBS and dehydration in methanol (MeOH), fixed tissue was stored in 100% MeOH at –20 °C. Rehydrated samples were transferred to PBS + 0.5% Triton X-100 (PBST). Samples were blocked overnight at 4 °C in PBST + 1% bovine serum albumin (BSA) or in blocking solution from the Mouse-on-Mouse (MOM) detection kit (Vector laboratories) if using mouse primary antibodies (For details of antibodies and concentrations use, please see Supplementary Table 1). Primary antibodies were incubated in PBST + 1% BSA or MOM diluent for 48 h at 4 °C. Samples were washed all day in PBST with a minimum of five washing buffer changes before addition of secondary antibodies and DNA staining dyes such as 4',6-diamidino-2-phenylindole (DAPI) or DRAQ5. Biotinylated antibodies and secondary antibodies were supplied in PBST + 1% BSA or MOM diluent for 48 h at 4 °C followed by tissue washing and dehydration to 100% MeOH. Samples were stored at –20 °C in 100% MeOH until imaging.

Sample clearing and imaging. For imaging of whole-mount stained pancreata and subsequent 3D reconstruction, samples were subjected to clearing, hereby reducing light scattering. Optical clearing was performed by submerging samples in a 1:2 solution of benzyl alcohol:benzyl benzoate (BABB). Cleared samples were subsequently mounted in glass concavity slides and submerged completely in BABB to maintain refractive index matching and sample transparency. Cleared samples were imaged using a Leica SP8 confocal microscope with a 20×/0.75 oil immersion objective at 1024 × 1024 resolution. Samples were imaged in an 8-bit format unless otherwise indicated.

In vivo clonal analysis. 4-OH tamoxifen (4-OHTm, Sigma, H6278) was prepared as a 10 mg/mL stock solution in 10% ethanol and 90% corn oil and

subsequently diluted in vehicle solution (10% ethanol, 90% corn oil) to obtain the desired concentration. For E9.5 to E14.5 clonal analyses, mice carrying *Hnf1b*^{*CreER*}, *mT/mG*, *Ptf1a*^{*CreER*}, *mT/mG* and *Rosa26*^{*CreER*}, *mT/mG* embryos received a single intraperitoneal injection of 4-OHTm at E9.5, at a concentration of 11.5 µg/g, 57.5 µg/g and 1.35 µg/g, respectively. The dosage of 4-OHTm required to reach labelling at clonal density was initially determined by performing dose titration of E9.5 injections and analysis of clone density at E10.5 by whole-mount immunostaining. The temporal accuracy of labelling was tested using the *Ptf1a*^{*CreER*}, *mT/mG*. As PTF1A expression starts at E9.5, we injected 57.5 µg/g 4-OHTm at E7.5 or E8.5 and observed no labelled cell at E14.5 in 9 embryos analysed in total (4 embryos from E7.5 injection and 5 from E8.5 injection). Using ImarisTM software, GFP⁺ cells were identified in 3D reconstructed pancreata, and the fate of cells determined by immunostaining for various pancreatic lineage markers. For the E9.5–E10.5 short-term clonal analysis, GFP⁺ labelled cells were considered to be of clonal origin if one cell was seen or if the distance between recombined cells was less than 30 µm after the tracing period, based on the estimates of cell migratory capacity from Kim et al.¹⁹. For the mapping of clone position in the E10.5 bud, embryos harbouring one labelled cell or two labelled sister cells were considered for the analysis.

Single-cell qRT-PCR. E9.5 gut tube regions spanning the pancreatic bud and proximal duodenum were isolated from *Hes1*-eGFP embryos and stored in PBS on ice until all gut regions had been collected. Embryonic *mT/mG* tissue was added as a tissue spike-in to generate a bulk pellet mass preventing loss of the scarce GFP⁺ cell population. The pooled *Hes1*-eGFP gut tubes and *mT/mG* embryonic tissue was dissociated in 0.05% trypsin-EDTA (Gibco) containing 200 U DNase I (Roche) for 15 min at 37 °C with manual trituration using a p1000 pipette. Following dissociation, PBS + 10% FCS was added to inactivate the trypsinisation, and the single-cell suspension was centrifuged and re-suspended in PBS + 10% FCS followed by another round of centrifugation. The single-cell suspension was re-suspended in PBS + 10% FCS + DAPI to allow exclusion of DAPI⁺ dead cells. 260 single GFP⁺ cells were sorted into 96-well plates containing 5 µL CellsDirect 2× reaction mix (Invitrogen) and 0.05 U SUPERase-InTM RNase inhibitor (Thermo Fischer). 96-well plates containing single cells in CellsDirect were stored at –80 °C until ready to perform single-cell qRT-PCR reaction.

Prior to single-cell qRT-PCR, all primer pairs (Supplementary Table 2) had been validated on E14.5 bulk pancreatic cDNA using standard qRT-PCR. A mix containing forward and reverse primers for all 96 target genes were prepared in TE-buffer, generating a final concentration of 500 nM for each primer. mRNA from single cells was next subjected to one-step reverse transcription and specific target amplification according to the Fluidigm protocol 'One-Step Single-Cell Gene Expression Using EvaGreen® SuperMix on the BioMarkTM HD system'. Upon loading of 96 × 96 chips, a 5-fold standard series of E14.5 bulk cDNA was added to five chip inlets, allowing identification of specific gene product detection by comparison of melt profiles of single-cell amplifications and bulk reactions. Using Fluidigm Real-Time PCR analysis software, data from three independent chip runs were combined, and individual reactions were passed (203)/failed (57) according to software peak detection and melt peak temperature being in range with bulk reactions. Expression of housekeeping genes was used as inclusion criterion for downstream analysis of individual cells. Single-cell qRT-PCR data were subsequently analysed using Fluidigm SINGuLARTM Analysis Toolset, while global gene correlation tSNE-mediated dimensionality reduction was performed using a Python-based script (code available upon request). For the analysis of *Pdx1*⁺ pancreatic progenitors, cells were categorised as being positive for *Pdx1* if displaying CT values <20 for *Pdx1*.

Model implementation. We started simulating each clone from a cell in a progenitor state and when comparing with clonal data (Figs. 1h and 4f), we only included clones with at least one progenitor. The model thus underestimates the number of clones fully committed to endocrine cells, but it does not affect our results since we only look at the acinar fractions. The algorithm follows the steps below.

First, start with cell in a progenitor state. Second, draw a cell cycle length, t_{cc} , from a Gamma-distribution from¹⁹. To account for the unknown start of the cell cycle for the first cell, choose a random start between 0 and t_{cc} . Third, after time counter reaches t_{cc} the cell divides and adopts one of three possible fates according to the diagram in Fig. 5a: Progenitor probability $1 - c$, acinar fate with probability $c(1 - f)$ and endocrine fate otherwise. Fourth, assign two new cell-cycle lengths from the Gamma distribution. Fifth, repeat steps 3–4 for all cells.

For model 1: $f = q$, while for model 2: $f(q, t) = q + 0.5(1 - q)(1 + \cos(2\pi t))$. To estimate a probability density function (PDF) for a distribution of discrete datapoints we use KDE. In effect it is a smoothening step, where each data point is represented by a kernel (in our case gaussian with $\sigma = 6.5$ for data sets from Supplementary Fig. 7d and $\sigma = 0.015$ for combined data set Figs. 1h and 4f, also referred to as bandwidth)⁴⁴.

We find the likelihood, L , of an observation, x_i , being consistent with the model by evaluating the PDF at x_i , $PDF(x_i)$. The likelihood that all datapoints in a data set are consistent with the model is a product of their individual likelihoods. If there are two different data sets, their likelihoods are thus

$$L_1 = \prod_i PDF_1(x_i); L_2 = \prod_j PDF_2(x_j), \text{ and the likelihood of both datasets}$$

agreeing with the model is $L = L_1 L_2$.

The AIC is a method for selecting among models. It does not give an absolute estimate of how well each of the models fits the data but $AIC = 2k - \ln(L)$ where k is the number of variables and L is the maximum likelihood, i.e., corresponding to the optimal parameter set⁴⁵. The model with the lowest AIC, is the preferred model. The relative probability that an inferior model is as good as the preferred model can be calculated by use of the equation $p_i = \exp(AIC_{\min} - AIC_i)$.

Label retention experiments and image analysis. Pregnant mice carrying *Pdx1-tTA; tetO-H2B-GFP* were subjected to oral gavage of 200 μ L of 2 mg/mL doxycycline hydrochloride (Dox, Sigma), 3.5% vol/vol sucrose in H₂O at E9.5 or E11.5, and subsequently this solution replaced ad libitum water supply to maintain repression of H2B-GFP expression. Following whole-mount immunostaining, cleared samples were imaged at 12-bit depth and subjected to 3D reconstruction and downstream analysis in ImarisTM (Bitplane). Progenitor cells were identified by SOX9 immunoreactivity, and the pancreatic epithelium was masked based on SOX9 staining in order to exclude label-retaining endocrine cells from further analysis. The xyz position of SOX9⁺ progenitors was obtained using the ImarisTM spot detection algorithm on the SOX9-masked channel, additionally enabling extraction of mean GFP immunostaining intensity signal from the volume of the spot encompassing SOX9⁺ nuclei. Kernel density estimation of SOX9⁺ and the top 10% of GFP cells was applied to estimate the 2D distribution of these two cell populations. For the one-dimensional analysis of GFP retaining cell distribution, the distal-most point of the dorsal pancreatic epithelium and the centre of the dorsal pancreatic epithelium just proximal to the turning of the ductal structure connecting the dorsal pancreas to the ventral was used to extract the equation for the diagonal line running between these two points along the length axis of the dorsal pancreas. Using standard trigonometry, the xy-coordinates of SOX9⁺ progenitors and the xy-coordinates of the top 10% GFP-retaining SOX9⁺ cells were used to project these cells onto the diagonal line and ultimately to calculate the xy-coordinates of the intersection between the diagonal and projection line. Finally, the distance between the intersection point and the distal landmark was calculated, allowing kernel density estimation of SOX9⁺ cells and the top 10% GFP-retaining SOX9⁺ cells along this one-dimensional length axis.

For the analysis of spatial distribution of clones according to total clone size, clones from *Hnf1b^{CreER}*-mediated E9.5–E14.5 lineage tracing amenable to analysis were classified as small and large so that both groups contained an equal number of clones.

Quantification of endocrine precursor cell ratios. Quantification of the ratio of endocrine precursors, namely, SOX9, NEUROG3, PAX6 and PTF1a, progenitors and acinar precursors obtained at E9.5, E10.5, E11.5, E12.5 and E14.5 (Supplementary Fig. 7). At E9.5, cells were manually counted. At E10.5–E12.5, cell numbers were determined using ImarisTM spot detection. For the quantification of putative acinar progenitors at E11.5 and E12.5, PTF1A^{High} cells were quantified based on mean intensity of nuclear PTF1A above 80 grey scale values from 8-bit format images. This pixel intensity threshold was selected based on the intensity of PTF1A⁺ cells segregated to the periphery at E11.5 and E12.5, although PTF1A⁺ displaying mean intensity values above the threshold are still found scattered within the central epithelium. At E14.5, absolute cell numbers were determined using a custom built image segmentation and analysis software.

Neighbour identification by Voronoi-Delaunay triangulation. The xyz coordinates of E9.5 pancreatic progenitors were obtained after manual spot detection of SOX9⁺ nuclei in 3D reconstructed images of E9.5 gut tubes. The

mean fluorescence intensity of the applied staining for pancreatic transcription factors were extracted from the spot volume. Voronoi-Delaunay triangulation was implemented using a python-based script, and a 10 μ m distance threshold was applied in order to identify nearest neighbours with biological meaning. The coefficient of variation, as well as the mean intensity of neighbours was computed from corresponding fluorescence intensities of neighbour-connected cells, in order to visualise spatial patterns of heterogeneity and regionalisation of transcription factor expression levels.

Code availability. Python and Python-Notebook code files, along with an explanatory Read-me file, linked to Fig. 5 are provided as Supplementary Software, under the GNU General Public Licence (GPL). Codes are available upon request for label retention experiment quantifications and Voronoi-Delaunay triangulations.

Data availability. The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request.

Received: 13 September 2016 Accepted: 15 June 2017

Published online: 19 September 2017

References

- Clayton, E. et al. A single type of progenitor cell maintains normal epidermis. *Nature* **446**, 185–189 (2007).
- Doupe, D. P., Klein, A. M., Simons, B. D. & Jones, P. H. The ordered architecture of murine ear epidermis is maintained by progenitor cells with random fate. *Dev. Cell* **18**, 317–323 (2010).
- Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
- Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822–825 (2010).
- Itzkovitz, S., Blat, I. C., Jacks, T., Clevers, H. & van Oudenaarden, A. Optimality in the development of intestinal crypts. *Cell* **148**, 608–619 (2012).
- Gomes, F. L. et al. Reconstruction of rat retinal progenitor cell lineages in vitro reveals a surprising degree of stochasticity in cell fate decisions. *Development* **138**, 227–235 (2011).
- He, J. et al. How variable clones build an invariant retina. *Neuron* **75**, 786–798 (2012).
- Gao, P. et al. Deterministic progenitor behavior and unitary production of neurons in the neocortex. *Cell* **159**, 775–788 (2014).
- Jorgensen, M. C. et al. An illustrated review of early pancreas development in the mouse. *Endocr. Rev.* **28**, 685–705 (2007).
- Larsen, H. L. & Grapin-Botton, A. The molecular and morphogenetic basis of pancreas organogenesis. *Semin. Cell Dev. Biol.* **66**, 51–68 (2017).
- Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl Acad. Sci. USA* **97**, 1607–1611 (2000).
- Gu, G., Dubauskaite, J. & Melton, D. A. Direct evidence for the pancreatic lineage: NGN3⁺ cells are islet progenitors and are distinct from duct progenitors. *Development* **129**, 2447–2457 (2002).
- Kesavan, G. et al. Cdc42-mediated tubulogenesis controls cell specification. *Cell* **139**, 791–801 (2009).
- Villasenor, A., Chong, D. C., Henkemeyer, M. & Cleaver, O. Epithelial dynamics of pancreatic branching morphogenesis. *Development* **137**, 4295–4305 (2010).
- Zhou, Q. et al. A multipotent progenitor domain guides pancreatic organogenesis. *Dev. Cell* **13**, 103–114 (2007).
- Pan, F. C. et al. Spatiotemporal patterns of multipotentiality in Ptf1a-expressing cells during pancreas organogenesis and injury-induced facultative restoration. *Development* **140**, 751–764 (2013).
- Solar, M. et al. Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. *Dev. Cell* **17**, 849–860 (2009).
- Kopinke, D. et al. Lineage tracing reveals the dynamic contribution of Hes1⁺ cells to the developing and adult pancreas. *Development* **138**, 431–441 (2011).
- Kim, Y. H. et al. Cell cycle-dependent differentiation dynamics balances growth and endocrine differentiation in the pancreas. *PLoS Biol.* **13**, e1002111 (2015).
- Gouzi, M., Kim, Y. H., Katsumoto, K., Johansson, K. & Grapin-Botton, A. Neurogenin3 initiates stepwise delamination of differentiating endocrine cells during pancreas development. *Dev. Dyn.* **240**, 589–604 (2011).

21. Kopp, J. L. et al. Sox9 + ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. *Development* **138**, 653–665 (2011).
22. Desgraz, R. & Herrera, P. L. Pancreatic neurogenin 3-expressing cells are unipotent islet precursors. *Development* **136**, 3567–3574 (2009).
23. Ventura, A. et al. Restoration of p53 function leads to tumour regression in vivo. *Nature* **445**, 661–665 (2007).
24. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
25. Hayashi, S. & McMahon, A. P. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev. Biol.* **244**, 305–318 (2002).
26. Chintinne, M. et al. Contribution of postnatally formed small beta cell aggregates to functional beta cell mass in adult rat pancreas. *Diabetologia* **53**, 2380–2388 (2010).
27. Schaffer, A. E., Freude, K. K., Nelson, S. B. & Sander, M. Nkx6 transcription factors and Ptf1a function as antagonistic lineage determinants in multipotent pancreatic progenitors. *Dev. Cell* **18**, 1022–1029 (2010).
28. Maestro, M. A. et al. Hnf6 and Tcf2 (MODY5) are linked in a gene network operating in a precursor cell domain of the embryonic pancreas. *Hum. Mol. Genet.* **12**, 3307–3314 (2003).
29. Bechard, M. E. et al. Precommitment low-level Neurog3 expression defines a long-lived mitotic endocrine-biased progenitor pool that drives production of endocrine-committed cells. *Genes Dev.* **30**, 1852–1865 (2016).
30. Bankaitis, E. D., Bechard, M. E. & Wright, C. V. Feedback control of growth, differentiation, and morphogenesis of pancreatic endocrine progenitors in an epithelial plexus niche. *Genes Dev.* **29**, 2203–2216 (2015).
31. Villasenor, A., Chong, D. C. & Cleaver, O. Biphasic Ngn3 expression in the developing pancreas. *Dev. Dyn.* **237**, 3270–3279 (2008).
32. Brennand, K., Huangfu, D. & Melton, D. All beta cells contribute equally to islet growth and maintenance. *PLoS Biol.* **5**, e163 (2007).
33. Krapp, A. et al. The bHLH protein PTF1-p48 is essential for the formation of the exocrine and the correct spatial organization of the endocrine pancreas. *Genes Dev.* **12**, 3752–3763 (1998).
34. Kawaguchi, Y. et al. The role of the transcriptional regulator Ptf1a in converting intestinal to pancreatic progenitors. *Nat. Genet.* **32**, 128–134 (2002).
35. Apelqvist, A. et al. Notch signalling controls pancreatic cell differentiation. *Nature* **400**, 877–881 (1999).
36. Murtaugh, L. C., Stanger, B. Z., Kwan, K. M. & Melton, D. A. Notch signaling controls multiple steps of pancreatic differentiation. *Proc. Natl Acad. Sci. USA* **100**, 14920–14925 (2003).
37. Fujikura, J. et al. Notch/Rbp-j signaling prevents premature endocrine and ductal cell differentiation in the pancreas. *Cell Metab.* **3**, 59–65 (2006).
38. Horn, S. et al. Mind bomb 1 is required for pancreatic beta-cell formation. *Proc. Natl Acad. Sci. USA* **109**, 7356–7361 (2012).
39. Shih, H. P., Panlasigui, D., Cirulli, V. & Sander, M. ECM signaling regulates collective cellular dynamics to control pancreas branching morphogenesis. *Cell Rep.* **14**, 169–179 (2016).
40. Lescroart, F. et al. Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development. *Nat. Cell Biol.* **16**, 829–840 (2014).
41. Holland, A. M., Hale, M. A., Kagami, H., Hammer, R. E. & MacDonald, R. J. Experimental control of pancreatic development and maintenance. *Proc. Natl Acad. Sci. USA* **99**, 12236–12241 (2002).
42. Tumber, T. et al. Defining the epithelial stem cell niche in skin. *Science* **303**, 359–363 (2004).
43. Klinck, R. et al. A BAC transgenic Hes1-EGFP reporter reveals novel expression domains in mouse embryos. *Gene Exp. Patterns* **11**, 415–426 (2011).
44. Bashtannyk, D. M. & Hyndman, R. J. Bandwidth selection for kernel conditional density estimation. *Comput. Stat. Data Anal.* **36**, 279–298 (2001).
45. Burnham, K. & Anderson, D. *Model Selection and Multimodel Inference* (Springer, 2002).

Acknowledgements

We thank Pau Rué, University of Cambridge, UK, for his initial guidance on modelling, Alfonso Martinez Arias, University of Cambridge, UK, for enabling the mathematical modelling training, Jutta Bulkescher for imaging guidance and Gopal Karemore for automated image segmentation. This project was supported by the Novo Nordisk Foundation and grant 12–126875 from Det Frie Forskningsråd-Sundhed og Sygdom to A.G.B. and the Danish National Research Foundation/Danmarks Grundforskningsfond, grant DNRF 116 to A.T. and A.G.B.

Author contributions

H.L.L., A.G.-B. and Y.H.K. designed the project. H.L.L. performed and analysed most experiments except those specified below. L.M.-C. performed and analysed the experiments leading to Fig. 1. A.V.N., guided by A.T. performed the *in silico* modelling leading to Fig. 5 and Supplementary Figs. 8–10. Y.H.K. guided experiments and contributed to single-cell qRT-PCR. C.V.W. provided the *PTF1a^{CreER}* mice. H.L.L. and A.G.-B. wrote the manuscript which was commented and approved by L.M.-C., Y.H.K., A.V.N., A.T. and C.V.W.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00258-4.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017